Discovering the Genetic Basis of Epilepsies through Computational Analysis of Clinical Phenotype

David Lewis-Smith

Doctor of Philosophy

Translational and Clinical Research Institute

Faculty of Medical Sciences

Newcastle University

March 2024

This page has been left blank intentionally.

Thesis Restriction

Chapter 5 of this thesis relates to my analysis of data belonging to the Epi25 (http://epi-25.org) as a secondary project approved by its committees. The work presented in this thesis is my own and for the benefit of this examination, I have avoided obtaining direct contributions towards it from the wider Epi25.

The Epi25 has a formal publication policy. The approval of its Publication Committee will be required before publication of work presenting Epi25 data. I will be continuing to prepare this work for submission to the Publication Committee and submission to a peer-reviewed journal in the coming months. This chapter should be embargoed until my manuscript has been approved for publication by the Epi25 Publication Committee.

This page has been left blank intentionally.

Abstract

The epilepsies are complex, primarily because of their diverse seizure types, neurodevelopmental features, and comorbidities. High throughput sequencing and genotyping has accelerated the pace of discovery for new genetic causes and risk factors. The rate at which these biological discoveries have been translated into clinically actionable knowledge has been slower because of the burden of collecting, harmonizing, and analyzing precise clinical data manually, which prohibits detailed phenotypic analysis at scale.

In this thesis, I report some of my contribution to the opening of this phenotyping bottleneck by modeling clinical neurological reasoning for computational scaling of phenotypic analysis to increase the clinical value of genetic discoveries.

I present a conceptual model for harmonizing seizure descriptions based on contemporary classifications. I critically appraise methods of associating clinical and genetic features in epileptology and propose a new phenotypic similarity measure, $\mathrm{Sim}_{\mathrm{MinTO}}$, based on the communication of clinical features between experts that models the traditional notion of recognizing a genetically associated syndrome.

Applying these tools in two studies of over 10,000 people, I characterize the clinical features of carriers of copy number variants and ultra-rare sequence variants with reference to other people with presumed genetic epilepsies, reflecting the situation in an epilepsy clinic. Finally, I present a detailed analysis of data from 67 adults with CDKL5 Deficiency Disorder, a developmental and epileptic encephalopathy in which prior knowledge has been limited to children. I show how harmonization of historical multisource data can enable retrospective natural history studies of rare disorders to inform prognostication and the design of expensive and lengthy prospective studies including precision medicine trials.

I hope that this work will introduce new standards, promoting more formal clinical characterization of complex disorders that will facilitate the rapid discovery and validation of associations, improving classification for stratified care and research both within neurology and beyond.

This page is left blank intentionally.

This work is dedicated to those people with epilepsy that I have cared for and learned from, and to those people with epilepsy that I will be privileged to care for and work with over the years to come.

Acknowledgements

I thank my primary supervisor Dr Rhys H. Thomas for encouraging me to embark on research into epilepsy to answer the questions that would arise in our clinics. I have been fortunate to be able to learn so much from him about the care of adults with epilepsy and neurodevelopmental disorders, the skills of collaboration and communication, and the attitudes necessary to ask the right research questions, and to instigate and deliver positive change for people with epilepsy.

I thank Dr Helbig for his supervision and external mentorship, and for welcoming me to join his laboratory and visit his clinics at Children's Hospital of Philadelphia. Dr Helbig pioneered formal computational analysis of clinical epilepsy data using the Human Phenotype Ontology (HPO). Now while advising me, he has allowed me to lead the collaboration revising the epilepsy domains of the HPO, develop and apply new methods in our projects within and beyond those described here, and to lead the Epi25 Collaborative phenotypic analyses.

Together Dr Thomas and Dr Helbig have provided professional, academic, and personal advice and encouragement from complementary perspectives over five busy years. They have facilitated my contributions to International League Against Epilepsy (ILAE) task forces, the HPO, the ILAE Consortium on Complex Epilepsies (genome-wide association studies), and the NIH Clinical Genome Resource (ClinGen) expert panels on epilepsy genes and sodium channel variants, all of which I continue to contribute to and learn from.

Dr Helbig has introduced me to his knowledgeable, generous, and hardworking colleagues. In particular, I thank Mr Peter D. Galer for guiding me through my first months of coding in R and familiarizing me with the techniques already established in the Helbig laboratory, and Mr Shiva Ganesan for assisting with genetic bioinformatics for my work with the Epi25 Collaborative. I also thank Ms Julie Xian, Mr Shridhar Parthasarathy, and Mr Michael C. Kaufman, for their collaboration, discussions, and assistance with the many figures we have prepared over the past few years.

I remain grateful to the Guarantors of Brain for generously awarded me their Entry Fellowship. This provided time to develop the ideas that led to this work, in particular their funding allowed me to attend the European Molecular Biology Organization workshop on 'Phenotyping neurological syndromes for systems genetics', arranged by Dr Roland Krause (Université du Luxembourg), at which discussions with Dr Krause and Prof. Peter N. Robinson (The Jackson Laboratory, Connecticut) inspired me to focus my work on formal phenotypic analysis using the HPO.

I thank Prof. Heather J. Cordell (Newcastle University) and Prof. Marcus Kaiser (University of Nottingham, formerly of Newcastle University) for their support, teaching, and advice as

secondary supervisors in the early stages of the fellowship. Furthermore, I remain grateful to Prof. James B. Rowe (University of Cambridge) for igniting my academic interest in neurological phenotypes and to Prof. Patrick F. Chinnery (University of Cambridge, formerly of Newcastle University) for introducing me to neurogenetics.

I have been fortunate to collaborate and develop valued friendships with many researchers across a large number of countries while undertaking this work.

First, in the revision of the HPO seizure domain, I am particularly grateful to Dr Ganna Balagura (Giannina Gaslini Institute, Genoa) and Dr Hugh Kearney (Royal College of Surgeons in Ireland), who helped with the drafting of over 300 new HPO terms, and shared their intellectual and clinical perspectives when discussing the literature of seizure classifications and how we might represent these classifications ontologically. I thank Dr Helbig and Dr Thomas for providing advice from expert paediatric and adult epileptological perspectives when required, and Prof. Robinson for his enthusiasm for our work, his technical advice, and his personal effort to incorporate the revisions we proposed on his weekends.

Second, my work relating to the Epi25 Collaborative builds upon the remarkable cooperative efforts of a large international team of epilepsy researchers, clinicians, and their teams, under the inspiring leadership of Prof. Daniel H. Lowenstein (University of California, San Francisco), Prof. David B. Goldstein (Columbia University), Prof. Holger Lerche (University of Tübingen), and Prof. Samuel F. Berkovic (University of Melbourne). I am grateful for the willingness of the Epi25 Collaborators to share their data and encourage my work, and for allowing me to join their strategy and phenotyping committees. I appreciate the efforts of the teams led by Dr Krause and Prof. Berkovic to collate and curate the raw phenotypic data and I thank the National Human Genome Research Institute and Centers for Common Disease Genomics (USA) for funding the exome sequencing of Epi25 participants at the Broad Institute. Without these considerable efforts, my analyses would not have been possible. With regards to contributions to the Epi25-related chapters of this thesis, I thank Dr Ludovica Montanucci and her colleagues under the supervision of Dr Costin Leu and Prof. Dennis Lal (University of Texas) and their collaborators led by Dr Ryan Collins at Prof. Michael Talkowski's research group (Broad Institute), who identified which Epi25 participants carry copy number variants and recognised the potential value of my contribution to their study. I am indebted to Dr Siwei Chen and Dr Benjamin Neale (Broad Institute) who provided the exome data from Epi25 participants for secondary analyses, and to Mr Ganesan for undertaking the large-scale preliminary annotation and filtering according to my criteria to provide me with a manageable dataset.

Third, the study of 67 adults with CDKL5 Deficiency Disorder would not have been possible without Dr Ángel Aledo-Serrano (Synaptia Madrid Epilepsy Center), Prof. Rikke S. Møller (University of Southern Denmark and Danish Epilepsy Centre), and their international network of clinical collaborators. Dr Aledo-Serrano collated the data in a REDCap database under my guidance and generously allowed me to lead the analyses using my new techniques.

Fourth, there have been a large number of epilepsy professionals and researchers beyond these, from within and beyond the UK, whom I have been honoured to work with or even just meet during this period of training. While they may not have contributed directly to this thesis, many have inspired and taught me, and made me feel valued. I look forward to future projects and collaborative successes in this generous, talented, and hardworking international environment.

I thank the Wellcome Trust for funding my work and training from 2019–2022 through the 4ward North Clinical PhD programme. I am grateful to the leaders of the ILAE and its branches for their support and opportunities to present my work and to develop leadership skills, and to the charities and patient groups that have invited me to speak at their congresses. I thank Dr Paula Rocktäschel (University of Oxford), Dr Gashirai Mbizvo (University of Liverpool), Dr Konrad Wagstyl (University College London), and Dr Stuart Smith (Kings College London). Together we launched the Young Epilepsy Section of the British Branch of the ILAE in 2019. They provided valuable peer support during the challenges of the Covid-19 pandemic.

Most importantly for enabling this work, I thank those thousands of people living with epilepsy who have shared personal details of their disease and genetics for research. I hope that this research, and what might in future develop from it, contributes to a future in which the burden of epilepsy might be eliminated or at least reduced for those living with the *Sacred Disease*.

Finally, I will forever be grateful to my family. Lizzy, Dan, and Anna, who have tolerated and supported me remarkably consistently over five years of long hours spanning time zones, and long periods of absence for international collaboration and conferences. I look forward to giving them back the time and attention they are owed, more than anything else. Similarly, I am grateful for the encouragement and sacrifices of my parents, Paul and Uca, and my sister Helena. I thank my wider family and friends, in particular my mother-in-law, Margie, for teaching me the unwritten curriculum of epileptology and sharing the proof-reading of this work with my father, and Ruth for her support and help alleviating the effects of my work on Dan, Anna, and Lizzy.

List of abbreviations

Abbreviation	Meaning
ACMG	American College of Medical Genetics and Genomics
ACTH	Adrenocorticotropin
AMP	Association for Molecular Pathology
ASM	Antiseizure medication
BP	Chromosome breakpoint
BPAN	Beta-propeller protein associated neurodegeneration
bMissense	Benign missense
CDE	Common data element
CDKL5	Cyclin-dependent kinase-like 5
CDS	CDKL5 Developmental Score
CHOP	Children's Hospital of Philadelphia
CNV	Copy number variant
CUI	Concept unique identifier
DEE	Developmental and epileptic encephalopathy
dMissense	Deleterious missense
DNA	Deoxyribonucleic acid
dPAV	Damaging protein altering variant
EMBO	European Molecular Biology Organization
EEG	Electroencephalogram
EE-SWAS	Epileptic encephalopathy with spike-and-wave activation in sleep
EMR	Electronic medical records
EPGP	Epilepsy Phenome/Genome Project
Epi25	Epi25 Collaborative
FDR	False discovery rate
FET	Fisher's exact test
GABA	Gamma-aminobutyric acid
GE	Genomics England
GGE	Genetic generalized epilepsy
GA4GH	Global Alliance for Genomics and Health
gnomAD	Genome Aggregation Database
GO	Gene Ontology
GWAS	Genome-wide association study
HPO	Human Phenotype Ontology
HPO 2017	HPO release 2017-12-12
HPO 2020	HPO release 2020-12-07
IC	Information content
ID	Intellectual disability
ILAE	International League Against Epilepsy
ILAE 2017	ILAE Classification of seizures 2017
IRB	Institutional review board
MICA	Most informative common ancestor

Abbreviation	Meaning
Mis-*z*	Missense *z*-score
MRI	Magnetic resonance imaging
NIH	National Institutes of Health
NHGRI	National Human Genome Research Institute
NHS	National Health Service
OMIM	Online Mendelian Inheritance in Man
OR	Odds ratio
ORDO	Orphanet Rare Disease Ontology
PAK1	p21-activated kinase 1
PAV	Protein altering variant
pHaplo	Probability of haploinsufficiency
pLI	Probability of being loss-of-function intolerant
possSplicing	Possible splicing
PPV	Positive predictive value
pTriplo	Probability of triplosensitivity
PTV	Protein truncating variant
PCA	Principal component analysis
pLI	Probability of being loss of function intolerant
RR	Relative risk
SFARI	Simons Foundation Autism Research Initiative
SNARE	Sensitive factor attachment protein receptor
SNOMED-CT	Systematized Nomenclature of Medicine – Clinical Terms
SNP	Single nucleotide polymorphism
UMLS	Unified Medical Language System
URV	Ultra-rare genetic (sequence) variant
VNS	Vagus nerve stimulation
VEP	Variant Effect Predictor
WRST	Wilcoxon rank sum test with continuity correction
95 % CI	95% confidence interval

I have written the body of this thesis using American English for consistency with the Human Phenotype Ontology in its standard form.

Table of Contents

Thesis Restriction ..... *i*

Abstract ..... *iii*

Acknowledgements ..... *vi*

List of abbreviations ..... *ix*

List of figures ..... *xvii*

List of tables ..... *xxii*

Chapter 1. An introduction to epilepsy, its genetics, and phenotypic analysis *1*

1.1 Summary **1**

1.2 The definitions and burden of epilepsy and seizures **2**

1.3 The classification of epilepsy and seizures **3**

1.4 The role of genetics in explaining why someone has epilepsy and stratifying their care **6**

1.5 The traditional role of phenotype in genetic studies of epilepsy **11**

1.5.1 Phenotypic considerations in gene discovery studies 11

1.5.2 Genotype-phenotype association 14

Chapter 2. Modeling seizures in the Human Phenotype Ontology according to contemporary ILAE concepts makes big phenotypic data tractable *16*

2.1 Preface **16**

2.1.1 Abstract 17

2.1.2 Objective 17

2.1.3 Methods 17

2.1.4 Results 17

2.1.5 Significance 17

2.2 Introduction **18**

2.3 Methods **21**

2.3.1 Cohorts providing seizure data 21

2.3.2 Harmonization of data using the HPO 21

2.3.3 Statistical procedures 22

2.3.4 The process of revising the HPO seizure subontology 22

2.4 Results **22**

2.4.1 HPO-based inference harmonizes seizure descriptions across different levels of precision, semiologies and classifications 22
2.4.2 Creation of a new subontology of contemporary seizure concepts for the HPO 24
2.4.3 The 2020 HPO seizure subontology captures more seizure information than the 2017 version . 27

2.5 Discussion 31

Chapter 3. Computational analysis of neurodevelopmental phenotypes — harmonization empowers clinical discovery 34

3.1 Preface 34

3.2 Abstract 34

3.3 Introduction 35

3.4 Biomedical ontologies capture the relationship between clinical annotations 38

3.5 Inference allows for harmonization across the depth of all phenotypes 39

3.6 Clinical phenotypic annotations translated from detailed raw data are sparse 42

3.7 Domain-specific biomedical ontologies can be improved through expert curation 43

3.8 The challenge of representing important absent phenotypes 44

3.9 Longitudinal phenotype data captures age-specific clinical features 49

3.10 Common Data Elements can be combined with biomedical ontology data 52

3.11 Digitization of rare disease cohorts creates phenotypic references 54

3.12 Distinct genetic etiologies have recognizable phenotypic constellations 55

3.13 Can computational assessment of phenotypic features aid variant interpretation? 57

Chapter 4. Interpretation of genetically stratified phenotypic data 59

4.1 Preface 59

4.2 Introduction 59

4.3 The case for an open-world assumption 59

4.4 Defining sets of annotations for phenotypic analysis 65

4.5 Association of genetic factors with phenotypes 67

4.6 Desirable features of a similarity measure for large epilepsy studies 71

4.7 Edge-based similarity measures 74

4.8 Set-based measures of similarity 76

4.8.1 Comparison according to the terms in each person's set of propagated annotations....76
4.8.2 Similarity based on the probability of encountering shared combinations of terms....82
4.9 Semantic similarity....84
4.9.1 Comparison of pairs of terms....84
4.9.2 Summarizing the similarity of a pair of people after pairwise comparison of their annotations..92
4.10 Measures used to prioritize candidate diagnoses....97
4.11 Refinement of annotations for semantic similarity analysis....101
4.12 Introducing minimal term overlap similarity: a node set-based similarity measure modeled on clinical communication....106
4.13 Phenotype first and genotype first approaches to similarity analysis....113
4.14 Interpretation of pairwise similarities within groups of people....114
4.15 Phenotypic similarity measures to be prioritized for an epilepsy study....121
Chapter 5. The phenotypic consequences of ultra-rare variants....130
5.1 Preface....130
5.2 Abstract....130
5.3 Introduction....131
5.4 Materials and methods....133
5.4.1 Participant recruitment and data collection....133
5.4.2 Informed consent....133
5.4.3 Phenotypic data collection and quality control....134
5.4.4 Code and statistics....134
5.4.5 Harmonization of categorical phenotypic data....134
5.4.6 Harmonization of quantitative phenotypic data....136
5.4.7 Harmonization of free-text phenotypic data....137
5.4.8 Whole exome sequencing data generation and quality control....138
5.4.9 Filtering of ultra-rare variants....140
5.4.10 Variant classification....142
5.4.11 Assigning gene-level classes by participant....144
5.4.12 Cohort definition....147
5.4.13 Genotype-phenotype analysis....147
5.4.14 Data availability....149
5.5 Results....149
5.5.1 Description of the study cohort....149
5.5.2 Clinical data harmonization....150

5.5.3 Ultra-rare variants 154
5.5.4 Phenotypic similarity of the cohort 157
5.5.5 Phenotypic similarity of dominant epilepsy genes 158
5.5.6 Phenotypic similarity beyond dominant epilepsy genes 161
5.5.7 Phenotypes associated with dominant epilepsy genes 165
5.5.8 Phenotypic similarity and associations of clinically validated recessive genes 204
5.5.9 Phenotypic similarity and associations of clinically validated neurodevelopmental genes 215
5.5.10 Phenotypic similarity and associations of genes not yet clinically validated in epilepsy or neurodevelopmental disorders 244

5.6 Discussion 265

Chapter 6. Genome-wide identification and phenotypic characterization of seizure-associated copy number variations in 741,075 individuals 270

6.1 Preface 270

6.2 Abstract 271

6.3 Introduction 271

6.4 Methods 273

6.4.1 Study cohorts 273
6.4.2 Individuals with clinically defined epilepsy - Epi25 Collaborative 273
6.4.3 CNV calling and quality control - Epi25 Collaborative 274
6.4.4 Individuals with seizures or neuropsychiatric phenotypes 275
6.4.5 Quality control - neuropsychiatric disorders cohort 275
6.4.6 Genome-wide association analysis 276
6.4.7 Meta-analysis and fine-mapping 276
6.4.8 Detailed HPO characterization of Epi25 participants 277
6.4.9 Phenome-wide association analysis of CNVs 278
6.4.10 Phenotypic similarity analysis 279
6.4.11 Data availability 280
6.4.12 Code availability 280

6.5 Results 281

6.5.1 Discovery of 25 genome-wide significant seizure-associated CNVs regions 281
6.5.2 Fine-mapping and candidate genes 282
6.5.3 Most of the 25 identified risk CNVs are pleiotropic 288
6.5.4 Characterization of the clinical phenotypes enriched in the carriers of each seizure-associated CNV in epilepsy patients with deep phenotypes 289
6.5.5 Syndromic similarity among carriers of the same CNV 292

6.6 Discussion 294

Chapter 7. The natural history of CDKL5 Deficiency Disorder into adulthood.....316

7.1 Preface.....316

7.2 Abstract.....316

7.3 Introduction.....317

7.4 Methods.....318

7.4.1 Participant recruitment and data collection.....318

7.4.2 Protocol approval and consent.....319

7.4.3 Mapping of phenotypic data.....319

7.4.4 Negated phenotypes.....320

7.4.5 Clinical outcomes.....321

7.4.6 Genetic stratification.....322

7.4.7 Statistical analysis.....323

7.5 Results.....324

7.5.1 Description of the study cohort.....324

7.5.2 Phenotypic annotation.....327

7.5.3 Seizures.....331

7.5.4 Electroencephalography.....335

7.5.5 Antiseizure treatment.....338

7.5.6 Neurodevelopmental, psychiatric, and other neurological features.....342

7.5.7 Neurodevelopmental evolution.....343

7.5.8 Comorbidities.....350

7.5.9 Neuroimaging.....350

7.5.10 Mortality.....351

7.5.11 Molecular diagnosis.....351

7.5.12 Genotype-phenotype associations.....353

7.5.13 Genotype-based phenotypic similarity analysis.....358

7.5.14 Sex-phenotype associations.....358

7.5.15 Phenotypic associations of neonatal seizures.....360

7.5.16 Predicting total maximum CDS.....364

7.6 Discussion.....368

Chapter 8. Epilogue.....372

Bibliography.....376

Appendix A. Professional accomplishments within the Wellcome Trust Clinical PhD Fellowship and preceding Guarantors of Brain Entry Fellowship.....434

8.1 Relevant publications.....434

8.1.1 Primary or senior author peer-reviewed publications 434
8.1.2 Co-authorship of peer-reviewed publications 435
8.1.3 Commissioned publications 438

8.2 Conference abstracts 438

8.2.1 Invited platform presentations 438
8.2.2 Other platform presentations 439
8.2.3 Poster presentations 439

8.3 Prizes and awards 441

8.4 Professional roles within epileptology 441

8.4.1 Membership of professional bodies and qualifications 441
8.4.2 Peer review 442
8.4.3 Grant review 442

Appendix B. Collaborators providing data contributing to this work 443

8.5 Members of the Epi25 Collaborative 443

8.5.1 Columbia University Institute for Genomic Medicine analysis group 443
8.5.2 Epi25 sequencing, analysis, project management, and browser development 443
8.5.3 Epi25 executive committee 443
8.5.4 Epi25 strategy, phenotyping, analysis, informatics, and project management committees 443
8.5.5 Authors from individual Epi25 cohorts: 443

8.6 Collaborators contributing data about adults with CDKL5 Deficiency Disorder 456

List of figures

Figure 2-1. Harmonization of clinical annotations using the HPO. 20

Figure 2-2. Increasing the number of concepts annotated to allow comparison of cohorts. 23

Figure 2-3. Comparison of the phenotypic information encoded from the same data from 791 individuals using the two seizure subontologies. 31

Figure 3-1. A visual representation of the complexity of HPO version 1.7.13 released 2021-10-10, focusing on seizure and related neurological phenotypes. 39

Figure 3-2. Interpretation using the *is_a* relationships of the HPO. 41

Figure 3-3. The proportion of individuals with *SCN2A*-related disorders with particular phenotypes using data from Crawford *et al.* (2021). 47

Figure 3-4. The longitudinal interrogation of HPO annotations from the EMR of patients with genetic epilepsies. 51

Figure 3-5. The distribution of seizure frequencies and prescription of various antiseizure treatments with age. 53

Figure 3-6. An example of phenotypic similarity analysis using the $\mathrm{Sim}_{\mathrm{Resnik\text{-}BMS}}$ algorithm. 55

Figure 3-7. HPO-based visualizations demonstrate the clinical features associated with de novo variants in *SCN1A* in published cohorts with developmental and epileptic encephalopathies. 57

Figure 4-1. Comparison of forms of adjustment of the *p*-value from Fisher's exact test. 70

Figure 4-2. In isolation, a term may be less informative than the sum of its parents, those most specific terms it implies. 109

Figure 4-3. Using $\mathrm{Sim}_{\mathrm{MinTO}}$ with monotonic adjustment of IC. 111

Figure 4-4. The distribution of pairwise similarities varies by similarity measure. 117

Figure 4-5. The effect of group size on the distribution of the median pairwise similarity of groups of participants. 119

Figure 4-6. Spearman's correlation between 27 similarity measures in Epi25. 123

Figure 4-7. Hierarchical cluster analysis of 27 similarity measures in Epi25. 124

Figure 4-8. Spearman's correlation between 26 similarity measures in a study of CDKL5 Deficiency Disorder. 125

Figure 4-9. Hierarchical cluster analysis of 26 similarity measures in cohort with CDKL5 Deficiency Disotder. 126

Figure 5-1. The intersections of 1,473 clinically validated genes by virtual panel. 141

Figure 5-2. The intersections of sets of genes that have not been clinically validated. 141

Figure 5-3. The relationship between variant classes. 144

Figure 5-4. The geographic distribution of the Epi25 cohort. 149

Figure 5-5. The number and informativeness of annotations per person. 151

Figure 5-6. The phenotypic repertoire of the cohort. 153

Figure 5-7. The distribution of HPO term frequency in Epi25, stratified by source of annotation with HPO terms ranked by frequency. 154

Figure 5-8. Quantile-quantile plots of the similarity of carriers of URVs in dominant epilepsy genes according to $\mathrm{Sim}_{\mathrm{MinTO}}$. 159

Figure 5-9. Quantile-quantile plots of the similarity of carriers of URVs in dominant epilepsy genes according to $\mathrm{Sim}_{\mathrm{Goodall3}}$. 160

Figure 5-10. Quantile-quantile plots of the similarity of carriers of URVs in the global set of genes of the study according to $\mathrm{Sim}_{\mathrm{MinTO}}$. 163

Figure 5-11. Quantile-quantile plots of the similarity of carriers of URVs in the global set of genes of the study according to $\mathrm{Sim}_{\mathrm{Goodall3}}$. 164

Figure 5-12. The associations of carriers of PAVs in *SCN1A*. 172

Figure 5-13. The associations of carriers of PAVs in *SCN1A* shown conceptually. 173

Figure 5-14. The associations of carriers of PAVs in *SLC6A1* 174

Figure 5-15. The associations of carriers of PAVs in *SLC6A1* shown conceptually. 175

Figure 5-16.The associations of carriers of PAVs in *RORA*. 176

Figure 5-17. The associations of carriers of PAVs in *RORA* shown conceptually. 177

Figure 5-18.The associations of carriers of PAVs in *CHD2* 178

Figure 5-19. The associations of carriers of PAVs in *CHD2* shown conceptually. 179

Figure 5-20 The associations of carriers of PAVs in *STX1B*. 180

Figure 5-21 The associations of carriers of PAVs in *STX1B* shown conceptually. 181

Figure 5-22. The associations of carriers of PTVs in *WDR45*. 186

Figure 5-23. The associations of carriers of PTVs in *WDR45* shown conceptually. 187
Figure 5-24. The associations of carriers of PTVs in *NEXMIF*. 188
Figure 5-25. The associations of carriers of PTVs in *NEXMIF* shown conceptually. 189
Figure 5-26. The associations of carriers of PTVs in *DEPDC5*.....190
Figure 5-27. The associations of carriers of PTVs in *DEPDC5* shown conceptually.....191
Figure 5-28. The associations of carriers of PTVs in *POLG*.....211
Figure 5-29. The associations of carriers of PAVs in *POLG* shown conceptually. 212
Figure 5-30. The associations of carriers of PAVs in *PIGG*. 213
Figure 5-31. The associations of carriers of PAVs in *PIGG* shown conceptually. 214
Figure 5-32. The associations of carriers of dPAVs in *GMMPA*. 240
Figure 5-33. The associations of carriers of dPAVs in *GMMPA* shown conceptually. 241
Figure 5-34. The associations of carriers of PAVs in *LZTR1*. 242
Figure 5-35. The associations of carriers of PAVs in *LZTR1* shown conceptually.....243
Figure 5-36. The associations of carriers of PAVs in *PAK1IP1*. 253
Figure 5-37. The associations of carriers of PAVs in *PAK1IP1* shown conceptually. 254
Figure 5-38. The associations of carriers of PAVs in *KANSL1L*. 255
Figure 5-39. The associations of carriers of PAVs in *KANSL1L* shown conceptually. 256
Figure 5-40. The associations of carriers of PAVs in *UTP14A*.....257
Figure 5-41. The associations of carriers of PAVs in *UTP14A* shown conceptually.....258
Figure 5-42. The associations of carriers of dPAVs in *SESTD1*. 259
Figure 5-43. The associations of carriers of dPAVs in *SESTD1* shown conceptually. 260
Figure 6-1. Genome-wide meta-analysis identifies 25 CNVs associated with seizure disorders. 282
Figure 6-2. Regional CNV distribution plot and fine-mapping of the 15q11-q13 duplication. 283
Figure 6-3. Associations of deletions at 1p36.33 [0.91-1.51 Mb] and [2.02–2.49 Mb].....303
Figure 6-4. Associations of deletions at 15q12–q13.1 and duplications at 15q11.2-q13.3....304
Figure 6-5. Associations of deletions at 2p21-p16.3 and 15q11.2. 305

Figure 6-6. Associations of deletions at 16p13.11 and 15q13.2-q13.3. 306

Figure 6-7. Associations of deletions at 22q11.21 and duplications at 16p11.2. 307

Figure 6-8. Associations of duplications at 9q34.3 and deletions at 15q13.3. 308

Figure 6-9. Associations of deletions at 22q11.21 [19.67-19.96 Mb] and [20.65-21.54 Mb]. 309

Figure 6-10. Associations of duplication at 22q11.21 [20.02-21.54 Mb] and 17q12. 310

Figure 6-11. Broad clinical profiles of carriers of carriers of CNVs. 311

Figure 7-1. Age at last follow up (A) and *CDKL5* variants (B) in the cohort. 325

Figure 7-2. The geographic distribution of the cohort. 325

Figure 7-3. Visualization of phenotypic data harmonization. 329

Figure 7-4. The phenotypic landscape of the cohort, including the phenotyping gap. 330

Figure 7-5. The distribution of age of onset of epileptic seizures of any type. 332

Figure 7-6. The evolution of seizures by participant with CDD. 333

Figure 7-7. (A) The evolution of seizure types by onset, and (B) 6-year intervals containing 6 months of seizure freedom. 334

Figure 7-8. The evolution of seizures in the cohort by seizure type. 336

Figure 7-9. (A) The distribution of total maximum CDS across the entire cohort and how this compares to the most recent total CDS per person and (B) the percentage of the cohort attaining and retaining each constituent skill. 344

Figure 7-10. The age of attainment of developmental milestones. 345

Figure 7-11. The association between the total number of CDS skills attained and the presence or absence of neonatal onset seizures. 347

Figure 7-12. The relationship between age of epilepsy onset and total maximum CDS. 347

Figure 7-13. Total maximum CDS according to the presence or absence of seizures in first 2 months of life. 348

Figure 7-14. The relationship between age at last collection and total maximum CDS. 348

Figure 7-15. The relationship between age at last collection and total recent CDS. 349

Figure 7-16. The total maximum CDS by vital status. 351

Figure 7-17. The relationship between age at diagnosis and year of diagnosis. 352

Figure 7-18. The phenotypic associations of carriers of missense *CDKL5* variants. 353

Figure 7-19. The phenotypic associations of carriers of missense *CDKL5* variants viewed conceptually. Associations for non-missense variants are identical with reversal of the direction (enrichment versus depletion). 354

Figure 7-20. The association between the total number of CDS skills attained and the broad class of *CDKL5* variant. 355

Figure 7-21. The phenotypic associations of 40 carriers of de novo *CDKL5* variants..... 356

Figure 7-22. The phenotypic associations of 10 carriers of large deletions of *CDKL5*. 357

Figure 7-23. The phenotypic associations of participant by sex. (A) presented from the perspective of males (the minority in this cohort)..... 359

Figure 7-24. Total maximum and total recent CDS by sex. 360

Figure 7-25. The phenotypic associations of those believed to have had neonatal seizures. . 361

Figure 7-26. The phenotypic associations of those believed not to have had neonatal seizures, viewed statistically. 362

Figure 7-27. The phenotypic associations of those believed not to have had neonatal seizures, viewed conceptually. 363

Figure 7-28. The total maximum CDS by the occurrence of neonatal seizures and whether the diagnostic *CDKL5* variant is missense. 364

Figure 7-29. Attainment of CDS skills by sex, the occurrence of neonatal seizures, and whether the diagnostic *CDKL5* variant is missense. 366

List of tables

Table 2-1. Classifications and definitions consulted to generate the HPO 2020-12-7 seizure subontology..... 25

Table 2-2. HPO 2017 seizure terms retired or subsumed in the HPO 2020. 27

Table 2-3. Description of the HPO 2020 and the HPO 2017 seizure subontologies and the frequency of the 32 most frequent seizure concepts in the combined cohort of 791 individuals according to each. 29

Table 2-4. Comparison of the amount of phenotypic information captured about the seizures of 791 individuals using the HPO 2017 and the HPO 2020. 30

Table 4-1. Contingency table for analysis of the association between a genetic factor and the annotation of a phenotype..... 68

Table 4-2. $\mathrm{Sim}_{\mathrm{Resnik}}$ with a duplicated annotation. 102

Table 4-3. The effect of including a duplicated annotation in semantic similarity measures based on $\mathrm{Sim}_{\mathrm{Resnik}}$. 102

Table 4-4: $\mathrm{Sim}_{\mathrm{Lin}}$ with a duplicated annotation. 103

Table 4-5. The effect of including a duplicated annotation in semantic similarity measures based on $\mathrm{Sim}_{\mathrm{Lin}}$. 103

Table 4-6. The ten most common intersections of a pair's propagated annotation sets. 118

Table 4-7. The sensitivity of median similarity of non-carriers to carrier group size by similarity measure. 121

Table 5-1. List of modifications to the HPO (version released 2022-02-14). 135

Table 5-2. Harmonization of quantitative phenotypes. 136

Table 5-3. The free-text variables and number of free-text entries on each REDCap form. . 137

Table 5-4. The classification of single variants. 143

Table 5-5. The demographics and broad clinical features of the cohort. 150

Table 5-6. The number of annotations and amount of information describing the phenotypes of the cohort and per participant. 150

Table 5-7. The number of genes carrying different URV classes by gene set. 155

Table 5-8. The number of carriers of different URV classes by gene set. 156

Table 5-9. The number of dominant epilepsy genes achieving phenotypic similarity $p < 0.05$ according to variant class and similarity measure. 158

Table 5-10. A summary of the associations of carriers of PAVs in dominant epilepsy genes. 171

Table 5-11. A summary of the associations of carriers of PTVs in dominant epilepsy genes. 185

Table 5-12. A summary of the associations of carriers of dPAVs in dominant epilepsy genes. 196

Table 5-13. A summary of the associations of carriers of dMissense variants in dominant epilepsy genes. 200

Table 5-14. A summary of the associations of carriers of missense variants in dominant epilepsy genes. 203

Table 5-15. A summary of the associations of carriers of PAV, dPAV, dMissense, and missense, variants in non-dominant epilepsy genes. 210

Table 5-16. A summary of the associations of carriers of PAV, dPAV, dMissense, and missense variants in clinical genes beyond the GE 'green' epilepsy genes. 239

Table 5-17. Clinical associations of carriers of ultra-rare variants in genes without a clinically validated role in epilepsy 252

Table 6-1. The percentage of 10,880 Epi25 participants included in the phenomic analysis cohort annotated with a clinically-prioritized selection of common epilepsy phenotypes. 278

Table 6-2. Genome-wide significantly associated CNV regions and credible intervals. 284

Table 6-3. Known disease genes in the credible intervals of the seizure-associated CNV regions. 285

Table 6-4. Significant individual CNV-HPO associations in the Epi25 phenomic cohort. ... 290

Table 6-5. List of cohorts for which CNV calls were made. 299

Table 6-6. Phenotypes of neuropsychiatric disorders cohorts used in this study. 300

Table 6-7. CNV frequencies in the cases & controls of the meta-analysis, DGV-Gold Standard, and DECIPHER databases. 301

Table 6-8. Resources used to collect information about CNVs and the affected genes. 302

Table 7-1. Modification of HPO version 1.7.14 (release 2022-02-14) for this study.Table 5-1 has been 320

Table 7-2. Description of the demographics, CDKL5 Developmental Scores, and variants of the cohort. 326

Table 7-3. Comparison of electroencephalographic abnormalities in childhood and adulthood. 337

Table 7-4. Antiseizure treatment across all ages..... 340

Table 7-5. Antiseizure treatment in adulthood..... 341

Table 7-6. The evolution of CDS into adulthood..... 346

Table 7-7. The relationship between age and loss of skills. 349

Table 7-8. Attainment of CDS skills by whether the diagnostic *CDKL5* variant is missense or not, among those with neonatal seizures. 367

Table 7-9. Attainment of CDS skills by whether the diagnostic *CDKL5* variant is missense or not, among those without neonatal seizures. 367

Chapter 1. An introduction to epilepsy, its genetics, and phenotypic analysis

1.1 Summary

The epilepsies are a group of collectively common, clinically heterogeneous disorders characterized by a predisposition to unprovoked seizures. Some rare epilepsies have both clinically and biologically distinctive features. However, many people with epilepsy have a disorder that appears common or hard to classify among existing syndromes. Additionally, some syndromes are clinically variable. The health and socioeconomic burden of epilepsies is considerable, and current, often empirical treatment remains inadequate for a third of people affected. Genetic causes of rare epilepsies have been discovered and a large genetic liability is suspected in the majority of currently unexplained cases.

Clinical genetic studies have traditionally relied upon meticulous manual analysis of phenotype in families or unrelated individuals with distinctive clinical features. However, the increasingly apparent genetic heterogeneity of common epilepsies has indicated a need for large cohorts in order to empower studies. Advances in high-throughput genetic technologies and in bioinformatic analysis pipelines have facilitated the identification of common and rare genetic factors that are associated with epilepsy risk. However, the challenges of gathering, harmonizing, and interpreting clinical data from collaborations at this scale has limited the phenotypic resolution of large studies. Assuming that a significant proportion of clinical similarity reflects genetic similarity, the clinical relevance of large genetic studies could be increased through more granular consideration of clinical phenotype.

The formal approaches to analyze large amounts of complex clinical phenotypic data are less established than those for genetic data. Conceptual representations of phenotypic features such as the Human Phenotype Ontology (HPO) are used by genetic diagnostic laboratories and academics to standardize phenotypic data for the interpretation of variants. However, they can also be used to model the phenotypic reasoning traditionally undertaken by clinical experts to evaluate whether people sharing a particular genetic feature are sufficiently similar clinically to suggest a shared disorder.

This doctoral work aims to discover genotype-phenotype associations in data sets too large and complex for manual analysis, at a resolution that can stratify care and clinical counselling for single patients and inform the design of stratified medicine trials. Associations between genetic and clinical features would inform future refinement of the classification of epilepsies to increase its biological validity, and thus facilitate advances in clinical management via a

genetically stratified precision medicine approach to prognostication, treatment, and explanation of a person's experience of epilepsy.

1.2 The definitions and burden of epilepsy and seizures

> *'An epileptic seizure is a transient occurrence of signs and/or symptoms due to abnormal excessive or synchronous neuronal activity in the brain… An epileptic seizure is transient, demarcated in time, with a clear start and finish… Epilepsy is a disorder of the brain characterized by an enduring predisposition to generate epileptic seizures and by the neurobiological, cognitive, psychological, and social consequences of this condition… Epilepsy is not one condition, but is a diverse family of disorders, having in common an abnormally increased predisposition to seizures.'*
>
> The International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (Fisher *et al.*, 2005).

The epilepsies are disorders primarily characterized by epileptic seizures (hereafter simply *seizures*). In order for a person to have epilepsy they must (i) have had a transient clinical event (either manifest or subjective) with documented epileptiform electroencephalographic changes or, more commonly, consistent with a recognized type of seizure, and (ii) continue to have an increased risk of seizures due to an enduring alteration of the brain (Fisher *et al.*, 2005). Operationally, this is indicated either by recurrent unprovoked seizures,[1] or by a single unprovoked seizure with other evidence of a continuing risk similar to that following two unprovoked seizures such as an epileptogenic structural brain lesion or metabolic disorder, or the recognition of a specific electroclinical epilepsy syndrome (Fisher *et al.*, 2014).

The global age-standardized prevalence of active epilepsy[2] in 2016 exceeded 0.5 %, split equally between cases with and those without a demonstrable cause, and the epilepsies as a whole were responsible for at least 13 million disability-adjusted life years, exceeding 0.5 % of all cause disability-adjusted life years globally (Global Burden of Disease Epilepsy Collaborators, 2019). Suspected seizures were the most common neurological reason why people attended National Health Service (NHS) emergency departments (accounting for 0.33 % of all coded diagnoses) or were admitted to hospital as an emergency in England between 2007 and 2013 (Dickson *et al.*, 2018). They accounted for 0.71 % of all unscheduled admissions.

[1] This indicates a risk of further recurrence of at least 60 % over 10 years (Hauser *et al.*, 1998).

[2] Defined as a history of at least two seizures with at least one seizure in the past 5 years

Additionally, 34 % of those people reattended within six years because of a further seizure. The combined annual direct and indirect costs of epilepsy in high income countries are in the thousands of pounds per person with epilepsy (Allers *et al.*, 2015).

A third of juveniles, adolescents and adults treated in a leading UK epilepsy center continue to have seizures despite trials of multiple treatments (Chen *et al.*, 2018d). This proportion with drug-resistant epilepsy (Kwan *et al.*, 2010) has remained static over the past 25 years despite a stream of new antiseizure medications (ASMs) and evidence from three large pragmatic trials that can guide the selection of medications based on broad phenotypic features (Kwan and Brodie, 2000; Marson *et al.*, 2007a; Marson *et al.*, 2007b; Marson *et al.*, 2021a; Marson *et al.*, 2021b). However, people with rare epilepsies can benefit from genetically stratified treatment decisions (Symonds *et al.*, 2019; Helbig and Ellis, 2020; Johannesen *et al.*, 2020), based on studies of cohorts defined by genetic and phenotypic features (de Vivo *et al.*, 1991; Chiron *et al.*, 2000; Brunklaus *et al.*, 2012; Boerma *et al.*, 2016; Devinsky *et al.*, 2017; Devinsky *et al.*, 2018; Lagae *et al.*, 2020). Hence, a classification that represents biological natural classes defined by cause may complement the current largely phenotypic classifications, with the hope that this may lead to more effective and efficient clinical care and associated improvements in the personal and societal burden of epilepsy.

1.3 The classification of epilepsy and seizures

'The study of the causes of things must be preceded by the study of things caused.'

'We have multitudes of facts, but we require, as they accumulate, organizations of them into higher knowledge; we require generalizations and working hypotheses.'

Both attributed to John Hughlings Jackson (1835-1911) by Beveridge (1957).

The father of modern epileptology stressed that to elucidate the mechanisms responsible for particular epilepsies we should begin by noting their complex phenotypic features and conceptualizing their relationships. The clinical and etiological complexity of the epilepsies suggests that models of their phenotypic repertoire should form the basis for interpreting genetic data.

Expert consensus classifications of seizures and epilepsies have been proposed by the Commission on Terminology of the ILAE (ILAE, 1964; Gastaut, 1969; ILAE,1981; ILAE, 1989). These were revised recently with the aim of incorporating modern concepts based on

scientific advances and clearer language (Trinka *et al.*, 2015; Fisher *et al.*, 2017b; Scheffer *et al.*, 2017; Pressler *et al.*, 2021; Hirsch *et al.*, 2022; Riney *et al.*, 2022; Specchio *et al.*, 2022; Wirrell *et al.*, 2022a; Zuberi *et al.*, 2022). However, alternatives have also been developed and are in common use, particularly in the surgical management of epilepsy (Luders *et al.*, 1998; Loddenkemper *et al.*, 2005). The importance of classification in such a varied group of disorders has led to several proposed changes and much debate (Engel, 1998; Everitt and Sander, 1999; Engel, 2001; Avanzini, 2003; Berg and Blackstone, 2003; Engel, 2003; Fisher, 2003; Lüders;Najm and Wyllie, 2003; Nicholl, 2003; Wolf, 2003; Engel, 2006; Iinuma *et al.*, 2006; Berg *et al.*, 2010; Panayiotopoulos, 2011; Luders *et al.*, 2012; French, 2014; Scheffer *et al.*, 2016; Palmini *et al.*, 2020; Rosenow *et al.*, 2020; Steriade *et al.*, 2022). Recalling Hughlings Jackson's comparison of the classifications used by botanists and gardeners, Peter Wolf reflected on how much of the debate stems from the tension between the needs of an *operational arrangement* of concepts to guide pragmatic clinical management within a range of health care settings, and a *taxonomy* of concepts to be used in research (Jackson, 1958; Wolf, 2003). The new classification of epilepsies aims to serve both purposes and recommends that epilepsies be primarily categorized into *epilepsy types* according to their predominant types of seizure, comorbidities, and broad cause (Scheffer *et al.*, 2017).

Seizures are the fundamental features that define epilepsy, and that differentiate the various epilepsies. According to the ILAE, seizures are primarily classified with regard to whether their *initial manifestation* indicates a *focal* or *generalized-onset*, and subsequently according to whether their initial manifestation is *motor* or *non-motor* (Fisher *et al.*, 2017a; Fisher *et al.*, 2017b).[3] *Focal-onset* seizures are conceptualized as 'originating within networks limited to one hemisphere' and *generalized* seizures as 'originating at some point within, and rapidly engaging, bilaterally distributed networks', practically meaning that both cerebral hemispheres are involved at the onset of clinical manifestations (Berg *et al.*, 2010). The current ILAE classification categorizes a seizure pragmatically by semiology: its initial clinical manifestations are interpreted in the context of historical video-electroencephalographic studies that associated particular clinical features with electroencephalographic paroxysms of either focal or generalized onset (Fisher *et al.*, 2017b). While all seizures must be of either focal or generalized onset, there are instances in which there is insufficient evidence to establish which.

[3] Others describe the components of a seizure (Luders *et al.*, 2019) as phases of the same clinical event that can be distinguished semiologically, arguing that asserting focal or generalized-onset from the seizure description conflates seizure and electroencephalographic components to the phenotypic characterization of a person's disorder.

Consequently, the seizure is classified as being of *unknown onset* and described according to its observed features. *Unknown* in these labels refers not to the seizure phenomenon intrinsically, but rather reflects our knowledge about that particular instance of the seizure being described. While the aim is to move towards a more biological classification, the ILAE Commission for Classification and Terminology acknowledges that much of the underlying biological basis for seizure types and their semiological variability remains unknown, and consequently the current ILAE seizure classification is primarily a pragmatic descriptive, operational model (Fisher *et al.*, 2017b; Fisher *et al.*, 2019). It seems that while many changes from the 1981 classification simply updated labels of shared concepts, the 2017 classification may increase the proportion of people whose seizures can be positively categorized (Gao *et al.*, 2018).

According to a person's seizure types, their epilepsy can be classified by *type* as *generalized* (seizures of exclusively generalized onset), *focal* (seizures of exclusively focal onset), *combined generalized and focal* (seizure types of both generalized and focal seizures), or in the presence of seizures of uncertain onset, *unknown* (Scheffer *et al.*, 2017). The relative frequency of epilepsy types is unclear and varies according to study methodology, geography, and age, and because in many cases the type is unknown, as is the rate of misclassification (Manford *et al.*, 1992; Banerjee;Filippi and Hauser, 2009; Global Burden of Disease Epilepsy Collaborators, 2019; Wigglesworth *et al.*, 2023).

Beyond epilepsy *types*, the ILAE emphasizes consideration of age, comorbidities, and etiology when classifying an epilepsy (Scheffer *et al.*, 2017). It recognizes a range of more specific age-related *electroclinical syndromes* and clinical constellations associated with specific causes, clinical features, and outcomes (Hirsch *et al.*, 2022; Riney *et al.*, 2022; Specchio *et al.*, 2022; Zuberi *et al.*, 2022)[4] and a concept of developmental and epileptic encephalopathies (DEEs, Scheffer *et al.*, 2017). DEEs are severe childhood-onset epilepsies of any type, characterized by seizures that are often drug-resistant, although in some cases seizures may spontaneously improve with age. There are prominent neurodevelopmental features such as developmental delay, regression, intellectual disability, and autism spectrum disorder. Challenging behavioral, motor, and sleep disorders and non-neurological comorbidities are common. The burden of these non-seizure comorbidities can often eclipse that of the seizures, particularly with

[4] The terminology used for specific types of seizure and epilepsy syndromes sometimes overlaps because some syndromes were named after their distinctive seizure type. In order to minimize ambiguity, I have capitalized the first word of syndromes. Thus, Febrile seizures is a syndrome in which a neurodevelopmentally normal child experiences febrile seizures in the acceptable age range of 3 months to 6 years.

advancing age. The differentiation of DEEs from neurodevelopmental disorders with epilepsy is made on clinical grounds based on the burden of seizures and the chronological relationship between the emergence and worsening of seizures and the emergence and progression of non-seizure phenotypes. An *epileptic encephalopathy* is a disorder in which an impairment in neurodevelopmental trajectory is thought to have occurred, at least in part, as a consequence of seizure burden on the developing brain.

The chronology of the epilepsies is also highly variable with a bimodal distribution of age of onset (Beghi and Giussani, 2018). People with onset in adulthood often have focal epilepsies secondary to acquired causes such as infection, cerebrovascular disease, and malignancy (Carpio *et al.*, 2024; Freiman *et al.*, 2024). The epilepsy of only a minority can be classified into historical ILAE syndromes (Manford *et al.*, 1992).

In summary, the heterogeneity of epilepsy requires classification on which to base research and clinical management. Classifications are evolving, refining existing phenotypic concepts but recognizing the need for integration of biomarkers of epileptogenic mechanisms in future.

1.4 The role of genetics in explaining why someone has epilepsy and stratifying their care[5]

One in 300 people has an idiopathic epilepsy, that is an epilepsy with no known or suspected cause other than a suspected hereditary predisposition (Global Burden of Disease Epilepsy Collaborators, 2019). The ILAE has recently recommended that in the context of the generalized epilepsies, the historic term *idiopathic* (meaning 'pertaining to itself') be replaced by the term *genetic* because the ILAE wishes to emphasize that the cause of these disorders can be presumed to have a large genetic component given their high heritability (Berkovic *et al.*, 1998; Miller *et al.*, 1998; Kjeldsen *et al.*, 2003; Marini *et al.*, 2004; Vadlamudi *et al.*, 2004; Peljto *et al.*, 2014; Vadlamudi *et al.*, 2014; Epi4K Consortium, 2017; Ellis *et al.*, 2019; Epi4K Consortium, 2019) and the absence of acquired brain insults (Scheffer *et al.*, 2017). In recent years the term *idiopathic* has increasingly been reserved for four classical syndromes: the Idiopathic Generalized Epilepsies, traditionally considered not to be associated with phenotypes beyond generalized-onset seizures and electroencephalographic features: Childhood Absence

[5] I provide a summary here for brevity. For a full review of the clinical application genetics in epilepsy, of which I was the primary author, see Lewis-Smith *et al.* (2020a).

Epilepsy, Juvenile Absence Epilepsy, Juvenile Myoclonic Epilepsy, and Epilepsy with Generalized Tonic-Clonic Seizures Alone (Hirsch *et al.*, 2022).

Studies of families with rare mendelian epilepsies (Afawi *et al.*, 2016) and sporadic DEEs (Epi4K Consortium and Epilepsy Phenome Genome Project, 2013) have shown that variants in over 400 genes can cause epilepsy (Oliver *et al.*, 2023; Genomics England, 2024a), whether transmitted (Steinlein *et al.*, 1995; Escayg *et al.*, 2000) or occurring de novo (Claes *et al.*, 2001; Epi4K Consortium and Epilepsy Phenome Genome Project, 2013). Surgical specimens from people with drug-resistant focal epilepsy have demonstrated somatic variants in some of these genes (López-Rivera *et al.*, 2022). However, beyond the DEEs, only a minority of cases of epilepsy can be attributed to single variants (mendelian or de novo) and the genetic architecture appears to be complex with contributions from both common and rare genetic variants.

Polygenic combinations of large numbers of common variants (single nucleotide polymorphisms, SNPs) appear to explain a third of the heritable liability to Idiopathic Generalized Epilepsies and a tenth of the liability to various focal epilepsies (ILAE Consortium on Complex Epilepsies, 2018; Leu *et al.*, 2019; ILAE Consortium on Complex Epilepsies, 2023). These may also contribute to the 'monogenic' epilepsies (Campbell *et al.*, 2022; Martins Custodio *et al.*, 2023).

Ultra-rare sequence variants (URVs, those that are not encountered in large population databases) and rare copy number variants (CNVs) are more common in people with epilepsies from across the phenotypic spectrum (Helbig *et al.*, 2009; Heinzen *et al.*, 2010; Mefford *et al.*, 2010; Mefford *et al.*, 2011; Mullen *et al.*, 2013; Epilepsy Phenome Genome Project and Epi4K Consortium, 2015; Lal *et al.*, 2015; Addis *et al.*, 2016; Epi4K Consortium and Epilepsy Phenome Genome Project, 2017; Addis *et al.*, 2018; Jabbari *et al.*, 2018; May *et al.*, 2018; Monlong *et al.*, 2018; Epi25 Collaborative, 2019; Epi25 Collaborative, 2021; Koko *et al.*, 2021; Epi25 Collaborative, 2023; Montanucci *et al.*, 2023). These are particularly enriched in genes already associated with rare monogenic epilepsies, genes involved in gamma-aminobutyric acid (GABA) neurotransmission, and genes in which haploinsufficiency appears to be poorly tolerated (Epi4K Consortium and Epilepsy Phenome Genome Project, 2017; May *et al.*, 2018; Epi25 Collaborative, 2019; Takata *et al.*, 2019; Epi25 Collaborative, 2023). Cation channel genes (some of which are established as monogenic epilepsy genes) have also been implicated in common epilepsies by URVs, but evidence for URVs in excitatory receptors (such as the acetylcholine receptor genes in which some variants can cause Autosomal Dominant Nocturnal Frontal Lobe Epilepsy) is more limited (Epi25 Collaborative, 2019; Epi25 Collaborative, 2023). Beyond the monogenic epilepsies, little is known of the role of URVs in causing specific

clinical features beyond an association with broad epilepsy types, and while some CNVs are associated with genetic syndromes, many appear to act as risk factors with incomplete penetrance. Furthermore, there appears to be pleiomorphism with variants in the same genes being found among a broad range of neurodevelopmental, neuropsychiatric, and neurodevelopmental disorders (Brainstorm *et al.*, 2018; Peng;Zhou and Wang, 2021).

Many recent discoveries have required large collaborative studies, made feasible by superficial categorization of participants into broad epilepsy classes. More precise phenotypic analysis of participants within large genetic studies has the potential to improve the resolution of phenotypic associations. Precise associations could inform future classifications that use both molecular and phenotypic markers to define distinct epilepsies that are closer to natural classes.

> *'In precision medicine, the ultimate end point is the selection of a subset of patients, with a common biological basis of disease, who are most likely to benefit from a drug or other treatment…'*
>
> National Research Council (2011)

Precision medicine aspires to exploit markers of natural classes of disease and of the relevant parts of an individual patient's biology to improve clinical management. Inherent in the identification of patients who could benefit from a particular treatment is the need to prognosticate accurately.

Prognostication begins with the prediction of a personalized risk of having a first seizure or developing epilepsy after a single seizure. Beyond the diagnosis of epilepsy is its classification by type and syndrome based on recognizing characteristic seizure types, ages of onset, electroencephalographic features, developmental trajectories, and comorbidities that can predict trajectory (Hirsch *et al.*, 2022; Riney *et al.*, 2022; Specchio *et al.*, 2022; Wirrell *et al.*, 2022b; Zuberi *et al.*, 2022). Some of these syndromes historically included terms such as *benign* (Loiseau *et al.*, 1988) and *malignant* (Coppola, 2009) in their names, emphasizing the role of syndromic diagnosis in prognostication. Estimates of seizure recurrence risk after medication withdrawal based on clinical features have been possible for over a decade (MRC Antiepileptic Drug Withdrawal Study, 1991), and recurrence can now be predicted more precisely using models based on meta-analysis of the clinical features of longitudinal cohorts (Lamberink *et al.*, 2017).

The success of medical treatment decisions for common epilepsies remains disappointing with pharmacoresistance remaining stubbornly at a third (Beyenburg;Stavem and Schmidt, 2010; Chen *et al.*, 2018d) and adverse events remaining frequent (Alsfouk *et al.*, 2020), despite hopes that newer ASMs and prescribing practices may have improved both. Epileptologists are able to stratify medications using evidence of tolerability and efficacy based on coarse phenotypic features at the level of broad epilepsy type (Marson *et al.*, 2007a; Marson *et al.*, 2007b; Marson *et al.*, 2021a; Marson *et al.*, 2021b) or common syndromes (Glauser *et al.*, 2013; Silvennoinen *et al.*, 2019), although a hierarchy of treatment options for particular syndromes has not always been forthcoming (Androsova *et al.*, 2017). Additionally, clinicians mitigate the risk of adverse events by contraindicating particular medications based on their known adverse effects and the patient's particular clinical circumstances.

A significant challenge of developing precision medicine for epilepsies is that seizures and epilepsy are *symptoms* of underlying structural or molecular lesions, and it has become apparent that in many cases it is these lesions that dictate prognosis and treatment response. While in some cases these may be amenable to resection (Chen *et al.*, 2018b), antimicrobials (Otte *et al.*, 2013), immunomodulation (Varley;Taylor and Irani, 2018), or metabolic (Kass *et al.*, 2016; van Karnebeek *et al.*, 2016; Aledo-Serrano *et al.*, 2023) or toxicological (Leach;Mohanraj and Borland, 2012) interventions; in 25 million people no cause is found beyond a presumed genetic predisposition (Global Burden of Disease Epilepsy Collaborators, 2019). Modern classifications of diseases with a genetic mechanism, such as brain tumors (Louis *et al.*, 2016), are increasingly incorporating a greater role for molecular findings in recognition that these may inform prognostication and treatment decisions, and allow a more biological classification of each individual's disease through which to advance research. The ILAE has recognized that future classifications of epilepsy should exploit increasing knowledge of its genetic substrates where these may facilitate more precise clinical management (Berg *et al.*, 2010; Scheffer *et al.*, 2017). The promise of molecular diagnosis to inform clinical decision-making and in particular for genetically stratified prognosis and treatment is increasing (Knowles *et al.*, 2022). However, there are a number of obstacles to overcome before we can integrate genetics into routine care for the benefit of most people with epilepsy.

The main obstacle is that we cannot at present identify a specific molecular contribution in the majority of people with presumed genetic epilepsies. A meta-analysis of 103 studies of high-throughput sequencing diagnostics in 32,331 people with neurodevelopmental disorders including epilepsy found a diagnostic yield of 24 % in epilepsy (27,923 people across 72 studies), similar to that of testing for autism spectrum disorder (17 %) and intellectual

disability (28 %, Stefanski *et al.*, 2021). Those people with either intellectual disability or early-onset seizures had greater yields (28 % and 37 % respectively) than those with epilepsy alone (9.3 %). The yield of testing varied by epilepsy type. Those with generalized-onset seizures, either alone or combined with focal, had a yield of 24 % while those with focal epilepsies had a yield of 16 %. There was also a difference in diagnostic yield by age of onset of seizures. Those with neonatal or infantile onset were more likely to obtain a molecular diagnosis (29 %) than those with childhood onset (15 %). However, the samples for age-stratified analysis were small and it is unclear to what extent the difference in yield by age of onset may be explained by the greater feasibility of trio testing (including both parents) for younger patients to demonstrate de novo status, which carries importance in the clinical diagnostic interpretation of variants (Richards *et al.*, 2015). A previous meta-analysis showed that whole exome sequencing has a higher diagnostic yield (45 %) than gene panel testing (23 %) in studies using a variety of inclusion criteria for people with epilepsy, and that CNV analysis has a much lower yield (8 %, Sanchez Fernandez *et al.*, 2019). The yield was greatest when using trio whole exome sequencing for DEE (70-78 %). The largest study of a clinical whole exome sequencing registry demonstrated a yield of 38 % among phenotypically unselected people with epilepsy (Helbig *et al.*, 2016). The epilepsy subtypes with the greatest yield were DEE (43 %), particularly if neonatal in onset (58 %), followed by generalized (42 %), fever-related (33 %), and focal epilepsies (31 %). Similarly, the yield fell with increasing age of seizure onset from 43 % in neonatal onset, to 37 % for onset in childhood, to 19 % for the few individuals with adolescent and adult onset. A smaller but population-based study of children in Scotland with early onset epilepsies (Symonds *et al.*, 2019) demonstrated a yield of 25 % despite not testing for common causes of early-onset epilepsy such as tuberous sclerosis. A similar yield has been found in smaller studies of adults with epilepsy and intellectual disability, with onset of a neurodevelopmental disorder in childhood (Borlot *et al.*, 2019; Johannesen *et al.*, 2020). In general, the availability of parental deoxyribonucleic acid (DNA) for segregation analysis increases the diagnostic yield and this (as well as the motivation of parents if present) may influence the decision to undertake genetic testing in adult patients, resulting in a selection bias and older adults being less likely to undergo comprehensive investigation (Aledo-Serrano *et al.*, 2020). However, it is encouraging that in addition to ending the diagnostic odyssey, informing reproductive and in some cases prognostic counselling, there may already be peer-reviewed evidence that inform treatment decisions in 33–80 % of cases where a molecular diagnosis is made (Symonds *et al.*, 2019; Truty *et al.*, 2019; Johannesen *et al.*, 2020; Bayat *et al.*, 2021; Bayat *et al.*, 2022; Krey *et al.*, 2022). Yet experience of genetically stratified

prescribing so far has been far from paradigm-shifting for the care of most people (Balestrini *et al.*, 2021).

Regarding SNPs, while different polygenic contributions to common epilepsy types and syndromes have emerged at cohort level, polygenic risk scores currently fall short of the threshold to contribute to clinical decision-making for a particular patient. We cannot yet use them to predict if a person will develop epilepsy or to predict what type, syndrome, or specific features it will take (ILAE Consortium on Complex Epilepsies, 2018; Leu *et al.*, 2019; Moreau *et al.*, 2020; ILAE Consortium on Complex Epilepsies, 2023). Furthermore, current polygenic risk scores are sensitive to ancestral genetic background (Leu *et al.*, 2019; Suzuki *et al.*, 2021).

1.5 The traditional role of phenotype in genetic studies of epilepsy

1.5.1 Phenotypic considerations in gene discovery studies

Clinical genetic discovery has traditionally taken a *phenotype first* approach, depending on experienced clinicians recognizing those patients who share particularly distinctive clinical features that are hypothesized to have a shared genetic basis. These features could be the occurrence of a rare single phenotypic feature (such as those which define their own syndromes: epileptic spasms, myoclonic-atonic seizures, and absence with eyelid myoclonia), an unusual presentation of a more common feature (such as an early onset of absence seizures in Glucose Transporter 1 Deficiency Syndrome or particularly prolonged or complex febrile seizures in Dravet Syndrome), or a complex combination of features (such as the coincidence of focal and generalized seizures, usually with febrile and clonic seizures but not epileptic spasms in Dravet Syndrome). Genetic data are then analyzed in an attempt to explain the phenotypic variance.

How the phenotypes of epilepsy are considered influences the design of genetic analyses and the interpretation of their results. At the most basic level, all forms of epilepsy can be considered as a single phenotype, thus generating a binary variable: a person either meets the diagnostic criteria for epilepsy or does not, with exclusion of uncertain cases (such as those with a single possibly provoked seizure, or with incidental epileptiform electroencephalographic abnormalities but no seizures). The genetic features of individuals with epilepsy would then be compared with those of a control cohort without epilepsy. This design places minimal phenotyping burden on contributing clinicians and participants because little phenotypic data is required. The benefit of this is high case ascertainment and the increase in power that is associated with increasing cohort size. Recent large association studies following this strategy have identified common (ILAE Consortium on Complex Epilepsies, 2018; ILAE Consortium

on Complex Epilepsies, 2023) and rare (Epi4K Consortium and Epilepsy Phenome Genome Project, 2017; Epi25 Collaborative, 2019; Epi25 Collaborative, 2021; Epi25 Collaborative, 2023) variants associated with a risk of epilepsy generically. However, isolated broadly defined case-control analyses cannot demonstrate that the statistically significant associations hold true in all forms of epilepsy. Apparent associations may be driven by some particular subtype of epilepsy and emerge or diminish as a function of the latent phenotypic mix of the case cohort. Furthermore, heterogeneity within the broad case group can result in variants that predispose only to particular subtypes being overlooked.

Classification of participants into more clinically homogeneous strata can identify the genetic associations of more nuanced phenotypic variation. Large genetic association studies typically stratify participants into the generalized and focal epilepsies and the DEEs. Comparisons between these groups and controls without epilepsy have identified statistically significant, or suggestive associations that predispose to the particular type of epilepsy in question but not necessarily to others (Kasperaviciute *et al.*, 2010; Epi4K Consortium and Epilepsy Phenome Genome Project, 2017; May *et al.*, 2018; McCormack *et al.*, 2018; Epi25 Collaborative, 2019; Leu *et al.*, 2019; Epi25 Collaborative, 2023). Phenotypic stratification beyond broad epilepsy types has been successful, revealing syndrome-specific collections of SNPs to which syndromic heritability can be attributed (ILAE Consortium on Complex Epilepsies, 2018; ILAE Consortium on Complex Epilepsies, 2023). Stratification for granular case-control studies has identified associations between SNPs and Juvenile Myoclonic Epilepsy (Roshandel *et al.*, 2023), and between rare genetic variants and Self-Limited Epilepsy with Centrotemporal Spikes (Lemke *et al.*, 2013; Addis *et al.*, 2018; Bobbili *et al.*, 2018; May *et al.*, 2018) and also epilepsies with absence seizures (Addis *et al.*, 2016), as well as DEEs characterized by Infantile Epileptic Spasms Syndrome (Epi4K Consortium and Epilepsy Phenome Genome Project, 2013).

However, this approach is not optimal for the identification of variants that steer some generic predisposition to seizures towards particular seizure types, epilepsy types, syndromes, and variability within syndromes. Such a study should compare cohorts of people defined by different epilepsies with each other rather than control populations presumed not to have epilepsy. Indeed, studies of common variants have compared results from people with different epilepsy syndromes with each other to identify their shared and distinct SNP-related heritability (ILAE Consortium on Complex Epilepsies, 2018; ILAE Consortium on Complex Epilepsies, 2023) However, clinicians recognize that there is great phenotypic variability within epilepsy types, and even within classical syndromes that are thought to have a large genetic contribution,

for example Juvenile Myoclonic Epilepsy (Martinez-Juarez *et al.*, 2006; Kasteleijn-Nolst Trenite *et al.*, 2013).

Early genetic research focused on families, whose phenotypes had been analyzed in great depth (Greenberg *et al.*, 1988a; Greenberg *et al.*, 1988b; Phillips *et al.*, 1995; Steinlein *et al.*, 1995; Scheffer and Berkovic, 1997; Scheffer *et al.*, 1998; Wallace *et al.*, 1998; Xiong *et al.*, 1999; Winawer *et al.*, 2000; Durner *et al.*, 2001; Kalachikov *et al.*, 2002; Winawer *et al.*, 2002; Pal *et al.*, 2003; Greenberg *et al.*, 2005). Even in studies of these sizes, the collection, harmonization, and analysis of clinical phenotypic data is time-consuming and requires clinical expertise.

The quantity and complexity of phenotypic epilepsy data raises questions about what precisely an analysis should attempt to link or associate genetic variants with (Winawer, 2006). The heterogeneity that can be seen between unrelated individuals with the same broad epilepsy type or syndrome, and even within multiplex families, places considerable weight on the question of whether to *lump or split* different phenotypic features such as specific seizure types or age of onset, and if one chooses to split, then according to what unit. Were phenotypic heterogeneity within families to reflect genetic heterogeneity then stratification by phenotypic features may increase the power of linkage analyses, but there is a considerable risk of reducing the power for discovery by excessive stratification that results in small sample sizes and the need to correct for multiple testing (Leal and Ott, 2000). Even when these studies have yielded positive results, these familial epilepsies are rare and the number of sporadic cases that can be attributed to their causes is small. In some cases, their findings have been hard to replicate or reconcile with the results of large association studies further raising doubts as to the benefits of focusing on what may be private or very rare and ancestry-specific causes (Pal *et al.*, 2003; Suzuki *et al.*, 2004; Cavalleri *et al.*, 2007; de Kovel *et al.*, 2007; Pal and Helbig, 2015; Subaran *et al.*, 2015; Bailey *et al.*, 2017; Bailey *et al.*, 2018; Pathak *et al.*, 2018; Lerche *et al.*, 2019; Schulz *et al.*, 2019).

While high-throughput genomic data from increasingly large collaborative cohorts is yielding incremental discoveries when analyzed according to broad phenotypic groups, promising candidate genes have frequently failed to reach genome-wide statistical significance (Epi4K Consortium and Epilepsy Phenome Genome Project, 2017; Epi25 Collaborative, 2019; Epi25 Collaborative, 2023). There have been concerns that this approach may be overlooking mechanistically important associations that might become apparent when considering more nuanced phenotypic features (Greenberg and Subaran, 2011; Greenberg and Stewart, 2012). Furthermore, when based on coarse inclusion and stratification criteria, genetic analyses of a disorder as heterogeneous as epilepsy are unlikely to lead to practical benefits in the short-term as their generalizability to the patient in clinic is questionable.

1.5.2 Genotype-phenotype association

Beyond providing evidence of causality, genotype-phenotype associations are clinically useful, facilitating more precise prognostication and selection of treatments based on studies of individuals with similar clinical and genetic features. The current paradigm for establishing genotype-phenotype associations is largely qualitative and may be susceptible to confirmation and ascertainment biases. A distinctive phenotype is identified and found to be associated with a particular genetic variant. This pairing is then validated in independent cohorts with the same phenotype and becomes established as the classical clinical presentation of the genetic disorder. A carrier with a different phenotype may then be found nested within a family with mendelian transmission of the classical phenotype suggesting variable expressivity, or a broader phenotypic spectrum may become apparent as clinicians collate individuals with unexpected chance findings from gene panel and exome sequencing. These may then be validated in phenotypically matched cohorts. The result of this process is that particular genes and genotypes become associated with a *classical* presentation and then through serendipity, the phenotypic spectrum may expand in a stepwise manner, yet we remain cognitively anchored to the seminal association. One does not know what other phenotypes may be associated with a particular variant were they to be tested.

A more granular analysis of phenotype in large genetic studies of people with presumed genetic, but otherwise relatively unselected epilepsies may be able to identify genotype-phenotype associations in a less biased manner. Genetic association studies increasingly benefit from increasing size facilitated by the falling cost of generating high-quality genomic data and through the increasing sophistication of bioinformatic analysis. This enables a *genotype first* approach, stratifying individuals into groups according to genetic rather than phenotypic features. Then phenotypic data are interrogated statistically for evidence of an association. The phenotypic data should ideally be in a harmonized format, complete, and ascertained blinded to genotype. However, the generation and analysis of phenotypic data have not kept pace with high-throughput genomic technologies. The epilepsies have few objective biomarkers to use as endophenotypes. Hence, the quantity and quality of phenotypic data depend upon a clinician's interpretation of the individual's history, electroclinical examination, and the results of any clinical investigations such as magnetic resonance imaging (MRI). It is hoped that advances in electronic medical record (EMR) structure, clinical coding, and natural language processing of digitally stored unstructured clinical data will make high quality phenotypic data more available for research in the future. Currently phenotypic data collection, harmonization to study clinical

record forms, and quality control checks are largely manual, and dependent on specialist clinical expertise and time. Additionally, participants may be required to attend study specific consultation for data collection. Thus, phenotypic data is hugely valuable, by virtue not only of its potential contribution to analyses but also in view of its cost.

Despite the challenges to be overcome for the incorporation of greater phenotypic granularity into large genomic analyses, such studies do typically collect clinical details and while these are often sparse and variably complete, the data are usually structured, managed, and submitted to quality control to confirm case eligibility and broad classification (Epi25 Collaborative, 2020). However, exploitation of the available data requires automated approaches using models of expert clinical knowledge and that can, ideally, accommodate incomplete data.

In the following two chapters, I shall introduce the HPO as a tool for handling large amounts of granular phenotypic data and then review how it has been applied and combined with other analyses to delineate the clinical features of epilepsies and neurodevelopmental disorders. In Chapter 4, I will critically review proposed methods of phenotypic analysis. In Chapters 5 and 6, I will report two studies I have undertaken using data from over 10,000 people with presumed genetic epilepsies to understand the phenotypic consequences of ultra-rare sequence variants and CNVs. I will conclude by reporting a detailed retrospective natural history study of 67 adults with CDKL5 Deficiency Syndrome and reflecting on what I have learned.

Chapter 2. Modeling seizures in the Human Phenotype Ontology according to contemporary ILAE concepts makes big phenotypic data tractable

2.1 Preface

This chapter describes the development of a new seizure subontology for the HPO, based on contemporary ILAE and other classifications. The chapter is modified from the resulting publication under CC-BY 4.0:

> *Lewis-Smith, D., Galer, P.D., Balagura, G., Kearney, H., Ganesan, S., Cosico, M., O'Brien, M., Vaidiswaran, P., Krause, R., Ellis, C.A., Thomas, R.H., Robinson, P.N. and Helbig, I. (2021) 'Modeling seizures in the Human Phenotype Ontology according to contemporary ILAE concepts makes big phenotypic data tractable',* Epilepsia*, 62(6), pp. 1293-1305.*

I led an international team of early career epilepsy researchers to undertake this project under the supervision of Dr Ingo Helbig (Children's Hospital of Philadelphia). The project emerged out of discussions at the European Molecular Biology Organization (EMBO) workshop on *Phenotyping neurological syndromes for systems genetics*, in Luxembourg, October 4–10th 2018, organized by Dr Roland Krause (Université du Luxembourg), to whom I am grateful as this meeting inspired much of the work that I have embarked upon since. I coordinated the collaborative team, chaired its meetings, created the majority of the new HPO terms and their relationships, mapped pre-existing HPO terms to the new concepts, edited the new ontology as it was developed, provided the new subontology to Prof. Peter N. Robinson (The Jackson Laboratory, Connecticut) for him to integrate it into the production version of the HPO, undertook the analyses, and drafted the manuscript and figures that form the publication and this chapter. I am a corresponding author of the resulting publication. Co-authors contributed to the generation of the intellectual content of new subontology (G.B., H.K.), provided data for the analyses (P.D.G., M.C., M.O., P.V., I.H.), provided supervision and advice (R.K., R.H.T., P.N.R, I.H.), and all co-authors were invited to provide comments on the manuscript. I thank *Epilepsia's* reviewers and editor (Prof. Alica Goldman) for their comments. Collaborators making particular intellectual contributions were Dr Ganna Balagura (Giannina Gaslini Institute, Genoa, Italy) and Dr Hugh Kearney (Royal College of Surgeons in Ireland, Dublin, Ireland), who assisted with review of pre-existing HPO terms, review of classification literature, and the generation of new terms and definitions. Dr Balagura (a training pediatrician) led development of the neonatal seizure subontology (Pressler *et al.*, 2021). Dr Kearney provided advice regarding the practical application of Prof. Hans Lüders' semiological classification (Luders *et al.*, 1998).

2.1.1 Abstract

2.1.2 Objective

The clinical features of epilepsy determine how it is defined, which in turn guides management. Therefore, consideration of the fundamental clinical entities that comprise an epilepsy is essential in the study of causes, trajectories, and treatment responses. The HPO is used widely in clinical and research genetics for concise communication and modeling of clinical features, allowing extracted data to be harmonized using logical inference. We sought to redesign the HPO seizure subontology to improve its consistency with current epileptological concepts, supporting the use of large clinical data sets in high-throughput clinical and research genomics.

2.1.3 Methods

We created a new HPO seizure subontology based on the 2017 ILAE Operational Classification of Seizure Types, and integrated concepts of status epilepticus, febrile, reflex, and neonatal seizures at different levels of detail. We compared the HPO seizure subontology prior to, and following our revision, according to the information that could be inferred about the seizures of 791 individuals from three independent cohorts: two previously published, and 150 newly recruited individuals. Each cohort's data were provided in a different format and harmonized using the two versions of the HPO.

2.1.4 Results

The new seizure subontology increased the number of descriptive concepts for seizures five-fold. The number of seizure descriptors that could be annotated to the cohort increased by 40 % and the total amount of information about individuals' seizures increased by 38 %. The most important qualitative difference was the relationship of focal to bilateral tonic-clonic seizure to generalized-onset and focal-onset seizures.

2.1.5 Significance

We have generated a detailed contemporary conceptual map for harmonization of clinical seizure data, implemented in the official 2020-12-7 HPO release and freely available at hpo.jax.org. This will help to overcome the phenotypic bottleneck in genomics, facilitate reuse of valuable data, and ultimately improve diagnostics and precision treatment of the epilepsies.

2.2 Introduction

Information about the clinical features of research participants is essential for diagnostic interpretation of genetic findings and discovery research in epilepsy genetics (Greenberg and Subaran, 2011; Helbig *et al.*, 2018). To fully exploit these data, information must be harmonized to allow reliable comparison of individuals to diagnostic criteria, or to each other (Lhatoo *et al.*, 2020). Interpretation of phenotypic data is challenging in epilepsy, particularly because the central clinical features — seizures — are diverse, dynamic and can occur in different combinations and in different developmental and comorbidity contexts (Blume *et al.*, 2001; Fisher *et al.*, 2014; Fisher *et al.*, 2017a; Fisher *et al.*, 2017b; Scheffer *et al.*, 2017; Luders *et al.*, 2019). Crucially, the same seizure can be classified at various levels of detail according to the clinical need, or to accommodate incomplete information (Fisher *et al.*, 2017a; Fisher *et al.*, 2017b; Luders *et al.*, 2019).

Even a single individual's clinical data can require harmonization between different formal classifications because of the preferences of their various health care providers, the design of study-specific research data collection forms, or because their history may span several decades (Commission on Classification and Terminology of the ILAE, 1981; Fisher *et al.*, 2017b; Luders *et al.*, 2019). For example, different sources may describe the same seizure as a focal aware non-motor seizure (Fisher *et al.*, 2017b), focal seizure (Blume *et al.*, 2001; Berg *et al.*, 2010; Fisher *et al.*, 2017b), aura (Luders *et al.*, 1998; Blume *et al.*, 2001; Luders *et al.*, 2019), sensory seizure (Blume *et al.*, 2001), or simple partial seizure (Commission on Classification and Terminology of the ILAE, 1981).

Harmonization traditionally occurs manually during data entry, dependent on the clinical expertise of the contributor completing a form specific to a particular diagnostic service or research study. Manual harmonization of data places a large burden on clinicians and risks overly constraining the depth and quality of the phenotypic details provided, or even case ascertainment. Additionally, the format of clinical data collected for one study may not map on to the variables required for another study, limiting the ease and yield of reuse. The need for rapid and accurate automated phenotypic harmonization of large multicenter data sets is increasingly important as diagnostic services and research studies grow to analyze data from tens of thousands of individuals (Akawi *et al.*, 2015; Helbig *et al.*, 2018; ILAE Consortium on Complex Epilepsies, 2018; Lindy *et al.*, 2018; Epi25 Collaborative, 2019; Truty *et al.*, 2019; Lhatoo *et al.*, 2020; NHS Health Education England, 2020; Turro *et al.*, 2020; Epi25 Collaborative, 2023; ILAE Consortium on Complex Epilepsies, 2023).

Ontologies formally model the concepts and relationships that exist in a particular expert domain (Gruber, 1993). The HPO is a tool that has become the lingua franca for clinical and research geneticists for concise communication and harmonization of clinical features (human-phenotype-ontology.org, Kohler *et al.*, 2021). The HPO is central to the data capture in the Deciphering Developmental Disorders study (Akawi *et al.*, 2015) and the National Institute of Health Research BioResource for the 100,000 Genomes Project (Turro *et al.*, 2020), and for the upcoming National Health Service England Genomic Medicine Service clinical genome sequencing service (NHS Health Education England, 2020). Through incremental development with contributions from experts across the breadth of clinical domains since 2008, the HPO has come to include 15,247 terms describing distinct clinical phenotypic concepts (Kohler *et al.*, 2021). Each concept has a name, unique identifier (e.g. *HP:0001250*) and definition, and may have synonyms (by which it may be searched for), comments, and cross-referenced concepts from other databases such as Online Mendelian Inheritance in Man (OMIM, omim.org, McKusick-Nathans Institute of Genetic Medicine, 2020) or Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT, snomed.org, SNOMED International, 2020) and PubMed identifiers of supporting references. Current projects include translation of the HPO into multiple languages and improvements in the cross-referencing to clinical diagnostic and research resources such as Orphanet (www.orphadata.org) and ontologies of model organism phenotypes (Kohler *et al.*, 2021).

Each relevant clinical feature of an individual is annotated as the most conceptually specific HPO terms necessary to describe it fully. To distinguish HPO terms from formal seizure types or semiological descriptions within the current manuscript, we refer to HPO concepts in italics by their capitalized name and HPO concept unique identifier (CUI). For example, if an individual has focal aware motor seizures, then these should be coded as *Focal aware motor seizure [HP:0020217].* Automated reasoning algorithms can then infer from the *is_a* relationship within the HPO that the individual has a type of *Focal aware seizure [HP:0002349]* and *Focal motor seizure [HP:0011153]*, and from either of these that they have a *Focal-onset seizure [HP:0007359]*, and from this a *Seizure [HP:0001250].* Hence, from a single input datum, an individual's clinical features can be automatically described across the breadth of applicable descriptive terms; terms which span phenotypic depth. Consequently, coupled with natural language processing and thesauruses of synonyms, these algorithms can translate medical records and research data into the same HPO format and compare the clinical features of individuals to each other for research (Shen;Wang and Liu, 2017; Son *et al.*, 2018; Helbig *et al.*, 2019; Galer *et al.*, 2020; Ganesan *et al.*, 2020), or to the clinical features associated with particular genetic disorder or syndrome in a database for variant prioritization (**Figure 2-1**; Kohler *et al.*, 2009; Girdea *et al.*, 2013;

Zemojtel *et al.*, 2014). As is the case with clinical records spanning classifications, data coded as HPO terms may span multiple versions because the HPO is regularly updated to incorporate advances in phenotypic conceptualization. Unique term identifiers map to consistent concepts to facilitate translation from one HPO version to another.

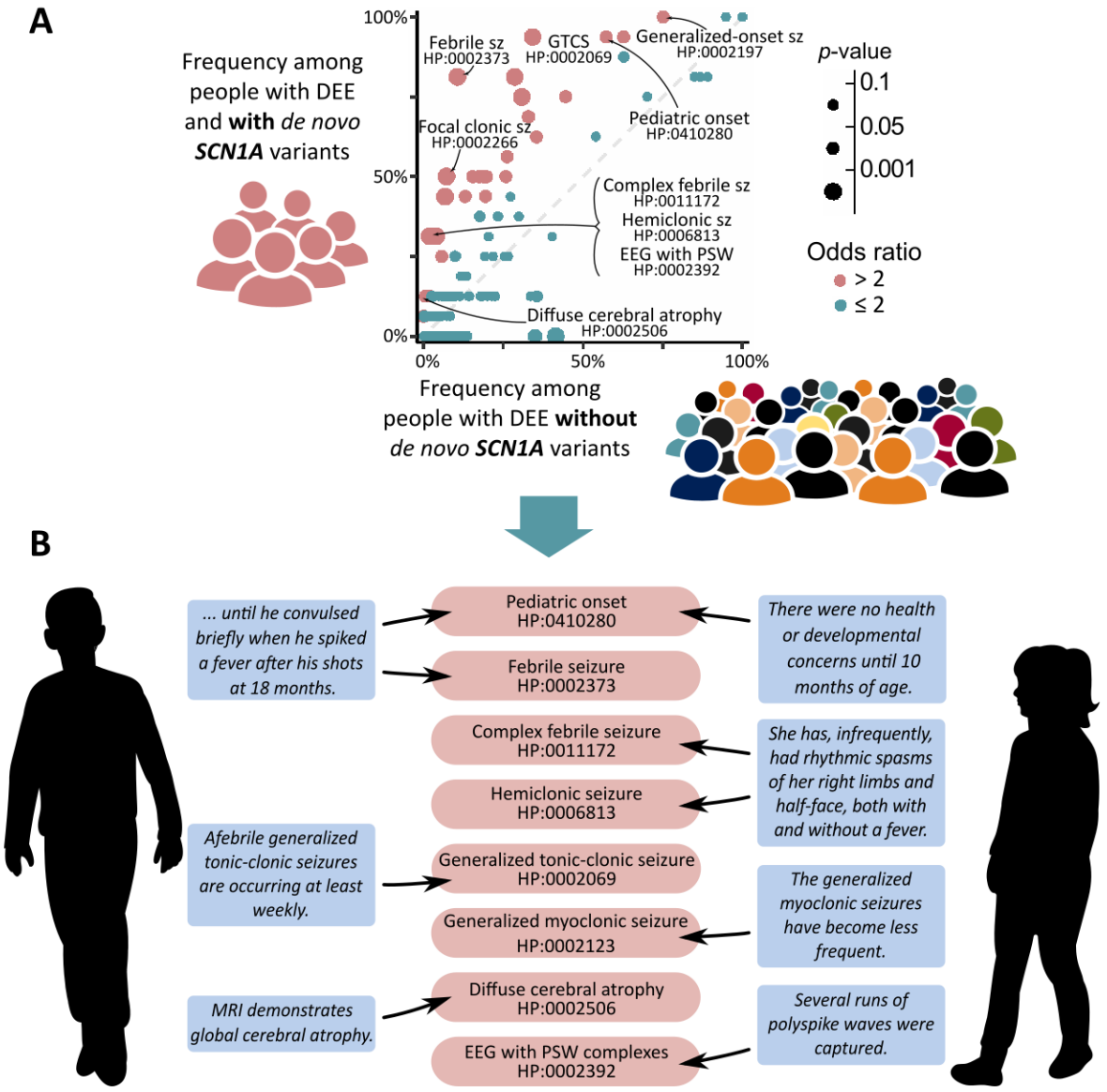


Figure 2-1. Harmonization of clinical annotations using the HPO. (A) The HPO can harmonize phenotypic data from large cohorts to identify phenotypic associations with a particular genetic (or other) categorical factor, in this case de novo variants in *SCN1A*. Data from Galer *et al.* (2020), *p*-values and odds ratio obtained using the two-sided Fisher's exact test. (B) Once the set of HPO terms associated with this form of epilepsy is known, the clinical features of patients can be translated into HPO terms to assess how closely they match phenotypically. GTCS, generalized tonic-clonic seizure; PSW, polyspike-wave; sz, Seizure.

Recognizing the importance of the HPO in genomics, the pertinence of seizure types to the conceptualization of epilepsies (Scheffer *et al.*, 2017; Luders *et al.*, 2019; Rosenow *et al.*,

2020), and the recent revision of seizure classification (ILAE 2017, Fisher *et al.*, 2017b), we have created a new ontology of contemporary seizure concepts for the HPO. Herein we demonstrate harmonization of three cohorts using the HPO, and compare the information captured by our new seizure subontology (release 2020-12-7) to that from a prior iteration (release 2017-12-12).

2.3 Methods

2.3.1 Cohorts providing seizure data

The local epilepsy cohort for this study comprises 150 individuals with genetic generalized epilepsy (GGE, $n = 65$), focal epilepsy ($n = 53$), and developmental and epileptic encephalopathy (DEE, $n = 32$) recruited from CHOP for the Epi25 Collaborative (Epi25) study.(Epi25 Collaborative, 2019) We used seizure data within the Epi25 format but did not use primary data of Epi25. Informed consent for participation was obtained from participants themselves or, where necessary, their parents in agreement with the Declaration of Helsinki, and the study was completed per protocol using de-identified data with local approval by the CHOP Institutional Review Board (IRB 15-12226). Data from two further and independent cohorts with DEE were reused to demonstrate harmonization of distinct data sets. These have been published previously: Helbig *et al.* (2019), comprising 306 individuals; and a subcohort from Galer *et al.* (2020), comprising 335 individuals from the Epilepsy Phenome/Genome Project (EPGP). We included only unique individuals documented as having any form of epileptic seizure.

The local cohort's data were in the form of categorical seizure variables from three Epi25 data collection forms mapping to each of GGE, focal epilepsy, and DEE (epi-25.org/epi25-forms). Data from Helbig *et al.* (2019) are based on HPO release 2017-12-12 (HPO 2017) and those from Galer *et al.* (2020) are based on HPO release 2018-11-8.

2.3.2 Harmonization of data using the HPO

For each cohort, we created a dictionary translating seizure data from their raw format into the specified version of the HPO. Once data had been translated into HPO annotations the *is_a* relationship between terms of the HPO were used to infer the other, conceptually less specific HPO seizure descriptors that must apply to the individual, as described with focal aware motor seizures above. We call this process of capturing all the terms applicable to an initial annotation, comprehensively describing a phenotypic feature at different levels of detail, propagation

(Schulz *et al.*, 2011). The set of propagated terms includes both the initial term annotated through translation using a dictionary and those inferred by logic within the ontology.

2.3.3 Statistical procedures

All statistical tests and ontology-based automatic inferences were performed using the R Statistical Framework, including the *tidyverse* collection of packages (R Core Team, 2017) and *ontologyIndex* (Greene;Richardson and Turro, 2017).

2.3.4 The process of revising the HPO seizure subontology

We set up an international collaboration to build a seizure subontology for the HPO based on contemporary seizure concepts at the EMBO workshop on *Phenotyping neurological syndromes for systems genetics*, in Luxembourg, October 4–10th 2018. Our group met fortnightly by videoconference to coordinate ontology design using the online tool *WebProtégé* (webprotege.stanford.edu, Tudorache *et al.*, 2013), under the supervision of I.H. and R.K. (Epilepsiome Task Force of the Genetics Commission of the ILAE) and P.N.R. (creator of the HPO). We considered feedback and comments provided by HPO users on the HPO GitHub page (github.com/obophenotype/human-phenotype-ontology).

2.4 Results

2.4.1 HPO-based inference harmonizes seizure descriptions across different levels of precision, semiologies and classifications

We translated categorical seizure data in Epi25 format from 150 people in our local cohort and HPO release 2019-11-8 seizure annotations of 335 participants in EPGP (Galer 2020 cohort, Galer *et al.*, 2020) into HPO 2017 terms. This harmonized their data with those already coded according to HPO 2017 from 306 individuals (Helbig 2019 cohort, Helbig *et al.*, 2019). We compared the number of seizure descriptors that could be annotated to each of these 791 individuals prior to and following propagation using this version of the HPO (**Figure 2-2A**). We found that the median number of terms by which each individual's seizures could be described increased from two to six as a result of propagation, with a median increase of three and Wilcoxon signed rank test two-sided p-value < 0.001.

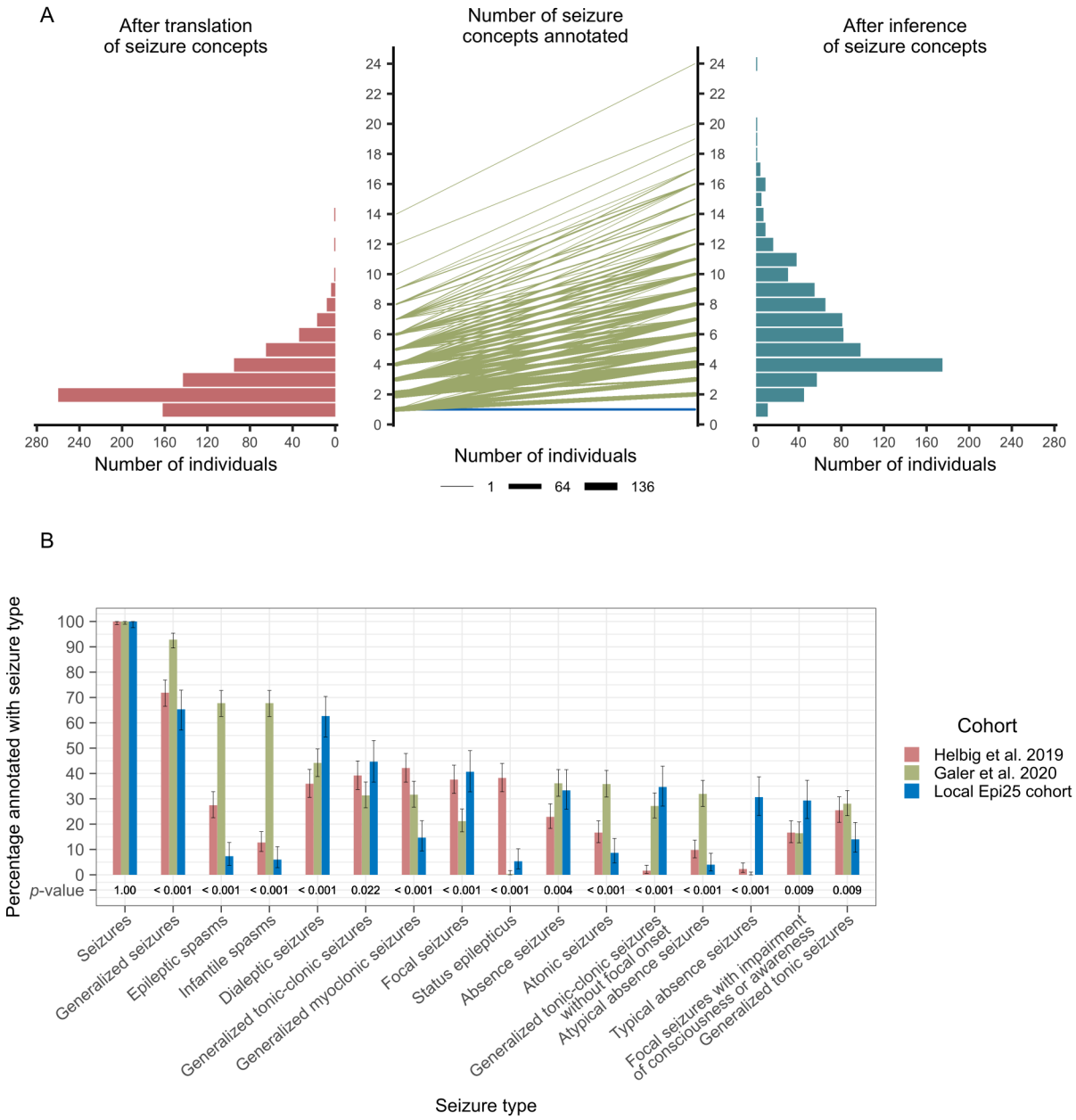


Figure 2-2. Increasing the number of concepts annotated to allow comparison of cohorts. (A) The increase in the number of seizure concepts by which data collected from 791 individuals can be described as a result of automated inference using the HPO 2017. Green lines indicate an increase in the number of seizure descriptors annotated by inference after translation. The blue line indicates eight individuals whose only HPO term translated from input data was *Seizure [HP:0001250]*, which is insufficient to infer descriptions. (B) Comparison of the three independent cohorts according to the percentage of individuals annotated with each of 16 HPO seizure terms, comprising the ten most common seizure concepts in each cohort. Note that in this HPO version *Generalized tonic-clonic seizure [HP:0002069]* is indifferent to onset. *p*-values are two-sided from Fisher's exact test with Holm-Bonferroni adjustment for 16 comparisons, 95 % standard error confidence intervals were calculated from the binomial distribution.

As a consequence of having harmonized data from these three independent cohorts onto the same conceptual model, we were able to compare them by the proportion of individuals

annotated with particular seizure description (**Figure 2-2B**). The relative abundance of seizure types in each cohort reflects the phenotypic focuses of each study. A greater proportion of individuals in the previously published DEE cohorts than in the local cohort had *Generalized myoclonic seizures [HP:0002123]* and *Generalized tonic seizures [HP:0010818]*. The Helbig 2019 cohort had a high burden of *Status epilepticus [HP:0002133]* and a similar prevalence of *Focal seizures [HP:0007359]* to the mixed cohort, but a low prevalence of *Generalized tonic-clonic seizures without focal onset [HP:0025190]*. *Epileptic spasms [HP:0011097]*, *Infantile spasms [HP:0012469]*, and *Atypical absences [HP:0007270]* were particularly frequently annotated in the Galer 2020 cohort. Conversely, more individuals in the mixed cohort had *Dialeptic seizures [HP:0011146]*, both *Typical absences [HP:0011147]* and *Focal seizures with impairment of awareness [HP:0002384]*.

2.4.2 Creation of a new subontology of contemporary seizure concepts for the HPO

At the time that the ILAE 2017 was published, the seizure subontology of the HPO lacked the lexicon to distinguish between some seizure types yet included redundant terms and poorly integrated concepts. For example, *Generalized tonic-clonic seizures with focal onset [HP:0007334]* (a focal to bilateral tonic-clonic seizure) was implied to be a type of generalized-onset rather than focal-onset seizure. Recognizing such existing discrepancies, we generated a new ontology of seizure types then mapped pre-existing HPO terms across to synonymous concepts where possible to allow historical HPO data to be mapped to modern concepts for reuse.

The core of the new ontology was based on the ILAE 2017 with subsequent incorporation of other classifications and commonly used concepts (**Table 2-1**). We selected preferred names for seizure concepts that describe their features concisely for visualization in ontology browsers such as hpo.jax.org. We generally favored terminology from ILAE classifications where applicable and followed precedence where not. However, we preferred the name ‘Bilateral tonic-clonic seizure with generalized onset’ to ‘Generalized tonic-clonic seizure’ to minimize misinterpretation by people who use the latter irrespective of onset, contrary to the ILAE 2017. HPO terms can be searched for by preferred name, identifier, or synonym. We provided ILAE and historical classification terminology as synonyms where necessary.

Domain	Reference
Seizure types and semiological descriptions	Operational classification of seizure types by the ILAE: Position Paper of the ILAE Commission for Classification and Terminology (Fisher *et al.*, 2017b). Instruction manual for the ILAE 2017 operational classification of seizure types (Fisher *et al.*, 2017a). EpilepsyDiagnosis.org (International League Against Epilepsy) Glossary of descriptive terminology for ictal semiology: Report of the ILAE task force on classification and terminology (Blume *et al.*, 2001). Semiological seizure classification (Luders *et al.*, 1998). Classification of paroxysmal events and the four-dimensional epilepsy classification system (Luders *et al.*, 2019).
Status epilepticus	A definition and classification of status epilepticus: Report of the ILAE Task Force on Classification of Status Epilepticus (Trinka *et al.*, 2015).
Neonatal seizure types	The ILAE classification of seizures and the epilepsies: Modification for seizures in the neonate. Position paper by the ILAE Task Force on Neonatal Seizures (Pressler *et al.*, 2021).
Febrile and mild gastroenteritis-associated seizures	Predictors of epilepsy in children who have experienced febrile seizures (Nelson and Ellenberg, 1976). Febrile Seizures: Clinical Practice Guideline for the Long-term Management of the Child With Simple Febrile Seizures (American Academy of Pediatrics, 2008). Generalized epilepsy with febrile seizures plus. A genetic disorder with heterogeneous clinical phenotypes (Scheffer and Berkovic, 1997). Clinical features of benign convulsions with mild gastroenteritis (Uemura *et al.*, 2002). Benign infantile seizures with mild gastroenteritis: Study of 22 patients (Caraballo *et al.*, 2009). Clinical outcome of recurrent afebrile seizures in children with benign convulsions associated with mild gastroenteritis (Chen *et al.*, 2018a).
Reflex seizures	A proposed diagnostic scheme for people with epileptic seizures and with epilepsy: Report of the ILAE Task Force on Classification and Terminology (Engel, 2001).

Table 2-1. Classifications and definitions consulted to generate the HPO 2020-12-7 seizure subontology.

While it is clinically important for classifications to include explicit terms to denote diagnostic uncertainty, such as unknown onset tonic-clonic seizure, this explicit uncertainty does not pertain to the seizure itself (the tonic-clonic seizure must either be generalized or focal in onset), rather to the clinician's ability to classify the seizure based on the available evidence: 'the term *unknown onset* is a placeholder – not a characteristic of the seizure, but of ignorance.' (Fisher *et al.*, 2017b). HPO terms represent concepts of phenotypes themselves, rather than the limitations of classification in a particular instance. Accordingly, within the HPO a seizure that is classified as unknown onset tonic-clonic seizure should be mapped to an HPO term that implies that this seizure is either a *Bilateral tonic-clonic seizure with generalized onset [HP:0025190]* or *Bilateral tonic-clonic seizure with focal onset [HP:0007334]* — generalized tonic-clonic seizure or focal to bilateral tonic-clonic seizure in ILAE 2017 — rather than a third, intrinsically different seizure type, such as an absence seizure. Hence, the seizure subontology includes additional terms such as *Bilateral tonic-clonic seizure [HP:0002069]* without specifying onset. This term applies to the ILAE 2017 concepts of unknown onset tonic-clonic seizure, generalized tonic-clonic seizure, and focal to bilateral tonic-clonic seizure. The onset-specific forms have their respective onset-specific *is_a* relationships to *Generalized-onset motor seizure [HP:0032677]* and *Focal-onset seizure [HP:0007359]*, and in addition have a second *is_a* relationship to this onset-agnostic term. Consequently, individuals with unknown onset tonic-clonic seizures are annotated with the onset-agnostic term in the absence of the onset-specific terms. This arrangement of concepts is analogous to the ILAE 2017 classification of focal-onset seizures,

which allows a particular instance of a focal-onset seizure to be classified by either awareness or motor versus non-motor onset independently without explicit reference to the other when necessary, rather than requiring all focal seizures to be classified according to both of these dimensions. Additionally, it allows optimal integration of the proposed classification of convulsive status epilepticus (Trinka *et al.*, 2015).

So-called *dual parentage* can also help to mitigate against misclassification in diagnostic decision support tools using HPO-based algorithms. In some circumstances, distinguishing between atypical absence seizures and focal impaired awareness seizures can be difficult, and non-specialists often mislabel focal impaired awareness seizures as 'absences' or even 'petit-mal'. The pre-existing HPO term *Dialeptic seizure [HP:0011146]* recognizes the semiological similarity between *Absence seizure [HP:0002121]* and *Focal impaired awareness seizure [HP:0002384]* (Luders *et al.*, 1998). Without a parent term in the HPO spanning electrographic onsets, these two concepts would appear no more similar to each than *Absence seizure [HP:0002121]* does to *Bilateral tonic-clonic seizure with focal onset [HP:0007334]*: their most specific common ancestor would be *Seizure [HP:0001250]*. Thus, were an individual with a focal impaired awareness seizure to be incorrectly annotated with *Absence seizure [HP:0002121]*, an algorithm comparing this to a reference of *Focal impaired awareness seizure [HP:0002384]* can use *fuzzy matching* to recognize that a correct (yet imprecise) match would have been made, were the individual to have been coded with a parent of their actual annotation: *Dialeptic seizure [HP:0011146]*.

We extended this strategy of creating dual parentage independent of electrographic onset beyond its clinical scope for mitigating uncertainty and misclassification to handle those semiologies that can occur with either focal or generalized onset. Researchers may be interested in identifying individuals whose seizures share semiological features independent of electrographic onset in order to investigate shared associations and mechanisms. Dual parentage allows the ontology to recognize that a focal tonic seizure is similar to other types of focal motor seizure but also that it is more similar to a generalized tonic seizure than it is to a generalized myoclonic seizure by introducing a second parent term, *Tonic seizure [HP:0032792]*, which is indifferent to onset. Users not wishing to use these, or other concepts need not, simply by overlooking them in their data dictionaries and sets of propagated terms.

A major component of the phenotypic bottleneck in clinical genetic diagnostics and research is data entry. We provide very specific seizure concepts and relationships that allow the ontology to infer as much information as possible about the seizures from the provision of the minimum number of input terms. For example, on provision of the single highly specific term *Focal aware*

cognitive seizure with deja vu/jamais vu [HP:0032883] a reasoning algorithm can use the ontology to infer the applicability of 12 other ontological seizure descriptors capturing its focal onset, cognitive ictal manifestation, and the retention of awareness throughout in different combinations and at different degrees of granularity.

Pre-existing HPO terms were retired only if they could not be mapped among current seizure concepts or were redundant (**Table 2-2**). Otherwise, terms were integrated into the new ontology, with modification of their definition or relationships if required, or as a synonym for a new term.

2.4.3 The 2020 HPO seizure subontology captures more seizure information than the 2017 version

After completing our integration of the new seizure subontology into the HPO release 2020-12-7 (HPO 2020), we compared this to the version released 2017-12-12 (HPO 2017) prior to commencement of our revision according to the number of terms by which seizures could be described and the relationships between these concepts (**Table 2-3**). The HPO 2020 contained 5-fold as many terms and 7-fold as many relationships.

Seizure term	Reason retired
Absence seizures with special features [HP:0011148]	While this concept has been suggested previously to group (i) myoclonic absences with (ii) eyelid myoclonia (Berg *et al.*, 2010), it has not been included in the most recent ILAE classification (Fisher *et al.*, 2017b).
Focal autonomic seizures with altered responsiveness [HP:0011155] Focal autonomic seizures without altered responsiveness [HP:0011156]	The ILAE recommends classification of focal seizures by the presence or absence of retained awareness throughout the seizure (Fisher *et al.*, 2017b). While similar concepts, responsiveness is not synonymous with awareness. These two terms, which combine autonomic and responsiveness descriptors, were thought likely to confuse, or distract HPO users from the preferred classification concepts.
Psychic auras [HP:0011162] Vegetative auras [HP:0011164]	These terms were retired when creating the backbone of the new subontology using current ILAE concepts (Fisher *et al.*, 2017b). Subsequently, we reintroduced them as Four-dimensional concepts (Luders *et al.*, 2019). Psychic epileptic aura [HP:0033346] is a type of Epileptic aura [HP:0033348]. Vegetative aura is subsumed by Focal autonomic seizure [HP:0011154].
Simple partial occipital seizures [HP:0025121]	The ILAE classifies seizure types by clinical manifestations rather than anatomical localization (Fisher *et al.*, 2017b), which would require assumptions about the symptomatogenic or epileptogenic zone and/or integration of electroencephalographic and imaging evidence, the availability and interpretation of which may inconsistent. For the encoding of data precisely and retaining the independence of clinical semiology, electrographic and brain morphological findings (which have independent subontologies within the HPO), it is preferable to avoid a separate seizure localization subontology. Ontologies that model localization of seizures are discussed in the manuscript discussion.
Segmental myoclonic seizures [HP:0025191]	This concept has been subsumed by the current formal ILAE concept of a focal myoclonic seizure (Fisher *et al.*, 2017b).
Focal seizures, afebrile [HP:0040168]	This concept is broad and applies to the majority of focal seizures. Yet it had no child terms (risking superficial coding if used in isolation) and there was no corresponding term for febrile focal seizures (or afebrile generalized seizures). Febrile seizures are usually classified as simple or complex (Nelson and Ellenberg, 1976). The majority of seizure phenotypes are afebrile, and the revised subontology assumes coded seizures to be afebrile unless explicitly specified as febrile. This is consistent with the usual use of seizure terminology and allows the majority of seizure phenotypes to be coded efficiently with a single term.

Table 2-2. HPO 2017 seizure terms retired or subsumed in the HPO 2020.

Applying this, we harmonized categorical data from our local cohort and HPO 2017 data from the previously published DEE cohorts into the HPO 2020 subontology then compared the annotated seizure data from the combined cohort of 791 individuals by HPO version (**Table 2-3**). Several new terms depict novel concepts, and thus were not captured by the HPO 2017. *Focal-onset seizures [HP:0007359]* appear more common and *Generalized-onset seizures [HP:0002197]* appear less so in the HPO 2020. This is a result of revised relationships to make *Bilateral tonic-clonic seizure with focal onset [HP:0007334]* a form of *Focal-onset seizure [HP:0007359]* rather than *Generalized-onset seizure [HP:0002197]*. The increase in frequency of *Bilateral tonic-clonic seizure [HP:0002069]*, which is indifferent to onset, results from the addition of a specific term for *Convulsive status epilepticus [HP:0032660]* as a prolonged form of the former.

Phenotypic depth can be conceptualized as the amount of information provided about an individual's or cohort's phenotypes. We compared the depth of phenotypic information about our combined cohort encoded using each seizure subontology (**Table 2-4**). The number of seizure terms annotated to an individual after translation is a measure of the concepts within the raw data that can be captured by the ontology without inference. The number of terms annotated after propagation measures the total number of potential ways that the seizures in the raw data can be described, or categorized, by the ontology after inference. We found a modest detectable increase in annotations to the cohort after translation: the HPO 2020 increased the number of annotations in 46 individuals and reduced the number of annotations in 13. Compared to the HPO 2017, the HPO 2020 increased the number of propagated annotations in the combined cohort by 40 %. The median increase in the number of propagated terms per individual was 50 % (**Table 2-4** and **Figure 2-3A**). However, the number of annotations does not distinguish between common and rare features of seizures. Types of seizure that are encountered more rarely are intuitively of greater value for distinguishing an individual from the rest of the cohort than common types of seizure. In other words, the presence of a rare seizure type is more informative than that of a common seizure type. In information theory this can be quantified as information content (IC), usually measured in bits and defined as the negative logarithm of their frequency in the cohort, $IC = -\log_2 f$. This gives annotations of rare seizure terms more bits of information than those which are more common. Because annotations are propagated from conceptually narrow to broader parent terms, IC tends to increase as one goes from broad to more detailed terms and thus is a good measure of specificity.

Comparison		HPO subontology	
		2017 release	2020 release
Metadata of the seizure subontologies			
Number of seizure terms		71	347
Number of is_a relationships between two seizure terms and		73	518
between a seizure and non-seizure term		3	4
Frequency of different seizure concepts captured from the cohort of 791 individuals			
Term identifier	**Seizure descriptor from 2020 release**	**Frequency in 2017 release**	**Frequency in 2020 release**
HP:0001250	Seizure	1	1
HP:0020219	Motor seizure	NA	0.8161
HP:0002197	Generalized-onset seizure	0.8033	0.7395
HP:0032677	Generalized-onset motor seizure	NA	0.5045
HP:0011146	Dialeptic seizure	0.4496	0.4496
HP:0011097	Epileptic spasm	0.4112	0.4112
HP:0002069	Bilateral tonic-clonic seizure	0.3729	0.3742
HP:0012469	Infantile spasms	0.3512	0.3512
HP:0033259	Non-motor seizure	NA	0.3384
HP:0007359	Focal-onset seizure	0.3155	0.3244
HP:0032794	Myoclonic seizure	NA	0.318
HP:0002123	Generalized myoclonic seizure	0.3282	0.3167
HP:0002121	Generalized non-motor (absence) seizure	0.3078	0.3078
HP:0032792	Tonic seizure	NA	0.2644
HP:0010818	Generalized tonic seizure	0.2465	0.2465
HP:0010819	Atonic seizure	0.235	0.235
HP:0002384	Focal impaired awareness seizure	0.1916	0.1916
HP:0025190	Bilateral tonic-clonic seizure with generalized onset	0.189	0.189
HP:0007270	Atypical absence seizure	0.1826	0.1814
HP:0002133	Status epilepticus	0.1609	0.1609
HP:0011153	Focal motor seizure	0.1328	0.1328
HP:0032892	Infection-related seizure	NA	0.1315
HP:0032894	Seizure precipitated by febrile infection	NA	0.1315
HP:0002373	Febrile seizure (within age range 3 months - 6 years)	0.1239	0.1239
HP:0020221	Clonic seizure	NA	0.0881
HP:0007334	Bilateral tonic-clonic seizure with focal onset	0.0677	0.0677
HP:0011147	Typical absence seizure	0.0677	0.0677
HP:0002266	Focal clonic seizure	0.0639	0.0639
HP:0011170	Generalized myoclonic-atonic seizure	0.0498	0.0498
HP:0032679	Focal non-motor seizure	NA	0.0447
HP:0011175	Focal motor seizure with version	0.0345	0.0345
HP:0002349	Focal aware seizure	0.0153	0.0307

Table 2-3. Description of the HPO 2020 and the HPO 2017 seizure subontologies and the frequency of the 32 most frequent seizure concepts in the combined cohort of 791 individuals according to each. Terms with frequency 'NA' were not conceptualized.

Comparison	HPO subontology		Difference (% change or median, *p*-value)
	2017 release	2020 release	
Number of input annotations			
Entire cohort (total)	2,275	2,308	Increase by 1.5 %
Per individual (median)	2	2	0, $p < 0.001$
Number of annotations after propagation (inference)			
Entire cohort (total)	5029	7063	Increase by 40 %
Per individual (median)	6	8	3, $p < 0.001$
Total information content (bits)			
Entire cohort (total)	8,280	11,403	Increase by 38 %
Per individual (median)	7.90	11.43	3.15, $p < 0.001$

Table 2-4. Comparison of the amount of phenotypic information captured about the seizures of 791 individuals using the HPO 2017 and the HPO 2020. Differences between total cohort measures are given as percentages of the HPO 2017 figures. Paired differences between individuals are given as median absolute differences, *p*-values are two-sided from a Wilcoxon signed rank test.

Accordingly, after propagation one can measure the true frequency of each seizure term and calculate the IC of each term annotated in a cohort. The total amount of information captured by the ontology describing an individual's seizure repertoire can be calculated as the sum of the IC of each of their seizure annotations. We found that the total amount of information about the cohort's seizures increased by 38 % when using HPO 2020 (**Figure 2-3B**). The information about each individual's seizures increased for 686 individuals and decreased for 86, demonstrating a large effect size ($|z/\sqrt{n}| = 0.80$). While there was a small reduction in the quantity of information about 86 individuals, this was attributable to the changes described above, correcting the relationships of the concept of *Bilateral tonic-clonic seizure with focal onset [HP:0007334]* and introducing *Convulsive status epilepticus [HP:0032660]* as a type of *Bilateral tonic-clonic seizure [HP:0002069]*. Hence, despite a reduction in quantity of information in a minority of individuals, the quality of information captured by HPO 2020 is greater.

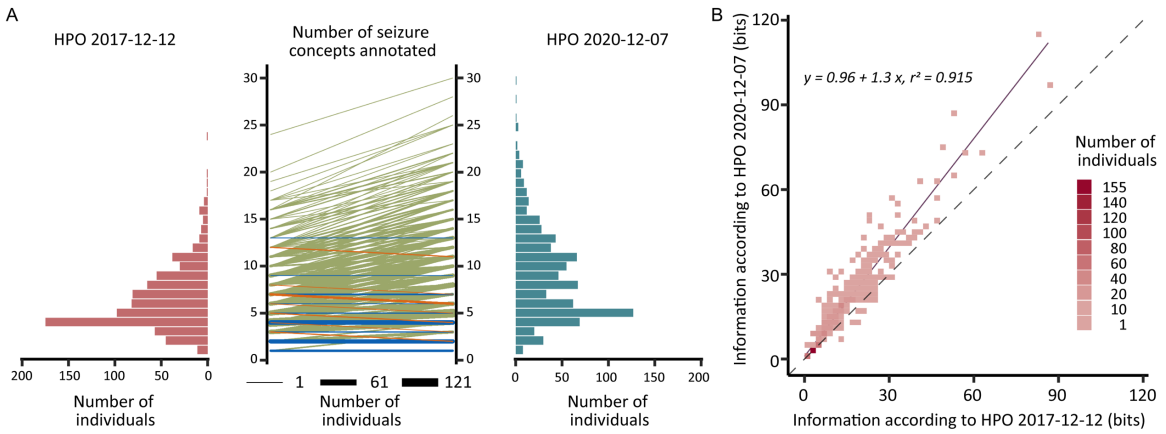


Figure 2-3. Comparison of the phenotypic information encoded from the same data from 791 individuals using the two seizure subontologies. (A) Green lines indicate an increase, red lines a decrease, and blue lines no difference in the number of seizure descriptors annotated after inference. (B) The total amount of information encoded about each individual's seizure types according to each HPO seizure subontology. Numbers of individuals shown correspond to those falling within 2 bit by 2 bit bins, the dashed grey line indicates equality, the regression line is shown in purple.

2.5 Discussion

We describe the revision of the seizure domain within the HPO to align it with contemporary classification concepts, and the resulting effect on precise seizure description in 791 individuals. We demonstrate how seizure data can be encoded in HPO terms, automatically classifying each seizure according to multiple conceptually broader descriptors, and how it can harmonize data provided in different formats. We show that compared to a previous iteration, the new subontology increased the amount of information that could be encoded in HPO terms about an individual's seizure repertoire thereby facilitating combination with other data sets, or analysis, based on automated inference of seizure concepts. Our newly developed seizure subontology is fully implemented in the official HPO release as of 2020-12-7 and freely available at hpo.jax.org.

The new subontology encompasses the breadth of epileptic seizure types encountered in clinical practice, taking evidence from formal classifications, consensus statements, and pertinent peer-reviewed literature describing particular seizure types that define clinical entities. It is not a clinical classification and serves a different purpose: to facilitate automated inference of clinical seizure data for harmonized analysis or comparison. While we did not amend classification concepts themselves, we used ontological approaches to retain as much information as possible about seizures that can be classified incompletely. This allows semiological descriptions of seizures in medical records that do not specify electroclinical onset

(for example, ‘she continues to have tonic seizures’) to be coded without assuming the electrographic onset or losing all description of the clinical manifestation. Similarly, the subontology includes semiological terms that can code seizures without 80 % certainty of onset (Fisher *et al.*, 2017b). For example, the term *Non-motor seizure [HP:0033259]* may be relevant for people with epilepsy who are unable to disclose sufficient subjective experience to confirm a specific type of focal non-motor seizure, and in whom insufficient supporting evidence can be obtained to differentiate between a focal and absence seizure, thus including the ILAE 2017 concept of an unknown onset behavior arrest seizure. The ability to code seizures informatively according to only a limited patient or witness history optimizes harmonization of data emanating from health care settings with limited access to resources such as video-electroencephalography.

The HPO continues to include older concepts such as *Epileptic aura [HP:0033348]* and *Dialeptic seizure [HP:0011146]* that some epileptologists find useful, and we have integrated these with ILAE 2017 concepts where possible (Luders *et al.*, 1998; Blume *et al.*, 2001; Fisher *et al.*, 2017a; Luders *et al.*, 2019; Rosenow *et al.*, 2020). This maximizes the precision of information from historical data that can be harmonized with current ILAE concepts using the HPO.

A priority when developing the new subontology was to try to make the encoding of seizure types as efficient as possible to reduce the burden of manual data entry and dictionary design. Single input terms allow harmonization to seizure types and descriptors at different levels of detail and also unambiguously bind features such as level of awareness and initial manifestation together for a particular instance of a seizure. An exception to this is the concept of a focal to bilateral tonic-clonic seizure, which the ILAE 2017 considers a propagation pattern rather than unitary seizure type (Fisher *et al.*, 2017a). Similarly, we have represented this concept as a single independent term, *Bilateral tonic-clonic seizure with focal onset [HP:0007334]*, which avoids making the number of terms in the subontology unwieldy.

Consistent with the ILAE 2017, the structure of the new HPO subontology was based on electrographic onset and prioritization of initial manifestations. An ontology-based clinical diagnostic support tool using ILAE 2017 terms to suggest syndromic or genetic diagnoses has been developed (Chiang *et al.*, 2020). While the list of seizure concepts in this ontology was published, the relationships between specific concepts are not openly available for comparison with our subontology.

The semiological evolution of a focal seizure adds granularity that is important for mapping its spread across eloquent cortex and can be recorded using descriptors from the ILAE 2017 (Fisher *et al.*, 2017a), ILAE glossary (Blume *et al.*, 2001), or a sequence of components within

the semiology classification of the Four-dimensional Epilepsy Classification (Luders *et al.*, 1998). Beyond loss of awareness and bilateral propagation, different semiological evolutions of the same unitary focal seizure type are indistinguishable in the HPO. This may be insufficient for detailed surgical and network analyses. However, the HPO is predominantly a tool of genomics, and phenotypic analysis in epilepsy genetics has focused on seizure types rather than differences in the evolution of seizures of the same type. The Epilepsy Syndrome Seizure Ontology (version 1.03, bioportal.bioontology.org/ontologies/ESSO) and Epilepsy and Seizure Ontology (version 1.0, bioportal.bioontology.org/ontologies/EPSO) predate the ILAE 2017 and their multidimensional representation of seizures (aligned to the Four-dimensional Epilepsy Classification) may be better suited to such tasks (Sahoo *et al.*, 2014; Luders *et al.*, 2019). Another ontological approach to semiology prior to ILAE 2017 has shown that analysis of combinations of semiological features can be used for classification of temporal lobe and extra-temporal lobe epilepsies (Kassahun *et al.*, 2014). Advances in understanding and classification of focal seizures may facilitate prioritization of important sequences of semiological components for inclusion in the HPO.

Currently, as in this study, HPO terms are usually annotated manually on review of the medical records, or through bulk translation of structured categorical data from research or clinical genetic clinical data forms. However, the burden of HPO coding will decrease as the availability of EMR systems with integrated HPO annotation modules increases, and with advances in natural language processing.

While this coordinated revision of the HPO seizure subontology has improved the integration of its terms and its consistency with current ILAE classifications, scope remains for refinement, including integration of concepts from other classifications to maximize its value for harmonization of non-ILAE 2017 data, such as that from historical cohorts. Continuing expert engagement (we encourage suggestions via https://github.com/obophenotype/human-phenotype-ontology/issues) will lead to incremental improvements to HPO domains pertinent to epilepsy that will benefit clinical diagnostic algorithms and research for people with epilepsy over years to come.

Chapter 3. Computational analysis of neurodevelopmental phenotypes — harmonization empowers clinical discovery

3.1 Preface

This chapter is a modified reproduction of a published review article of methods and the state of computational analysis of phenotypes in neurodevelopmental disorders, serving as an introduction to the application of the HPO and other computational tools to the analysis of epilepsy data.

Lewis-Smith, D., Parthasarathy, S., Xian, J., Kaufman, M.C., Ganesan, S., Galer, P.D., Thomas, R.H. and Helbig, I. (2022) 'Computational analysis of neurodevelopmental phenotypes: Harmonization empowers clinical discovery', Hum Mutat*, 43(11), pp. 1642-1658.*

I drafted and decided on the content of the manuscript. co-authors assisting with the drafting of new, or reproduction of previously published figures (S.P., J.X., M.C., S.G., P.D.G.), provided supervision and advice (R.H.T., I.H.), and all co-authors were invited to provide comments on the manuscript. I thank *Human Mutation's* reviewers and editors (Prof. Mark H. Paalman, Prof. Nancy Spinner, and Dr Kai Wang) for their comments. This reproduction is based on the author accepted manuscript, available through PubMed Central, I have also obtained permission from John Wiley and Sons for its inclusion in this thesis.

3.2 Abstract

Making a specific diagnosis in neurodevelopmental disorders is traditionally based on recognizing clinical features of a distinct syndrome, which guides testing of its possible genetic etiologies. Scalable frameworks for genomic diagnostics, however, have struggled to integrate meaningful measurements of clinical phenotypic features. While standardization has enabled generation and interpretation of genomic data for clinical diagnostics at unprecedented scale, making the equivalent breakthrough for clinical data has proven challenging. However, increasingly clinical features are being recorded using controlled dictionaries with machine readable formats such as the HPO, which greatly facilitates their use in the diagnostic space. Improving the tractability of large-scale clinical information will present new opportunities to inform genomic research and diagnostics from a clinical perspective. Here, we describe novel approaches for computational phenotyping to harmonize clinical features, improve data translation through revising domain-specific dictionaries, quantify phenotypic features, and determine clinical relatedness. We demonstrate how these concepts can be applied to longitudinal phenotypic information, which represents a critical element of developmental

disorders and pediatric conditions. Finally, we expand our discussion to clinical data derived from EMR, a largely untapped resource of deep clinical information with distinct strengths and weaknesses.

3.3 Introduction

The critical role of deep, reliable clinical information in interpreting genetic findings is beyond doubt. The traditional approach to define disease entities in clinical medicine is based on recognizing shared clinical features in individuals, ranging from constellations of symptoms and signs to laboratory values, imaging abnormalities, and physiological findings. Furthermore, clinical presentations are not static but dynamic, with diseases differing by their clinical trajectories. In the history of neurogenetics, delineating phenotypic features in detail and identifying such distinctive patterns has also been the route to identifying monogenic etiologies. A classic example of this approach is the initial discovery of Rett Syndrome. Dr Andreas Rett recognized that two girls in his pediatric practice frequently exhibited unusual handwashing motions and reasoned that this stereotypy, in combination with clinical features such as regression and lack of speech development, may represent a novel clinical syndrome (Rett, 1966; Hagberg *et al.*, 1983). This insight eventually led to the delineation of Rett Syndrome and the identification of *MECP2* as the causative genetic etiology (Amir *et al.*, 1999).

Sometimes it is the age at which a clinical feature manifests in the context of typical developmental or aging trajectories, rather than the feature itself, that is sufficiently distinctive to identify a monogenic disorder. For example, Genetic Epilepsy with Febrile Seizures Plus is a mendelian disorder most frequently attributable to variants in *SCN1A*, which encodes a voltage-gated sodium channel (Escayg *et al.*, 2000). In this case, the occurrence of febrile seizures, which occur in up to 5 % of preschool children, outside of this usual age range serves as a strong indicator.

Sometimes it is the characteristic combination of multiple phenotypic features, and their sequence, which has led to successful gene discovery. Dravet Syndrome, initially described by Charlotte Dravet as Severe Myoclonic Epilepsy in Infancy, represents the most prominent example in pediatric epileptology. Also caused by pathogenic variants in *SCN1A* (Claes *et al.*, 2001), Dravet Syndrome is characterized by typical development in the first six months of life, followed by febrile seizures, often febrile status epilepticus, and hemiclonic seizures (Dravet, 2011). In isolation, few of these features would necessarily raise high suspicion for Dravet Syndrome, emphasizing how it is often the overall clinical ‘gestalt’ of a genetic disease rather than a single tell-tale clinical feature that is distinctive. This is especially the case in

neurogenetics, where clinical manifestations are complex combinations of features that may be interwoven with childhood development or aging.

The historical and examination skills of a master clinician consist of correctly eliciting and interpreting clinical features in context, ideally on a background of encyclopedic clinical knowledge. However, what exactly does it mean to recognize clinical features and to have the quintessential epiphany of recognizing a typical genetic syndrome — particularly when individuals do not manifest all classical features and many present with atypical clinical features? Trainee clinicians can be taught Bayesian reasoning, yet the formal application of this to rare diseases is challenging in clinical practice. First, beyond exclusion of the more frequently encountered conditions such as Trisomy 21, 22q Deletion Syndrome, Fragile X Syndrome, and Angelman Syndrome, the baseline prior probability of a patient having some disease (prevalence of the disease), marginal probability (prevalence of some clinical feature in the general population), and likelihood (the proportion of people with the disease who have the clinical feature) are often insufficiently determined for differential diagnoses to be compared accurately. Take the diagnosis of tuberous sclerosis as a relatively common and well-studied example, in which diagnosis enables genetically guided surveillance and treatment. Current estimates of the incidence of tuberous sclerosis range from 1:6,000 to 1:10,000 live births and approximately 5 % of cases cannot yet be explained by variants in *TSC1* or *TSC2* (Northrup *et al.*, 2021). If you have identified an elliptical hypopigmented skin lesion that you think is an ash leaf spot, how much more likely does this make it that the 2-year-old child before you carries a diagnostic variant in *TSC1* or *TSC2*, and how much should you modify your prediction based on the absence of developmental delay but the presence of seizures at this age? In practice, the cognitive processes that contribute to clinical pattern recognition for genetic diagnosis are not well-defined but heuristic. What are the informative clinical features, how informative are they, and when is a combination of shared clinical features distinctive? Formalizing these concepts for big-data computational approaches in the age of high-throughput genomics and clinical data repositories would allow us to generate the reference data and frameworks needed to incorporate detailed phenotypic perspectives into personalized genomic medicine.

Within this review, we will revisit these concepts and discuss how they can be described or approximated using computational phenotyping approaches, which analyze clinical data statistically at scale while retaining as much phenotypic information as possible to empower discovery and maximize interpretability for translation to the care of individual patients. The pressing need to develop these analytic approaches is reflected in the current paucity of

phenotypic resolution of high-throughput genomic studies, which limits the clinical interpretability of advances for personalized genomic medicine. While the last two decades have led to paradigm-changing approaches in genomic sequencing that can be deployed at large scale with sophisticated quality control pipelines that empower interpretation, clinical phenotyping often remains a manual, non-scalable task. Accordingly, in the past, most genomic datasets had little or no relevant phenotypic information attached, often limited to diagnostic codes for disorders, for example, from the International Classification of Disease identifiers (World Health Organization, 1992), or to descriptive names (sometimes with a subtype number) such as those referred to as 'phenotypes' and 'phenotypic series' in OMIM (Bodenreider, 2004). However, the last 5 years have seen a renewed interest in integrating detailed clinical data into genomic approaches — aiming to provide phenotypes at the resolution of single clinical features (such as 'focal-onset motor seizures') rather than disease labels (such as 'epilepsy'). This has led to novel gene discoveries using large biorepositories. For example, the discovery of recessive genetic etiologies such as *KIAA0586*, *HACE1*, *PRMT7*, or *MMP2* in the Deciphering Developmental Disorders project was achieved through a statistical framework combining genomic and phenotypic evidence (Akawi *et al.*, 2015). Even regarding established disease genes, questions remain about the full clinical spectrum, genotype-specific phenotypic associations and treatment responses, and whether 'lumping or splitting' would be more meaningful for biological understanding and, more importantly, for clinical care. Large research repositories and routine clinical records that link detailed clinical and genomic data are likely to contain many answers. Yet if we are to exploit the opportunities these resources offer, we must address the challenge of making phenotypic data — typically highly dimensional and heterogeneous — tractable, without compromising its content.

Given the need for approaches to fill the widening gap between the increasing amount of genomic information and the limited bandwidth of traditional clinical studies, novel frameworks have been developed to overcome the phenotyping bottleneck. The HPO, a controlled dictionary of more than 16,000 clinical concepts, is the most widely adopted model used to capture people's clinical features in a format that is both human and machine readable (Core Human Phenotype Ontology/Monarch Initiative Team; Robinson *et al.*, 2008; Kohler *et al.*, 2014; Kohler *et al.*, 2017; Kohler *et al.*, 2021). The HPO is adapted and used by diagnostic laboratories, the National Institutes of Health (NIH) Undiagnosed Disease Network (Gahl *et al.*, 2016), the 100,000 Genomes Project (Turro *et al.*, 2020) and NHS England Genomic Medicine Service (NHS Health Education England, 2020), and a large number of other initiatives around the globe. The HPO has emerged as the *lingua franca* for phenotypic features and is appealing through its simplicity, which encourages clinical domain experts to engage in

its revision and expansion, with updated versions released several times per year. Our team has already mapped clinical data from major epilepsy genetics projects, including the Epi4K project, the EuroEPINOMICS project, and the Epi25 project, as well from our local EMR into HPO annotations, demonstrating the versatility of this framework to harmonize clinical information from heterogeneous data sources (Helbig *et al.*, 2019; Galer *et al.*, 2020; Ganesan *et al.*, 2020; Crawford *et al.*, 2021; Lewis-Smith D *et al.*, 2021; Xian *et al.*, 2022). In summary, the HPO provides a generally accepted framework that can be used to harmonize clinical data and has emerged as one of the major vehicles to make large-scale phenotypic data available for analytic approaches.

In the following sections, we discuss insights from the emerging fields of computational phenotyping and EMR genomics, identifying possibilities and pitfalls as well as highlighting novel paradigms that exploiting these approaches. We cover data harmonization, IC, phenotypic depth, the value of annotating negative (explicitly absent) phenotypes, and concepts relevant to the interpretation of longitudinal phenotypic data when studying clinical trajectories and outcomes.

3.4 Biomedical ontologies capture the relationship between clinical annotations

Mapping clinical information to structured dictionaries such as the HPO makes clinical information available for computational analyses harmonized in both human- (for example, *Seizure*) and machine-readable (for example, *[HP:0001250]*) form. While human-readable names may evolve with clinical preference, their machine-readable identifiers should correspond to consistent concepts, future-proofing data so that they can be reinterpreted using subsequent versions of the HPO. The HPO not only represents clinical features as a dictionary, but it arranges these concepts a directed acyclic graph that provides information about *is_a* relationships between phenotypic features (**Figure 3-1**).

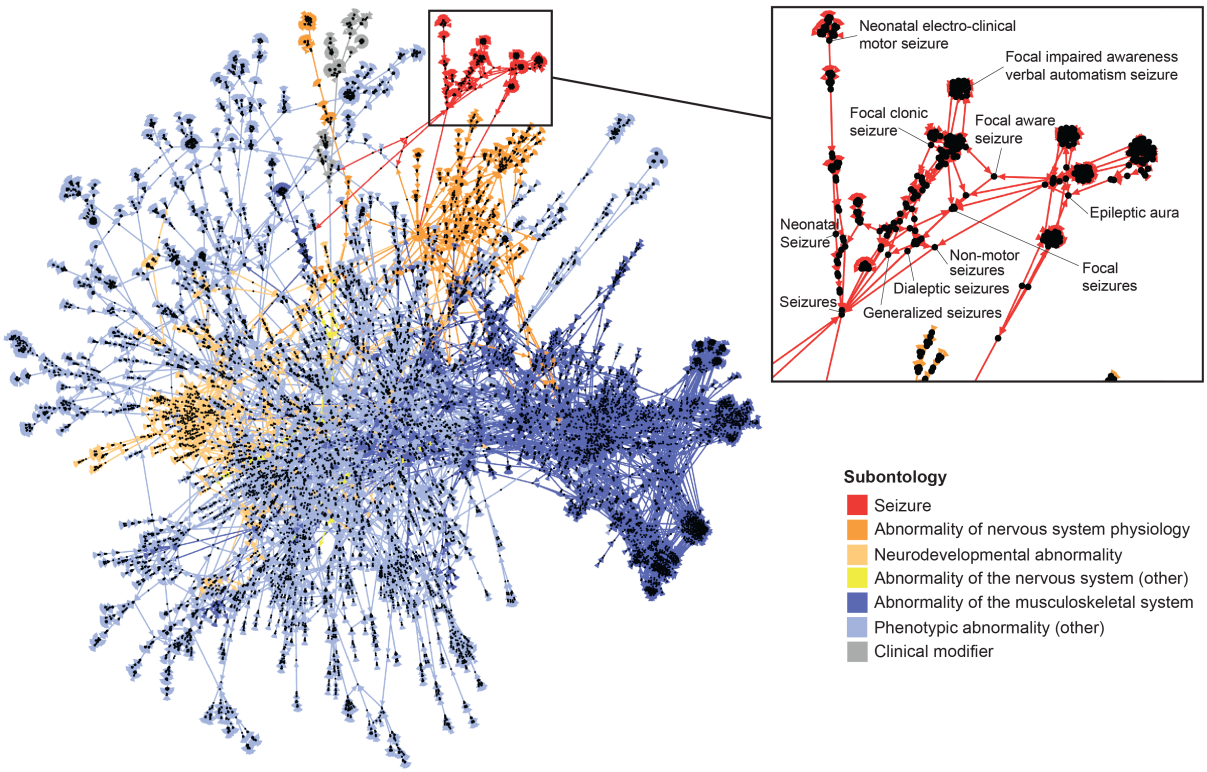


Figure 3-1. A visual representation of the complexity of HPO version 1.7.13 released 2021-10-10, focusing on seizure and related neurological phenotypes. This version contains 16,290 terms and 20,529 *is_a* relationships.

3.5 Inference allows for harmonization across the depth of all phenotypes

When raw clinical data are translated into HPO annotations, they become harmonized in that equivalent concepts are unified using a controlled vocabulary. For example, synonyms 'generalized tonic-clonic seizure', 'grand mal', and 'generalized convulsion' might be unified under the HPO concept of *Bilateral tonic-clonic seizure with generalized onset [HP:0025190]*. This reduces arbitrary heterogeneity in the data. However, a further step of harmonization — this time across phenotypic depth — allows better estimation of the true frequency of a clinical feature in the cohort. For example, generic phenotypic descriptors represented by higher-level HPO terms such as *Abnormality of digestive system morphology [HP:0025033]* or *Abnormality of the Nervous System [HP:0000707]* are unlikely to be translated from raw data because clinicians would tend to document the clinical features more precisely than this. However, many specific clinical terms imply the presence of higher-level concepts via the *is_a* relationships of the HPO. These implied terms are *ancestors* of the terms translated from the text, which we call *base terms*. For instance, the applicability of the two high-level HPO terms above to an individual should be inferred from the presence of annotations *Pyloric stenosis [HP:0002021]* (narrowing of the outflow from the stomach) and *Epileptic spasm [HP:0011097]* (a type of motor seizure most

frequently occurring in clusters during infancy) respectively, even if these high-level terms were not present among those translated from the raw data. We refer to the process of inferring more generic, higher-level phenotypic terms (and adding these to an individual's annotations that have been translated from the raw data) as propagation. Accordingly, sets of patients' HPO annotations are referred to as base annotations if they are those translated from raw data and propagated annotations if these base terms have been supplemented by all applicable terms inferred according to the *is_a* relationships of the HPO (**Figure 3-2A**).

Estimates of the frequency of more generic phenotypic concepts in a cohort are more accurate and internally consistent after propagation. For example, in a study of 413 people with *SCN2A*-related disorders, translation of raw phenotypic data sources assigned *Neurodevelopmental abnormality [HP:0012759]* to only 0.2 % of the cohort, whereas counterintuitively *Intellectual disability [HP:0001249]* (a type of neurodevelopmental abnormality) was annotated to 7 %. However, after propagation, the frequency of *Neurodevelopmental abnormality [HP:0012759]* increased to 63 % and the frequency of *Intellectual disability [HP:0001249]* to 46 % as subtypes of intellectual disability were present among the base annotations of many participants. This is now consistent with the logical intuition that a generic phenotypic description should not be applicable to fewer individuals than one of its subtypes, and precise descriptions should be more informative than generic descriptions of the same phenotypic feature.

People can be compared by their propagated annotations, as they will have some in common (the intersection of their propagated annotation sets), even if generic (such as *Phenotypic abnormality [HP:0000118]*, or even the HPO root: *All [HP:0000001]*). Conversely, two people may not share any base terms and an automatic reasoner without knowledge of epilepsy would not be able to quantify the similarity further than recognizing that they have disjoint sets of annotations, even if these are conceptually closely related. For example, consider four people whose raw data are translated into four disjoint base annotations (**Figure 3-2B**). After propagation we see that they all share a generic predisposition to Seizure [HP:0001250]. Alan and Barbara have further similarity because they share types of Focal-onset seizure [HP:0007359], and Barbara's seizure type is also a Tonic seizure [HP:0032792] identifying a similarity to Charlotte. Charlotte's is not of Focal [HP:0007359] but of Generalized-onset [HP:0002197], identifying a similarity to Dennis. Similarly, this concept can be applied to comparison of individuals by sets of multiple HPO terms (**Figure 3-2C**).

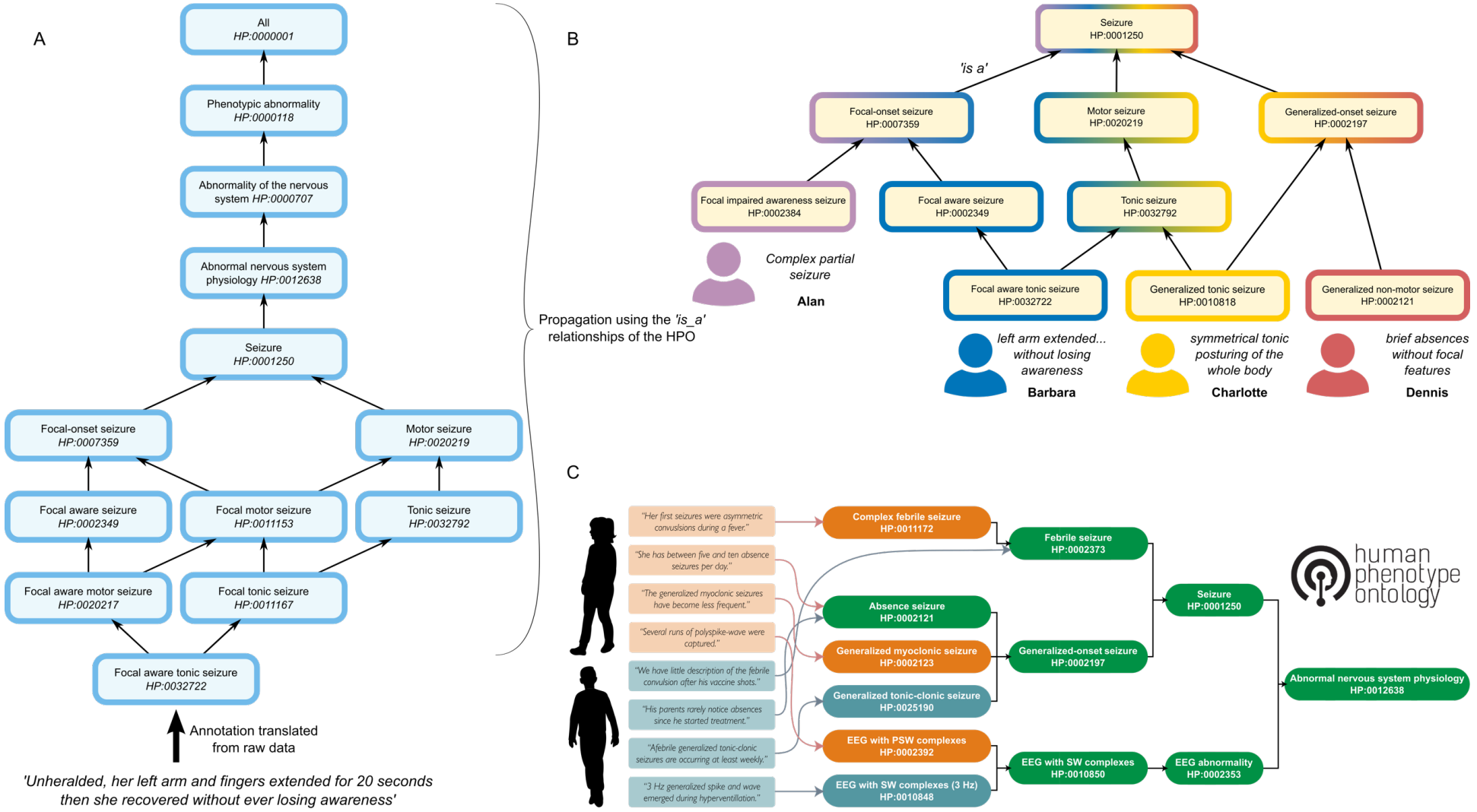


Figure 3-2. Interpretation using the *is_a* relationships of the HPO. (A) Translation of raw clinical data typically results in precise phenotypic annotations that should be propagated to infer the presence of less conceptually specific phenotypic concepts. (B) A simplified subontology of the HPO, comparing four individuals, each annotated with a single phenotypic term. (C) Individuals can be compared according to sets of HPO annotations.

3.6 Clinical phenotypic annotations translated from detailed raw data are sparse

While all clinical features can be linked through such an approach (with the ultimate root concept *All [HP:0000001]*), some aspects of clinical information require dedicated consideration. Inherent features of clinical datasets typically become obvious once translation has been performed and some of the properties in clinical data are often surprising and counterintuitive.

First, a frequently underappreciated feature of large clinical datasets is sparseness, where many phenotypic terms are only assigned to a small number of individuals. For example, in a study of 846 people with developmental and epileptic encephalopathies, we found that despite propagation, 679 of 1,616 terms were annotated to only one individual and only 331 of 1,616 terms were annotated to > 10 % of the cohort (Galer *et al.*, 2020). Second, the available clinical information per individual varies significantly within and between datasets. When comparing seizure descriptions regarding 791 individuals from three cohorts, we found a median of 8 clinical annotations per individual after propagation (Lewis-Smith *et al.*, 2021a). However, the range of annotations per individual after propagation extended from 1–30, emphasizing that available clinical information per individual is not uniform across cohorts, but widely distributed. Third, availability of standardized clinical information allows for operational definitions of concepts such as phenotypic depth. However, the number of clinical annotations is only one method of quantifying the available information and needs to be assessed alongside other measures. For example, we have used IC per clinical annotation, calculated as $-\log_2$ of the frequency with which the given term is annotated in the cohort after propagation and measured in bits. Consequently, just as the frequency with which terms are annotated should increase monotonically as one follows the *is_a* relationships towards the root of the HPO, the IC should decrease monotonically. The result of this is that conceptually specific phenotypic features tend to have a low frequency and high IC (they are more informative), while more generic phenotypic features have a high frequency with correspondingly low IC (less informative). This captures how clinicians typically not only consider the breadth of a constellation of clinical phenotypes but also weigh each element by its individual specificity.

The amount of information per individual can be summed across all their propagated phenotypic annotations, providing a measure of the total information available as HPO annotations about that person's clinical features. Consequently, the total information per individual quantifies how much information is available in the data to distinguish them within the context of the overall cohort being studied (which dictates the IC of each annotation). This can be useful for assessing the amount of information available for phenotypic analysis or to establish how much variation there is between cohorts recruited from different centers in a

multicenter study when deciding where to focus efforts on further data collection. The number of clinical descriptors that can be inferred about an individual depends on three factors: (i) the complexity of their real-world clinical features, (ii) how thoroughly these have been annotated to HPO terms, and (iii) how precisely these phenotypic concepts are represented in the HPO, which limits the precision of annotation. Hence, a large amount of information may be available about an individual because they have one or a few rare phenotypic annotations or because they have many, particularly if these common features span multiple anatomical and functional systems — modeling the distinctiveness of classical multisystemic genetic syndromes for which a single feature is seldom pathognomonic.

3.7 Domain-specific biomedical ontologies can be improved through expert curation

The HPO is a relatively simple ontology where phenotypic concepts are connected through a single, *is_a* relationship. Therefore, in comparison to other more complex frameworks such as the SNOMED-CT or the Unified Medical Language System (UMLS), revision of entire branches of the HPO is feasible and can be accomplished by domain-specific expert panels (Kohler *et al.*, 2021). In 2018–2020, we revised the seizure subontology of the HPO on behalf of the Genetics Commission of the ILAE (Lewis-Smith *et al.*, 2021a). This revision, led by a dedicated Task Force, was aligned with the most recent seizure classifications, allowing us to measure the impact of revising the dictionary on the amount and quality of data captured. The number of concepts by which seizures can be described increased five-fold, despite removing several redundant terms. Importantly, the number of *is_a* relationships between phenotypic terms increased seven-fold, as a result of our efforts to harmonize concepts within and between formal classifications so that the HPO can be used to combine data from sources using different classifications with minimal loss of information or arbitrary distinction. Application of the seizure subontology of the HPO (using versions from before and after our revision) to data from a cohort of 791 individuals, demonstrated that the number of annotations according to which the cohort's seizures could be analyzed, and the amount of information about each participant's seizures both increased by 40 % following the revision (**Figure 2-3**) (Lewis-Smith *et al.*, 2021a). Moreover, we concluded that for the 86 participants for whom the revised seizure subontology appeared to capture less information than its predecessor, much of this lost information would be considered inaccurate according to current epilepsy classifications.

Furthermore, our findings address a more general question: does a new framework such as revised nomenclature or classification hold value for real-word clinical care or research?

Translating this question into a comparison of the amount of information captured through two competing frameworks allows this to be answered in a data-driven manner. In this instance, integration of recent classifications to build a harmonized model of clinical domain knowledge suggest that this is the case.

3.8 The challenge of representing important absent phenotypes

In some disorders, it is the absence of particular clinical features that is characteristic, and potentially important for clinical reasons. With regards to the two *SCN1A*-related epilepsies discussed in the introduction, in people with Dravet Syndrome, developmental delay typically occurs after seizures begin, usually progressing to some degree of intellectual disability, whereas in those with Febrile Seizures Plus, developmental delay is unusual. When phenotypes are coded as HPO annotations or even lists of clinical features, the relevance of absent clinical information is not immediately clear. For example, if an individual's propagated annotations do not include the term *Intellectual Disability [HP: 0001249]* is it safe to assume that this phenotypic feature is truly absent (that the individual did not have intellectual disability of any sort)? Alternatively, might it be present but not annotated the clinician did not consider it relevant at the time of data collection, so it was not documented in the raw data, or because some manual or automated data capture from the raw data missed it? In formal systems of logic used for knowledge representation, a difference is made between *open world* and *closed world* assumptions.

The more conservative assumption, used in the majority of HPO-based phenotypic studies, is open world, meaning that the absence of an annotation is not interpreted as the corresponding phenotype being absent. In this case, only those phenotypes that are explicitly recorded as being present are used for analysis. Consequently, one cannot distinguish between *absence of evidence* of the presence of a phenotype and explicit *absence of the phenotype itself*.

If a closed world assumption is made, the lack of phenotypic annotation represents the true absence of the feature. However, with regards to HPO annotations, a closed world assumption can typically only be made in a few scenarios. For example, in a study of individuals with developmental disorders, clinicians were required to complete a standardized data set recording each feature as explicitly present or absent, thus the absence an HPO term among an individual's annotations could be interpreted as meaning that the phenotype was actively sought and not found at the time of assessment (Andrews *et al.*, 2015). The more common scenario is that phenotypic data are provided in a less standardized, typically opportunistic manner, and limited conclusions can be drawn from the lack of assigned clinical features, especially in

dataset with relatively shallow phenotypic annotations. The Phenopacket format has been developed recently as part of the Global Alliance for Genomics and Health (GA4GH), and its *PhenotypicFeature* element allows clinical features to be recorded as being *present* or *excluded* using an ontology such as the HPO as a controlled dictionary (Jacobsen *et al.*, 2022).

We were able to evaluate the value of including explicitly absent phenotypes in a study of 413 individuals with *SCN2A*-related neurodevelopmental disorders, which are associated with a range of developmental outcomes that are clinically important to capture (Crawford *et al.*, 2021). After annotating individuals with positive HPO terms (explicitly present), we duplicated the HPO, modifying each term by adding *No* as a prefix to each human-readable name, and changing *HP:* machine-readable labels to *NP:*. We use this modified HPO as a notation of convenience for coding clinical features that are absent: 'negative phenotypes'. We then assigned negative phenotypes when clinical reports unequivocally reported that the feature was absent. In this way, we were able to give 260 of 413 individuals a total of 475 negative base annotations. This enabled us to confirm that not only was the presence of *Autism [HP:0000717]* associated with protein-truncating rather than missense variants in *SCN2A*, but also that the explicit absence of autism (*No Autism [NP:0000717]*) was associated with missense variants. We found the converse with regards to seizures. Yet addition of explicit negative annotations generated the power to detect a more nuanced relationship between seizures and missense variants. In contrast to the positive association of missense variants with *Seizure [HP:0001250]*, when discriminating by domain, missense variants in the S5-6 pore loop were associated with *No Seizure [NP:0001250]* suggesting that carries of missense variants in this domain may manifest phenotypes more typical of protein truncating variants.

We were also able to assess the phenotyping gap at the resolution of single HPO terms, meaning the proportion of individuals in the cohort for which the evidence was insufficient to assert the presence or absence of the given clinical feature under an open world assumption (**Figure 3-3**). After propagation, 1,064 of 1,268 HPO concepts were annotated exclusively as present or absent in this cohort. Other than phenotypes that were ubiquitous in this cohort (*All [HP:0000001]*, *Phenotypic abnormality [HP:0000118]*, *Abnormality of the nervous system [HP:0000707]* and *Abnormal nervous system physiology [HP:0012638]*), these have very large phenotyping gaps (> 50 %) so for 1,059 annotations, we do not know if the phenotype was present. 204 HPO concepts were annotated as being present in at least one individual and absent in another. For example, the presence or absence of *Delayed speech and language development [HP:0000750]* was only documented in 23 % of individuals (present in 10 % and absent in 13 %), leaving us blind to the language development of the majority of people with *SCN2A* disorders without further raw

data. The median phenotypic gap across these 204 phenotypic concepts was 80 %. This suggests that in general we should strongly favor an open world assumption where data are missing. However, depending on how raw data were collected (which may in turn depend on academic focus or clinical context), there are circumstances in which one might hypothesize that a closed world assumption may be reasonable. In this study, the presence of seizures might be expected to be reported whenever present, as enquiring about the occurrence of seizures would be expected as part of routine clinical assessment of patients presenting with neurodevelopmental concerns. Thus, in the absence of the explicit presence of seizures it might be reasonable to assume that they were absent. Regardless, we found that only 3 % of the cohort were annotated with neither *Seizure [HP:0001250]* nor *No Seizure [NP:0001250]*, indicating that authors of previous *SCN2A* case series and users of our EMR (local cohort) recognized the importance of assessing the occurrence of seizures and reporting their absence explicitly. Consequently, even in this case, the potential benefit of making a closed word assumption to fill missing data is minimal.

A particular technical matter needs consideration when using negative HPO annotations. Whereas propagation of positive phenotypes follows the *is_a* relationships to conceptually more generic, ancestral terms all the way to the HPO root node *All [HP:0000001]*, the opposite is true for negative phenotypes, which require 'downward propagation' so that the absence of generic concepts is used to infer the absence of their specific subtypes. For example, *No Global developmental delay [NP:0001263]* implies that there is *No Mild global developmental delay [NP:0011342]* and *No Severe global developmental delay [HP:0011344]*. Downward propagation, however, results in an unreasonably large number of inferred negative HPO annotations that could be analyzed but which are redundant for clinical interpretation. For example, in the *SCN2A* study downward propagation from *No Morphological abnormality of the central nervous system [NP:0002011]* resulted in a total of 695 negative annotations, corresponding to all descendants of this high-level term. In order to curb the number of negative annotations while allowing for meaning data harmonization, we developed a pruning method that eliminates terms that add no further information beyond their parent term.

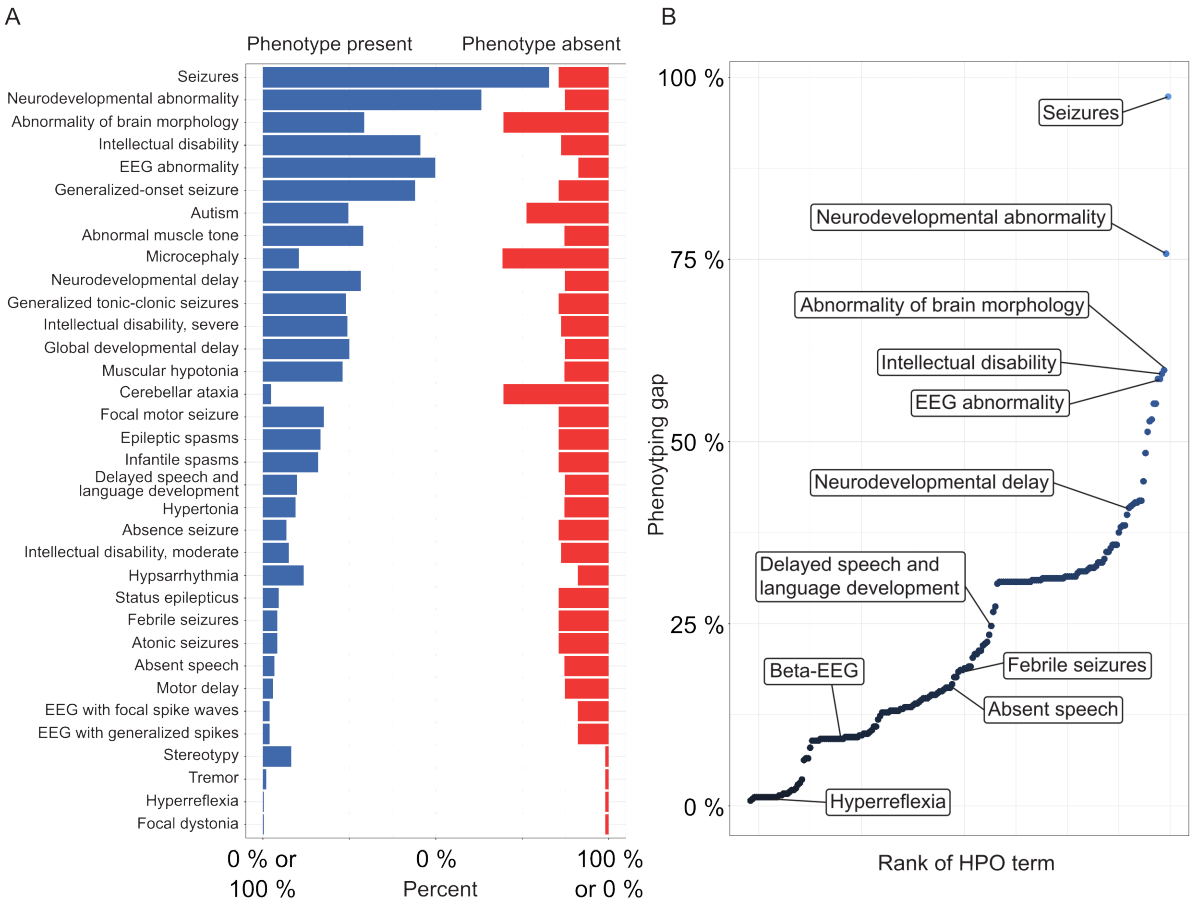


Figure 3-3. The proportion of individuals with *SCN2A*-related disorders with particular phenotypes using data from Crawford *et al.* (2021). (A) The percentage of individuals coded as having (blue) or not having (red) a selection of 204 phenotypes coded as present and absent in this cohort after propagation, as well as the phenotypic gap: the percentage of individuals in whom we cannot annotate the presence or absence of the phenotype. The six HPO concepts with the smallest phenotyping gap and a representative selection of the remainder are shown. (B) The distribution of the phenotypic gap for all 204 phenotypes that were coded as present in at least one and absent in at least one member of the cohort, ranked by phenotyping gap with only a selection labelled for clarity.

While the phenotyping gap might be expected to be large for clinical features that are rare, this can also be the case for phenotypes that are common in the general population and overlooked in the EMR. For example, stuttering (also known as stammering) is common. However, Pruett *et al.* (2021) found it to be under-reported in EMR billing code data from the Vanderbilt University Medical Center, given the 1–3 % prevalence in the general population. We too have found *Stuttering [HP:0025268]* to be conspicuous by its scarcity in the HPO annotations of several neurodevelopmental cohorts. To overcome the challenge of identifying people who stammer from using EMR data, Pruett *et al.* (2021) trained a supervised classifier to use sets of Phecodes to predict stammering in their cohort. Phecodes are a representation of clinical phenotypes that can be used to identify cases and controls based on the mapping of International Classification

of Diseases codes present in individuals' EMR (Bastarache, 2021). The classifier enabled identification of 9,239 individuals predicted to have stammering from their research biorepository, empowering a common variant genome-wide association study (GWAS) that identified two loci (Shaw *et al.*, 2021).

A further related consideration should be borne in mind if our interest is in knowing whether a genetic variant is associated with the presence or absence of a phenotype that is dynamic. If the phenotype is age-dependent, the age of the patient at the time of the clinical assessment will matter. For example, if none of a hypothetical cohort has *Delayed speech and language development [HP:0000750]* annotated, we might be more confident that their genetic variants are not positively associated with language delay if the entire cohort was assessed up to a minimum age of 10 years old than if half of the cohort were neonates in whom this phenotype cannot be assessed. This is a lesser concern for congenital phenotypes such as *Primary microcephaly [HP:0011451]* but greater for phenotypes that may not manifest until late adulthood such as *Parkinsonism [HP:0001300]*. One should also be careful before asserting that a phenotype that is transient and hard to capture is truly absent without caveats. Such scenarios may include the assessment in an adult of early life phenotypes that resolve in childhood (for example, many adults do not know for certain that they did not have febrile seizures as a child and their parents may no longer be available to confirm this). Furthermore, some dynamic phenotypes are not apparent to the patient or an onlooker without technical assistance and might be missed despite a complete medical history. For example, not all patients with seizures may have had electroencephalography (EEG), an investigation necessary to be able to assert that they have an *EEG abnormality [HP:000235]*. Even were this clinical investigation performed and no abnormality found, this may not be sufficient to conclude that the patient would never have had an abnormality were continuous EEG feasible. Given the dynamic and stochastic nature of cerebral electrophysiology, some EEG abnormalities may have been missed due to sampling.

In summary, inclusion of explicitly absent phenotypic annotations allows additional information to be captured without the risk of assuming that a phenotype is absent, which is relevant for phenotypic features where both presence and absence are meaningful when interpreting clinical data. In the future, the explicit presence, absence, or uncertain status of core phenotypes of particular interest for a disease domain (for example, *Neurodevelopmental delay [HP:0012758]* and *Seizure [HP:0001250]* in neurodevelopmental disorders) may become established within reporting standards. However, negative HPO annotations play by their own rules and require distinct data harmonization techniques and critical interpretation given their conceptual difference from explicitly positive assigned phenotypic annotations.

Some phenotypes, particularly those based on laboratory assays, are quantitative and consequently require binning into ordinal categories that can be represented in the HPO, for example *Hyperphenylalaninemia [HP:0004923]* and *Hypophenylalaninemia [HP:0500141]*. Additionally, one could use negative annotations such as *No abnormal circulating phenylalanine concentration [NP:0010893]* to represent normal values. However, information within bins is lost and the location of thresholds must be decided to partition the range of values optimally for the relevant phenotype, and for some phenotypes such as serum testosterone concentration, these need to vary by age and sex. While ordinal annotations (low, normal, or high values) may be sufficient for gross cross-sectional phenotypic analysis of groups of individuals, they may be insufficiently powerful for longitudinal analysis of dynamic phenotypes or identifying relationships at high resolution, such as the 1.3–3.1 point reduction in intelligence quotient per 100 μmol/l increase in serum phenylalanine concentration in phenylketonuria (Waisbren *et al.*, 2007). The Phenopacket format allows for documentation of quantitative measurements using standards such as the Logical Observation Identifiers Names and Codes (LOINC, (LOINC, Vreeman;McDonald and Huff, 2010)) with time stamps that could be used for quantitative analyses (Jacobsen *et al.*, 2022).

3.9 Longitudinal phenotype data captures age-specific clinical features

Understanding the temporal relationship of clinical features is critical, not only for diagnosis but also for prognosis and treatment decision-making. However, other than for more frequently encountered examples such as Trisomy 21 or 22q Deletion Syndrome, the phenotypic trajectories of neurodevelopmental disorders are largely unexplored at scale because of the efforts required to perform longitudinal studies in rare diseases and the lack of conceptual frameworks to analyze longitudinal, often piecemeal phenotypic data.

We have previous examined longitudinal phenotypic data by extracting features from EMR (Ganesan *et al.*, 2020; Lewis-Smith *et al.*, 2021b; Xian *et al.*, 2022). The widespread adoption of EMR mandated by the American Recovery and Reinvestment Act of 2009 is an unprecedented opportunity to leverage clinical data generated as a byproduct of healthcare for genomic research. Large national and international initiatives have started to systematically link biorepositories to EMR data, including the NHGRI-funded eMERGE consortium (eMerge, 2007) and the Geisinger MyCode Community Healthcare Initiative (Geisinger Healthcare System, 2021). These combined resources have been used to identify the protective effect of *ANGPTL4* for coronary artery disease and type 2 diabetes (Dewey *et al.*, 2016; Gusarova *et al.*, 2018), and a reduced risk of chronic liver disease in patients with variants in *HSD17B13* (Abul-

Husn *et al.*, 2018). The systematic use of EMR data allowed these studies to include up to 80,000 cases and 500,000 controls. Extraction of detailed clinical phenotypes from EMR data for genetic studies requires robust phenotyping algorithms, whether these are to be used for genotype-phenotype analyses or for inference or validation of affectation state regarding a specific disease (sometimes called a 'phenotype') or trait in a genetic association study. As of 2018, more than 60 phenotyping algorithms are available, demonstrating that methods based on EMR data can be successfully developed and applied for de-identified data across institutions (Kirby *et al.*, 2016). Similarly, national biobank projects such as the UK Biobank periodically update the phenotypic data available to researchers from hospital and primary care EMR codes, supplementing study-specific questionnaires and investigations and both the UK Biobank core team and collaborating researchers contribute algorithms for identification of participants with a wide number of phenotypes (UK Biobank, 2007). However, few algorithms specifically interrogate longitudinal clinical information.

In our longitudinal EMR-based study of children with genetic epilepsies, we developed an approach to analyze longitudinal clinical data taking advantage of HPO-based data harmonization and time stamps from the EMR (Ganesan *et al.*, 2020; Lewis-Smith *et al.*, 2021b). With clinical features translated into HPO annotations, we binned the annotations into 3-month age epochs using each participant's date of birth and the date of the clinical encounter recording the phenotype. We were able to map longitudinal phenotype data from 62,104 clinical encounters pertaining to 658 individuals across a cumulative 3,251 patient-years, generating 286,085 age-specific annotations based on a repertoire of 528 unique HPO terms. These data allowed us to map how phenotypic features associated with distinct genetic etiologies over time, such as the frequency of *Status Epilepticus [HP:0002133]* in 29 individuals with disease-causing variants in *SCN1A* compared to the remainder of the patient cohort (**Figure 3-4**). The approach developed in this study demonstrates how longitudinal phenotypic data can be harmonized and mapped to provide information about natural history and prognosis.

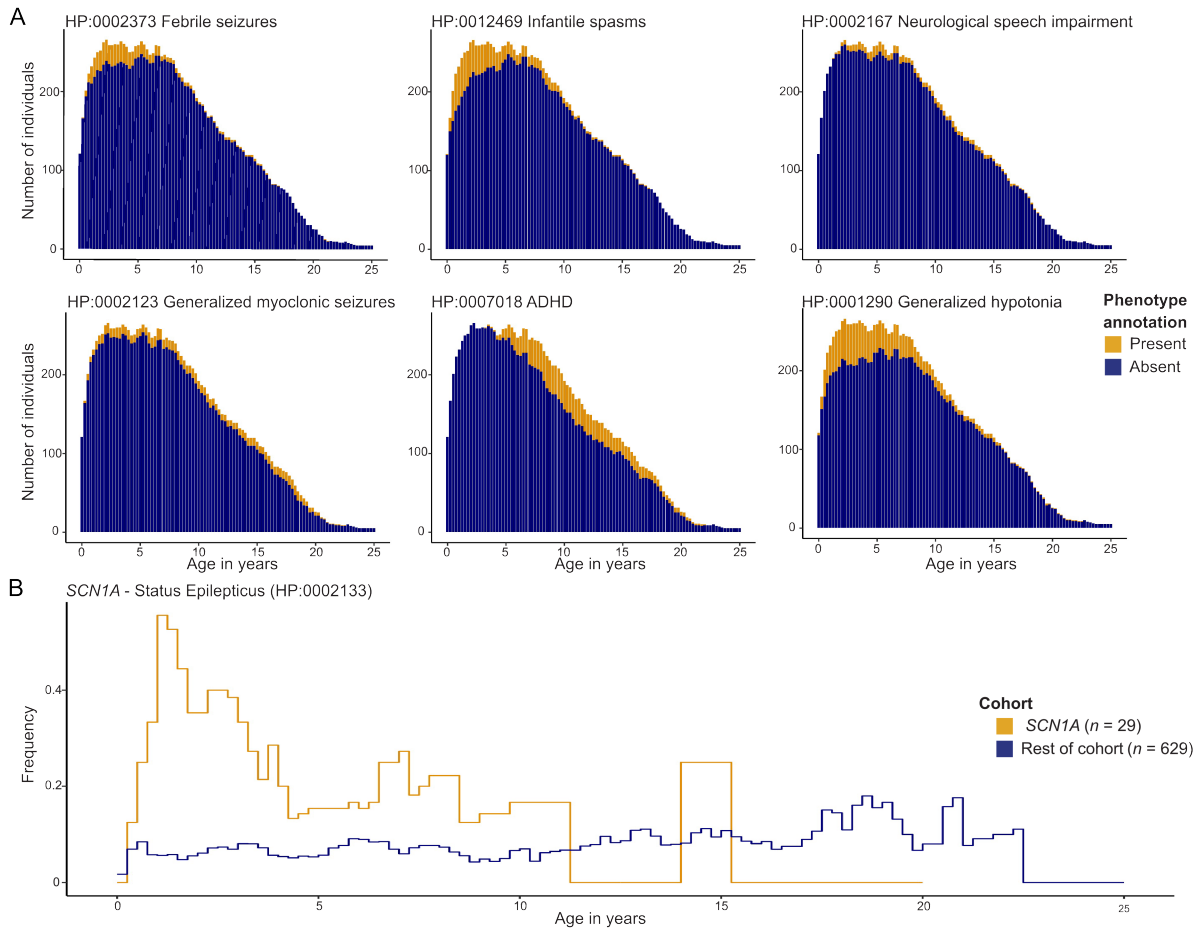


Figure 3-4. The longitudinal interrogation of HPO annotations from the EMR of patients with genetic epilepsies. (A) Stacked bar charts demonstrating how the number of patients with the given phenotype recorded and without the phenotype recorded varies with age. For example, *Febrile seizures [HP:0002373]* is coded most frequently in children aged 1–7 years of age but only in a minority of individuals with clinical encounters over this age range. (B) How status epilepticus is particularly common in children under the age of 5 years with diagnostic *SCN1A* variants compared to those without. Reproduced and adapted under CC-BY from Ganesan *et al.* (2020).

Several key considerations are relevant with respect to longitudinal mapping of phenotypic data. First, selection of time interval bin width is often arbitrary. While we compared the effect of various age bins, the use of constant time intervals across the entire age range can be reasonably questioned and the use of logarithmic scales have been suggested to better reflect human biology over time. Second, application of longitudinal phenotypic data emphasizes the concept of EMR usage. EMR usage refers to the interval between the first and most recent EMR encounter, reflecting the time frame where clinical information on an individual can be retrieved and analyzed — it is the chronological perspective from which we are limited when *observing* a person via their EMR. The effect of EMR usage is understudied but likely to play an important role in the interpretation of longitudinal clinical data. For example, in the EMR cohort discussed

above, no more than 250 of 658 individuals had EMR usage at any given age. In fact, for some disease groups as defined by genetic etiology, groups were temporally separated in the EMR data; individuals within these groups never overlapped chronologically, restricting our ability to make age-matched phenotypic comparisons. For example, all individuals with *PRRT2*-related diseases only had EMR usage up to age 5 years and those with *GRIN2A*-related diseases did not have any date before age 6 years so the clinical features of these two disorders were not directly compared (Ganesan *et al.*, 2020; Lewis-Smith *et al.*, 2021b). Future studies could compare the ages at which HPO terms can be inferred from the EMR to mitigate non-overlapping EMR usage. Third, individuals have often been cared for at multiple institutions (they may not have reached our center for a specialist assessment until they had had phenotypes for some time or they may be discharged to a local healthcare provider for follow up) and unless EMR data are visible across institutional boundaries (for example national and regional services and networks), a comprehensive perspective of each individual's dynamic phenotype requires collection of data from each institution within data protection regulations. Companies such as Ciitizen are now helping patient advocacy groups with such efforts (Ciitizen corporation). In summary, EMR usage is a critically relevant phenomenon in real-world EMR data with consequences for data analysis that may be underestimated.

3.10 Common Data Elements can be combined with biomedical ontology data

In pediatric epileptology, initiatives such as the Epilepsy Foundation-championed Epilepsy Learning Health System and the Pediatric Epilepsy Learning Health System are developing standardized methods to assess disease severity using Common Data Elements (CDEs) that can be directly integrated into the EMR across different healthcare networks, facilitating not only local but harmonized national quality evaluation and improvement (Grinspan *et al.*, 2021). In a study by Fitzgerald and collaborators, we reported the first analysis of seizure frequency using an EMR-implemented CDEs form to document disease severity for 1,696 patient encounters involving 1,038 patients (Fitzgerald *et al.*, 2021). This allowed us to describe how seizure frequency varied as the COVID-19 pandemic necessitated transition from predominantly face-to-face to telemedicine encounters, and by combining clinical data with routine demographic data we were able to assess racial and socioeconomic healthcare disparities.

We have recently combined our HPO- and CDEs- based methodologies to analyze how clinical features including both phenotypes and treatment responses vary in a single genetically defined disorder (Xian *et al.*, 2022). Within our international study of 534 people with *STXBP1*-related disorders, we reconstructed the clinical features of 62 individuals recruited from our center

across 4,433 monthly intervals, demonstrating a dynamic seizure pattern in the first year of life that was previously underappreciated. Combined with prescriptions for anti-seizure medication extracted from the EMR, this allowed us to outline the medication landscape of *STXBP1*-related disorders and assess the comparative effectiveness of various ASM. This method enables reconstruction of disease histories and easily be adapted to other conditions. **Figure 3-5** demonstrates a comparable analysis in 13 individuals with *SCN8A*-related epilepsies followed at CHOP, showing a gradual increase of seizure frequency in infancy compared to those with *STXBP1*-related disorders (Xian *et al.*, 2022) and differences in their responses to specific treatments. Seizures tend to become more common over the first year of life among people with *SCN8A*-related disorders, and those taking oxcarbazepine (a sodium channel blocker) are most likely to experience an improvement in seizure frequency. Regarding *STXBP1*-related disorders, seizures tend to become less common over the first year of life, typically responding well to the ketogenic diet and clobazam. However, those requiring ASM into adulthood commonly take levetiracetam rather than alternative treatments.

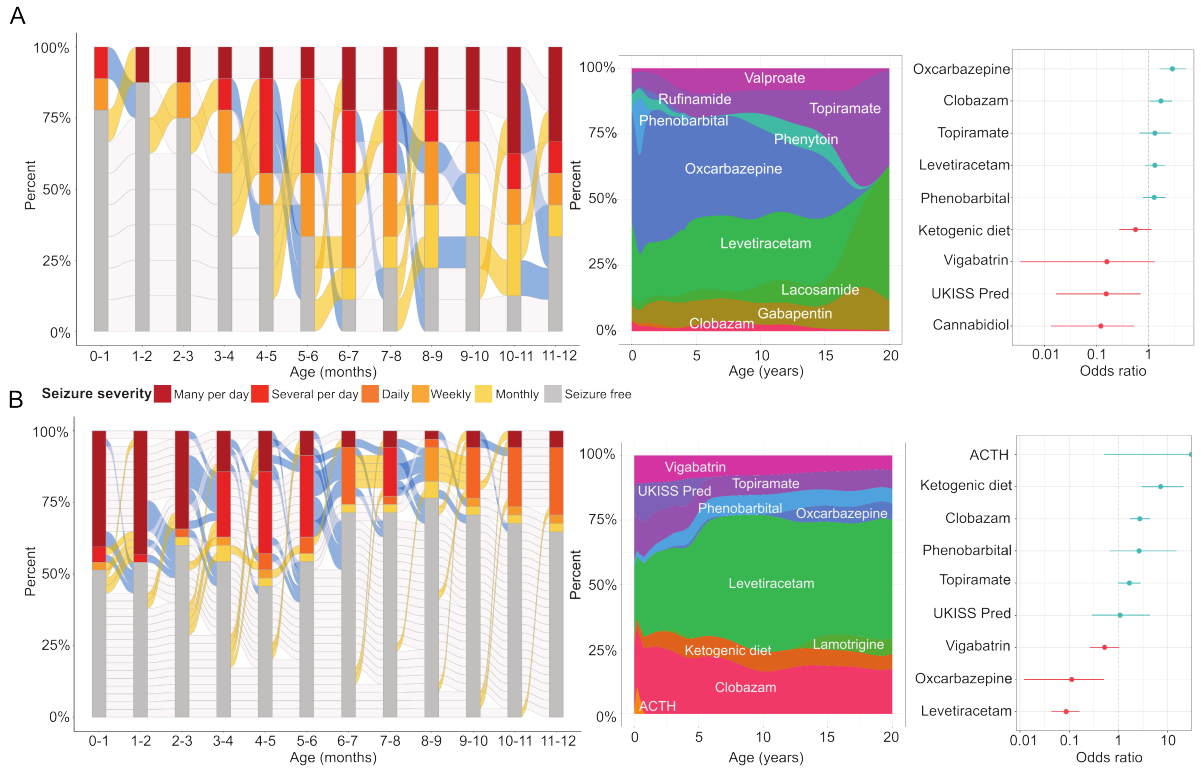


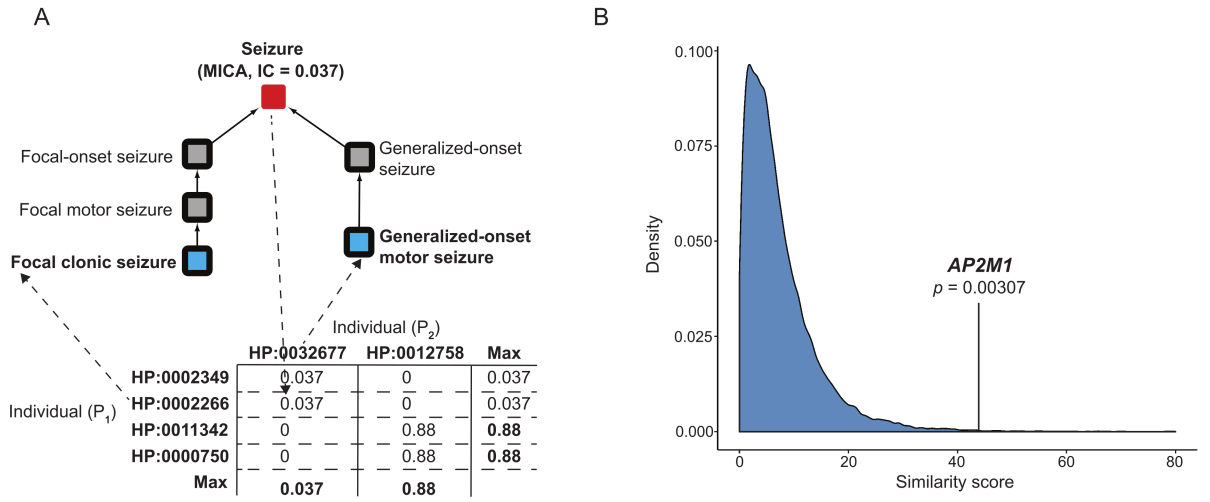
Figure 3-5. The distribution of seizure frequencies and prescription of various antiseizure treatments with age. The righthand panels show the ORs of achieving a reduction in seizure frequency or maintaining seizure freedom for a selection of medications and the ketogenic diet in patients at our center with (A) *SCN8A*-related and (B) *STXBP1*-related disorders. (B) reproduced and adapted under CC-BY from Xian *et al.* (2022).

3.11 Digitization of rare disease cohorts creates phenotypic references for genomic diagnostics

As many neurodevelopmental disorders include a broad phenotypic range and various known and as yet unidentified subgroups, it is desirable to describe the entire phenotypic landscape of a disorder or genetic etiology. In 2020, we undertook this for *SCN2A*, making it the first of the established monogenic neurodevelopmental disorder genes to have its phenotypic landscape represented digitally, i.e., the clinical features of all individuals reported in the literature supplemented by those followed at our center were systematically coded in HPO terms and analyzed jointly, akin to a GWAS mega-analysis (Crawford *et al.*, 2021). This combined data set included 10,860 HPO annotations in 413 individuals, allowing a us to examine the genotype-phenotype relationships of a gene-defined disorder quantitatively in unprecedented detail. By adapting the semantic similarity approach from natural language analytics to clinical phenotypes, we found that in the case of eight recurrent variants, people sharing the same variant were more similar clinically than expected by chance. We used principal component analysis (PCA) to identify the distinct phenotypic dimensions that contribute to clinical constellations seen in *SCN2A*-related disorders. This study was followed by our study of *STXBP1*-related disorders (Xian *et al.*, 2022) and preceded by a joint analysis of all prior major multicenter genetic studies of developmental and epileptic encephalopathies with a total of 31,742 HPO terms in 846 individuals (Galer *et al.*, 2020). Given that all three studies use the HPO as a joint phenotyping framework, information can easily be combined and jointly analyzed (Lewis-Smith *et al.*, 2021a) or combined with even larger datasets, such as the Deciphering Developmental Disorders study that currently provides more than 600,000 harmonized clinical annotations in more than 9,000 individuals with genomic data (Firth;Wright and Study, 2011). Accordingly, these harmonized datasets not only provide unprecedented insight into the range of phenotypic features in neurodevelopmental disorders, but also provide a critical reference framework for future diagnostic approaches in genomic medicine, particularly where the data collected prior to any potential bias from knowledge of the genetic diagnosis can be identified. One can hope that in the phenotypic data will increasingly be made available in the GA4GH *Phenopackets* format, which includes not only present and explicitly absent phenotypes but additionally, allow the use of annotations of each *PhenotypicFeature* with elements that can be used to modify the phenotypic feature with indicators of severity and chronological data such as the age of onset and resolution (Jacobsen *et al.*, 2022). A universally accepted format capturing such phenotypic richness would empower reuse of valuable data that might have already been published in future analyses.

3.12 Distinct genetic etiologies have recognizable phenotypic constellations

A harmonized genomic and clinical dataset enables a joint analysis of clinical and genomic data. In 2019, we used harmonized clinical and genetic data to identify *AP2M1* as a novel genetic etiology through a method referred to as phenotypic similarity analysis (Helbig *et al.*, 2019). Algorithms assessing phenotypic similarity aim to formalize an intuitive concept in clinical practice: that those individuals whose overall constellations of clinical features are remarkably similar are likely to share the same etiology (**Figure 3-6**).



Similarity score (sim_{max} = $Sim_{Resnik\text{-}BMS}$) for Individual P_1 and P_2 = $\frac{(0.037 + 0.037 + 0.88 + 0.88) + (0.037 + 0.88)}{2}$

Similarity score (sim_{max} = $Sim_{Resnik\text{-}BMS}$) for Individual P_1 and P_2 = 1.3755

Figure 3-6. An example of phenotypic similarity analysis using the $Sim_{Resnik\text{-}BMS}$ algorithm. (A) The annotations of two individuals are compared to each other. For each pairwise comparison of a phenotype from P_1 and P_2 the most informative common ancestor (MICA) is found. The MICA is the term that is an ancestor of the two terms being compared with the highest IC. The similarity of the two terms being compared is defined as the IC of their MICA. Once this has been completed for all pairwise comparisons of phenotypes, the overall similarity of P_1 and P_2 is calculated as the sum of highest similarity of each of P_1's annotations and each of P_2's annotations. The denominator of 2 helps with comparison of the similarity score calculated using this algorithm to those obtained using other similarity algorithms. (B) The median similarity of individuals grouped according to a genetic feature such as de novo variants in *AP2M1* is compared to the null distribution of median similarity scores generated by Monte Carlo simulation for groups of the same number of individuals (in this case $n = 2$) to yield an empirical *p*-value, estimating the probability of having observed a similarity this great due to chance in this cohort. (B) created using data from Helbig *et al.* (2019).

Historically, such reasoning has led to the discovery of the underlying cause of many monogenic neurodevelopmental disorders, such as Rett Syndrome and Dravet Syndrome. Yet these discoveries depended on a single expert clinician being fortunate enough to assess

sufficient patients with these disorders to recognize that these patients had characteristic patterns of clinical features. Phenotypic similarity algorithms aim to replicate this discovery process through a formal comparison of phenotypic features across an entire cohort. Once such algorithms can be successfully deployed on adequate and harmonized data, the distribution of similarities for individuals grouped by shared genetic features compared to a Monte Carlo generated empirical null distribution can identify novel genetic etiologies based on phenotypic evidence. We demonstrated that individuals with de novo variants in *AP2M1* were more similar than expected by chance in a cohort with developmental and epileptic encephalopathies, emphasizing that such a framework can be generalized to larger datasets, even if these cohorts have broadly similar clinical disorders. In our subsequent study combining data from three cohorts with developmental and epileptic encephalopathies, we used the same approach to show that the HPO data available were sufficient to demonstrate that 11 of 41 genetic etiologies (including prominent genes such as *SCN1A*, *STXBP1*, and *KCNB1*) had distinctive phenotypic patterns (Galer *et al.*, 2020). In this same article, we introduced novel visualizations to assess the contribution of harmonized phenotypic terms referred to as 'phenograms'. For example, the phenogram of those with de novo *SCN1A* variants showed a strong pattern consistent with the features observed in Dravet Syndrome, even though none of these individuals had been given this diagnosis clinically (**Figure 3-7**). This example demonstrates the power of computational phenotyping approaches, known disease entities can be reconstructed from clinical information sufficiently sparse to preclude manual diagnosis in clinic. Additionally, such approaches can be adapted using binned longitudinal EMR data to map the ages at which individuals sharing the same genetic etiology are more or less clinically homogeneous (Lewis-Smith *et al.*, 2021b).

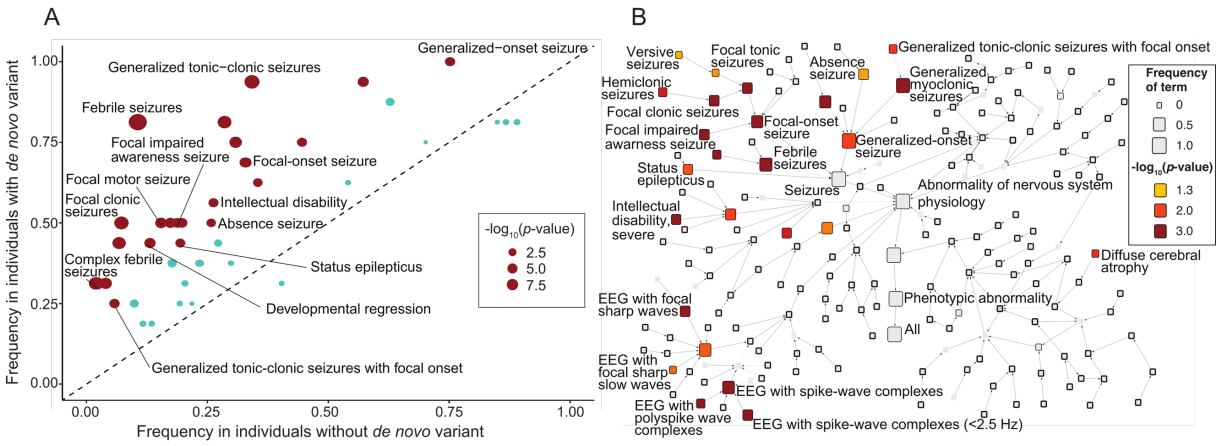


Figure 3-7. HPO-based visualizations demonstrate the clinical features associated with de novo variants in *SCN1A* in published cohorts with developmental and epileptic encephalopathies. (A) The frequency of annotation of HPO terms in carriers of *SCN1A* de novo variants versus non-carriers regardless of age. (B) The same data presented to demonstrate the conceptual relationships between associated features within the structure of the HPO. *p*-values were calculated using Fisher's exact test. Reproduced and adapted under CC-BY from Lewis-Smith *et al.* (2021a) using data from Galer *et al.* (2020).

3.13 Can computational assessment of phenotypic features aid variant interpretation?

Quantitative phenotyping based on widely accepted models such as the HPO can provide a framework through which to operationalize concepts such as phenotypic overlap at scale, including measures of significance and uncertainty. We previously demonstrated that existing HPO annotations from relatively modestly sized patient cohorts enables prediction of underlying genetic etiologies with high positive predictive values (PPVs). For example, the combination of *Febrile seizures [HP:0002373]*, *Generalized tonic-clonic seizures [HP:0002069]*, *Infantile onset [HP:0003593]*, *EEG with spike-wave complexes [HP:0010850]*, and *Focal Motor Seizures [HP:0011153]* has a positive predictive value of 0.9 for *SCN1A* de novo variants in children with unexplained yet suspected genetic developmental and epileptic encephalopathies (Galer *et al.*, 2020). Thus, even if individual terms appear non-specific, specific combinations allow for sufficiently powered predictions. Rare disease syndromes from Orphanet (Institut National de la Santé Et de la Recherche Médicale, 1997) are now being annotated with HPO terms, each with an estimate of the frequency with which it is observed in the disorder through the HPO – Orphanet Rare Disease Ontology (ORDO) Module, which integrates the ORDO with the HPO (Kohler *et al.*, 2019). Similar efforts are being made with monogenic disorders in OMIM (McKusick-Nathans Institute of Genetic Medicine, 2020) and sets of HPO terms are available on the HPO website (Core Human Phenotype Ontology/Monarch Initiative Team).

The value of these resources as reference data that can be compared to the features of a particular individual during variant prioritization will increase as manually curated gene-specific phenomic results become available from international efforts to collate clinical data from large cohorts such as the digitization of the clinical features of *SCN2A* and *STXBP1* discussed above. For example, Genomics England (GE) now provide researchers with the frequency of HPO phenotypes among 100,000 Genomes Project participants grouped by gene and variant classification (Turro *et al.*, 2020). The development of algorithms optimized for comparison of an individual's HPO annotations to a reference set continues (Robinson *et al.*, 2020). At the time of writing, the means by which such frameworks can be meaningfully integrated into American College of Medical Genetics and Genomics/the Association for Molecular Pathology (ACMG/AMP) criteria is unclear given that phenotypic features are often gene-specific rather than variant-specific and detailed phenotypic considerations contribute only supporting evidence (Richards *et al.*, 2015). If validated, increased weighting of phenotypic evidence may be particularly helpful in the molecular diagnosis of people without both parents available for demonstration of de novo status (a not infrequent scenario for clinicians caring for adults with neurodevelopmental disorders), particularly if this unlocks access to precision treatments. Even if phenomic tools provide evidence outside of formal variant interpretation criteria, they are likely to reduce lag to genetic diagnosis if successfully implemented within clinical decision support and learning health systems.

Chapter 4. Interpretation of genetically stratified phenotypic data

4.1 Preface

This chapter consists of my own fully independent critical review of the literature of phenotypic and biomedical similarity measures with consideration of their applicability to large clinical epilepsy data sets, and the use and interpretation of phenome-wide association tests.

4.2 Introduction

When comparing people with epilepsy by the clinical features that they manifest, each feature can be considered in isolation (for example, are people with variants in *SCN1A* more likely to have febrile seizures than people without *SCN1A* variants?) or people can be compared by their syndromes (for example, do people with variants in *SCN1A* tend to have similar clinical features?). Importantly, one should consider the context — the degree of phenotypic variability seen in the cohort or population of reference.

In this chapter, I will first discuss the reasons for favoring an open-world interpretation of clinical phenotypic data and how these should affect our interpretation of genotype-phenotype studies. Following this, I will outline the desirable characteristics of a phenotypic similarity measure in a large epilepsy research study. I will then discuss the formal analysis of genotype-phenotype associations before summarizing existing similarity measures, focusing on those that have been applied to the HPO and other biological ontologies.[6] I will use examples based on the annotation of HPO annotations to the Epi25 and CDKL5 Deficiency Disorder cohorts I will analyze in subsequent chapters. This chapter concludes with the proposal of a new similarity measure that models how a clinician might concisely summarize the clinical features that two people share, and the prioritization of similarity measures for subsequent analyses.

4.3 The case for an open-world assumption

A challenge in the analysis of clinical phenotypes is the incompleteness of data. People are not models on which experiments can be repeated until all features of interest have been comprehensively assessed. The information available is constrained by factors inherent in the

[6] Previous authors have used different labeling conventions, causing the labels of some similarity measures to be ambiguous in the context of the wider literature. Consequently, while I used authors' original nomenclature where possible, I have attempted to harmonize labels descriptively, referencing their author's original labels.

healthcare context and the dynamic phenotypes that can evolve over years and decades. Furthermore, many epilepsy phenotypes are paroxysmal, making them challenging to observe and diagnose. Accordingly, in studying the effect of genotype on clinical manifestations, it is critical to consider how to interpret the phenotypic data available for analysis.

Representations of knowledge interpret the absence of a particular piece of information differently depending on whether their formal logic makes an open-world or closed-world assumption. The *closed*-world assumption takes what is true to be known to be true, and what is not known to be true is assumed to be false. While the *open*-world assumption also accepts what is declared to be true to be true, it concedes that what is left undeclared remains unknown (Keet, 2013). For example, if we are told only that 'Michael's myoclonic seizures continue despite increasing his medication dose' then either interpretation would lead us to believe that 'Michael has myoclonic seizures'. Beyond this, a *closed*-world interpretation would lead us to conclude that 'Michael does *not* have intellectual disability', because were Michael to have intellectual disability, we would have been told this. Conversely, based on an *open*-world assumption, we would have insufficient data to conclude either that 'Michael has intellectual disability' or that 'Michael does not have intellectual disability'.

In the analysis of clinical data, we are limited to the data that have been collected in the clinical environment. A strict closed-world assumption implies the unrealistic expectation that the clinician was able to reliably exclude, or otherwise to confirm and document the presence of *every* possible phenotype of interest. Furthermore, it would assume that this information was captured comprehensively in annotations available for analysis. Furthermore, if we are interested in knowing whether the carrier of a given genetic variant is predisposed to a particular phenotype, we should base our interpretation not only on whether they have manifested the phenotype yet, but also on whether there remains a significant possibility that they may manifest it in future. This is rarely explicitly addressed in genotype-phenotype studies. Yet in most circumstances, the more conservative, open-world assumption is necessary for reasons one can consider intrinsic to the patient and phenotype, factors relating to the physician and clinical interaction, and factors affecting data capture from clinical records.

The patient may carry a variant conferring high risk of phenotype but may not have manifested this by the age at data collection. The phenotype may be age-dependent and not present by the time of data collection despite the patient going on to develop this phenotype subsequently. For example, intellectual disability cannot be diagnosed in a neonate. Some phenotypes may only manifest in people of a particular sex, genetic background, or in a different clinical state. For example, cryptorchidism cannot be diagnosed in females; latent epistasis could mask

manifestation of a phenotype positively associated with a genetic variant of interest; a person with profound intellectual disability may not be able to describe their subjective experience of a focal sensory seizure sufficiently for this to be diagnosed or classified; and a person may have latent photosensitive seizures but not yet been challenged with the necessary photic stimulus to provoke a seizure.

The phenotypes that define epilepsy are paroxysmal and this limits how readily they are observed and diagnosed, and the precision with which they can be classified. Clinicians rarely observe their patient's seizures. The diagnosis and classification of a seizure may benefit from detailed electroclinical interrogation of video-telemetry data on the epilepsy monitoring unit for one patient. Yet for another patient, with a latently identical seizure, diagnosis and classification may rest solely on interpretation of a non-expert witness description of an incompletely observed event. Additionally, neurophysiological phenotypes such as generalized spike and wave can only be observed at times when electroencephalography is being performed. Even if electroencephalography has been performed once and failed to capture a particular phenotype, this does not exclude the phenotype forming part of the patient's disorder (Hasan and Tatum, 2021).

The absence of some phenotypes can be established more confidently, particularly if they are congenital anatomical abnormalities that are reliably assessed. Preterm birth, low birth weight, and head size are usually well documented in developed healthcare systems. Dramatic events in the vulnerable neonatal period are reliably assessed, recorded, and often recalled by parents. Similarly, some phenotypes only present in a particular developmental phase of life. Consequently, they can be excluded if a patient has been assessed and a history of the phenotype excluded at an appropriate age. For example, a history of typical febrile seizures can be excluded in a child who reaches seven years of age without any febrile seizures. Similarly, any neurodevelopmental delay can be excluded if not manifest by adulthood. However, many phenotypes can present for the first time in adulthood. Consequently, even clinical records documenting their absence at a recent clinical encounter cannot be interpreted to mean that the person will not manifest the phenotype in future.

Within clinical encounters, the recognition of phenotypes that are, or have been, present depends on the thoroughness of the clinical history, examination, and investigations. Some phenotypes may not be volunteered in the history unless the clinician actively seeks these out. For example, a patient may have had several seizures on exposure to a specific stimulus but may not have recognized that these seizures were provoked (Lewis-Smith;Jaiser and Thomas, 2022). Another may be aware of a café au lait spot but consider it an irrelevant birthmark and

consequently not report it. A clinician may not test for hyperventilation-induced absences. Sometimes the patient cannot provide a history, and this depends on the availability of a contemporaneous historian (such as a parent) or documentation. For example, it is unlikely that an infantile febrile convulsion can be confirmed or excluded if the patient attends alone without notes. Contemporaneous documentation is particularly important for the precise age of events. For example, adults whose epilepsy started in childhood may not recall the exact age (even in years) when their seizures started. The nature of remote transient phenotypes may be uncertain. For example, a child may have had paroxysms that in retrospect may have been seizures, but which were not diagnosed at the time. The time of onset may differ from the time of clinical assessment, diagnosis, or classification. Commonly, a person may have had traits suggestive of autism spectrum disorder for some time before assessment and diagnosis, raising the question of when the phenotype should be considered to have begun (Lai and Baron-Cohen, 2015).

At a health service level, the detection of some phenotypes depends on the access to clinicians and investigations. Clearly, the expertise and culture of, and resources (including time) afforded to, the clinician and healthcare center matter. For example, there is variation among clinicians and healthcare centers in their preferred classification of seizures (Fisher *et al.*, 2017b; Luders *et al.*, 2019; Rosenow *et al.*, 2020), the precision at which they value classifying a seizure, the core sets of clinical data recorded in their clinical records and how these are structured, and the choice of investigations that might be performed for a given clinical presentation.

Furthermore, the recording of phenotypes that have been observed may be incomplete. Clinicians frequently identify incidental clinical features that they consider insufficiently relevant to document. Conversely, even if the phenotypes are core features that traditionally guide management of seizures, clinicians may consider these implicit given the context and consequently not document them (Lewis-Smith, 2022).[7] Beyond this, features that have been documented in medical records may not be captured during annotation and structuring of clinical record data for research purposes. This could be because of the limited expertise of the person collating the data for research or prioritization because of the burden of data entry. Sometimes research data have been imported from a historical research database or previous study for reuse but do not include the richness of clinical information desired for the present study. Furthermore, there can be information lost or inaccuracies introduced when coding data. While clean categorical raw data may make the analyst's task easier technically, analysis of weakened or distorted data detracts from their end. The person entering data may not be able to

[7] In some cases, even the patient's name may be sufficient for a clinician to recall many of their clinical features.

categorize a patient's phenotype within the categorical options provided in the data capture instrument. If free-text is not accepted, they are left with an inappropriately forced choice. For example, a clinical researcher may have to indicate if the participant has had 'absence seizures' or 'no absences seizures'. How are they to do this if they cannot classify the patient's dialeptic seizures as absences or focal impaired awareness seizures with reasonable confidence? Some phenotypes may not map well to any of the over 16,000 concepts available within the HPO. They may not have been represented in the HPO because they are new, because the field of study is a new application of the HPO, or because of limited domain expertise among the HPO development team. For example, to facilitate this work I had to introduce *Sudden unexpected death in epilepsy [HP:0033258]* and *Hippocampal sclerosis [HP:0033715]* as new HPO concepts (Gargano *et al.*, 2023). The former had been proposed to the core HPO team previously but rejected as it had not been understood to be a formal clinical entity. The latter appears to have been too epilepsy-specific for others to have integrated it into a tool developed primarily for clinical genetics.

Even if all of a person's phenotypes can be captured in the research data, these may not be at the precision required to be clinically meaningful. A person with juvenile myoclonic epilepsy could be annotated precisely with *Generalized myoclonic seizure [HP:0002123]*, *Bilateral tonic-clonic seizure with generalized onset [HP:0025190]*, and *EEG with polyspike wave complexes [HP:0002392]*. Alternatively, all three of these phenotypes could be captured less precisely but no less accurately with the two annotations *Seizure [HP:0001250]* and *EEG abnormality [HP:0002353]*. Such shallow phenotyping is less informative. Clearly this person must have some specific type of seizure and of EEG abnormality. We cannot assume that they have none of the subtypes of these two phenotypic descriptions. Instead, we are uncertain as to which one or more subtypes of each of *Seizure [HP:0001250]* and *EEG abnormality [HP:0002353]* the person has. Using the HPO to harmonize phenotypic annotations by inferring generic phenotypes from specific ones enables simultaneous analysis of clinical features at different degrees of detail. This maximizes the proportion of a cohort with data pertaining to generic phenotypic concepts. However, many phenotypes, particularly those that are conceptually precise and conceptually informative, are sparsely annotated in the cohort. Analysis of these is likely to include significant uncertainty, whether acknowledged or not.

In studies with thorough and systematic data collection, it may be possible to confirm the absence of a phenotype by the age of data collection, and greater confidence can be placed in this if the phenotype is early-onset, permanent or persistent for years, and if it is screened for

comprehensively by clinicians (Lewis-Smith *et al.*, 2022).[8] In these circumstances, after considering those in whom the phenotype has been excluded, it can be enlightening to measure the *phenotyping gap* (Crawford *et al.*, 2021). This is defined as the proportion of the cohort for which, from the data available, it is impossible to conclude with confidence if they have or have not manifested the phenotype. In these cases, for a given phenotype the data should be interpreted as representing up to three groups: (i) those people known to have manifested the phenotype, (ii) those who can confidently be said to have not manifested the phenotype *yet*, and (iii) the remainder. However, this is practically difficult in the study of largely sparse and heterogeneous opportunistic real-world clinical data from large cohorts of mixed or unknown ages for age-agnostic interpretation.

For some phenotypes, a person who has not manifested the phenotype may yet do so if given sufficient time and the necessary environmental and biologic conditions. Many typically early-onset phenotypes can occasionally first manifest or be detected late (for example, absence seizures and autistic traits). For an age-agnostic analysis, it may be prudent to avoid negating any potentially late-onset phenotype occurring, even after being actively excluded. In this case the phenotyping gap is the complement of the proportion annotated with the phenotype. Conversely, it could be argued that where an analysis is qualified as being conducted across data from participants followed up to the same age, early-onset phenotype can be negated.

Finally, we return to clinical phenotypes that are paroxysmal with short duration, difficult to identify, or that would require impractical continuous investigation to exclude, such as a paroxysmal EEG feature. It would rarely be appropriate to interpret an annotation negating such a phenotype to eliminate the possibility that the person has had the phenotype completely. Conceivably, the phenotype may have occurred without being observed due to chance or sampling bias. However, assuming that not detecting the phenotype *despite seeking it* contributes no information over the case in which it has *not been sought* seems overly conservative. It would be more fitting to consider the negation as conditional on some known or assumed sensitivity of capture as the test has some sensitivity greater than zero. Pragmatism is required given the opportunistic nature of real-world clinical data. For example, the presence or absence of a photoparoxysmal response can vary in the same person from one EEG study to another, presumably influenced by medication, time of day, season, and other complex factors (Danesi, 1988). Additionally, there is variability in the protocol to elicit the phenomenon and

[8] Some phenotypes may be demonstrated with more objective evidence resulting from biochemical assays or imaging, the results of which may be reliably archived for critical re-interpretation.

in clinician reporting. Thus, a single EEG may not be sufficient to completely exclude a photoparoxysmal responsiveness in an adolescent with clinically generalized-onset seizures.[9] One assumes that those with a stronger predisposition to a photoparoxysmal response will be more likely to exhibit it during a single EEG. Thus, the absence of a photoparoxysmal response despite at least a single EEG may be sufficiently informative to identify association in large samples.

Ultimately, the characteristics of clinical data suggest that an open-world assumption is to be preferred for analysis of the effects of genetic variants on clinical phenotypes. Clinicians are trained to be conservative, setting a high threshold to apply a diagnostic label. Accordingly, we can usually accept their assertions that a research participant has manifested a phenotype.[10] Consequently, the proportion of a cohort annotated as having a particular phenotype should be interpreted as a lower limit of those believed to have manifested the phenotype. Even where people have been annotated as having *not* manifested a particular phenotype, we should remember that its absence is dependent on many unknowns or at best conditional on a known age of data collection.

4.4 Defining sets of annotations for phenotypic analysis

Different sets of HPO annotations can be defined. Formal measures of the phenotypic similarity of a pair of people are derived from operations on these:

$\mathrm{anc}(t)$ is a function that returns the set of *terms propagated from term* t by following every directed *is_a* edge from t to the root term (*All [HP:0000001]*), including t and the root term. This set represents the complete set of descriptors that can be inferred to be applicable to t according to the ontology applied.

A *person's set of base annotations* (denoted by A or B in a pair) consists of the set of terms annotated to a person before propagation, typically translated from raw data. These are the input for semantic similarity measures (Semantic similarity). As we will see in 4.11, given that some of a person A's base annotations can be conceptually redundant (duplicated or implied by another base annotation of the same person), one can distinguish between a raw base set that

[9] Photoparoxysmal responses have consistently been found to be most prevalent in young people with a generalized epilepsy, suggesting a higher prior probability (Rathore;Prakash and Makwana, 2020).

[10] For example, a clinician should have at least 80 % confidence when classifying a seizure (Fisher *et al.*, 2017b).

contains all annotations translated from the raw data and a refinement of these to leave a *minimal set*, those terms necessary and sufficient to recreate D by propagation.

A *person's set of propagated annotations* (denoted by D) consists of the complete set of terms that can be said to apply to them, given their base annotations, by inference according to the ontology structure. Modeling each person to be compared as their set of base terms, their propagated set of terms consists of the union of the propagated sets of each of their base terms (**Equation 4-1**).

$$D_A = \bigcup_{t_A \in A} \mathrm{anc}(t_\mathrm{A}) \qquad \text{Equation 4-1}$$

Note that D_A is invariant to the addition of further terms to be propagated, so long as these are ontologically redundant (among the ancestors of another term in A). Consequently, the propagated set of D_A is identical to D_A. Propagated sets are used to harmonize a person's annotations across all more generic descriptors. Consider formal association analysis of single phenotypes (to be discussed in 4.5). Propagation of each person's annotations before this increases its validity, ensuring that they have all applicable phenotypic labels. For example, ensuring that a person whose raw data are mapped to *Focal impaired awareness seizure [HP:0002384]* is included in the nominator of the fraction of those with *Focal-onset seizure [HP:0007359]* despite this not being explicit in their translated data. Similarly, propagated sets are used for set-based measures of similarity (as we will see in 4.8).

Propagation as discussed above is *orthodromic*, following the directed edges towards the root of the ontology. This is useful for inferring applicable more generic phenotypic labels that can be said to be (of have been) manifested by a person A: each term t_A being propagated can be said to be a type of every one of its propagated terms $\mathrm{anc}(t_\mathrm{A})$. However, one can propagate in the opposite direction, *antidromic* to the directed edges of the HPO, towards the ontology's leaf nodes rather than its root. If interested in negated phenotypes, given the considerations above, antidromic propagation can be used to infer all those phenotypes that could be validly negated given that at least one of their ancestors (hypernym) has been explicitly excluded (negative annotations, Crawford *et al.*, 2021). This allows harmonization of data for association analysis of negated phenotypes.

Care should be taken to ensure that there are no internal inconstancies between positive and negative propagated sets. A person cannot be said to have, and also to have never had, some

phenotype. This facilitates quality control, allowing contractions to be clarified with clinical collaborators providing data. In my experience, the error usually arises in the terms negated because of misalignment of the relationship between broad classes in the HPO and the clinician's phenotypic schema: clinicians entering data may not be aware that some generic terms have clinically unexpected descendants. For example, in the following study of CDKL5 Deficiency Disorder (Chapter 7), a clinical contributor conceptualized stereotypies as *behavioral* disorders rather than *movement* disorders, negating the latter in people without other movement disorders but with stereotypies. However, in the applicable HPO version (2022-2-14), *Stereotypy [HP:0000733]* is a grandchild of *Abnormality of movement [HP:0100022]*. Consequently, initially it appeared that some participants were inferred both *to have had, and to have not had* stereotypies. This is clearly untenable, and annotations negating *Abnormality of movement [HP:0100022]* were censored.[11]

4.5 Association of genetic factors with phenotypes

Many quantitative variables in this work, such as the age of onset of a phenotype, or ordinal variables, such as degrees of its severity, have non-Gaussian or sufficiently discrete distributions to question the validity of parametric tests. Comparison of these variables across different groups of people is typically undertaken using non-parametric tests, such as the Wilcoxon rank sum test (Mann-Whitney U test) for two groups, the Kruskal-Wallis test (and its post-hoc extensions) by ranks for three or more groups, or where paired (for example change in serial measurements of severity in the same cohort) by the Wilcoxon signed rank test.

The prevalence of a single phenotype across groups lends itself to tests of independence between the two categorical variables of genetically and phenotypically defined groups, each with two levels. Consequently, the data for all individuals in the cohort can be placed in the mutually exclusive cells of a 2 × 2 contingency table (**Table 4-1**)

[11] In subsequent HPO release 2023-04-05, the concept of *Stereotypy [HP:0000733]* has been renamed *Abnormal repetitive mannerisms [HP:0000733]* and made an *Atypical behavior [HP:0000708]* and is no longer an *Abnormality of movement [HP:0100022]*.

		Phenotypic factor	
Genetic factor		Phenotype present	Phenotype absent
	Carrier	*a*	*b*
	Non-carrier	*c*	*d*

Table 4-1. Contingency table for analysis of the association between a genetic factor and the annotation of a phenotype.

For convenience, the levels of the two categorical variants may be labelled ‘phenotype present’ versus ‘phenotype absent’ and ‘carrier’ versus ‘non-carrier’. However, for the reasons described previously (4.3), the group labelled as phenotype absent should be considered critically to be those for whom we have been provided with insufficient evidence of having manifested the phenotype. Similarly, if stringent thresholds are set for the filtering of variants when establishing which individuals carry a particular variant, then carrier status should be interpreted likewise. The genomic analysis pipeline and specified filters will identify some people who can confidently be said to carry a variant of interest surpassing the threshold of evidence required for the research study. The remainder (the ‘non-carriers’) will include individuals in whom no variants of potential interest were detected as well as individuals in whom some evidence of a variant of interest was found, but this was insufficient to meet the criteria to pass the filters. Hence, with less conservative filters or with more evidence supporting the variant from better sequencing technologies or better-quality DNA, some of these individuals may be classified as ‘carriers’ rather than ‘non-carriers’. As a result, not only should the marginal $a + c$ be considered a lower limit, but if stringent thresholds have been set in the filtering of variants, then the marginal value of $a + b$ should also be considered a lower limit.

The association can be expressed as an odds ratio (OR). Its significance is interpreted using a two-sided Fisher’s exact test (FET) given the small number of individuals falling into some cells analysis of rare phenotypes and genetic features. Ideally, to calculate the relative risk (RR) of a phenotype given the presence of a genetic variant, one would recruit balanced sample of carriers and non-carriers. As we will see, Epi25 recruitment is relatively independent of phenotype and generates a cohort grossly representative of people seen in clinic with

unexplained, presumed genetic, epilepsies.[12] Hence, it is reasonable to express genotype-phenotype associations as RRs to aid interpretation by clinicians and patients.

An advantage of harmonization of large data sets is the large number of clinical features for which associations can be tested in each genetic group. While we are primarily interested in discovering clinical associations with genetic factors that might inform clinical care and have typically reported raw *p*-values previously (Helbig *et al.*, 2019; Galer *et al.*, 2020; Ganesan *et al.*, 2020; Crawford *et al.*, 2021; Xian *et al.*, 2022), the large number of tests can complicate interpretation of statistical significance. In the longer run, discoveries from this study can be considered hypothesis-generating, for testing and validation in independent cohorts. Within a large study such as Epi25, established gene-phenotype and genotype-phenotype associations (for example, febrile seizures and *SCN1A*, lissencephaly and *DCX*, and periventricular nodular heterotopia and *FLNA*) can be used as benchmarks. Furthermore, *p*-values can be adjusted to take account of multiple testing. However, common methods of controlling the family-wise error rate such as Bonferroni's correction, or even the more powerful Bonferroni-Holm's stepdown adjustment are overly conservative as they fail to allow for the complex dependency of HPO annotations (**Figure 4-1**). For example, one cannot have *Focal impaired awareness motor seizure [HP:0032712]* without also having its ancestors such as *Focal impaired awareness seizure [HP:0002384]* and *Focal-onset seizure [HP:000735]*. A Monte Carlo simulation allows one to maintain the dependency structure in the data, while assessing how frequently one should expect to observe the empirical *p*-values from FET due to chance. I have elected to use the *minP stepdown* (*minPsd*) and *single-step* (*minP*) algorithms to obtain *p*-values in reference to these simulations (Westfall and Wolfinger, 1997). I have found in the analyses described in this thesis that the stepdown variant of *minP* adds negligible power at the cost of greatly elevated time to run each analysis. Despite the false discovery rate (FDR) being less stringent than the family-wise error rate (Benjamini and Hochberg, 1995), I have found the *minP single step* procedure provides similar benefits in power over Bonferroni adjustment with the benefit of controlling the family-wise error rate. One would ideally simulate the analyses of all groups together to allow adjustment across not only all phenotypes by *group size* but by all *specific groups*, as like the phenotypic structure of the data, the genetic structure contains dependency (for example, the protein-truncating variants in *SCN1A* are a subset of the protein-altering variants in *SCN1A*). However, this would carry the computational burden of repeating the complete analysis across

[12] The Epi25 Collaborative pools research cohorts meeting inclusion criteria. Some cohorts might have been collected based on an epilepsy type or syndrome, yet overall recruitment to Epi25 is not based on the presence of absence of any particular epilepsy phenotype.

every genetically defined groups and phenotypes many tens of thousands of times for debatable benefit. Consequently, I have chosen to simulate the association of each phenotype with groups of participants of each genetically defined group size in each study, without adjusting across genetic groups. This is equivalent to considering each gene's association analyses as an independent study performed on the same dataset.

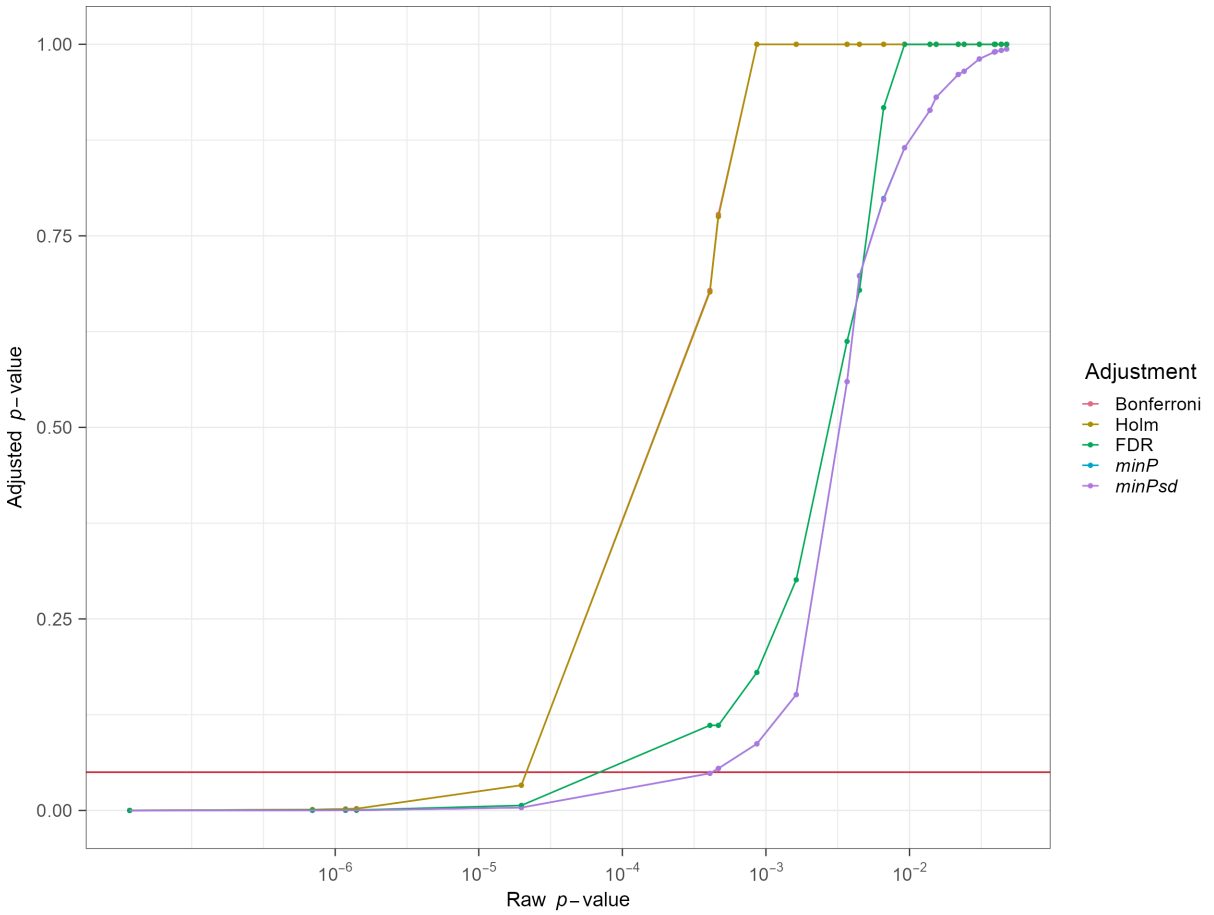


Figure 4-1. Comparison of forms of adjustment of the *p*-value from Fisher's exact test. For this example, the raw *p*-values are taken from the associations of 24 carriers of deletions at 15q13.2-q13.3 [31.06-32.51 Mb] in a cohort of 10,880 people with 1,667 informative phenotypes (see Chapter 6). Bonferroni's *p*-values are obscured by Holm's *p*-values and *minP* single step (*minP*) values are largely obscured by *minP stepdown* (*minPsd*). FDR, false discovery rate. *minP* and *minPsd* result in one further result reaching an adjusted *p*-value < 0.05 than then other methods.

Testing of the association between specific hypothesized combinations of features and a genetic grouping factor would be possible using FET. One would group individuals phenotypically not by the presence or absence of an annotation of a single phenotype, but by the presence of all of a necessary set of phenotypes. This can be considered formalization of the association between a genetic factor and a syndrome. However, the susceptibility to missing phenotypic data may increase as phenotypic dichotomization becomes dependent on combinations of greater numbers of phenotypes. Even when testing a specific hypothesis, one would have the challenge

of selecting the features that define the hypothesized phenotypic dichotomization and the precision at which to define them. Furthermore, the testing (and subsequent interpretation) of the vast number of associations between genetic factors and comprehensive combinations of different numbers of phenotypes would challenge an unbiased hypothesis-free analysis.

An alternative approach to the comparison of groups of people according to their sets of clinical annotations is to measure the similarity (or inversely the distance) between pairs of people according to their sets of clinical features. Descriptive statistics of the similarity between people sharing the genetic feature can then be interpreted in comparison to those of people without the genetic feature, or the overall context of the cohort being studied. Semantic similarity analyses quantify the degree of similarity between two entities based on the conceptual relationships between their annotated features (Pesquita *et al.*, 2009). In this work, the entities are people with epilepsy, and the analysis compares the sets of HPO terms (nodes) annotated to each according to their *is_a* relationships (edges) in the HPO (Kohler *et al.*, 2021). As we will see, multiple algorithms based on the nodes or edges of an ontology have been proposed for phenotypic similarity analysis and molecular bioinformatics. Often these have been repurposed from natural language processing. Consequently, we should first consider the features of an ideal phenotypic similarity algorithm given the nuances of clinical research data.

4.6 Desirable features of a similarity measure for large epilepsy studies

The HPO provides a controlled dictionary to which one can map raw data to reduce arbitrary dimensionality, particularly when evaluating unstructured free-text data. HPO data lend themselves to semantic and lexical similarity measures and selection of the most appropriate options requires consideration of what the similarity measure seeks to achieve. Broadly, the similarity algorithm should aim to imitate an expert clinician's considered judgement of how alike two people are clinically.

A formal similarity measure should distinguish those people sharing a greater number of clinical features from those who share only one. The overall similarity of a pair of people should tend to increase as the number of distinct phenotypes that they share increases.

Not all phenotypes are equally informative. A clinician is more struck by the similarity of two people who share a rarely encountered phenotype (such as myoclonic-absence seizures) than if they share only a common phenotype (such as bilateral tonic-clonic seizures).

The measure should be explainable, allowing interrogation of the grounds on which it has considered two people to be similar. More generally, one should be able to interpret,

generically, how it models the subjective phenotypically similarity used by clinicians to recognize syndromes: the process that has traditionally formed the backbone of diagnostics in clinical neurology and genetics. This increases the chance of acceptance by clinicians, and clinical influence of study results (Rajpurkar *et al.*, 2022).

The measure should be computationally efficient to allow analysis of large cohorts or large longitudinal data sets in a timely manner with proportionate computational resources.

The measure should allow comparison of phenotypes recorded at different degrees of precision because the granularity at which seizures can be classified is variable. This would remove the need to exclude people with insufficiently precise phenotypic data or alternatively to discard information beyond the degree of detail at which most of the cohort has been annotated.[13] Sematic similarity measures can achieve this using the *is_a* relationships of the HPO to modeling reasoning based on formal seizure classification and semiological description, and set-based similarity approaches can be applied after propagation of mapped annotations (Lewis-Smith *et al.*, 2021a).

An HPO-based similarity measure should have minimal sensitivity to arbitrary features of an ontology that has been developed through largely ad hoc piecemeal contributions from academic and clinical human geneticists. In general, the HPO should be considered a tool for harmonization of clinical descriptors rather than a representation of natural classes of phenotype whose relationships and divisions are based on robust scientific evidence. Even where coordinated efforts have been made to design a subontology according to formal classifications, these subontologies are dependent on the classifications on which they are based (Kohler *et al.*, 2021). For example, seizures, the core phenotypes of epilepsies, are not classified according to scientific evidence but rather according to clinical pragmatism (Fisher, 2017; Fisher *et al.*, 2017a), and it may still be more appropriate to consider the formal classification more akin to a gardener's classification than that of a botanist (Jackson, 1958; Wolf, 2003). The presence or absence of an *is_a* relationship within the HPO is sometimes arbitrary. For example, in revising the seizure subontology, we had to decide how closely to link seizure types with similar semiology between three ILAE seizure classifications covering different clinical scenarios: the current ILAE classification applicable to seizures beyond the neonatal period (Fisher 2017, Fisher, Cross et al. 2017), the physiologically different brains of neonates with their distinct seizure repertoire (Pressler *et al.*, 2021), and the seizures of status epilepticus (Trinka *et al.*, 2015), classification of which predated the core concepts of the other two (Lewis-Smith *et al.*,

[13] For example, simplifying all nuanced focal semiologies to *Focal-onset seizure [HP:0007359]*.

2021a). Because of these differences, we usually opted for the conservative option to represent these three classifications in three distinct subontologies, setting a high threshold when considering whether to link two concepts between these.

Given the paucity of scientific evidence for some of the HPO structure, analyses should avoid putting too much confidence in the biological validity of nuanced distinctions within the HPO, but rather, focus on shared, more generic descriptors. Consider the comparison of focal impaired awareness emotional seizures with laughing and focal impaired awareness seizures with visual features. It may be preferrable to focus on the fact that both could be called focal impaired awareness non-motor seizures more generically, rather than focus on the difference in their initial manifestations. Furthermore, As discussed above, two people should be considered more similar clinically if they share (i) the same precise clinical annotation, such as *Focal impaired awareness seizure [HP:0002384]* than if they share (ii) only a more generic term (*Focal-onset seizure [HP:000735]*) as their most specific annotation. However, the similarity in a third scenario, in which one person has the more precise annotation (*Focal impaired awareness seizure [HP:0002384]*) and the other the more generic (*Focal-onset seizure [HP:000735]*) should not be overly penalized compared to the second. The evidence provided indicates that each pair has a shared tendency towards focal-onset seizures and does not indicate that the third pair has any less of this tendency than the second. The difference in the precision of annotation seen in the third pair will rarely be due to a difference in their true phenotypes, more often it will be caused by one of the factors affecting phenotypic data capture discussed above. A phenotypic similarity measure should recognize this, rather than focus on how close two sets of annotations are to being identical.

Epilepsy is a complex disorder and people with epilepsy are likely to have comorbidities beyond the core features of seizures (Scheffer *et al.*, 2017). A person with epilepsy will often have multiple phenotypes. A similarity measure designed to compare people should take account of these to compare people by their complete clinical constellation as it is often combinations of clinical features rather than single phenotypes that characterize disorders. Not all individuals with the same recognizable disorder will be annotated with the same complete syndrome of characteristic clinical features. While Dravet syndrome typically presents with febrile seizures and these may be hemiclonic, with subsequent myoclonic seizures (Severe myoclonic epilepsy in infancy), up to 45 % present with afebrile seizures, a third have no history of myoclonic seizures, and a third have no history of hemiclonic seizures (Li;Schneider and Scheffer, 2021). This may be due to variable expressivity, age at last data capture, or variability in phenotypic

observation and reporting. A measure that takes account of all available phenotypic annotations may partially mitigate against variability in manifestation or capture of a single phenotype.

4.7 Edge-based similarity measures

Purely edge-based methods are straightforward to understand in an ontology such as the HPO with only a single type of relationship. They return a similarity that is a function of the number of edges between nodes representing the pair of concepts being compared, and in some measures, the root node and/or the lowest common ancestor (the most conceptually specific term to subsume both terms being compared[14]). They emerged from the analysis of words and language (Rada *et al.*, 1989; Wu and Palmer, 1994; Pekar and Staab, 2002). The similarity returned by these depends only on the ontology structure and the terms compared. This can be advantageous as the similarity of two terms within the same ontology version is externally consistent across cohorts.

The simplest measures that could define the similarity of two terms, t_1 and t_2, are based on the path length (number of edges) between them (Rada *et al.*, 1989). The reciprocal of the shortest path length (distance) could represent their maximum similarity. The reciprocal of the mean path length of all possible paths between t_1 and t_2 could represent their average similarity, according to all the ways in which they can be said to be alike. However, these measures fail to take account of how conceptually specific the terms are. For example, the conceptual similarity of fine-grained terms such as *Focal impaired awareness clonic seizure [HP:0032725]* and *Focal impaired awareness myoclonic seizure [HP:0032730]* should be greater than that between generic concepts such as *Abnormality of the skeletal system [HP:0000924]* and *Abnormality of the musculature [HP:0003011]*, despite both the distance within both pairs being two edges. This difference can be represented by the depths (number of edges from the ontology root) of their most conceptually specific common subsuming concept(s). The most similar that *Focal impaired awareness clonic seizure [HP:0032725]* and *Focal impaired awareness myoclonic seizure [HP:0032730]* can be said to be is that both are types of *Focal impaired awareness motor seizure [HP:0032712]*, which is seven edges from the root. Conversely, *Abnormality of the musculoskeletal system [HP:0033127]*, the most specific concept subsuming both *Abnormality of the skeletal system [HP:0000924]* and *Abnormality of the musculature [HP:0003011]*, is two edges from the root, indicating that it is conceptually less specific than *Focal impaired awareness motor seizure*

[14] In linguistics this would be the most specific hypernym applicable to two words being compared.

[HP:0032712]. Several authors have proposed distance or similarity metrics that incorporate the depth of the lowest common ancestor as well as the distances between t_1 and t_2 and this ancestor (Wu and Palmer, 1994; Pekar and Staab, 2002). Within analyses of Gene Ontology (GO, http://geneontology.org, Ashburner *et al.*, 2000; The Gene Ontology Consortium *et al.*, 2023) data, some authors have compared all the possible paths through the directed acyclic graph of the HPO from t_1 to the root to those from t_2 to the root. They define the similarity of t_1 to t_2 as the depth of their lowest common ancestor, represented by the maximum number of terms reached along the longest path from the lowest common ancestor to the root (Wu *et al.*, 2005). This approach has been further adapted to also consider the descendants of t_1 and t_2, those terms that they subsume conceptually (in linguistics described as hyponyms). Furthermore, the distance from each of t_1 and t_2 to the leaf nodes of their respective subontologies (their most conceptually specific descendants) may offer complementary information that allows a similarity measure that attempts to capture the relative conceptual depth of t_1 and t_2 between their ancestors and descendants (Wu *et al.*, 2006).

However, there is a general weakness when applying these path-based measures to HPO data. The structure of the HPO is opportunistic. It predominantly reflects the relationships between those phenotypic concepts that clinicians can typically observe and distinguish. It can potentially over-represent clinically nuanced distinctions of academic interest. Thus, it is difficult to quantify the conceptual distance represented by any edge and consequently to compare edges. For example, *Focal impaired awareness clonic seizure [HP:0032725]* is one edge from *Focal impaired awareness motor seizure [HP:0032712]*, and *Abnormality of the skeletal system [HP:0000924]* is one edge from *Abnormality of the musculoskeletal system [HP:0033127]*. How valid is it to say that the difference (or similarity) between the first pair of concepts is equivalent to the that of the second pair? As an ontology developed by a community of interested users, the HPO has developed in a largely disjointed manner. As a result, the complexity and precision of subontologies will be biased by the clinical domains of interest to the contributing experts. This means that the HPO is unlikely to meet the assumptions required for edge-based approaches to be reliable (Pesquita *et al.*, 2009). In some subontologies the conceptual distinction between nodes will be more nuanced (a particular phenotype has been subdivided to a high resolution), resulting in a high density of nodes separated by edges that represent short conceptual distances. In others, a phenotype may not have been subdivided into as many subtypes. Even the distance between two children of the same term from that parent term cannot be assumed to be similar, despite each child being a path of length one from the term. Consider *Phenotypic abnormality [HP:0000118]*, near the HPO root where concepts are defined most generically. It has children defined anatomically (for example *Abnormality of the nervous system [HP:0000707]*), and others

defined pathologically (for example, *Abnormality of metabolism/homeostasis [HP:0001939]* and *Neoplasm [HP:0002664]*). How confident can one be that the single edges between each of these children and their parent represent the same semantic distance, despite being the same distance from the root? Consequently, edge-based similarity measures could bias similarity values according to variations in the richness of relevant phenotypic domains within the HPO.

A potential solution proposed for the GO is to weight edges by their conceptual distance based on investigator judgement (Cheng *et al.*, 2004). However, this would be difficult to undertake and appraise critically in a large ontology, particularly given that alternative paths between each term and the root would need to be reconciled.

Furthermore, the HPO is developed iteratively with new versions released several times per year. The addition of new nodes (dividing the existing edges) would be sufficient for the similarity value of phenotypes being compared to vary across HPO versions, potentially for arbitrary reasons.

4.8 Set-based measures of similarity

4.8.1 Comparison according to the terms in each person's set of propagated annotations

An alternative to edge-based similarity measures is to use approaches based on nodes representing phenotypic terms that a pair of people do or do not share. As we will see, these can be given data-derived weights to account for their discriminatory value.

The simplest form of set-based measure is the *total term overlap measure* ($\mathrm{Sim}_{\mathrm{TOc}}$), which has been applied to GO data for the comparison of gene products (**Equation 4-2**; Lee *et al.*, 2004).

$$\mathrm{Sim}_{\mathrm{TOc}}(A, B) = |D_A \cap D_B| \qquad \text{Equation 4-2}$$

Here, after propagation to describe each person's phenotypes using all applicable HPO terms, the set of terms intersecting the propagated annotations of A and B is found. The cardinality of this set, the number of unique intersecting terms, gives an unweighted similarity of the pair of people. Pairs whose shared phenotypic features can be described using more HPO terms are thus given a greater similarity than those whose propagated sets have little overlap. This models how a clinician might consider the similarity of two patients to depend on how many phenotypic descriptors could be used to describe how the two are similar. Similarity by term overlap cardinality takes the value of an integer, resulting in a very discrete distribution of values and

consequently many ties are encountered when comparing the relative similarities of multiple pairs of individuals in large cohorts. These ties reduce the ability to rank pairwise comparisons of individuals by similarity. The overlap similarity can be normalized according to the number of annotations in the smaller of the two sets being compared (**Equation 4-3**; Mistry and Pavlidis, 2008).

$$\mathrm{Sim}_{\mathrm{NTOc}}(A,B) = \frac{|D_A \cap D_B|}{\min(|D_A|, |D_B|)} \qquad \text{Equation 4-3}$$

The *normalized term overlap similarity* ($\mathrm{Sim}_{\mathrm{NTOc}}$) adjusts $\mathrm{Sim}_{\mathrm{TOc}}$ by the size of the smaller of the two sets of terms being compared (it does not take account of the larger of the two). The *Jaccard index* takes account of the number of distinct propagated annotations of A and B, penalizing discordant annotations (Jaccard, 1901). It divides $\mathrm{Sim}_{\mathrm{TOc}}$ by the cardinality of the union of a pair's propagated annotations, It has been defined as *union-intersection similarity* ($\mathrm{Sim}_{\mathrm{UI}}$) in GOstats (Falcon and Gentleman, 2006) and as Jaccard similarity ($\mathrm{Sim}_{\mathrm{J}}$) in OWLSim (Washington *et al.*, 2009; Chen *et al.*, 2012), two of the earliest packages for calculating similarity in biological ontologies (**Equation 4-4**).

$$\mathrm{Sim}_{\mathrm{UI}}(A,B) = \mathrm{Sim}_{\mathrm{J}}(A,B) = \frac{|D_A \cap D_B|}{|D_A \cup D_B|} \qquad \text{Equation 4-4}$$

In a clinical setting, this measure models a clinician considering the proportion of the phenotypic descriptors applicable to at least one of two patients that are applicable to both. Measures based on Jaccard's index are limited to the closed unit interval [0–1]. They are less discrete, reducing the number of ties encountered using term overlap similarity by considering those pairs with fewer discordant annotations to be more similar than those with a greater number of discordant annotations. A further variation on this approach is the *Sørensen-Dice coefficient* (SimSDC, Dice, 1945; Sørensen, 1948) which puts greater weight on the number of intersecting terms, penalizing discordant less than Jaccard's index (**Equation 4-5**).

$$\mathrm{Sim}_{\mathrm{SDC}}(A,B) = \frac{2 \times |D_A \cap D_B|}{|D_A| + |D_B|} \qquad \text{Equation 4-5}$$

The similarity measures above consider all nodes of the ontology to be of equal intrinsic similarity value, regardless of the discriminatory potential of the phenotypic concepts that they represent. It is desirable for a similarity measure to recognize that the shared presence of some phenotypes is more informative than others. This would discriminate between the many pairs of individuals whose similarities would be identical by simple node-counting measures and allow a pair of individuals sharing a sufficiently informative phenotype (such as *Early onset absence seizures [HP:0011152]*) to appear more similar than a pair sharing multiple very uninformative phenotypes (such as unspecified *Focal-onset seizure [HP:0007359]* and *EEG with focal epileptiform discharges [HP:0011185]*). Where large data sets are to be analyzed, a natural data-derived measure of the informativeness or distinguishing value of single term is its IC. The IC of each term, t in the HPO annotations of a data set is given by the negative logarithmic transformation of the probability, P of encountering it among the annotations of each person in the cohort (**Equation 4-6**).

$$IC_t = -\log_2 P(t) | t \in HPO \qquad \text{Equation 4-6}$$

The minimum value of IC_t is 0, where t is annotated ubiquitously and thus uninformative, and the maximum possible value, where t is annotated to a single person, is the reciprocal of the size of the cohort. As a function of the frequency of each phenotypic concept in the cohort, absolute IC-based similarity values are not externally transferable from one cohort to another. Edge or node counting measures are given that they are intrinsic to the HPO graph structure. Thus, in isolated comparisons of small numbers of people or where cases are compared to phenotypic diagnostic criteria, measures that are either unweighted or weighted according to values generated from external reference data or the structure of the HPO may be preferred. For instance, when comparing a single person to a catalogue of reference phenotypic sets such as the diagnostic criteria of syndromes, the relative discriminatory value of each phenotypic concept may be known or estimated with each phenotype weighted according to mandatory, major, and minor diagnostic criteria.

Many different node-based IC-weighted similarity measures have been proposed for biomedical ontologies (see Pesquita *et al.* (2009), Pesquita (2017), Deng *et al.* (2015) for

summaries). Often these were first developed to compare the similarity of texts in the context of a corpus (Jiang, 1997; Resnik, 1999), or to compare molecules according to features defined in the GO (Lord *et al.*, 2003a; Schlicker *et al.*, 2006).

For set-based approaches, like the term overlap measures above, the acyclic graph structure of the HPO is used only for propagation along *is_a* relationships, to generate an exhaustive list of phenotypic descriptors that apply to a person based on their raw data. Set-based analyses without this harmonization recognize only direct hits. Consequently, for the shared nature of a phenotype to have been captured would require it to have been annotated at the same degree of specificity and described according to the same features. For example, the similarity of person A with *Focal motor seizure [HP:0011153]* to a person B with *Focal aware seizure [HP:0002349]* due them both having unannotated focal-onset seizures would be missed. Indeed, the similarity of A to another person who has been *Focal-onset seizure [HP:0007359]* would be missed. Once each person's sets of annotations have been propagated, most set-based measures compare people using set operations without needing to consider the relationships between the phenotypic concepts that the annotations represent. This has a lower computational burden than the ontology-based inference required for every pairwise comparison of terms within semantic similarity analysis.

Galer *et al.* (2020) defined a set-based *IC-weighted total term overlap similarity* ($\mathrm{Sim}_{\mathrm{cm}}$) in their analysis of the phenotypes of people with DEE (**Equation 4-7**).

$$\mathrm{Sim}_{\mathrm{cm}}(A,B) = \sum_{t \in D_A \cap D_B} IC_t \qquad \text{Equation 4-7}$$

This models how a clinician might consider the similarity of two patients to depend solely on all the phenotypic concepts that they can be said to share, while taking account of the discriminatory value of each in some context (their epilepsy clinic, broader experience, or published epidemiology). Once the IC of each term has been calculated, $\mathrm{Sim}_{\mathrm{cm}}$ is computationally simple, making it appropriate for intensive tasks such as the analysis of large cohorts or longitudinal studies with repeated phenotypic similarity analyses (Lewis-Smith *et al.*, 2021b). As a sum of IC, $\mathrm{Sim}_{\mathrm{cm}}$ can take non-negative values with an infinite upper limit.

Graph-IC similarity ($\mathrm{Sim}_{\mathrm{GIC}}$) is an IC-weighted adaptation of the Jaccard index, developed for analysis of GO data (**Equation 4-8**; Pesquita *et al.*, 2007; Pesquita *et al.*, 2008).

$$\mathrm{Sim}_{\mathrm{GIC}}(A,B) = \frac{\sum_{t \in D_A \cap D_B} IC_t}{\sum_{x \in D_A \cup D_B} IC_x}$$ Equation 4-8

$\mathrm{Sim}_{\mathrm{GIC}}$ was implemented in the initial release of PhenomeNET, a tool for identifying orthologous genes and pathways implicated in similar sets of phenotypes by comparing phenotypes across species and models (Hoehndorf;Schofield and Gkoutos, 2011) and in OWLSim (as SimIC, Washington *et al.*, 2009).[15]

In keeping with the Jaccard approach, this is equivalent to $\mathrm{Sim}_{\mathrm{cm}}$ divided by the sum of the IC of each term in the union of the propagated annotations of A and B. Both $\mathrm{Sim}_{\mathrm{UI}}$ and $\mathrm{Sim}_{\mathrm{GIC}}$ take values in the unit interval. As a proportion, these Jaccard measures have some shortcomings in this context. Their values reflect the relative concordance of two people's annotations, penalizing pairs with discordant annotations. This may not be desirable when an open-world interpretation is favored. Similarly, were a clinical feature that both of a pair of people share to be annotated more precisely in one than the other, the calculated similarity of the pair would be lower than if it were annotated uniformly at lower precision. For example, a pair of people with identical phenotypes would appear more similar if annotated identically with *Seizure [HP:0001250]* and *EEG abnormality [HP:0002353]* than if one of the pair were annotated more precisely with *Seizure [HP:0001250]* and *Interictal EEG abnormality [HP:0025373]*. Furthermore, because these measures reflect concordance rather than the absolute informativeness of each shared feature, a pair of people with identical common annotations of little discriminatory value would appear more similar than another pair sharing several remarkable phenotypes with only a minor discordance. For example, the pair annotated identically with *Seizure [HP:0001250]* and *EEG abnormality [HP:0002353]* would appear more similar to another pair, both of whom annotated with *Myoclonic absence seizure [HP:0011150]* and *Telangiectasia [HP:0001009]*, but with one additionally annotated with *Deviation of the 4th finger [HP:0009273]* while the other was annotated its subtype, *Ulnar deviation of the 4th finger [HP:0009278]*.

A common measure of similarity between two sets of categorical data is *cosine similarity*, which considers the Euclidean distance between entities across their annotated features in vector space (Singhal, 2001). It has been applied comparing proteins according to GO

[15] Note that this differs from $\mathrm{Sim}_{\mathrm{IC}}$ discussed as a modification of Resnik's similarity in 4.9.1.

annotations (Ning;Jansen and Gerstein, 2002; Chabalier;Mosser and Burgun, 2007). Cosine similarity can be calculated with ($\mathrm{Sim}_{\mathrm{cosIC}}$) or without ($\mathrm{Sim}_{\mathrm{cos}}$) weighting the value of each term by IC (**Equation 4-9**).

$$\mathrm{Sim}_{\cos}(A,B) = \frac{\vec{A}.\vec{B}}{\left|\vec{A}\right|\left|\vec{B}\right|} = \frac{\sum_{t\in HPO} v_{At} \times v_{Bt}}{\sqrt{\sum_{t\in HPO} {v_{At}}^2 \times {v_{Bt}}^2}} \qquad \text{Equation 4-9}$$

given

$$v_{Et} = \begin{cases} 1 \times w_t & \text{if } t \text{ is annotated to } E \\ 0 \times w_t & \text{if } t \text{ is not annotated to } E \end{cases}$$

$$w_t = \begin{cases} 1 & \text{if unweighted} \\ IC_t & \text{if weighted by IC} \end{cases}$$

In the context of a study of epilepsy phenotypes, one can consider each person to be represented by a vector v of length equal to the number of HPO concepts annotated to the cohort after propagation. The value of each element, v_{Et} indicates whether t is annotated to individual E or not, multiplied by IC_{t} in the information-weighted form. The similarity of two people is given by the dot product of their vectors with cosine adjustment for the number of annotations made to each person.

Regardless of weighting, cosine similarity takes values in the positive closed unit interval and thus shares some of the limitations of Jaccard measures discussed above. The cosine adjustment could be omitted (removing the denominator of **Equation 4-9**) to avoid this scaling. Alone, the dot product of the vectors in the unweighted form would be equivalent to $\mathrm{Sim}_{\mathrm{TOc}}$. In the IC-weighted form, the dot product can be conceptualized as comparable to $\mathrm{Sim}_{\mathrm{cm}}$ but with augmentation weighting due to squaring of IC.

An alternative measure, *Goodall3* (Boriah;Chandola and Kumar, 2008), was used by Andrews *et al.* (2015) in a study of people with developmental disorders. This quantifies the similarity of a pair of people according to concordance not only in the terms with which they have been annotated but also those that they have not been annotated. This generates a vector, G of length equal to the number of (informative) HPO concepts annotated to the cohort after propagation. The value of each element, G_{t} depends on whether t is annotated to both or neither of a pair, or whether its annotation was discordant. Where annotation of t is concordant, G_{t} is weighted by the complement of the square of the probability of observing this concordance. Where

annotation was discordant, the term is censored. The similarity of each pair of individuals is given by the sum of their weighted similarity across all HPO terms (**Equation 4-10**).

$$\mathrm{Sim}_{\mathrm{Goodall3}}(A,B) = \sum_{t \in HPO} G_t \qquad \text{Equation 4-10}$$

given

$$G_t = \begin{cases} 1 - P(t)^2 & \text{if } t \in \mathrm{A} \cap \mathrm{B} \\ 1 - (1 - P(t))^2 & \text{if } t \in (A \cup B)^c \\ 0 & \text{otherwise} \end{cases}$$

While the similarity of two people according to a single term is in the unit interval, the value of their overall pairwise $\mathrm{Sim}_{\mathrm{Goodall3}}$ is non-negative with an upper bound equal to the number of terms annotated to the cohort.

Taking account of concordant lack of a phenotype could be valuable. Sometimes it is the absence of a common phenotypic feature that is distinctive. For example, the presence of generalized tonic-clonic seizures from wakefulness precludes a diagnosis of Self-limited epilepsy with centrotemporal spikes (Specchio *et al.*, 2022). An analysis may attempt to identify (potential protective) factors associated with the absence of a clinically important phenotype such as regression, or intellectual disability.

Andrews *et al.* (2015) justified a closed-world assumption in their analysis on the grounds that all participants in their study had been assessed for the presence of the same set of HPO terms. Their cohort was annotated with 1,350 terms. Even were clinicians able to indicate the presence or absence of all these,[16] this assumption may not be appropriate for phenotypes that are dynamic, late onset, or that require a particular infrequently performed investigation to be detected.

4.8.2 Similarity based on the probability of encountering shared combinations of terms

GO functional similarity ($\mathrm{Sim}_{\mathrm{GOFun}}$) had been defined as the probability of observing a pair of genes whose set of common ancestors includes those of the pair of genes being compared (Lin *et al.*, 2004). Applied to the HPO annotations of people with epilepsy this is equivalent to

[16] Furthermore, they do not unambiguously indicate that these were the only phenotypes sought and excluded.

calculating the proportion of all pairwise combinations of people who share all the clinical features common to the pair being compared. These reference pairs may also share further terms not shared by the pair being compared. It can be converted to the information carried by a combination of terms so that its value increases with greater similarity to yield $\mathrm{Sim}_{\mathrm{GOFun}}$ (**Equation 4-11**).

$$\mathrm{Sim}_{\mathrm{GOFun}}(A,B) = -\log_2 P\left(D_A \bigcap D_B\right) \qquad \text{Equation 4-11}$$

This measure is statistically simple and attractive for a phenotypic analysis because it allows for possible latent associations between phenotypes due to systematic biases in data capture or true associations between phenotypes. For example, it could recognize that the probability of a pair of people having the most informative common ancestors (MICA) of *EEG with generalized epileptiform discharges [HP:0011198]* and *Generalized-onset seizure [HP:0002197]* is likely to be greater than the product of their individual probabilities because they are unlikely to be distributed independently, but are likely co-occur in those people with generalized epilepsies. Given the rules of propagation, the quantity of data to be analyzed can be reduced by refining the intersection of propagated terms of the pair being compared to its minimal set.

Yu *et al.* (2007) proposed a *total ancestry similarity* method ($\mathrm{Sim}_{\mathrm{TA}}$) to compare the GO annotations of proteins. They define total ancestry similarity in the text of their manuscript thus:

> *'Given two proteins that share a specific set of lowest common ancestor nodes in the classification structure, one can count the total number of protein pairs n that also have the exact same set of lowest common ancestors.' (section 2.3.1 in Yu et al., 2007)*

This suggests $\mathrm{Sim}_{\mathrm{TA}}$ is based upon exact matches of lowest common ancestors. Yet, their examples leave this unclear (figures 1F and 4 in Yu *et al.*, 2007). In figure 1F, they explicitly find seven pairs of genes sharing the same set of lowest common ancestors. The calculation provided uses 16 instead, which is in the number of pairs whose complete intersection (rather than set of lowest common nodes) would include the nodes in question. This is in keeping with $\mathrm{Sim}_{\mathrm{GOFun}}$ rather than their definition of $\mathrm{Sim}_{\mathrm{TA}}$. Intuitively, using only exact matches of lowest common ancestors would underestimate the prevalence with which the combination of features of the pair of interest occurs.

A concern for both $\mathrm{Sim}_{\mathrm{GOFun}}$ and $\mathrm{Sim}_{\mathrm{TA}}$ (whether this is interpreted as identical to $\mathrm{Sim}_{\mathrm{GOFun}}$ or not) is the potential to find ties at ceiling in a cohort with variable phenotyping depth and sparse

coverage of auxiliary clinical features. For example, in Epi25 there are 239 phenotypes annotated to only one pair of participants each. Thus, pairs sharing these phenotypes have similarity constrained at the maximum possible in this cohort of 12,611 people, 26.2 bits ($-\log_2(1/{}^{12,611}C_2)$). This is regardless of their other shared annotations. These measures also require calculation, storage, and looking up of the set of (at least the lowest) common ancestors of all pairs of individuals in a cohort, which is computationally demanding for large cohorts. Hereafter, I will use $\mathrm{Sim}_{\mathrm{TA}}$ to refer to the measure described as $\mathrm{Sim}_{\mathrm{GOFun}}$. While $\mathrm{Sim}_{\mathrm{GOFun}}$ was described first and more clearly, we are moving away from the context of the 'functional similarity of protein pairs according to the GO', and 'total ancestry' seems an apt generic description for the measure $\mathrm{Sim}_{\mathrm{GOFun}}$ given that it is a function of the probability of observing the total shared *total ancestry* of pair among pairs in a cohort.

4.9 Semantic similarity

4.9.1 Comparison of pairs of terms

A further node-based approach that can be adapted for the symmetric comparison of two people by their base annotations is semantic similarity. The similarity of A and B is based on a pairwise comparison of each t_i of the m HPO terms from A to each t_j of those n from B, generating a $m \times n$ matrix, $\boldsymbol{S}$. Following this pairwise comparison of terms, various methods of have been used to summarize this matrix as a single value indicating the overall similarity of A and B.

Whereas set-based approaches are based on the sets of propagated annotations of the entities being compared, semantic similarity compares entities by their sets of base annotations without propagation. Instead of propagation being used to harmonize the applicable descriptors of each entity, it is used to infer the descriptors applicable to each annotation: its ancestors, $\mathrm{anc}(t)$. Recall that $\mathrm{anc}(t)$ returns the set of nodes encountered following all directed edges from t to the root of the HPO, including both t and the root.

Semantic similarity methods depend on identification of common ancestors of the two terms being compared, conceptually subsuming them (**Equation 4-12**).

$$\mathrm{CommonAncestors}(t_1, t_2) = \mathrm{anc}(t_1) \cap \mathrm{anc}(t_2) \qquad \text{Equation 4-12}$$

The most established node-based measure of the similarity of two terms is *Resnik's similarity* (SimResnik, Resnik, 1995; Resnik, 1999). $\mathrm{Sim}_{\mathrm{Resnik}}$ defines the similarity of two terms, t_1 and

t_2 as the IC of their MICA. It can be conceptualized as the information that these terms share. The MICA is the term with the greatest IC of those terms intersecting the set of ancestors of term t_1 and the set ancestors of t_2 (**Equation 4-13**).

$$\mathrm{Sim}_{\mathrm{Resnik}}(t_1, t_2) = \mathrm{Share}_{\mathrm{Resnik}}(t_1, t_2) = \mathrm{IC}_{\mathrm{MICA}(t_1,t_2)} \qquad \text{Equation 4-13}$$

given

$$IC_{\mathrm{MICA}(t_1,t_2)} = \max_{x \in \mathrm{anc}(t_1) \cap \mathrm{anc}(t_2)} IC\,(x)$$

The dependence structure of terms linked by *is_a* relationships in the HPO, results in the IC of a term decreases monotonically as one moves from conceptually specific terms to their more generic subsuming ancestors nearer the root. Hence, $\mathrm{Sim}_{\mathrm{Resnik}}$ models how a clinician might consider the similarity of two phenotypes to be equivalent to the informativeness of the most informative of the phenotypic descriptors that can be said to apply to both. For example, when considering how focal visual seizures are similar to focal cognitive seizures a clinician would recognize that both can be considered types of focal non-motor seizure. They may overlook the less conceptually specific applicable descriptors (focal-onset seizure, seizure, etc.) because these are less precise and are implicit once we know that both seizure types compared can be said to be focal non-motor seizures. When quantifying the similarity of a pair of phenotypes by $\mathrm{Sim}_{\mathrm{Resnik}}$, context becomes important: how common each relevant phenotype is in the cohort of reference. Absence seizures are similar to generalized myoclonic seizures in that they are both types of generalized-onset seizure. A pediatric neurologist may conclude that focal visual seizures and focal cognitive seizures are more similar to each other than absence seizures are similar to generalized myoclonic seizures. This is because they encounter more children with generalized-onset seizures than they do children with focal non-motor seizures. An adult neurologist may conclude the opposite. In adult practice, one may have encountered more patients describing some form of non-motor onset to a focal-onset seizure than adults with generalized-onset seizures.

$\mathrm{Sim}_{\mathrm{Resnik}}$ was the first information-based measure of similarity to become established for the pairwise comparison of two terms from a biomedical ontology (Lord *et al.*, 2003a; Washington *et al.*, 2009). It has been used extensively in the HPO-based analyses of phenotype within cohorts discussed previously (Akawi *et al.*, 2015; Westbury *et al.*, 2015; Helbig *et al.*, 2019; Galer *et al.*, 2020; Crawford *et al.*, 2021), and Phenomizer: a popular algorithm (we will return

to in 4.10) for assisting in the prioritization of candidate genetic diagnoses based on a patient's set of HPO annotations (Kohler *et al.*, 2009).

While $\mathrm{Sim}_{\mathrm{Resnik}}$ focuses on the amount of information that indicates t_1 and t_2 are similar, other measures take account of the semantic distance between the terms being compared. These penalize $\mathrm{Sim}_{\mathrm{Resnik}}$ increasingly as the evidence that t_1 and t_2 can be distinguished increases. *Lin's similarity* ($\mathrm{Sim}_{\mathrm{Lin}}$) divides $\mathrm{Sim}_{\mathrm{Resnik}}$ by the sum of the IC of the two terms being compared (**Equation 4-14**; Lin, 1998).

$$\mathrm{Sim}_{\mathrm{Lin}}(t_1, t_2) = \frac{2 \times IC_{\mathrm{MICA}(t_1,t_2)}}{IC_{t_1} + IC_{t_2}} = \frac{2 \times \mathrm{Share}_{\mathrm{Resnik}}(t_1, t_2)}{IC_{t_1} + IC_{t_2}} \qquad \text{Equation 4-14}$$

$\mathrm{Sim}_{\mathrm{Lin}}$ was combined with a regression analysis of HPO terms characterizing the complex phenotype associated with variants in *DIAPH* in bleeding disorders (Greene *et al.*, 2016; Stritt *et al.*, 2016) and is the phenotypic similarity measure used within the Heterogeneous Association Network for Rare Diseases (HANRD, Rao *et al.*, 2018).

Jiang (1997) defined a distance metric from which a similarity can be calculated: $\mathrm{Sim}_{\mathrm{JC}}$ (**Equation 4-15**).

$$\mathrm{Sim}_{\mathrm{JC}}(t_1, t_2) = \frac{1}{IC_{t_1} + IC_{t_2} - 2 \times IC_{\mathrm{MICA}(t_1,t_2)} + 1} \qquad \text{Equation 4-15}$$
$$= \frac{1}{IC_{t_1} + IC_{t_2} - 2 \times \mathrm{Share}_{\mathrm{Resnik}}(t_1, t_2) + 1}$$

Both $\mathrm{Sim}_{\mathrm{Lin}}$ and $\mathrm{Sim}_{\mathrm{JC}}$ penalize $\mathrm{Sim}_{\mathrm{Resnik}}$ of terms being compared by their semantic distance in the subontology of their MICA. Their values span the unit interval because they represent the distance between t_1 and t_2. Consider a clinician comparing the similarity of a pair of individuals, for both of whom there exists only sufficient evidence to diagnose focal-onset seizures, with another pair that can be diagnosed more precisely, one with focal impaired awareness emotional seizures and the other with focal aware clonic seizures. It is possible that, were more evidence available for classification, the first pair might be found to have exactly the same precise subtype of focal-onset seizure. Conversely, we know that the only similarity between the seizures of the second pair is that they are focal-onset seizures. $\mathrm{Sim}_{\mathrm{Lin}}$ and $\mathrm{Sim}_{\mathrm{JC}}$

would capture this difference, returning a greater similarity for the first pair (where $t_1 = t_2$, these measures take the maximum value = 1), unlike $\mathrm{Sim}_{\mathrm{Resnik}}$, which would return the same value for both pairwise comparisons (the IC of *Focal-onset seizure [HP:0007359]*).

That $\mathrm{Sim}_{\mathrm{Lin}}$ and $\mathrm{Sim}_{\mathrm{JC}}$ measure a different concept from $\mathrm{Sim}_{\mathrm{Resnik}}$ becomes clear on considering how they would apply to an extreme example. According to $\mathrm{Sim}_{\mathrm{Lin}}$ and $\mathrm{Sim}_{\mathrm{JC}}$, a pair of people annotated identically with only a generic term such as *Seizure [HP:0001250]* would appear identical with the maximum possible similarity value, 1. This would make them appear more similar than another pair of people annotated with related concepts *Focal aware sensory seizure [HP:0032754]* and *Focal aware emotional seizure [HP:0032734]*, despite the second pair sharing a common ancestor (both could be said to have *Focal aware non-motor seizure [HP:0032682]*) that is conceptually far more informative than the first. Conversely, the second pair would be more similar than the first by $\mathrm{Sim}_{\mathrm{Resnik}}$.

Some measures attempt to capture the semantic distance between t_1 and t_2 in the context of the entire ontology by considering the specificity of the MICA in the body of information (or cohort) of reference. Schlicker *et al.* (2006) developed a measure called *relevance similarity* ($\mathrm{Sim}_{\mathrm{Rel}}$) for the analysis of GO data. In the context of a phenotypic study this would weight $\mathrm{Sim}_{\mathrm{Lin}}$ by the complement of the proportion of individuals annotated with the MICA (**Equation 4-16**).

$$\begin{aligned}\mathrm{Sim}_{\mathrm{Rel}}(t_1, t_2) &= \mathrm{Sim}_{\mathrm{Lin}} \times \Big(1 - P\big(\mathrm{MICA}(t_1, t_2)\big)\Big) \\ &= \mathrm{Sim}_{\mathrm{Lin}} \times \left(1 - \frac{1}{2^{IC_{\mathrm{MICA}(t_1,t_2)}}}\right)\end{aligned} \qquad \text{Equation 4-16}$$

The *information coefficient similarity* ($\mathrm{Sim}_{\mathrm{IC}}$, **Equation 4-17**) and *adjusted-Resnik similarity* ($\mathrm{Sim}_{\mathrm{adjResnik}}$, **Equation 4-18**) have been proposed as alternative measures for capturing a similar concept of similarity to $\mathrm{Sim}_{\mathrm{Rel}}$ for GO data, weighting $\mathrm{Sim}_{\mathrm{Lin}}$ by another measurement of the specificity of the MICA (Li *et al.*, 2010).

$$\mathrm{Sim}_{\mathrm{IC}}(t_1, t_2) = \mathrm{Sim}_{\mathrm{Lin}} \times \left(1 - \frac{1}{1 + IC_{\mathrm{MICA}(t_1,t_2)}}\right) \qquad \text{Equation 4-17}$$

$$\mathrm{Sim}_{\mathrm{adjResnik}}(t_1, t_2) = \mathrm{Sim}_{\mathrm{Lin}} \times \mathrm{Sim}_{\mathrm{Resnik}} = \mathrm{Sim}_{\mathrm{Lin}} \times IC_{\mathrm{MICA}(t_1,t_2)} \quad \text{Equation 4-18}$$

$\mathrm{Sim}_{\mathrm{IC}}$ is scaled over a unit range for the entire set of potential pairwise term comparisons whereas $\mathrm{Sim}_{\mathrm{adjResnik}}$ is bound only by cohort size and the sparsity of phenotypes. As Li *et al.* (2010) suggest, the same adjustment could be applied to $\mathrm{Sim}_{\mathrm{JC}}$.

Whether the penalization of the shared information is at all desirable for a measure of similarity between people with epilepsy by phenotype is doubtful. Phenotypes may be coded at different degrees of specificity due to the challenges of classifying a paroxysmal, often stochastic phenotype or because of diagnostic confidence or arbitrary variability of contributors' depth of phenotyping. A person's focal aware clonic seizure is no less of a focal-onset seizure for being described in more detail than a focal motor, focal aware, or even a focal-onset seizure without further specification. Penalization of more precise annotation is likely to be of benefit only if the depth of phenotyping represents something systematic that is biologically or clinically meaningful. A study of the performance of algorithms in the prioritization of candidate genes based on phenotypic annotations found a $\mathrm{Sim}_{\mathrm{Resnik}}$ to perform best (Slater *et al.*, 2022). In fact, in the molecular realm, $\mathrm{Sim}_{\mathrm{Lin}}$ and $\mathrm{Sim}_{\mathrm{JC}}$ have frequently been outperformed by $\mathrm{Sim}_{\mathrm{Resnik}}$ in tests using genetic sequence, expression profile, and functional data, (Lord *et al.*, 2003b; Sevilla *et al.*, 2005; Guo *et al.*, 2006; Parrini *et al.*, 2006; Mistry and Pavlidis, 2008).

The node-based similarities described above focus on a single MICA to define their shared information. Couto;Silva and Coutinho (2007) argued that when comparing proteins using the GO, it would be preferable to consider all the independent ways in which two terms can be said to be similar when considering their shared information. Their *graph-based similarity measure* ($\mathrm{Share}_{\mathrm{GraSM}}$) seeks to consider all possible paths along the directed edges between the root node of the GO and each of the terms for which the similarity is sought, t_1 and t_2. Ancestors x_1 and x_2 of t are said to be disjunctive ancestors of t if the path from t to x_1 does not contain x_2 and the path from t to x_2 does not contain x_1, Disjunctive ancestors of t capture independent features of t. The common ancestors of t_1 and t_2 are examined to identify the common disjunctive ancestors, that is the MICAs of the union of the disjunctive ancestors of t_1 and t_2. These common disjunctive ancestors represent all the independent ways in which t_1 and t_2 can be said to be similar according to the GO. The final step is to define the shared information of t_1 and t_2 as the mean IC of these disjunctive common ancestors (**Equation 4-19**).

$$\mathrm{Share}_{\mathrm{GraSM}}(t_1, t_2) = \overline{\{IC_x | x \in CommonDisjAnc(t_1, t_2)\}} \qquad \text{Equation 4-19}$$

$\mathrm{Share}_{\mathrm{GraSM}}$ derived shared information of t_1 and t_2 can be substituted for $\mathrm{Share}_{\mathrm{Resnik}}$ (**Equation 4-13**) to modify the similarity measures for $\mathrm{Sim}_{\mathrm{Resnik}}$, $\mathrm{Sim}_{\mathrm{Lin}}$, and $\mathrm{Sim}_{\mathrm{JC}}$. This allows them to take account of the multiple ways (represented by paths in the directed acyclic graph of the ontology) that t_1 and t_2 can be compared independently. Regardless of which of these three similarity measures is modified, using $\mathrm{Share}_{\mathrm{GraSM}}$ in place of $\mathrm{Share}_{\mathrm{Resnik}}$ decreases the similarity of t_1 and t_2 monotonically (Couto;Silva and Coutinho, 2007). This occurs because the IC of additional disjunctive common ancestors beyond the MICA will decrease monotonically. The authors suggest that this is desirable when comparing two gene products because the similarity of two gene products should be lower when one considers all the independent ways in which they can be said to be similar rather than only the single dimension in which they are most alike. This may be the case when analyzing the similarity of thoroughly characterized molecules by their annotations using the GO, an ontology built around objective experimental findings that may represent natural classes more closely than might be expected of the HPO. Couto;Silva and Coutinho (2007) concluded that $\mathrm{Share}_{\mathrm{GraSM}}$ improved the performance of $\mathrm{Sim}_{\mathrm{Resnik}}$, $\mathrm{Sim}_{\mathrm{Lin}}$, and $\mathrm{Sim}_{\mathrm{JC}}$ from their original forms on GO annotations according to the correlation of protein family membership of 500 proteins.

However, when two people are compared according to HPO annotations, one must consider whether differences in their annotated terms and the paths from these to the root via their ancestors reflect real differences in phenotype rather than only differences in how the phenotypes were detected and annotated, and the opportunistic, sometimes arbitrary structure of the HPO. The following examples indicate why in a study of epilepsy using the HPO, it may be preferable to focus the shared information purely on the MICA. Applying $\mathrm{Share}_{\mathrm{GraSM}}$ to HPO annotations, one person with unknown-onset tonic-clonic seizures (annotated with *Bilateral tonic-clonic seizure [HP:0002069]*) will appear more similar to another with unknown-onset tonic-clonic seizures than they do to a third person with focal to bilateral tonic-clonic seizures (annotated as *Bilateral tonic-clonic seizure with focal onset [HP:0007334]*) because of the addition of a second, less informative disjunctive common ancestor term of *Bilateral tonic-clonic seizure with focal onset [HP:0007334]*, (which is *Seizure [HP:0001250]* reached via disjunctive ancestor *Focal-onset seizure [HP:0007359]*). In this case, the difference could be large in a cohort with epilepsy, as the additional common disjunctive ancestor is *Seizure [HP:0001250]* , which would be nearly ubiquitous, carrying little information. Similarly, a person annotated with *Focal impaired awareness seizure [HP:0002384]* would appear much less similar to a person with *Focal impaired*

awareness motor seizure [HP:0032712] than they do to another person with *Focal impaired awareness seizure [HP:0002384].* This is because $\mathrm{Share}_{\mathrm{Resnik}}$ depends on the IC of *Focal impaired awareness seizure [HP:0002384]* alone, while $\mathrm{Share}_{\mathrm{GraSM}}$ would be the mean IC of this, and of notably less informative terms *Focal-onset seizure [HP:0007359]* (considering the path via *Focal motor seizure [HP:0011153]*) and *Seizure [HP:0001250]* (considering the path via *Motor seizure [HP:0020219]*).

Topology-based clustering semantic similarity has been proposed to quantify the similarity of gene products in GO based on the IC of their MICA and weighting this according to its position within the GO (Jain and Bader, 2010). The GO is first divided into subgraphs based on an arbitrary topological threshold IC (level of conceptual specificity). This defines which GO terms become roots of subgraphs within the complete (meta)graph of the ontology. The similarity of pairs of terms belonging to the same subgraph is then augmented in comparison with the similarity of pairs not coinciding in the same subgraph. While the authors report favorable performance using GO annotations of various examples, even when compared with $\mathrm{Sim}_{\mathrm{Resnik}}$, the arbitrary nature of dividing the ontology into subgraphs would be difficult to justify in the HPO and require significant testing for optimization.

A simple measure of similarity based on GO topology was suggested by Zhang *et al.* (2006). In developing a *gene functional similarity search tool*, they defined 'distribution values' for each term. The 'distribution value' of the root term is defined as 1. The distribution value of each term with children is defined as the sum of the distribution values of its children, having allocated each leaf term (those terms that have no children) the same 'distribution value'. In this way, the distribution values of terms increase monotonically as they approach the root. The distribution value (similarity) of two terms being compared is defined as the minimum distribution value of their shared ancestors, or 0 if a term is compared to itself. The distribution value of a pair of genes is asymmetric and would thus need modification for pairwise phenotypic comparisons.[17]

Wang *et al.* (2007a) suggested another measure for calculating the similarity of GO two terms based on the semantic value of the terms and the contribution of their common ancestors to each ($\mathrm{Sim}_{\mathrm{Wang}}$, **Equation 4-20**).

[17] Zhang *et al.* (2006) envisage their measure being used to find target gene products annotated similarly to a set of queried GO annotations. It is defined as the mean of the pairwise comparisons of the query set of GO terms that include only one comparison for each query term. It avoids repeated pairing of the target terms unless these are fewer than the query terms, while minimizing the term-term distribution values.

$$\mathrm{Sim}_{\mathrm{Wang}}(t_1, t_2) = \frac{\sum_{x \in \mathrm{anc}(t_1) \cap \mathrm{anc}(t_2)} \left(\mathrm{S}_{t_1}(x) + \mathrm{S}_{t_2}(x) \right)}{SV_{t_1} + SV_{t_2}} \qquad \text{Equation 4-20}$$

given

$$\begin{cases} \mathrm{S}_t(x) = 1 & \text{if } x = t \\ \mathrm{S}_t(x) = \max\{w_e \times \mathrm{S}_t(x') | x' \in \mathrm{children}(x)\} & \text{if } x \neq t \end{cases}$$

$$SV_t = \sum_{x \in \mathrm{anc}(t)} \mathrm{S}_t(x)$$

$$0 < < 1$$

First, for each of the two terms being compared, the directed acyclic graph from the term to the ontology root is obtained by following all paths to the root. In each directed acyclic graph, the semantic contribution of each of the ancestors, x, to the term being compared, t is calculated according to the minimum number of edges that separate x from t, giving $\mathrm{S_t}(x)$. The semantic contribution of t to itself is defined as 1. Each edge connecting a term to its parent attenuates the semantic contribution of the parent term by a fixed proportion. Thus, the contribution to t of an ancestor, x, decreases with distance from t. The attenuation of semantic contribution with each edge, w_e is defined arbitrarily. If multiple paths from t converge on an ancestor, then the ancestor's contribution is based on the maximum contribution from any one of its children on one of these paths. The semantic value of t, assigned SV_t, is defined as the sum of the semantic contributions of all x to t. The similarity of t_1 and t_2 is defined as the sum of the contributions of the ancestors common to t_1 and t_2 to each of these, divided by the sum of the semantic contribution of all ancestors of t_1 to t_1 and the contribution of all ancestors of t_2 to t_2

Two different values of w_e are used for analysis of similarity using the GO because unlike the HPO, GO has two relationship types: 0.8 was used for *is_a* and 0.6 was used for *part_of*. Subsequently, Zhao and Wang (2018) modified the value of w_e, penalizing the contribution of ancestors with a large number of children.

Mazandu and Mulder (2012) suggested a *GO topology-based information measure* that calculated the information of each term recursively from that of its parents and the number of child terms these parents have. This assumes that if there is only one child of a parent term, then all cases annotated with parent term can also be annotated with the child. This would not be valid in an ontology that does not explicitly capture all natural classes distinctly.

However, these last three measures suffer from the same limitations as discussed for other measures that depend on purely ontology-intrinsic rather than information-based weighting.

4.9.2 Summarizing the similarity of a pair of people after pairwise comparison of their annotations

Pairwise comparison of each combination of two people's terms yields a matrix of similarities, $\boldsymbol{S}$. To provide a single value representing the overall similarity between the two people's sets of HPO annotations, some method of summarizing $\boldsymbol{S}$ must be chosen. Resnik (1995), when studying similarity between words, used the single best match to quantify the similarity of two words as this selects the *maximum similarity* ($\mathrm{Sim}_{\mathrm{Max}}$) among the combinatorial comparisons of the possible concepts represented by those two words (**Equation 4-21**).

$$\mathrm{Sim}_{\mathrm{Max}}(A,B) = \max_{\substack{1\leq i\leq m \\ 1\leq j\leq n}} s_{ij} \qquad \text{Equation 4-21}$$

This is desirable for the comparison of two words which could potentially represent different concepts, but which usually only represent a single concept in each instance. Resnik suggested that the greatest similarity between the sets of concepts can usually be presumed to be the most relevant similarity between two words. Similarly, when assessing the functional similarity of gene products, some authors have used $\mathrm{Sim}_{\mathrm{Max}}$ on the grounds that this is likely to be most relevant (Sevilla *et al.*, 2005; Guo *et al.*, 2006). Applied with $\mathrm{Sim}_{\mathrm{Resnik}}$, this $\mathrm{Sim}_{\mathrm{Max\text{-}Resnik}}$ approach is equivalent to defining the similarity of two annotated entities as the single MICA taken from the intersection of their propagated sets. This is how Cho *et al.* (2007) defined the weights between terms in a network of GO terms. A single best match across all of each organism's phenotypes might be desirable if one expects more variable pleiotropism and expressivity of variants in orthologous genes across species. Applied to HPO annotations, this would be equivalent to defining the overall similarity of two people as the informativeness of the single most specific clinical feature that they share. Akawi *et al.* (2015) applied $\mathrm{Sim}_{\mathrm{Max\text{-}Resnik}}$, combining the results with probabilistic analysis of genotype to identify four genes with significant evidence of causing autosomal recessive developmental disorders.

Note that $\mathrm{Sim}_{\mathrm{Max}}$ ignores all of a pair's clinical features other than the single best match. This can overlook important secondary features and potentially cost power because its distribution of similarities across a cohort could be more discrete than more complex measures. Assuming

systematic differences in annotation between research centers, such extreme weighting of the similarity of two people on a single feature may also augment the similarity of people recruited from the same center if the center were to annotate a particular rare phenotype preferentially. For example, a center might have an academic interest in sleep, and therefore provide thorough phenotypic annotation of sleep phenotypes that are underreported by other centers.

Lord *et al.* (2003a) took an *average overall similarity* ($\mathrm{Sim}_{\mathrm{Avg}}$) when comparing molecules. In contrast to $\mathrm{Sim}_{\mathrm{Max}}$, this takes account of all pairwise comparisons of GO annotations and adjusts for the number of pairwise comparisons of annotations by taking the overall mean (**Equation 4-22**).

$$\mathrm{Sim}_{\mathrm{Avg}}(A,B) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} s_{ij} \qquad \text{Equation 4-22}$$

The authors reasoned that unlike words, which can have different meanings in different instances, all GO annotations of a particular molecule can be said to apply to it in every instance and thus should be included. However, it is unlikely to be optimal to weight all possible pairwise combinations of two people's phenotypes equally. Consider a pair in which one person had *Focal motor seizure [HP:0011153]*, *Neurodevelopmental delay [HP:0012758]*, and *Pes cavus [HP:0001761]* and the other had *Focal motor seizure [HP:0011153]*, *Autism [HP:0000717]*, and *Clinodactyly [HP:0030084]*. In this case, it might be preferable to consider their overall similarity on the basis of only a shared propensity to *Focal motor seizure [HP:0011153]*, a *Neurodevelopmental abnormality [HP:0012759]*, and *Abnormality of limbs [HP:0040064]*. Alternatively, placing equal weighting on all 3^2 pairwise combinations would lead to six generic concepts carrying more weight than these three intuitive similarities.[18] As the number of phenotypically distinct annotations in A or B increases and consequently the size of $\boldsymbol{S}$ increases, $\mathrm{Sim}_{\mathrm{Avg}}(A, B)$ will fall if the additional annotations are phenotypically very different from those already considered (for example, adding shared phenotypic breadth from other organ systems). This could be the case even where an additional annotation to each person is conceptually identical and rare. For example, consider adding *Hepatomegaly [HP:0002240]* to both people in

[18] Comparisons of *Focal motor seizure [HP:0011153]* with each of *Autism [HP:0000717]* and *Neurodevelopmental delay [HP:0012758]* result in a lowest common ancestor (also MICA) of *Abnormal nervous system physiology [HP:0012638]*. The lowest common ancestor/MICA for the remaining four is even less informative, *Phenotypic abnormality [HP:0000118]*.

the pair considered above. This behavior is counter to that desired for replicating the clinical concept of overall phenotypic similarity.

Asymmetric *Best match* algorithms ($\mathrm{Sim}_{\mathrm{BM}}$) compare a set of query terms to a set of target terms. They require all annotations of the query set of terms to contribute exactly once within the asymmetric comparison. The maximum similarity value obtained for each query term is taken, and these are summed (**Equation 4-23**).

$$\mathrm{Sim}_{\mathrm{BM}}(A \rightarrow B) = \sum_{i=1}^{m} \max_{1 \leq j \leq n} s_{ij}$$

$$\mathrm{Sim}_{\mathrm{BM}}(B \rightarrow A) = \sum_{j=1}^{n} \max_{1 \leq i \leq m} s_{ij}$$

Equation 4-23

If comparing a person with epilepsy to a target set of phenotypes (representing another person or diagnostic criteria), this models how a clinician can consider each clinical feature of the person in turn, identifying what is the most similar it can be said to be to any of the target phenotypes. $\mathrm{Sim}_{\mathrm{BM}}$ does not penalize the similarity according to the presence of dissimilar phenotypic features nor need it adjust the similarity for the number of annotations considered. However, without adjusting for the number of annotations provided, a pair of people compared in this way will appear increasingly similar as additional annotations are added to one or both of the pair, so long as the additional annotation has an informative common ancestor (IC > 0) with at least one annotation of the other individual.

For the purpose of generating symmetric pairwise similarities of people in a cohort, we need to summarize the two asymmetric similarities as a single value. Different approaches to summarizing asymmetric $\mathrm{Sim}_{\mathrm{BM}}$ have been proposed. One option, used by Schlicker *et al.* (2006) for the comparison of molecules, has been described as the *best match maximum* (SimBMM, Zhao and Wang, 2018). $\mathrm{Sim}_{\mathrm{BMM}}$ defines the overall symmetric similarity as the greater of the two asymmetric similarities after adjustment by the number of query annotations it was derived from, which can be considered asymmetric *best match averages* (**Equation 4-24** and **Equation 4-25**).

$$\mathrm{Sim}_{\mathrm{BMA}}(A \rightarrow B) = \frac{\mathrm{Sim}_{\mathrm{BM}}(A \rightarrow B)}{m} \qquad \text{Equation 4-24}$$

$$\mathrm{Sim}_{\mathrm{BMM}}(A, B) = \max\begin{cases}\mathrm{Sim}_{\mathrm{BMA}}(A \rightarrow B)\\ \mathrm{Sim}_{\mathrm{BMA}}(\mathrm{B} \rightarrow \mathrm{A})\end{cases} = \max\begin{cases}\dfrac{\mathrm{Sim}_{\mathrm{BM}}(A \rightarrow B)}{m}\\ \dfrac{\mathrm{Sim}_{\mathrm{BM}}(B \rightarrow A)}{n}\end{cases} \qquad \text{Equation 4-25}$$

The Phenomizer tool for prioritizing candidate genetic diagnoses also adjusts for the number of pairwise annotations considered, taking a *best match average* ($\mathrm{Sim}_{\mathrm{BMA}}$) to prioritize candidate diagnoses for a patient (Kohler *et al.*, 2009). In its default setting, it uses an asymmetric $\mathrm{Sim}_{\mathrm{BMA}}$ to compare each of query set of m annotations of a patient, to the n of each disorder from a reference database such as the OMIM (McKusick-Nathans Institute of Genetic Medicine, 2020) or Orphanet (Institut National de la Santé Et de la Recherche Médicale, 1997), adjusting for the m annotations of the patient (**Equation 4-24**). However, Phenomizer can also be used with the symmetric mean of the two asymmetric $\mathrm{Sim}_{\mathrm{BMA}}$, and this is advised where there are more than 5 query terms (**Equation 4-26**).

$$\mathrm{Sim}_{\mathrm{BMA}}(A \leftrightarrow B) = \frac{1}{2}\left(\frac{\mathrm{Sim}_{\mathrm{BM}}(A \rightarrow B)}{m} + \frac{\mathrm{Sim}_{\mathrm{BM}}(B \rightarrow A)}{n}\right) \qquad \text{Equation 4-26}$$

Hereafter, unless otherwise specified, it is this symmetric form of the $\mathrm{Sim}_{\mathrm{BMA}}$ that I refer to as this is more applicable to the unweighted comparison of a pair of people in a research cohort.

Phenomizer is typically used with $\mathrm{Sim}_{\mathrm{Resnik}}$. The application of $\mathrm{Sim}_{\mathrm{BMA}}$ to summarize $\boldsymbol{S}$ of $\mathrm{Sim}_{\mathrm{Resnik}}$ ($\mathrm{Sim}_{\mathrm{Resnik\text{-}BMA}}$) has become common practice in phenotypic analyses. Westbury *et al.* (2015) used the $\mathrm{Sim}_{\mathrm{Resnik\text{-}BMA}}$ to measure the similarity of people with bleeding disorders and converted this into a distance matrix for cluster analysis. $\mathrm{Sim}_{\mathrm{Resnik\text{-}BMA}}$ is also the more recent measure used by PhenomeNET for the mapping of phenotypes across different phenotype and disorder ontologies (Rodriguez-Garcia *et al.*, 2017). Helbig *et al.* (2019) used the $\mathrm{Sim}_{\mathrm{BMA}}$ in a supplementary analysis demonstrating de novo genetic variants in *AP2M1* as a cause of DEE (referring to it as $\mathrm{sim}_{\mathrm{av}}$).

Wang *et al.* (2007a) suggested taking the mean of all the pairwise best matches (*average best match*, $\mathrm{Sim_{ABM}}$) rather than the mean of the means of two asymmetric best matches in the comparison of sets of GO annotations of two gene products (**Equation 4-27**).

$$\mathrm{Sim_{ABM}}(A,B) = \frac{\mathrm{Sim_{BM}}(A \rightarrow B) + \mathrm{Sim_{BM}}(B \rightarrow A)}{m+n} \qquad \text{Equation 4-27}$$

Whereas $\mathrm{Sim_{BMA}}$ weights each asymmetric similarity (each *person*) equally in quantifying the overall symmetric similarity, $\mathrm{Sim_{ABM}}$ weights the best match of each *annotation* of a pair equally.

In their primary analysis, Helbig *et al.* (2019) presented an alternative to $\mathrm{Sim_{BMA}}$ without adjustment for the number of terms annotated to each individual, calling it $\mathrm{sim_{max}}$. This can be considered a *best match sum* ($\mathrm{Sim_{BMS}}$, **Equation 4-28**).

$$\mathrm{Sim_{BMS}}(A,B) = \frac{\mathrm{Sim_{BM}}(A \rightarrow B) + \mathrm{Sim_{BM}}(B \rightarrow A)}{2} \qquad \text{Equation 4-28}$$

Helbig *et al.* (2019) reported $\mathrm{Sim_{BMS}}$ to be more conservative than the $\mathrm{Sim_{BMA}}$ in their study. We used $\mathrm{Sim_{BMS}}$ for the primary analysis within a further study of DEE demonstrating that people with de novo diagnostic variants in the same established DEE gene were significantly similar by this measure (Galer *et al.*, 2020).

An important difference between $\mathrm{Sim_{BMS}}$ and the other methods of summarizing $\boldsymbol{S}$ described is that in $\mathrm{Sim_{BMS}}$ there is no adjustment according to m or n. Instead, best match sum increases monotonically as the number of annotations ($m + n$) increases. The calculated similarity will increase if the additional term shares any informative common ancestor with any of the annotations of the other person. It may seem undesirable for the similarity to increase due to the number annotations. Yet, an advantage of the $\mathrm{Sim_{BMS}}$ is that unlike the normalized measures, it cannot penalize the overall similarity of two people for the addition of further phenotypic annotations, even if these are more generic than the remainder. As more evidence of shared phenotypic features is provided $\mathrm{Sim_{BMS}}$ increases. Thus, it may be considered to have some similarities to the IC-weighted set-based measures that are not penalized by discordance.

Some applications combine multiple similarity measures described above. PhenoDigm (Smedley *et al.*, 2013) is a method for mapping phenotypes between human diseases and existing animal models of disease. It compares a set of human diseases to a library of animal models. For each of these comparisons, PhenoDigm start can start with the $\mathrm{Sim}_{\mathrm{J}}$ (unweighted) or $\mathrm{Sim}_{\mathrm{Resnik}}$ (referred to as 'IC' from OWLSim by the authors) for the pairwise comparison of phenotypic annotations of the human disease and the animal model to generate $\boldsymbol{S}$. It then summarizes $\boldsymbol{S}$ using $\mathrm{SimM}_{\mathrm{ax}}$ and $\mathrm{Sim}_{\mathrm{ABM}}$ (described as the 'Avg score'), standardizing these relative to a hypothetical best possible model of a disease (one replicating the human disease HPO annotations as closely as possible with the analogous Mammalian Phenotype Ontology (MPO) annotations). The overall similarity is the average of the standardized $\mathrm{Sim}_{\mathrm{Max}}$ and $\mathrm{Sim}_{\mathrm{ABM}}$ expressed as a percentage.

4.10 Measures used to prioritize candidate diagnoses

We will now review the measures developed or adapted for quantitative comparison of HPO annotations, and briefly revisit some discussed above. Few have been used for the comparison of people with each other by their HPO annotations. More commonly phenotypic similarity measures have been designed to rank candidate diagnoses. These compare a set of query terms from a patient with references containing the phenotypes associated with these disorders, such as OMIM. In this context, without a reference population annotated with diagnoses and phenotypes, IC_{t} is often calculated as the negative logarithm of the proportion of *disorders* or *genes* annotated with t.

The most established tool for this purpose is Phenomizer, which uses $\mathrm{Sim}_{\mathrm{Resnik}}$ of pairwise combinations of annotations of the patient and of the candidate disorder, summarized using the asymmetric $\mathrm{Sim}_{\mathrm{BMA}}$ in the default setting (Kohler *et al.*, 2009). It also allows a symmetric $\mathrm{Sim}_{\mathrm{BMA}}$. Phenomizer compares a person's HPO annotations with those of a list of *disorders*, then ranks these in order of the similarity of the patient to each. The ranked list of disorders is then used to generate a ranked list of genes. The same similarity measure has been used to compare a person's HPO annotations directly with phenotypic annotations of *genes* (Masino *et al.*, 2014).

Phenomizer was implemented as PhenIX (Zemojtel *et al.*, 2014) within Exomiser (Smedley *et al.*, 2015), which integrates the genetic and phenotypic evidence to prioritize variants in monogenic disease genes as candidate explanations for patient's disorder from exome and HPO

data. It is used to generate candidate gene lists and rank these according to Bayesian analysis of known interactions between gene products in Phen-Gen (Javed;Agrawal and Ng, 2014).

Gong *et al.* (2018) proposed *relative best pair* ($\mathrm{Sim_{RBP}}$) to rank candidate genetic diagnoses based on their annotated phenotypes, again using a database of the HPO terms annotated to a database of monogenic disorders. $\mathrm{Sim_{RBP}}$ is presented as an asymmetric measure of the similarity of a person's set of propagated annotations to the set of propagated annotations of a disorder, Dis. As in Phenomizer, the similarity of the person to each disorder in the database is calculated, enabling these candidate diagnoses to be ranked (**Equation 4-29**).

$$\mathrm{Sim_{RBP}}(A \rightarrow Dis) = \sum_{i=1}^{m} \min\left(\alpha, \mathrm{S}(Dis|t_i)\right) \qquad \text{Equation 4-29}$$

given

$$\mathrm{S}(Dis|t) = \begin{cases} \frac{1}{N_t} & \text{if } t \text{ is among the terms annotated to } Dis \\ 0 & \text{otherwise} \end{cases}$$

Here, N_t is the number of disorders annotated with t after propagation and α is a parameter that limits the maximum weight that can be attributed to a single phenotype. The authors arbitrarily recommend setting α as 0.01, meaning that the weight of any term cannot exceed that of a term annotated to 100 of the disorders in the corpus. However, $\mathrm{Sim_{RBP}}$ can be seen as a term overlap measure, in that it takes account of only the intersection of the two sets of propagated annotations being compared. It does not penalize annotation of terms to a single entity. Consequently, it could be adapted to compare two people A and B (**Equation 4-30**).

$$\mathrm{Sim_{RBP}}(A, B) = \sum_{t \in D_A \cap D_B} \min\left(\alpha, \frac{1}{N(\mathrm{t})}\right) \qquad \text{Equation 4-30}$$

$\mathrm{Sim_{RBP}}$ weights each HPO term by the reciprocal of the number of entities in the corpus it is annotated to, which in any single analysis is proportional to the reciprocal of the proportion of entities. This augments the weighting of very sparsely annotated terms compared to IC, potentially explaining the need for the parameter α to limit the weighting of very rarely encountered annotations. However, the arbitrary nature of α is a weakness of this approach.

Phrank score is another similarity measure developed for the ranking of candidate genes according to a set of query phenotypes. It is based on a Bayesian network built on the intersection of propagated annotation sets for the patient and a gene-defined disorder, taking account of the number of genes annotated with each term (SimPhrank, Jagadeesh *et al.*, 2019). It defines the similarity of two sets of HPO terms as the sum of the negative logarithms of the number of candidate genes annotated with each term t in the intersecting set, conditional on the number of genes annotated with any of its parent terms, $\mathrm{parents}(t)$. As it is symmetric, it could be applied to the calculation of the similarity between individuals in a large cohort based on propagated term sets (**Equation 4-31**).

$$\mathrm{Sim}_{\mathrm{Phrank}}(A,B) = \sum_{t \in D_A \cap D_B} -\log_2 \frac{P(t)}{P\big(\mathrm{parents}(t)\big)} \qquad \text{Equation 4-31}$$

$\mathrm{Sim}_{\mathrm{Phrank}}$ can be thought of as a development of $\mathrm{Sim}_{\mathrm{cm}}$ that can correct for the density of nodes in the HPO by discounting from the IC of each term the information carried by any of its parents.

The *LIkelihood Ratio Interpretation of Clinical AbnormaLities (LIRICAL)* algorithm also uses a Bayesian approach to compare a person's HPO annotations with those annotated to gene-defined disorders taken from OMIM and annotated to syndromic disorders from Orphanet (Robinson *et al.*, 2020). The likelihood ratio that a person has each disorder according to only phenotypic features is quantified according to the person's annotations (categorical data) and the *proportion* of people with and people without that disorder annotated with each phenotype (continuous data) in a reference database. It can consider phenotypes absent among people with the disorder in this calculation. *LIRICAL* is based on a number of assumptions, most notably the independence of phenotypes and that the person being tested will have some latent monogenic disorder to be discovered. The former permits likelihood ratios based on single phenotypes to be multiplied to generate an overall likelihood. The latter allows the baseline proportion of people without the disorder but with the phenotype to be estimated from the pooled frequencies of the phenotype across all disorders in the corpus. *LIRICAL* weights the pairwise similarity of phenotypes using a set of rules that are based on their intrinsic relationship within the HPO. Robinson *et al.* (2020) reported that *LIRICAL* outperformed Phenomizer (their previous tool) for the ranking of candidate genetic diagnoses from published case reports.

PhenoVar is another algorithm that ranks candidate genetic diagnoses for a person according to the *proportion* of people with each phenotype in a reference database for that disorder

(Trakadis *et al.*, 2014). In this case, a simulated cohort with each disorder is generated based on frequencies from reference databases. The algorithm then generates a score based on the number of simulated individuals with exact and fuzzy matches of the patient's phenotypes.

Although *LIRICAL* and *PhenoVar* may perform well for relative quantification of the similarity that a person has to groups of people with defined disorders according to HPO annotations, these measures are not readily applicable to the pairwise comparison of the sets of annotations of single individuals. However, they demonstrate the potential value of large standardized genomic and phenotypic reference data sets such as that being created by Epi25 to enable algorithms such as these to help clinicians to prioritize candidate genetic diagnoses for their patients.

Phevor (Singleton *et al.*, 2014) uses annotations of genes to a number of linked biomedical ontologies including the HPO within a database to rank and expand lists of candidate genetic diagnoses for a person according to fuzzy matching of their raw HPO annotations. The person's annotations are reverse propagated (towards descendants) as well as forwards (towards the root) from their input HPO annotations, setting scores to each node according to the number of edges traversed. The score halves with each edge traversed. At nodes where edges converge, the scores imparted by all edges they receive are summed. The genes annotated to the nodes reached by this bidirectional propagation from the person's annotations are then ranked by their maximum score across the nodes to which they are annotated in a reference database.

One could conceivably compare a person with person in a cohort by substituting the genes in this algorithm with members of the cohort. The propagation backwards towards descendants could potentially mitigate against shallow annotation of some members of the cohort. However, this assumes that each child term of the most specific term annotated to a person has an equal probability of being a valid (latent) description of their clinical feature. For example, consider how if a person were annotated with *Generalized-onset motor seizure [HP:0032677]*, this term would be scored 1 and according to *Phevor*, each of its eight child terms would be weighted equally with ½ regardless of how probable it was that the person had each of these phenotypes. This is likely to perform less well than weighting by IC as some types of generalized-onset motor seizure such as *Bilateral tonic-clonic seizure with generalized onset [HP:0025190]* and *Generalized myoclonic seizure [HP:0002123]* are far more common than others such as *Generalized myoclonic-atonic seizure [HP:0011170]*. Furthermore, with the pairwise similarity of two people based on the maximum single node score (like $\mathrm{Sim}_{\mathrm{Max}}$), it is likely that there would be many ties due to the discrete nature of data collected using categorical variables in a structured data collection sheet. For example, the majority of participants in Epi25 (7,689/12,611 = 61 %) were

annotated with *Focal-onset seizure [HP:0007359]* translated directly from their categorical data. Thus, while in rare cases it is possible that a pair of individuals would have a *Phevor* similarity greater than 1 if multiple paths from the annotations of one individual converge on the annotation of another individual, it is expected that the similarity of many pairs of participants in Epi25 would be based on the exact match of *Focal-onset seizure [HP:0007359]* alone, reducing the discriminatory power between many pairs.

4.11 Refinement of annotations for semantic similarity analysis

Semantic similarity approaches are based on pairwise comparison of each of one person's set of m annotations to each of another's n annotations B generating matrix $\boldsymbol{S}$. Authors often summarize how phenotypic annotations were made by clinicians or how features were extracted from clinical records by automated pipelines. However, publications often lack complete descriptions of how these sets of annotations were defined for application in a semantic similarity calculation. This can be important for semantic similarity measures.

Consider how a person may be annotated with *Focal-onset seizure [HP:0007359]* several times from different elements of clinical notes (diagnostic lists, problem lists, indications within a medication list, and narrative descriptions). Similarly, in the mapping of research data, it may be repeatedly annotated from different variables.[19] Epi25 data collection tools include 51 possible categorical data values spanning syndromes, seizure classification, and summaries of the evidence to support a diagnosis of focal epilepsy that can be mapped to *Focal-onset seizure [HP:0007359]*. Furthermore, some individuals have free-text descriptions of focal-onset seizures that cannot be further classified for more precise annotation. Thus, duplicate annotations can reflect raw data structure and its translation into HPO terms rather than carrying any phenotypic meaning. Hence, some authors have explicitly removed duplicate annotations (Helbig *et al.*, 2019; Galer *et al.*, 2020; Crawford *et al.*, 2021). The inclusion or exclusion of duplicated base set annotations affects the semantic similarity of two people (**Table 4-2** and **Table 4-3**).

[19] Particularly if redundancy is built into data dictionaries to mitigate incomplete data entry.

$\mathrm{Sim}_{\mathrm{Resnik}}$		Person *B*		
		Bilateral tonic-clonic seizure with focal onset [HP:0007334]	*Abnormality of brain morphology [HP:0012443]*	*Focal non-motor seizure [HP:0032679]*
Person *A*	*Focal-onset seizure [HP:0007359]*	0.68686	0.00000	0.68686
	Cortical dysplasia [HP:0002539]	0.00000	1.6355	0.00000
	Focal-onset seizure [HP:0007359]	0.68686	0.00000	0.68686

Table 4-2. $\mathrm{Sim}_{\mathrm{Resnik}}$ with a duplicated annotation.

Term sets	**$\mathrm{Sim}_{\mathrm{Resnik}}$**					
Summarization	**SimMax**	**SimAvg**	**SimBMM**	**SimBMA**	**SimABM**	**SimBMS**
Unique terms in *A* and *B*	1.6355	0.50153	1.1612	1.0821	1.0663	2.6658
With addition of a single duplicated term in *A*	1.6355	0.48699	1.0031	1.0031	1.0031	3.0092

Table 4-3. The effect of including a duplicated annotation (*Focal-onset seizure [HP:0007359]*) in semantic similarity measures based on $\mathrm{Sim}_{\mathrm{Resnik}}$.

One can see how $\mathrm{Sim}_{\mathrm{Max}}$ does not change as the maximum s_{ij} of the $m \times n$ pairwise comparisons is unaltered by duplicating a row or column of $\boldsymbol{S}$. $\mathrm{Sim}_{\mathrm{BMS}}$ increases monotonically as m or n increases because it does not make any adjustment for the number of annotations. However, the other measures change as m and n appear in their denominators. In this example based on $\mathrm{Sim}_{\mathrm{Resnik}}$, these decrease with inclusion of the duplicated *Focal-onset seizure [HP:0007359]*. This results from the modest increase in their numerators relative to their denominators. Conversely, these values increase with inclusion of the same duplicated annotation if the same comparison is made according to $\mathrm{Sim}_{\mathrm{Lin}}$ (**Table 4-4** and **Table 4-5**).

$\mathrm{Sim}_{\mathrm{Lin}}$		Person *B*		
		Bilateral tonic-clonic seizure with focal onset [HP:0007334]	*Abnormality of brain morphology [HP:0012443]*	*Focal non-motor seizure [HP:0032679]*
Person *A*	*Focal-onset seizure [HP:0007359]*	0.54695	0.00000	0.44203
	Cortical dysplasia [HP:0002539]	0.00000	0.48769	0.00000
	Focal-onset seizure [HP:0007359]	0.54695	0.00000	0.44203

Table 4-4: $\mathrm{Sim}_{\mathrm{Lin}}$ with a duplicated annotation.

Term sets	$\mathbf{Sim}_{\mathbf{Lin}}$					
Summarization	**SimMax**	**SimAvg**	**SimBMM**	**SimBMA**	**SimABM**	**SimBMS**
Unique terms in *A* and *B*	0.54695	0.24611	0.51732	0.50477	0.50226	1.2557
With addition of a single duplicated term in *A*	0.54695	0.27396	0.52720	0.50971	0.50971	1.5291

Table 4-5. The effect of including a duplicated annotation (*Focal-onset seizure [HP:0007359]*) in semantic similarity measures based on $\mathrm{Sim}_{\mathrm{Lin}}$.

This is because of the short semantic distance (the difference in IC between a term and its ancestor) between the additional annotation, *Focal-onset seizure [HP:0007359]* and its MICA, in comparison with *Bilateral tonic-clonic seizure with focal onset [HP:0007334]* and *Focal non-motor seizure [HP:0032679]*. Recall that in contrast to $\mathrm{Sim}_{\mathrm{Resnik}}$, it is the semantic distance between annotations and their MICA that $\mathrm{Sim}_{\mathrm{Lin}}$ reflects, rather than the shared information of annotations.

Extending this beyond duplicated base terms, some authors have explicitly refined each person's annotations to the minimal set necessary and sufficient for capturing all of the phenotypic concepts that can be said to apply to them by propagation (Greene *et al.*, 2016). Thus, in refining the minimal set A' of A, a term t_a from A is retained only if it is not implied by any other term in A (**Equation 4-32**).

$$\mathrm{A}' \subseteq A \qquad \text{Equation 4-32}$$

such that

$$\mathrm{A}'\{t \in D_A \backslash \mathrm{parents}(D_A)\}$$

$$\mathrm{A}'\{t \in D_A : t \notin \mathrm{parents}(D_A)\}$$

$\mathrm{parents}(t)$ returns the parents of t, those terms represented by nodes reached by following a single directed edge from t.

For example, *Focal motor seizure [HP:0011153]* would be excluded from the terms representing a person if that person were also annotated with *Focal automatism seizure [HP:0032898]* or any of its other descendants. In this way, assuming that the HPO represents human clinical phenotypes as physicians conceptualize them, the minimal set of terms of a person is analogous to the set of descriptors a physician would use to describe a patient to an expert colleague as succinctly as possible. Removal of implicitly duplicated annotations affects semantic similarity measures analogously to explicitly duplicated annotations. Phenomizer appears to refine the annotations provided. The query set of *Aplasia of the distal phalanx of 2nd finger [HP:0009565]*, *Renal hypoplasia [HP:0000089]*, *Focal seizures [HP:0007359]*, *Mild intellectual disability [HP:0001256], Abnormality of the helix [HP:0011039]*, *Abnormality of the finger [HP:0001167]*, *Abnormal renal morphology [HP:0012210]*, *Seizures [HP:0001250]*, *Intellectual disability [HP:0001249]*, and *Abnormality of the ear [HP:0000598]* consists of specific descriptions of five phenotypes and five more generic descriptors of the same phenotypes. Submission of this using the symmetric Resnik option without specifying a mode of inheritance returns the set of results and *p*-values[20] regardless of whether the five conceptually redundant terms are included or manually excluded (Kohler, 2009).

Some studies, including Helbig *et al.* (2019); Galer *et al.* (2020); Crawford *et al.* (2021), define m base annotations in A, as the unique HPO terms translated from the raw data, regardless of whether some t_i in A is implied by another term in A, $t_{i'}$. There are situations in which retention of duplicated base annotations or those that are ancestors of other base annotations may be desirable. Their retention could capture distinct subtypes of a phenotype in a person, in cases

[20] Five X-linked 'Mental retardation' disorders associated with *ZNF711* and one autosomal recessive 'Mental retardation' disorder associated with *METTL23*, all returning Benjamini-Hochberg $p = 0.1335$.

where one cannot be annotated more precisely as a conceptually independent term. Conceivably this can occur because a concrete phenotypic concept is not yet represented as a distinct class in the HPO. Alternatively, the need to map a phenotype to a term that can be assumed to be correct may make it impossible to capture all the information in cases of clinical uncertainty. In cases of uncertainty, a clinician may have to accept an annotation representing a broader concept than they feel applies. Consider a person with two types of focal-onset seizure. Half of their focal-onset seizures can be classified precisely as a *Focal impaired awareness sensory seizure with visual features [HP:0032806]* because the patient reports seeing rapidly expanding blue balls before losing awareness. Yet in their other focal-onset seizures awareness is lost without any reported symptom or observable motor onset. The clinician may not feel it appropriate to describe these any more precisely than *Focal impaired awareness non-motor seizure [HP:0032716]*. Thus, the patient has two distinct types of *Focal impaired awareness non-motor seizure [HP:0032716]*. In this case, were base terms to represent independent phenotypes of the person rather than independent concepts, one can argue that the base term set for semantic similarity analysis should include both *Focal impaired awareness sensory seizure with visual features [HP:0032806]* and *Focal impaired awareness non-motor seizure [HP:0032716]* despite the latter being an ancestor of the former. However, applying such an approach consistently may require overwhelming manual review of large quantities of raw data. The cost of excluding a possible additional base term for semantic similarity analysis is the opportunity to identify further similarity. Depending on the terms annotated and compared, and the method used to summarize two people's similarities, exclusion may even result in greater rather than reduced similarity, as discussed above.

The simpler approach of using the minimal set of base annotations offers an alternative pragmatic and systematically consistent strategy. Furthermore, using the minimal set is valid if interpreting a person's annotation to indicate *a tendency to manifest* the relevant class of phenotype, *regardless of the number of distinct subtypes of this phenotype* the person may have manifested. This is how set-based similarity analyses interpret HPO annotations: sets of propagated annotations do not include duplicate terms. Furthermore, there is not a one-to-one relationship of clinical features in the raw data and HPO annotations. Not infrequently a single clinical feature is mapped to multiple HPO phenotypic terms, each describing a different component or dimension of the symptom or sign. For example, there are no single HPO terms that allow encoding of different types of nocturnal seizure. Capture of a tendency to nocturnal tonic-clonic seizures requires coding of both *Bilateral tonic-clonic seizure [HP:0002069]* and *Nocturnal seizures [HP:0031951]*.

Refinement to the minimal set offers a means of standardizing the selection of HPO annotations to be used for analyses efficiently regardless of the process of raw data capture without losing any qualitative information about the phenotypes they can be said to have manifest. However, this is particular to a constant version of the HPO. The HPO is in continuing development. It is possible to lose information when a person's HPO annotations translated from raw data are summarized to a minimal set using one version of the HPO and then this set is reanalyzed using a different HPO version. Consider how two versions of the HPO could have different *is_a* edges between the person's translated HPO annotations so that refinement according to each version resulted in different minimal sets. Using a different version for propagation to that used for refinement could fail to recall all the translated phenotypic annotations that would have been obtained by propagation from the original raw data with the initial HPO version. Hence, while minimal sets of annotations are useful for standardization of semantic similarity measures and may be desirable for conceptual reasons, it is useful to store the HPO annotations translated from raw data should a future analysis using an improved version of the HPO be desired.[21] At such a time, refinement and propagation should be undertaken using a single HPO version.

4.12 Introducing minimal term overlap similarity: a node set-based similarity measure modeled on clinical communication

I propose a new phenotypic similarity measure, *minimal term overlap similarity* ($\mathrm{Sim}_{\mathrm{MinTO}}$). Like set-based measures (and semantic similarity measures based on minimal sets of annotations) $\mathrm{Sim}_{\mathrm{MinTO}}$ treats each annotation as a liability to a particular class of phenotype. As a term overlap measure, it models how a clinician comparing two individuals (or a single person with a set of diagnostic criteria) might consider all the ways in which they can be described as having a phenotype in common by taking the intersection of their propagated HPO annotations, A and B. This set of shared phenotypic descriptors is refined to only the minimal set of HPO terms necessary and sufficient to infer the common ancestors of A and B by propagation. $\mathrm{Sim}_{\mathrm{MinTO}}$ is given by summation of the weights of this minimal set of the term overlap (**Equation 4-33**). In the first instance, we shall take each term's IC as a natural data-derived weighting of informativeness.

[21] Ideally, HPO terms translated from raw data should be stored with the data they were annotated to. This facilitates review of the evidence for a particular annotation.

$$\mathrm{Sim}_{\mathrm{MinTO}}(A,B) = \sum_{t \in D_A \cap D_B \setminus \mathrm{parents}(D_A \cap D_B)} \mathrm{w}_t \qquad \text{Equation 4-33}$$

given

$$D_A \cap D_B \setminus \mathrm{parents}(D_A \cap D_B) = \{t \in D_A \cap D_B : t \notin \mathrm{parents}(D_A \cap D_B)\}$$

$$w_t = IC_t$$

$\mathrm{Sim}_{\mathrm{MinTO}}$ simulates how a clinician, possessing expertise in clinical phenotypes and the ontological relationships between them, might summarize the similarities as concisely as possible when communicating with a colleague possessing the same external knowledge of clinical phenotypes and the relationships between their applicable labels. For example, when referring two siblings, one with focal clonic seizures and the other with focal tonic seizures, to a colleague, a clinician might note that both share a phenotype of 'focal motor seizures'. However, they are unlikely to explicitly state that both have also manifested 'focal-onset seizures' or 'seizures' as to an expert recipient, communication of these is redundant. The recipient is able to infer these shared phenotypes if required (for example, when considering if the siblings are eligible for a new treatment licensed for 'focal-onset seizures').

Like other measures based on comparison of people's propagated sets of annotations, $\mathrm{Sim}_{\mathrm{MinTO}}$ benefits from neither being sensitive to the degree of refinement of a person's HPO annotations after translation of raw data nor requiring calculation of matrix $\boldsymbol{S}$ of pairwise comparisons by semantic similarity.

Furthermore, like $\mathrm{Sim}_{\mathrm{cm}}$, $\mathrm{Sim}_{\mathrm{MinTO}}$ mitigates penalization of discordant annotations to suit an open-world assumption about a person's current or past undocumented phenotypes or those that they might manifest in future. Rather than reducing the similarity value of two people based on a discordant HPO annotation on the assumption that one never has nor will manifest it because it has not been annotated, the discordance is penalized only in that it is a potential missed opportunity to identify further phenotypic similarity. Practically, this dependence on only shared phenotypes means that all HPO terms annotated to the propagated set of only one person, which may be many given the sparsity of precise phenotypes (553 of 1,738 phenotypes, 31.2 %, in Epi25), can be excluded from analyses to improve efficiency.

A benefit of $\mathrm{Sim}_{\mathrm{MinTO}}$ is that unlike total overlap measures, the density of nodes in the HPO affects only the resolution at which shared phenotypes are captured. Consider the comparison of a pair of people sharing a rare HPO annotation that has many ancestors with another pair

who share an identically rare annotation with few ancestors. Total overlap measures will tend to exaggerate the similarity of the first pair, that with an annotation from a subontology of the HPO with a higher density of nodes. The variable capture of non-neurological phenotypes in epilepsy studies could contribute by augmenting the difference in non-neurological subontologies with high nodal density. For example, consider *Camptodactyly of finger [HP:0100490]*, a precise phenotype with 32 ancestors in the HPO. It is annotated to two of 12,611 people in the Epi25 cohort, a frequency of 0.016 %, but has a prevalence of 1.5 % in the general population of children (Ekblom;Laurell and Arner, 2010). As an appendicular and sometimes subtle or surgically corrected musculoskeletal condition, it is likely to be overlooked or considered irrelevant by those collating phenotypic data, be this for clinical epilepsy management or a research study focusing on epilepsy.[22] The two people in Epi25 sharing documentation of *Camptodactyly of finger [HP:0100490]* would have a $\mathrm{Sim}_{\mathrm{cm}}$ of 292.7 bits, purely from this annotation. This is greater than the $\mathrm{Sim}_{\mathrm{cm}}$ of 243.6 bits from 42 shared phenotypic descriptors of a pair of people sharing a complex clinical constellation of *Atonic seizure [HP:0010819]*, *Focal hemiclonic seizure [HP:0006813]*, *EEG with photoparoxysmal response [HP:0010852]*, *Polymicrogyria [HP:0002126]*, *Dystonia [HP:0001332]*, *Abnormal ear morphology [HP:0031703]*, and *Hypospadias [HP:0000047]*. However, using IC-weighted $\mathrm{Sim}_{\mathrm{MinTO}}$, the first pair would share 12.62 bits and the second pair 52.61 bits, recognizing that the combination of seven conceptually independent phenotypes spanning multiple organ systems is more clinically remarkable than the single shared finding of a rare finger phenotype.

However, there is a risk of removing the contribution of conceptually redundant terms from the similarity value. Occasionally, a pair of people sharing a rare HPO term t with multiple parent terms could appear less similar to each other than a pair who share some of these parents in the absence of t itself. For example, consider some types of posterior fossa malformation (**Figure 4-2**).

[22] While possible, it seems less likely that there is an inverse association between Camptodactyly and epilepsy in the general population given that people with epilepsy, and in particular genetic epilepsies, have a higher prevalence of morphological phenotypes, leading to the guidance for epilepsy clinicians to consider genetic testing in those with dysmorphology (Krey *et al.*, 2022).

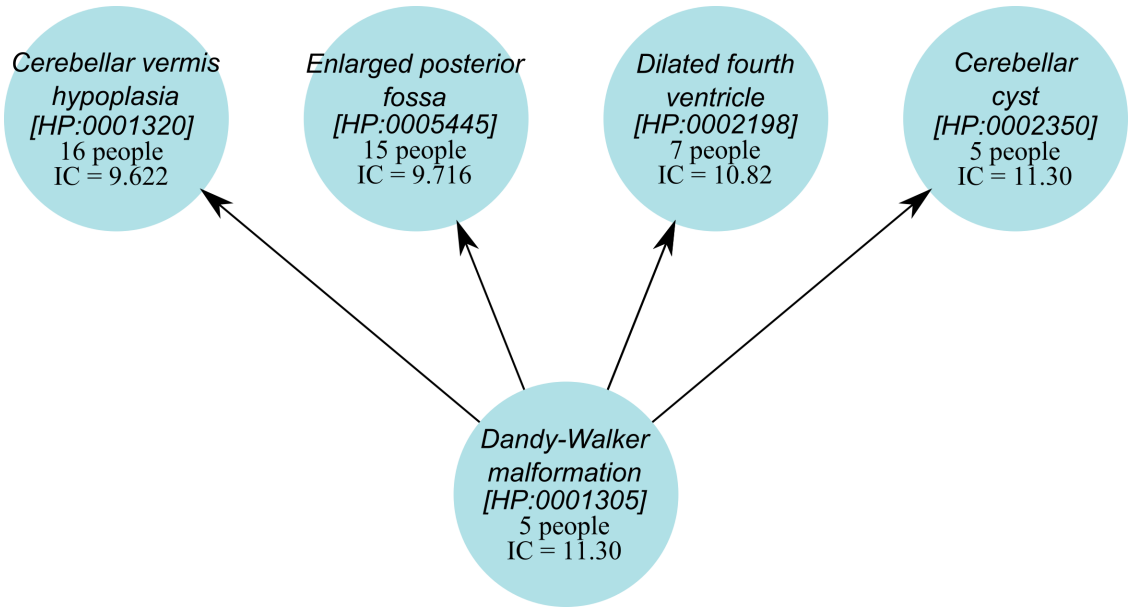


Figure 4-2. In isolation, a term may be less informative than the sum of its parents, those most specific terms it implies. In this example, the IC of *Dandy-Walker malformation [HP:0001305]* is 11.30, whereas the sum of the IC of four of its mandatory component phenotypes is 41.46 (= 9.622 + 9.716 + 10.82 + 11.30).

Dandy-Walker malformation [HP:0001305] is a conceptually specific form of posterior fossa malformation with multiple features (Klein *et al.*, 2003). After propagation it is annotated to five people in Epi25, resulting in an IC of 11.30 bits. This term has four parents reflecting the mandatory features of this malformation: *Cerebellar vermis hypoplasia [HP:0001320]*, *Enlarged posterior fossa [HP:0005445]*, *Dilated fourth ventricle [HP:0002198]*, and *Cerebellar cyst [HP:0002350]*. One can see how two people who share any two (or more) parents of *Dandy-Walker malformation [HP:0001305]* in their propagated annotation sets without being annotated with *Dandy-Walker malformation [HP:0001305]* itself could have a greater $\mathrm{Sim}_{\mathrm{MinTO}}$ than two other people, who have *Dandy-Walker malformation [HP:0001305]*.[23] This is undesirable. The addition of a more conceptually specific description of the clinical features two people share should result in an increase in their similarity value (as it would for $\mathrm{Sim}_{\mathrm{cm}}$), or at least not a reduction. This can be avoided by applying a monotonicity adjustment to the frequency-derived IC of each term. In this manner we define adjusted-IC recursively as the maximum of its unadjusted IC and the sum of the adjusted-IC across its parent terms (**Equation 4-34**).

[23] It is not uncommon to find variation in how radiologists report their findings and willingness to use eponyms. One can envisage how the example could arise where a radiology report describes the features of Dandy-Walker syndrome without using the eponymous label.

$$\mathrm{Sim}_{\mathrm{MinTO}}(A,B) = \sum_{t \in D_A \cap D_B \setminus \mathrm{parents}(D_A \cap D_B)} adjusted\text{-}IC_t \qquad \text{Equation 4-34}$$

given

$$adjusted\text{-}IC_t = IC_t \qquad \text{if } |\mathrm{parents}(t)| = 0$$

$$adjusted\text{-}IC_t = \max\left(IC_t, \sum_{x \in \mathrm{parents}(t)} adjusted\text{-}IC_x\right) \qquad \text{if } |\mathrm{parents}(t)| \neq 0$$

The adjustment does not affect only rare, highly specific terms with several parents. The adjustment applied to 363 of 1,181 informative terms annotated to multiple people (30.7 %) in a $\mathrm{Sim}_{\mathrm{MinTO}}$ analysis of 12,611 participants of Epi25. The median adjustment among these was 7.58 bits (range 0.000250–165 bits). These terms included common epilepsy phenotypes with only two parents. The most common of these was *Focal impaired awareness seizure [HP:0002384]*. Over half of people who can be said to have *Focal-onset seizure [HP:0007359]* and over half of those with *Dialeptic seizure [HP:0011146]* have their child term, *Focal impaired awareness seizure [HP:0002384]*, the subtype of focal-onset seizure with a dialeptic component. Consequently, if both *Focal-onset seizure [HP:0007359]* and *Dialeptic seizure [HP:0011146]* are in the minimal set summarizing the shared phenotypes of a pair of individuals, then without adjustment of IC, this pair of individuals would have greater $\mathrm{Sim}_{\mathrm{MinTO}}$ than they would if the data enabled annotation of the more specific *Focal impaired awareness seizure [HP:0002384]*. Similarly, adjustment was required for *Focal aware non-motor seizure [HP:0032682]*, and for common generalized-onset seizures: *Generalized-onset motor seizure [HP:0032677]* and *Bilateral tonic-clonic seizure with generalized onset [HP:002519]*, with the latter being affected by the adjustment of the former (**Figure 4-3**).

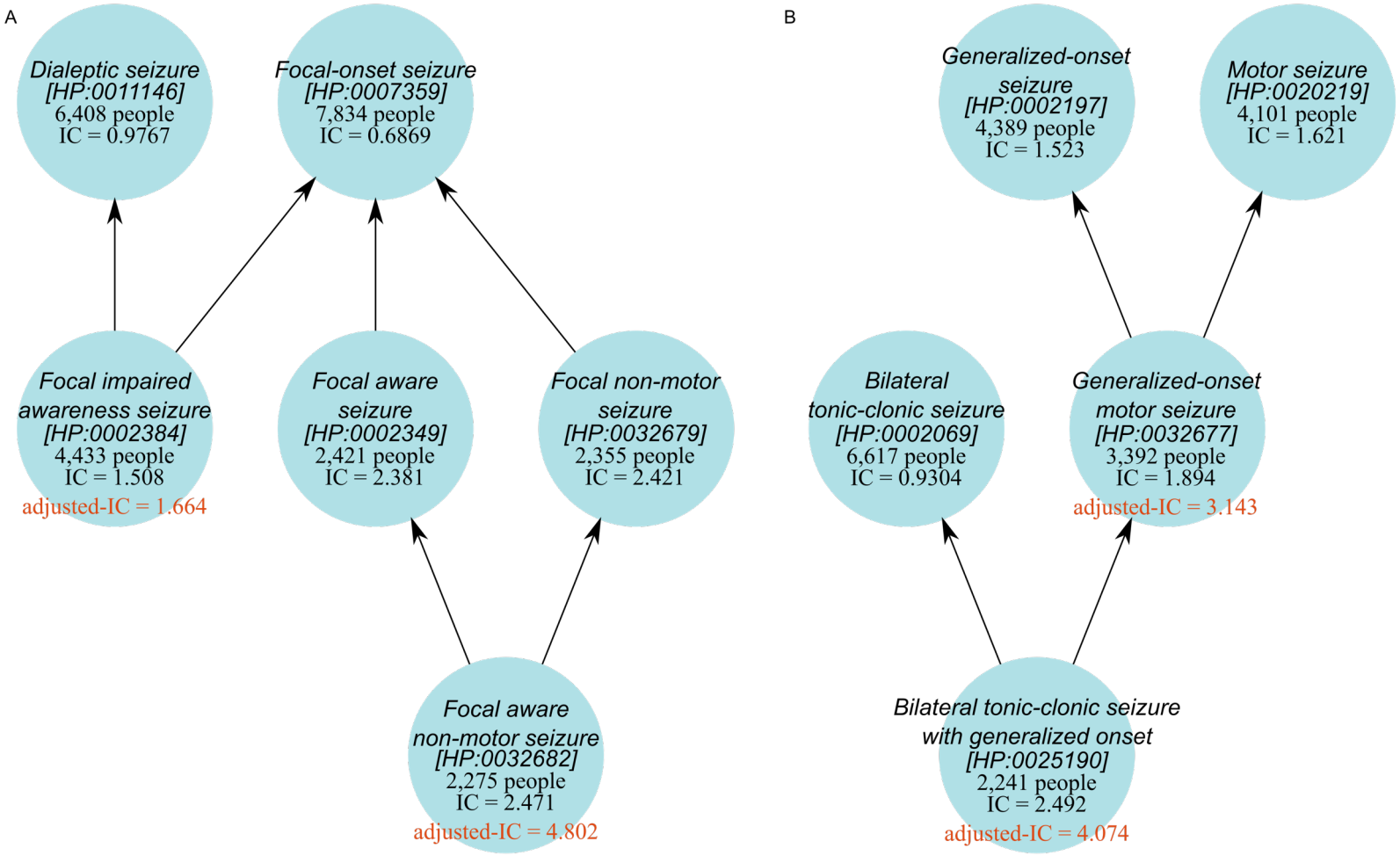


Figure 4-3. Using $\mathrm{Sim}_{\mathrm{MinTO}}$ with monotonic adjustment of IC. Monotonic adjustment ensures that two people do not appear less similar based on sharing a more precise type of seizure than they would by sharing its multiple (less specific) parents in its absence. Common phenotypes for which this is significant in Epi25 include *Focal impaired awareness seizure [HP:0002384]*, 1.508 < 1.664 (= 0.9767 + 0.6869); *Focal aware non-motor seizure [HP:0032682]*, 2.471 < 4.802 (= 2.381 + 2.421); *Generalized-onset motor seizure [HP:0032677]*, 1.894 < 3.143 (= 1.523 + 1.621); and these can propagate *Bilateral tonic-clonic seizure with generalized onset [HP:0025190]*, 2.492 < 4.074 (= 0.9304 + 3.143, rather than 0.9304 + 1.894).

$\mathrm{Sim}_{\mathrm{MinTO}}$ has some undesirable behaviors. According to the properties of logarithms, without adjustment of IC, $\mathrm{Sim}_{\mathrm{MinTO}}$ is the logarithm of the product of the proportions of the cohort with each non-redundant shared phenotype of the pair. This is analogous to calculating the probability of observing two individuals sharing that set of non-redundant phenotypes, assuming that those phenotypes were annotated independently. Consequently, using monotonically adjusted-IC in $\mathrm{Sim}_{\mathrm{MinTO}}$ is equivalent to assuming that annotation of a phenotype with multiple parents should be considered at least as informative as it would be should those parent phenotypes occur independently of each other. For example, the adjusted IC of *Focal aware non-motor seizure [HP:0032682]* considers the probability of two people sharing this seizure type to be equivalent to the probability of two people sharing both *Focal aware seizure [HP:0002349]* and *Focal non-motor seizure [HP:0032679]*, while assuming that the probability of being annotated with either of these two is independent of the probability of being annotated with the other (**Figure 4-3**). The effect of this is that *Focal aware non-motor seizure [HP:0032682]* is weighted as if it were annotated to 3.6 % rather than 18 % of the cohort. The difference reflects that its parents are annotated far from independently: only 226 of 2,501 people (9.0 %) annotated with either *Focal aware seizure [HP:0002349]* or *Focal non-motor seizure [HP:0032679]* are not annotated with *Focal aware non-motor seizure [HP:0032682]*.[24] This reflects the frequently annotated categorical variable in the Epi25 Focal epilepsy REDCap form, *aura_seizures*, with the human label 'Aura (simple partial)', the pre-2017 ILAE term for the concept now best subsumed within 'focal aware non-motor seizure' (Commission on Classification and Terminology of the ILAE, 1981; Fisher *et al.*, 2017a).

We see that the cost of monotonic adjustment of IC is that some specific terms that have many paths along the *is_a* edges to the HPO root (implying that ancestors may have multiple parentage) may have their IC considerably inflated. The adjustment of IC will be particularly notable for sparsely annotated terms distal to subontologies with a high density of nodes with multiple parentage. Of the terms annotated to multiple participants in Epi25, this applies to the terms describing musculoskeletal abnormalities of the extremities, most notably in

[24] Jaccard's index of those with *Focal aware seizure [HP:0002349]* and *Focal non-motor seizure [HP:0032679]* greater than this would suggest (2,276/2,500 = 0.91), albeit with the intersection only increased by one person because not all people with these two terms will necessarily have *Focal aware non-motor seizure [HP:0032682]*. The person annotated with *Focal aware seizure [HP:0002349]* and *Focal non-motor seizure [HP:0032679]* without *Focal aware non-motor seizure [HP:0032682]* was also annotated with *Epileptic aura [HP:0033348],* from their free-text description. The HPO interprets 'epileptic aura' conservatively because it cannot be assumed to be 'simple' (i.e., to remain a focal aware non-motor seizure) because awareness may become impaired subsequently within the same seizure.

Camptodactyly of finger [HP:0100490], which has an adjusted-IC of 177.3 bits rather than its naïve IC of 12.6 bits. While this weight is large compared to the 12.6 bits that would apply to $\mathrm{Sim}_{\mathrm{Resnik}}$, it is 177.3 bits is less than the 292.7 bits that it would be sufficient to attract according to $\mathrm{Sim}_{\mathrm{cm}}$.

While there may be capacity for further optimization of term weighting in $\mathrm{Sim}_{\mathrm{MinTO}}$, it remains attractive as a model of how clinicians conceptualize similarity and forms the primary measure used in the chapters that follow.

4.13 Phenotype first and genotype first approaches to similarity analysis

The similarity of a pair of people can be used for phenotype first approaches such as cluster analysis to identify individuals with similar phenotypic features who might be more likely to share an etiology. This phenotype first strategy formalizes the traditional paradigm of clinicians identifying phenotypically similar people, defining a syndromic diagnosis, hypothesizing a shared genetic basis, and *subsequently* performing genetic assays to identify a shared genetic feature to which the syndrome can be attributed based on association. The traditional approach is based on manual assignment of people to phenotypically defined groups and potentially liable to confirmation or observer bias. Cluster analyses takes a formal data-driven, potentially less humanly biased approach.

However, in clustering of people into groups from which to develop and test hypotheses using genomic data, one has to decide on parameters such as the number clusters to define, often arbitrarily, and a person's membership of a cluster of interest may be sensitive to the number of clusters selected by the investigator. Furthermore, with a phenotype first paradigm, it can remain challenging to overcome biases resulting from structured data collection and variation in annotation of phenotypes between centers. For example, for the benefit of minimizing the burden of data entry, Epi25 uses three different data collection tools spanning generalized and mixed epilepsies, focal epilepsies, and DEEs. Clinicians entering data choose the most appropriate of these three based on their own broad categorization of the participant. Consequently, clustering of individuals according to their phenotypes has a strong tendency to prioritize partitioning by data collection form. Furthermore, they would be expected to tend to cluster by center because the data were often imported from local research databases or clinical record systems. Consequently, the annotations are likely to reflect the structure and focus of these.

Genomic variants are also expected to cluster geographically due to the relationship between geography and ancestry. Consequently, if insufficient care is taken to avoid misclassification of ancestry-specific polymorphisms as rare variants of interest, associations between apparently rare variants and phenotypes due to geographic confounding may be misinterpreted as causal. Beyond the biases of recruiting centers, there are other indirect explanations for associations between phenotypes and rare variants associated with geography. For example, the ancestry-specific rare variant may be in linkage disequilibrium with unmeasured ancestrally stratified polygenic liability to some epilepsy subtypes (Leu *et al.*, 2019) or geographically stratified environmental risk factors for brain infections causing focal-onset seizures (Ndimubanzi *et al.*, 2010).

In the modern age, a genotype first approach may be preferred. High throughput sequencing now has consistently high genomic coverage and standardized quality control pipelines are based on data-derived thresholds and large reference databases of genomic variants. Epi25 has generated genetic data for all participants according to a single highly standardized pipeline (Epi25 Collaborative, 2019; Epi25 Collaborative, 2023). Consequently, a genotype first approach allows classification of individuals by more standardized genomic factors that can undergo quality control more readily than phenotypic data. Phenotypic data can then be interrogated for evidence of genotype-phenotype associations.

Helbig *et al.* (2019) used Ward's D2 hierarchical clustering algorithm to demonstrate how individuals carrying de novo variants in the same gene were among the most phenotypically similar in their DEE cohort. Beyond using cluster analysis to search for genetic causes, one can also generate clusters and explore the phenotypes that characterize these clusters. Westbury *et al.* (2015) used the partition around medoids algorithm on distances converted from similarity (Kaufman and Rousseeuw, 2005) to distinguish phenotypic clusters. They analyzed the statistical enrichment of HPO annotations within each cluster using FET.

4.14 Interpretation of pairwise similarities within groups of people

After assigning pairwise similarities to all pairs in a cohort and identifying those carrying a genetic variant that indicates membership of a group of interest, one has to decide on a statistical

test that can identify whether membership of the genetically defined group is associated with particularly homogeneous constellations of clinical features.[25]

One could compare the mean of pairwise similarities among group members with that of the remainder of the cohort using the unpaired Student's t-test. However, the distribution of pairwise similarities is skewed and discrete for at least many similarity measures, invalidating parametric tests. The non-parametric alternative, Wilcoxon's rank sum test, could be used to test whether the ranks of pairwise similarities of group members is greater than that of non-members. However, Wilcoxon's rank sum test requires all observations to be independent of each other. This is not the case when comparing exhaustive pairwise comparisons of a phenotypes in cohort. The dependency structure can exaggerate the differences in location between carriers and non-carriers. For example, if $\mathrm{Sim}(A, B)$ and $\mathrm{Sim}(A, C)$ are large, then $\mathrm{Sim}(B, C)$ is more likely to be large than small.

An alternative is to select a point estimate of group similarity for a group of size n. Both the absolute mean and absolute median of ${}^{n}\mathrm{C}_2$ pairwise similarities have been used. Akawi *et al.* (2015) used the sum and Westbury *et al.* (2015) used the mean. These are equivalent as the latter is the former divided by ${}^{n}\mathrm{C}_2$. Helbig *et al.* (2019) selected the median because similarities are rarely distributed normally. Furthermore, absolute similarity values may not be linearly related to the abstract concept of clinical similarity and may not possess the metric property of triangle inequality. This suggests that the median, or at least a rank-based statistic, should be preferred for its robustness to the absolute similarities of outliers.

Groups containing the same number of individuals, n can be compared directly by their median similarity. However, it is difficult to interpret the significance of different median similarities across groups of different sizes because the distribution of median similarities varies with group size. One expects to observe extreme median similarities more frequently when n is small than when n is large. Consider how if $n = 3$, the distribution of the median pairwise similarities will be broader than if $n = 100$. While the median of ${}^{3}\mathrm{C}_2 = 3$ pairwise similarities is bounded by the second largest and second smallest pairwise similarity values in the cohort, the median of ${}^{100}\mathrm{C}_2 = 4{,}950$ pairwise similarities is bounded by the 2,475$^{\text{th}}$ largest and 2,475$^{\text{th}}$ smallest.

[25] In addition to evaluating the measure of location of a genetically defined group, one can examine the distribution of pairwise similarities within the group. This may suggest clustering of people in phenotypically similar subgroups that could be explored for evidence of stratification by more granular variant classification.

Consequently, one must generate distinct null distributions of the mean pairwise similarity for each value of n using the observed data by Monte Carlo simulation. Median pairwise similarities observed for each genetically group can then be ranked in this distribution to calculate an empirical one-sided p-value.

The distribution of pairwise similarity is discrete, with many ties in a cohort as large as Epi25 in which many participants will have been annotated with identical categorical data (**Figure 4-4**, **Supplementary Figure 4-1**). Clearly, as a range of integers, the distribution of $\mathrm{Sim}_{\mathrm{TOc}}$ is very discrete (not shown). The distribution is made less discrete by adding a denominator as occurs with an unweight Jaccard index ($\mathrm{Sim}_{\mathrm{UI}}$), particularly if weighting is added ($\mathrm{Sim}_{\mathrm{GIC}}$). Comparison across a variety of measures calculated for Epi25 confirms that those that are unweighted ($\mathrm{Sim}_{\mathrm{UI}}$ and $\mathrm{Sim}_{\mathrm{Cos}}$) or probabilistic ($\mathrm{Sim}_{\mathrm{TA}}$), or that summarize semantic similarity by a single term's IC ($\mathrm{Sim}_{\mathrm{Resnik\text{-}Max}}$), tend to be more discrete than those applying term weights, particularly where the complexity is added by summation ($\mathrm{Sim}_{\mathrm{Goodall3}}$, $\mathrm{Sim}_{\mathrm{CosIC}}$, $\mathrm{Sim}_{\mathrm{TO}}$, $\mathrm{Sim}_{\mathrm{MinTO}}$, $\mathrm{Sim}_{\mathrm{Phrank}}$, $\mathrm{Sim}_{\mathrm{Resnik\text{-}BMS}}$, $\mathrm{Sim}_{\mathrm{Rel\text{-}BMS}}$) or averaging ($\mathrm{Sim}_{\mathrm{Resnik\text{-}BMA}}$, $\mathrm{Sim}_{\mathrm{Resnik\text{-}Avg}}$, $\mathrm{Sim}_{\mathrm{Rel\text{-}BMA}}$, $\mathrm{Sim}_{\mathrm{Rel\text{-}Avg}}$) of these.

Measures weighted by IC consistently demonstrate large number of pairs with modest or no (informative) similarity and a positive skew. For measures based on summed IC $\mathrm{Sim}_{\mathrm{TO}}$, $\mathrm{Sim}_{\mathrm{MinTO}}$, $\mathrm{Sim}_{\mathrm{Resnik\text{-}BMS}}$, and $\mathrm{Sim}_{\mathrm{Rel\text{-}BMS}}$, the number of pairs falls exponentially with increasing similarity, suggestive of Zipf's law. This is also seen with $\mathrm{Sim}_{\mathrm{Phrank}}$, which can be considered a conditional information-weighted measure, and which correlates closely with $\mathrm{Sim}_{\mathrm{MinTO}}$.

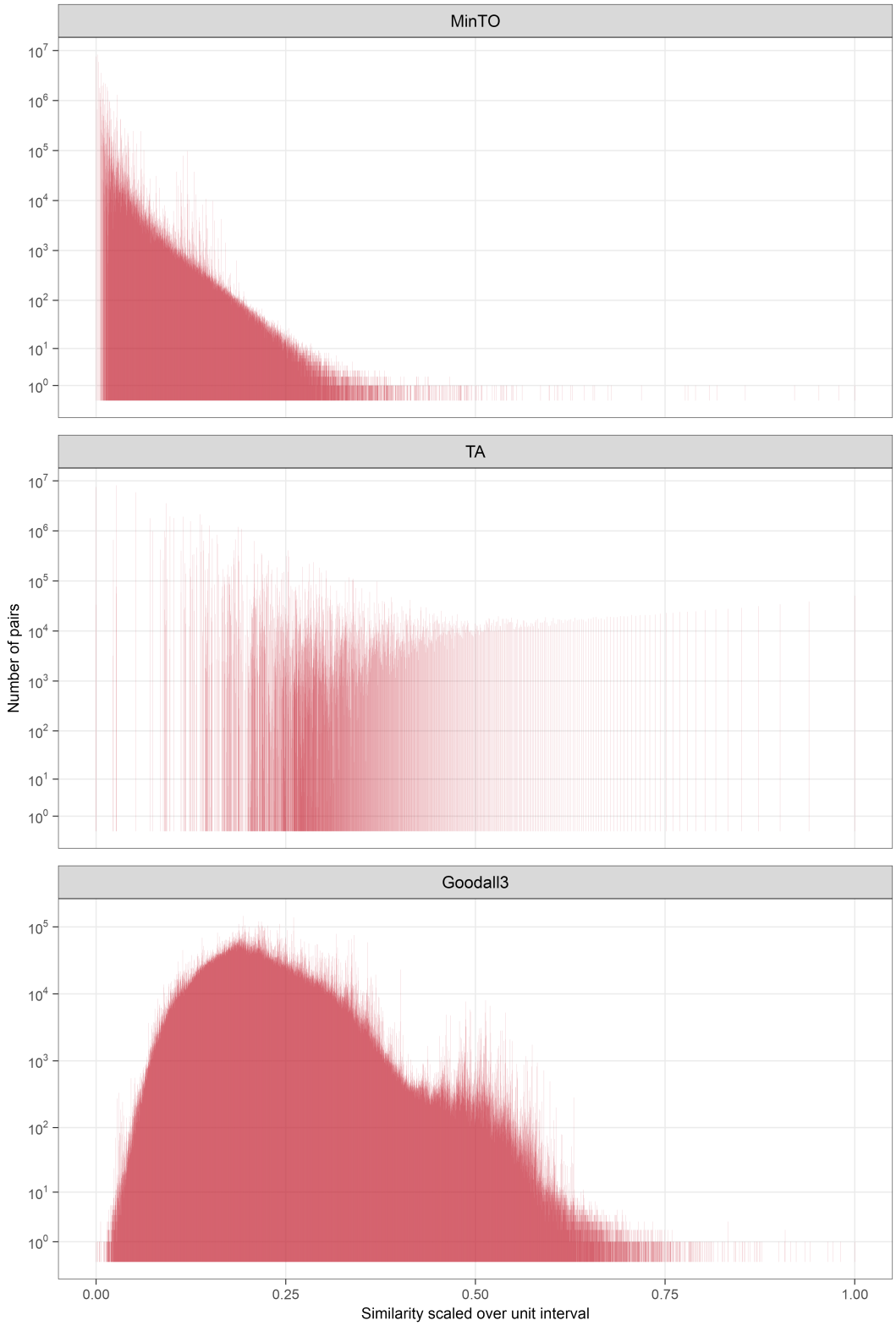


Figure 4-4. The distribution of pairwise similarities varies by similarity measure. The similarity of each of the 79,512,355 pairs that can be formed from 12,611 Epi25 participants are shown according to $\mathrm{Sim}_{\mathrm{MinTO}}$, $\mathrm{Sim}_{\mathrm{TA}}$, and $\mathrm{Sim}_{\mathrm{Goodall3}}$ as examples (distributions of further measures are included as Supplementary Figure 4-1). Similarity values have been unit-scaled and binned with width 10^{-4} to aid visualization. The probability of observing a particular $\mathrm{Sim}_{\mathrm{MinTO}}$ falls exponentially as similarity increases with sparse scaled similarities > 0.50. The distribution of $\mathrm{Sim}_{\mathrm{TA}}$ appears more uniform across a range of discrete values with low and high values more widely separated than those scaled to 0.20–0.35. The distribution of $\mathrm{Sim}_{\mathrm{Goodall3}}$ has a more continuous, central tendency with right skew and a second peak or plateau between 0.40–0.60.

Some measures do not show as marked a floor effect as these. The distribution of unweighted set-based measures Sim_{UI} and Sim_{Cosine} might be explained by phenotypic quality control that required all participants to meet electroclinical diagnostic criteria for epilepsy (Fisher *et al.*, 2014), resulting in at least four propagated annotations *All [HP:0000001]*, *Phenotypic abnormality [HP:0000118]*, *Abnormality of the nervous system [HP:0000707]*, and *Abnormal nervous system physiology [HP:0012638]*. All will have had additional annotations (either *Seizure [HP:0001250]*, annotated to 99.9 % and/or three terms describing *EEG abnormality [HP:0002353]*, annotated identically in 81.5 %). As no weighting was applied, even these uninformative and near ubiquitous annotations contributed to these pairwise similarities. $Sim_{Goodall3}$ is unique in this study, applying a close-world assumption that results in pairs of people with very few positive annotations still obtaining a range of similarities based on the shared absence of annotations. Sim_{TA} was peculiar as the only probabilistic measure. Many pairs of people with common combinations of only a few, more generic, annotations receive relatively small Sim_{TA} (**Table 4-6**). Beyond the scaled Sim_{TA} of 0.3, the number with each Sim_{TA} value stabilizes as combinations include increasingly sparse terms (with exponentially increasing intervals between Sim_{TA} due to the logarithmic transformation to convert proportions of the total set of ${}^{12,611}C_2$ combinations to IC).

HPO terms in minimal set of intersection		Pairs with terms as minimal set of intersection		Pairs with intersection containing terms		Sim_{TA} (bits)	
CUI	**Name**	**Number**	**Proportion**	**Number**	**Proportion**	**Raw**	**Scaled**
HP:0012638	Abnormal nervous system physiology	33541	0.000422	79,512,355	1.00	0.00	0.00
HP:0001250	Seizure	7,547,436	0.0949	79,323,310	0.998	0.00343	0.000139
HP:0002353	EEG abnormality	3,214	0.000404	52,813,503	0.664	0.590	0.0239
HP:0001250 HP:0002353	Seizure EEG abnormality	662,263	0.00833	52,659,453	0.662	0.594	0.0241
HP:0025373	Interictal EEG abnormality	76,841	0.000966	49,436,596	0.622	0.686	0.0278
HP:0001250 HP:0025373	Seizure Interictal EEG abnormality	56,062	0.000705	49,287,556	0.620	0.690	0.028
HP:0011182	Interictal epileptiform activity	22,333	0.000281	49,059,465	0.617	0.697	0.0283
HP:0001250 HP:0011182	Seizure Interictal epileptiform activity	8,123,629	0.102	49,009,950	0.616	0.698	0.0283
HP:0007359	Focal-onset seizure	5,903,393	0.0742	30,681,861	0.386	1.37	0.0557
HP:0002069	Bilateral tonic-clonic seizure	1,788,642	0.0225	21,889,036	0.275	1.86	0.0755

Table 4-6. The ten most common intersections of a pair's propagated annotation sets. The intersections are summarized by their minimal sets.

The distributions of group median similarity for each measure reflect those of single pairwise similarities (**Figure 4-5**, **Supplementary Figure 4-2**).

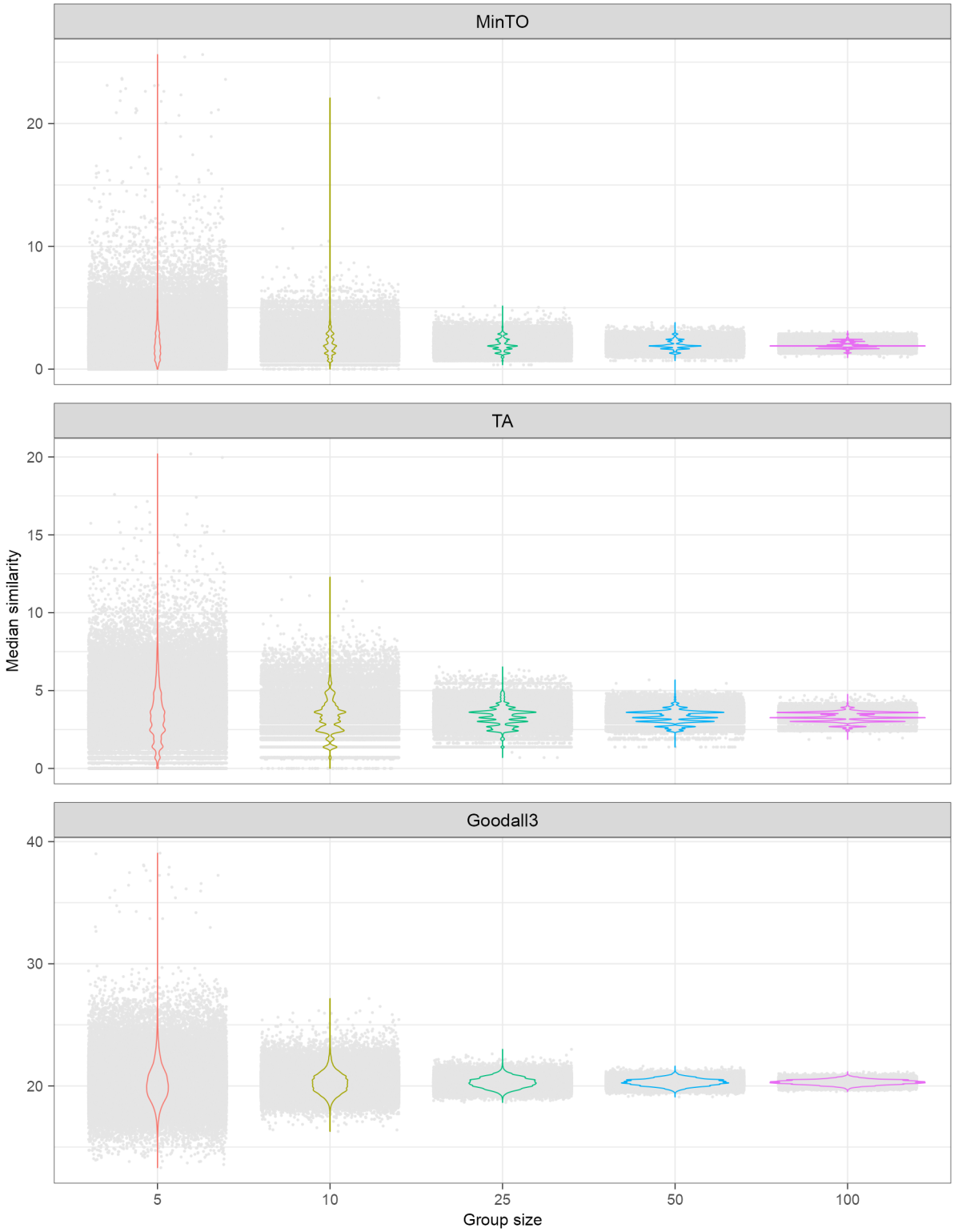


Figure 4-5. The effect of group size on the distribution of the median pairwise similarity of groups of participants. 10^5 groups of 5, 10, 25, 50, and 100 people selected at random from 12,611 Epi25 participants are plotted as grey points with density functions overlaid. $\mathrm{Sim}_{\mathrm{MinTO}}$, $\mathrm{Sim}_{\mathrm{TA}}$, and $\mathrm{Sim}_{\mathrm{Goodall3}}$ are shown as examples to show how the distribution of median similarity also varies by similarity measure (distributions of further measures are included as Supplementary Figure 4-2). The distribution of median similarity is broader for small group sizes (n = 5) than large group sizes (n = 100), reflecting the law of large numbers. The shape of the distribution is more discrete for $\mathrm{Sim}_{\mathrm{MinTO}}$ and $\mathrm{Sim}_{\mathrm{TA}}$ than for $\mathrm{Sim}_{\mathrm{Goodall3}}$ reflecting the underlying distribution of pairwise similarities (Figure 4-4).

As explained above, the range of median similarities becomes more restricted, and the distribution becomes more discrete as group size increases because of the robustness of the median. This is most marked for $\mathrm{Sim}_{\mathrm{RBP}}$. The more discrete distributions are seen among the simpler measures, those based on the proportion of the cohort sharing a single annotation ($\mathrm{Sim}_{\mathrm{Resnik\text{-}Max}}$) or set of annotations ($\mathrm{Sim}_{\mathrm{TA}}$), those based on summation, and those that are unweighted. Smoother distributions are seen for the more arithmetically complex measures, such as those with denominators and $\mathrm{Sim}_{\mathrm{Goodall3}}$.

The consequences of this are that separate Monte Carlo simulations are required for each group size and the more discrete measures may lack discriminatory power. Ties are treated conservatively given that the definition of a one-sided *p*-value in this context is the probability of observing a median pairwise similarity *at least* as great as that observed in the group of interest. This loss of power is greatest for large groups because the distribution of the median of ${}^{k}\mathrm{C}_2$ similarity values becomes increasingly discrete as group size increases.

Analysis of the absolute median of pairs of carriers is a one-sample test using the entire cohort (carriers and non-carriers) as a consistent frame of reference across tests. It asks how surprised we should be at the similarity of carriers in the context of the entire cohort. As an alternative, one could consider using the difference in medians between carriers and non-carriers as the test statistic. This treats carriers and non-carriers as two independent groups (as in a two-sample Student's *t*-test or Wilcoxon rank sum test). Accordingly, it asks a subtly different question: how surprised should we be by the difference in similarity between carriers and non-carriers? In this paradigm, the frame of reference changes for each genetic group and may be harder to translate into a clinically useful interpretation.

In the case of some of the more discrete similarity measures there may be no difference in the results of these two designs. If many pairs are tied around the cohort median similarity, then a genetically defined group would need to be large and clinically similar to affect the median of the remainder. For example, in the study of 12,611 participants in Epi25 there are ${}^{12{,}611}\mathrm{C}_2 = 79{,}512{,}355$ pairs. Of 27 similarity measures calculated, 11 had over 10,000 pairs tied on the median similarity (**Table 4-7**). $\mathrm{Sim}_{\mathrm{Resnik\text{-}Max}}$ had the greatest number of pairs tied on the median similarity while ranking below the median, 1,756,889. Even were the top ${}^{n}\mathrm{C}_2$ pairs to result from a group of n people, which is unlikely, n could be as large as 281 without causing the median $\mathrm{Sim}_{\mathrm{Resnik\text{-}Max}}$ of the remaining $12{,}611 - n$ to shift from the median $\mathrm{Sim}_{\mathrm{Resnik\text{-}Max}}$ of the entire cohort. The measure with the smallest number of pairs tied with, but ranked below, the median to demonstrate this robustness was $\mathrm{Sim}_{\mathrm{TA}}$ with 77,467 pairs tied with but ranked

below the cohort median Sim_{TA}. This tolerates a group size of at least 12 before the median of the remainder would fall below that of the complete cohort.

In conclusion, the distribution of absolute median pairwise similarity, adjusted for group size by one-sided *p*-value is preferable to that of the mean, or assessment of pseudo-median difference in ranks due to its robustness and absence of assumptions.

Similarity measure	Number of pairs with		Maximum group size to never cause the median similarity of the remaining cohort to differ from that of the complete cohort [a]
	The cohort median similarity	The cohort median similarity and ranked below the median	
Resnik-Max	3,107,326	1,756,889	281
MinTO	1,810,827	1,704,069	273
MinTO with raw IC	1,810,827	1,180,627	188
JC-Max	1,291,415	1,097,837	175
Phrank	834,908	713,409	113
TO	1,057,418	597,280	95
Lin-Max	4,773,496	579,546	92
RBP	464,488	238,056	37
SimUI	318,256	182,772	29
Resnik-BMS	188,549	88,389	14
TA	121,285	77,467	12
Rel-BMS	7,668	5,343	0
Cosine	9,3821	2,137	0
Rel-BMA	1,397	1,394	0
Rel-Max	1,181	794	0
CosIC	1,950	678	0
JC-BMA	5,310	589	0
SimGIC	336	330	0
Lin-BMA	2,103	193	0
JC-BMS	97	62	0
Lin-BMS	41	35	0
Resnik-BMA	18	16	0

Table 4-7. The sensitivity of median similarity of non-carriers to carrier group size by similarity measure. Only similarity measures giving multiple pairs the median similarity are shown. [a] Under the extreme conservative assumption that these groups contribute the most similar pairs.

4.15 Phenotypic similarity measures to be prioritized for an epilepsy study

Intuitively, Sim_{MinTO}, Sim_{Phrank}, and Sim_{TA} seem most appropriate among the set-based measures for a cohort with epilepsy in which an open-world assumption is prioritized. Sim_{MinTO} is designed to mimic clinical communication. The consideration by Sim_{Phrank} of only the additional information carried by each additional annotation given its parents avoids the inflation of IC due to nodal density when applying Sim_{TO}, and to a lesser extent when applying Sim_{MinTO}. Sim_{TA} benefits from a probabilistic strategy that is robust to potentially complex

latent phenotype-phenotype associations due to either the true biological nature of epilepsy or biases of data capture.

Set-based measures $\mathrm{Sim}_{\mathrm{CosIC}}$, $\mathrm{Sim}_{\mathrm{GIC}}$, and particularly $\mathrm{Sim}_{\mathrm{Goodall3}}$ can test the validity of the closed-world assumption in large incomplete epilepsy datasets. The null distributions of their medians are less discrete, potentially providing greater power to discriminate between groups, but like the term overlap measures, by considering all annotated terms they retain some sensitivity to nodal density.

Among the semantic similarity approaches, $\mathrm{Sim}_{\mathrm{Resnik\text{-}BMS}}$ seems most appropriate, given the need for an open-world assumption. However, $\mathrm{Sim}_{\mathrm{Resnik\text{-}BMA}}$ has become established, largely through the popularity of Phenomizer, despite its sensitivity to the addition or omission of common phenotypes. $\mathrm{Sim}_{\mathrm{BMA}}$ measures have a less discrete distribution than $\mathrm{Sim}_{\mathrm{BMS}}$, approaching that of the set-based measures, potentially providing a compromise. Semantic similarity measures based on $\mathrm{Sim}_{\mathrm{Rel}}$, $\mathrm{Sim}_{\mathrm{IC}}$, or $\mathrm{Sim}_{\mathrm{adjResnik}}$ may be of interest to see whether consideration of differences in annotation depth is of value. However, $\mathrm{Sim}_{\mathrm{Lin}}$ and $\mathrm{Sim}_{\mathrm{JC}}$ seem inappropriate to the task of measuring the similarity, meaning shared information, of two people. Rather they represent the potential phenotypic distance between them.

The correlations between a variety of measures reflect these similarities between measures, demonstrating that $\mathrm{Sim}_{\mathrm{MinTO}}$, $\mathrm{Sim}_{\mathrm{Phrank}}$, $\mathrm{Sim}_{\mathrm{TA}}$, $\mathrm{Sim}_{\mathrm{RBP}}$ (with α set to limit the weighting of terms to those annotated to 0.1 % of the cohort), $\mathrm{Sim}_{\mathrm{TO}}$, $\mathrm{Sim}_{\mathrm{Resnik\text{-}BMS}}$, and $\mathrm{Sim}_{\mathrm{Rel\text{-}BMS}}$ are very closely correlated in 12,611 participants in Epi25 (Spearman's $\rho \geq 0.91$, **Figure 4-6** and **Figure 4-7**) and a cohort of 67 adults with CDKL5 Deficiency Disorder (**Figure 4-8** and **Figure 4-9**). $\mathrm{Sim}_{\mathrm{Resnik\text{-}Max}}$, $\mathrm{Sim}_{\mathrm{Rel\text{-}Max}}$, $\mathrm{Sim}_{\mathrm{Resnik\text{-}BMA}}$, and $\mathrm{Sim}_{\mathrm{Rel\text{-}BMA}}$ are moderately well correlated with these, as well as with the independent cluster of weighted measures that take account of a large number of annotations with weighting: $\mathrm{Sim}_{\mathrm{Avg}}$, $\mathrm{Sim}_{\mathrm{CosIC}}$, and $\mathrm{Sim}_{\mathrm{GIC}}$. $\mathrm{Sim}_{\mathrm{Goodall3}}$ is independent of other measures and may offer a complementary perspective based on the closed-world assumption.

In the following chapters, I use HPO annotations to describe the phenotypic landscape of cohorts and delineate the phenotypic repertoires of carriers of specific genetic variants, and to perform genotype-first tests of phenome-wide association and overall phenotypic similarity analysis using $\mathrm{Sim}_{\mathrm{MinTO}}$ as the primary similarity measure with others to offer complementary perspectives. I provide my own function for calculating semantic similarity, and various set-based similarity measures (**Algorithm 4-1**).

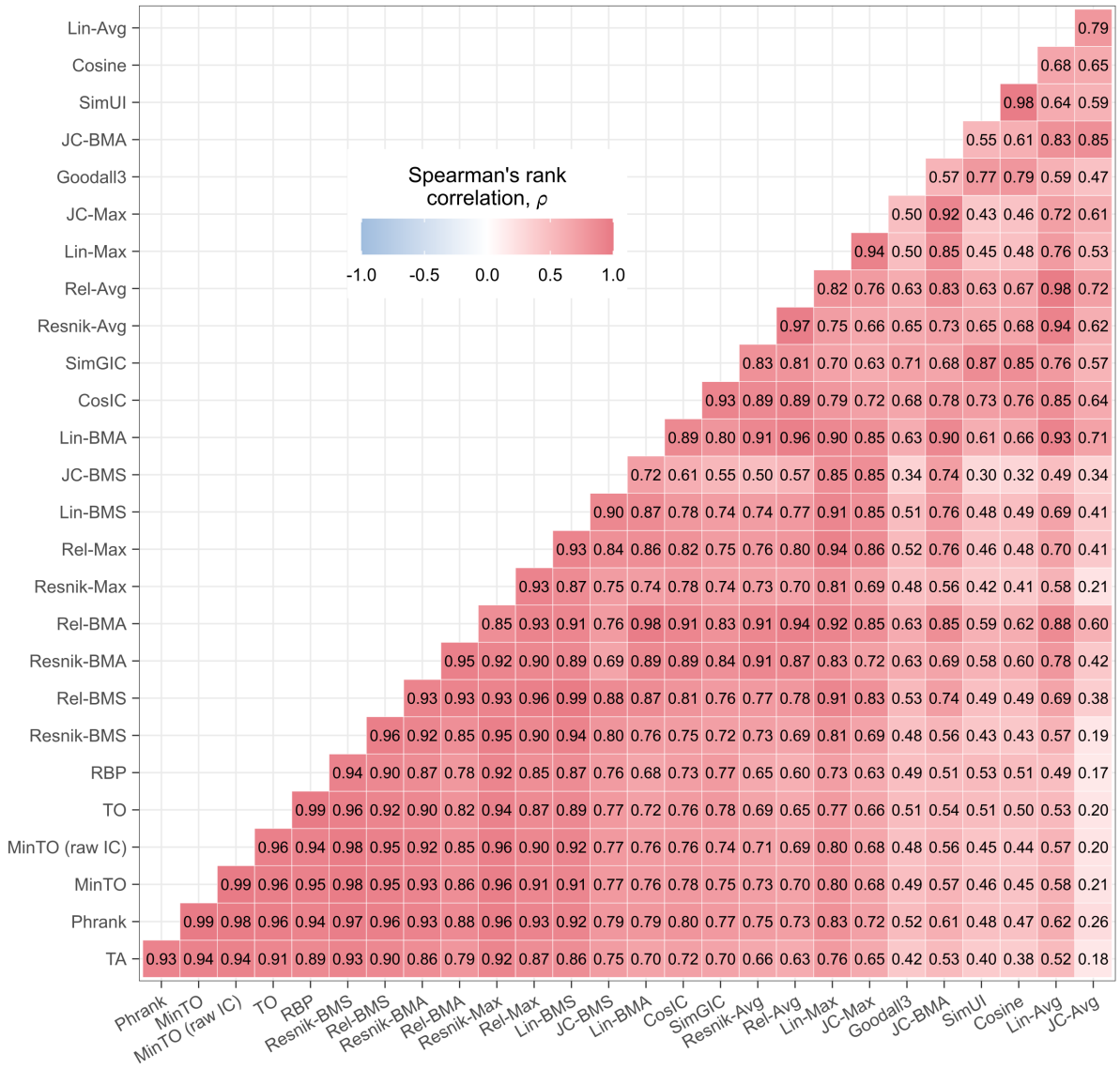


Figure 4-6. Spearman's correlation between 27 similarity measures in Epi25. The similarity of all 79,512,355 pairs of individuals formed from 12,611 participants in Epi25 (Chapter 5) was calculated according to 27 similarity measures. Spearman's rank correlation ρ is shown for each pairwise combination of similarity measures across these 79,512,355 pairs. ρ is positive across all comparisons, indicating that the prioritized measures are not grossly divergent in this data set. $\text{Sim}_{\text{MinTO}}$ (using monotonic IC) ranks pairs most similarly to $\text{Sim}_{\text{Phrank}}$ ($\rho = 0.99$), to $\text{Sim}_{\text{MinTO}}$ using raw IC ($\rho = 0.99$), and to $\text{Sim}_{\text{Resnik-BMS}}$ ($\rho = 0.98$). $\text{Sim}_{\text{Phrank}}$ ranks pairs most similarly to $\text{Sim}_{\text{MinTO}}$ both with monotonic ($\rho = 0.99$) and raw IC ($\rho = 0.98$). Sim_{TO} and Sim_{RBP} ($\rho = 0.99$), $\text{Sim}_{\text{Rel-Avg}}$ and $\text{Sim}_{\text{Lin-Avg}}$ ($\rho = 0.98$), and Sim_{UI} and $\text{Sim}_{\text{Cosine}}$ ($\rho = 0.98$) yield very similar ranks. The probabilistic measure Sim_{TA} is most strongly correlated to the semantic similarity measures $\text{Sim}_{\text{Resnik}}$, Sim_{Lin}, and Sim_{Rel} summarized by best match sum or maximum, and to $\text{Sim}_{\text{MinTO}}$, $\text{Sim}_{\text{Phrank}}$, Sim_{RBP}, and Sim_{TO} (all $\rho \geq 0.86$). While other similarity measures correlate with at least one other measure with $\rho \geq 0.85$, $\text{Sim}_{\text{Goodall3}}$ appears distinct (all $\rho < 0.80$) reflecting its conceptual difference, incorporating the concordant exclusion of a phenotype.

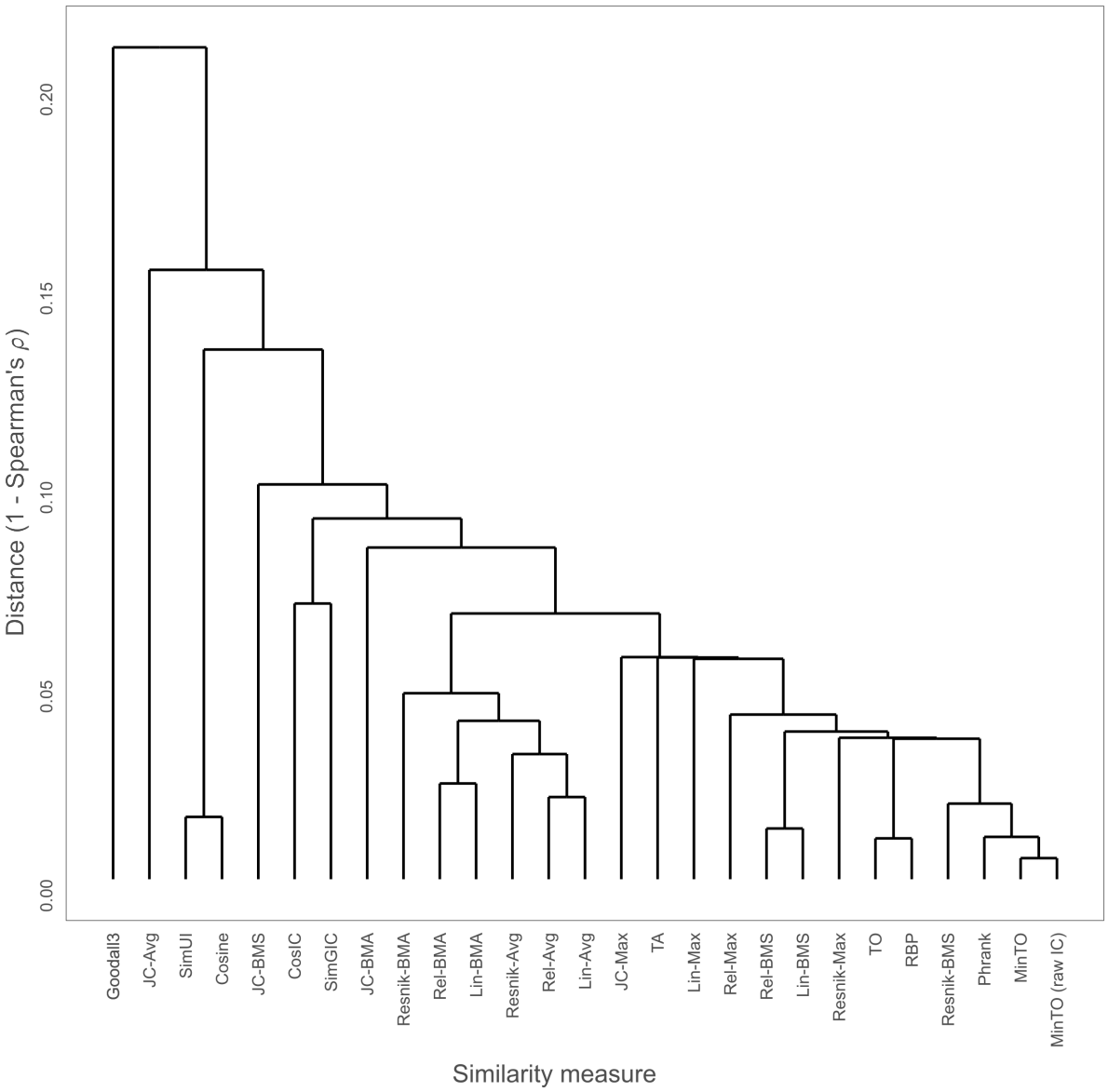


Figure 4-7. Hierarchical cluster analysis of 27 similarity measures in Epi25. Distances between measures were calculated as the complement of the Spearman's ρ from Figure 4-6. Clustering was performed by single linkage (*ggdendro::ggdendrogram(hclust(as.dist(1-sm), method = 'single'))*, version 0.2.0, where *sm* is a matrix representation of Figure 4-6). This results in the repeated agglomeration of the two clusters of similarity measure(s) containing the most closely correlated similarity measures that are not yet in the same cluster. In addition to the correlations emphasized in the caption to Figure 4-6, one can see the high correlation between semantic similarity measures based on $\mathrm{Sim}_{\mathrm{Lin}}$ and $\mathrm{Sim}_{\mathrm{Rel}}$, and according to the summarization (average, best match average) of matrix $\boldsymbol{S}$ of term-term similarities for a pair of people. $\mathrm{Sim}_{\mathrm{UI}}$ and $\mathrm{Sim}_{\mathrm{Cosine}}$, the measures based on the unweighted counts of concordant annotations, are correlated. Across all four summarization methods (average, best match average, best match sum, and maximum), $\mathrm{Sim}_{\mathrm{JC}}$-based measures tend to be most distant from the other semantic similarity measures ($\mathrm{Sim}_{\mathrm{Resnik}}$, $\mathrm{Sim}_{\mathrm{Lin}}$, $\mathrm{Sim}_{\mathrm{Rel}}$). While Figure 4-6 demonstrates that $\mathrm{Sim}_{\mathrm{TA}}$ is correlated to a number of semantic similarity measures summarized by best match sum or maximum, and to $\mathrm{Sim}_{\mathrm{MinTO}}$, $\mathrm{Sim}_{\mathrm{Phrank}}$, $\mathrm{Sim}_{\mathrm{RBP}}$, and $\mathrm{Sim}_{\mathrm{TO}}$ (here shown to its right), its probabilistic perspective appears to capture a concept distinct from each of these as it is not highly correlated with any of these. As in Figure 4-6, $\mathrm{Sim}_{\mathrm{Goodall3}}$ appears very distinct from other measures.

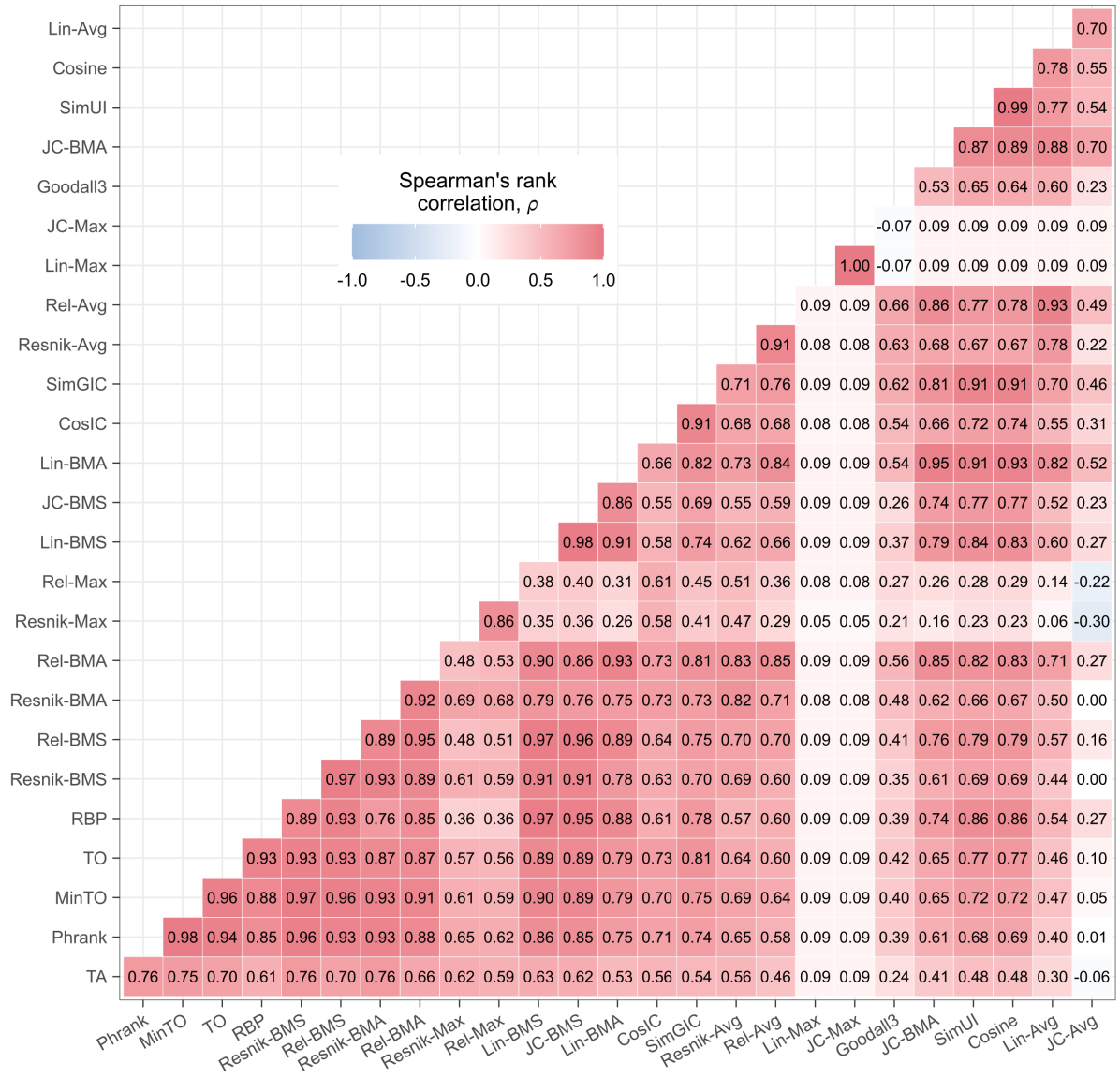


Figure 4-8. Spearman's correlation between 26 similarity measures in a study of CDKL5 Deficiency Disorder. The similarity of all 2,211 pairs of individuals formed from 67 participants in the CDKL5 Deficiency Disorder natural history study (Chapter 7) was calculated according to 26 similarity measures. Spearman's rank correlation ρ is shown for each pairwise combination of similarity measures across these 2,211 pairs. ρ is positive across all comparisons, except for some including $\text{Sim}_{\text{JC-Avg}}$, $\text{Sim}_{\text{JC-Max}}$, $\text{Sim}_{\text{Lin-Max}}$, or $\text{Sim}_{\text{Goodall3}}$. $\text{Sim}_{\text{JC-Max}}$ and $\text{Sim}_{\text{Lin-Max}}$ were highly correlated to each other ($\rho = 1.00$) but otherwise these four measures were distinct from each other and other measures ($\rho \leq 0.70$). As in Epi25, Sim_{UI} and $\text{Sim}_{\text{Cosine}}$ ($\rho = 0.99$), and $\text{Sim}_{\text{MinTO}}$ (using monotonic IC) and $\text{Sim}_{\text{Phrank}}$ ($\rho = 0.98$) were highly correlated. $\text{Sim}_{\text{JC-BMS}}$ and $\text{Sim}_{\text{Lim-BMS}}$ were also highly correlated ($\rho = 0.98$).

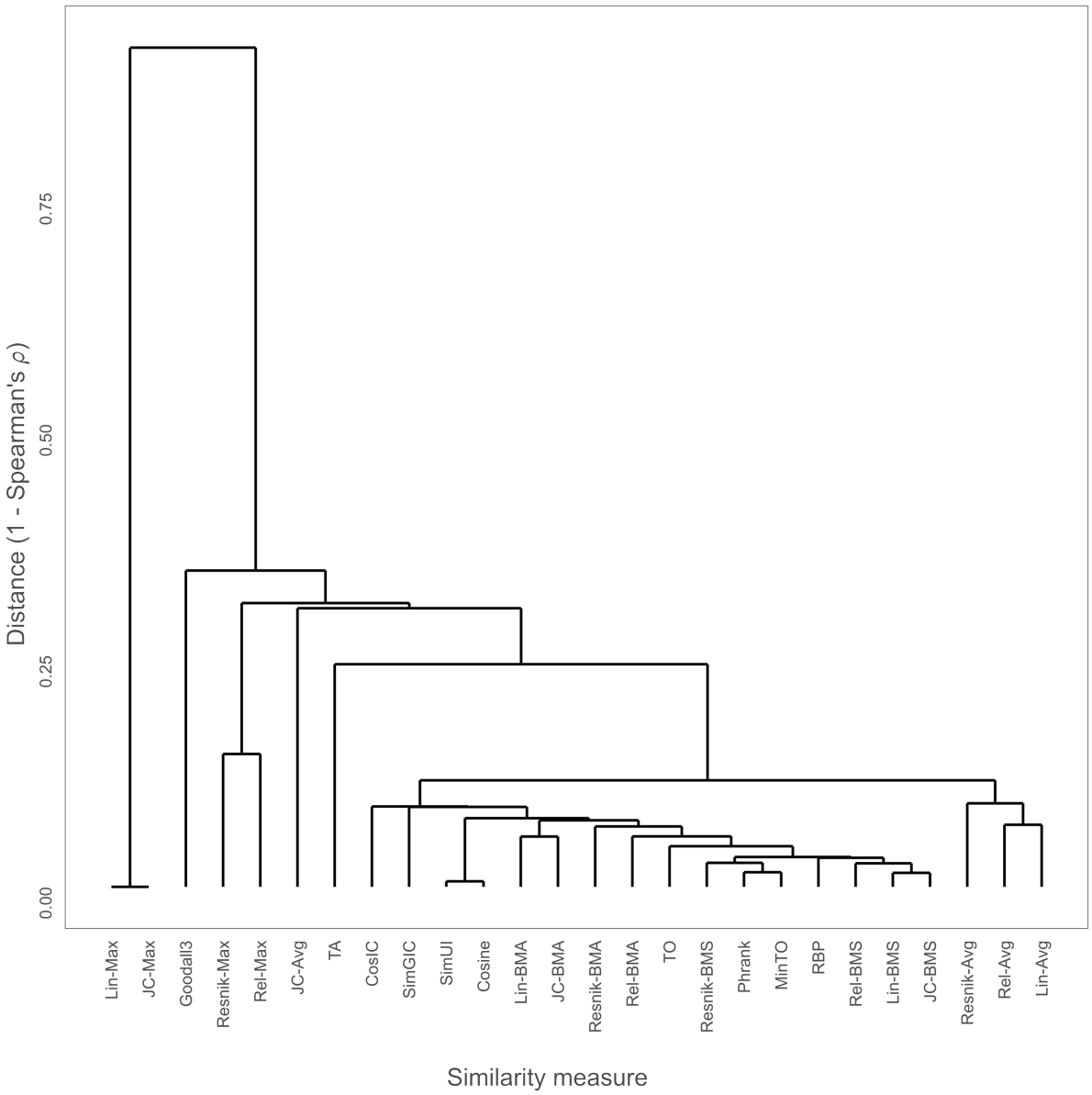


Figure 4-9. Hierarchical cluster analysis of 26 similarity measures in cohort with CDKL5 Deficiency Disotder. Distances between measures were calculated as the complement of the Spearman's correlation of all 2,211 pairwise combinations of 67 participants in the natural history study. As in Figure 4-7, clustering was performed by single linkage (*ggdendro::ggdendrogram(hclust(as.dist(1-sm), method = 'single'))*, version 0.2.0, where *sm* is a matrix representation of Figure 4-8). Similar clusters of similarity measures emerge to those in Figure 4-7, particularly at short distances.

```
# A base R function to calculate similarity according to multiple measures
#
# IDENTIFERS
# id_1, identifier for first person in pair being compared
# id_2, identifier for second person in pair being compared
#
# PHENOTYPIC DATA
# prop, named list of propagated annotations as HPO CUIs, names match identifiers
# id_2 and id_2
# base, named list of base annotations as HPO CUIs, names match identifiers id_2 and
# id_2
#
# HPO
# A named list of each HPO term's ancestors as CUIs, names are the corresponding
# terms' CUIs
#
# NAMED VECTORS OF WEIGHTS OF EACH TERM, NAMES ARE HPO CUIS
# ic, a named double vector of the information content of each term
# ic.mono, a monotonically adjusted form of double vector ic
# rbp.wt, a named double vector of the RBP weight of each term
# phrank.wt, a named double vector of the Phrank weight of each term
# ppt, a named double vector of the proportion of cohort annotated with each term
#
# A CHARACTER VECTOR OF THE SIMILARITY MEASURES TO BE CALCULATED
# sim.measures <- c(
#   "Resnik.Avg", "Resnik.BMA", "Resnik.BMS", "Resnik.Max",
#   "Rel.Avg", "Rel.BMA", "Rel.BMS", "Rel.Max",
#   "Lin.Avg", "Lin.BMA", "Lin.BMS", "Lin.Max",
#   "JC.Avg", "JC.BMA", "JC.BMS", "JC.Max",
#   "MinTO", "TO", "Phrank", "RBP",
#   "Cosine", "CosIC", "SimUI", "SimGIC",
#   "Goodall"
# )

sim_f <- function(id_1, id_2) {

  # Construct an array of 1 x 1 x the number of similarity measures to be populated
  res <- rep(NA, length(sim.measures))
  res <- array(dim = c(1,1,length(sim.measures)))
  dimnames(res)[[1]] <- id_1
  dimnames(res)[[2]] <- id_2
  dimnames(res)[[3]] <- sim.measures

  # Obtain the propagated annotations of each person from a named list
  prop_1 <- prop[[id_1]]
  prop_2 <- prop[[id_2]]

  # Obtain the intersection of their propagated annotations
  i.terms <- intersect(prop[[id_1]], prop[[id_2]])

  # Obtain the union of their propagated annotations
  if (any(c("SimUI", "SimGIC", "Goodall") %in% sim.measures)) {
    u.terms <- union(prop_1, prop_2)
  }

  # Calculate SimMinTO using monotonically adjusted IC vector
  if ("MinTO" %in% sim.measures) {
    redundant <- unlist(use.names = FALSE,
                        lapply(i.terms,
                               function(x)
                                 setdiff(hpo[[x]], x)))
    res[1,1,"MinTO"] <- sum(ic.mono[setdiff(i.terms, redundant)])
  }

  # Calculate SimTO using raw IC vector
  if ("TO" %in% sim.measures) {
    res[1,1,"TO"] <- sum(ic[i.terms])
  }
```

```
# Calculate SimUI from the intersection and union
if ("SimUI" %in% sim.measures) {
  res[1,1,"SimUI"] <- length(i.terms) / length(u.terms)
}

# Calculate SimGIC by weighting these by the raw IC
if ("SimGIC" %in% sim.measures) {
  res[1,1,"SimGIC"] <-
    sum(ic[i.terms]) / sum(ic[u.terms])
}

# Calculate SimCosine without weights
if ("Cosine" %in% sim.measures) {
  res[1,1,"Cosine"] <-
    length(i.terms) / sqrt(length(prop_1) * length(prop_2))
}

# Calculate the SimCosIC by weighting these by the raw IC
if ("CosIC" %in% sim.measures) {
  res[1,1,"CosIC"] <-
    sum(ic[i.terms] ^ 2) / (sqrt(sum(ic[prop_1] ^ 2) * sum(ic[prop_2] ^ 2)))
}

# Calculate the SimRBP by weighting these using a vector of the RBP specific
# weights, limited by alpha
if ("RBP" %in% sim.measures) {
  res[1,1,"RBP"] <-
    sum(rbp.wt[i.terms]) # Precalculated minimum
}

# Calculate the SimPhrank by weighting these using a vector of the Phrank specific
# weights, calculated recursively from the root based on the members of the cohort
# annotate with each term and annotated with any of its parents
if ("Phrank" %in% sim.measures) {
  res[1,1,"Phrank"] <-
    sum(phrank.wt[i.terms])
}

# Calculate the SimGoodall3 by based on the square of the proportion of the cohort
# with or without each phenotype in concordance with the pair's concordant
# annotations (both present or both absent)
if ("Goodall" %in% sim.measures) {
  res[1,1,"Goodall"] <-
    sum(
      (1 - (ppt[i.terms])^2),
      (1 - (1 - ppt[setdiff(names(ppt), u.terms)])^2)
    )
}

# Create a pairwise similarity matrix for each semantic similarity measure
if (any(grepl("(Resnik)|(Lin)|(JC)|(Rel)", sim.measures))) {
  # Generate list of matrices
  matrices <- rep(NA, 4)
  matrices <- lapply(matrices,
                     function(xx) {
                       matrix(nrow = length(base[[id_1]]),
                              ncol = length(base[[id_2]]))
                     })
  names(matrices) <- c("Resnik", "Lin", "JC", "Rel")

  # Calculate semantic similarity of each pair of terms, populating these matrices
  for (i in 1:length(base[[id_1]])) {
    for (j in 1:length(base[[id_2]])) {
      # Identify common ancestors in the study dataset
      ca <-
        intersect(hpo[[base[[id_1]][i]]],
                  hpo[[base[[id_2]][j]]])
      ca <- ca[ca %in% names(ic)]
      # SimResnik is the maximum IC of the common ancestors of each pair of terms
      if (sum(grepl("Resnik", sim.measures)) > 0) {
        matrices$Resnik[i, j] <- max(ic[ca])
```

```
        }

        if (any(grepl("(Lin)|(Rel)", sim.measures))) {
          # SimLin is double the maximum IC of the common ancestors / the sum of
          # the sum of IC of the terms being compared
          Lin <-
            2 * max(ic[ca]) / (ic[base[[id_1]][i]] + ic[base[[id_2]][j]])
        }
        if (any(grepl("Lin", sim.measures))) {
          matrices$Lin[i, j] <-
            Lin
        }
        # SimRelevance is the SimLin weighted by the complement of the proportion
        # of the cohort with the most informative common ancestor
        if (any(grepl("Rel", sim.measures))) {
          matrices$Rel[i, j] <-
            Lin * (1 - (1 / (2 ^ max(ic[ca]))))
        }
        # SimJC is the reciprocal of the JC distance: the sum of the information
        # content of the terms being compared minus double the information content
        # of their most informative common ancestor
        if (any(grepl("JC", sim.measures))) {
          matrices$JC[i, j] <-
            1 / ((ic[base[[id_1]][i]] + ic[base[[id_2]][j]] - 2 * max(ic[ca])) + 1)
        }
      }
    }

    # Summarize semantic similarity measures by single maximum, average,
    # best match average, or best match sum
    for (i in sim.measures) {
      if (grepl("(Avg)|(Max)", i)) {
        for (j in c(".Avg", ".Max")) {
          if (grepl(j, i)) {
            sim.measure <- gsub(j, "", i)
            if (j == ".Avg") {
              res[1, 1, i] <- mean(matrices[[sim.measure]])
            }
            if (j == ".Max") {
              res[1, 1, i] <- max(matrices[[sim.measure]])
            }
          }
        }
      }
      if (grepl("(BMA)|(BMS)", i)) {
        for (j in c(".BMA", ".BMS")) {
          if (grepl(j, i)) {
            sim.measure <- gsub(j, "", i)
            max1.2 <- apply(matrices[[sim.measure]], 1, max)
            max2.1 <- apply(matrices[[sim.measure]], 2, max)
            if (j == ".BMA") {
              res[1, 1, i] <- 0.5 * sum(mean(max1.2), mean(max2.1))
            }
            if (j == ".BMS") {
              res[1, 1, i] <- 0.5 * sum(max1.2, max2.1)
            }
          }
        }
      }
    }
  }
  # Return the results array (for combination with results from other pairs)
  res
}
```

Algorithm 4-1. A function for calculating multiple semantic and set-based similarity measures. Commencing 2 pages previously.

Chapter 5. The phenotypic consequences of ultra-rare variants

5.1 Preface

In this chapter, I describe my analysis of data that I harmonized from an international cohort of 12,611 people undergoing whole exome sequencing within the Epi25 Collaboration. The aim was to use the methods discussed in Chapter 4 to establish whether URVs have sufficient biological effect to contribute not only to a risk of epilepsy generically, but also to shape the precise clinical phenotypes of the epilepsy. This finding would have consequences for clinical practice, informing prognostication and offering a biological explanation for some elements of a person's experience of epilepsy. Furthermore, it would demonstrate the need for consideration of genetic stratification in research cohorts and the development of increasingly personalized medicine.

I created the dictionary to translate categorical Epi25 data into HPO annotations, translated all the phenotypic data into HPO terms (with manual review of every free-text entry), undertook quality control of phenotypic data and matching of DNA sample and participant clinical record identifiers, and designed and undertook the similarity and association analyses using my own scripts, including all coding of these. I created the panels of genes tested based on clinical and research resources and decided on annotation and preliminary variant filtering parameters for filtering by my collaborator Shiva Ganesan (Children's Hospital of Philadelphia). I undertook the final variant filtering and gene-collapsing classification of qualifying variants. I drafted this chapter and its figures independently.

5.2 Abstract

In up to 5 % of people with epilepsy, the disorder can be attributed to a monogenic diagnosis related to monoallelic or biallelic variants in one of over 400 epilepsy genes. URVs, those that are too rare to be found in large population databases are more common in people with unexplained epilepsies than in the general population. Yet beyond the minority of cases in which a variant meets formal diagnostic criteria for a monogenic disorder, the clinical relevance of these variants to individual people is unknown.

An international collaboration collated clinical summaries of a cohort of over 12,611 people prior to an exome sequencing study. I mapped the phenotypic details to the HPO and grouped participants according to the presence of URVs 262 genes in which monoallelic variants may cause epilepsy. I calculated phenotypic similarity and phenotypic associations of each group

and extended the analysis to a further 5,148 genes with less reason to hypothesize a detectable association with phenotype.

Clinical data generated 291,432 phenotypic annotations (a median of 20 per person), spanning 1,738 phenotypic concepts. 5,517 people (43.7 %) carried an ultra-rare protein-altering variant (PAV) in one of the 262 dominant epilepsy genes, and 12,538 (99.4 %) carried a PAV in any gene in the study. 257 of 260 (98.8 %) of dominant genes had at least one association with $p < 0.05$, and 32/260 (12.3 %) had an association robust to the 1,734 phenotypic tests undertaken per genetic group. For example, *SCN1A* ($n = 116$) with *Focal hemiclonic seizure [HP:0006813]* (OR = 13.9) and *Febrile seizure (within the age range of 3 months to 6 years) [HP:0002373]* (OR = 2.92). The clinical constellations of people with variants in the same gene tended to be more similar to each other than expected by chance with carriers of variants in *SLC6A1*, *RORA*, and *SCN1A*, and those carrying protein-truncating variants (PTVs) in *WDR45* and *NEXMIF* appearing particularly similar. Exploratory analyses identified genes with associations and high phenotypic similarity that can be hypothesized to reflect their effects on clinical phenotype.

This study demonstrates that URVs in many validated epilepsy genes are associated with characteristic clinical repertoires, even in individuals in whom these associations might not have been recognized by experienced clinicians and in whom the variants might not reach a diagnostic threshold. Furthermore, phenotypic evidence implicates some genes previously not associated with epilepsy as new candidates with phenotypic consequences. Hence, the relevance of URVs shows potential to extend from the biological to the clinical domain, beyond explaining why somebody has epilepsy, towards biomarker status for stratification to inform classification, prognostication, and better treatment decisions.

5.3 Introduction

The epilepsies are clinically broad disorders characterized by an increased predisposition to seizures (Fisher *et al.*, 2014). Clinicians recognize clinically useful broad types of epilepsy (Scheffer *et al.*, 2017) and carefully defined syndromes (Hirsch *et al.*, 2022; Riney *et al.*, 2022; Specchio *et al.*, 2022; Wirrell *et al.*, 2022b; Zuberi *et al.*, 2022) but even within these, manifestations, treatment responses, and developmental outcomes vary. This clinical heterogeneity is thought to reflect pathophysiological heterogeneity and it is proposed that increasing the biological validity of clinical classification by etiology may yield long awaited improvements in clinical care and outcomes (Fisher *et al.*, 2017b; Scheffer *et al.*, 2017).

The enrichment of URVs, those absent from large population databases, in monogenic epilepsy genes among people with common epilepsies suggests that common, genetically complex epilepsies have molecular causes that overlap with those of monogenic epilepsies (Thomas and Berkovic, 2014; Epi4K Consortium and Epilepsy Phenome Genome Project, 2017; May *et al.*, 2018; Epi25 Collaborative, 2019; Epi25 Collaborative, 2021), and may explain some of the heritability not accounted for by common variants (Gorlov *et al.*, 2011; Speed *et al.*, 2014; Zuk *et al.*, 2014; Bomba;Walter and Soranzo, 2017; ILAE Consortium on Complex Epilepsies, 2023). However, beyond the minority of people in whom there is sufficient evidence to assign a monogenic diagnosis (Richards *et al.*, 2015), the complexity of the genetic architecture and phenotypic repertoire of the epilepsies has impeded efforts to translate genetic discoveries into information that contributes to routine clinical care.

While many of the private de novo variants we each carry will have little or no clinical consequences, the absence of URVs in increasingly large and ancestrally diverse population databases suggests that these are under negative selection: the effect size of an individual URV on reproductively detrimental phenotypes is much greater than those of common variants that conversely, are tolerated in the population (Schoech *et al.*, 2019; Whiffin;Ware and O'Donnell-Luria, 2019). Like common variants, URVs are thought to make a sizable pleiotropic contribution to the risk of a broad range of neurodevelopmental and psychiatric disorders (O'Roak *et al.*, 2011; O'Roak *et al.*, 2012; de Rubeis *et al.*, 2014; Ganna *et al.*, 2018; Singh *et al.*, 2022).

From a clinical perspective, we would like to identify the genetic basis of particular clinical features and constellations with sufficient resolution to maximize their clinical usefulness for single patients. While URVs may partially explain why a person has epilepsy, it remains unclear whether they shape the particular manifestations of the person's epilepsy beyond different contributions to epilepsy type, and thus to focal-onset versus generalized-onset seizures (Epi4K Consortium and Epilepsy Phenome Genome Project, 2017; May *et al.*, 2018; Epi25 Collaborative, 2019; Epi25 Collaborative, 2021; Koko *et al.*, 2021; Epi25 Collaborative, 2023). There are likely to be independent genetic mechanisms contributing to the heritability of clinical features (Winawer *et al.*, 2005; Tobochnik *et al.*, 2017; Gramm *et al.*, 2020) and age of onset (Ellis *et al.*, 2019). Even in the *monogenic* epilepsies, there is phenotypic heterogeneity, presumed to be at least partly due to genetic complexity as the consequences of the principal variant interact with those of both rare (Takata *et al.*, 2019; de Lange *et al.*, 2020) and common (Campbell *et al.*, 2022; Martins Custodio *et al.*, 2023) polygenic factors. However, were a single URV to have a distinguishable effect on clinical phenotypes despite this, the finding of a

qualifying variant in clinic would move its relevance from explanation of what is usually already known, that the person has a generic predisposition to seizures, to more precise decision making with genetically stratified care.

Here, applying my computational phenotyping approaches from Chapter 4 to a large international cohort undergoing whole exome sequencing (Epi25 Collaborative, 2021), I show that a person's URV may influence their clinical features beyond a generic predisposition to epilepsy. Using phenome-wide association and phenotypic similarity analyses, I quantify the phenotypic evidence for each of 5,418 genes having a role in epilepsy, generating phenotypically derived hypotheses for future investigation and a phenotypic reference resource.

5.4 Materials and methods

5.4.1 Participant recruitment and data collection

In years 1–3 of the Epi25 Collaboration, we collected DNA from and phenotypic data about unrelated individuals with unexplained epilepsy from 41 epilepsy centers spanning Europe, Australasia, North America, and Asia (Epi25 Collaborative, 2021). Case-control analysis of their whole-exome data according to broad epilepsy type has been published previously (Epi25 Collaborative, 2019; Epi25 Collaborative, 2021). While prior genetic testing was not an exclusion criterion, collaborators were instructed not to recruit individuals with known molecular diagnoses (Dravet, 2011). For each participant, phenotypic data were submitted prior to return of genetic results to collaborators, blinding clinical data entry from the results of the genetic analysis. Phenotypic analyses of those carrying a qualifying variant in a given gene were tested with reference to the remainder of the cohort, without external controls.

5.4.2 Informed consent

Consent was granted by adult participants, or the legal guardians of child participants according to local ethical requirements at participating centers, as previously described (Epi25 Collaborative, 2019; Epi25 Collaborative, 2021). For individuals recruited after 2015-1-25, consent forms were required to meet the standards of the National Institutes of Health's Genomic Data Sharing policy. Data and DNA collected before this were only accepted if consent did not exclude data sharing.

5.4.3 Phenotypic data collection and quality control

Epilepsy was diagnosed by experienced epileptologists on clinical grounds consistent with ILAE criteria (Fisher *et al.*, 2014). Deidentified phenotyping data for each participant were collected from clinical records and entered into a REDCap database (Harris *et al.*, 2009), using one of three study-specific online case record forms selected depending on the collaborator's classification tailored to the participant's type of epilepsy: generalized epilepsy (including mixed), focal epilepsy, or DEE (Epi25 Collaborative, 2020). Forms were hyperlinked to an online reference resource International League Against Epilepsy definitions of seizure types for reference (International League Against Epilepsy). These data were managed in the Epi25 Data Repository (hosted at the Luxembourg Centre for Systems Biomedicine). Where participants had been part of the Epilepsy Phenome/Genome Project (Nesbitt *et al.*, 2013), deidentified data were accessed and mapped to the Epi25 REDCap database.

5.4.4 Code and statistics

Analyses and statistical tests were conducted using the R statistical language (R-project.org, version 4.3.2, R Core Team, 2017) using the *tidyverse* suite of packages (version 2.0.0), *data.table* (1.14.10), *ontologyIndex* (2.11), and *IRanges* (2.18.3) to manipulate data. *parallel* (4.3.2), *doParallel* (1.0.17), *foreach* (1.5.2), and *parallelly* (1.37.1) were used for local parallelization of code, and *rslurm* (0.6.2) to facilitate remote high-performance computing. Visualizations were generated using *ggplot2* (3.5.0), *ggrepel* (0.9.1), *ggVennDiagram* (0.3.3), *ggraph* (2.2.0), and *circlize* (0.4.15), followed by optimization of labels using Inkscape (version 1.3.1). Similarity and association analyses were run using my own functions and scripts. Further packages and functions are specified in the text.

5.4.5 Harmonization of categorical phenotypic data

This study used HPO version 1.7.14 (release date 2022 02 14, Gargano *et al.*, 2023) modified to suit this analysis of epilepsy (Table 5-1).

HPO CUI	HPO name	Rationale for exclusion or modified use
HP:0000001	All	Excluded as ubiquitous after propagation in this cohort.
HP:0000118	Phenotypic abnormality	Excluded as ubiquitous after propagation in this cohort.
HP:0000707	Abnormality of the nervous system	Excluded as ubiquitous after propagation in this cohort.
HP:0012638	Abnormal nervous system physiology	Excluded as ubiquitous after propagation in this cohort.
HP:0020174	Refractory drug response	Made a child of Phenotypic abnormality [HP:0000118] only because in the context of this analysis, we do not consider this to be a form an Abnormality of metabolism/homeostasis [HP:0001939]. Abnormal drug response [HP:0020169] was excluded from this modification because no individual was annotated with this without also being annotated with Refractory drug response [HP:0020174], making the former uninformative relative to the latter.
HP:0033349	Seizure cluster	Made a child of Seizure [HP:0001250] only as a description of a phenotypic feature of seizures.
HP:0040006	Mortality/aging	I used this term to code those individuals who were recorded as deceased.
HP:0001699	Sudden death	Made a child of Mortality/Aging [HP:0040006] only, a term I used to code those individuals who were recorded as deceased. Sudden unexpected death in epilepsy [HP:0033258] remained a child of Sudden death [HP:0001699].
HP:0025142	Constitutional symptom	Types of Constitutional symptom [HP:0025142] were grossly underreported compared to the general population (Fatigue [HP:0012378] in two and Myalgia [HP:0003326] in one). Hence, I omitted the term Constitutional symptom [HP:0025142] to avoid overweighting of the similarity between what are phenotypically only loosely related symptoms, making Fatigue [HP:0012378] and Pain [HP:0012531] (parent of Myalgia [HP:0003326]) children of Phenotypic abnormality [HP:0000118] only.
HP:0002372	Normal interictal EEG	While the absence of electroencephalographic abnormalities can be characteristic of some epilepsies, particularly in their early stages, I excluded this term because in HPO release 2022-02-14 it is considered a type of Interictal EEG abnormality [HP:0025373] and I did not wish the presence of a normal EEG to imply an abnormal interictal EEG.

Table 5-1. List of modifications to the HPO (version released 2022-02-14).

I created a dictionary mapping the possible responses given to 92 categorical phenotypic REDCap data fields across the three data entry tools to HPO terms (**Supplementary Table 5-1**). I reviewed the REDCap data in the context of the REDCap documented syndrome and fully harmonized HPO annotations from non-syndrome REDCap fields (**Supplementary Table 5-2**). I found that some people documented as having a given syndrome had not been documented as having clinical features expected for that syndrome in other fields. Furthermore, these people often had comparable numbers of propagated HPO terms to those who had the expected phenotypes coded separately, suggesting no difference in thoroughness of data entry, and indicating that conservative inference of core phenotypes from syndromic fields might add information. For example, there was no categorical REDCap field that would capture myoclonic-atonic seizures other than the syndrome field *Epilepsy_with_myoclonic_atonic_seizures*. Similarly myoclonic absence seizures could only be captures with the syndrome field *Epilepsy_with_Myoclonic_Absences* or free-text. This encouraged me to make use of syndromic data, mapping these conservatively to HPO terms of features one would expect to be present given that the syndrome had been diagnosed at an academic epilepsy center (guided beyond formal diagnostic criteria through discussion with Dr Thomas and Dr Helbig).

5.4.6 Harmonization of quantitative phenotypic data

The REDCap DEE form contained free-text data regarding quantitative variables of gestational age, weight at birth, and head circumference at birth. I cleaned and mapped these to HPO definitions conservatively, based on population distributions from the INTERGROWTH-21st Project (Villar *et al.*, 2014) using the R *growthstandards* package (version 0.1.4, **Table 5-2**). I calculated lower and higher estimates for these three measures for each individual with sufficient data. I only applied HPO annotations if both lower and higher estimates met the annotation threshold for the phenotype. This allowed us to include raw data that were qualitative (for example, 'normal') or ambiguous (for example, *gestational_age* = '42.6', which could mean 42 weeks and 6 days, or 42 weeks and 4 days rounded to one decimal place, and 'VIII months' could span 32–35 weeks). With regards to gestational age, which is required to calculate z-scores for birth weight and head circumference at birth, I conservatively interpreted integer values as spanning 0.9 weeks from the number of weeks completed (for example, *gestational_age* = '33' was mapped to a lower estimate of 33.0 weeks and a higher estimate of 33.9 weeks). Higher estimates of gestational age were used to calculate lower *z*-score estimates of birth weight and head circumference at birth, and vice versa.

Variable	Criterion	HPO term and CUI	HPO concept definition
gestational_age	Higher estimate < 37.0	*Premature birth [HP:0001622]*	The birth of a baby of less than 37 weeks gestational age.
gestational_age	Lower estimate ≥ 42.0.	*Postterm pregnancy [HP:0031169]*	A pregnancy that extends to 42 weeks of gestation or beyond.
birth_weight	Higher estimate of z-score < -1.28, based on the reported sex and lower estimate of gestational age	*Small for gestational age [HP:0001518]*	Smaller than normal size according to sex and gestational age-related norms, defined as a weight below the 10th percentile for the gestational age.
birth_weight	Lower estimate of z-score > +1.28, based on the reported sex and higher estimate of gestational age	*Large for gestational age [HP:0001520]*	The term large for gestational age applies to babies whose birth weight lies above the 90th percentile for that gestational age.
head_circumference	Higher estimate of z-score < -2.00, based on the reported sex and lower estimate of gestational age	*Primary microcephaly [HP:0011451]*	Head circumference below 2 standard deviations below the mean for age and gender at birth.
head_circumference	Lower estimate of z-score > +2.00, based on the reported sex and higher estimate of gestational age	*Macrocephaly at birth [HP:0004488]* child of *Macrocephaly [HP:0000256]*	The presence of an abnormally large skull with onset at birth. Occipitofrontal (head) circumference greater than 97th centile compared to appropriate, age matched, sex-matched normal standards.

Table 5-2. Harmonization of quantitative phenotypes.

5.4.7 Harmonization of free-text phenotypic data

I annotated all free-text data with HPO terms manually, taking advice from adult (Dr Thomas) and a pediatric (Dr Helbig) epileptologists when required. I annotated 11 free-text phenotypic data fields from the three REDCap forms with HPO terms (**Table 5-3**). Free-text data were provided for 6,408/12,611 members (50.8 %) of the analysis cohort: 1,002/1,738 people with data entered using the DEE form, 1,540/3,736 with data entered via the generalized/mixed epilepsy form, and 3,866/7,137 with data entered via the focal epilepsy form.

Variable	REDCap form			Total
	DEE	Focal	Generalized/mixed	
neuroimaging_specify	476	2,234	327	3,037
other_seizures_specify	53	538	267	858
neurological_specify	52	378	132	562
syndrome_notes	-	2,602	1,077	3,679
eeg_epicom	139	-	250	389
other_abnormalities_comment	248	-	-	248
neonatal_comments	237	-	-	237
movement_disorder_comment	227	-	-	227
ee_syndrome_notes	166	-	-	166
dysmorphic_comment	150	-	-	150
multiple_syndromes	57	-	-	57
Total	1,805	5,752	2,053	9,610

Table 5-3. The free-text variables and number of free-text entries on each REDCap form. The number of individuals (or in the case of *eeg_findings_epicom*, the number of entries across three instances) with raw data in the cohort of 12,611 individuals. Dash entries indicate that the variable was absent from the form.

I elected to do this manually rather than relying on natural language processing because the corpus was not sufficiently large to train an automated pipeline to approach the expected performance of a single expert annotator, particularly given the linguistic intricacies of a data provided by international collaborators and nuances of seizure classification. I handled each free-text field regarding a specific domain separately taking the following steps to increase efficiency and facilitate error checking:

1. I used *tidytext::unnest_tokens()* (version 0.3.0) to tokenize free-text to single words and used *stopwords::stopwords(language = 'en', source = 'snowball')* (version 2.1) to remove stopwords except negations and epistemic qualifiers, and also removed numeric digits.
2. I ranked tokens by their frequency and built a thesaurus to harmonize frequent character strings referring to synonyms, spelling variations, spelling errors, and acronyms.

3. I reunited the tokens to produce cleaned approximations of the raw text with reduced dimensionality.
4. I proceeded to annotate every unique raw text entry manually in Microsoft Excel. I sorted the data to be annotated according to the cleaned approximations allowing me to work in thematically related batches that were likely to require similar repertoires of HPO concepts. This reduced the burden and potential error of looking up disparate HPO terms. In addition to the use of a single annotator, contemporaneous annotation of data relating to the same epilepsy themes is likely to have reduced the risk of arbitrary differences in interpretation. For example, all entries with a cleaned approximation relating to 'dysplasia' were annotated in a single batch.
5. Raw text was annotated with human readable HPO term names (rather than their CUIs) concatenated using semi-colon separators. Human-readable annotations facilitated error checking and referral back to check how thematically similar free-text had been annotated previously.

To assist manual review for quality control, each thematic batch was checked in R for quality control. The HPO term name annotations were separated into different rows using *tidyr::separate_rows(sep = ';')* and merged with a data frame representation of the HPO using *dplyr::left_join(annot.df, hpo.df, by = 'term.name')* to add CUIs to the annotations where the text value entered manually matched a complete term name in the HPO. Typographic errors in data entry resulting in a failure to match were identified using *dplyr::anti_join(annot.df, hpo.df, by = 'term.name')*. The frequency and subontologies of annotations was reviewed to assist manual quality control.

I reviewed the categorical data dictionary to make the translation of free-text consistent with the categorical data. Phenotypic concepts modified by confident qualifiers 'probable' or 'likely' were annotated as present, whereas those with less certain qualifiers such as 'presumed', 'possible', or 'suggestive of' were considered on a case-by-case basis.

5.4.8 Whole exome sequencing data generation and quality control

Whole exome sequencing was performed at the Broad Institute of Harvard and the Massachusetts Institute of Technology on the Illumina HiSeq X and or NovaSeq 6000 platforms with 150 bp paired-end reads, with reads aligned and called using the GATK pipeline (Van der Auwera *et al.*, 2013) including Variant Quality Score Recalibration, as described previously (Epi25 Collaborative, 2023). Variant calls were decomposed, and normalized using

VT (Tan;Abecasis and Kang, 2015) and annotated using Variant Effect Predictor (VEP, version 107, McLaren *et al.*, 2010) including annotation of Genomic Aggregation Database (gnomAD) exome coverage (Broad Institute), CADD (Rentzsch *et al.*, 2019), MPC (Samocha *et al.*, 2017), and SpliceAI (Jaganathan *et al.*, 2019) using dbNSFP (version 4.1a). Further processing was performed using the R statistical language.

We selected only variants absent from (all ancestral groups in) gnomAD versions 2.1.1 and 3.1. These include exomes from 125,748 people and whole genomes from at least 76,156 people. I discovered unexpectedly large numbers of variants that appeared to be ultra-rare clustering in gene regions with low gnomAD coverage. I suspect that many of these will not be truly ultra-rare but appear so because of the better coverage of modern platforms and capture kits used in our study than in the historic exome reference data within gnomAD. I decided to limit the analysis to genomic coordinates with median coverage ≥ 25 among gnomAD 2.1 exomes because I wanted to include only truly URVs. In brief, we annotated every variant with the minimum of median coverage among exomes included in gnomAD 2.1 for the coordinates it spanned plus the flanking coordinates. For example, a single nucleotide substitution at position 5 would be annotated with the minimum of the median gnomAD coverage of positions 4–6, an insertion between positions 5 and 6 would be annotated with the median gnomAD coverage of positions 5 and 6, and a deletion of positions 5–6 would be annotated with the median gnomAD coverage of positions 4–7. Each variant was then assigned the minimum of the median gnomAD coverages annotated to it. Where no coverage was provided for a coordinate in gnomAD 2.1, I assumed that coordinate to have median coverage of 0. In this way, we were able to select only variants affecting regions with median coverage ≥ 25 among all the exomes included in gnomAD 2.1, reducing the risk of misinterpreting as a variant as being rarer than it truly is because of mismatches between historic and current exome sampling bias.

We included only heterozygous and X chromosome hemizygous variant calls; the latter inferred conservatively as automated homozygous calls in males. To reduce the risk of including false positive variant calls we included only those with genotype quality ≥ 20 and read depth ≥ 10, and we excluded insertions and deletions ≥ 6 nucleotides in length. To further minimize the risk of including variants that are more common in some regions sharing ancestry than might be recognized, we excluded variants recurring with an allele count greater ≥ 3 in the complete study cohort and alleles recurring among individuals recruited from the same country. The latter filter reduces the risk of misattributing the similarity of individuals to a shared URV when they

are cryptically related and their similarity may be due to unappreciated shared risk due to SNPs, CNVs, and environmental exposures.

5.4.9 Filtering of ultra-rare variants

In the primary analysis, I focused on dominant epilepsy genes (those in which monoallelic variants have been validated as sufficient to cause epilepsy) because these are those already known to be relevant to epilepsy. I selected those 263 listed as causes of monoallelic and X-linked epilepsies in the Genomics England R59 Genetic Epilepsy Syndromes virtual clinical testing panel, version 4.178 (GE epilepsy panel, Genomics England, 2024a), with 'green' status, indicating validation for clinical use by the NHS England Genomic Medicine Service (**Supplementary Table 5-3**).

Then I sought to explore the phenotypic evidence supporting a role for a further 5,436 genes (total = 5,699, **Supplementary Table 5-3**). I selected the remainder of the 'green' genes from the GE epilepsy panel (non-dominant, $n = 293$) and from the GE panel for Intellectual Disability R29 (n = 1,363, version 5.493, Genomics England, 2024b), and those assessed as having a non-refuted[26] role in Epilepsy, Intellectual Disability, or Autism by NIH ClinGen Gene Curation Expert Panels (n = 357, accessed 2024-03-06, Helbig *et al.*, 2018; Clinical Genome Resource, 2024). These genes formed the clinically validated set (**Figure 5-1**). Beyond these genes with sufficient gene-disease association to be included in these clinical panels ($n = 1,473$), I selected 4,226 other genes that might be suspected to contain variants of clinical relevance (**Figure 5-2**). These included those not appearing in the clinical gene sets classified as 'amber' in the GE panels (epilepsy $n = 78$, intellectual disability $n = 331$), considered to be implicated in autism spectrum disorders by the Simons Foundation with 'High confidence' or as 'Strong candidates' (SFARI, n = 600, accessed 2024-03-06, Basu;Kollu and Banerjee-Basu, 2009; Simons Foundation, 2024), those predicted to be loss of function intolerant (n = 2,479, probability of being loss-of-function intolerant, pLI > 0.90, Lek *et al.*, 2016), missense-constrained (n = 623, Missense z-score, Mis-z > 3.09, Samocha *et al.*, 2014), or sensitive to gene dosage (n = 1,896, probability of haploinsufficiency, pHaplo > 0.90, or n = 1,597, triplosensitivity, pTriplo > 0.90, Collins *et al.*, 2022). For further confidence that the variants used in analysis affect biologically relevant transcripts of the gene, I limited our analyses to variants affecting 'MANE Select' or 'MANE Plus Clinical' consensus transcripts (Morales *et*

[26] 'Definitive', 'Strong', 'Moderate', 'Limited', and 'Disputed'.

al., 2022). Analysis of the membership of these gene sets demonstrated little redundancy (**Figure 5-2**).

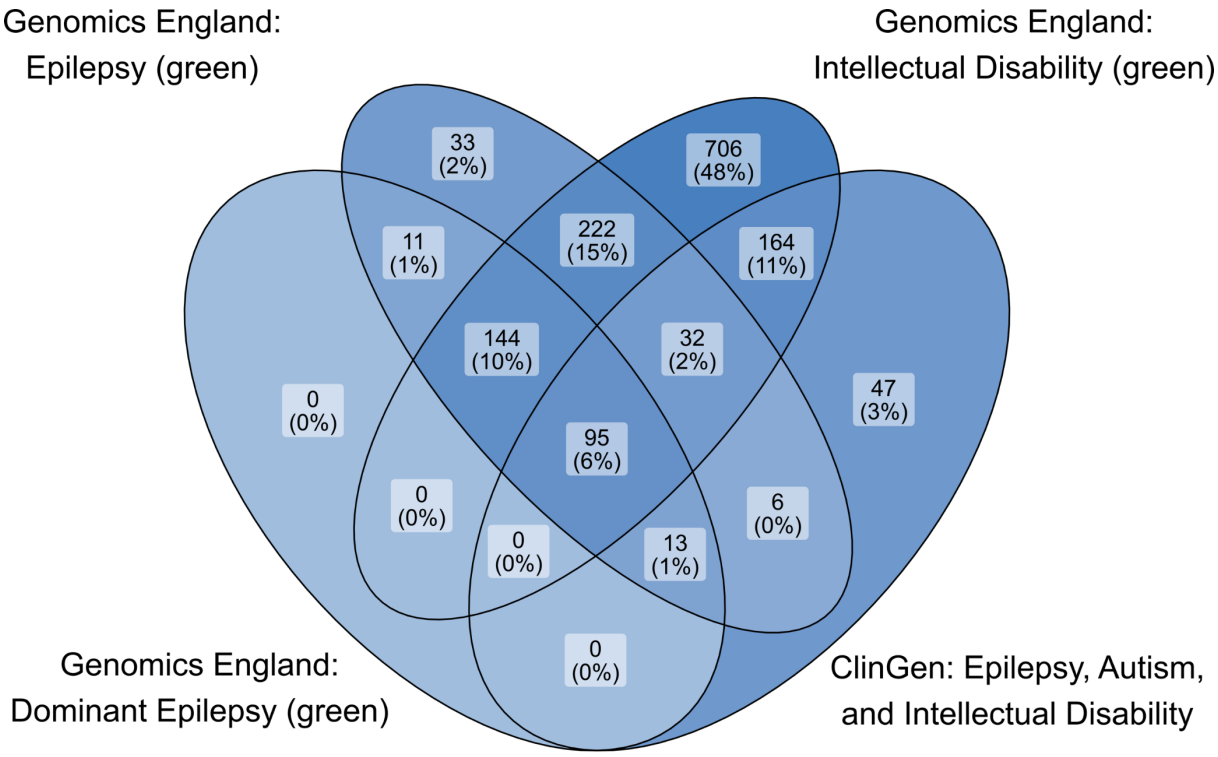


Figure 5-1. The intersections of 1,473 clinically validated genes by virtual panel.

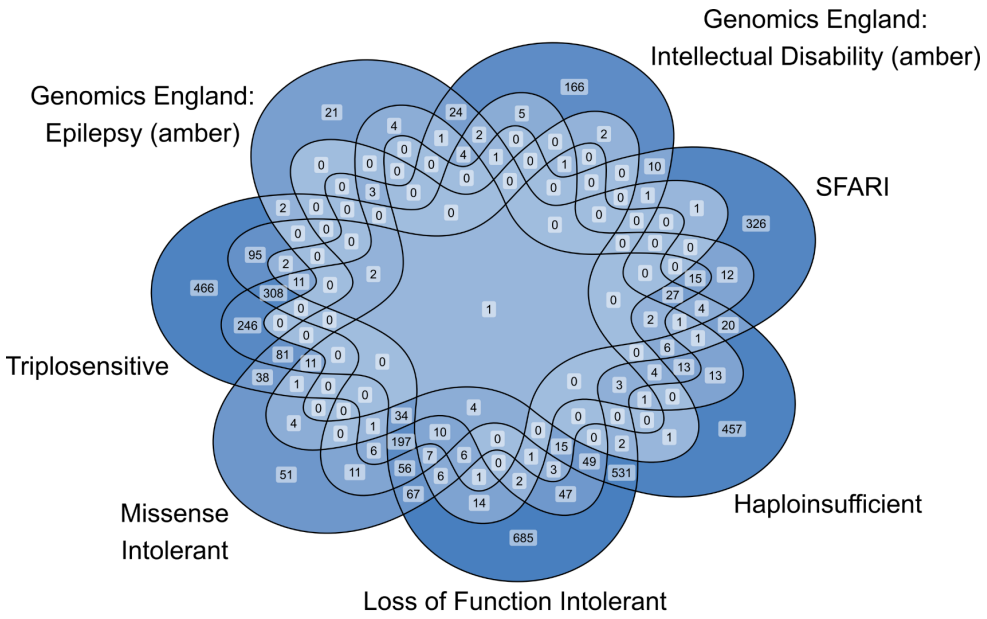


Figure 5-2. The intersections of sets of genes that have not been clinically validated.

I annotated those genes with any brain expression in adults (GTEx release 2017 06 05, version 8, Lonsdale *et al.*, 2013; Adult Genotype Tissue Expression (GTEx) Project, 2017).

5.4.10 Variant classification

I classified variants hierarchically by the severity of their predicted consequences (**Table 5-4**). Variants predicted to be in splicing regions and with any SpliceAI delta prediction ≥ 0.5 were interpreted as affecting splicing (Jaganathan *et al.*, 2019) and therefore were designated as protein truncating variants (PTVs) together with variants causing frameshift, loss of a start or stop codon, or gain of a stop codon. The presence of a CADD score ≥ 20 (in the top 1 % of variants) or MPC score ≥ 2 (Epi25 Collaborative, 2019) was considered sufficient for a missense variant to be considered deleterious (dMissense). Both CADD score < 20 and MPC < 2 were necessary to categorize a missense variant as benign (bMissense).

I took a conservative approach to any variant suspected to affect splicing because I wished to avoid misclassifying a variant as synonymous (implying that it is unlikely to have significant molecular consequences) when in fact it caused defective splicing (which can have severe molecular consequences). Variants annotated as residing in splicing regions but without SpliceAI ≥ 0.5, and those outside of splicing regions with SpliceAI ≥ 0.5, were interpreted as either deleterious protein-altering variants (dPAVs) if they also met criteria to be dMissense or possible splicing (possSplicing) if not. Consequently, variants were only classified as synonymous if this was their only consequence predicted by VEP and there was no suggestion of an effect on splicing. possSplicing variants were not presumed to be PAVs (or dPAVs). Those variants predicted as inframe deletions or insertions, or protein altering were classified only as protein-altering variants (PAV) because their precise consequences can be difficult to predict. The interpretation of possSplicing and PAV at gene-level classification was contingent on the other variants present (see 5.4.11). Some missense variants with CADD score < 20 had missing MPC scores and these were not further classified as dMissense or bMissense. All missense and PTVs were inferred to also be PAVs. Additionally, dMissense and PTVs were inferred to be dPAVs (**Figure 5-3**). The outcome of these variant classification rules can be confirmed in **Supplementary Table 5-4**.

Hierarchy	Variant class assigned (inferred)	VEP consequence	Other necessary annotations
1	PTV (dPAV, PAV)	*ANY* frameshift_variant start_loss stop_gained stop_lost transcript_ablation	
		OR ANY splice_acceptor_site splice_donor_variant splice_donor_5th_base_variant splice_region_variant splice_donor_region_variant splice_polypyrimidine_tract	*ANY* SpliceAI prediction ≥ 0.5
2	dPAV (dPAV)	*ANY* splice_acceptor_site splice_donor_variant splice_donor_5th_base_variant splice_region_variant splice_donor_region_variant splice_polypyrimidine_tract	
		AND missense_variant	CADD ≥ 20 *OR* MPC ≥ 2
3	PAV-only (PAV) [a]	*ANY* inframe_deletion inframe_insertion protein_altering_variant	
		OR missense_variant	*ANY* SpliceAI prediction ≥ 0.5
		OR ANY splice_acceptor_site splice_donor_variant splice_donor_5th_base_variant splice_region_variant splice_donor_region_variant splice_polypyrimidine_tract	
		AND missense_variant	
4	possSplicing [a]	*ANY* splice_acceptor_site splice_donor_variant splice_donor_5th_base_variant splice_region_variant splice_donor_region_variant splice_polypyrimidine_tract	
		OR synonymous_variant	*ANY* SpliceAI prediction ≥ 0.5
5	dMissense (Missense, dPAV, PAV)	missense_variant	CADD ≥ 20 *OR* MPC ≥2
6	bMissense (Missense, PAV)	missense_variant	CADD < 20 *AND* MPC <2
7	Missense (PAV)	missense_variant	(MPC missing)
8	Synonymous	synonymous_variant	

Table 5-4. The classification of single variants. [a] When collapsing each participant's variants at gene level, variants classed as only protein-altering variants or possible splicing variants were interpreted conditionally on the other URV classes in the gene.

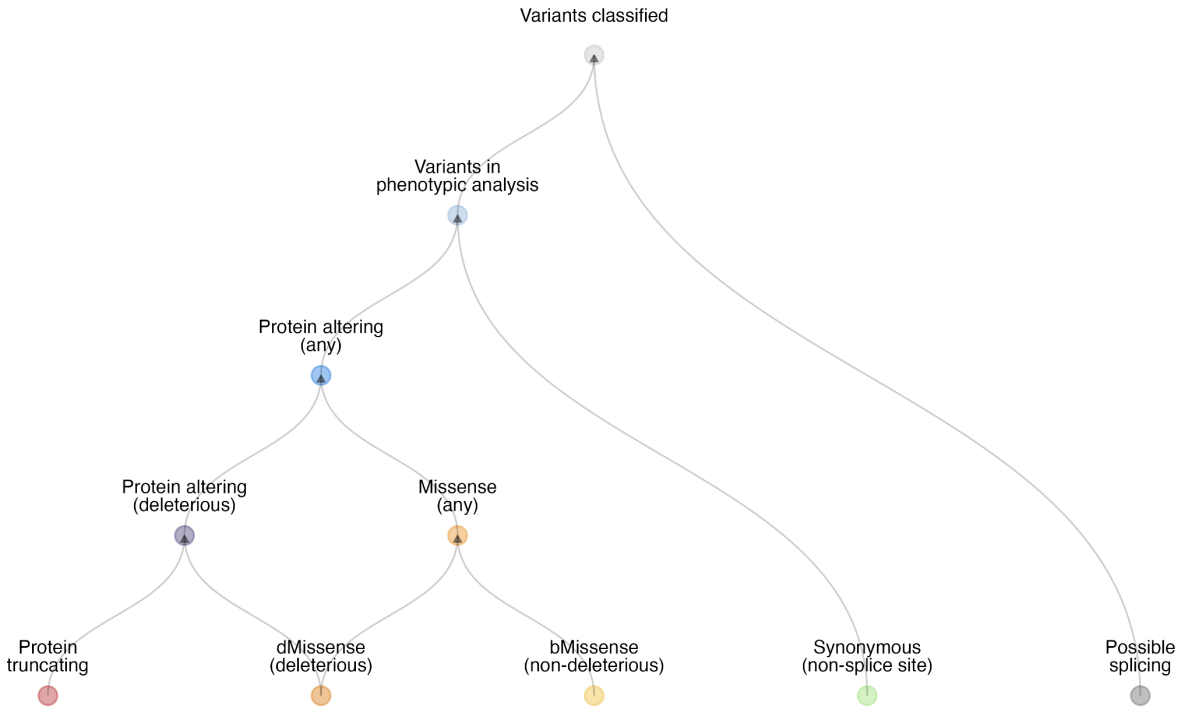


Figure 5-3. The relationship between variant classes. Interpretation of possible splicing variants at gene-level depended on the other variant classes the carried in the gene.

5.4.11 Assigning gene-level classes by participant

I collapsed the URVs of each person A at gene level hierarchically, assigning to them a set of variant classes for each gene G_A according to the most damaging class identified among their URVs in that gene, V_{AG} (**Equation 5-1**). Consider a person annotated with two URVs in the same gene. Were one variant classified as a PTV due to being annotated a *splice_donor_variant* with SpliceAI = 0.8, and another classified as bMissense due to being annotated as *missense_variant* with CADD = 12 and MPC = 1.2, the person would be annotated as having PTV for that gene. From PTV one would infer the more generic classes of dPAV and PAV regarding this gene for this participant.

$$G_A = \{\text{PTV, dPAV, PAV}\} | \text{PTV} \in V_{AG}$$

$$G_A = \{\text{dPAV, PAV}\} \left| \text{dPAV} \in V_{AG} \bigwedge \text{PTV} \notin V_{AG} \right.$$

$$G_A = \text{PAV} \left| \left(\text{PAV-only} \in V_{AG} \bigvee \{\text{possSplicing, PAV}\} \subseteq V_{AG} \right) \bigwedge \text{dPAV} \notin V_{AG} \right.$$

$$G_A = \text{possSplicing} \left| \text{possSplicing} \in V_{AG} \bigwedge \text{PAV} \notin V_{AG} \right.$$

$$G_A = \{\text{dMissense, Missense, dPAV, PAV}\} \left| \text{dMissense} \in V_{AG} \bigwedge \text{possSplicing, PAV-only, PTV} \notin V_{AG} \right.$$

$$G_A = \{\text{bMissense, Missense, PAV}\} \left| \text{bMissense} \in V_{AG} \bigwedge \text{possSplicing, PAV-only, PTV, dMissense} \notin V_{AG} \right.$$

$$G_A = \{\text{Missense, PAV}\} \left| \text{Missense} \in V_{AG} \bigwedge \text{possSplicing, PAV-only, PTV, dMissense, bMissense} \notin V_{AG} \right.$$

$$G_A = \{\text{Synonymous}\} \left| \text{Synonymous} \in V_{AG} \bigwedge \{\text{possSplicing, PAV}\} \notin V_{AG} \right.$$

Equation 5-1. Interpretation sets of variant class annotations when collapsing participant data to gene level annotations.

I interpreted consequences of possSplicing and PAV-only variants conservatively at gene level as their deleteriousness is hard to predict. If within a gene, a person had any URV annotated as possSplicing or PAV-only, they could not be annotated as any form of missense or synonymous variant for that gene as there remains a risk of an alternative, more severe consequence. The presence of possSplicing and PAV-only did not affect the primacy of PTVs, because I assumed that the loss of gene product would dominating over any other possible consequence. Where other variants in a gene were classified as dPAV, the presence of a possSplicing or PAV-only caused the gene to be annotated as dPAV. Where only synonymous and non-deleterious missense URVs were identified with PAV-only, the gene was annotated as PAV in the expectation of some uncertain consequence for the protein. Similarly, where only non-deleterious PAVs and possSplicing variants were reported, the gene was annotated as PAV. However, where only synonymous URVs and possSplicing variants were reported, the gene was annotated as possSplicing, and the carrier was excluded from any analysis groups related to that gene because of the extreme uncertainty as to the most severe consequence of their URVs. Consequently, for a person to be annotated as synonymous at gene level, they must have had no URV class from those above other than *synonymous_variant* and no variant with any SpliceAI ≥ 0.5.

As a result of this process, each person was annotated for each gene with one of the following set of annotations:

1. PAV.
2. dPAV, PAV.
3. PTV, dPAV, PAV.
4. dMissense, Missense, dPAV, PAV.
5. bMissense, Missense, PAV.
6. Missense, PAV.
7. Synonymous.
8. No qualifying variant.

This enabled analysis according to seven sets of participants per gene according to the most severe predicted consequence of their URVs in that gene, related as per **Equation 5-2**.

1. PAV: participants that I predict carry a URV that in alters the resulting protein in some way. This forms the primary analysis group.
2. dPAV: participants with predicted PTVs or dMissense variants.
3. PTV: participants with predicted PTVs.
4. Missense: participants with predicted missense variants but unlikely to carry PTVs.

5. dMissense: participants with predicted dMissense variants but unlikely to carry PTVs.
6. bMissense: participants with predicted bMissense variants but unlikely to carry PTVs or dMissense variants.
7. Synonymous: participants with only predicted synonymous variants that are unlikely to affect splicing. This is expected to act as a negative control on the assumption that the majority of these variants will be biologically neutral.

$$\mathrm{dPAV} \subseteq \mathrm{PAV}$$
$$\mathrm{PTV} \subseteq \mathrm{dPAV}$$
$$\mathrm{dMissense} \subseteq \mathrm{dPAV}$$
$$P(\mathrm{Missense}|\mathrm{PTV}) = 0$$
$$\mathrm{dMissense} \subseteq \mathrm{Missense}$$
$$\mathrm{bMissense} \subseteq \mathrm{Missense}$$
$$P(\mathrm{dMissense}|\mathrm{bMissense}) = 0$$
$$P(\mathrm{Synonymous}|\mathrm{PAV}) = 0$$

Equation 5-2. The relationship between sets of participants annotated to each gene.

5.4.12 Cohort definition

After harmonizing available phenotypic data available for all participants, I included only those whose clinical data indicated a history of a clinical seizure or a DEE with Spike-and-Wave Activation in Sleep, special epilepsy syndromes in which clinical seizures are not mandatory (Specchio *et al.*, 2022). I excluded those documented as having an uncertain diagnosis of epilepsy, and those with evidence of an acquired case for epilepsy including significant head trauma that was not clearly documented as postdating epilepsy onset, and CNS infective or inflammatory illnesses. I did not exclude those with brain tumors or cerebrovascular disease as these may be phenotypes of an underlying genetic epilepsy. Dissociative/non-epileptic seizures were not an exclusion criterion unless doubts as to whether the participant had comorbid epilepsy were documented. I excluded those whose genetically imputed sex contradicted that reported in the phenotypic data in case of data mismatch.

5.4.13 Genotype-phenotype analysis

I undertook two genotype-stratified analyses of phenotypic data. I used $\mathrm{Sim}_{\mathrm{MinTO}}$ as the primary measure, with 14 further symmetric similarity measures to provide further validity and different perspectives: $\mathrm{Sim}_{\mathrm{TO}}$, $\mathrm{Sim}_{\mathrm{Resnik\text{-}BMS}}$, $\mathrm{Sim}_{\mathrm{Resnik\text{-}BMA}}$, $\mathrm{Sim}_{\mathrm{Resnik\text{-}Max}}$, $\mathrm{Sim}_{\mathrm{Phrank}}$, $\mathrm{Sim}_{\mathrm{RBP}}$ (setting

$\alpha = 0.001$, which caps the informativeness of a single phenotypic concept to that of a phenotype annotated to 0.1 % of the cohort), $\mathrm{Sim}_{\mathrm{Goodall3}}$ (making a closed-world assumption about those phenotypes that are not annotated), $\mathrm{Sim}_{\mathrm{Rel\text{-}BMS}}$, $\mathrm{Sim}_{\mathrm{Rel\text{-}BMA}}$, $\mathrm{Sim}_{\mathrm{TA}}$ (which is based on the probability of encountering pairs with the same combination of phenotypes), $\mathrm{Sim}_{\mathrm{Cosine}}$, $\mathrm{Sim}_{\mathrm{CosIC}}$, $\mathrm{Sim}_{\mathrm{UI}}$, and $\mathrm{Sim}_{\mathrm{GIC}}$.

I generated null distributions of the median of the ${}^{n}C_2$ pairwise similarities for each group size n in the cohort by Monte Carlo simulation of 1,000,000 randomly selected groups of n unique individuals. The rank of the observed median similarity of the group of interest among the simulations provides a one-sided p-value, allowing comparison of point estimates of group similarity across groups of different sizes in relation to the phenotypic landscape of the cohort.

The distribution of the observed p-values of a set of m genetically defined groups of sizes n_1, n_2, … n_m was interpreted by Monte Carlo simulation of the expected distribution of p-values from 1,000 randomly selected groups for each size n_1, n_2, … n_m. While the Kolmogorov–Smirnov test allows comparison of the distribution of p-values, it sensitive to the discrete nature of similarity values and takes account of the entire unit interval of p-values, whereas I am interested in those groups of sufficiently high similarity to be of potential clinical significance. Therefore, I counted the number (and percentage) of gene-defined groups reaching similarity p-values < 0.05 (selected arbitrarily) to evaluate the effect of URV class and similarity measure. While I primarily interpret each similarity p-value as the degree of similarity relative standardized for group size in the study cohort, rather than statistical significance, I provide Holm's adjusted p-values (*stats::p.adjust(method='holm')*) for the primary analysis of PAVs in dominant epilepsy genes by $\mathrm{Sim}_{\mathrm{MinTO}}$.

The second analysis quantified the statistical association of single HPO concepts to each group of participants. The frequency of annotation of each HPO term was calculated in carriers and non-carriers and described by odds ratio (OR) with two-sided p-values calculated using Fisher's exact test (*stats::fisher.test(alternative='two.sided')*). To aid interpretation of the large number of associations tested for each group while allowing for the dependency structure of HPO annotations, I also report *minP* single-step p-values (Westfall and Wolfinger, 1997) based on Monte Carlo simulation using 100,000 random samples of n individuals. This retains the dependency structure of HPO terms. I report relative risks (RR) and positive predictive values (PPV) to aid interpretation. While globally, participants were not recruited due to having any

particular type of presumed genetic epilepsy[27] it should be noted that the validity of these precise OR, RR, and PPVs applies to the phenotypes annotated to participants in a research study rather than the features a clinician may identify during thorough clinical review.

5.4.14 Data availability

Investigators wishing to apply for access to the Epi25 Collaborative genetic and clinical data used for this analysis should follow the guidance at epi-25.org/epi25-data-application. A subset of the raw data is summarized on dbGaP: phs001489 and deidentified genetic results are released periodically on the Epi25 WES Browser (epi25.broadinstitute.org).

5.5 Results

5.5.1 Description of the study cohort

I matched clinical data about participants within years 1–3 of Epi25 available on 2022-1-25 to exome data (Epi25 Collaborative, 2019; Epi25 Collaborative, 2021). After applying the inclusion and exclusion criteria (5.4.12), this generated a cohort of 12,611 individuals spanning 41 epilepsy centers across 21 countries (**Figure 5-4** and **Table 5-5**).

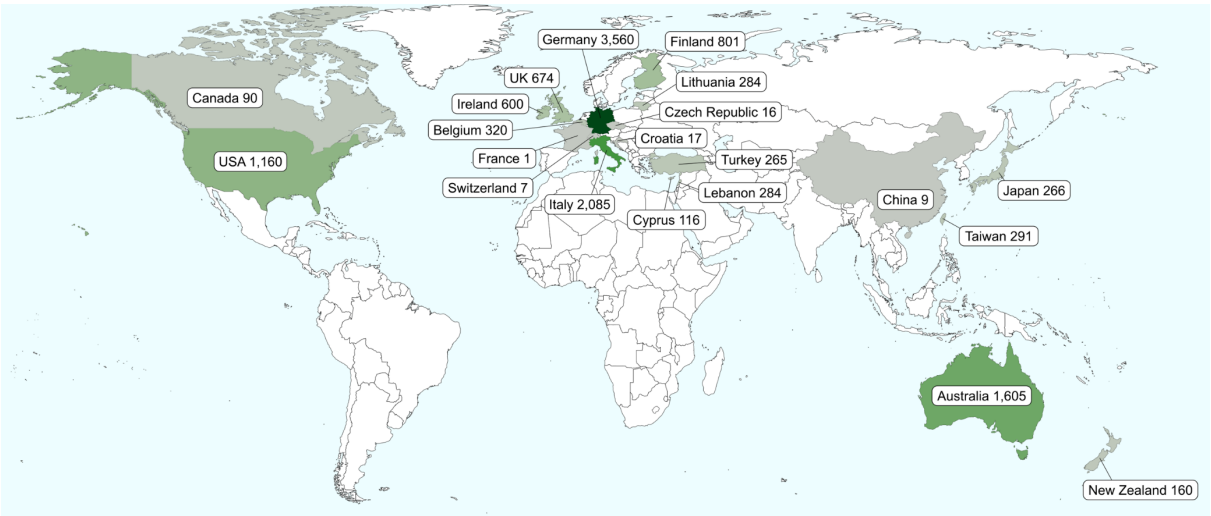


Figure 5-4. The geographic distribution of the Epi25 cohort.

[27] Although well-established and relatively common genetic epilepsies such as Dravet syndrome and tuberous sclerosis should be under-represented as collaborators may have made these diagnoses clinically and were asked not to recruit patients with valid molecular diagnoses.

Feature *[HPO CUI]*	Number and percentage of people or median, interquartile range, number
Sex	
Female	6,610 (52.4 %)
Male	5,767 (45.7 %)
Age at last data collection (years)	29 (16–46, *n* = 10,562)
Gestational age (weeks)	39.5 (38.5–40.5, *n* = 864)
Birth weight (gestational age and sex-adjusted z-score)	-0.12 (-0.82–0.53, *n* = 687)
Head circumference (gestational age and sex-adjusted z-score)	+0.50 (-0.41–1.51, *n* = 218)
Seizure *[HP:0001250]*	12,596 (99.9 %)
Focal-onset seizure *[HP:0007359]*	7,834 (62.1 %)
Generalized-onset seizure *[HP:0002197]*	4,389 (34.8 %)
Both focal-onset *[HP:0007359]* and generalized-onset seizures *[HP:0002197]*	378 (3.0 %)
Bilateral tonic-clonic seizure *[HP:0002069]*	6,617 (52.5 %)
Refractory drug response to seizures *[HP:0020174]*	2,713 (21.5 %, 47.8 % [a])
Neurodevelopmental abnormality *[HP:0012759]*	2,119 (16.8 %)
EEG abnormality *[HP:0002353]*	10,278 (81.5 %)
Abnormal cerebral morphology *[HP:0002060]*	2,780 (22.0 %)

Table 5-5. The demographics and broad clinical features of the cohort. Percentages provided refer to the complete cohort (*n* = 12,611) except where a different denominator is indicated. [a] With reference to only those with the presence or absence of drug-resistant epilepsy (Kwan *et al.*, 2010) indicated explicitly (*n* = 5,672, 45.0 % of the cohort).

5.5.2 Clinical data harmonization

Translation and inference of the clinical features of the cohort generated 291,432 annotations (**Table 5-6**) with a median of 20 annotations per individual (**Figure 5-5**). Categorical and quantitative data yielded 256,441 annotations after propagation. After propagation, free-text contributed 34,991 annotations beyond those that could be inferred from categorical and quantitative data (13.6 % increase), despite most individuals gaining none. Yet, when present, free-text annotations tended to be highly informative: one person gained 95 propagated annotations and the maximum amount of information gained was 897 bits. Overall, the pooled information about the cohort increased by 199,523 bits (52.0 %) on additional of free-text data.

Data source	Number of participants with data (% of cohort)	Total number and *information* of annotations pooled from cohort			Number and *information (bits)* of annotations per participant [b] Median (range)		
		Input	Minimal	Propagated	Input	Minimal	Propagated
Categorical data	12,605 (99.95 %)	77,319 *179,414*	49,489 *145,142*	255,426 *377,268*	5 (1–28) *9.57 (0.002–106)*	3 (1–20) *7.88 (0.002–89.0)*	19 (5–73) *19.6 (0.002–252)*
Quantitative data	263 (2.09 %)	290 *2,222*	290 *2,222*	1,718 *7,463*	1 (1–3) *7.25 (6.98–24.4)*	1 (1–3) *7.25 (6.98–24.4)*	6 (4–27) *27.0 (13.4–123)*
Free-text data	6,152 (48.8 %)	13,076 *69,778*	15,468 *76,897*	94,102 *267,764*	1 (1–19) *7.54 (0–196)*	2 (1–21) *9.06 (0–198)*	13 (3–105) *26.9 (0–903)*
Combined [a]	12,611 (100 %)	56,001 *202,272*	88,144 *248,603*	291,432 *583,524*	4 (1–35) *9.78 (0.002–268)*	6 (1–45) *12.8 (0.002–295)*	20 (5–132) *25.4 (0.002–968)*

Table 5-6. The number of annotations and amount of information describing the phenotypes of the cohort and per participant. [a] Sets of annotations from different sources intersect. [b] Excluding individuals without any annotations of the specified provenance. Information is shown in italics and calculated summing the (unadjusted) IC calculated from the full harmonized (propagated) data as the best estimate of the ground truth.

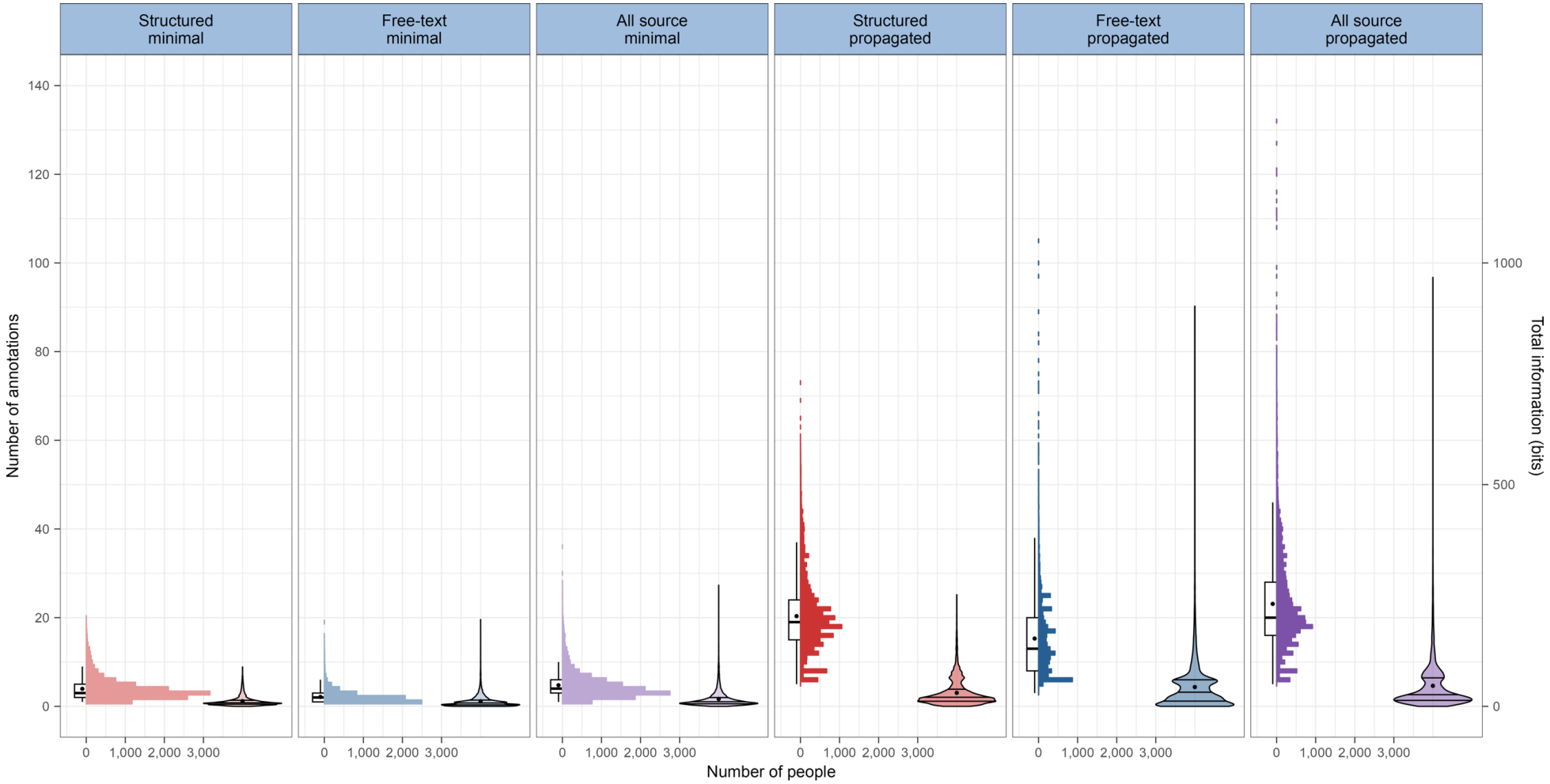


Figure 5-5. The number and informativeness of annotations per person.The left-hand three plots show this before (minimal sets, removing redundant annotations) and the right-hand three after inference (propagated sets) according to source. Structured data include categorical and quantitative variables. Bar and boxplots relate to the number of terms annotated. The boxplots indicate the quartiles with whiskers indicating the locations of the values within 1.5-fold of the interquartile range. Violin plots indicate the distribution of information per person. Black points within boxplots and violin plots indicate mean values.

The number of non-redundant input annotations regarding each participant (the size of each participant's minimal set) correlated with both their number of propagated annotations (Kendall's rank correlation, *stats::cor.test(method = "Kendal")*, $\tau = 0.83$, $z = 137.08$, $p < 2.2 \times 10^{-16}$) and their total information after propagation ($\tau = 0.70$, $z = 110.52$, $p < 2.2 \times 10^{-16}$).

After propagation the clinical repertoire of this cohort consisted of 1,738 clinical terms (**Figure 5-6, Figure 5-7,** and **Supplementary Table 5-5**). 1,107 concepts (63.7 %) were explicitly captured by translation of raw data without inference. Most diversity arose from free-text (1,094 terms, 62.9 %) rather than categorical and quantitative fields (120 terms, 6.90 %). 107 terms (6.16 %) were captured in both structured and free-text data. 631 terms (36.3 %) were annotated purely by inference using the HPO. Propagation of categorical and quantitative field data yielded 223 terms (12.8 %), whereas free-text yielded 1,727 terms (99.4 %), of which 1,515 (87.2 % of the total 1,738 concepts) would not have been captured solely by propagation from categorical and quantitative data. The clinical data are sparse with only 157/1,738 HPO terms (9.03 %) annotated to more than 126 people, 1 % of the cohort. Four terms reflect the inclusion criteria were annotated ubiquitously and are thus uninformative (*All [HP:0000001]*, *Phenotypic abnormality [HP:0000118]*, *Abnormality of the nervous system [HP:0000707]*, and *Abnormal nervous system physiology [HP: 0012638]*). 553 terms (30.7 %) were annotated to only one person.

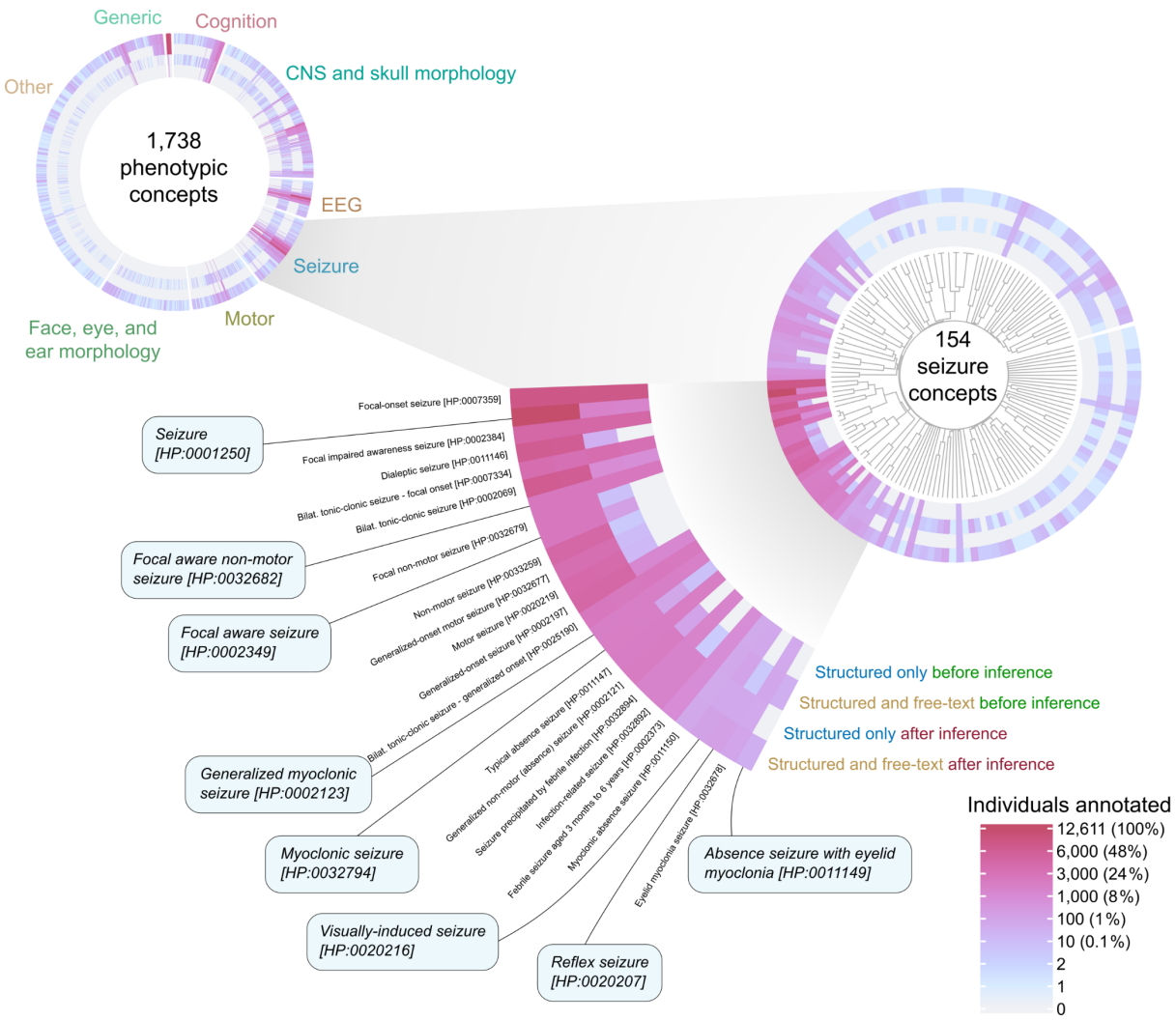


Figure 5-6. The phenotypic repertoire of the cohort. The four tracks of the circular heatmap indicate the number of individuals annotated with each HPO phenotype based on either structured (including categorical and quantitative) source data only or all available data (also including free-text), before and after inference using the HPO *is_a* relationships. Six of 12,611 individuals met inclusion criteria only after analysis of free-text. Illustration created using the *circlize* package (version 0.4.15). Terms are arranged by hierarchical clustering of their annotation to the cohort after propagation from all data (*stats::hclust(stats::dist(method='binary'), method='average')*).

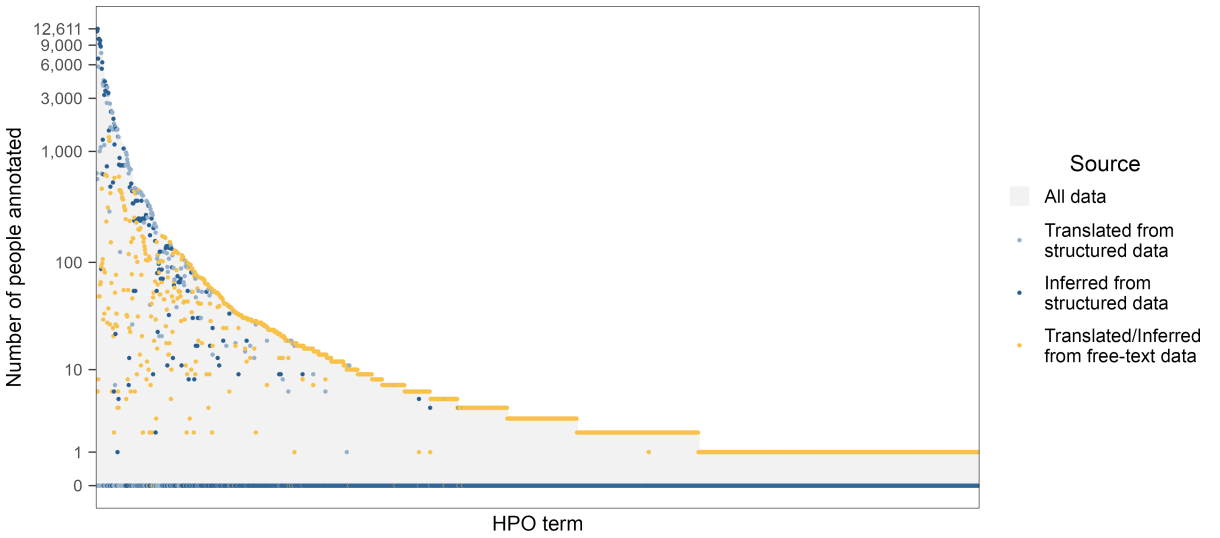


Figure 5-7. The distribution of HPO term frequency in Epi25, stratified by source of annotation with HPO terms ranked by frequency. 1,738 HPO terms are ordered along the x-axis by descending frequency of annotation after propagation of all data sources. The number of participants annotated with each term according to all sources (gray bars), translated structured source data (light blue points), inferred from structured data (dark blue points), and translation or inference from free-text only (yellow points) are shown on the y-axis plotted on $\log_{10}(y + 1)$ scale. A small number of HPO terms are annotated very frequently because of inference from structured data (mostly far left, where dark blue points often approach to the 'all data' gray value). A small number of HPO terms are annotated less frequently, largely from direct translation of structured data (left, light blue). The majority of HPO terms are annotated purely because of translation of, or inference from, free-text data and these are annotated more sparsely (central and right, where yellow points are equal or close to the 'all data' gray value). Some terms are annotated to some participants on the basis of structured data and to others on the basis of free-text data, suggesting heterogeneity in the completion of case record forms by collaborators (left, where all points are plotted below the 'all data' gray value).

5.5.3 Ultra-rare variants

We were able to assess 5,418 genes of the intended 5,699 that met my panel-based inclusion criteria (**Table 5-7**). The only GE epilepsy panel dominant gene we did not capture was *WNK3* (X chromosome). Nine from the GE epilepsy panel were not captured: *AFG2A*, *AFG2B*, *BLTP1*, *C2orf69*, *C12orf57*, *GBA1*, *MADD*, *NRROS*, and *SNORD118*; and nine further from the GE intellectual disability panel: *C12orf4*, *CPLANE1*, *HYCC1*, *MTRFR*, *POLR2A*, *RNU7-1*, *SDHAF1*, *SHANK3*, and *SKIC3*. Two genes from the ClinGen panel were not captured: *CNTN2* and *SHANK3*.[28]

We identified 5,517 participants (43.7 % of the cohort) with at least one PAV in at least one of the 262 dominant genes from the GE epilepsy panel for the primary analysis (**Table 5-8**). Two

[28] *C2orf69*, *C12orf57*, *SNORD118*, *C12orf4*, and *RNU7-1* have not yet been assigned MANE transcripts.

thirds (8,360, 66.3 %) carried PAVs in 546 genes from the GE epilepsy panel, and 11,826 (93.8 %) carried PAVs in one of the 1,453 clinically validated epilepsy, intellectual disability, or autism spectrum genes. 31 participants did not carry any URV in any of the 5,418 genes in the study.

Genes	Selected	Included in analysis	Any variant	PAV	dPAV	PTV	dMissense	Missense	bMissense	Synonymous	possSplicing
All genes	5,699	5,418	5,410	5,392	5,382	4,314	5,348	5,383	4,197	5,405	5,057
Clinical gene panels											
All clinical	1,473	1,454	1,453	1,448	1,446	1,255	1,438	1,446	1,174	1,450	1,384
GE Epilepsy	556	546	546	544	543	478	539	543	439	544	525
Dominant	263	262	262	260	259	211	257	259	199	262	248
Others	293	284	284	284	284	267	282	284	240	282	277
GE ID	1,363	1,346	1,345	1,340	1,338	1,163	1,333	1,338	1,083	1,342	1,283
ClinGen	357	355	355	354	353	289	352	354	292	355	336
Other gene sets											
All other	4,770	4,712	4,705	4,687	4,678	3,658	4,651	4,680	3,604	4,703	4,387
GE Epi amber	194	149	149	149	149	125	148	149	115	149	138
GE ID amber	372	182	182	181	181	149	179	181	138	182	174
SFARI	934	922	922	919	918	797	913	918	745	921	882
pHaplo > 0.90	2,298	2,279	2,273	2,266	2,265	1,799	2,260	2,263	1,758	2,272	2,151
pLI > 0.90	3,029	2,998	2,998	2,989	2,983	2,279	2,976	2,986	2,323	2,998	2,844
Mis-z > 3.09	951	941	941	937	936	757	935	936	683	941	914
pTriplo > 0.90	1,958	1,937	1,937	1,928	1,922	1,491	1,907	1,923	1,416	1,937	1,813

Table 5-7. The number of genes carrying different URV classes by gene set. GE = Genomics England Clinical Testing Panel; Epi = Epilepsy; ID = Intellectual disability. Genes intersecting multiple sets are counted for each.

Exploring the distribution of URVs further, 8,179 people (64.9 %) carried URVs in multiple genes in our primary analysis set of dominant epilepsy genes (median = 3, mean = 3.43, mode = 1, range = 0–34 genes). 1,563 people (12.4 %) carried PAVs in multiple genes from the 262 (median = 0, mean = 0.590, mode = 0, range = 0–8), 1,018 of whom carried dPAVs in multiple genes (8.07 % of cohort, mean = 0.465, range = 0–7), and 49 carried PTVs in multiple genes (0.390 % of cohort, mean = 0.0875, range = 0–5), and 656 carried dMissense in multiple genes (5.20 % of cohort, mean = 0.365, range = 0–6). 3,688 people carried synonymous variants in multiple genes from the primary panel (29.2 %, median = 1, mean = 1.11, mode = 0, range = 0–15 genes). Regarding the study global set of 5,418 genes, nearly all (12,557/12,611, 99.6 %) carried URVs in multiple genes, and similarly when looking at only PAVs (12,460/12,611 = 98.8 %, median = 11, mean = 11.3, mode = 10, range = 0–77) or synonymous URVs (12,425/12,611 = 98.5 %, median = 16, mean = 18.5, mode = 14,

range = 0–209). Half carried PTVs in multiple genes (6,190/12,611 = 49.1 %, median = 1, mean = 1.67, mode = 1, range = 0–36).

Genes	Any variant	PAV	dPAV	PTV	dMissense	Missense	bMissense	Synonymous	possSplicing
All genes	12,580 (99.8 %)	12,538 (99.4 %)	12,514 (99.2 %)	10,033 (79.6 %)	12,387 (98.2 %)	12,464 (98.8 %)	10,473 (83 %)	12,506 (99.2 %)	12,355 (98 %)
Clinical gene panels									
All clinical	12,485 (99 %)	11,826 (93.8 %)	11,234 (89.1 %)	5,331 (42.3 %)	10,151 (80.5 %)	11,099 (88 %)	5,503 (43.6 %)	12,123 (96.1 %)	10,232 (81.1 %)
GE Epilepsy	11,856 (94 %)	8,360 (66.3 %)	7,281 (57.7 %)	2,302 (18.3 %)	5,958 (47.2 %)	7,058 (56 %)	2,095 (16.6 %)	9,941 (78.8 %)	6,054 (48 %)
Dominant	10,470 (83 %)	5,517 (43.7 %)	4,668 (37 %)	1,050 (8.33 %)	3,850 (30.5 %)	4,606 (36.5 %)	1,014 (8.04 %)	7,925 (62.8 %)	3,768 (29.9 %)
Others	9,660 (76.6 %)	5,253 (41.7 %)	4,284 (34 %)	1,369 (10.9 %)	3,179 (25.2 %)	4,096 (32.5 %)	1,186 (9.4 %)	6,329 (50.2 %)	3,483 (27.6 %)
GE ID	12,476 (98.9 %)	11,717 (92.9 %)	11,057 (87.7 %)	5,098 (40.4 %)	9,935 (78.8 %)	10,934 (86.7 %)	5,265 (41.7 %)	12,065 (95.7 %)	9,977 (79.1 %)
ClinGen	11,372 (90.2 %)	7,202 (57.1 %)	6,055 (48 %)	1,421 (11.3 %)	5,113 (40.5 %)	6,202 (49.2 %)	1,902 (15.1 %)	9,202 (73 %)	4,901 (38.9 %)
Other gene sets									
All other	12,576 (99.7 %)	12,530 (99.4 %)	12,490 (99 %)	9,162 (72.7 %)	12,338 (97.8 %)	12,445 (98.7 %)	10,014 (79.4 %)	12,503 (99.1 %)	12,286 (97.4 %)
GE Epi amber	7,805 (61.9 %)	3,447 (27.3 %)	2,732 (21.7 %)	615 (4.88 %)	2,147 (17 %)	2,804 (22.2 %)	723 (5.73 %)	4,879 (38.7 %)	2,291 (18.2 %)
GE ID amber	8,682 (68.8 %)	3,865 (30.6 %)	3,058 (24.2 %)	661 (5.24 %)	2,458 (19.5 %)	3,160 (25.1 %)	827 (6.56 %)	5,775 (45.8 %)	2,682 (21.3 %)
SFARI	12,490 (99 %)	11,437 (90.7 %)	10,594 (84 %)	4,318 (34.2 %)	9,520 (75.5 %)	10,667 (84.6 %)	4,994 (39.6 %)	12,071 (95.7 %)	9,015 (71.5 %)
pHaplo > 0.90	12,535 (99.4 %)	12,323 (97.7 %)	12,083 (95.8 %)	5,748 (45.6 %)	11,670 (92.5 %)	12,057 (95.6 %)	7,342 (58.2 %)	12,377 (98.1 %)	11,468 (90.9 %)
pLI > 0.90	12,548 (99.5 %)	12,438 (98.6 %)	12,287 (97.4 %)	6,265 (49.7 %)	11,969 (94.9 %)	12,259 (97.2 %)	8,209 (65.1 %)	12,424 (98.5 %)	11,876 (94.2 %)
Mis-z > 3.09	12,440 (98.6 %)	10,776 (85.4 %)	10,028 (79.5 %)	3,033 (24.1 %)	9,062 (71.9 %)	9,910 (78.6 %)	3,142 (24.9 %)	11,885 (94.2 %)	9,422 (74.7 %)
pTriplo > 0.90	12,504 (99.2 %)	11,971 (94.9 %)	11,487 (91.1 %)	4,728 (37.5 %)	10,812 (85.7 %)	11,483 (91.1 %)	5,666 (44.9 %)	12,280 (97.4 %)	10,774 (85.4 %)

Table 5-8. The number of carriers of different URV classes by gene set. GE = Genomics England Clinical Testing Panel; Epi = Epilepsy; ID = Intellectual disability.

There was a significant difference in the number of people carrying an URV in any of the 262 primary analysis genes by country (1×10^5 permutation estimated two-sided FET $p = 1 \times 10^{-5}$). Focusing on PAVs, there were differences for both the 262 dominant ($p = 1 \times 10^{-5}$). The difference was evident across all the other seven variant classes in both the dominant and global sets (all $p = 1 \times 10^{-5}$, except for PTV, where $p = 0.0175$). A similar pattern was seen for the global study gene set and on repeating the analysis on the number of people carrying multiple URVs. Excluding the countries with fewer than 50 participants, the countries with the greatest proportion of participants carrying PAVs in any dominant epilepsy gene were Cyprus (71/116 = 61.2 %), Turkey (161/265 = 60.8 %), Japan (157/266 = 59.0 %), and Lebanon (159/284 = 56.0 %), and those with the lowest were Finland (239/801 = 29.8 %), Ireland (187/600 = 31.2%), the United Kingdom (258/674 = 38.3 %), Germany

(1,499/3,560 = 42.2 %), and the United States of America (490/1,160 = 42.2 %). The countries with the greatest proportion of participants with synonymous variants were Taiwan (250/291 = 85.9 %), Lebanon (236/284 = 84.1 %), and Cyprus (94/116 = 91 %), and the those with the lowest were Finland (207/801 = 25.8 %), Ireland (335/600 = 55.8 %), and Germany (2,055/3,560 = 57.7%). These differences could reflect geographical differences in variant frequency. Yet they may also be a consequence of the bias of gnomAD reference data used to define URV status. Historical gnomAD data emanates predominantly from North American and other populations with Western European ancestry, while efforts are increasing to improve representation of other ancestral groups after the past 15 years focusing on increasing the representation of people with Finnish ancestry because of their genetic bottleneck (Lim *et al.*, 2014; Lek *et al.*, 2016).

5.5.4 Phenotypic similarity of the cohort

I calculated the overall clinical similarity between each of the 79,512,355 possible pairwise combinations of participants according to $\mathrm{Sim_{MinTO}}$ and the secondary similarity measures. The median $\mathrm{Sim_{MinTO}}$ among the 79,512,355 pairs in this cohort was 1.89 and the distribution was positively skewed (mean = 3.11, mode = 0.350, range = 0.00–120, interquartile range = 0.687–3.85, the unit scaled distribution is shown in **Figure 4-4**).

I found high correlation (Spearman's $\rho \geq 0.935$) among $\mathrm{Sim_{MinTO}}$, $\mathrm{Sim_{Phrank}}$, $\mathrm{Sim_{RBP}}$, and $\mathrm{Sim_{Resnik\text{-}BMS}}$, and $\mathrm{Sim_{TO}}$, and closely related to these were $\mathrm{Sim_{Resnik\text{-}Max}}$ ($\rho \geq 0.921$), $\mathrm{Sim_{Rel\text{-}BMS}}$ ($\rho \geq 0.899$), and $\mathrm{Sim_{TA}}$ ($\rho \geq 0.894$, **Figure 4-6** and **Figure 4-7**). A second group consisting of unweighted cosine and Jaccard and IC-weighted measures were closely correlated as pairs ($\rho = 0.984$ for $\mathrm{Sim_{Cosine}}$ and $\mathrm{Sim_{UI}}$, and $\rho = 0.931$ for $\mathrm{Sim_{CosIC}}$ and $\mathrm{Sim_{GIC}}$) and moderately well correlated as a group ($\rho \geq 0.733$) but appear separate from the first group (greatest $\rho = 0.810$ for $\mathrm{Sim_{CosIC}}$ with $\mathrm{Sim_{Rel\text{-}BMS}}$, and least $\rho = 0.426$ for $\mathrm{Sim_{Cosine}}$ with $\mathrm{Sim_{Resnik\text{-}BMS}}$). Best match average measures ($\mathrm{Sim_{Rel\text{-}BMA}}$, $\mathrm{Sim_{Resnik\text{-}BMA}}$) correlated well with each other ($\rho = 0.953$) and appear to sit between the two former groups. The greatest correlation with the Cosine/Jaccard group was $\mathrm{Sim_{Rel\text{-}BMA}}$ with $\mathrm{Sim_{CosIC}}$ ($\rho = 0.909$) and least was $\mathrm{Sim_{Resnik\text{-}BMA}}$ with $\mathrm{Sim_{UI}}$ ($\rho = 0.582$). $\mathrm{Sim_{Goodall3}}$ appears independent of other measures ($\rho \leq 0.791$), but closer to the cosine/Jaccard group ($0.677 \leq \rho \leq 0.791$), than the best match average ($\rho = 0.633$ with $\mathrm{Sim_{Rel\text{-}BMA}}$ and $\rho = 0.632$ with $\mathrm{Sim_{Resnik\text{-}BMA}}$), and the first group ($0.416 \leq \rho \leq 0.532$).

5.5.5 Phenotypic similarity of dominant epilepsy genes

For dominant epilepsy genes with multiple carriers of PAVs (similarity requires a pair), I calculated empirical one-sided *p*-values for observing median $\mathrm{Sim}_{\mathrm{MinTO}}$ similarities at least this great within the cohort. Of 258 dominant epilepsy genes with multiple carriers of PAVs, 33 (12.8 %) had $\mathrm{Sim}_{\mathrm{MinTO}}$ $p < 0.05$, that is median pairwise similarities in the top 5 % of those expected in the cohort (**Figure 5-8**, **Supplementary Table 5-6**, and **Supplementary Visualizations**), most remarkably *SLC6A1* (Holm's-adjusted $p = 5.16 \times 10^{-4}$), *RORA* (Holm's-adjusted $p = 0.00514$), *SCN1A* (Holm's-adjusted $p = 0.00794$), and *STX1B* (Holm's-adjusted $p = 0.0500$). In contrast, 12/261 (4.60 %) groups consisting of those with synonymous, presumed biologically neutral, variants in the same gene had $p < 0.05$.

I found a positive correlation between the predicted deleteriousness of the gene-collapsed variant classes and the proportion of dominantly epilepsy genes reaching $p < 0.05$ (PTV > dMissense > Missense > bMissense > Synonymous, **Table 5-9**). I found similar results across the similarity measures that correlate closely with $\mathrm{Sim}_{\mathrm{MinTO}}$, as well as with $\mathrm{Sim}_{\mathrm{Goodall3}}$ (**Figure 5-9** and **Supplementary Visualizations**), but less remarkable results when using the best match average, cosine, and Jaccard measures. This gradient according to the predicted severity of variants was reflected not only by the number of genes reaching the arbitrary threshold of $p < 0.05$ but also the point and gradient at which the distribution of observed *p*-values deflects from the simulated null distribution, with earlier and steeper takeoff for PTVs than other PAVs (**Figure 5-8** and **Figure 5-9**).

Similarity measure	Number (%) of genes reaching similarity $p < 0.05$ by ultra-rare variant class						
	PAV	dPAV	PTV	dMissense	Missense	bMissense	Synonymous
$\mathrm{Sim}_{\mathrm{MinTO}}$	33/258 (12.8 %)	29/254 (11.4 %)	30/167 (18.0 %)	24/254 (9.45 %)	24/257 (9.34 %)	9/158 (5.70 %)	12/261 (4.60 %)
$\mathrm{Sim}_{\mathrm{Phrank}}$	30/258 (11.6 %)	28/254 (11.0 %)	30/167 (18.0 %)	23/254 (9.06 %)	22/257 (8.56 %)	11/158 (6.96 %)	9/261 (3.45 %)
$\mathrm{Sim}_{\mathrm{Rel\text{-}BMS}}$	28/258 (10.9 %)	28/254 (11.0 %)	28/167 (16.8 %)	18/254 (7.09 %)	21/257 (8.17 %)	7/158 (4.43 %)	17/261 (6.51 %)
$\mathrm{Sim}_{\mathrm{Resnik\text{-}Max}}$	27/258 (10.5 %)	34/254 (13.4 %)	31/167 (18.6 %)	24/254 (9.45 %)	18/257 (7.00 %)	9/158 (5.70 %)	7/261 (2.68 %)
$\mathrm{Sim}_{\mathrm{Goodall3}}$	26/258 (10.1 %)	31/254 (12.2 %)	23/167 (13.8 %)	27/254 (10.6 %)	23/257 (8.95 %)	13/158 (8.23 %)	10/261 (3.83 %)
$\mathrm{Sim}_{\mathrm{Resnik\text{-}BMS}}$	26/258 (10.1 %)	26/254 (10.2 %)	30/167 (18.0 %)	21/254 (8.27 %)	21/257 (8.17 %)	8/158 (5.06 %)	12/261 (4.60 %)
$\mathrm{Sim}_{\mathrm{RBP}}$	25/258 (9.69 %)	26/254 (10.2 %)	29/167 (17.4 %)	21/254 (8.27 %)	14/257 (5.45 %)	10/158 (6.33 %)	15/261 (5.75 %)
$\mathrm{Sim}_{\mathrm{TO}}$	25/258 (9.69 %)	24/254 (9.45 %)	28/167 (16.8 %)	20/254 (7.87 %)	16/257 (6.23 %)	12/158 (7.59 %)	16/261 (6.13 %)
$\mathrm{Sim}_{\mathrm{TA}}$	22/258 (8.53 %)	27/254 (10.6 %)	29/167 (17.4 %)	22/254 (8.66 %)	19/257 (7.39 %)	11/158 (6.96 %)	11/261 (4.21 %)
$\mathrm{Sim}_{\mathrm{Rel\text{-}BMA}}$	20/258 (7.75 %)	20/254 (7.87 %)	20/167 (12.0 %)	17/254 (6.69 %)	10/257 (3.89 %)	8/158 (5.06 %)	24/261 (9.20 %)
$\mathrm{Sim}_{\mathrm{CosIC}}$	15/258 (5.81 %)	18/254 (7.09 %)	19/167 (11.4 %)	13/254 (5.12 %)	12/257 (4.67 %)	7/158 (4.43 %)	8/261 (3.07 %)
$\mathrm{Sim}_{\mathrm{Resnik\text{-}BMA}}$	13/258 (5.04 %)	19/254 (7.48 %)	20/167 (12.0 %)	16/254 (6.30 %)	12/257 (4.67 %)	4/158 (2.53 %)	7/261 (2.68 %)
$\mathrm{Sim}_{\mathrm{UI}}$	9/258 (3.49 %)	8/254 (3.15 %)	13/167 (7.78 %)	10/254 (3.94 %)	9/257 (3.50 %)	9/158 (5.70 %)	9/261 (3.45 %)
$\mathrm{Sim}_{\mathrm{Cosine}}$	8/258 (3.10 %)	8/254 (3.15 %)	16/167 (9.58 %)	11/254 (4.33 %)	8/257 (3.11 %)	8/158 (5.06 %)	8/261 (3.07 %)
$\mathrm{Sim}_{\mathrm{GIC}}$	8/258 (3.10 %)	11/254 (4.33 %)	23/167 (13.8 %)	11/254 (4.33 %)	11/257 (4.28 %)	9/158 (5.70 %)	11/261 (4.21 %)

Table 5-9. The number of dominant epilepsy genes achieving phenotypic similarity $p < 0.05$ according to variant class and similarity measure.

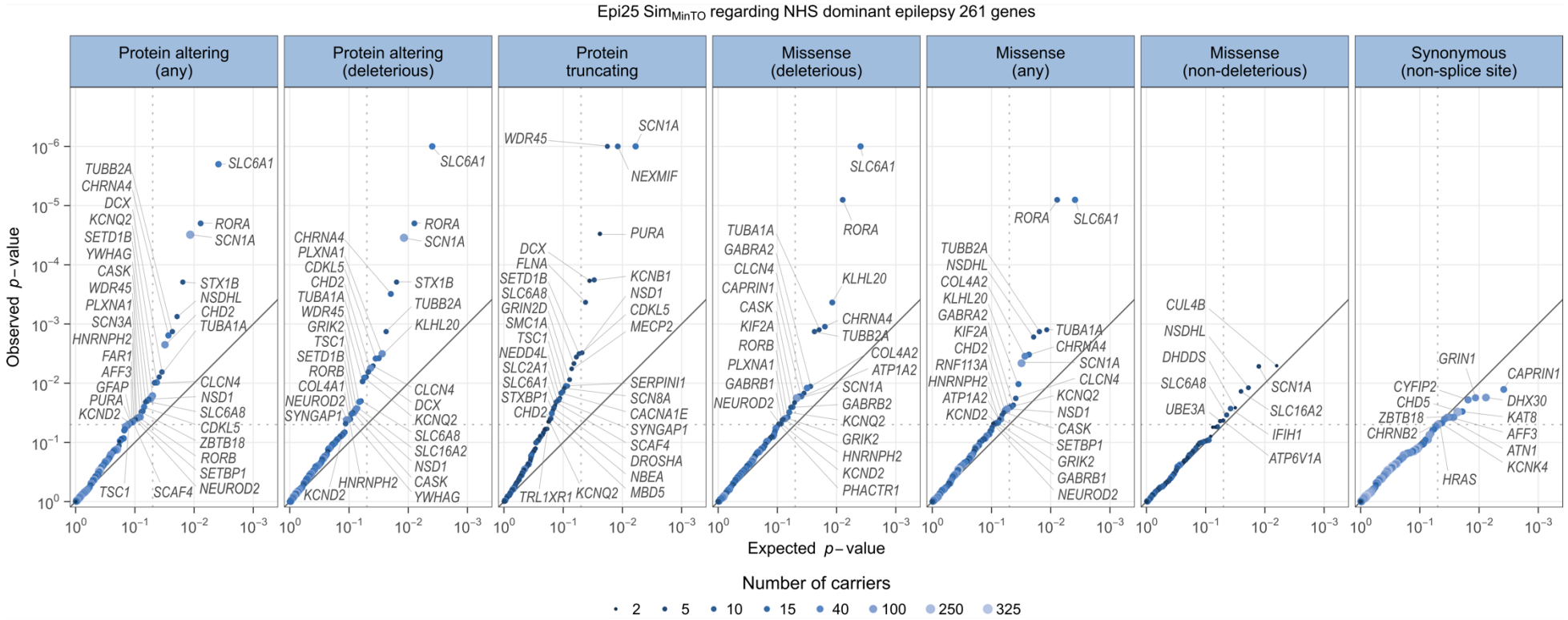


Figure 5-8. Quantile-quantile plots of the similarity of carriers of URVs in dominant epilepsy genes according to $\mathrm{Sim}_{\mathrm{MinTO}}$. Both axes are negative log scaled.

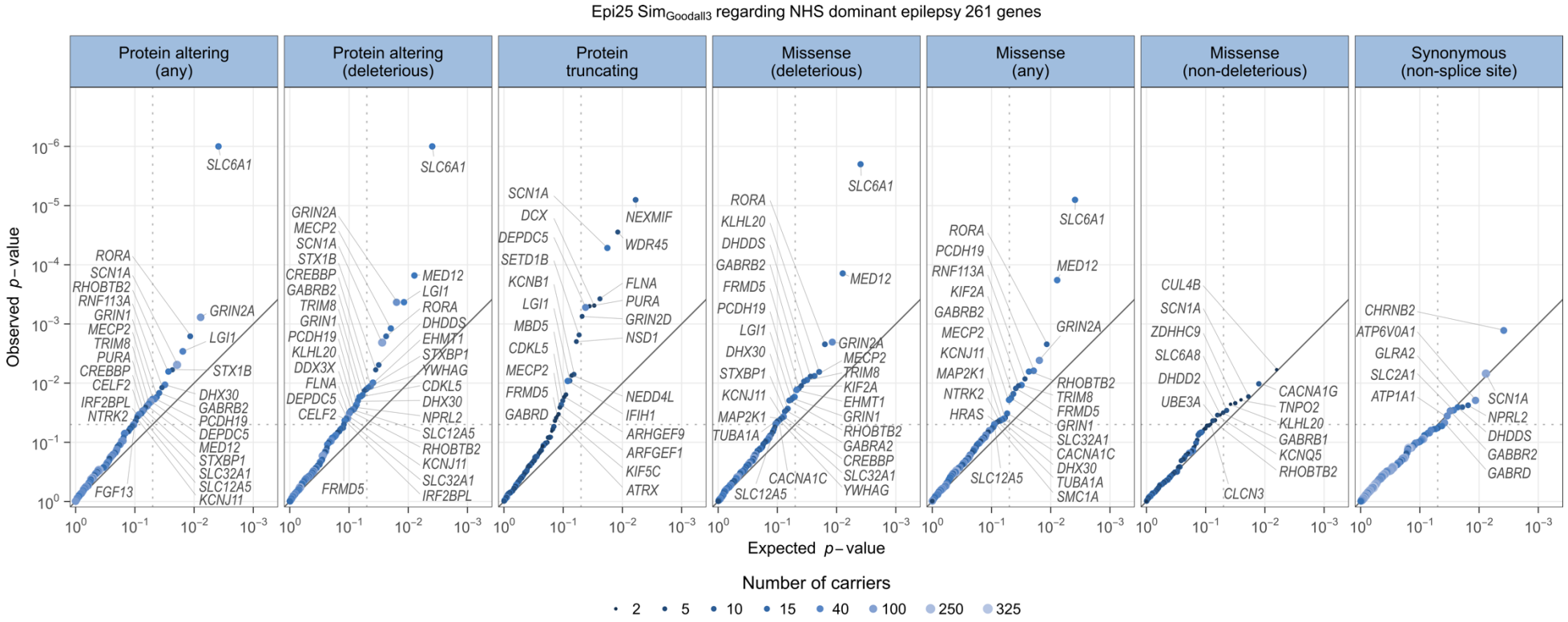


Figure 5-9. Quantile-quantile plots of the similarity of carriers of URVs in dominant epilepsy genes according to $\mathrm{Sim}_{\mathrm{Goodall3}}$.

$\mathrm{Sim}_{\mathrm{MinTO}}$ and $\mathrm{Sim}_{\mathrm{Goodall3}}$ provide complementary open and closed-world perspectives of the data. One would expect that those groups that are similar by $\mathrm{Sim}_{\mathrm{Goodall3}}$ but not $\mathrm{Sim}_{\mathrm{MinTO}}$ owe their similarity to the shared lack of annotations of phenotypes that are common in the remainder of the cohort whereas those that are similar according to $\mathrm{Sim}_{\mathrm{MinTO}}$ are likely to share informative phenotypic annotations. Groups of PAV carriers in five genes attained $p < 0.05$ for both measures, 28 for $\mathrm{Sim}_{\mathrm{MinTO}}$ only, and 21 for $\mathrm{Sim}_{\mathrm{Goodall3}}$ only (**Supplementary Table 5-7**). Similar proportions of genes were concordant for classes of PAV, except for PTV for which greater concordance was seen with 14 genes reached $p < 0.05$ for both, 16 for $\mathrm{Sim}_{\mathrm{MinTO}}$ only, and nine for $\mathrm{Sim}_{\mathrm{Goodall3}}$ only.

Some genes had a dPAV group equal to the union of the PTV and dMissense groups (there were no carriers of undifferentiable dPAVs, see **Table 5-4**). Among these, some had a more remarkable *p*-value for dPAV than for both PTV and dMissense groups: *CHRNA4*, *CLCN4*, *GRIK2*, and *RORB*. This indicates that carriers of dMissense variants have similar clinical features to carriers of PTVs in the same gene and one can hypothesize that the dMissense variants may have similar biological consequences (presumably loss of function) to the PTVs in the same gene. Alternatively, for some genes with dPAV carriers of only distinguished PTV and dMissense subclasses this was not the case: *DCX*, *SLC6A8*, *SETD1B*, and *CDKL5*. In these genes, we lack evidence from phenotypic similarity to suggest that a similar biological consequence of the PTVs and dMissense variants.

5.5.6 Phenotypic similarity beyond dominant epilepsy genes

I extended the similarity analyses to genes beyond the primary dominant epilepsy set (**Figure 5-10**, **Figure 5-11**, **Supplementary Table 5-6**, and **Supplementary Visualizations**). Focusing on $\mathrm{Sim}_{\mathrm{MinTO}}$ of PAV groups, *LZTR1* (Autosomal dominant Noonan syndrome 10 [OMIM:616564], Autosomal recessive Noonan syndrome 2 [OMIM:605275], and Susceptibility to Schwannomatosis-2 [OMIM:615670]), *PCGF2* (Turnpenny-Fry syndrome [OMIM:618371]), and *KDM5C* (X-linked syndromic intellectual developmental disorder, Claes-Jensen type [OMIM:300534]), were the top ranking genes, and all recognized GE Intellectual Disability 'green' genes. Beyond the genes validated for clinical testing of mendelian neurodevelopmental disorders, carriers of PAVs in *STRAP* (serine/threonine kinase receptor associated protein, not associated with mendelian disease), *MYRF* (reported in two Japanese families with Mild encephalitis/encephalopathy with reversible myelin vacuolization [OMIM:618113], an autosomal dominant childhood-onset relapsing/remitting fever-sensitive

leukoencephalopathy with seizures, and Cardiac-urogenital syndrome [OMIM:618280]), and *TEAD2* (Transcriptional enhancer factor 4, not associated with mendelian disease) were the top ranking genes not to feature in the clinical testing panels. $\mathrm{Sim}_{\mathrm{Goodall3}}$ ranked *LRRC7* (Leucine-rich repeat-containing protein 7, a preferentially brain-expressed scaffold protein not associated with mendelian disease) top among the PAV groups.

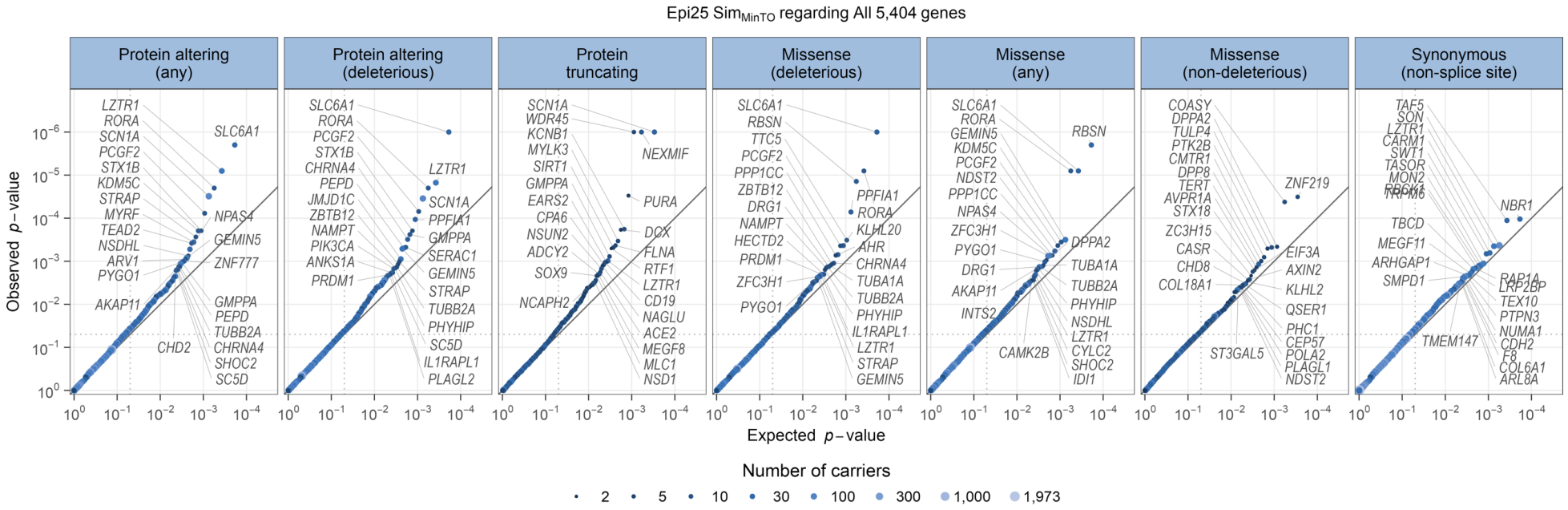


Figure 5-10. Quantile-quantile plots of the similarity of carriers of URVs in the global set of genes of the study according to $\mathrm{Sim}_{\mathrm{MinTO}}$.

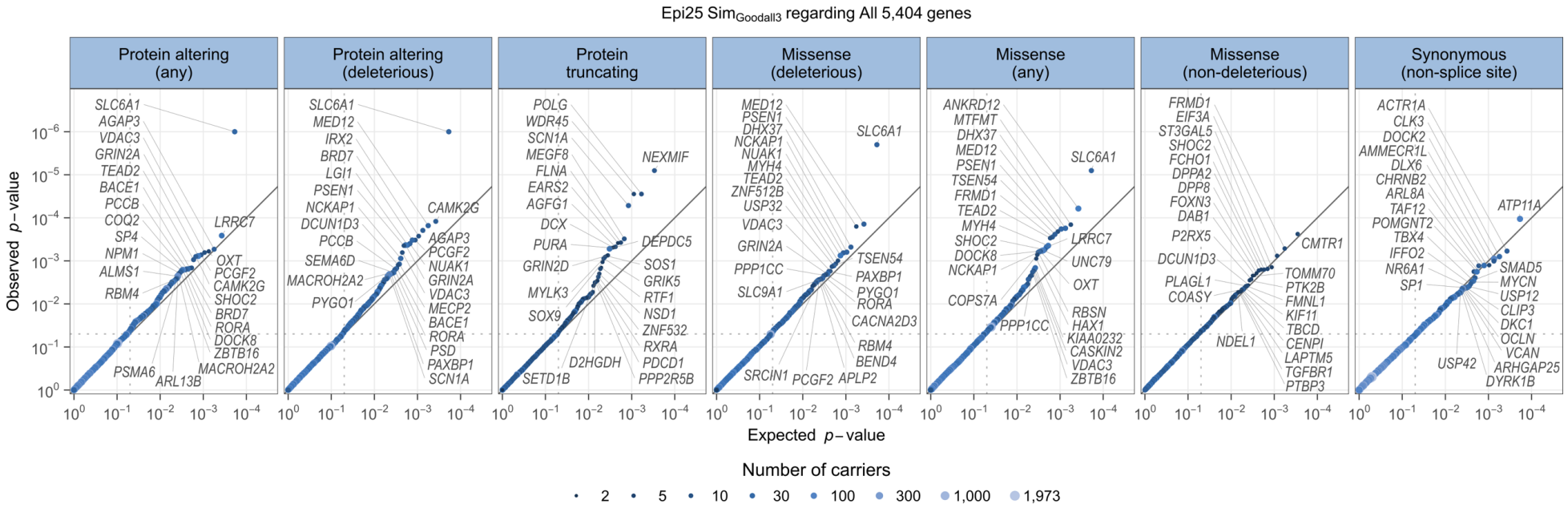


Figure 5-11. Quantile-quantile plots of the similarity of carriers of URVs in the global set of genes of the study according to $\mathrm{Sim}_{\mathrm{Goodall3}}$.

5.5.7 Phenotypes associated with dominant epilepsy genes

I calculated the prevalence of each phenotypic annotation among carriers and non-carriers of URVs of each class in each gene, calculating their association using the two-sided FET, and performed the *minP* single step procedure within each genetically defined group level to assist in interpretation of the large number of dependent tests. I shall focus on the statistically interesting associations of the primary gene set of dominant epilepsy genes before summarizing those of non-dominant epilepsy genes from the Genomics England panel, and those genes validated for neurodevelopmental disorders but not the Genomics England panel, then explore the clinically unvalidated genes. Given the large number of results, I shall present data for genetically defined group in a series of tables listing the minimal set of the phenotypes that reach $minP < 0.05$, that is those that are robust to multiple testing and that represent clinically independent phenotypic concepts. For context, the tables also summarize the number of terms associated with $p < 0.05$, how many of these are clinically independent (the greater the number conceptually independent terms, the broader the phenotypic spectrum), and for convenience, the $\mathrm{Sim}_{\mathrm{MinTO}}$ and $\mathrm{Sim}_{\mathrm{Goodall3}}$ as open and closed-world similarity measures of the complete clinical constellations.

Additionally, for selected examples, I shall illustrate the repertoire of associated terms statistically, and conceptually. On the polar plots (panel one of each se), the percentage of carriers of the URV with each broad phenotype is shown by the height of bars arranged on a polar axis, with two-sided 95 % CI error bars for these percentages derived from the binomial distribution. The dots indicate the percentage of the entire cohort of 12,611 people with each broad phenotype and are shown for reference only (representing the prior probability of a person having the phenotype without genetic stratification). The corresponding 95 % CI of the complete phenomic cohort are no wider than 1.9 % (and would not extend beyond the dot so are omitted for clarity). 'Craniofacial or skeletal dysmorphism' includes individuals with either *Abnormality of the head [HP:0000234]* (which excludes isolated brain structural abnormalities) or *Abnormal skeletal morphology [HP:0011842]*. 'Motor, movement or muscular disorder' includes individuals with any of *Abnormal central motor function [HP:0011442]*, *Abnormality of movement [HP:0100022]*, or *Abnormality of the musculature [HP:0003011]*, but not *Motor delay [HP:0001270]*, which is included in 'Neurodevelopmental abnormality'. While 'Neurodevelopmental abnormality' includes those with 'Intellectual disability', the latter is shown additionally as it is a neurodevelopmental outcome with particularly important socioeconomically important consequences.

The statistical plots show the proportion of carriers and non-carriers with each phenotype with log(1+x) and log(1+y) scaling of axes to optimize visualization of the majority of associated phenotypes which affect < 50 % of the cohort.

The conceptual plots summarize the HPO to show only the most and least conceptually specific terms associated with $p < 0.05$ and annotated to multiple or no carriers of the specified variant (censoring those associations attributable to single carriers). The comprehensive set of association tests reaching unadjusted $p < 0.05$ is provided in **Supplementary Table 5-8**.

257 of 260 (98.8 %) dominant epilepsy genes in which at least one person carried a PAV had at least one association reaching $p < 0.05$ when considering carriers of any ultra-rare PAV. Thirty-two (12.3 %) had associations reaching *minP* < 0.05, robust to the 1,734 phenotypic tests undertaken per genetic group. 169 of 211 (80.1 %) genes harboring by PTVs had an association reaching $p < 0.05$ and in 22 (10.4 %), an association reached *minP* < 0.05.

Despite a known genetic diagnosis being an exclusion criterion for Epi25 and most contributing sites being early adopters of genetic testing, carriers of variants in many clinically validated epilepsy genes tended to exhibit their classical clinical features (PAVs: **Table 5-10**, Figure 5-12, Figure 5-13, Figure 5-14, Figure 5-15, Figure 5-16, Figure 5-17, Figure 5-18, **Figure 5-19**, Figure 5-20, Figure 5-21; PTVs: **Table 5-11**, Figure 5-22, Figure 5-23, **Figure 5-25**, **Figure 5-26**, **Figure 5-27**; dPAVs: **Table 5-12**; dMissense: **Table 5-13**; Missense: **Table 5-14**).

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
SCN1A PAV	116	89 (37)	13 (5)	3.10×10^{-5}	0.00491	Focal hemiclonic seizure [HP:0006813]	8/116 (6.90 %)	13.9	3.65×10^{-7}	7.00×10^{-5}	13.1	0.108
						Epileptic encephalopathy [HP:0200134]	35/116 (30.2 %)	3.03	6.05×10^{-7}	9.00×10^{-5}	2.42	0.0220
						Intellectual disability [HP:0001249]	34/116 (29.3 %)	2.93	1.35×10^{-6}	2.30×10^{-4}	2.37	0.0215
						Febrile seizure (within the age range of 3 months to 6 years) [HP:0002373]	19/116 (16.4 %)	2.93	1.21×10^{-4}	0.0321	2.62	0.0237
						Developmental regression [HP:0002376]	17/116 (14.7 %)	3.00	1.93×10^{-4}	0.0430	2.71	0.0245
IRF2BPL PAV	32	37 (18)	11 (5)	0.611	0.0362	Large for gestational age [HP:0001520]	4/32 (12.5 %)	34.3	1.14×10^{-5}	0.00226	30.2	0.0714
						Cleft lip [HP:0410030]	2/32 (6.25 %)	829	1.87×10^{-5}	0.00499	786.	0.667
						Epileptic spasm [HP:0011097]	9/32 (28.1 %)	6.29	6.90×10^{-5}	0.0147	4.80	0.0121
						Cleft palate [HP:0000175]	2/32 (6.25 %)	166	1.30×10^{-4}	0.0234	157	0.286
						Epileptic encephalopathy [HP:0200134]	12/32 (37.5 %)	4.17	3.14×10^{-4}	0.0445	2.98	0.00753
MECP2 PAV	32	81 (39)	6 (4)	0.602	0.0186	Neurodevelopmental delay [HP:0012758]	10/32 (31.2 %)	7.17	1.14×10^{-5}	0.00225	5.24	0.0132
						Epileptic encephalopathy [HP:0200134]	13/32 (40.6 %)	4.76	6.74×10^{-5}	0.0144	3.23	0.00816
						Decreased head circumference [HP:0040195]	5/32 (15.6 %)	10.7	2.03×10^{-4}	0.0299	9.23	0.0229
						Intellectual disability [HP:0001249]	12/32 (37.5 %)	4.21	2.93×10^{-4}	0.0423	3.00	0.00759

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
RORA PAV (dPAV)	14	38 (15)	6 (4)	2.00×10^{-5}	0.00162	Developmental regression [HP:0002376]	6/14 (42.9 %)	13.0	5.53×10^{-5}	0.00677	7.86	0.00866
						Atypical absence seizure [HP:0007270]	4/14 (28.6 %)	23.7	6.93×10^{-5}	0.00824	17.2	0.0188
						EEG with generalized epileptiform discharges [HP:0011198]	12/14 (85.7 %)	10.6	1.96×10^{-4}	0.0220	2.38	0.00264
						Focal-onset seizure [HP:0007359]	*2/14 (14.3 %)*	*0.101*	*3.34×10^{-4}*	*0.0312*	*0.230*	*0.00251*
STX1B PAV	7	48 (25)	5 (3)	1.96×10^{-4}	0.00596	Intellectual disability, severe s[HP:0010864]	3/7 (42.9 %)	33.1	3.62×10^{-4}	0.0222	19.4	0.0106
						Epileptic encephalopathy [HP:0200134]	5/7 (71.4 %)	17.3	5.40×10^{-4}	0.0306	5.67	0.00314
						Generalized atonic seizure [HP:0032887]	3/7 (42.9 %)	28.5	5.55×10^{-4}	0.0313	16.7	0.00920
SLC6A1 PAV	27	52 (23)	8 (3)	2.00×10^{-6}	1.00×10^{-6}	*Focal impaired awareness seizure [HP:0002384]*	*1/27 (3.70 %)*	*0.0708*	*1.71×10^{-4}*	*0.0247*	*0.105*	*0.00318*
(dPAV)						EEG with spike-wave complexes [HP:0010850]	17/27 (63.0 %)	6.74	1.56×10^{-6}	1.90×10^{-4}	3.13	0.00667
						Early onset absence seizures [HP:0011152]	5/27 (18.5 %)	31.1	1.63×10^{-6}	2.10×10^{-4}	25.6	0.0521
PURA PAV	6	58 (19)	8 (3)	0.0476	0.0255	Neonatal hypotonia [HP:0001319]	2/6 (33.3 %)	477	1.98×10^{-5}	0.00229	323	0.133
	6	58 (19)	8 (3)	0.0476	0.0255	Developmental regression [HP:0002376]	4/6 (66.7 %)	34.6	1.24×10^{-4}	0.00972	12.2	0.00577
						Neurodevelopmental delay [HP:0012758]	4/6 (66.7 %)	31.3	1.78×10^{-4}	0.0119	11.1	0.00526

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
TUBA1A PAV	7	17 (7)	3 (3)	0.00786	0.145	Abnormal cortical gyration [HP:0002536]	3/7 (42.9 %)	42.1	1.81×10^{-4}	0.0133	24.6	0.0135
						Multifocal epileptiform discharges [HP:0010841]	4/7 (57.1 %)	19.8	4.82×10^{-4}	0.0283	9.05	0.00500
						Intellectual disability [HP:0001249]	5/7 (71.4 %)	17.5	5.21×10^{-4}	0.0300	5.71	0.00316
CHD2 PAV	65	108 (46)	2 (2)	0.00224	0.589	EEG with spike-wave complexes [HP:0010850]	28/65 (43.1 %)	3.01	3.03×10^{-5}	0.00846	2.14	0.0110
						Abnormal skeletal morphology [HP:0011842]	9/65 (13.8 %)	4.97	2.04×10^{-4}	0.0439	4.42	0.0224
SCN2A PAV	60	65 (26)	2 (2)	0.649	0.178	Neonatal seizure [HP:0032807]	7/60 (11.7 %)	8.28	4.89×10^{-5}	0.00934	7.43	0.0343
						EEG with burst suppression [HP:0010851]	4/60 (6.67 %)	15.6	1.97×10^{-4}	0.0405	14.7	0.0656
KCNH1 PAV	32	71 (21)	4 (2)	0.205	0.681	Seizure precipitated by febrile infection [HP:0032894]	12/32 (37.5 %)	5.00	6.77×10^{-5}	0.0144	3.50	0.00882
						Downslanted palpebral fissures [HP:0000494]	2/32 (6.25 %)	166	1.30×10^{-4}	0.0234	157	0.286
WDR45 PAV	13	56 (19)	2 (2)	0.0296	0.884	Diffuse cerebral atrophy [HP:0002506]	3/13 (23.1 %)	139	3.42×10^{-6}	7.90×10^{-4}	108	0.100
						Global developmental delay [HP:0001263]	5/13 (38.5 %)	14.3	1.28×10^{-4}	0.0142	9.18	0.00938
KCNT1 PAV	59	50 (18)	11 (2)	0.649	0.538	Cleft lip [HP:0410030]	3/59 (5.08 %)	Inf	9.73×10^{-8}	7.00×10^{-5}	Inf	1.00
						Cleft palate [HP:0000175]	3/59 (5.08 %)	167	3.36×10^{-6}	5.50×10^{-4}	160	0.429
KCNA2 PAV	21	42 (16)	3 (2)	0.172	0.940	Convulsive status epilepticus [HP:0032660]	4/21 (19.0 %)	20.7	8.62×10^{-5}	0.0121	17.0	0.0276
						Intellectual disability, mild [HP:0001256]	5/21 (23.8 %)	9.57	4.33×10^{-4}	0.0483	7.53	0.0124

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
CDKL5 PAV	23	33 (15)	2 (2)	0.0251	0.0943	Epileptic spasm [HP:0011097]	7/23 (30.4 %)	7.01	2.62×10^{-4}	0.0334	5.18	0.00938
						Neurodevelopmental delay [HP:0012758]	7/23 (30.4 %)	6.87	2.94×10^{-4}	0.0373	5.09	0.00921
YWHAG PAV	11	33 (14)	3 (2)	0.0187	0.210	Neonatal electro-clinical myoclonic seizure [HP:0032815]	2/11 (18.2 %)	538	1.45×10^{-5}	0.00214	458	0.286
						Intellectual disability, mild [HP:0001256]	4/11 (36.4 %)	17.5	2.84×10^{-4}	0.0277	11.5	0.00993
STXBP1 PAV	19	20 (12)	2 (2)	0.377	0.0281	Epileptic spasm [HP:0011097]	7/19 (36.8 %)	9.35	6.63×10^{-5}	0.00984	6.28	0.00938
						Neonatal seizure [HP:0032807]	4/19 (21.1 %)	16.5	2.13×10^{-4}	0.0250	13.3	0.0196
KCNT2 PAV	38	27 (9)	2 (2)	0.837	0.797	Benign neoplasm of the central nervous system [HP:0100835]	6/38 (15.8 %)	9.02	1.20×10^{-4}	0.0237	7.75	0.0229
						Intracranial epidermoid cyst [HP:0012096]	2/38 (5.26 %)	139	1.84×10^{-4}	0.0345	132	0.286
A*LG13* PAV	21	38 (21)	1 (1)	0.575	0.566	EEG with focal spike waves [HP:0011197]	3/21 (14.3 %)	99.1	7.88×10^{-6}	0.00114	85.6	0.125
NTRK2 PAV	15	56 (21)	1 (1)	0.386	0.0436	Feeding difficulties in infancy [HP:0008872]	2/15 (13.3 %)	76.9	4.56×10^{-4}	0.0437	67.2	0.0741
*GRIN2*A PAV (dPAV)	73	40 (19)	1 (1)	0.664	7.74×10^{-4}	EEG with central epileptiform discharges [HP:0033718]	15/73 (20.5 %)	7.06	4.21×10^{-8}	1.00×10^{-5}	5.82	0.0328
CHRNA4 PAV	23	42 (18)	1 (1)	0.00156	0.289	Motor seizure [HP:0020219]	16/23 (69.6 %)	4.76	4.15×10^{-4}	0.0496	2.14	0.00390
KCNQ2 PAV	29	33 (18)	1 (1)	0.00970	0.530	EEG with burst suppression [HP:0010851]	3/29 (10.3 %)	24.9	3.60×10^{-4}	0.0484	22.4	0.0492
NSDHL PAV	8	51 (17)	3 (1)	7.49×10^{-4}	0.559	Myoclonus [HP:0001336]	2/8 (25.0 %)	149	1.52×10^{-4}	0.0151	113	0.0667

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
SLC2A1 PAV	25	30 (13)	2 (1)	0.236	0.311	EEG with spike-wave complexes (<2.5 Hz) [HP:0010847]	5/25 (20.0 %)	13.4	8.61×10^{-5}	0.0150	10.9	0.0212
LGI1 PAV	35	24 (12)	1 (1)	0.199	0.00290	Bilateral tonic-clonic seizure [HP:0002069]	29/35 (82.9 %)	4.39	2.58×10^{-4}	0.0387	1.58	0.00438
PA*K1* PAV	19	37 (12)	1 (1)	0.747	0.0615	Abnormality of the musculoskeletal system [HP:0033127]	6/19 (31.6 %)	7.76	4.60×10^{-4}	0.0488	5.62	0.00842
DCX PAV (dPAV)	10	19 (11)	6 (1)	0.00648	0.578	Subcortical band heterotopia [HP:0032409]	5/10 (50.0 %)	585	6.20×10^{-12}	1.00×10^{-5}	300	0.192
CASK PAV	7	30 (10)	1 (1)	0.0208	0.602	Epileptic spasm [HP:0011097]	4/7 (57.1 %)	21.3	3.68×10^{-4}	0.0225	9.71	0.00536
FLNA PAV	58	22 (8)	2 (1)	0.136	0.0504	Periventricular heterotopia [HP:0007165]	5/58 (8.62 %)	10.0	2.37×10^{-4}	0.0493	9.25	0.0410
TUBB2B PAV (dPAV)	6	23 (8)	1 (1)	0.812	0.846	Lissencephaly [HP:0001339]	2/6 (33.3 %)	115	2.87 x 10-4	0.0170	77.8	0.0357
SCN1B PAV	12	6 (3)	1 (1)	0.250	0.151	Febrile seizure (within the age range of 3 months to 6 years) [HP:0002373]	5/12 (41.7 %)	10.6	5.55 x 10-4	0.0463	6.59	0.00624

Table 5-10. A summary of the associations of carriers of PAVs in dominant epilepsy genes. Those italicized rows pertain to inverse associations (the phenotype is less common in the genetically specified group than the remainder of the cohort) and their predictive values indicate the proportion of the cohort without (rather than with) the phenotype that are members of this genetic group. Commenced 4 pages previously.

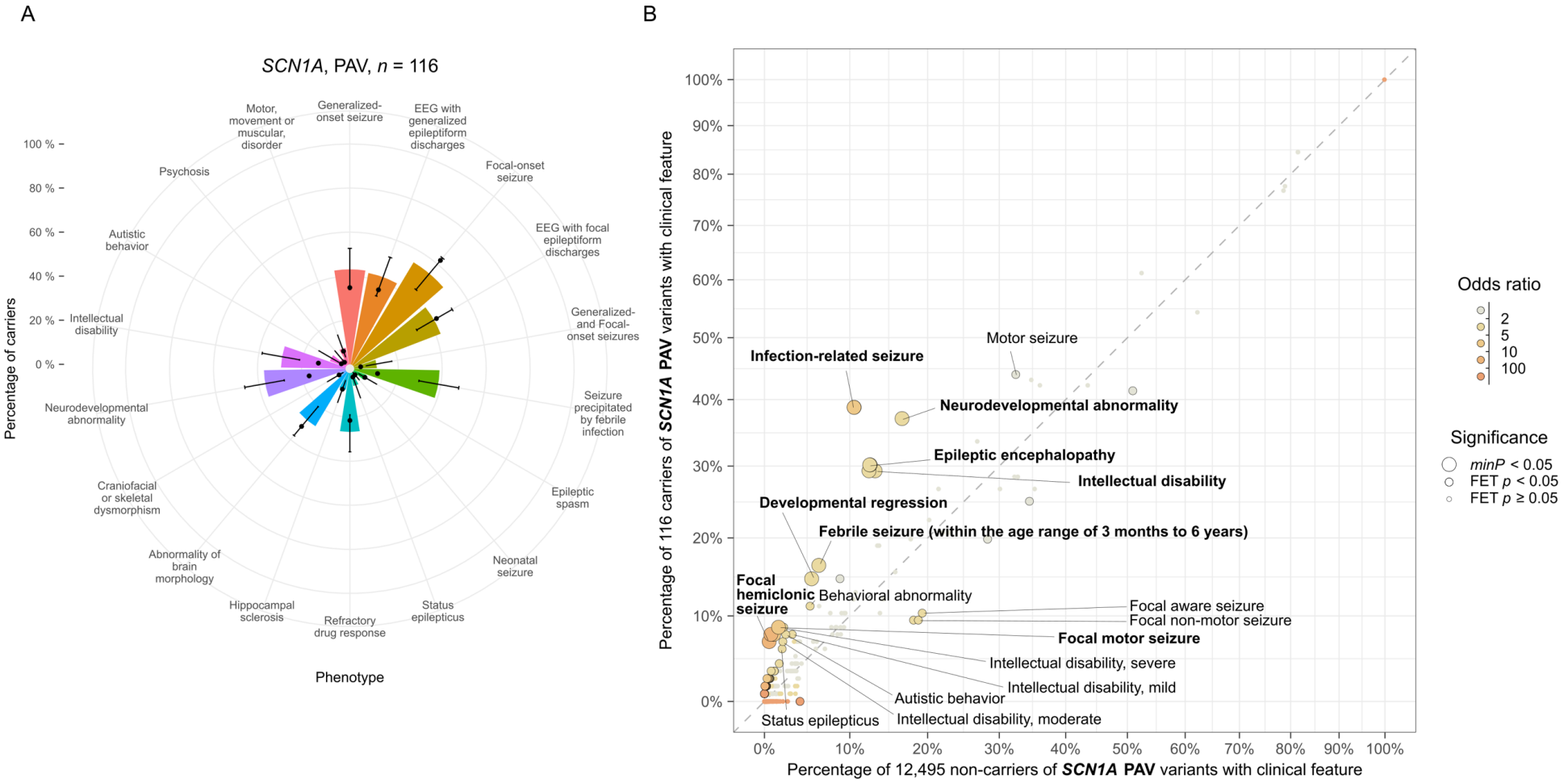


Figure 5-12. The associations of carriers of PAVs in *SCN1A*.

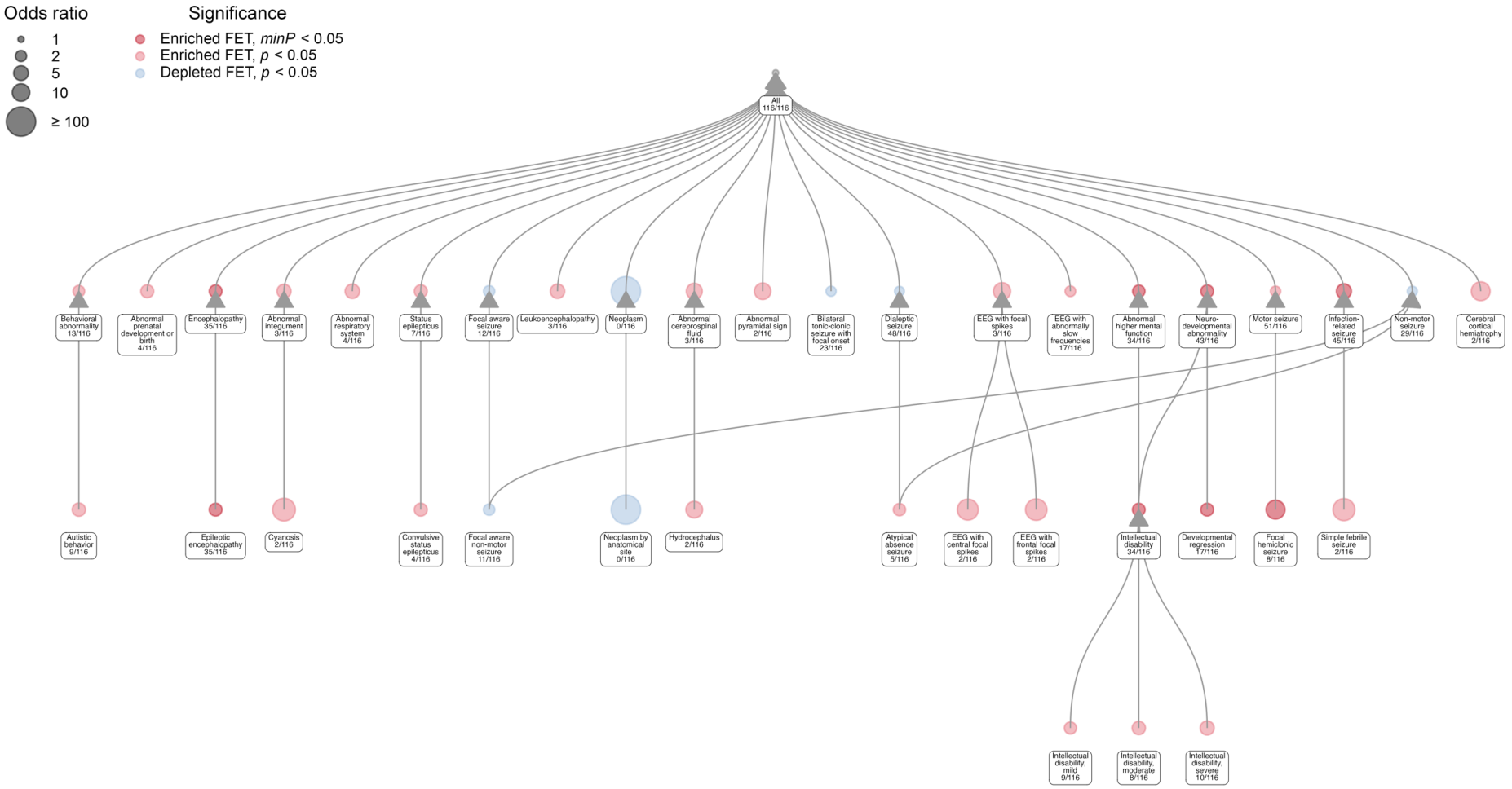


Figure 5-13. The associations of carriers of PAVs in *SCN1A* shown conceptually.

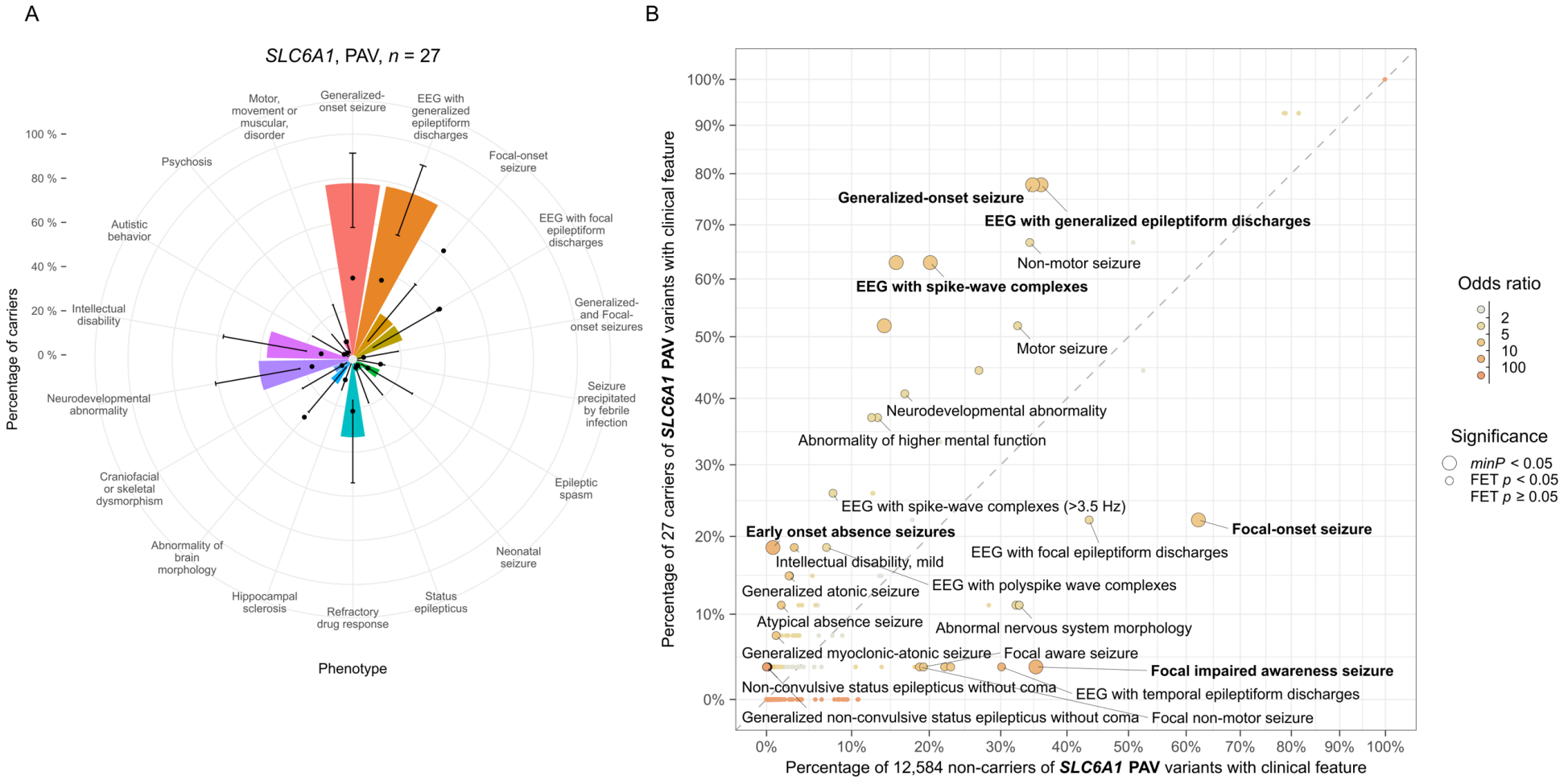


Figure 5-14. The associations of carriers of PAVs in *SLC6A1*.

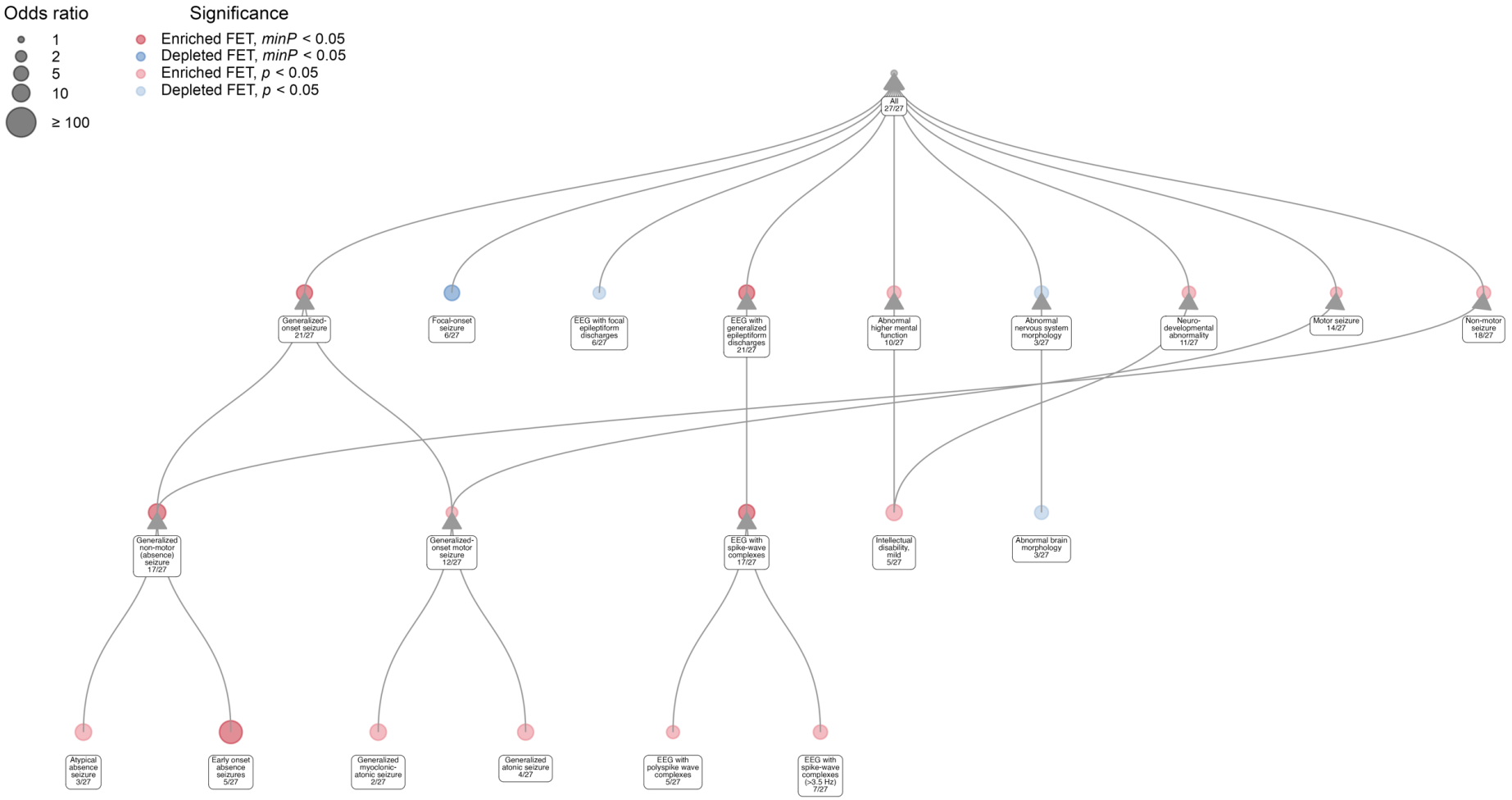


Figure 5-15. The associations of carriers of PAVs in *SLC6A1* shown conceptually.

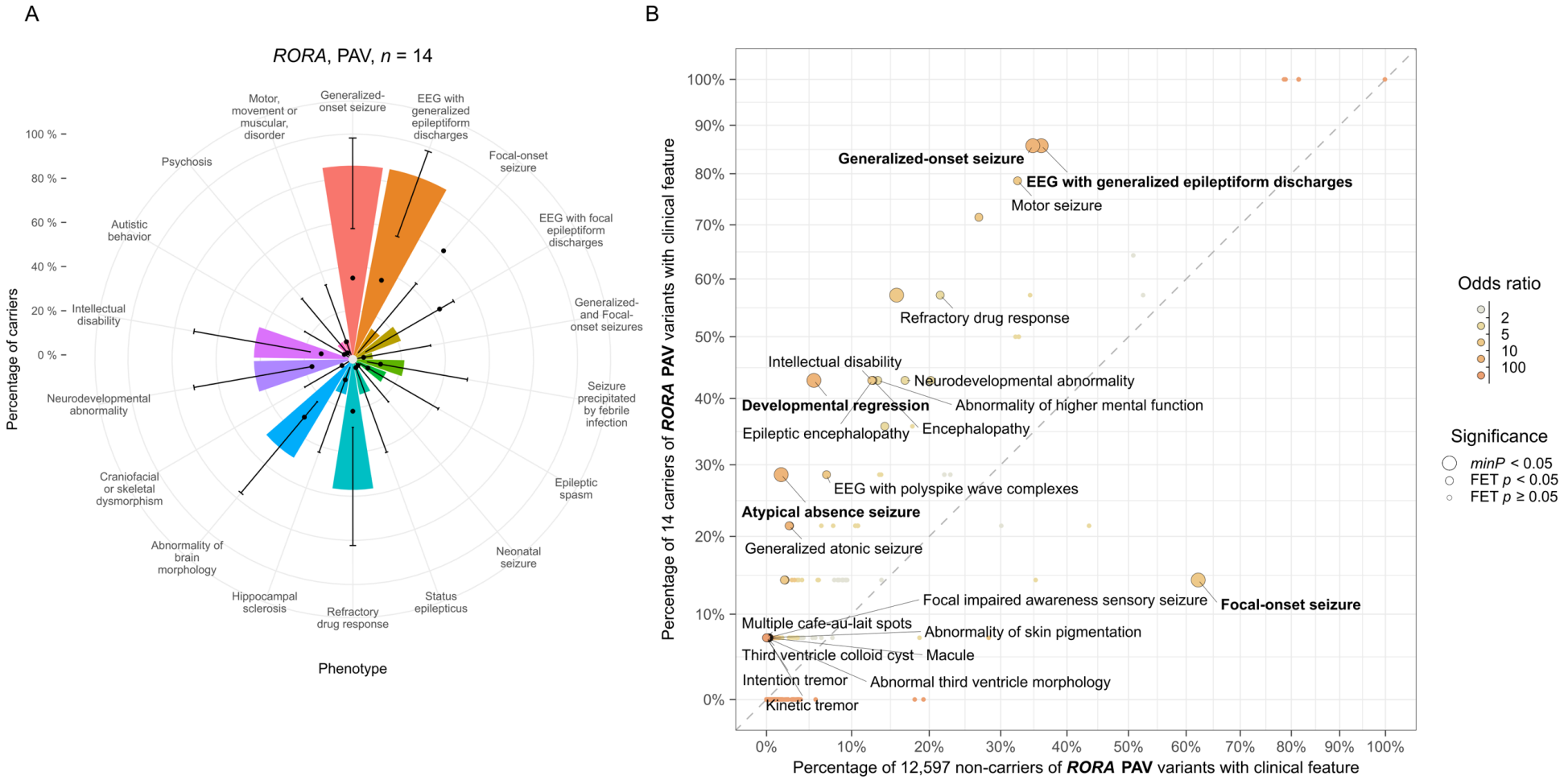


Figure 5-16.The associations of carriers of PAVs in *RORA*.

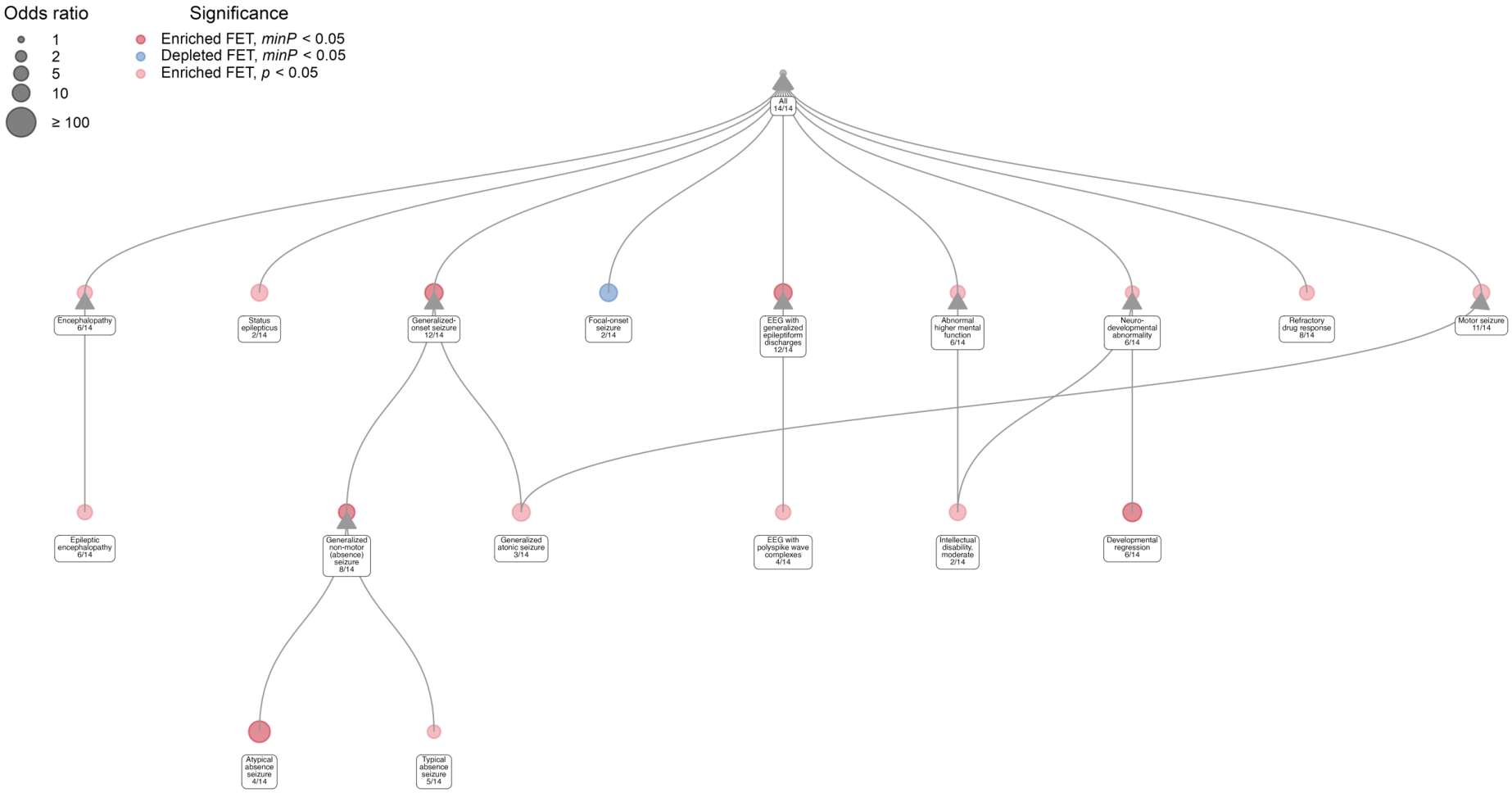


Figure 5-17. The associations of carriers of PAVs in *RORA* shown conceptually.

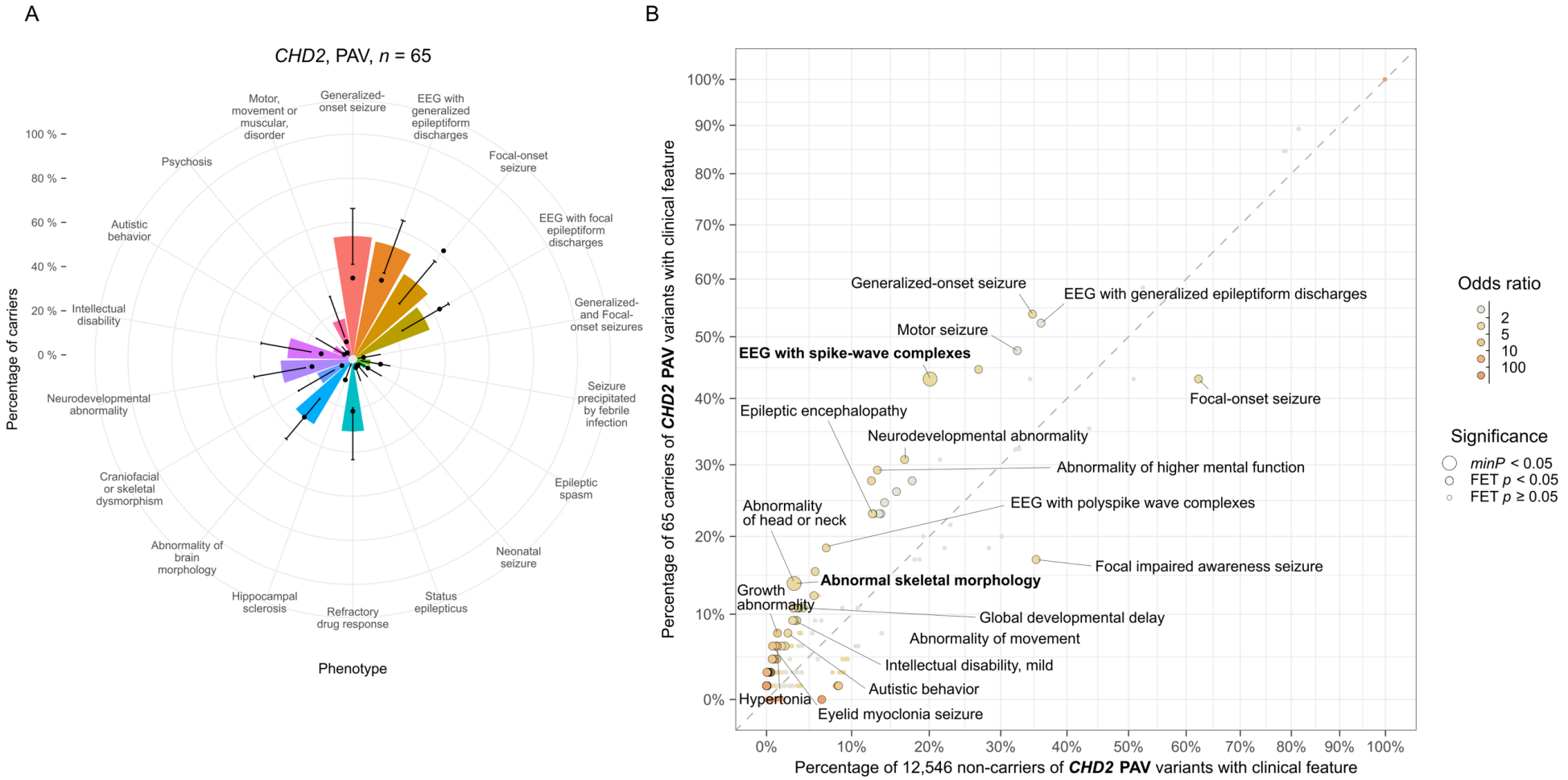


Figure 5-18.The associations of carriers of PAVs in *CHD2*.

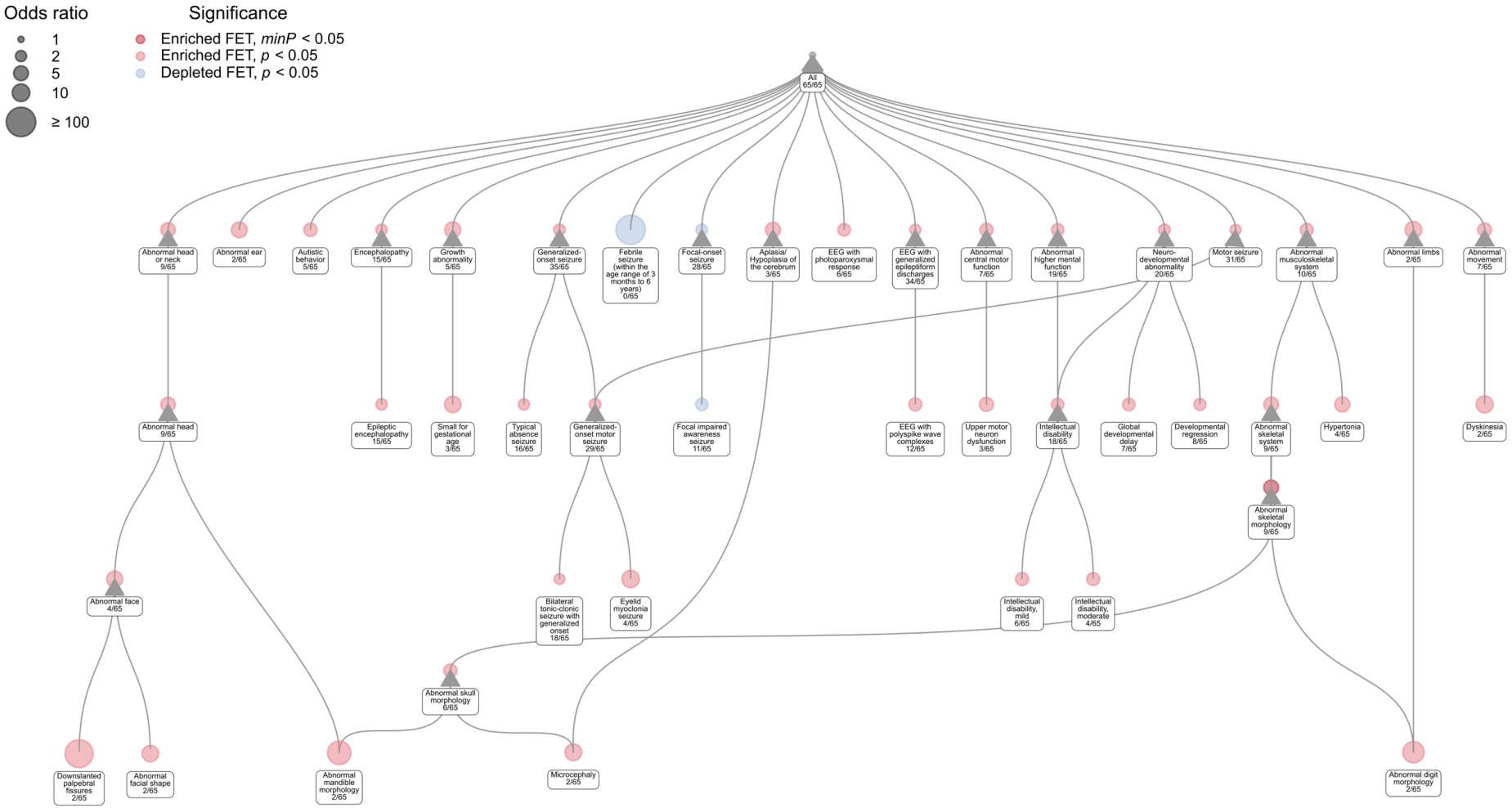


Figure 5-19. The associations of carriers of PAVs in *CHD2* shown conceptually.

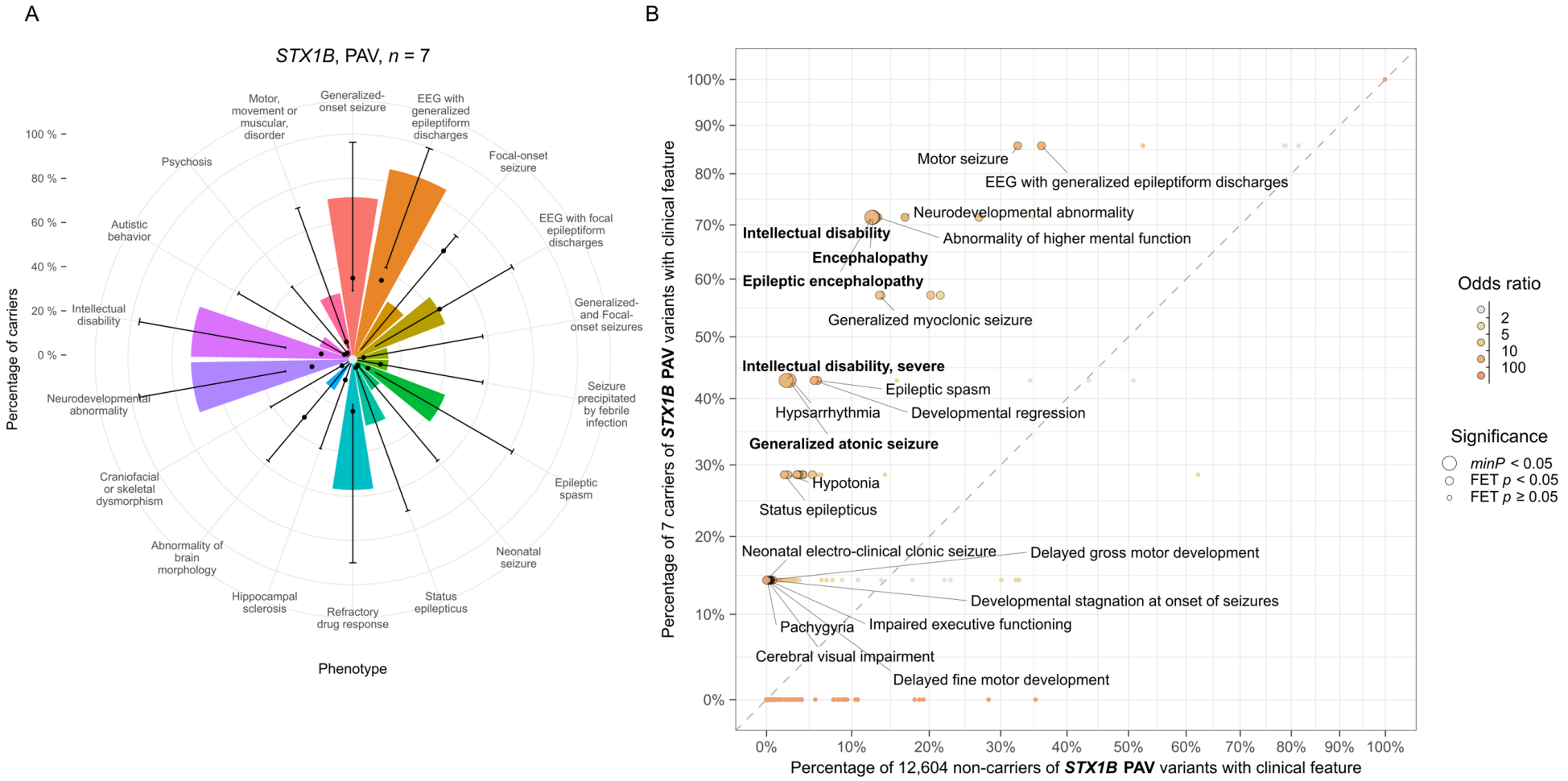


Figure 5-20 The associations of carriers of PAVs in *STX1B*.

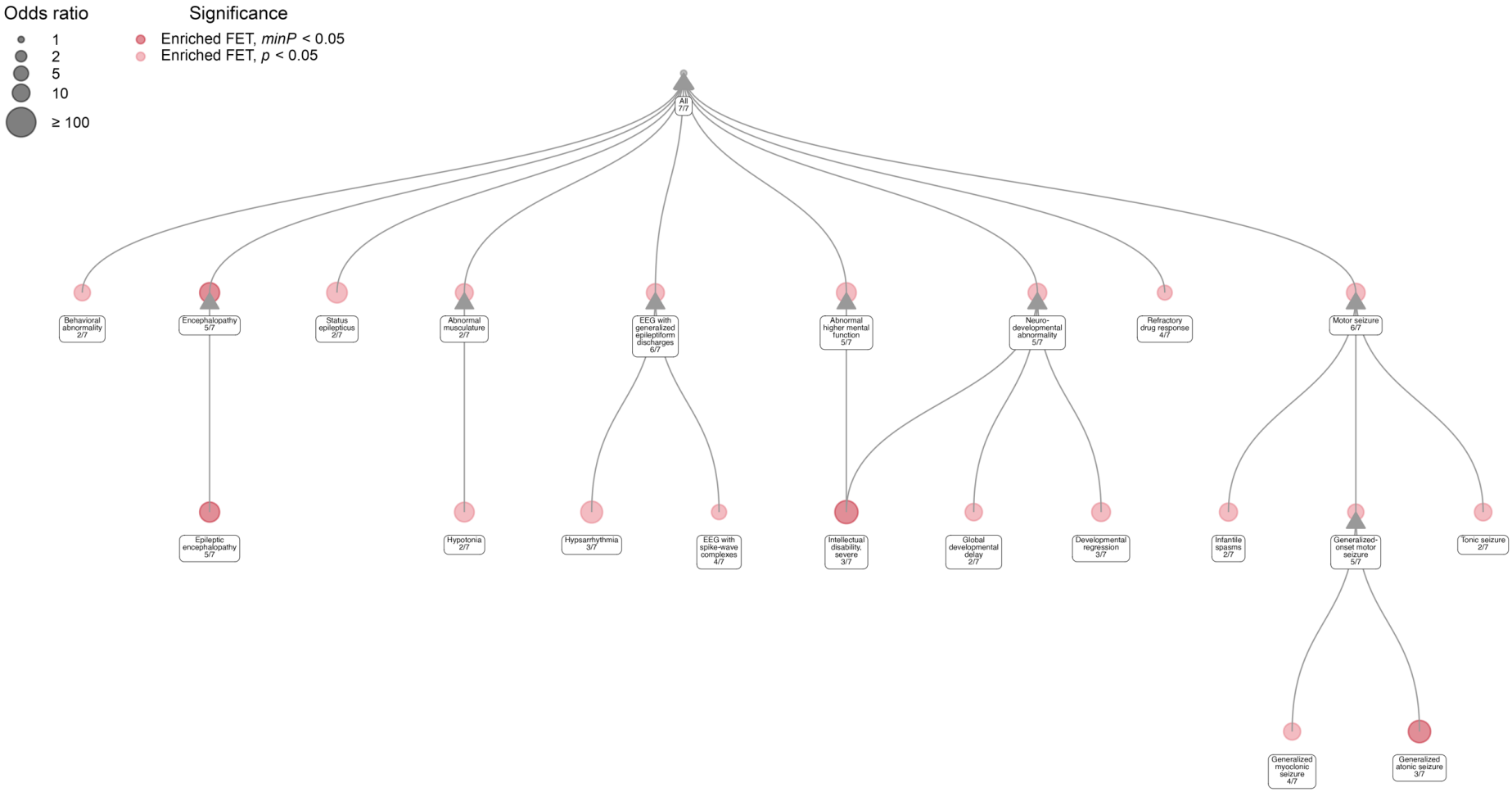


Figure 5-21 The associations of carriers of PAVs in *STX1B* shown conceptually.

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
WDR45 PTV	6	63 (22)	15 (6)	1.00×10^{-6}	2.80×10^{-5}	Diffuse cerebral atrophy [HP:0002506]	3/6 (50.0 %)	457	2.42×10^{-7}	1.10×10^{-4}	233	0.100
						Global developmental delay [HP:0001263]	5/6 (83.3 %)	114	7.67×10^{-7}	2.20×10^{-4}	19.9	0.00938
						Abnormal muscle tone [HP:0003808]	4/6 (66.7 %)	56.2	2.02×10^{-5}	0.00230	19.4	0.00915
						Epileptic spasm [HP:0011097]	4/6 (66.7 %)	32.0	1.66×10^{-4}	0.0113	11.3	0.00536
						Epileptic encephalopathy [HP:0200134]	5/6 (83.3 %)	34.6	1.72×10^{-4}	0.0117	6.61	0.00314
						Intellectual disability, severe [HP:0010864]	3/6 (50.0 %)	44.1	2.10×10^{-4}	0.0126	22.6	0.0106
NEXMIF PTV	13	31 (16)	12 (6)	1.00×10^{-6}	8.00×10^{-6}	Epileptic encephalopathy [HP:0200134]	10/13 (76.9 %)	23.2	2.02×10^{-7}	1.00×10^{-5}	6.12	0.00627
						Global developmental delay [HP:0001263]	6/13 (46.2 %)	19.6	7.38×10^{-6}	0.00147	11.0	0.0113
						Generalized myoclonic-atonic seizure [HP:0011170]	4/13 (30.8 %)	40.9	9.36×10^{-6}	0.00156	28.7	0.0288
						Intellectual disability, severe [HP:0010864]	4/13 (30.8 %)	19.7	1.49×10^{-4}	0.0164	13.9	0.0142
						Developmental regression [HP:0002376]	5/13 (38.5 %)	10.8	4.39×10^{-4}	0.0416	7.04	0.00722
						EEG with generalized epileptiform discharges [HP:0011198]	11/13 (84.6 %)	9.76	4.70×10^{-4}	0.0452	2.35	0.00242

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
SCN1A PTV	22	35 (19)	8 (4)	1.00×10^{-6}	5.20×10^{-5}	Seizure precipitated by febrile infection [HP:0032894]	13/22 (59.1 %)	12.1	4.93×10^{-8}	2.00×10^{-5}	5.52	0.00956
						Epileptic encephalopathy [HP:0200134]	13/22 (59.1 %)	10.1	3.29×10^{-7}	5.00×10^{-5}	4.71	0.00816
						Intellectual disability [HP:0001249]	12/22 (54.5 %)	8.42	2.80×10^{-6}	3.30×10^{-4}	4.37	0.00759
						Developmental regression [HP:0002376]	7/22 (31.8 %)	8.10	1.21×10^{-4}	0.0176	5.84	0.0101
KCNB1 PTV	7	33 (10)	6 (4)	1.80×10^{-4}	0.00712	Infantile spasms [HP:0012469]	4/7 (57.1 %)	36.9	4.83×10^{-5}	0.00498	16.4	0.00903
						Abnormality of skull size [HP:0000240]	3/7 (42.9 %)	32.0	3.97×10^{-4}	0.0243	18.8	0.0103
						Autistic behavior [HP:0000729]	3/7 (42.9 %)	30.0	4.78×10^{-4}	0.0280	17.6	0.00968
						Epileptic encephalopathy [HP:0200134]	5/7 (71.4 %)	17.3	5.40×10^{-4}	0.0306	5.67	0.00314

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
PURA PTV	3	55 (17)	8 (3)	3.00×10^{-5}	5.01×10^{-4}	Hypotonia [HP:0001252]	3/3 (100 %)	Inf	1.47×10^{-5}	0.00107	41.1	0.00968
						Global developmental delay [HP:0001263]	3/3 (100 %)	Inf	7.51×10^{-5}	0.00401	23.8	0.00563
						Developmental regression [HP:0002376]	3/3 (100 %)	Inf	1.65×10^{-4}	0.00592	18.3	0.00433
NSDHL PTV	1	36 (7)	5 (3)	-	-	Almond-shaped palpebral fissure [HP:0007874]	1/1 (100 %)	Inf	7.93×10^{-5}	0.0200	Inf	1.00
						Long eyelashes [HP:0000527]	1/1 (100 %)	Inf	7.93×10^{-5}	0.0200	Inf	1.00
						Thick eyebrow [HP:0000574]	1/1 (100 %)	Inf	7.93×10^{-5}	0.0200	Inf	1.00
MECP2 PTV	4	58 (24)	2 (2)	0.00571	0.0258	Abnormal emotion/affect behavior [HP:0100851]	2/4 (50.0 %)	373	4.47×10^{-5}	0.00343	191	0.0571
						Autistic behavior [HP:0000729]	3/4 (75.0 %)	120	5.78×10^{-5}	0.00370	30.8	0.00968
DEPDC5 PTV	44	27 (11)	2 (2)	0.283	5.26×10^{-4}	Focal-onset seizure [HP:0007359]	40/44 (90.9 %)	6.12	2.88×10^{-5}	0.00555	1.47	0.00511
						EEG with frontal epileptiform discharges [HP:0033716]	12/44 (27.3 %)	4.56	8.07×10^{-5}	0.0185	3.59	0.0124
SLC6A8 PTV	5	18 (13)	1 (1)	0.00466	0.251	Multifocal epileptiform discharges [HP:0010841]	4/5 (80.0 %)	59.4	7.63×10^{-5}	0.00605	12.7	0.00500
TSC2 PTV	12	21 (12)	1 (1)	0.0928	0.304	Behavioral abnormality [HP:0000708]	5/12 (41.7 %)	12.9	2.37×10^{-4}	0.0226	7.93	0.00750
ANKRD11 PTV	9	42 (11)	1 (1)	0.647	0.176	Generalized myoclonic-atonic seizure [HP:0011170]	3/9 (33.3 %)	45.7	1.05×10^{-4}	0.0103	30.9	0.0216
CHD2 PTV	9	25 (11)	1 (1)	0.0395	0.180	Abnormal facial shape [HP:0001999]	2/9 (22.2 %)	70.0	6.12×10^{-4}	0.0395	54.9	0.0377

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
CDKL5 PTV	6	14 (8)	1 (1)	0.00361	0.0199	Epileptic spasm [HP:0011097]	4/6 (66.7 %)	32.0	1.66×10^{-4}	0.0113	11.3	0.00536
SCN8A PTV	4	18 (8)	3 (1)	0.0121	0.110	Intellectual disability, moderate [HP:0002342]	3/4 (75.0 %)	139	3.74×10^{-5}	0.00297	35.7	0.0112
DCX PTV	3	20 (7)	6 (1)	1.86×10^{-4}	4.89×10^{-4}	Subcortical band heterotopia [HP:0032409]	3/3 (100 %)	Inf	7.78×10^{-9}	1.00×10^{-5}	548	0.115
SETD1B PTV	6	13 (7)	1 (1)	0.00318	0.00153	Typical absence seizure [HP:0011147]	5/6 (83.3 %)	30.3	3.03×10^{-4}	0.0173	5.89	0.00279
FLNA PTV	6	11 (5)	4 (1)	4.30×10^{-4}	3.76×10^{-4}	Periventricular nodular heterotopia [HP:0032388]	4/6 (66.7 %)	246	6.72×10^{-8}	2.00×10^{-5}	83.2	0.0381
KCNQ2 PTV	4	11 (5)	2 (1)	0.0443	0.100	Abnormal cerebral white matter morphology [HP:0002500]	3/4 (75.0 %)	77.2	2.04×10^{-4}	0.0102	20.1	0.00634
SLC16A2 PTV	1	12 (5)	4 (1)	-	-	Corticospinal tract atrophy [HP:0007117]	1/1 (100 %)	Inf	7.93×10^{-5}	0.0200	Inf	1.00
GRIN1 PTV	1	11 (4)	1 (1)	-	-	Asterixis [HP:0012164]	1/1 (100 %)	Inf	1.59×10^{-4}	0.0330	12600	0.500
GRIN2A PTV	14	16 (4)	1 (1)	0.564	0.113	EEG with central epileptiform discharges [HP:0033718]	7/14 (50.0 %)	26.9	2.19×10^{-7}	3.00×10^{-5}	14.0	0.0153
GABRA1 PTV	1	3 (1)	1 (1)	-	-	Abducens palsy [HP:0006897]	1/1 (100 %)	Inf	7.93×10^{-5}	0.0200	Inf	1.00

Table 5-11. A summary of the associations of carriers of PTVs in dominant epilepsy genes. Commenced 3 pages previously.

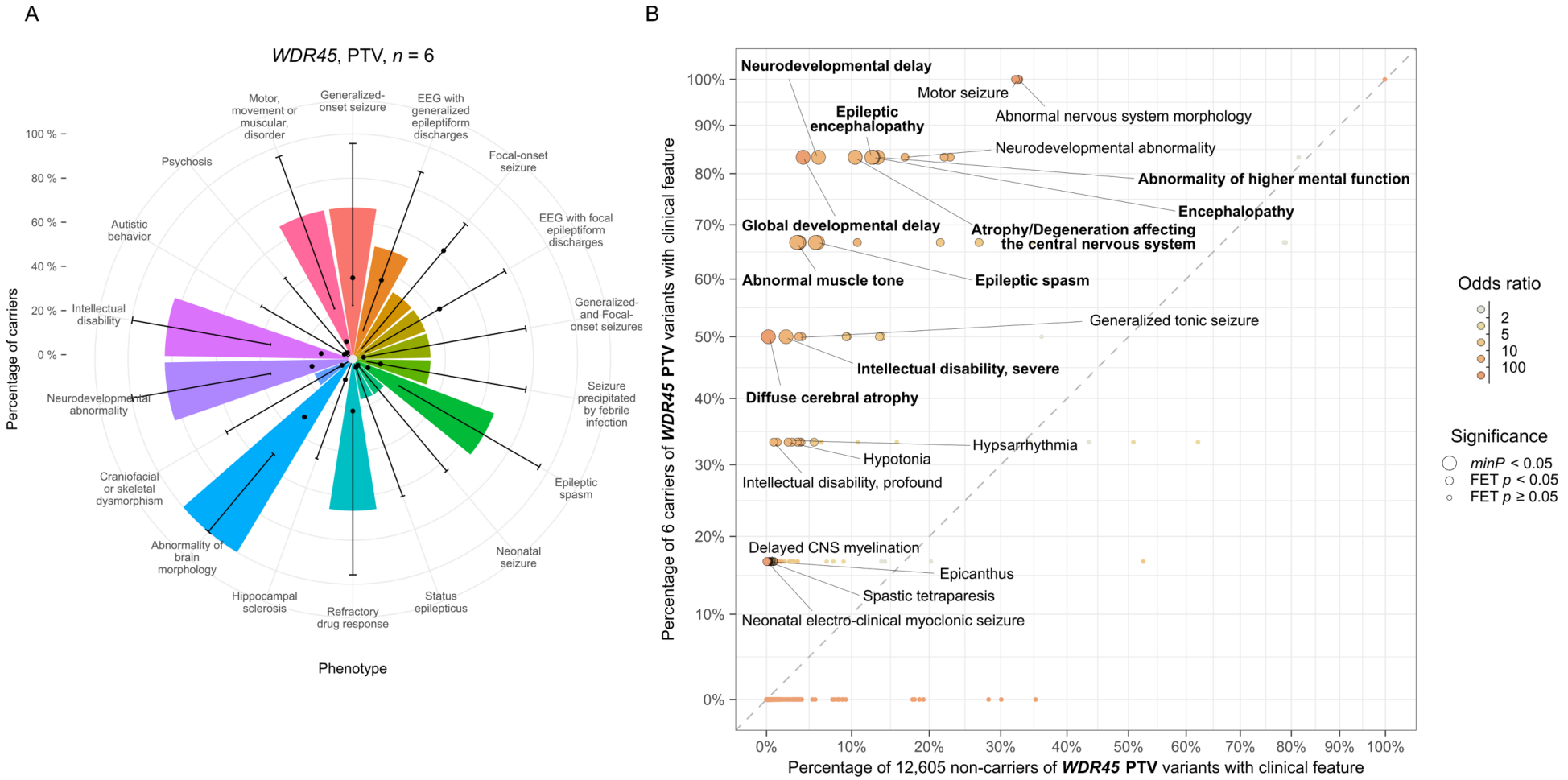


Figure 5-22. The associations of carriers of PTVs in *WDR45*.

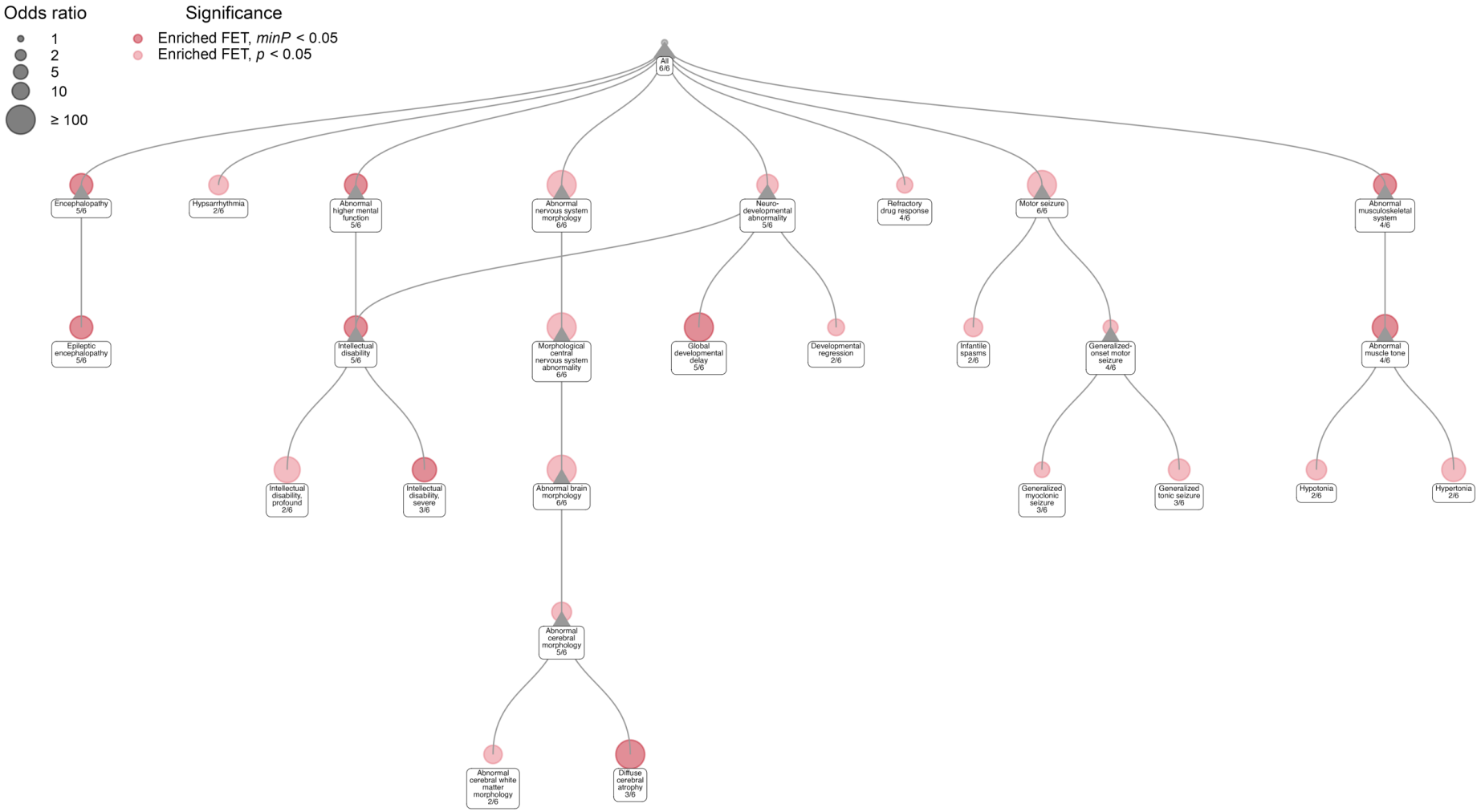


Figure 5-23. The associations of carriers of PTVs in *WDR45* shown conceptually.

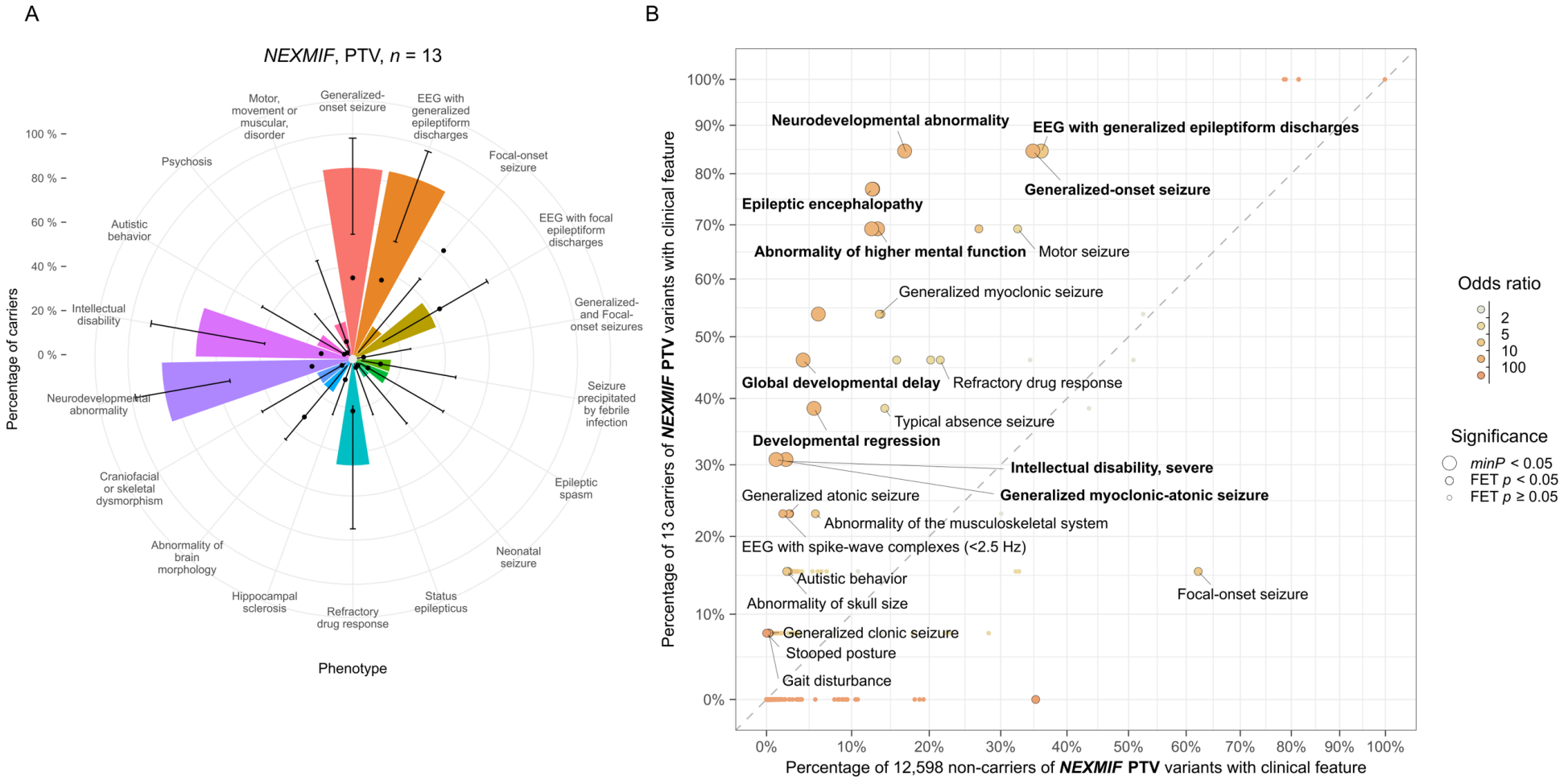


Figure 5-24. The associations of carriers of PTVs in *NEXMIF*.

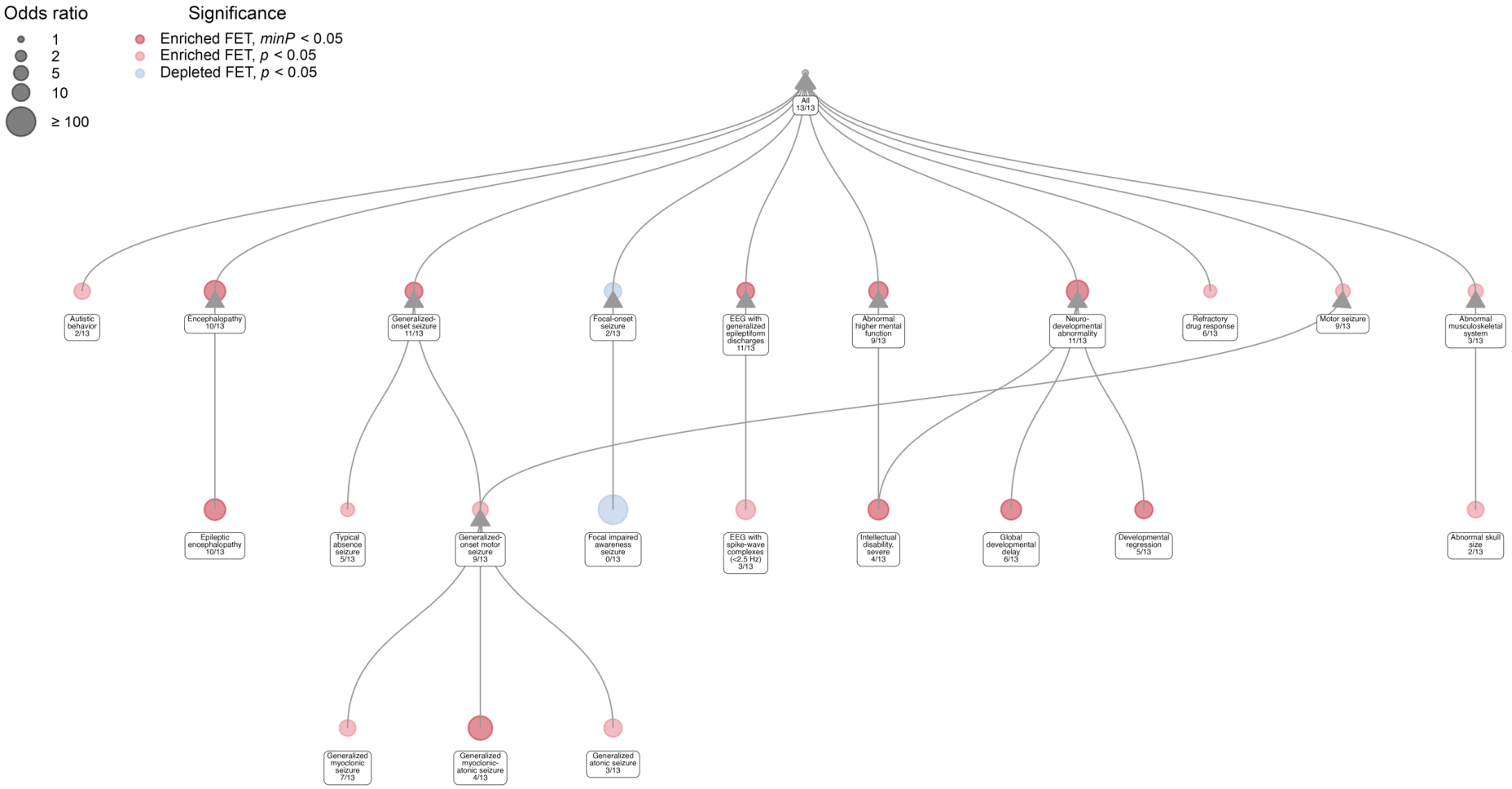


Figure 5-25. The associations of carriers of PTVs in *NEXMIF* shown conceptually.

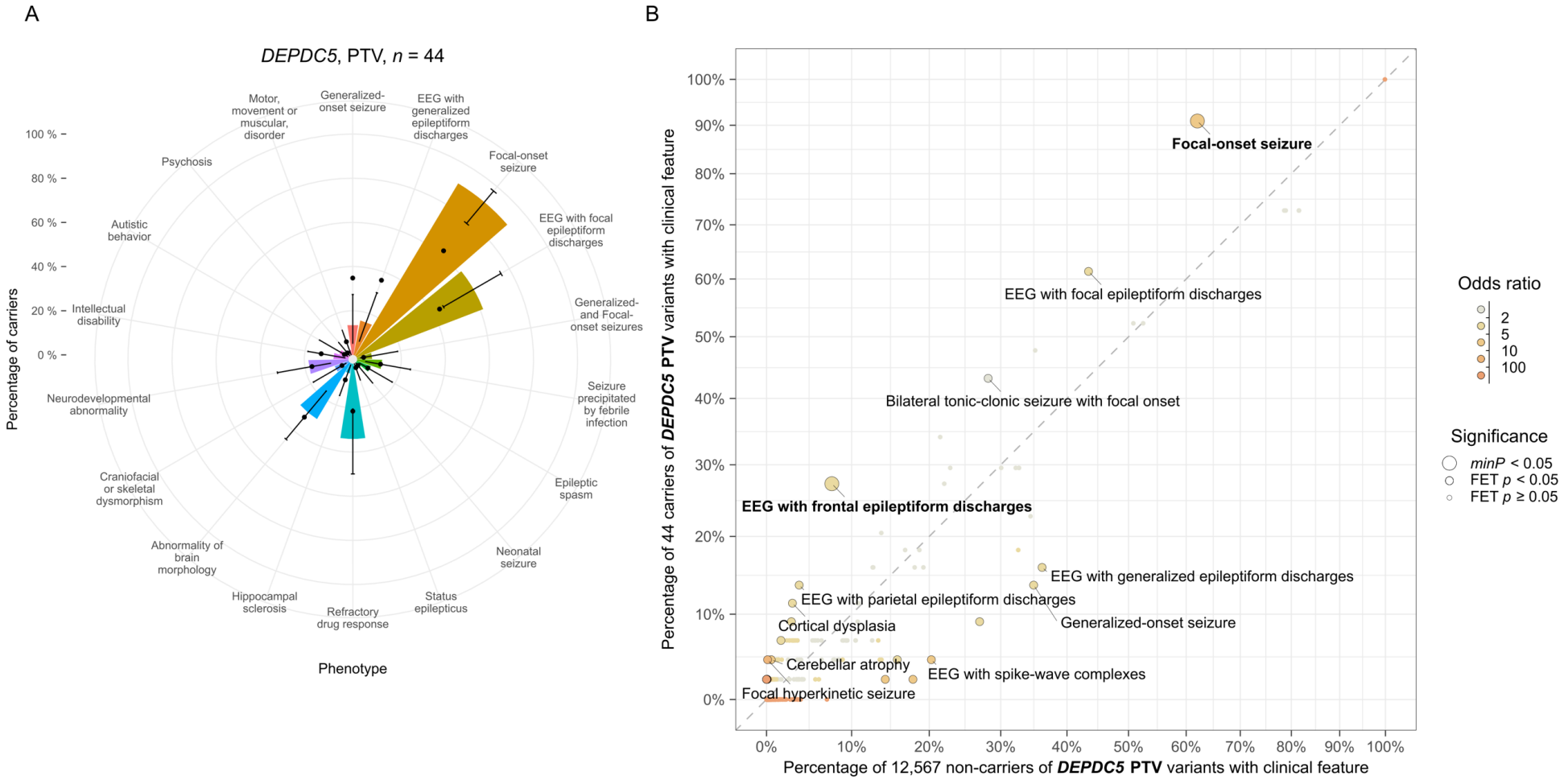


Figure 5-26. The associations of carriers of PTVs in *DEPDC5*.

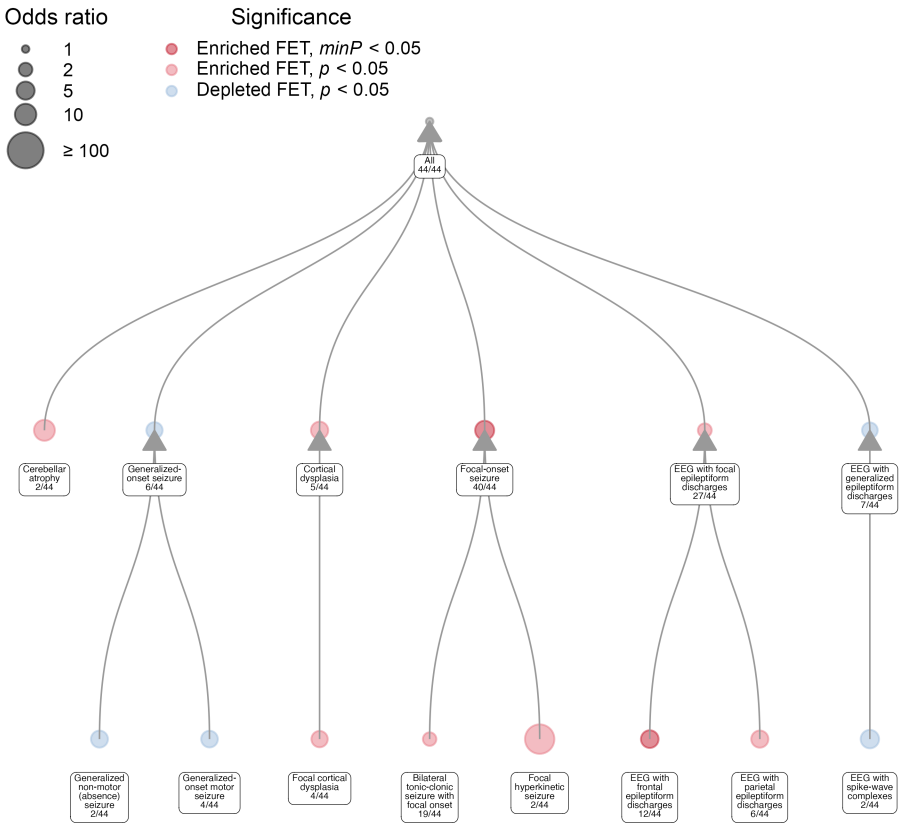


Figure 5-27. The associations of carriers of PTVs in *DEPDC5* shown conceptually.

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
IRF2BPL dPAV	28	43 (18)	12 (5)	0.594	0.0429	Large for gestational age [HP:0001520]	4/28 (14.3 %)	40.1	6.60×10^{-6}	0.00139	34.6	0.0714
						Cleft lip [HP:0410030]	2/28 (7.14 %)	959	1.42×10^{-5}	0.00351	899	0.667
						Cleft palate [HP:0000175]	2/28 (7.14 %)	191	9.91×10^{-5}	0.0180	180	0.286
						Epileptic encephalopathy [HP:0200134]	11/28 (39.3 %)	4.50	3.48×10^{-4}	0.0450	3.12	0.00690
						Infantile spasms [HP:0012469]	6/28 (21.4 %)	7.58	3.54×10^{-4}	0.0455	6.17	0.0135
SCN1A dPAV	112	66 (29)	12 (4)	3.50×10^{-5}	0.00206	Focal hemiclonic seizure [HP:0006813]	8/112 (7.14 %)	14.5	2.79×10^{-7}	4.00×10^{-5}	13.5	0.108
						Epileptic encephalopathy [HP:0200134]	33/112 (29.5 %)	2.93	2.01×10^{-6}	3.30×10^{-4}	2.36	0.0207
						Intellectual disability [HP:0001249]	31/112 (27.7 %)	2.70	1.74×10^{-5}	0.00337	2.23	0.0196
						Febrile seizure (within the age range of 3 months to 6 years) [HP:0002373]	19/112 (17.0 %)	3.06	7.44×10^{-5}	0.0108	2.71	0.0237
SLC6A1 dPAV	25	56 (23)	10 (4)	1.00×10^{-6}	1.00×10^{-6}	EEG with spike-wave complexes [HP:0010850]	17/25 (68.0 %)	8.43	3.05×10^{-7}	6.00×10^{-5}	3.38	0.00667
						Early onset absence seizures [HP:0011152]	5/25 (20.0 %)	34.3	1.08×10^{-6}	1.30×10^{-4}	27.7	0.0521
						Focal impaired awareness seizure [HP:0002384]	*0/25 (0.00 %)*	*0.00*	*2.52×10^{-5}*	*0.00578*	*0.00*	*0.00306*
						EEG with temporal epileptiform discharges [HP:0033717]	*0/25 (0.00 %)*	*0.00*	*2.30×10^{-4}*	*0.0322*	*0.00*	*0.00283*

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
RORA dPAV	14	38 (15)	6 (4)	2.00×10^{-5}	0.00162	Developmental regression [HP:0002376]	6/14 (42.9 %)	13.0	5.53×10^{-5}	0.00677	7.86	0.00866
						Atypical absence seizure [HP:0007270]	4/14 (28.6 %)	23.7	6.93×10^{-5}	0.00824	17.2	0.0188
						EEG with generalized epileptiform discharges [HP:0011198]	12/14 (85.7 %)	10.6	1.96×10^{-4}	0.0220	2.38	0.00264
						Focal-onset seizure [HP:0007359]	*2/14 (14.3 %)*	*0.101*	*3.34×10^{-4}*	*0.0312*	*0.230*	*0.00251*
SCN2A dPAV	54	68 (26)	3 (3)	0.402	0.270	Neonatal seizure [HP:0032807]	7/54 (13.0 %)	9.34	2.43×10^{-5}	0.00573	8.26	0.0343
						EEG with burst suppression [HP:0010851]	4/54 (7.41 %)	17.5	1.31×10^{-4}	0.0287	16.3	0.0656
						Abnormality of the musculoskeletal system [HP:0033127]	11/54 (20.4 %)	4.32	1.79×10^{-4}	0.0399	3.64	0.0154
STX1B dPAV	7	48 (25)	5 (3)	1.96×10^{-4}	0.00596	Intellectual disability, severe [HP:0010864]	3/7 (42.9 %)	33.1	3.62×10^{-4}	0.0222	19.4	0.0106
						Epileptic encephalopathy [HP:0200134]	5/7 (71.4 %)	17.3	5.40×10^{-4}	0.0306	5.67	0.00314
						Generalized atonic seizure [HP:0032887]	3/7 (42.9 %)	28.5	5.55×10^{-4}	0.0313	16.7	0.00920
WDR45 dPAV	12	59 (21)	4 (3)	0.00934	0.868	Diffuse cerebral atrophy [HP:0002506]	3/12 (25.0 %)	153	2.63×10^{-6}	6.40×10^{-4}	117	0.100
						Global developmental delay [HP:0001263]	5/12 (41.7 %)	16.3	8.19×10^{-5}	0.00903	9.94	0.00938
						Abnormal muscle tone [HP:0003808]	4/12 (33.3 %)	14.0	5.64×10^{-4}	0.0465	9.70	0.00915

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
NEXMIF dPAV	28	26 (16)	4 (3)	0.0890	0.275	Epileptic encephalopathy [HP:0200134]	12/28 (42.9 %)	5.21	6.88×10^{-5}	0.0120	3.41	0.00753
						Generalized myoclonic-atonic seizure [HP:0011170]	4/28 (14.3 %)	15.4	2.36×10^{-4}	0.0330	13.3	0.0288
						Neurodevelopmental abnormality [HP:0012759]	13/28 (46.4 %)	4.31	2.51×10^{-4}	0.0335	2.77	0.00613
TUBA1A dPAV	7	17 (7)	3 (3)	0.00786	0.145	Abnormal cortical gyration [HP:0002536]	3/7 (42.9 %)	42.1	1.81×10^{-4}	0.0133	24.6	0.0135
						Multifocal epileptiform discharges [HP:0010841]	4/7 (57.1 %)	19.8	4.82×10^{-4}	0.0283	9.05	0.00500
						Intellectual disability [HP:0001249]	5/7 (71.4 %)	17.5	5.21×10^{-4}	0.0300	5.71	0.00316
MECP2 dPAV	22	57 (25)	4 (2)	0.437	0.00119	Neurodevelopmental delay [HP:0012758]	8/22 (36.4 %)	8.99	2.51×10^{-5}	0.00508	6.09	0.0105
						Epileptic encephalopathy [HP:0200134]	10/22 (45.5 %)	5.79	1.54×10^{-4}	0.0217	3.61	0.00627
KCNT1 dPAV	51	51 (20)	11 (2)	0.638	0.492	Cleft lip [HP:0410030]	3/51 (5.88 %)	Inf	6.23×10^{-8}	3.00×10^{-5}	Inf	1.00
						Cleft palate [HP:0000175]	3/51 (5.88 %)	194	2.16×10^{-6}	3.90×10^{-4}	185	0.429
STXBP1 dPAV	17	21 (12)	2 (2)	0.457	0.0119	Epileptic spasm [HP:0011097]	7/17 (41.2 %)	11.2	2.84×10^{-5}	0.00513	7.02	0.00938
						Neonatal seizure [HP:0032807]	4/17 (23.5 %)	19.0	1.34×10^{-4}	0.0169	14.8	0.0196
CHD2 dPAV	52	101 (39)	1 (1)	0.00559	0.393	EEG with spike-wave complexes [HP:0010850]	24/52 (46.2 %)	3.40	2.98×10^{-5}	0.00658	2.29	0.00941
KCNA2 dPAV	19	47 (20)	2 (1)	0.268	0.971	Convulsive status epilepticus [HP:0032660]	4/19 (21.1 %)	23.5	5.68×10^{-5}	0.00844	18.8	0.0276
GRIN2A dPAV	59	50 (19)	1 (1)	0.649	4.31×10^{-4}	EEG with central epileptiform discharges [HP:0033718]	14/59 (23.7 %)	8.48	1.69×10^{-8}	1.00×10^{-5}	6.71	0.0306

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
NTRK2 dPAV (dMissense)	12	43 (19)	3 (1)	0.303	0.125	Feeding difficulties in infancy [HP:0008872]	2/12 (16.7 %)	99.8	2.88×10^{-4}	0.0267	84.0	0.0741
PURA dPAV	5	55 (17)	2 (1)	0.0970	0.0602	Hypotonia [HP:0001252]	3/5 (60.0 %)	60.0	1.42×10^{-4}	0.00819	24.6	0.00968
CDKL5 dPAV	19	32 (16)	1 (1)	0.00384	0.0168	Epileptic spasm [HP:0011097]	7/19 (36.8 %)	9.35	6.63×10^{-5}	0.00984	6.28	0.00938
KCNH1 dPAV	25	68 (16)	3 (1)	0.263	0.697	Downslanted palpebral fissures [HP:0000494]	2/25 (8.00 %)	216	7.88×10^{-5}	0.0147	201	0.286
KCNQ2 dPAV	23	36 (16)	1 (1)	0.00809	0.518	EEG with burst suppression [HP:0010851]	3/23 (13.0 %)	32.3	1.78×10^{-4}	0.0249	28.3	0.0492
ASXL3 dPAV	70	38 (15)	2 (1)	0.662	0.410	Increased inflammatory response [HP:0012649]	2/70 (2.86 %)	182	1.81×10^{-4}	0.0453	179	0.500
KLHL20 dPAV	17	36 (15)	1 (1)	0.00379	0.0280	Neurodevelopmental abnormality [HP:0012759]	10/17 (58.8 %)	7.10	1.09×10^{-4}	0.0130	3.51	0.00472
NSDHL dPAV	3	43 (12)	1 (1)	0.0833	0.156	Myoclonus [HP:0001336]	2/3 (66.7 %)	889	1.64×10^{-5}	0.00130	300	0.0667
ATN1 dPAV	31	18 (11)	2 (1)	0.246	0.0804	EEG with focal slow activity [HP:0010843]	9/31 (29.0 %)	6.95	3.47×10^{-5}	0.00557	5.22	0.0127
DCX dPAV	10	19 (11)	6 (1)	0.00648	0.578	Subcortical band heterotopia [HP:0032409]	5/10 (50.0 %)	585	6.20×10^{-12}	1.00×10^{-5}	300	0.192
TSC1 dPAV	44	24 (11)	1 (1)	0.0270	0.628	Multifocal cerebral white matter abnormalities [HP:0007052]	5/44 (11.4 %)	10.8	1.83×10^{-4}	0.0378	9.65	0.0327
DEPDC5 dPAV	80	16 (10)	1 (1)	0.672	0.0392	EEG with parietal epileptiform discharges [HP:0033719]	11/80 (13.8 %)	4.15	1.89×10^{-4}	0.0406	3.72	0.0232
DHDDS dPAV	11	15 (9)	1 (1)	0.126	0.00493	*Focal-onset seizure* *[HP:0007359]*	*1/11* *(9.09 %)*	*0.0609*	*4.36×10^{-4}*	*0.0402*	*0.146*	*0.00209*
KCNT2 dPAV	29	30 (9)	1 (1)	0.798	0.604	Benign neoplasm of the central nervous system [HP:0100835]	5/29 (17.2 %)	9.98	2.94×10^{-4}	0.0411	8.44	0.0191

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
RORB dPAV	18	22 (9)	1 (1)	0.0330	0.270	EEG with photoparoxysmal response [HP:0010852]	5/18 (27.8 %)	10.8	2.94×10^{-4}	0.0341	8.06	0.0114
YWHAG dPAV	8	25 (9)	1 (1)	0.0405	0.0134	Intellectual disability, mild [HP:0001256]	4/8 (50.0 %)	30.5	6.49×10^{-5}	0.00654	15.8	0.00993
FLNA dPAV	43	26 (8)	3 (1)	0.173	0.0320	Periventricular heterotopia [HP:0007165]	5/43 (11.6 %)	14.0	5.60×10^{-5}	0.0108	12.5	0.0410
SLC2A1 dPAV	20	14 (8)	1 (1)	0.296	0.503	EEG with spike-wave complexes (<2.5 Hz) [HP:0010847]	4/20 (20.0 %)	13.3	4.58×10^{-4}	0.0481	10.9	0.0169
TUBB2B dPAV	6	23 (8)	1 (1)	0.812	0.846	Lissencephaly [HP:0001339]	2/6 (33.3 %)	115	2.87×10^{-4}	0.0170	77.8	0.0357
MED12 dPAV	24	21 (7)	1 (1)	0.584	1.52×10^{-4}	Bilateral tonic-clonic seizure with generalized onset [HP:0025190]	12/24 (50.0 %)	4.65	3.13×10^{-4}	0.0411	2.82	0.00535
SATB2 dPAV (dMissense)	9	13 (6)	1 (1)	0.473	0.242	Benign neoplasm of the central nervous system [HP:0100835]	3/9 (33.3 %)	23.8	6.79×10^{-4}	0.0422	16.2	0.0115
ATP6V0A1 dPAV	29	5 (4)	1 (1)	0.672	0.462	Hot water-induced seizure [HP:0020209]	2/29 (6.90 %)	103	2.77×10^{-4}	0.0405	96.4	0.182
SCN1B dPAV	11	5 (3)	1 (1)	0.258	0.277	Febrile seizure (within the age range of 3 months to 6 years) [HP:0002373]	5/11 (45.5 %)	12.4	3.42×10^{-4}	0.0323	7.20	0.00624

Table 5-12. A summary of the associations of carriers of dPAVs in dominant epilepsy genes. Commenced 4 pages previously.

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
IRF2BPL dMissense	26	42 (18)	9 (4)	0.596	0.0956	Cleft lip [HP:0410030]	2/26 (7.69 %)	1040	1.22×10^{-5}	0.00297	968	0.667
						Cleft palate [HP:0000175]	2/26 (7.69 %)	208	8.53×10^{-5}	0.0152	194	0.286
						Epileptic spasm [HP:0011097]	8/26 (30.8 %)	7.13	8.69×10^{-5}	0.0153	5.25	0.0107
						Large for gestational age [HP:0001520]	3/26 (11.5 %)	30.8	2.01×10^{-4}	0.0278	27.4	0.0536
SLC6A1 dMissense	21	54 (24)	8 (3)	1.00×10^{-6}	2.00×10^{-6}	EEG with spike-wave complexes [HP:0010850]	14/21 (66.7 %)	7.93	5.05×10^{-6}	9.00×10^{-4}	3.31	0.00549
						Early onset absence seizures [HP:0011152]	4/21 (19.0 %)	31.9	1.71×10^{-5}	0.00380	26.1	0.0417
						Focal impaired awareness seizure *[HP:0002384]*	*0/21* *(0.00 %)*	*0.00*	*1.34×10^{-4}*	*0.0185*	*0.00*	*0.00257*
RORA dMissense	13	39 (16)	5 (3)	8.00×10^{-6}	0.00221	Developmental regression [HP:0002376]	6/13 (46.2 %)	14.9	3.32×10^{-5}	0.00425	8.46	0.00866
						Atypical absence seizure [HP:0007270]	4/13 (30.8 %)	26.3	5.02×10^{-5}	0.00616	18.5	0.0188
						EEG with generalized epileptiform discharges [HP:0011198]	11/13 (84.6 %)	9.76	4.70×10^{-4}	0.0452	2.35	0.00242
KLHL20 dMissense	16	36 (14)	4 (3)	4.33×10^{-4}	0.00647	Neurodevelopmental abnormality [HP:0012759]	10/16 (62.5 %)	8.28	5.26×10^{-5}	0.00694	3.73	0.00472
						EEG with generalized epileptiform discharges [HP:0011198]	13/16 (81.2 %)	7.69	2.89×10^{-4}	0.0315	2.26	0.00286
						Epileptic encephalopathy [HP:0200134]	8/16 (50.0 %)	6.94	3.21×10^{-4}	0.0340	3.97	0.00502

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
TUBA1A dMissense	6	19 (7)	4 (3)	0.00125	0.0447	Abnormal cortical gyration [HP:0002536]	3/6 (50.0 %)	56.2	1.05×10^{-4}	0.00850	28.6	0.0135
						Intellectual disability [HP:0001249]	5/6 (83.3 %)	34.9	1.66×10^{-4}	0.0114	6.66	0.00316
						Multifocal epileptiform discharges [HP:0010841]	4/6 (66.7 %)	29.7	2.17×10^{-4}	0.0131	10.6	0.00500
SCN2A dMissense	50	67 (25)	2 (2)	0.637	0.370	Neonatal seizure [HP:0032807]	7/50 (14.0 %)	10.2	1.45×10^{-5}	0.00198	8.93	0.0343
						EEG with burst suppression [HP:0010851]	4/50 (8.00 %)	19.1	9.66×10^{-5}	0.0233	17.6	0.0656
KCNT1 dMissense	42	59 (21)	11 (2)	0.388	0.160	Cleft lip [HP:0410030]	3/42 (7.14 %)	Inf	3.44×10^{-8}	1.00×10^{-5}	Inf	1.00
						Cleft palate [HP:0000175]	3/42 (7.14 %)	239	1.19×10^{-6}	1.80×10^{-4}	224	0.429
SCN1A dMissense	88	44 (15)	6 (2)	0.0179	0.233	Seizure precipitated by febrile infection [HP:0032894]	30/88 (34.1 %)	4.35	4.16×10^{-9}	1.00×10^{-5}	3.21	0.0221
						Focal hemiclonic seizure [HP:0006813]	6/88 (6.82 %)	13.4	1.23×10^{-5}	0.00228	12.6	0.0811
SMARCA2 dMissense	28	27 (13)	2 (2)	0.221	0.792	Hypsarrhythmia [HP:0002521]	6/28 (21.4 %)	9.26	1.26×10^{-4}	0.0198	7.49	0.0164
						Infantile spasms [HP:0012469]	6/28 (21.4 %)	7.58	3.54×10^{-4}	0.0455	6.17	0.0135
KCNQ2 dMissense	18	30 (12)	2 (2)	0.0310	0.219	EEG with burst suppression [HP:0010851]	3/18 (16.7 %)	43.1	8.34×10^{-5}	0.0112	36.2	0.0492
						Neonatal seizure [HP:0032807]	4/18 (22.2 %)	17.7	1.70×10^{-4}	0.0192	14.0	0.0196
GABRB2 dMissense	14	67 (21)	1 (1)	0.0253	0.00936	Neonatal electro-clinical tonic seizure [HP:0032821]	2/14 (14.3 %)	77.1	4.57×10^{-4}	0.0424	66.7	0.0690
KMT2E dMissense	68	36 (20)	1 (1)	0.144	0.164	Impaired executive functioning [HP:0033051]	3/68 (4.41 %)	33.9	1.60×10^{-4}	0.0304	32.6	0.150

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
RELN dMissense	86	32 (20)	1 (1)	0.797	0.180	Generalized clonic seizure [HP:0011169]	4/86 (4.65 %)	17.4	1.38×10^{-4}	0.0375	16.6	0.103
KCNA2 dMissense	18	48 (19)	3 (1)	0.297	0.942	Convulsive status epilepticus [HP:0032660]	4/18 (22.2 %)	25.2	4.53×10^{-5}	0.00720	19.8	0.0276
NTRK2 dMissense	12	43 (19)	3 (1)	0.303	0.125	Feeding difficulties in infancy [HP:0008872]	2/12 (16.7 %)	99.8	2.88×10^{-4}	0.0267	84.0	0.0741
KCNH1 dMissense	22	74 (18)	3 (1)	0.328	0.797	Downslanted palpebral fissures [HP:0000494]	2/22 (9.09 %)	249	6.07×10^{-5}	0.0109	229	0.286
TFE3 dMissense	45	37 (16)	1 (1)	0.890	0.302	Low-set ears [HP:0000369]	2/45 (4.44 %)	116	2.59×10^{-4}	0.0497	112	0.286
GABRB1 dMissense	7	38 (15)	1 (1)	0.0418	0.625	Neurodevelopmental delay [HP:0012758]	4/7 (57.1 %)	20.9	3.96×10^{-4}	0.0242	9.53	0.00526
AP1G1 dMissense	19	54 (14)	1 (1)	0.159	0.945	Abnormal palate morphology [HP:0000174]	2/19 (10.5 %)	91.8	3.24×10^{-4}	0.0368	82.8	0.111
CAPRIN1 dMissense	12	49 (14)	1 (1)	0.0167	0.310	Abnormal skeletal muscle morphology [HP:0011805]	2/12 (16.7 %)	86.1	3.80×10^{-4}	0.0370	72.4	0.0645
CUX2 dMissense	31	34 (13)	1 (1)	0.489	0.130	Abnormal facial shape [HP:0001999]	3/31 (9.68 %)	26.8	2.90×10^{-4}	0.0418	24.3	0.0566
ASXL3 dMissense	43	31 (11)	2 (1)	0.626	0.330	Increased inflammatory response [HP:0012649]	2/43 (4.65 %)	301	6.78×10^{-5}	0.0160	292	0.500
ATN1 dMissense	22	14 (11)	2 (1)	0.328	0.127	EEG with focal slow activity [HP:0010843]	8/22 (36.4 %)	9.70	1.50×10^{-5}	0.00274	6.54	0.0113
KCNT2 dMissense	22	38 (9)	3 (1)	0.654	0.450	Benign neoplasm of the central nervous system [HP:0100835]	5/22 (22.7 %)	14.1	7.34×10^{-5}	0.0116	11.1	0.0191
TUBB2B dMissense	4	26 (9)	1 (1)	0.404	0.949	Lissencephaly [HP:0001339]	2/4 (50.0 %)	230	1.16×10^{-4}	0.00625	117	0.0357
YWHAG dMissense	7	25 (9)	1 (1)	0.0545	0.0424	Febrile seizure (within the age range of 3 months to 6 years) [HP:0002373]	4/7 (57.1 %)	19.7	4.84×10^{-4}	0.0286	9.04	0.00499

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
DCX dMissense	7	13 (8)	4 (1)	0.251	0.874	Subcortical band heterotopia [HP:0032409]	2/7 (28.6 %)	208	8.53×10^{-5}	0.00735	150	0.0769
MECP2 dMissense	18	16 (8)	1 (1)	0.794	0.00758	Neurodevelopmental delay [HP:0012758]	6/18 (33.3 %)	7.85	4.66×10^{-4}	0.0471	5.57	0.00789
RORB dMissense	14	20 (8)	2 (1)	0.0374	0.0772	EEG with photoparoxysmal response grade IV [HP:0011214]	4/14 (28.6 %)	14.7	4.10×10^{-4}	0.0370	10.8	0.0118
MED12 dMissense	20	19 (7)	1 (1)	0.149	1.40×10^{-4}	Bilateral tonic-clonic seizure with generalized onset [HP:0025190]	11/20 (55.0 %)	5.68	1.88×10^{-4}	0.0232	3.11	0.00491
SATB2 dMissense	9	13 (6)	1 (1)	0.473	0.242	Benign neoplasm of the central nervous system [HP:0100835]	3/9 (33.3 %)	23.8	6.79×10^{-4}	0.0422	16.2	0.0115
NSDHL dMissense	2	8 (4)	1 (1)	0.400	0.969	Intention tremor [HP:0002080]	1/2 (50.0 %)	Inf	1.59×10^{-4}	0.0448	Inf	1.00

Table 5-13. A summary of the associations of carriers of dMissense variants in dominant epilepsy genes. Commenced 3 pages previously.

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
IRF2BPL Missense	27	40 (17)	11 (5)	0.489	0.0553	Cleft lip [HP:0410030]	2/27 (7.41 %)	999	1.32×10^{-5}	0.00333	932	0.667
						Cleft palate [HP:0000175]	2/27 (7.41 %)	198	9.21×10^{-5}	0.0164	186	0.286
						Epileptic spasm [HP:0011097]	8/27 (29.6 %)	6.76	1.17×10^{-4}	0.0182	5.05	0.0107
						Large for gestational age [HP:0001520]	3/27 (11.1 %)	29.5	2.25×10^{-4}	0.0303	26.4	0.0536
						Epileptic encephalopathy [HP:0200134]	11/27 (40.7 %)	4.78	2.38×10^{-4}	0.0309	3.24	0.00690
SLC6A1 Missense	23	48 (19)	7 (3)	8.00×10^{-6}	8.00×10^{-6}	EEG with spike-wave complexes [HP:0010850]	14/23 (60.9 %)	6.16	2.35×10^{-5}	0.00496	3.02	0.00549
						Early onset absence seizures [HP:0011152]	4/23 (17.4 %)	28.6	2.50×10^{-5}	0.00510	23.8	0.0417
						Focal-onset seizure *[HP:0007359]*	*5/23* *(21.7 %)*	*0.169*	*1.11×10^{-4}*	*0.0156*	*0.350*	*0.00377*
RORA Missense (dMissense)	13	39 (16)	5 (3)	8.00×10^{-6}	0.00221	Developmental regression [HP:0002376]	6/13 (46.2 %)	14.9	3.32×10^{-5}	0.00425	8.46	0.00866
						Atypical absence seizure [HP:0007270]	4/13 (30.8 %)	26.3	5.02×10^{-5}	0.00616	18.5	0.0188
						EEG with generalized epileptiform discharges [HP:0011198]	11/13 (84.6 %)	9.76	4.70×10^{-4}	0.0452	2.35	0.00242
TUBA1A Missense (dMissense)	6	19 (7)	4 (3)	0.00125	0.0447	Abnormal cortical gyration [HP:0002536]	3/6 (50.0 %)	56.2	1.05×10^{-4}	0.00850	28.6	0.0135
						Intellectual disability [HP:0001249]	5/6 (83.3 %)	34.9	1.66×10^{-4}	0.0114	6.66	0.00316
						Multifocal epileptiform discharges [HP:0010841]	4/6 (66.7 %)	29.7	2.17×10^{-4}	0.0131	10.6	0.00500

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
SCN2A Missense	56	63 (24)	2 (2)	0.643	0.272	Neonatal seizure [HP:0032807]	7/56 (12.5 %)	8.96	3.10×10^{-5}	0.00715	7.97	0.0343
						EEG with burst suppression [HP:0010851]	4/56 (7.14 %)	16.8	1.51×10^{-4}	0.0325	15.7	0.0656
SCN1A Missense	92	75 (23)	6 (2)	0.00459	0.336	Seizure precipitated by febrile infection [HP:0032894]	31/92 (33.7 %)	4.28	3.17×10^{-9}	1.00×10^{-5}	3.17	0.0228
						Focal hemiclonic seizure [HP:0006813]	6/92 (6.52 %)	12.8	1.59×10^{-5}	0.00257	12.0	0.0811
KCNT1 Missense	46	61 (21)	11 (2)	0.631	0.317	Cleft lip [HP:0410030]	3/46 (6.52 %)	Inf	4.54×10^{-8}	2.00×10^{-5}	Inf	1.00
						Cleft palate [HP:0000175]	3/46 (6.52 %)	217	1.57×10^{-6}	2.20×10^{-4}	205	0.429
KCNH1 Missense	28	73 (20)	4 (2)	0.221	0.766	Seizure precipitated by febrile infection [HP:0032894]	11/28 (39.3 %)	5.39	8.30×10^{-5}	0.0148	3.66	0.00809
						Downslanted palpebral fissures [HP:0000494]	2/28 (7.14 %)	191	9.91×10^{-5}	0.0180	180	0.286
KCNA2 Missense	20	46 (18)	3 (2)	0.184	0.863	Convulsive status epilepticus [HP:0032660]	4/20 (20.0 %)	22.0	7.04×10^{-5}	0.0105	17.9	0.0276
KCNA2 Missense	20	46 (18)	3 (2)	0.184	0.863	Intellectual disability, mild [HP:0001256]	5/20 (25.0 %)	10.2	3.39×10^{-4}	0.0380	7.91	0.0124
SMARCA2 Missense	33	28 (13)	2 (2)	0.204	0.774	Hypsarrhythmia [HP:0002521]	7/33 (21.2 %)	9.16	3.64×10^{-5}	0.00649	7.43	0.0191
						Infantile spasms [HP:0012469]	7/33 (21.2 %)	7.49	1.22×10^{-4}	0.0214	6.12	0.0158
KCNT2 Missense	27	39 (11)	4 (2)	0.793	0.450	Benign neoplasm of the central nervous system [HP:0100835]	6/27 (22.2 %)	13.7	1.56×10^{-5}	0.00349	10.9	0.0229
KCNT2 Missense	27	39 (11)	4 (2)	0.793	0.450	Intracranial epidermoid cyst [HP:0012096]	2/27 (7.41 %)	198	9.21×10^{-5}	0.0164	186	0.286
CUX2 Missense	47	85 (28)	2 (1)	0.341	0.0875	Abnormality of the head [HP:0000234]	8/47 (17.0 %)	5.68	2.15×10^{-4}	0.0437	4.88	0.0179

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
MECP2 Missense	26	52 (21)	1 (1)	0.596	0.0192	Neurodevelopmental delay [HP:0012758]	8/26 (30.8 %)	6.99	9.90×10^{-5}	0.0162	5.15	0.0105
NTRK2 Missense	15	56 (21)	1 (1)	0.386	0.0436	Feeding difficulties in infancy [HP:0008872]	2/15 (13.3 %)	76.9	4.56×10^{-4}	0.0437	67.2	0.0741
PLXNA1 Missense	56	33 (20)	2 (1)	0.243	0.765	Intellectual disability [HP:0001249]	18/56 (32.1 %)	3.33	1.03×10^{-4}	0.0207	2.58	0.0114
ALG13 Missense	18	29 (17)	1 (1)	0.636	0.349	EEG with focal spike waves [HP:0011197]	3/18 (16.7 %)	119	4.85×10^{-6}	7.50×10^{-4}	99.9	0.125
KCNQ2 Missense	23	30 (15)	1 (1)	0.0251	0.417	EEG with burst suppression [HP:0010851]	3/23 (13.0 %)	32.3	1.78×10^{-4}	0.0249	28.3	0.0492
KLHL20 Missense	18	37 (15)	1 (1)	0.0104	0.113	Neurodevelopmental abnormality [HP:0012759]	10/18 (55.6 %)	6.21	2.07×10^{-4}	0.0238	3.32	0.00472
AP1G1 Missense	21	54 (14)	1 (1)	0.156	0.966	Abnormal palate morphology [HP:0000174]	2/21 (9.52 %)	82.3	3.98×10^{-4}	0.0471	74.9	0.111
MED12 Missense	28	28 (13)	1 (1)	0.234	1.82×10^{-4}	*Focal-onset seizure [HP:0007359]*	*7/28 (25.0 %)*	*0.203*	*1.00×10^{-4}*	*0.0181*	*0.402*	*0.00440*
ATN1 Missense	32	15 (12)	1 (1)	0.239	0.570	EEG with focal slow activity [HP:0010843]	8/32 (25.0 %)	5.66	2.99×10^{-4}	0.0428	4.49	0.0113
BRAF Missense	7	26 (11)	1 (1)	0.128	0.661	Abnormality of skin morphology [HP:0011121]	2/7 (28.6 %)	82.1	5.08×10^{-4}	0.0295	59.0	0.0317
RORB Missense	18	15 (9)	1 (1)	0.166	0.380	EEG with photoparoxysmal response [HP:0010852]	5/18 (27.8 %)	10.8	2.94×10^{-4}	0.0341	8.06	0.0114
TUBB2B Missense (dMissense)	4	26 (9)	1 (1)	0.404	0.949	Lissencephaly [HP:0001339]	2/4 (50.0 %)	230	1.16×10^{-4}	0.00625	117	0.0357
DCX Missense (dMissense)	7	13 (8)	4 (1)	0.251	0.874	Subcortical band heterotopia [HP:0032409]	2/7 (28.6 %)	208	8.53×10^{-5}	0.00735	150	0.0769

Table 5-14. A summary of the associations of carriers of missense variants in dominant epilepsy genes. Commenced 2 pages previously.

5.5.8 Phenotypic similarity and associations of clinically validated recessive genes

I expected URVs in dominant epilepsy genes to have the greatest effects on the clinical phenotypes a person manifests because sometimes these URVs are sufficient to cause a monogenic epilepsy. However, some monogenic epilepsies occur because of biallelic variants, meaning that a single URV in one of these genes is not thought sufficient to *cause* epilepsy. However, the relevance of the gene to epilepsy means that a single variant might still have detectable phenotypic consequences. I hypothesize that it might shape the features of an epilepsy that the person has developed for reasons (at least in part) beyond that variant itself. **Table 5-15** shows the associations that summarize the clinical features characterizing groups of carriers of PAV, dPAV, PTV, dMissense, and missense variants in non-dominant epilepsy genes. Associations of carriers of PTVs in *POLG* are shown in **Figure 5-28** and **Figure 5-29**, and of carriers of PAVs in *PIGG* in **Figure 5-30** and **Figure 5-31**.

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
PIGG PAV	39	35 (12)	3 (3)	0.185	0.606	EEG with burst suppression [HP:0010851]	4/39 (10.3 %)	25.1	3.59 x 10-5	0.00801	22.6	0.0656
						Abnormal emotion/affect behavior [HP:0100851]	3/39 (7.69 %)	32.6	1.67 x 10-4	0.0313	30.2	0.0857
						Neonatal electro-clinical motor seizure [HP:0032813]	3/39 (7.69 %)	28.2	2.50 x 10-4	0.0401	26.1	0.0750
PIGG Missense	27	47 (20)	5 (3)	0.489	0.825	EEG with burst suppression [HP:0010851]	4/27 (14.8 %)	38.1	8.00 x 10-6	0.00160	32.7	0.0656
						Abnormal emotion/affect behavior [HP:0100851]	3/27 (11.1 %)	48.8	5.47 x 10-5	0.0103	43.7	0.0857
						Neonatal electro-clinical motor seizure [HP:0032813]	3/27 (11.1 %)	42.3	8.20 x 10-5	0.0140	37.8	0.0750
POLG PTV	10	14 (6)	5 (3)	0.00422	2.80 x 10-5	EEG with generalized epileptiform discharges [HP:0011198]	10/10 (100 %)	Inf	3.72 x 10-5	0.00454	2.77	0.00220
						Generalized-onset motor seizure [HP:0032677]	9/10 (90.0 %)	24.5	5.54 x 10-5	0.00594	3.35	0.00265
						Focal-onset seizure [HP:0007359]	*0/10 (0.00 %)*	*0.00*	*6.05 x 10-5*	*0.00661*	*0.00*	*0.00209*
HERC2 Missense	114	30 (15)	1 (1)	0.484	0.0507	Generalized-onset seizure [HP:0002197]	59/114 (51.8 %)	2.02	2.26 x 10-4	0.0451	1.49	0.0134
HERC2 dMissense	88	37 (18)	2 (2)	0.184	0.164	EEG with spike-wave complexes [HP:0010850]	33/88 (37.5 %)	2.39	1.57 x 10-4	0.0405	1.87	0.0129
						Generalized-onset seizure [HP:0002197]	48/88 (54.5 %)	2.26	1.74 x 10-4	0.0439	1.57	0.0109
UFM1 PTV	2	70 (16)	2 (2)	0.0962	0.521	Acrocyanosis [HP:0001063]	1/2 (50.0 %)	Inf	1.59 x 10-4	0.0448	Inf	1.00
						Coarctation of aorta [HP:0001680]	1/2 (50.0 %)	Inf	1.59 x 10-4	0.0448	Inf	1.00

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
ABAT PTV	2	17 (5)	4 (2)	0.895	0.800	Abnormal glucose-6-phosphate dehydrogenase level [HP:0410176]	1/2 (50.0 %)	Inf	1.59×10^{-4}	0.0448	Inf	1.00
						Thyroid carcinoma [HP:0002890]	1/2 (50.0 %)	Inf	1.59×10^{-4}	0.0448	Inf	1.00
CACNA2D2 PAV	43	95 (39)	3 (1)	0.626	0.693	Visual impairment [HP:0000505]	3/43 (6.98 %)	30.3	2.05×10^{-4}	0.0377	28.3	0.0882
CACNA2D2 PTV	7	53 (18)	2 (1)	0.0210	0.745	Visual impairment [HP:0000505]	2/7 (28.6 %)	156	1.47×10^{-4}	0.0124	113	0.0588
EIF2B5 PAV	21	56 (23)	1 (1)	0.468	0.283	Growth delay [HP:0001510]	2/21 (9.52 %)	324	3.95×10^{-5}	0.00762	300	0.333
EIF2B5 *d*PAV	16	66 (28)	1 (1)	0.139	0.434	Growth delay [HP:0001510]	2/16 (12.5 %)	437	2.26×10^{-5}	0.00405	394	0.333
EIF2B5 Missense	18	60 (26)	1 (1)	0.576	0.186	Growth delay [HP:0001510]	2/18 (11.1 %)	383	2.88×10^{-5}	0.00529	350	0.333
EIF2B5 dMissense	13	68 (29)	1 (1)	0.0677	0.276	Growth delay [HP:0001510]	2/13 (15.4 %)	550	1.47×10^{-5}	0.00261	485	0.333
SZT2 PAV	115	48 (22)	2 (1)	0.0588	0.502	Focal cortical dysplasia [HP:0032046]	12/115 (10.4 %)	4.08	1.04×10^{-4}	0.0280	3.76	0.0334
FOXRED1 PAV	26	49 (21)	1 (1)	0.234	0.885	Upper motor neuron dysfunction [HP:0002493]	4/26 (15.4 %)	15.6	2.29×10^{-4}	0.0316	13.4	0.0268
FOXRED1 *d*PAV	17	57 (22)	1 (1)	0.0315	0.891	Localized skin lesion [HP:0011355]	2/17 (11.8 %)	79.4	4.26×10^{-4}	0.0437	70.6	0.0870
FOXRED1 Missense	18	43 (22)	1 (1)	0.297	0.941	Localized skin lesion [HP:0011355]	2/18 (11.1 %)	74.4	4.78×10^{-4}	0.0492	66.6	0.0870
FOXRED1 dMissense	10	44 (21)	1 (1)	0.262	0.871	Localized skin lesion [HP:0011355]	2/10 (20.0 %)	148	1.42×10^{-4}	0.0137	120	0.0870
TRAK1 PAV	30	28 (14)	1 (1)	0.604	0.577	Growth delay [HP:0001510]	2/30 (6.67 %)	223	8.16×10^{-5}	0.0163	210	0.333
TRAK1 Missense	21	46 (19)	1 (1)	0.575	0.237	Growth delay [HP:0001510]	2/21 (9.52 %)	324	3.95×10^{-5}	0.00762	300	0.333

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
PPIL1 PTV	7	29 (17)	1 (1)	0.0770	0.855	EEG with burst suppression [HP:0010851]	2/7 (28.6 %)	84.6	4.76×10^{-4}	0.0275	61.0	0.0328
ARV1 PAV	8	57 (12)	1 (1)	8.99×10^{-4}	0.947	Neurodevelopmental abnormality [HP:0012759]	6/8 (75.0 %)	14.9	4.60×10^{-4}	0.0277	4.47	0.00283
ARV1 *d*PAV	6	48 (15)	2 (1)	0.00609	0.682	Abnormality of skin morphology [HP:0011121]	2/6 (33.3 %)	102	3.64×10^{-4}	0.0205	68.9	0.0317
PARS2 Missense	18	53 (15)	1 (1)	0.297	0.766	Neurodevelopmental delay [HP:0012758]	6/18 (33.3 %)	7.85	4.66×10^{-4}	0.0471	5.57	0.00789
KCTD7 PAV (dPAV)	9	15 (12)	2 (1)	0.823	0.654	Developmental stagnation at onset of seizures [HP:0006834]	2/9 (22.2 %)	169	1.14×10^{-4}	0.0107	133	0.0870
KCTD7 Missense (dMissense)	7	17 (14)	2 (1)	0.535	0.537	Developmental stagnation at onset of seizures [HP:0006834]	2/7 (28.6 %)	237	6.64×10^{-5}	0.00620	171	0.0870
SLC13A5 *d*PAV	14	25 (13)	3 (1)	0.906	0.851	Hearing impairment [HP:0000365]	2/14 (14.3 %)	77.1	4.57×10^{-4}	0.0424	66.7	0.0690
SLC13A5 PTV	1	26 (5)	1 (1)	-	-	Bilateral sensorineural hearing impairment [HP:0008619]	1/1 (100 %)	Inf	1.59×10^{-4}	0.0330	12600	0.500
PPFIBP1 *d*PAV	25	31 (13)	2 (1)	0.587	0.756	Neonatal respiratory distress [HP:0002643]	3/25 (12.0 %)	25.0	3.60×10^{-4}	0.0459	22.2	0.0423
DOCK7 PAV	70	33 (10)	1 (1)	0.662	0.150	Posterior fossa cyst [HP:0007291]	3/70 (4.29 %)	46.6	7.10×10^{-5}	0.0142	44.8	0.200
DOCK7 *d*PAV	59	36 (10)	1 (1)	0.752	0.317	Posterior fossa cyst [HP:0007291]	3/59 (5.08 %)	55.8	4.25×10^{-5}	0.00928	53.2	0.200
DOCK7 Missense	55	37 (10)	1 (1)	0.645	0.193	Posterior fossa cyst [HP:0007291]	3/55 (5.45 %)	60.1	3.44×10^{-5}	0.00760	57.1	0.200
DOCK7 dMissense	48	41 (12)	3 (1)	0.726	0.208	Posterior fossa cyst [HP:0007291]	3/48 (6.25 %)	69.4	2.28×10^{-5}	0.00533	65.4	0.200

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
CARS2 PAV	17	30 (12)	1 (1)	0.0474	0.00542	EEG with spike-wave complexes (>3.5 Hz) [HP:0010849]	7/17 (41.2 %)	8.36	1.63×10^{-4}	0.0198	5.33	0.00714
CARS2 Missense	10	13 (8)	1 (1)	0.0432	0.0472	EEG with spike-wave complexes (>3.5 Hz) [HP:0010849]	6/10 (60.0 %)	17.9	3.47×10^{-5}	0.00395	7.76	0.00612
PAH Missense	16	29 (12)	1 (1)	0.00352	0.549	Abnormality of the cerebral subcortex [HP:0010993]	5/16 (31.2 %)	11.2	2.86×10^{-4}	0.0313	7.98	0.0100
PAH dMissense	11	23 (10)	2 (1)	0.0764	0.409	Abnormality of the basal ganglia [HP:0002134]	2/11 (18.2 %)	103	2.77×10^{-4}	0.0269	84.8	0.0690
PIGB PAV	22	15 (9)	1 (1)	0.940	0.694	Dysplastic corpus callosum [HP:0006989]	2/22 (9.09 %)	114	2.24×10^{-4}	0.0295	104	0.154
PIGB Missense	17	20 (12)	1 (1)	0.271	0.567	Dysplastic corpus callosum [HP:0006989]	2/17 (11.8 %)	151	1.32×10^{-4}	0.0168	135	0.154
PIGB dMissense	8	10 (6)	1 (1)	0.239	0.822	Myoclonic absence seizure [HP:0011150]	2/8 (25.0 %)	73.1	5.92×10^{-4}	0.0352	55.3	0.0339
ADPRS dMissense	5	34 (11)	2 (1)	0.514	0.557	Abnormal blood vessel morphology [HP:0033353]	2/5 (40.0 %)	121	3.09×10^{-4}	0.0161	73.1	0.0282
RNASEH2A PTV	8	31 (11)	2 (1)	0.00851	0.274	Typical absence seizure [HP:0011147]	6/8 (75.0 %)	18.2	1.75×10^{-4}	0.0168	5.30	0.00335
CLN8 Missense	13	36 (11)	1 (1)	0.00424	0.270	Abnormality of the skull base [HP:0002693]	2/13 (15.4 %)	78.3	4.49×10^{-4}	0.0442	66.8	0.0645
WDR73 PAV	18	23 (11)	1 (1)	0.576	0.746	Intellectual disability, profound [HP:0002187]	3/18 (16.7 %)	24.7	4.06×10^{-4}	0.0437	20.8	0.0288
MMACHC dMissense	6	33 (10)	1 (1)	0.284	0.997	Hypsarrhythmia [HP:0002521]	3/6 (50.0 %)	33.6	4.54×10^{-4}	0.0231	17.4	0.00820
RARS2 Missense	15	18 (9)	1 (1)	0.967	0.0548	*EEG with focal epileptiform discharges [HP:0011185]*	*0/15 (0.00 %)*	*0.00*	*2.67×10^{-4}*	*0.0274*	*0.00*	*0.00210*
TSEN54 PAV	30	17 (9)	1 (1)	0.0121	0.00551	EEG with generalized epileptiform discharges [HP:0011198]	21/30 (70.0 %)	4.15	1.89×10^{-4}	0.0281	1.94	0.00461
TSEN54 Missense	20	16 (6)	1 (1)	0.0266	2.42×10^{-4}	EEG with generalized epileptiform discharges [HP:0011198]	16/20 (80.0 %)	7.10	7.57×10^{-5}	0.0108	2.22	0.00352

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
SURF1 *d*PAV	17	17 (8)	1 (1)	0.0118	0.999	Periventricular nodular heterotopia [HP:0032388]	3/17 (17.6 %)	26.2	3.50×10^{-4}	0.0379	21.8	0.0286
SURF1 Missense	14	17 (8)	1 (1)	0.00815	0.615	EEG with spike-wave complexes (>3.5 Hz) [HP:0010849]	6/14 (42.9 %)	8.95	3.78×10^{-4}	0.0342	5.54	0.00612
CPLX1 PAV	8	13 (8)	1 (1)	0.0370	0.731	Status epilepticus [HP:0002133]	3/8 (37.5 %)	29.1	4.34×10^{-4}	0.0259	18.6	0.0117
PLAA PAV	30	17 (6)	1 (1)	0.217	0.700	Neonatal asphyxia [HP:0012768]	3/30 (10.0 %)	30.9	1.95×10^{-4}	0.0318	28.0	0.0625
PLAA *d*PAV	28	18 (7)	1 (1)	0.221	0.641	Neonatal asphyxia [HP:0012768]	3/28 (10.7 %)	33.4	1.59×10^{-4}	0.0233	30.0	0.0625
ALG11 Missense	16	11 (7)	1 (1)	0.0584	0.104	Intellectual disability, severe [HP:0010864]	4/16 (25.0 %)	14.8	3.60×10^{-4}	0.0360	11.3	0.0142
ALG11 dMissense	14	10 (5)	1 (1)	0.0286	0.147	Intellectual disability, severe [HP:0010864]	4/14 (28.6 %)	17.7	2.05×10^{-4}	0.0226	12.9	0.0142
MTHFS *d*PAV (dMissense)	1	9 (6)	1 (1)	-	-	EEG with frontal sharp slow waves [HP:0011290]	1/1 (100 %)	Inf	2.38×10^{-4}	0.0447	6300	0.333
UGP2 Missense (dMissense)	8	16 (6)	1 (1)	0.0883	0.111	Gray matter heterotopia [HP:0002282]	3/8 (37.5 %)	37.3	2.14×10^{-4}	0.0189	23.7	0.0149
FBXL4 *d*PAV	17	8 (5)	1 (1)	0.726	0.227	Complex febrile seizure [HP:0011172]	2/17 (11.8 %)	98.0	2.89×10^{-4}	0.0343	87.2	0.105
FBXL4 dMissense	14	8 (3)	1 (1)	0.760	0.257	Complex febrile seizure [HP:0011172]	2/14 (14.3 %)	122	1.94×10^{-4}	0.0220	106	0.105
ALKBH8 PTV	9	24 (5)	1 (1)	0.816	0.881	Abnormality of limbs [HP:0040064]	2/9 (22.2 %)	72.9	5.67×10^{-4}	0.0366	57.2	0.0392
APC2 PTV	2	21 (5)	1 (1)	0.284	0.241	Inferior vermis hypoplasia [HP:0007068]	1/2 (50.0 %)	Inf	1.59×10^{-4}	0.0448	Inf	1.00
PIGH PTV	1	10 (5)	1 (1)	-	-	Abnormal midbrain morphology [HP:0002418]	1/1 (100 %)	Inf	2.38×10^{-4}	0.0447	6300	0.333

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
PCDH12 dMissense	24	10 (5)	1 (1)	0.807	0.218	Polymicrogyria [HP:0002126]	4/24 (16.7 %)	16.1	2.07×10^{-4}	0.0290	13.6	0.0253
ASAH1 Missense	22	15 (4)	2 (1)	0.940	0.125	Cerebral calcification [HP:0002514]	2/22 (9.09 %)	96.1	3.01×10^{-4}	0.0379	88.0	0.133

Table 5-15. A summary of the associations of carriers of PAV, dPAV, dMissense, and missense, variants in non-dominant epilepsy genes. Commenced 5 pages previously.

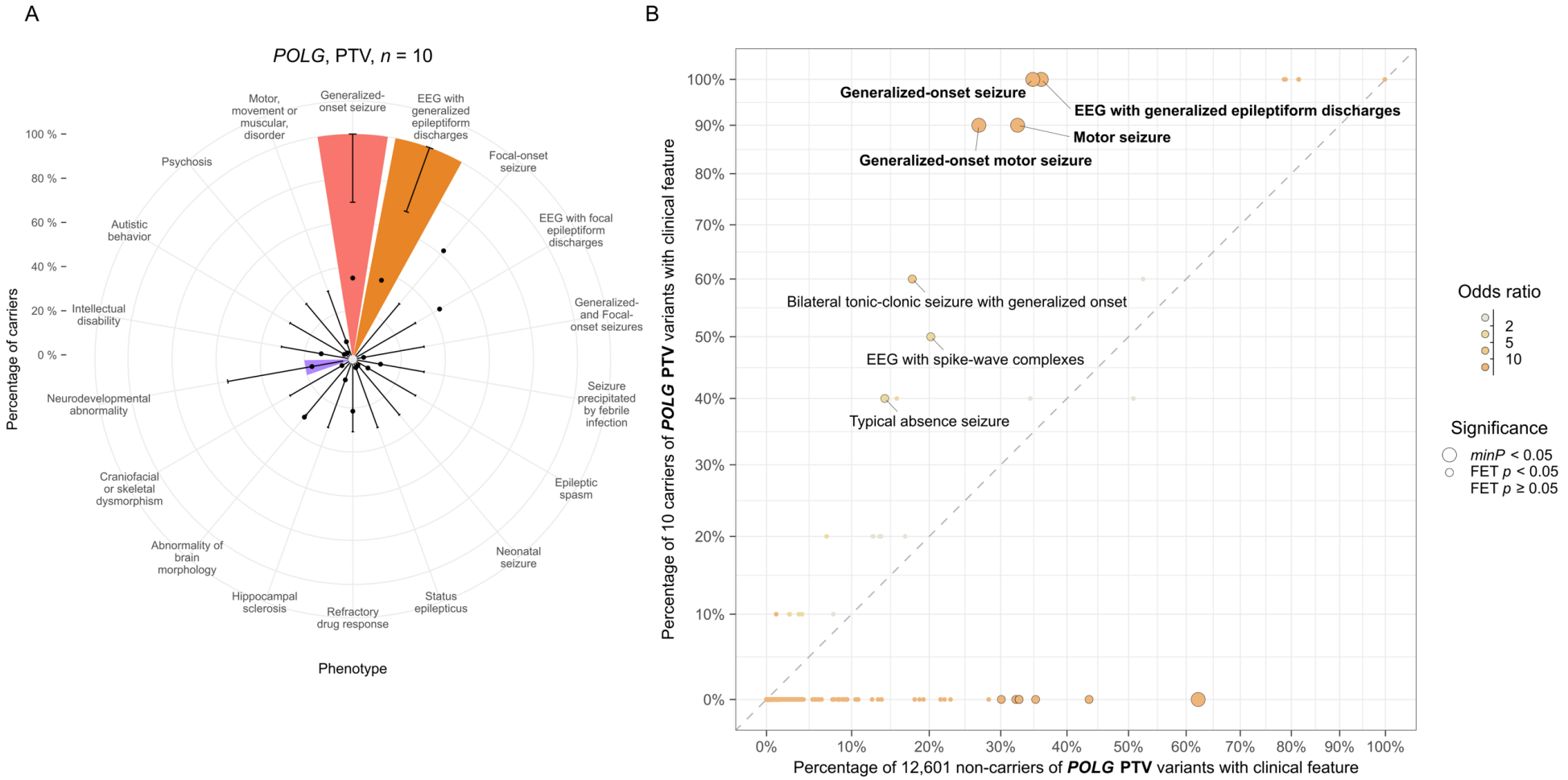


Figure 5-28. The associations of carriers of PTVs in *POLG*.

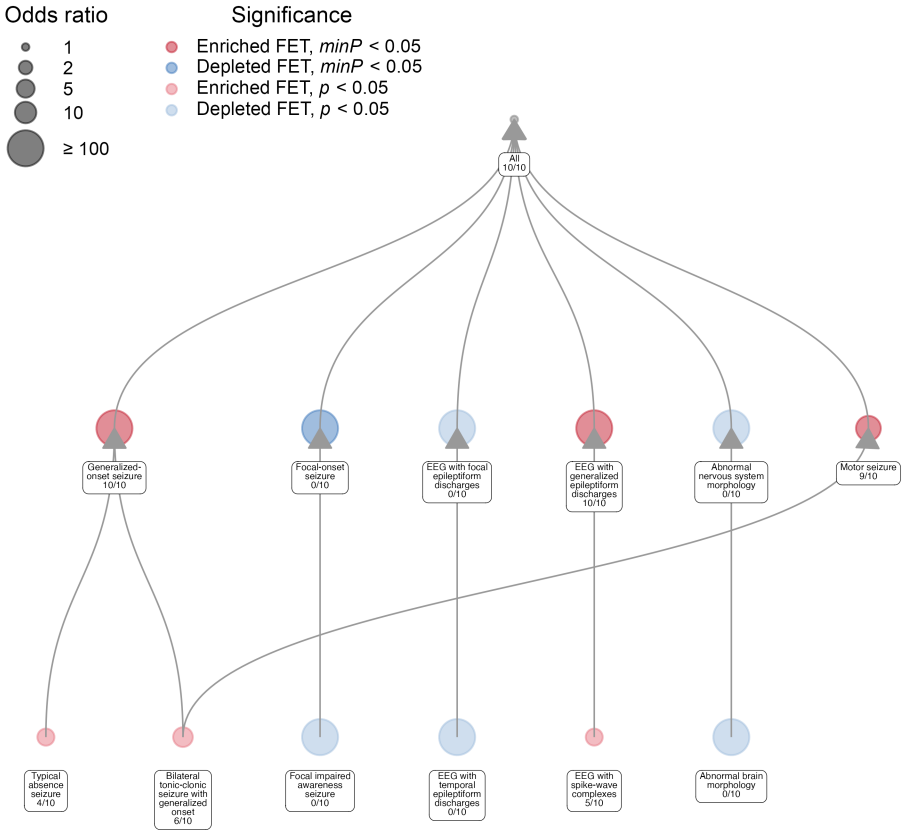


Figure 5-29. The associations of carriers of PAVs in *POLG* shown conceptually.

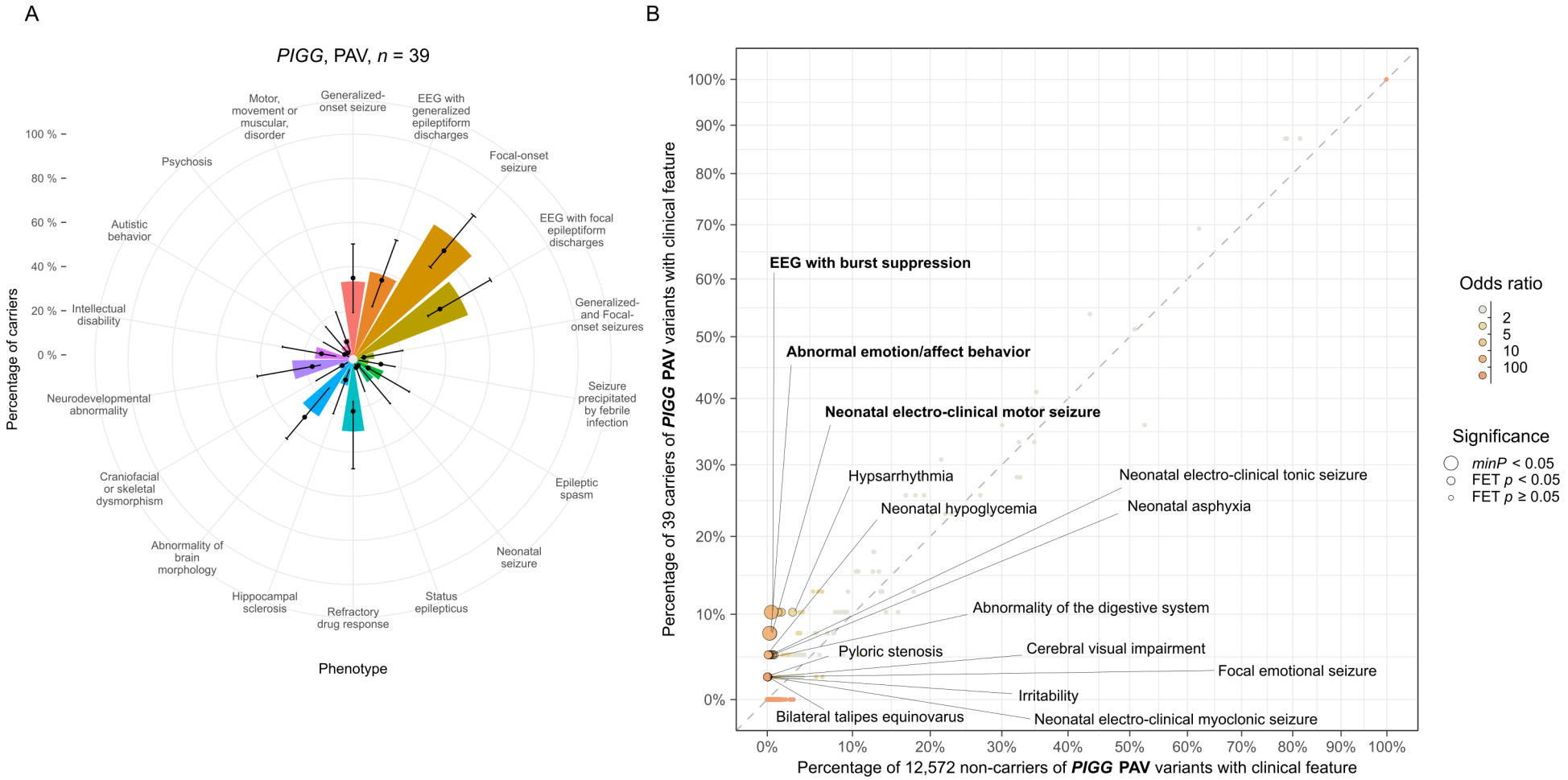


Figure 5-30. The associations of carriers of PAVs in *PIGG*.

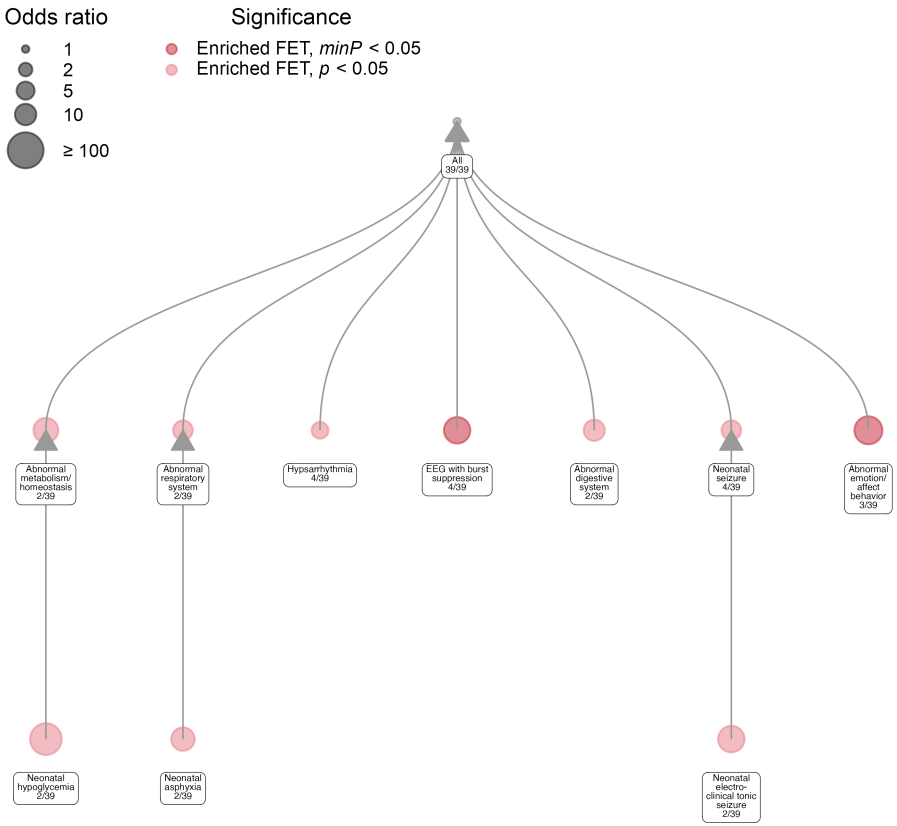


Figure 5-31. The associations of carriers of PAVs in *PIGG* shown conceptually.

5.5.9 Phenotypic similarity and associations of clinically validated neurodevelopmental genes

Monogenic disorders characterized predominantly by developmental, cognitive, behavioral, and neurological phenotypes beyond seizures sometimes cause comorbid seizures. Consequently, I assessed the phenotypic similarity and associations of these genes as URVs in these genes might influence the developing brain and how it expresses abnormal neurophysiological activity. **Table 5-16** shows the associations that summarize the clinical features characterizing groups of carriers of PAV, dPAV, PTV, dMissense, and missense variants in neurodevelopmental genes beyond the GE Epilepsy gene set. These include the ClinGen epilepsy, autism, and intellectual disability, and the GE intellectual disability 'green' gene sets. The associations of carriers of dPAVs in *GMMPA* are shown in **Figure 5-32** and **Figure 5-33**, and of carriers of PAVs in *LZTR1* in **Figure 5-34** and **Figure 5-35**.

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
ZNF699 PAV	18	33 (20)	4 (4)	0.0465	0.330	Delayed fine motor development [HP:0010862]	2/18 (11.1 %)	104	2.58×10^{-4}	0.0302	93.3	0.118
						Delayed gross motor development [HP:0002194]	2/18 (11.1 %)	91.8	3.24×10^{-4}	0.0386	82.3	0.105
						Impaired executive functioning [HP:0033051]	2/18 (11.1 %)	86.9	3.60×10^{-4}	0.0404	77.7	0.100
						Developmental stagnation at onset of seizures [HP:0006834]	2/18 (11.1 %)	74.4	4.78×10^{-4}	0.0492	66.6	0.0870
ZNF699 Missense	13	24 (13)	2 (2)	0.0756	0.289	Delayed fine motor development [HP:0010862]	2/13 (15.4 %)	151	1.32×10^{-4}	0.0151	129	0.118
						Delayed gross motor development [HP:0002194]	2/13 (15.4 %)	134	1.66×10^{-4}	0.0183	114	0.105
MYO5A dPAV	48	77 (29)	5 (3)	0.0592	0.772	Intellectual disability [HP:0001249]	16/48 (33.3 %)	3.51	1.52×10^{-4}	0.0315	2.67	0.0101
						Epileptic encephalopathy [HP:0200134]	16/48 (33.3 %)	3.48	1.67×10^{-4}	0.0336	2.65	0.0100
						Abnormal muscle tone [HP:0003808]	8/48 (16.7 %)	5.65	2.18×10^{-4}	0.0455	4.88	0.0183

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
KMT2B PAV	71	107 (41)	1 (1)	0.0105	0.951	Coarse facial features [HP:0000280]	3/71 (4.23 %)	273	1.70×10^{-6}	3.50×10^{-4}	265	0.600
KMT2B dPAV	41	74 (29)	3 (2)	0.0791	0.878	Hypertrichosis [HP:0000998]	2/41 (4.88 %)	160	1.53×10^{-4}	0.0325	153	0.333
						Coarse facial features [HP:0000280]	3/41 (7.32 %)	487	3.18×10^{-7}	8.00×10^{-5}	460	0.600
KMT2B Missense	56	83 (33)	1 (1)	0.0204	0.932	Coarse facial features [HP:0000280]	3/56 (5.36 %)	350	8.24×10^{-7}	1.60×10^{-4}	336	0.600
KMT2B dMissense	33	68 (26)	6 (3)	0.0480	0.871	Hypertrichosis [HP:0000998]	2/33 (6.06 %)	201	9.90×10^{-5}	0.0197	191	0.333
						Abnormal cerebral white matter morphology [HP:0002500]	7/33 (21.2 %)	6.99	1.83×10^{-4}	0.0284	5.73	0.0148
						Coarse facial features [HP:0000280]	3/33 (9.09 %)	622	1.63×10^{-7}	4.00×10^{-5}	572	0.600
KDM5C PAV	16	39 (18)	6 (3)	1.97×10^{-4}	0.0785	Epileptic encephalopathy [HP:0200134]	9/16 (56.2 %)	8.93	4.00×10^{-5}	0.00588	4.47	0.00565
						Neurodevelopmental delay [HP:0012758]	6/16 (37.5 %)	9.42	2.23×10^{-4}	0.0248	6.26	0.00789
						Intellectual disability [HP:0001249]	8/16 (50.0 %)	7.00	3.05×10^{-4}	0.0327	4.00	0.00506
KDM5C Missense	13	42 (17)	3 (2)	3.58×10^{-4}	0.184	Neurodevelopmental abnormality [HP:0012759]	8/13 (61.5 %)	7.95	3.63×10^{-4}	0.0321	3.67	0.00378
						Epileptic encephalopathy [HP:0200134]	7/13 (53.8 %)	8.09	4.36×10^{-4}	0.0413	4.27	0.00439

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
TRAPPC9 PAV	40	37 (15)	3 (2)	0.269	0.413	Epileptic encephalopathy [HP:0200134]	14/40 (35.0 %)	3.75	2.36×10^{-4}	0.0411	2.78	0.00878
						Developmental regression [HP:0002376]	9/40 (22.5 %)	5.04	2.55×10^{-4}	0.0422	4.14	0.0130
TRAPPC9 Missense	30	38 (16)	6 (3)	0.217	0.591	Developmental regression [HP:0002376]	9/30 (30.0 %)	7.45	2.18×10^{-5}	0.00461	5.52	0.0130
						Intellectual disability [HP:0001249]	12/30 (40.0 %)	4.67	1.43×10^{-4}	0.0230	3.21	0.00759
						Epileptic encephalopathy [HP:0200134]	12/30 (40.0 %)	4.63	1.53×10^{-4}	0.0251	3.18	0.00753
GMPPA PAV	14	23 (11)	4 (2)	0.00106	0.338	Status epilepticus [HP:0002133]	4/14 (28.6 %)	19.5	1.44×10^{-4}	0.0160	14.2	0.0156
						Intellectual disability, moderate [HP:0002342]	4/14 (28.6 %)	18.7	1.69×10^{-4}	0.0189	13.6	0.0149
GMPPA dPAV	13	25 (11)	6 (3)	2.40×10^{-4}	0.0958	Status epilepticus [HP:0002133]	4/13 (30.8 %)	21.7	1.04×10^{-4}	0.0120	15.3	0.0156
						Intellectual disability, moderate [HP:0002342]	4/13 (30.8 %)	20.7	1.23×10^{-4}	0.0138	14.7	0.0149
						Refractory drug response [HP:0020174]	9/13 (69.2 %)	8.23	2.97×10^{-4}	0.0272	3.23	0.00332
GMPPA PTV	7	22 (9)	2 (1)	7.74×10^{-4}	0.0544	Intellectual disability, moderate [HP:0002342]	3/7 (42.9 %)	34.9	3.12×10^{-4}	0.0194	20.4	0.0112

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
SERAC1 PAV	25	120 (42)	4 (2)	0.0347	0.682	Abnormal muscle physiology [HP:0011804]	6/25 (24.0 %)	8.42	2.26×10^{-4}	0.0314	6.64	0.0130
						Cerebellar atrophy [HP:0001272]	3/25 (12.0 %)	25.8	3.31×10^{-4}	0.0411	22.9	0.0435
SERAC1 dPAV	14	95 (35)	2 (1)	4.75×10^{-4}	0.595	Cerebellar atrophy [HP:0001272]	3/14 (21.4 %)	51.6	5.47×10^{-5}	0.00669	40.9	0.0435
SERAC1 Missense	19	111 (39)	1 (1)	0.0601	0.737	Abnormality of the musculoskeletal system [HP:0033127]	6/19 (31.6 %)	7.76	4.60×10^{-4}	0.0488	5.62	0.00842
SERAC1 dMissense	8	88 (27)	1 (1)	0.00494	0.457	Intellectual disability, moderate [HP:0002342]	3/8 (37.5 %)	27.9	4.91×10^{-4}	0.0298	17.8	0.0112
RBSN PAV	32	67 (31)	3 (2)	0.00950	0.0142	Neurodevelopmental abnormality [HP:0012759]	16/32 (50.0 %)	4.98	1.52×10^{-5}	0.00274	2.99	0.00755
						Epileptic encephalopathy [HP:0200134]	12/32 (37.5 %)	4.17	3.14×10^{-4}	0.0445	2.98	0.00753
RBSN Missense	26	80 (33)	5 (2)	2.00×10^{-6}	6.34×10^{-4}	Epileptic encephalopathy [HP:0200134]	12/26 (46.2 %)	5.96	2.79×10^{-5}	0.00610	3.67	0.00753
						Intellectual disability [HP:0001249]	11/26 (42.3 %)	5.14	1.49×10^{-4}	0.0224	3.39	0.00695
						Neurodevelopmental abnormality [HP:0012759]	10/15 (66.7 %)	9.94	2.32×10^{-5}	0.00373	3.98	0.00472
						Abnormality of higher mental function [HP:0011446]	8/15 (53.3 %)	7.48	2.60×10^{-4}	0.0269	4.02	0.00477

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
LZTR1 PAV	57	64 (29)	3 (2)	8.00×10^{-6}	0.591	Epileptic encephalopathy [HP:0200134]	18/57 (31.6 %)	3.21	1.46×10^{-4}	0.0324	2.52	0.0113
						EEG with frontal sharp waves [HP:0011294]	2/57 (3.51 %)	151	1.99×10^{-4}	0.0447	147	0.400
LZTR1 dPAV	53	68 (31)	3 (2)	1.50×10^{-5}	0.608	EEG with frontal sharp waves [HP:0011294]	2/53 (3.77 %)	163	1.72×10^{-4}	0.0385	158	0.400
						Epileptic encephalopathy [HP:0200134]	17/53 (32.1 %)	3.29	1.79×10^{-4}	0.0387	2.55	0.0107
LZTR1 PTV	24	63 (26)	5 (2)	9.79×10^{-4}	0.606	Abnormal muscle tone [HP:0003808]	6/24 (25.0 %)	9.40	1.32×10^{-4}	0.0199	7.30	0.0137
						Tonic seizure [HP:0032792]	6/24 (25.0 %)	7.94	3.17×10^{-4}	0.0412	6.21	0.0117
NFASC PAV	36	84 (27)	4 (2)	0.0408	0.204	Aplasia/Hypoplasia of the optic nerve [HP:0008058]	2/36 (5.56 %)	363	4.74×10^{-5}	0.0118	349	0.500
						Abnormality of the thyroid gland [HP:0000820]	2/36 (5.56 %)	105	2.82×10^{-4}	0.0449	99.8	0.222
NFASC dPAV	29	71 (25)	1 (1)	0.310	0.556	Abnormality of the thyroid gland [HP:0000820]	2/29 (6.90 %)	132	1.82×10^{-4}	0.0288	124	0.222
NFASC PTV	7	46 (13)	1 (1)	0.0133	0.330	Autistic behavior [HP:0000729]	3/7 (42.9 %)	30.0	4.78×10^{-4}	0.0280	17.6	0.00968
NFASC Missense	25	59 (23)	4 (1)	0.443	0.364	Aplasia/Hypoplasia of the optic nerve [HP:0008058]	2/25 (8.00 %)	543	2.26×10^{-5}	0.00539	503	0.500
TENM3 dPAV	93	66 (23)	1 (1)	0.00389	0.877	Hyperintensity of cerebral white matter on MRI [HP:0030890]	6/93 (6.45 %)	8.07	1.75×10^{-4}	0.0464	7.62	0.0536
TENM3 Missense	89	70 (25)	4 (2)	0.00512	0.941	Hyperintensity of cerebral white matter on MRI [HP:0030890]	6/89 (6.74 %)	8.46	1.37×10^{-4}	0.0265	7.96	0.0536
						EEG abnormality [HP:0002353]	85/89 (95.5 %)	4.86	1.67×10^{-4}	0.0420	1.17	0.00827
TENM3 dMissense	78	74 (26)	1 (1)	0.00769	0.950	Hyperintensity of cerebral white matter on MRI [HP:0030890]	6/78 (7.69 %)	9.77	6.55×10^{-5}	0.0146	9.10	0.0536

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
TPP2 PAV	42	70 (24)	2 (2)	0.301	0.136	*Focal impaired awareness seizure [HP:0002384]*	*4/42 (9.52 %)*	*0.193*	*2.56 x 10^{-4}*	*0.0444*	*0.270*	*0.00465*
						EEG with spike-wave complexes [HP:0010850]	19/42 (45.2 %)	3.28	2.91 x 10^{-4}	0.0478	2.25	0.00745
CLN6 PAV	13	84 (22)	2 (2)	0.00951	0.402	Intellectual disability [HP:0001249]	7/13 (53.8 %)	8.16	4.15 x 10^{-4}	0.0385	4.31	0.00442
						Abnormality of movement [HP:0100022]	6/13 (46.2 %)	27.0	1.28 x 10^{-6}	2.70 x 10^{-4}	15.0	0.0152
CLN6 dPAV	9	88 (24)	1 (1)	0.0228	0.412	Abnormality of movement [HP:0100022]	4/9 (44.4 %)	25.0	1.04 x 10^{-4}	0.0102	14.4	0.0102
POGZ PAV	33	63 (20)	4 (2)	0.106	0.765	Tetraparesis [HP:0002273]	3/33 (9.09 %)	48.1	5.70 x 10^{-5}	0.0112	44.0	0.103
						Abnormal corpus callosum morphology [HP:0001273]	4/33 (12.1 %)	14.3	2.92 x 10^{-4}	0.0403	12.7	0.0323
POGZ dPAV	25	29 (8)	1 (1)	0.102	0.645	Abnormal corpus callosum morphology [HP:0001273]	4/25 (16.0 %)	19.8	9.60 x 10^{-5}	0.0152	16.8	0.0323
POGZ Missense	30	50 (18)	4 (2)	0.0825	0.924	Tetraparesis [HP:0002273]	3/30 (10.0 %)	53.5	4.26 x 10^{-5}	0.00886	48.4	0.103
						Abnormal corpus callosum morphology [HP:0001273]	4/30 (13.3 %)	16.0	2.00 x 10^{-4}	0.0319	14.0	0.0323
POGZ dMissense	22	27 (11)	1 (1)	0.119	0.657	Abnormal corpus callosum morphology [HP:0001273]	4/22 (18.2 %)	23.1	5.68 x 10^{-5}	0.00963	19.1	0.0323

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
CDC6 PAV	20	50 (22)	3 (1)	0.256	0.579	Aplasia/Hypoplasia of the corpus callosum [HP:0007370]	3/20 (15.0 %)	36.1	1.34×10^{-4}	0.0184	31.0	0.0469
CDC6 dPAV	9	61 (19)	9 (2)	0.0143	0.140	Epileptic encephalopathy [HP:0200134]	6/9 (66.7 %)	13.9	2.41×10^{-4}	0.0223	5.29	0.00376
						Hypoplasia of the corpus callosum [HP:0002079]	2/9 (22.2 %)	71.4	5.89×10^{-4}	0.0373	56.0	0.0385
CDC6 PTV	3	44 (12)	1 (1)	0.0745	0.201	Thin corpus callosum [HP:0033725]	2/3 (66.7 %)	335	1.07×10^{-4}	0.00481	114	0.0263
CDC6 dMissense	6	41 (13)	2 (1)	0.0564	0.477	Hypoplasia of the corpus callosum [HP:0002079]	2/6 (33.3 %)	125	2.48×10^{-4}	0.0145	84.0	0.0385
CNOT2 PAV	29	44 (19)	3 (2)	0.433	0.445	Global developmental delay [HP:0001263]	7/29 (24.1 %)	7.29	1.60×10^{-4}	0.0247	5.77	0.0131
						Epileptic spasm [HP:0011097]	8/29 (27.6 %)	6.11	2.04×10^{-4}	0.0304	4.70	0.0107
CNOT2 dPAV	29	44 (19)	3 (2)	0.433	0.445	Global developmental delay [HP:0001263]	7/29 (24.1 %)	7.29	1.60×10^{-4}	0.0247	5.77	0.0131
						Epileptic spasm [HP:0011097]	8/29 (27.6 %)	6.11	2.04×10^{-4}	0.0304	4.70	0.0107
CNOT2 Missense (dMissense)	21	28 (12)	2 (1)	0.754	0.628	Global developmental delay [HP:0001263]	6/21 (28.6 %)	9.15	1.75×10^{-4}	0.0244	6.83	0.0113
MANBA PAV	54	36 (20)	1 (1)	0.0359	0.922	Coarse facial features [HP:0000280]	2/54 (3.70 %)	160	1.78×10^{-4}	0.0398	155	0.400
MANBA dPAV	38	41 (19)	2 (2)	0.0580	0.970	Coarse facial features [HP:0000280]	2/38 (5.26 %)	230	8.79×10^{-5}	0.0198	221	0.400
						Hypertrichosis [HP:0000998]	2/38 (5.26 %)	172	1.32×10^{-4}	0.0278	165	0.333
MANBA PTV	16	25 (12)	1 (1)	0.154	0.988	EEG with generalized slow activity [HP:0010845]	5/16 (31.2 %)	13.0	1.42×10^{-4}	0.0167	9.28	0.0117

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
COASY Missense	16	43 (16)	2 (1)	0.00730	0.226	Generalized tonic seizure [HP:0010818]	5/16 (31.2 %)	12.0	2.08×10^{-4}	0.0239	8.56	0.0108
NDUFA2 PAV	8	24 (11)	2 (2)	0.00298	0.0286	Generalized myoclonic-atonic seizure [HP:0011170]	3/8 (37.5 %)	54.9	7.05×10^{-5}	0.00680	34.8	0.0216
						Neurodevelopmental abnormality [HP:0012759]	6/8 (75.0 %)	14.9	4.60×10^{-4}	0.0277	4.47	0.00283
NR2F1 PAV (dPAV)	9	23 (11)	2 (2)	0.00908	0.746	Infantile spasms [HP:0012469]	4/9 (44.4 %)	22.1	1.64×10^{-4}	0.0149	12.8	0.00903
						Refractory drug response [HP:0020174]	7/9 (77.8 %)	12.8	5.04×10^{-4}	0.0349	3.62	0.00258
NR2F1 Missense (dMissense)	8	22 (10)	1 (1)	0.0334	0.899	Infantile spasms [HP:0012469]	4/8 (50.0 %)	27.7	9.40×10^{-5}	0.00851	14.4	0.00903
PDZD8 PAV	36	69 (27)	2 (1)	0.0535	0.903	Spinal cord lesion [HP:0100561]	2/36 (5.56 %)	719	2.37×10^{-5}	0.00639	699	0.667
PDZD8 PTV	2	50 (9)	7 (2)	0.109	0.485	Asymmetry of the thorax [HP:0001555]	1/2 (50.0 %)	Inf	1.59×10^{-4}	0.0448	Inf	1.00
						Short fifth metatarsal [HP:0004704]	1/2 (50.0 %)	Inf	1.59×10^{-4}	0.0448	Inf	1.00
DAG1 PTV	1	24 (7)	8 (2)	-	-	Neonatal intestinal obstruction [HP:0005234]	1/1 (100 %)	Inf	7.93×10^{-5}	0.0200	Inf	1.00
						Tracheoesophageal fistula [HP:0002575]	1/1 (100 %)	Inf	7.93×10^{-5}	0.0200	Inf	1.00
CTCF dPAV	13	18 (5)	2 (1)	0.870	0.403	Abnormality of body weight [HP:0004323]	3/13 (23.1 %)	24.9	4.58×10^{-4}	0.0449	19.4	0.0196
CTCF dMissense	11	21 (7)	3 (2)	0.729	0.295	Abnormality of body weight [HP:0004323]	3/11 (27.3 %)	31.1	2.69×10^{-4}	0.0251	22.9	0.0196
						Abnormality of metabolism/homeostasis [HP:0001939]	2/11 (18.2 %)	77.2	4.78×10^{-4}	0.0414	63.6	0.0526

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
GRM1 PTV	1	19 (6)	2 (2)	-	-	Corpus callosum atrophy [HP:0007371]	1/1 (100 %)	Inf	1.59×10^{-4}	0.0330	12600	0.500
						Small pituitary gland [HP:0012506]	1/1 (100 %)	Inf	2.38×10^{-4}	0.0447	6300	0.333
CTR9 PTV	2	22 (5)	4 (2)	0.229	0.388	Decreased circulating IgA level [HP:0002720]	1/2 (50.0 %)	Inf	1.59×10^{-4}	0.0448	Inf	1.00
						Decreased circulating IgG level [HP:0004315]	1/2 (50.0 %)	Inf	1.59×10^{-4}	0.0448	Inf	1.00
TAFAZZIN PAV	4	22 (4)	6 (2)	0.00515	0.0138	Abnormal cerebral vascular morphology [HP:0100659]	3/4 (75.0 %)	82.8	1.67×10^{-4}	0.00813	21.5	0.00679
						Abnormal blood vessel morphology [HP:0033353]	2/4 (50.0 %)	180	1.86×10^{-4}	0.00914	91.4	0.0282
TAFAZZIN Missense	3	24 (5)	3 (1)	0.289	0.852	Abnormal blood vessel morphology [HP:0033353]	2/3 (66.7 %)	359	9.34×10^{-5}	0.00434	122	0.0282
NDST1 PAV	31	106 (35)	4 (1)	0.0949	0.235	High palate [HP:0000218]	2/31 (6.45 %)	123	2.08×10^{-4}	0.0328	116	0.222
NDST1 dPAV (dMissense)	24	68 (21)	1 (1)	0.0405	0.00850	EEG with polyspike wave complexes [HP:0002392]	8/24 (33.3 %)	6.72	1.46×10^{-4}	0.0211	4.81	0.00909
NDST1 Missense	31	106 (35)	4 (1)	0.0949	0.235	High palate [HP:0000218]	2/31 (6.45 %)	123	2.08×10^{-4}	0.0328	116	0.222
MCPH1 Missense	44	84 (34)	1 (1)	0.988	0.168	Hypertrichosis [HP:0000998]	2/44 (4.55 %)	148	1.77×10^{-4}	0.0373	143	0.333
SRCAP PAV	117	76 (33)	1 (1)	0.258	0.463	EEG with generalized slow activity [HP:0010845]	13/117 (11.1 %)	3.63	1.69×10^{-4}	0.0406	3.34	0.0303
CDK13 Missense	31	80 (29)	1 (1)	0.214	0.346	Behavioral abnormality [HP:0000708]	8/31 (25.8 %)	6.29	1.56×10^{-4}	0.0246	4.93	0.0120

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
PTCH1 PAV	34	71 (29)	1 (1)	0.0782	0.752	Status epilepticus with prominent motor symptoms [HP:0032658]	5/34 (14.7 %)	14.0	6.01×10^{-5}	0.0118	12.1	0.0316
PTCH1 Missense	25	71 (25)	2 (1)	0.0775	0.796	Convulsive status epilepticus [HP:0032660]	4/25 (16.0 %)	16.8	1.76×10^{-4}	0.0258	14.3	0.0276
PEPD PAV	20	66 (24)	2 (1)	0.00134	0.0630	Abnormality of the pituitary gland [HP:0012503]	2/20 (10.0 %)	99.1	2.83×10^{-4}	0.0344	89.9	0.125
PEPD dPAV	16	74 (28)	2 (1)	4.93×10^{-4}	0.00978	Abnormality of the pituitary gland [HP:0012503]	2/16 (12.5 %)	127	1.79×10^{-4}	0.0217	112	0.125
TCF20 PAV	79	58 (28)	1 (1)	0.0110	0.473	EEG with occipital sharp waves [HP:0011292]	2/79 (2.53 %)	Inf	3.87×10^{-5}	0.0122	Inf	1.00
TCF20 dPAV	52	68 (28)	1 (1)	0.147	0.287	EEG with occipital sharp waves [HP:0011292]	2/52 (3.85 %)	Inf	1.67×10^{-5}	0.00478	Inf	1.00
TCF20 Missense	65	57 (23)	1 (1)	0.0494	0.332	EEG with occipital sharp waves [HP:0011292]	2/65 (3.08 %)	Inf	2.62×10^{-5}	0.00738	Inf	1.00
TCF20 dMissense	50	64 (25)	1 (1)	0.154	0.294	EEG with occipital sharp waves [HP:0011292]	2/50 (4.00 %)	Inf	1.54×10^{-5}	0.00448	Inf	1.00
FAT4 PAV	190	64 (27)	1 (1)	0.272	0.446	Aplasia/Hypoplasia of the cerebrum [HP:0007364]	10/190 (5.26 %)	6.39	1.05×10^{-5}	0.00221	6.11	0.0855
FAT4 Missense	178	65 (28)	1 (1)	0.212	0.364	Aplasia/Hypoplasia of the cerebrum [HP:0007364]	10/178 (5.62 %)	6.85	5.91×10^{-6}	0.00138	6.53	0.0855
PI4KA Missense	57	54 (24)	1 (1)	0.746	0.163	Coarse facial features [HP:0000280]	2/57 (3.51 %)	151	1.99×10^{-4}	0.0447	147	0.400
KCNH5 PAV	22	78 (23)	2 (1)	0.328	0.694	Cerebellar vermis hypoplasia [HP:0001320]	2/22 (9.09 %)	89.4	3.44×10^{-4}	0.0433	81.7	0.125
KCNH5 Missense	22	78 (23)	2 (1)	0.328	0.694	Cerebellar vermis hypoplasia [HP:0001320]	2/22 (9.09 %)	89.4	3.44×10^{-4}	0.0433	81.7	0.125
GEMIN5 Missense	54	56 (23)	1 (1)	3.17×10^{-4}	0.995	Abnormality of skull size [HP:0000240]	7/54 (13.0 %)	6.43	2.27×10^{-4}	0.0456	5.73	0.0241

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
BCAS3 PAV	25	44 (22)	1 (1)	0.587	0.872	Triangular face [HP:0000325]	2/25 (8.00 %)	357	3.76×10^{-5}	0.00844	336	0.400
BCAS3 dPAV	19	50 (22)	1 (1)	0.578	0.508	Triangular face [HP:0000325]	2/19 (10.5 %)	481	2.14×10^{-5}	0.00438	442	0.400
BCAS3 Missense	18	38 (16)	1 (1)	0.846	0.835	Triangular face [HP:0000325]	2/18 (11.1 %)	507	1.92×10^{-5}	0.00370	466	0.400
BCAS3 dMissense	13	44 (16)	1 (1)	0.547	0.624	Triangular face [HP:0000325]	2/13 (15.4 %)	735	9.79×10^{-6}	0.00200	646	0.400
DOCK6 PAV	91	44 (22)	1 (1)	0.188	0.372	Focal clonic seizure [HP:0002266]	6/91 (6.59 %)	9.43	7.82×10^{-5}	0.0203	8.88	0.0606
DOCK6 dPAV	69	53 (21)	2 (1)	0.143	0.384	Focal clonic seizure [HP:0002266]	6/69 (8.70 %)	12.7	1.62×10^{-5}	0.00244	11.7	0.0606
DOCK6 Missense	59	47 (19)	2 (1)	0.239	0.459	Focal clonic seizure [HP:0002266]	6/59 (10.2 %)	15.2	6.47×10^{-6}	0.00107	13.7	0.0606
DOCK6 dMissense	43	55 (20)	3 (1)	0.266	0.367	Focal clonic seizure [HP:0002266]	6/43 (14.0 %)	21.7	9.68×10^{-7}	1.60×10^{-4}	18.9	0.0606
LRPPRC dPAV	47	44 (21)	1 (1)	0.162	0.391	Infantile spasms [HP:0012469]	8/47 (17.0 %)	5.72	2.05×10^{-4}	0.0436	4.92	0.0181
VPS53 Missense	22	49 (15)	2 (1)	0.766	0.738	Abnormality of the curvature of the vertebral column [HP:0010674]	2/22 (9.09 %)	104	2.61×10^{-4}	0.0336	95.4	0.143
VPS53 dMissense	20	57 (20)	2 (1)	0.971	0.691	Abnormality of the curvature of the vertebral column [HP:0010674]	2/20 (10.0 %)	116	2.15×10^{-4}	0.0267	105	0.143

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
GMPPB PAV	17	54 (17)	1 (1)	0.118	0.961	Growth delay [HP:0001510]	2/17 (11.8 %)	411	2.56×10^{-5}	0.00497	370	0.333
GMPPB dPAV	15	55 (18)	1 (1)	0.139	0.931	Growth delay [HP:0001510]	2/15 (13.3 %)	477	1.98×10^{-5}	0.00345	420	0.333
GMPPB Missense	15	50 (17)	1 (1)	0.0865	0.953	Growth delay [HP:0001510]	2/15 (13.3 %)	477	1.98×10^{-5}	0.00345	420	0.333
GMPPB dMissense	13	56 (20)	1 (1)	0.112	0.836	Growth delay [HP:0001510]	2/13 (15.4 %)	550	1.47×10^{-5}	0.00261	485	0.333
FMN2 Missense	50	39 (20)	1 (1)	0.488	0.790	EEG with frontal sharp waves [HP:0011294]	2/50 (4.00 %)	174	1.53×10^{-4}	0.0342	167	0.400
SOX9 PTV	4	56 (19)	4 (1)	0.00219	0.00106	Hypotonia [HP:0001252]	3/4 (75.0 %)	120	5.78×10^{-5}	0.00370	30.8	0.00968
IFT172 PAV	54	52 (17)	1 (1)	0.0537	0.881	Vascular skin abnormality [HP:0011276]	2/54 (3.70 %)	160	1.78×10^{-4}	0.0398	155	0.400
IFT172 dPAV	46	57 (19)	2 (1)	0.165	0.784	Vascular skin abnormality [HP:0011276]	2/46 (4.35 %)	188	1.29×10^{-4}	0.0286	182	0.400
IFT172 Missense	38	39 (19)	1 (1)	0.322	0.679	Vascular skin abnormality [HP:0011276]	2/38 (5.26 %)	230	8.79×10^{-5}	0.0198	221	0.400
IFT172 dMissense	31	42 (18)	1 (1)	0.214	0.449	Vascular skin abnormality [HP:0011276]	2/31 (6.45 %)	284	5.82×10^{-5}	0.0127	271	0.400
TTC5 Missense	12	45 (17)	2 (1)	0.0676	0.0747	EEG with photoparoxysmal response grade IV [HP:0011214]	4/12 (33.3 %)	18.3	2.12×10^{-4}	0.0198	12.6	0.0118
TTC5 dMissense	10	57 (19)	2 (1)	3.24×10^{-4}	0.0134	EEG with photoparoxysmal response grade IV [HP:0011214]	4/10 (40.0 %)	24.5	9.37×10^{-5}	0.00985	15.1	0.0118
DLGAP2 dMissense	30	64 (19)	1 (1)	0.967	0.0325	EEG with spike-wave complexes (>3.5 Hz) [HP:0010849]	9/30 (30.0 %)	5.12	3.20×10^{-4}	0.0445	3.89	0.00918
SHANK2 PAV	58	34 (19)	1 (1)	0.165	0.963	Tonic seizure [HP:0032792]	10/58 (17.2 %)	4.99	1.02×10^{-4}	0.0214	4.30	0.0195
SHANK2 Missense	51	30 (18)	1 (1)	0.152	0.970	Tonic seizure [HP:0032792]	9/51 (17.6 %)	5.12	1.88×10^{-4}	0.0372	4.40	0.0175

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
THUMPD1 PAV	21	57 (18)	2 (1)	0.933	0.151	Abnormal esophagus physiology [HP:0025270]	2/21 (9.52 %)	101	2.74×10^{-4}	0.0341	92.2	0.133
TAT PAV	15	41 (17)	1 (1)	0.139	0.380	Intellectual disability, moderate [HP:0002342]	4/15 (26.7 %)	17.0	2.26×10^{-4}	0.0260	12.7	0.0149
TAT Missense	14	41 (18)	1 (1)	0.0987	0.223	Intellectual disability, moderate [HP:0002342]	4/14 (28.6 %)	18.7	1.69×10^{-4}	0.0189	13.6	0.0149
LSS dMissense	21	33 (18)	1 (1)	0.164	0.404	Seizure cluster [HP:0033349]	2/21 (9.52 %)	131	1.73×10^{-4}	0.0244	120	0.167
SPG11 dMissense	67	32 (18)	1 (1)	0.271	0.749	Abnormal neck morphology [HP:0025668]	2/67 (2.99 %)	192	1.66×10^{-4}	0.0413	187	0.500
SRRM2 PAV	179	34 (17)	1 (1)	0.279	0.826	Focal autonomic seizure with epigastric sensation/nausea/vomiting/other gastrointestinal phenomena [HP:0011159]	3/179 (1.68 %)	42.3	1.49×10^{-4}	0.0304	41.7	0.375
SRRM2 dPAV	93	31 (17)	1 (1)	0.176	0.928	Focal autonomic seizure with epigastric sensation/nausea/vomiting/other gastrointestinal phenomena [HP:0011159]	3/93 (3.23 %)	83.1	2.12×10^{-5}	0.00344	80.8	0.375
SRRM2 Missense	135	31 (18)	1 (1)	0.305	0.704	Focal autonomic seizure with epigastric sensation/nausea/vomiting/other gastrointestinal phenomena [HP:0011159]	3/135 (2.22 %)	56.5	6.46×10^{-5}	0.0123	55.4	0.375
SRRM2 dMissense	86	34 (16)	1 (1)	0.122	0.963	Focal autonomic seizure with epigastric sensation/nausea/vomiting/other gastrointestinal phenomena [HP:0011159]	3/86 (3.49 %)	90.3	1.67×10^{-5}	0.00264	87.4	0.375
CLN5 Missense	11	39 (15)	2 (1)	0.729	0.704	Cerebellar hypoplasia [HP:0001321]	2/11 (18.2 %)	81.6	4.29×10^{-4}	0.0397	67.4	0.0556
ALDH3A2 dPAV	13	39 (16)	2 (1)	0.0821	0.212	Intellectual disability [HP:0001249]	7/13 (53.8 %)	8.16	4.15×10^{-4}	0.0385	4.31	0.00442
IL1RAPL1 dMissense	20	31 (16)	1 (1)	0.00202	0.0191	Generalized non-motor (absence) seizure [HP:0002121]	10/20 (50.0 %)	5.38	3.66×10^{-4}	0.0414	3.19	0.00504

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
RPGRIP1L PTV	15	28 (16)	1 (1)	0.568	0.181	Motor delay [HP:0001270]	3/15 (20.0 %)	56.8	4.04×10^{-5}	0.00565	45.8	0.0517
POMT2 PAV	25	49 (14)	1 (1)	0.0347	0.923	Pachygyria [HP:0001302]	2/25 (8.00 %)	90.7	3.38×10^{-4}	0.0447	83.9	0.143
POMT2 dPAV	16	45 (15)	1 (1)	0.0826	0.728	Abnormality of the digestive system [HP:0025031]	3/16 (18.8 %)	41.2	9.87×10^{-5}	0.0113	33.7	0.0411
POMT2 Missense	17	49 (14)	1 (1)	0.0692	0.986	Pachygyria [HP:0001302]	2/17 (11.8 %)	138	1.54×10^{-4}	0.0194	123	0.143
SIL1 dPAV	12	43 (15)	1 (1)	0.331	0.533	Infantile spasms [HP:0012469]	4/12 (33.3 %)	13.8	5.94×10^{-4}	0.0483	9.57	0.00903
MED13L PAV	66	40 (15)	1 (1)	0.0200	0.628	Focal impaired awareness sensory seizure with cephalic sensation [HP:0032878]	2/66 (3.03 %)	195	1.61×10^{-4}	0.0406	190	0.500
MED13L dPAV	56	39 (13)	1 (1)	0.0204	0.799	Focal impaired awareness sensory seizure with cephalic sensation [HP:0032878]	2/56 (3.57 %)	230	1.16×10^{-4}	0.0294	224	0.500
EZH2 PAV	23	37 (15)	1 (1)	0.657	0.799	Focal autonomic seizure with epigastric sensation/nausea/vomiting/other gastrointestinal phenomena [HP:0011159]	2/23 (8.70 %)	197	8.85×10^{-5}	0.0140	182	0.250
EZH2 dPAV	21	37 (15)	1 (1)	0.575	0.773	Focal autonomic seizure with epigastric sensation/nausea/vomiting/other gastrointestinal phenomena [HP:0011159]	2/21 (9.52 %)	218	7.35×10^{-5}	0.0115	200	0.250
EZH2 Missense (dMissense)	18	37 (13)	1 (1)	0.636	0.889	Focal autonomic seizure with epigastric sensation/nausea/vomiting/other gastrointestinal phenomena [HP:0011159]	2/18 (11.1 %)	259	5.36×10^{-5}	0.00828	233	0.250
TRRAP dPAV	96	35 (15)	1 (1)	0.485	0.0568	Abnormality of body weight [HP:0004323]	7/96 (7.29 %)	6.66	1.63×10^{-4}	0.0315	6.25	0.0458
B4GALT7 Missense	11	28 (14)	2 (1)	0.0764	0.761	Stereotypy [HP:0000733]	2/11 (18.2 %)	107	2.58×10^{-4}	0.0245	88.1	0.0714

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
CEP152 PAV	75	35 (14)	1 (1)	0.779	0.177	Neoplasm by histology [HP:0011792]	7/75 (9.33 %)	6.65	1.70×10^{-4}	0.0350	6.13	0.0354
EED Missense	11	30 (14)	1 (1)	0.0460	0.941	Neurodevelopmental delay [HP:0012758]	5/11 (45.5 %)	13.1	2.67×10^{-4}	0.0247	7.59	0.00658
PSPH PAV	12	24 (14)	1 (1)	0.123	0.261	EEG with generalized slow activity [HP:0010845]	4/12 (33.3 %)	14.3	5.26×10^{-4}	0.0455	9.88	0.00932
PSPH Missense	8	30 (14)	1 (1)	0.0236	0.330	EEG with generalized slow activity [HP:0010845]	4/8 (50.0 %)	28.6	8.29×10^{-5}	0.00756	14.8	0.00932
ARSB Missense	12	21 (12)	1 (1)	0.0702	0.571	Asymmetric ventricles [HP:0100960]	2/12 (16.7 %)	92.7	3.32×10^{-4}	0.0316	77.8	0.0690
ARSB dMissense	10	26 (14)	1 (1)	0.227	0.672	Asymmetric ventricles [HP:0100960]	2/10 (20.0 %)	116	2.27×10^{-4}	0.0214	93.3	0.0690
TRIO dPAV	77	23 (14)	1 (1)	0.669	0.242	Myoclonic seizure [HP:0032794]	23/77 (29.9 %)	2.70	1.86×10^{-4}	0.0378	2.19	0.0133
TELO2 Missense	28	26 (13)	1 (1)	0.120	0.382	Encephalopathy [HP:0001298]	11/28 (39.3 %)	4.47	3.64×10^{-4}	0.0465	3.11	0.00687
TAOK1 PAV	26	22 (13)	1 (1)	0.788	0.690	Motor delay [HP:0001270]	3/26 (11.5 %)	29.7	2.23×10^{-4}	0.0314	26.4	0.0517
TAOK1 Missense	23	21 (12)	1 (1)	0.777	0.420	Motor delay [HP:0001270]	3/23 (13.0 %)	34.1	1.53×10^{-4}	0.0217	29.9	0.0517
JAM3 PTV	1	43 (12)	4 (1)	-	-	Aortic root aneurysm [HP:0002616]	1/1 (100 %)	Inf	7.93×10^{-5}	0.0200	Inf	1.00
MGAT2 dPAV	14	38 (11)	*3 (1)*	0.00227	0.0181	Abnormality of brain morphology [HP:0012443]	12/14 (85.7 %)	12.7	5.51×10^{-5}	0.00674	2.67	0.00296
MGAT2 dMissense	12	41 (12)	4 (1)	0.00281	0.0104	Abnormal cerebral cortex morphology [HP:0002538]	7/12 (58.3 %)	8.82	3.84×10^{-4}	0.0375	4.26	0.00404
NT5C2 dPAV	20	29 (12)	2 (1)	0.464	0.993	Abnormal respiratory system physiology [HP:0002795]	3/20 (15.0 %)	23.9	4.30×10^{-4}	0.0458	20.5	0.0316

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
DDX11 PAV	55	33 (11)	1 (1)	0.884	0.0471	Joint laxity [HP:0001388]	2/55 (3.64 %)	234	1.11×10^{-4}	0.0268	228	0.500
DDX11 Missense	31	37 (12)	1 (1)	0.810	0.0668	Joint laxity [HP:0001388]	2/31 (6.45 %)	425	3.50×10^{-5}	0.00781	406	0.500
RFX4 PAV	16	33 (12)	1 (1)	0.718	0.745	Facial asymmetry [HP:0000324]	2/16 (12.5 %)	198	8.25×10^{-5}	0.0103	175	0.182
RFX4 dPAV	13	33 (12)	1 (1)	0.691	0.899	Facial asymmetry [HP:0000324]	2/13 (15.4 %)	251	5.37×10^{-5}	0.00709	215	0.182
RFX4 Missense	15	36 (12)	1 (1)	0.718	0.742	Facial asymmetry [HP:0000324]	2/15 (13.3 %)	214	7.22×10^{-5}	0.00932	187	0.182
RFX4 dMissense	12	33 (12)	1 (1)	0.614	0.908	Facial asymmetry [HP:0000324]	2/12 (16.7 %)	274	4.54×10^{-5}	0.00598	233	0.182
INTS11 PAV	31	34 (12)	1 (1)	0.0551	0.585	Focal sensory seizure with cephalic sensation [HP:0032810]	2/31 (6.45 %)	143	1.62×10^{-4}	0.0264	135	0.250
INTS11 dPAV	25	31 (10)	1 (1)	0.0745	0.582	Focal sensory seizure with cephalic sensation [HP:0032810]	2/25 (8.00 %)	180	1.05×10^{-4}	0.0169	168	0.250
INTS11 Missense	20	36 (8)	1 (1)	0.184	0.898	Focal sensory seizure with cephalic sensation [HP:0032810]	2/20 (10.0 %)	231	6.65×10^{-5}	0.0100	210	0.250
INTS11 dMissense	16	41 (9)	1 (1)	0.139	0.821	Focal sensory seizure with cephalic sensation [HP:0032810]	2/16 (12.5 %)	293	4.21×10^{-5}	0.00638	262	0.250
IDUA PAV	20	26 (12)	1 (1)	0.396	0.151	Dilation of lateral ventricles [HP:0006956]	2/20 (10.0 %)	99.1	2.83×10^{-4}	0.0344	89.9	0.125
ADK PAV	13	24 (12)	1 (1)	0.0304	0.0370	EEG with photoparoxysmal response grade 1 [HP:0011211]	2/13 (15.4 %)	134	1.66×10^{-4}	0.0183	114	0.105
ADK dPAV	11	22 (11)	1 (1)	0.0252	0.149	EEG with photoparoxysmal response grade 1 [HP:0011211]	2/11 (18.2 %)	163	1.17×10^{-4}	0.0126	135	0.105
TTC8 dMissense	12	24 (12)	1 (1)	0.360	0.833	Involuntary movements [HP:0004305]	3/12 (25.0 %)	25.2	4.60×10^{-4}	0.0416	19.2	0.0180

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
TRNT1 dMissense	16	23 (12)	1 (1)	0.418	0.736	Abnormality of the cerebral subcortex [HP:0010993]	5/16 (31.2 %)	11.2	2.86×10^{-4}	0.0313	7.98	0.0100
GALT PAV (dPAV)	9	16 (10)	1 (1)	0.529	0.385	Short attention span [HP:0000736]	2/9 (22.2 %)	93.8	3.48×10^{-4}	0.0277	73.7	0.0500
GALT PTV	3	21 (12)	1 (1)	0.463	0.998	Short attention span [HP:0000736]	2/3 (66.7 %)	642	2.94×10^{-5}	0.00193	221	0.0500
CARS1 PAV	34	29 (11)	1 (1)	0.611	0.0929	Abnormality of immune system physiology [HP:0010978]	2/34 (5.88 %)	97.7	3.13×10^{-4}	0.0465	92.5	0.200
CARS1 dPAV	23	24 (9)	2 (1)	0.810	0.0662	Abnormality of immune system physiology [HP:0010978]	2/23 (8.70 %)	148	1.42×10^{-4}	0.0211	137	0.200
CARS1 Missense	27	29 (11)	2 (1)	0.850	0.0755	Abnormality of immune system physiology [HP:0010978]	2/27 (7.41 %)	125	1.97×10^{-4}	0.0290	117	0.200
CARS1 dMissense	17	28 (10)	2 (1)	0.999	0.0753	Abnormality of immune system physiology [HP:0010978]	2/17 (11.8 %)	208	7.65×10^{-5}	0.0107	185	0.200
MVK PAV	14	20 (8)	2 (1)	0.617	0.909	Absent septum pellucidum [HP:0001331]	2/14 (14.3 %)	673	1.14×10^{-5}	0.00227	600	0.400
MVK Missense	11	26 (11)	2 (1)	0.0413	0.929	Absent septum pellucidum [HP:0001331]	2/11 (18.2 %)	915	6.91×10^{-6}	0.00128	764	0.400
CPA6 PTV	8	25 (11)	2 (1)	0.00121	0.124	Intellectual disability, severe [HP:0010864]	3/8 (37.5 %)	26.5	5.70×10^{-4}	0.0326	16.9	0.0106
AUTS2 PAV	48	22 (11)	2 (1)	0.158	0.0919	Abnormal basal ganglia MRI signal intensity [HP:0012751]	2/48 (4.17 %)	135	2.11×10^{-4}	0.0440	131	0.333
AUTS2 dPAV	40	22 (9)	1 (1)	0.179	0.218	Abnormal basal ganglia MRI signal intensity [HP:0012751]	2/40 (5.00 %)	165	1.46×10^{-4}	0.0307	157	0.333
AUTS2 Missense	44	22 (11)	2 (1)	0.168	0.253	Abnormal basal ganglia MRI signal intensity [HP:0012751]	2/44 (4.55 %)	148	1.77×10^{-4}	0.0373	143	0.333
AUTS2 dMissense	37	24 (9)	1 (1)	0.188	0.470	Abnormal basal ganglia MRI signal intensity [HP:0012751]	2/37 (5.41 %)	178	1.25×10^{-4}	0.0254	170	0.333

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
SFXN4 PAV	19	29 (11)	1 (1)	0.268	0.638	Abnormal glucose homeostasis [HP:0011014]	2/19 (10.5 %)	91.8	3.24×10^{-4}	0.0368	82.8	0.111
SFXN4 Missense	11	24 (9)	3 (1)	0.322	0.0890	Abnormal glucose homeostasis [HP:0011014]	2/11 (18.2 %)	174	1.05×10^{-4}	0.0113	143	0.111
ADAM22 dPAV	23	24 (11)	1 (1)	0.0161	0.139	Focal impaired awareness seizure [HP:0002384]	17/23 (73.9 %)	5.24	2.18×10^{-4}	0.0298	2.11	0.00383
STT3A PAV	29	22 (11)	1 (1)	0.256	0.370	Intellectual disability, mild [HP:0001256]	8/29 (27.6 %)	11.7	2.41×10^{-6}	3.40×10^{-4}	8.79	0.0199
STT3A dPAV	28	22 (11)	1 (1)	0.221	0.314	Intellectual disability, mild [HP:0001256]	8/28 (28.6 %)	12.3	1.80×10^{-6}	1.80×10^{-4}	9.10	0.0199
STT3A Missense	27	17 (10)	1 (1)	0.461	0.363	Intellectual disability, mild [HP:0001256]	7/27 (25.9 %)	10.8	1.65×10^{-5}	0.00358	8.24	0.0174
STT3A dMissense	26	18 (10)	1 (1)	0.381	0.303	Intellectual disability, mild [HP:0001256]	7/26 (26.9 %)	11.3	1.26×10^{-5}	0.00299	8.56	0.0174
COG6 PAV	35	30 (10)	9 (1)	0.419	0.567	Unilateral facial palsy [HP:0012799]	2/35 (5.71 %)	189	1.11×10^{-4}	0.0237	180	0.333
COG6 Missense	24	26 (5)	10 (1)	0.587	0.731	Unilateral facial palsy [HP:0012799]	2/24 (8.33 %)	283	5.18×10^{-5}	0.0102	262	0.333
PCGF2 PAV	13	24 (9)	3 (1)	7.70×10^{-5}	7.26×10^{-4}	Focal aware non-motor seizure [HP:0032682]	9/13 (69.2 %)	10.3	7.06×10^{-5}	0.00876	3.85	0.00396
PCGF2 dPAV	12	27 (10)	3 (1)	7.00×10^{-5}	3.29×10^{-4}	Focal aware non-motor seizure [HP:0032682]	9/12 (75.0 %)	13.7	2.59×10^{-5}	0.00347	4.17	0.00396
PCGF2 dMissense	8	26 (9)	1 (1)	4.38×10^{-4}	0.00279	Focal impaired awareness seizure [HP:0002384]	8/8 (100 %)	Inf	2.32×10^{-4}	0.0193	2.85	0.00180
RBM10 Missense	19	33 (10)	2 (1)	0.268	0.564	Abnormal skeletal morphology [HP:0011842]	5/19 (26.3 %)	11.0	2.58×10^{-4}	0.0311	8.35	0.0124

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
TLK2 PAV	14	29 (10)	2 (1)	0.0528	0.296	Nystagmus [HP:0000639]	2/14 (14.3 %)	80.3	4.26×10^{-4}	0.0397	69.2	0.0714
TLK2 dPAV (dMissense)	12	27 (9)	2 (1)	0.380	0.330	Nystagmus [HP:0000639]	2/12 (16.7 %)	96.3	3.09×10^{-4}	0.0290	80.8	0.0714
TLK2 Missense	13	27 (9)	2 (1)	0.0649	0.323	Nystagmus [HP:0000639]	2/13 (15.4 %)	87.5	3.65×10^{-4}	0.0341	74.5	0.0714
SHOC2 PAV	12	20 (9)	2 (1)	0.00156	0.00146	Generalized-onset motor seizure [HP:0032677]	10/12 (83.3 %)	13.6	7.41×10^{-5}	0.00801	3.10	0.00295
SHOC2 Missense	11	21 (10)	1 (1)	0.00273	5.53×10^{-4}	Generalized-onset motor seizure [HP:0032677]	9/11 (81.8 %)	12.3	2.32×10^{-4}	0.0214	3.05	0.00265
SIN3A PAV	31	20 (10)	1 (1)	0.979	0.00513	Limb dysmetria [HP:0002406]	2/31 (6.45 %)	123	2.08×10^{-4}	0.0328	116	0.222
PEX26 bMissense	5	20 (10)	1 (1)	0.234	0.0724	Abnormality of movement [HP:0100022]	3/5 (60.0 %)	46.8	2.89×10^{-4}	0.0150	19.3	0.00761
ACTG1 PAV (dPAV)	8	16 (10)	1 (1)	0.00578	0.493	Refractory drug response [HP:0020174]	7/8 (87.5 %)	25.6	1.38×10^{-4}	0.0131	4.08	0.00258
L1CAM PAV	21	16 (10)	1 (1)	0.0353	0.407	Fixation-off epileptiform discharges [HP:0025644]	2/21 (9.52 %)	218	7.35×10^{-5}	0.0115	200	0.250
L1CAM Missense	18	16 (9)	1 (1)	0.115	0.546	Visually-induced seizure [HP:0020216]	3/18 (16.7 %)	24.2	4.30×10^{-4}	0.0454	20.4	0.0283
CC2D1A PTV	9	38 (9)	2 (1)	0.176	0.950	Cerebellar hypoplasia [HP:0001321]	2/9 (22.2 %)	105	2.82×10^{-4}	0.0248	82.4	0.0556
GPT2 dMissense	10	29 (9)	2 (1)	0.262	0.338	Intellectual disability [HP:0001249]	6/10 (60.0 %)	10.5	5.17×10^{-4}	0.0387	4.80	0.00379
BBS4 PAV	36	24 (9)	2 (1)	0.610	0.208	Skeletal muscle atrophy [HP:0003202]	2/36 (5.56 %)	242	7.88×10^{-5}	0.0179	233	0.400
BBS4 dPAV	28	26 (8)	2 (1)	0.594	0.154	Skeletal muscle atrophy [HP:0003202]	2/28 (7.14 %)	321	4.73×10^{-5}	0.0104	300	0.400

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
OFD1 dPAV	15	30 (9)	1 (1)	0.769	0.596	Feeding difficulties in infancy [HP:0008872]	2/15 (13.3 %)	76.9	4.56×10^{-4}	0.0437	67.2	0.0741
ESCO2 PAV	28	20 (9)	1 (1)	0.376	0.183	Focal polymicrogyria [HP:0032471]	2/28 (7.14 %)	96.3	3.09×10^{-4}	0.0432	89.9	0.167
ESCO2 dPAV	17	15 (8)	1 (1)	0.383	0.282	Focal polymicrogyria [HP:0032471]	2/17 (11.8 %)	165	1.12×10^{-4}	0.0149	148	0.167
ESCO2 PTV	8	12 (7)	1 (1)	0.473	0.693	Focal polymicrogyria [HP:0032471]	2/8 (25.0 %)	413	2.32×10^{-5}	0.00253	315	0.167
SET dMissense	2	20 (9)	1 (1)	0.895	0.999	Mitral valve prolapse [HP:0001634]	1/2 (50.0 %)	Inf	1.59×10^{-4}	0.0448	Inf	1.00
GRID2 PAV	43	9 (6)	1 (1)	0.286	0.772	Asymmetric ventricles [HP:0100960]	3/43 (6.98 %)	36.1	1.27×10^{-4}	0.0262	33.7	0.103
GRID2 dPAV	34	14 (8)	1 (1)	0.374	0.912	Asymmetric ventricles [HP:0100960]	3/34 (8.82 %)	46.6	6.24×10^{-5}	0.0124	42.7	0.103
GRID2 Missense	40	11 (6)	1 (1)	0.617	0.796	Asymmetric ventricles [HP:0100960]	3/40 (7.50 %)	39.0	1.02×10^{-4}	0.0232	36.3	0.103
GRID2 dMissense	32	15 (9)	1 (1)	0.602	0.971	Asymmetric ventricles [HP:0100960]	3/32 (9.38 %)	49.8	5.19×10^{-5}	0.0106	45.4	0.103
MN1 dPAV	25	14 (9)	1 (1)	0.587	0.704	Apraxia [HP:0002186]	3/25 (12.0 %)	34.1	1.51×10^{-4}	0.0219	30.2	0.0566
MN1 Missense	29	11 (7)	1 (1)	0.300	0.816	Apraxia [HP:0002186]	3/29 (10.3 %)	28.9	2.37×10^{-4}	0.0356	26.0	0.0566
MN1 dMissense	24	14 (9)	1 (1)	0.414	0.786	Apraxia [HP:0002186]	3/24 (12.5 %)	35.7	1.33×10^{-4}	0.0203	31.5	0.0566
TAF2 PTV	12	13 (9)	1 (1)	0.243	0.984	EEG with photoparoxysmal response [HP:0010852]	4/12 (33.3 %)	14.0	5.74×10^{-4}	0.0477	9.65	0.00911
PEX1 dPAV	42	20 (8)	2 (1)	0.205	0.470	Hyperreflexia [HP:0001347]	3/42 (7.14 %)	34.4	1.45×10^{-4}	0.0284	32.1	0.0968

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
B4GALNT1 PAV	20	8 (6)	2 (1)	0.0379	0.652	EEG with generalized slow activity [HP:0010845]	5/20 (25.0 %)	9.56	4.52×10^{-4}	0.0476	7.42	0.0117
B4GALNT1 dPAV	18	9 (7)	1 (1)	0.115	0.645	EEG with generalized slow activity [HP:0010845]	5/18 (27.8 %)	11.0	2.64×10^{-4}	0.0305	8.25	0.0117
B4GALNT1 Missense	17	9 (7)	2 (1)	0.0523	0.460	EEG with generalized slow activity [HP:0010845]	5/17 (29.4 %)	12.0	1.96×10^{-4}	0.0229	8.74	0.0117
B4GALNT1 dMissense	15	11 (8)	2 (1)	0.139	0.304	EEG with generalized slow activity [HP:0010845]	5/15 (33.3 %)	14.3	1.01×10^{-4}	0.0120	9.90	0.0117
LINS1 PAV	36	22 (8)	1 (1)	0.358	0.161	Gray matter heterotopia [HP:0002282]	5/36 (13.9 %)	10.1	2.53×10^{-4}	0.0393	8.87	0.0248
YY1 PAV	11	17 (8)	1 (1)	0.195	0.597	Periventricular leukomalacia [HP:0006970]	2/11 (18.2 %)	99.4	2.97×10^{-4}	0.0297	81.8	0.0667
SOX5 dMissense	18	17 (8)	1 (1)	0.144	0.248	Visually-induced seizure [HP:0020216]	3/18 (16.7 %)	24.2	4.30×10^{-4}	0.0454	20.4	0.0283
PPP1R21 dPAV	26	15 (8)	1 (1)	0.668	0.863	Dysarthria [HP:0001260]	2/26 (7.69 %)	116	2.22×10^{-4}	0.0310	108	0.182
GATAD2B PAV (dPAV, Missense, dMissense)	14	14 (8)	1 (1)	0.488	0.770	Perisylvian polymicrogyria [HP:0012650]	2/14 (14.3 %)	99.1	2.86×10^{-4}	0.0275	85.7	0.0870
PIK3CA PAV	16	22 (7)	1 (1)	0.00277	0.302	Perisylvian polymicrogyria [HP:0012650]	2/16 (12.5 %)	85.1	3.76×10^{-4}	0.0371	75.0	0.0870
PIK3CA dPAV	13	23 (7)	1 (1)	0.00150	0.599	Perisylvian polymicrogyria [HP:0012650]	2/13 (15.4 %)	108	2.45×10^{-4}	0.0230	92.3	0.0870
PNPLA6 PTV	13	15 (7)	1 (1)	0.798	0.566	Abnormality of extrapyramidal motor function [HP:0002071]	2/13 (15.4 %)	226	6.44×10^{-5}	0.00848	194	0.167
CDK10 PTV	7	14 (7)	1 (1)	0.577	0.196	EEG with burst suppression [HP:0010851]	2/7 (28.6 %)	84.6	4.76×10^{-4}	0.0275	61.0	0.0328
TAF6 PTV	1	10 (7)	1 (1)	-	-	Myoclonic status epilepticus [HP:0032667]	1/1 (100 %)	Inf	2.38×10^{-4}	0.0447	6300	0.333

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
SLC5A6 Missense	18	10 (7)	1 (1)	0.198	0.872	Continuous spike and waves during slow sleep [HP:0031491]	3/18 (16.7 %)	26.0	3.52×10^{-4}	0.0398	21.9	0.0303
PQBP1 PAV	11	15 (6)	2 (1)	0.600	0.266	Nystagmus [HP:0000639]	2/11 (18.2 %)	107	2.58×10^{-4}	0.0245	88.1	0.0714
PQBP1 dPAV	10	15 (6)	3 (1)	0.594	0.277	Nystagmus [HP:0000639]	2/10 (20.0 %)	120	2.12×10^{-4}	0.0197	96.9	0.0714
ZBTB7A dPAV (dMissense)	8	9 (6)	1 (1)	0.0387	0.00478	*Focal-onset seizure [HP:0007359]*	*0/8 (0.00 %)*	*0.00*	*4.22×10^{-4}*	*0.0254*	*0.00*	*0.00167*
BUB1 dMissense	24	9 (6)	1 (1)	0.947	0.137	EEG with photoparoxysmal response grade IV [HP:0011214]	5/24 (20.8 %)	9.68	3.75×10^{-4}	0.0476	7.87	0.0148
SPR PTV	1	12 (6)	1 (1)	-	-	Abnormal midbrain morphology [HP:0002418]	1/1 (100 %)	Inf	2.38×10^{-4}	0.0447	6300	0.333
WAC Missense (dMissense)	11	12 (6)	1 (1)	0.126	0.499	Postterm pregnancy [HP:0031169]	2/11 (18.2 %)	84.3	4.05×10^{-4}	0.0381	69.4	0.0571
AP4S1 dPAV	2	12 (3)	2 (1)	0.651	0.873	Circumoral cyanosis [HP:0032556]	1/2 (50.0 %)	Inf	1.59×10^{-4}	0.0448	Inf	1.00
AP4S1 Missense	2	12 (3)	2 (1)	0.479	0.622	Circumoral cyanosis [HP:0032556]	1/2 (50.0 %)	Inf	1.59×10^{-4}	0.0448	Inf	1.00
AP4S1 dMissense	1	17 (5)	3 (1)	-	-	Circumoral cyanosis [HP:0032556]	1/1 (100 %)	Inf	7.93×10^{-5}	0.0200	Inf	1.00
AGO2 PAV (dPAV, Missense, dMissense)	10	8 (5)	1 (1)	0.192	0.404	Non-motor seizure [HP:0033259]	9/10 (90.0 %)	17.2	4.60×10^{-4}	0.0354	2.62	0.00208
IVD dPAV	7	8 (5)	1 (1)	0.112	0.00608	Myoclonic absence seizure [HP:0011150]	2/7 (28.6 %)	87.6	4.45×10^{-4}	0.0266	63.2	0.0339

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
TTC19 Missense	2	4 (3)	1 (1)	0.678	0.630	Oromotor apraxia [HP:0007301]	1/2 (50.0 %)	Inf	1.59×10^{-4}	0.0448	Inf	1.00
CRPPA PAV	14	2 (2)	1 (1)	0.998	0.448	Complex febrile seizure [HP:0011172]	2/14 (14.3 %)	122	1.94×10^{-4}	0.0220	106	0.105
CRPPA dPAV	11	6 (5)	1 (1)	1.00	0.333	Complex febrile seizure [HP:0011172]	2/11 (18.2 %)	163	1.17×10^{-4}	0.0126	135	0.105
CRPPA Missense	10	5 (4)	1 (1)	0.990	0.512	Complex febrile seizure [HP:0011172]	2/10 (20.0 %)	183	9.61×10^{-5}	0.0106	148	0.105
CRPPA dMissense	8	7 (4)	1 (1)	0.915	0.471	Complex febrile seizure [HP:0011172]	2/8 (25.0 %)	245	5.99×10^{-5}	0.00634	185	0.105
SMAD4 PAV	10	17 (5)	1 (1)	0.0915	0.456	Abnormality of the diencephalon [HP:0010662]	2/10 (20.0 %)	97.7	3.13×10^{-4}	0.0289	78.8	0.0588
SMAD4 PTV	1	15 (2)	2 (1)	-	-	Focal aware emotional seizure with laughing [HP:0032745]	1/1 (100 %)	Inf	7.93×10^{-5}	0.0200	Inf	1.00
PEX16 dPAV	10	17 (5)	1 (1)	0.0319	0.819	Abnormal amygdala morphology [HP:0033668]	2/10 (20.0 %)	111	2.43×10^{-4}	0.0232	90.0	0.0667
PIGS PAV	27	11 (4)	1 (1)	0.850	0.440	Trigonocephaly [HP:0000243]	2/27 (7.41 %)	Inf	4.41×10^{-6}	0.00108	Inf	1.00
PIGS Missense	20	13 (5)	1 (1)	0.764	0.713	Trigonocephaly [HP:0000243]	2/20 (10.0 %)	Inf	2.39×10^{-6}	6.20×10^{-4}	Inf	1.00
MTFMT Missense	8	11 (5)	1 (1)	0.00406	1.44×10^{-4}	Generalized-onset motor seizure [HP:0032677]	7/8 (87.5 %)	19.1	6.20×10^{-4}	0.0368	3.26	0.00206
ASXL1 PTV	5	7 (4)	2 (1)	0.126	0.966	Memory impairment [HP:0002354]	2/5 (40.0 %)	267	6.61×10^{-5}	0.00567	163	0.0606
CA2 PAV	7	5 (4)	1 (1)	0.0537	0.624	Multifocal epileptiform discharges [HP:0010841]	4/7 (57.1 %)	19.8	4.82×10^{-4}	0.0283	9.05	0.00500
AASS dPAV	26	5 (4)	1 (1)	0.436	0.721	Intellectual disability, moderate [HP:0002342]	5/26 (19.2 %)	11.1	1.90×10^{-4}	0.0272	9.20	0.0187
TMEM94 PTV	10	9 (3)	2 (1)	0.348	0.495	Focal impaired awareness sensory seizure [HP:0032787]	2/10 (20.0 %)	208	7.65×10^{-5}	0.00852	168	0.118

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
OPA3 dPAV	5	8 (3)	1 (1)	0.560	0.515	Nocturnal seizures [HP:0031951]	2/5 (40.0 %)	117	3.27×10^{-4}	0.0171	71.0	0.0274
OPA3 dMissense	3	7 (2)	1 (1)	0.238	0.111	Nocturnal seizures [HP:0031951]	2/3 (66.7 %)	349	9.88×10^{-5}	0.00464	118	0.0274
CEP55 PTV	1	9 (2)	2 (1)	-	-	Underdeveloped nasolabial fold [HP:0010801]	1/1 (100 %)	Inf	2.38×10^{-4}	0.0447	6300	0.333
NPC2 PTV	3	4 (2)	1 (1)	0.100	0.194	EEG with focal slow activity [HP:0010843]	3/3 (100 %)	Inf	1.76×10^{-4}	0.00654	17.9	0.00424
TUBB Missense (dMissense)	1	16 (2)	1 (1)	-	-	Oculomotor apraxia [HP:0000657]	1/1 (100 %)	Inf	1.59×10^{-4}	0.0330	12600	0.500

Table 5-16. A summary of the associations of carriers of PAV, dPAV, dMissense, and missense variants in clinical genes beyond the GE 'green' epilepsy genes. Commenced 23 pages previously.

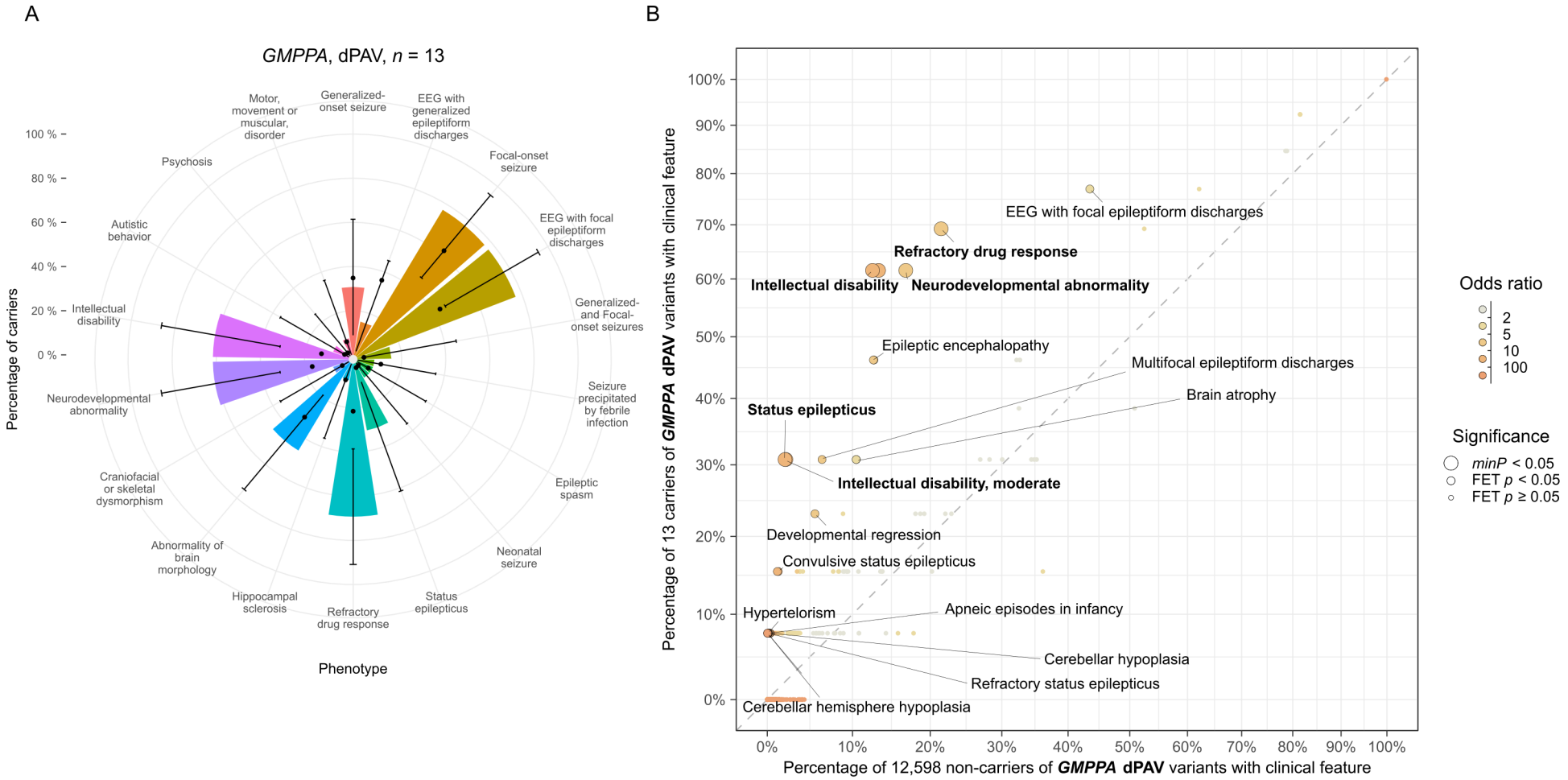


Figure 5-32. The associations of carriers of dPAVs in *GMMPA*.

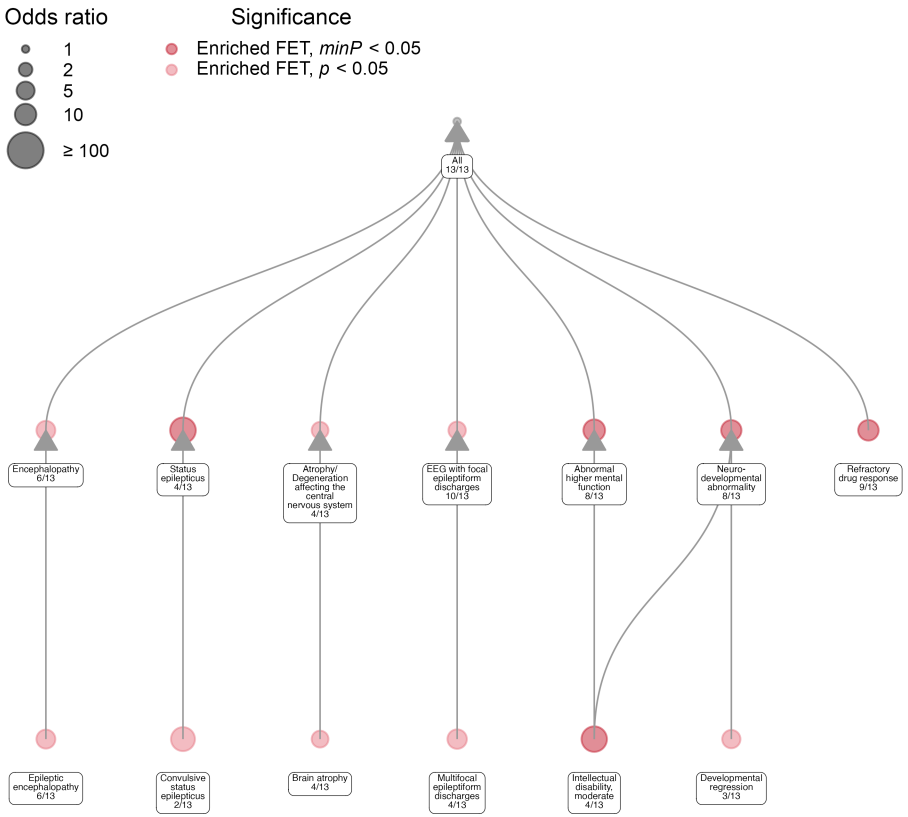


Figure 5-33. The associations of carriers of dPAVs in *GMMPA* shown conceptually.

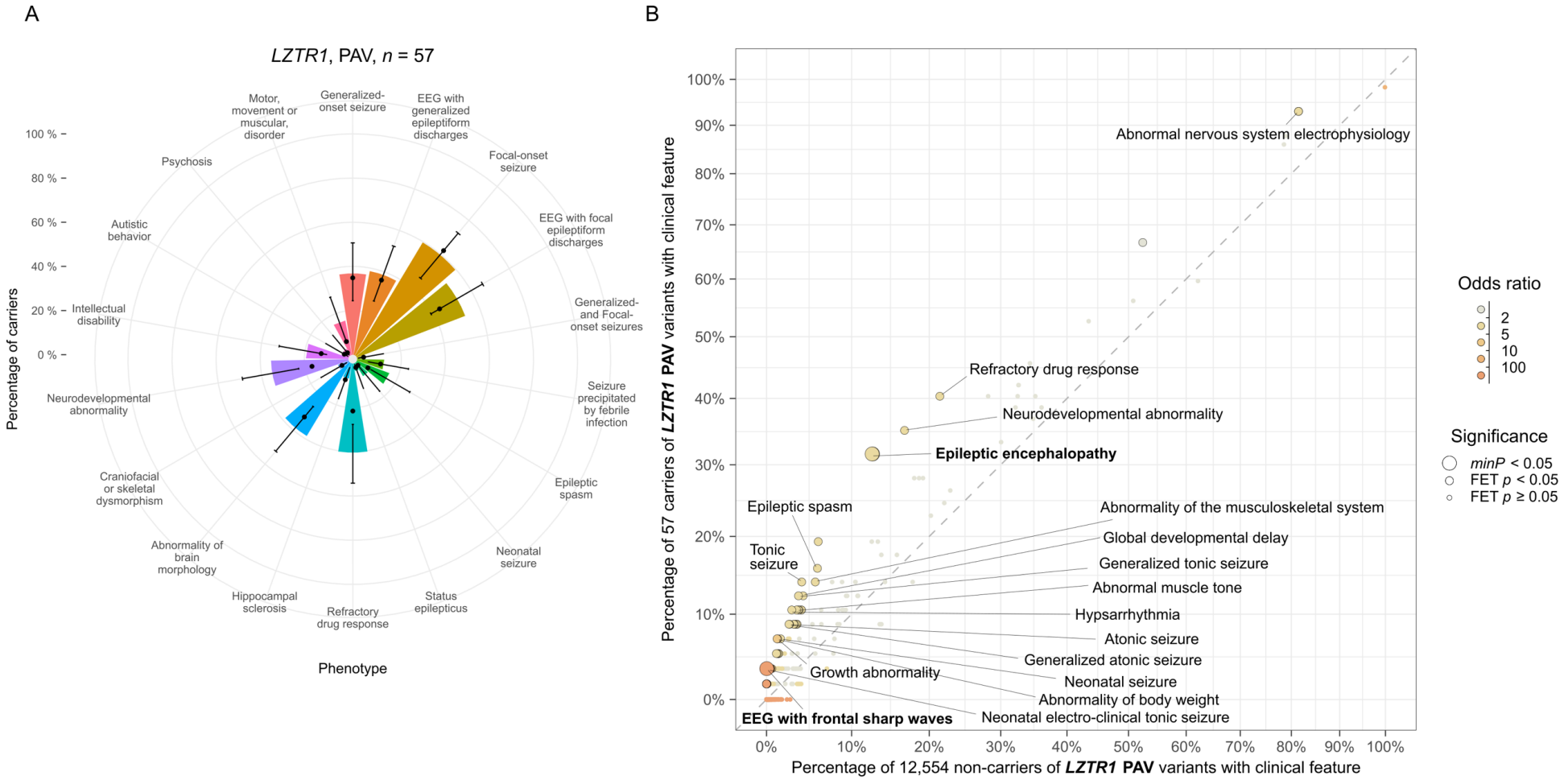


Figure 5-34. The associations of carriers of PAVs in *LZTR1*.

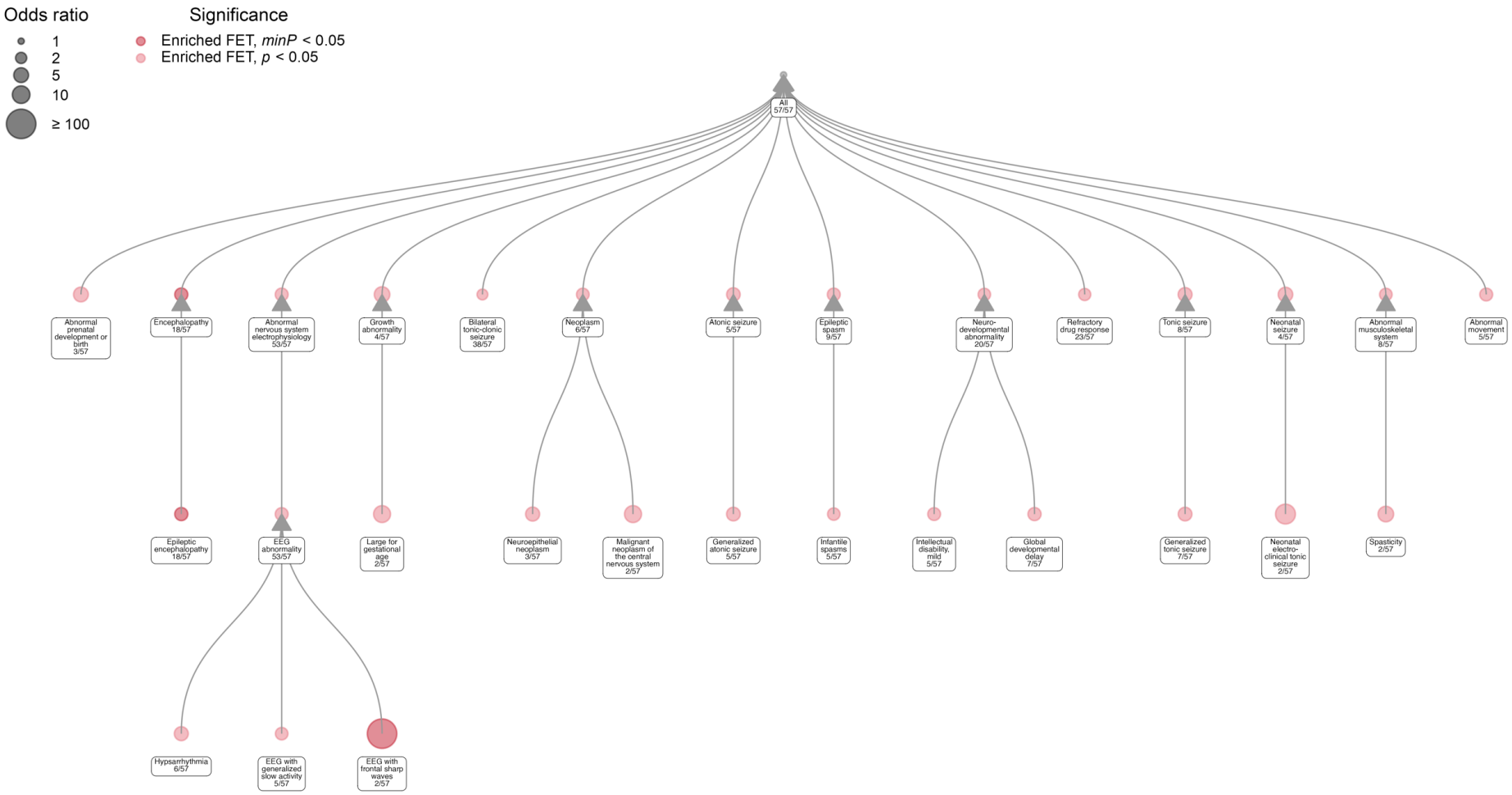


Figure 5-35. The associations of carriers of PAVs in *LZTR1* shown conceptually.

5.5.10 Phenotypic similarity and associations of genes not yet clinically validated in epilepsy or neurodevelopmental disorders

Now I shall present the results of a preliminary exploration of the similarities and associations of groups carrying URVs in genes that have not been validated as causes of epilepsy or neurodevelopmental disorders. A few have been implicated by case reports, association studies, and non-clinical research but currently have insufficient evidence for validation. These genes include genes with high pLI, Missense-*z*, pHaplo, pTriplo scores, the SFARI genes, and 'amber' genes from the GE panels for epilepsy and intellectual disability that are neither within the GE 'green' gene lists for epilepsy or intellectual disability nor the ClinGen epilepsy, autism, and intellectual disability gene set. Here, I present those PAV, dPAV, PTV, dMissense, and missense groups of at least five carriers, with at least three conceptually independent phenotypes associated with $minP < 0.05$ as candidates that might influence the clinical features of, or potentially cause, epilepsy (**Table 5-17**). The genes discussed are all brain expressed, and I will summarize the published literature of potential relevance to epilepsy and neurodevelopmental disorders for these genes and existing mouse models and their phenotypes from the Mouse Genome Database (www.informatics.jax.org, accessed 2024-03-17, Blake *et al.*, 2021).

The associations of carriers of PAVs in *PAK1IP* are shown in Figure 5-36 and Figure 5-37, of carriers of PAVs in *KANSL1L* in Figure 5-38 and Figure 5-39, of carriers of PAVs in *UTP14A* in **Figure 5-40** and Figure 5-41, and of carriers of dPAVs in *SESTD1* in **Figure 5-42** and **Figure 5-43.**

I will not report the associations with missense variants (1,208 apparent carriers, 9.58 % of the cohort) and dMissense variants (976 apparent carriers, 7.74 % of the cohort) in *TTN*, as these are likely to be false positive URVs. *TTN* is a large gene with repeating domains to which exome reads can misalign. Without a separate cohort processed through the same pipeline to control for platform or batch effects we were unable to discern the true URVs. No other genes reaching the threshold for reporting were among those identified as likely to report false positive rare variants in large exome studies (Shyr *et al.*, 2014).

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
PAK1IP1 PAV	14	116 (35)	8 (7)	0.0946	0.440	Abnormal circulating metabolite concentration [HP:0032180]	2/14 (14.3 %)	173	1.03×10^{-4}	0.0125	150	0.143
						Delayed fine motor development [HP:0010862]	2/14 (14.3 %)	138	1.54×10^{-4}	0.0179	120	0.118
						Delayed gross motor development [HP:0002194]	2/14 (14.3 %)	122	1.94×10^{-4}	0.0220	106	0.105
						Intellectual disability, severe [HP:0010864]	4/14 (28.6 %)	17.7	2.05×10^{-4}	0.0226	12.9	0.0142
						Impaired executive functioning [HP:0033051]	2/14 (14.3 %)	116	2.15×10^{-4}	0.0230	100	0.100
						Extra-axial cerebrospinal fluid accumulation [HP:0012510]	2/14 (14.3 %)	104	2.61×10^{-4}	0.0258	90.0	0.0909
						Developmental stagnation at onset of seizures [HP:0006834]	2/14 (14.3 %)	99.1	2.86×10^{-4}	0.0275	85.7	0.0870
PAK1IP1 dPAV	7	56 (27)	13 (9)	0.00213	0.0884	Delayed fine motor development [HP:0010862]	2/7 (28.6 %)	329	3.58×10^{-5}	0.00391	240	0.118
						Delayed gross motor development [HP:0002194]	2/7 (28.6 %)	294	4.50×10^{-5}	0.00477	212	0.105
						Impaired executive functioning [HP:0033051]	2/7 (28.6 %)	277	4.99×10^{-5}	0.00509	200	0.100
						Developmental stagnation at onset of seizures [HP:0006834]	2/7 (28.6 %)	237	6.64×10^{-5}	0.00620	171	0.0870
						Language impairment [HP:0002463]	3/7 (42.9 %)	54.7	8.53×10^{-5}	0.00707	31.8	0.0173
						Memory impairment [HP:0002354]	2/7 (28.6 %)	160	1.38×10^{-4}	0.0119	116	0.0606
						Short attention span [HP:0000736]	2/7 (28.6 %)	131	2.04×10^{-4}	0.0147	94.8	0.0500
						Intellectual disability, severe [HP:0010864]	3/7 (42.9 %)	33.1	3.62×10^{-4}	0.0222	19.4	0.0106

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
						Multifocal epileptiform discharges [HP:0010841]	4/7 (57.1 %)	19.8	4.82×10^{-4}	0.0283	9.05	0.00500
PAK1IP1 Missense	11	114 (37)	7 (5)	0.00430	0.400	Abnormal circulating metabolite concentration [HP:0032180]	2/11 (18.2 %)	230	6.26×10^{-5}	0.00676	191	0.143
						Delayed fine motor development [HP:0010862]	2/11 (18.2 %)	184	9.34×10^{-5}	0.0100	153	0.118
						Delayed gross motor development [HP:0002194]	2/11 (18.2 %)	163	1.17×10^{-4}	0.0126	135	0.105
						Extra-axial cerebrospinal fluid accumulation [HP:0012510]	2/11 (18.2 %)	139	1.58×10^{-4}	0.0151	115	0.0909
						Developmental stagnation at onset of seizures [HP:0006834]	2/11 (18.2 %)	132	1.73×10^{-4}	0.0164	109	0.0870
PAK1IP1 dMissense	6	47 (25)	5 (3)	0.00456	0.415	Delayed fine motor development [HP:0010862]	2/6 (33.3 %)	411	2.56×10^{-5}	0.00293	280	0.118
						Delayed gross motor development [HP:0002194]	2/6 (33.3 %)	363	3.21×10^{-5}	0.00366	247	0.105
						Developmental stagnation at onset of seizures [HP:0006834]	2/6 (33.3 %)	297	4.75×10^{-5}	0.00453	200	0.0870

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
OPCML Missense	8	70 (29)	11 (8)	0.00536	0.0611	Obsessive-compulsive behavior [HP:0000722]	2/8 (25.0 %)	457	1.93×10^{-5}	0.00209	350	0.182
(the *p*-value thresholds are met for the same individuals as nominators plus one more to give denominator of 9 with PAVs)						Aggressive behavior [HP:0000718]	2/8 (25.0 %)	277	4.77×10^{-5}	0.00506	210	0.118
						EEG with generalized slow activity [HP:0010845]	4/8 (50.0 %)	28.6	8.29×10^{-5}	0.00756	14.8	0.00932
						Developmental stagnation at onset of seizures [HP:0006834]	2/8 (25.0 %)	197	8.85×10^{-5}	0.00802	150	0.0870
						EEG with focal spike waves [HP:0011197]	2/8 (25.0 %)	188	9.65×10^{-5}	0.00928	143	0.0833
						Hyperactivity [HP:0000752]	2/8 (25.0 %)	173	1.14×10^{-4}	0.0104	131	0.0769
						EEG with focal sharp waves [HP:0011196]	2/8 (25.0 %)	166	1.23×10^{-4}	0.0116	126	0.0741
						Abnormal eye morphology [HP:0012372]	2/8 (25.0 %)	104	2.99×10^{-4}	0.0228	78.8	0.0476
KANSL1L PAV	45	63 (33)	5 (3)	0.0262	0.430	Tonic seizure [HP:0032792]	9/45 (20.0 %)	5.98	6.80×10^{-5}	0.0126	4.99	0.0175
						Abnormal muscle tone [HP:0003808]	8/45 (17.8 %)	6.12	1.36×10^{-4}	0.0287	5.21	0.0183
						Epileptic spasm [HP:0011097]	10/45 (22.2 %)	4.59	2.38×10^{-4}	0.0439	3.79	0.0134
KANSL1L PTV	16	49 (21)	9 (4)	0.00345	0.601	Epileptic spasm [HP:0011097]	6/16 (37.5 %)	9.61	2.01×10^{-4}	0.0225	6.38	0.00804
						Epileptic encephalopathy [HP:0200134]	8/16 (50.0 %)	6.94	3.21×10^{-4}	0.0340	3.97	0.00502
						Neurodevelopmental abnormality [HP:0012759]	9/16 (56.2 %)	6.39	3.86×10^{-4}	0.0372	3.36	0.00425
						Hypotonia [HP:0001252]	4/16 (25.0 %)	13.4	5.16×10^{-4}	0.0474	10.3	0.0129

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
KCNJ1 PAV	23	50 (20)	8 (3)	0.827	0.322	Cleft lip [HP:0410030]	2/23 (8.70 %)	1200	9.54×10^{-6}	0.00237	1090	0.667
						Cleft palate [HP:0000175]	2/23 (8.70 %)	237	6.64×10^{-5}	0.0120	219	0.286
						Abnormality of prenatal development or birth [HP:0001197]	4/23 (17.4 %)	18.8	1.19×10^{-4}	0.0180	15.7	0.0280
KCNJ1 Missense	20	45 (21)	9 (3)	0.857	0.122	Cleft lip [HP:0410030]	2/20 (10.0 %)	1300	7.16×10^{-6}	0.00174	1260	0.667
						Cleft palate [HP:0000175]	2/20 (10.0 %)	277	4.99×10^{-5}	0.00847	252	0.286
						Abnormality of prenatal development or birth [HP:0001197]	4/20 (20.0 %)	22.4	6.67×10^{-5}	0.0100	18.1	0.0280
KCNJ1 dMissense	13	25 (11)	7 (4)	0.915	0.134	Cleft palate [HP:0000175]	2/13 (15.4 %)	447	2.05×10^{-5}	0.00334	388	0.286
						Abnormality of prenatal development or birth [HP:0001197]	3/13 (23.1 %)	26.7	3.76×10^{-4}	0.0345	20.8	0.0210
						Neonatal electro-clinical tonic seizure [HP:0032821]	2/13 (15.4 %)	84.1	3.92×10^{-4}	0.0377	71.8	0.0690
						Cleft lip [HP:0410030]	2/13 (15.4 %)	2100	2.94×10^{-6}	6.90×10^{-4}	1940	0.667

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
NECTIN2 dPAV	13	28 (14)	6 (3)	0.766	0.400	Cleft palate [HP:0000175]	2/13 (15.4 %)	447	2.05×10^{-5}	0.00334	388	0.286
						Neonatal electro-clinical tonic seizure [HP:0032821]	2/13 (15.4 %)	84.1	3.92×10^{-4}	0.0377	71.8	0.0690
						Cleft lip [HP:0410030]	2/13 (15.4 %)	2100	2.94×10^{-6}	6.90×10^{-4}	1940	0.667
NECTIN2 Missense	13	46 (22)	7 (3)	0.497	0.285	Cleft palate [HP:0000175]	2/13 (15.4 %)	447	2.05×10^{-5}	0.00334	388	0.286
						Neonatal electro-clinical tonic seizure [HP:0032821]	2/13 (15.4 %)	84.1	3.92×10^{-4}	0.0377	71.8	0.0690
						Cleft lip [HP:0410030]	2/13 (15.4 %)	2100	2.94×10^{-6}	6.90×10^{-4}	1940	0.667
NECTIN2 dMissense	9	40 (18)	10 (3)	0.650	0.176	Cleft palate [HP:0000175]	2/9 (22.2 %)	684	9.49×10^{-6}	0.00148	560	0.286
						Neonatal electro-clinical tonic seizure [HP:0032821]	2/9 (22.2 %)	132	1.82×10^{-4}	0.0174	104	0.0690
						Cleft lip [HP:0410030]	2/9 (22.2 %)	3170	1.36×10^{-6}	3.40×10^{-4}	2800	0.667
NOP14 Missense	22	47 (22)	6 (3)	0.0465	0.0179	Epileptic encephalopathy [HP:0200134]	11/22 (50.0 %)	6.95	2.36×10^{-5}	0.00487	3.98	0.00690
						Neurodevelopmental abnormality [HP:0012759]	12/22 (54.5 %)	5.97	6.02×10^{-5}	0.00978	3.26	0.00566
						Neonatal electro-clinical seizure [HP:0032809]	4/22 (18.2 %)	18.3	1.35×10^{-4}	0.0201	15.2	0.0258
UTP14A PAV	19	52 (21)	3 (3)	0.0601	0.0774	*Dialeptic seizure [HP:0011146]*	*2/19 (10.5 %)*	*0.114*	*3.31×10^{-4}*	*0.0370*	*0.207*	*0.00274*
						Complex febrile seizure [HP:0011172]	2/19 (10.5 %)	86.6	3.62×10^{-4}	0.0405	78.0	0.105
						Non-motor seizure [HP:0033259]	*0/19 (0.00 %)*	*0.00*	*4.22×10^{-4}*	*0.0459*	*0.00*	*0.00230*

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		$p < 0.05$	$minP < 0.05$	Sim_{MinTO}	$Sim_{Goodall3}$							
STX12 PAV	7	50 (17)	4 (3)	0.188	0.950	Postterm pregnancy [HP:0031169]	2/7 (28.6 %)	151	1.56×10^{-4}	0.0129	109	0.0571
						Decreased head circumference [HP:0040195]	3/7 (42.9 %)	43.1	1.69×10^{-4}	0.0133	25.1	0.0138
						Hypotonia [HP:0001252]	3/7 (42.9 %)	30.0	4.78×10^{-4}	0.0280	17.6	0.00968
STX12 dPAV	6	51 (18)	4 (3)	0.529	0.858	Decreased head circumference [HP:0040195]	3/6 (50.0 %)	57.4	9.81×10^{-5}	0.00803	29.3	0.0138
						Postterm pregnancy [HP:0031169]	2/6 (33.3 %)	189	1.11×10^{-4}	0.00938	127	0.0571
						Hypotonia [HP:0001252]	3/6 (50.0 %)	40.0	2.78×10^{-4}	0.0166	20.5	0.00968
STX12 Missense (dMissense)	5	53 (18)	9 (3)	0.560	0.692	Decreased head circumference [HP:0040195]	3/5 (60.0 %)	86.2	4.97×10^{-5}	0.00404	35.2	0.0138
						Postterm pregnancy [HP:0031169]	2/5 (40.0 %)	253	7.44×10^{-5}	0.00604	153	0.0571
						Hypotonia [HP:0001252]	3/5 (60.0 %)	60.0	1.42×10^{-4}	0.00819	24.6	0.00968
PPFIA1 dMissense	19	42 (18)	7 (3)	7.20×10^{-5}	0.379	Intellectual disability [HP:0001249]	10/19 (52.6 %)	7.79	2.96×10^{-5}	0.00479	4.22	0.00632
						Generalized atonic seizure [HP:0032887]	5/19 (26.3 %)	13.6	9.66×10^{-5}	0.0139	10.3	0.0153
						Epileptic encephalopathy [HP:0200134]	9/19 (47.4 %)	6.25	2.26×10^{-4}	0.0271	3.76	0.00565
HCFC2 Missense	16	55 (17)	3 (3)	0.139	0.464	Premature birth [HP:0001622]	3/16 (18.8 %)	29.7	2.51×10^{-4}	0.0278	24.3	0.0300
						Abnormal foot morphology [HP:0001760]	2/16 (12.5 %)	81.2	4.10×10^{-4}	0.0417	71.6	0.0833
						Refractory drug response [HP:0020174]	10/16 (62.5 %)	6.10	4.60×10^{-4}	0.0431	2.91	0.00369

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
SESTD1 dPAV	22	33 (13)	4 (3)	0.0194	0.0665	Refractory drug response [HP:0020174]	13/22 (59.1 %)	5.29	1.41×10^{-4}	0.0209	2.76	0.00479
						Epileptic encephalopathy [HP:0200134]	10/22 (45.5 %)	5.79	1.54×10^{-4}	0.0217	3.61	0.00627
						Hypsarrhythmia [HP:0002521]	5/22 (22.7 %)	9.96	3.51×10^{-4}	0.0442	7.93	0.0137
VPS26B PAV	11	24 (10)	10 (3)	0.554	0.687	Cleft palate [HP:0000175]	2/11 (18.2 %)	538	1.45×10^{-5}	0.00214	458	0.286
						Neonatal electro-clinical tonic seizure [HP:0032821]	2/11 (18.2 %)	103	2.77×10^{-4}	0.0269	84.8	0.0690
						Cleft lip [HP:0410030]	2/11 (18.2 %)	2510	2.07×10^{-6}	5.70×10^{-4}	2290	0.667
VPS26B dPAV	10	27 (10)	10 (3)	0.330	0.561	Cleft palate [HP:0000175]	2/10 (20.0 %)	607	1.19×10^{-5}	0.00183	504	0.286
						Neonatal electro-clinical tonic seizure [HP:0032821]	2/10 (20.0 %)	116	2.27×10^{-4}	0.0214	93.3	0.0690
						Cleft lip [HP:0410030]	2/10 (20.0 %)	2800	1.70×10^{-6}	4.10×10^{-4}	2520	0.667
VPS26B Missense (dMissense)	9	30 (11)	10 (3)	0.128	0.692	Cleft palate [HP:0000175]	2/9 (22.2 %)	684	9.49×10^{-6}	0.00148	560	0.286
						Neonatal electro-clinical tonic seizure [HP:0032821]	2/9 (22.2 %)	132	1.82×10^{-4}	0.0174	104	0.0690
						Cleft lip [HP:0410030]	2/9 (22.2 %)	3170	1.36×10^{-6}	3.40×10^{-4}	2800	0.667

Gene variant class	Number of carriers	Number of phenotypic associations (conceptually independent)		Similarity *p*-value		Conceptually independent phenotypic associations reaching (*minP* < 0.05)	Frequency among carriers	Odds ratio	*p*-value	*minP*	Relative risk	Positive predictive value
		p < 0.05	*minP* < 0.05	Sim_{MinTO}	$Sim_{Goodall3}$							
POU2F2 dMissense	13	28 (10)	6 (3)	0.0418	0.853	Cleft palate [HP:0000175]	2/13 (15.4 %)	447	2.05×10^{-5}	0.00334	388	0.286
						Neonatal electro-clinical tonic seizure [HP:0032821]	2/13 (15.4 %)	84.1	3.92×10^{-4}	0.0377	71.8	0.0690
						Cleft lip [HP:0410030]	2/13 (15.4 %)	2100	2.94×10^{-6}	6.90×10^{-4}	1940	0.667

Table 5-17. Clinical associations of carriers of ultra-rare variants in genes without a clinically validated role in epilepsy. The associations of groups of at least five carriers of PAVs, dPAVs, PTVs, dMissense, and missense variants in genes without a clinically validated role in epilepsy and that have three conceptually independent phenotypic associations with *minP* < 0.05. Commenced 7 pages previously.

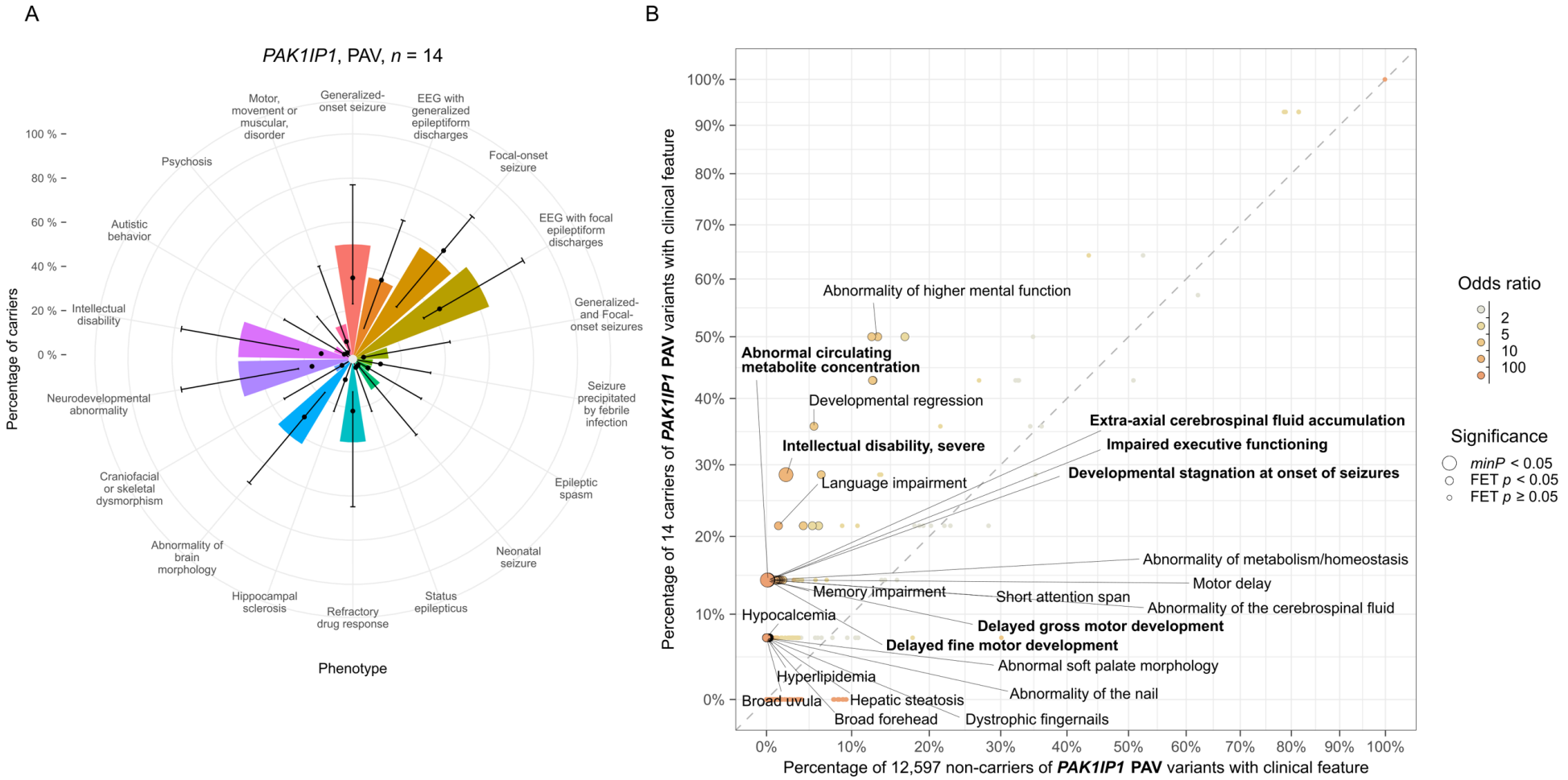


Figure 5-36. The associations of carriers of PAVs in *PAK1IP1*.

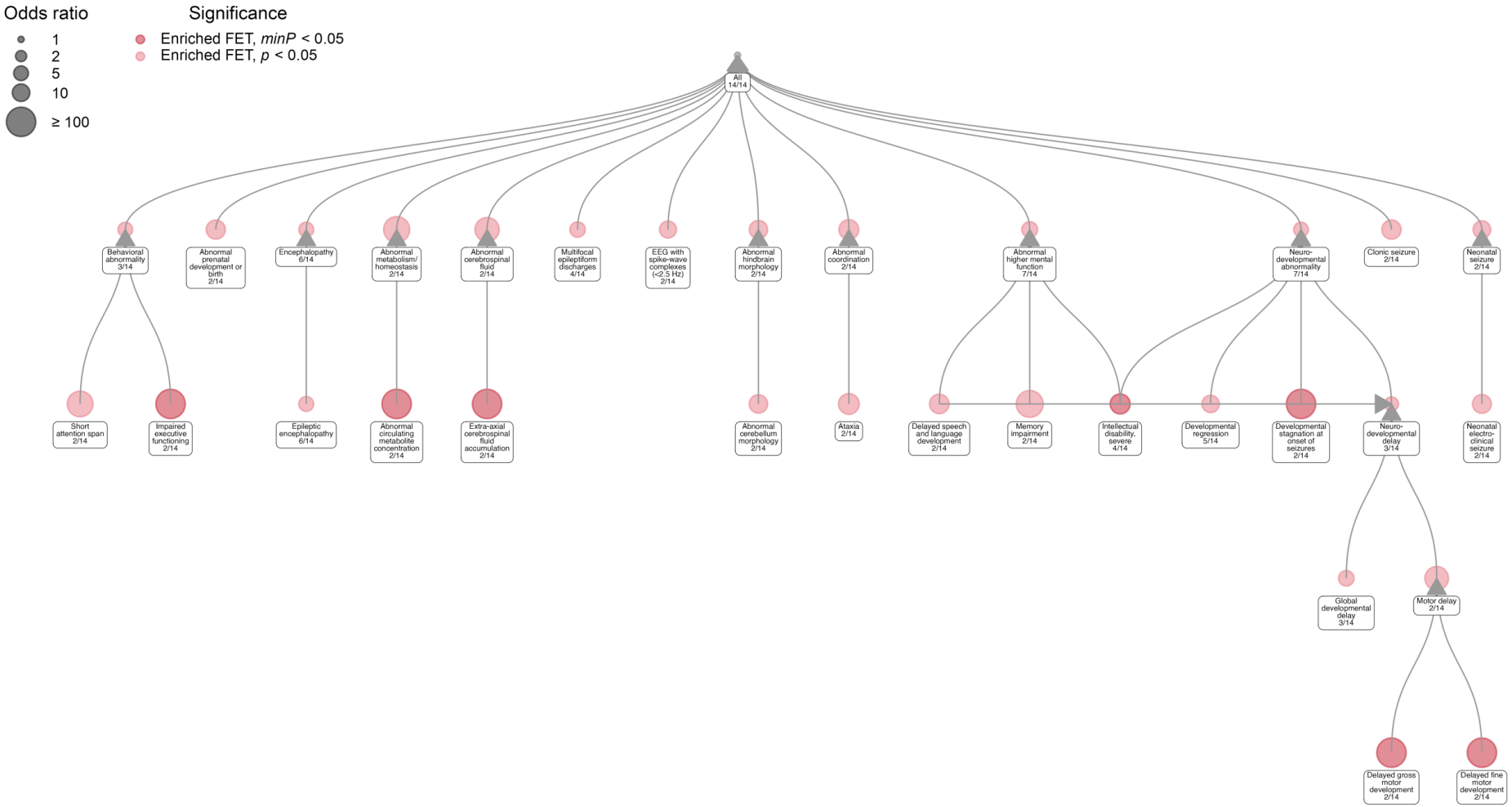


Figure 5-37. The associations of carriers of PAVs in *PAK1IP1* shown conceptually.

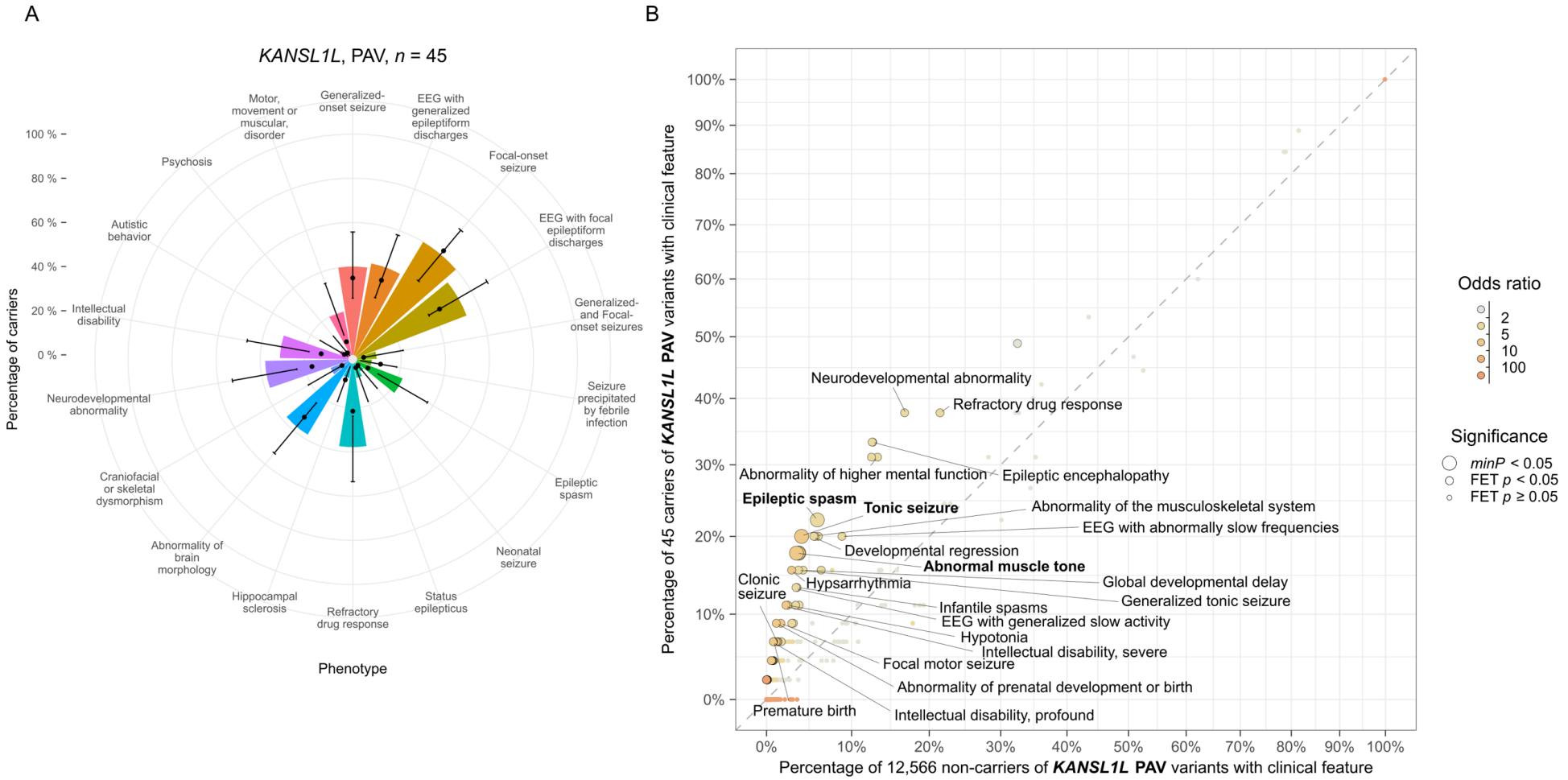


Figure 5-38. The associations of carriers of PAVs in *KANSL1L*.

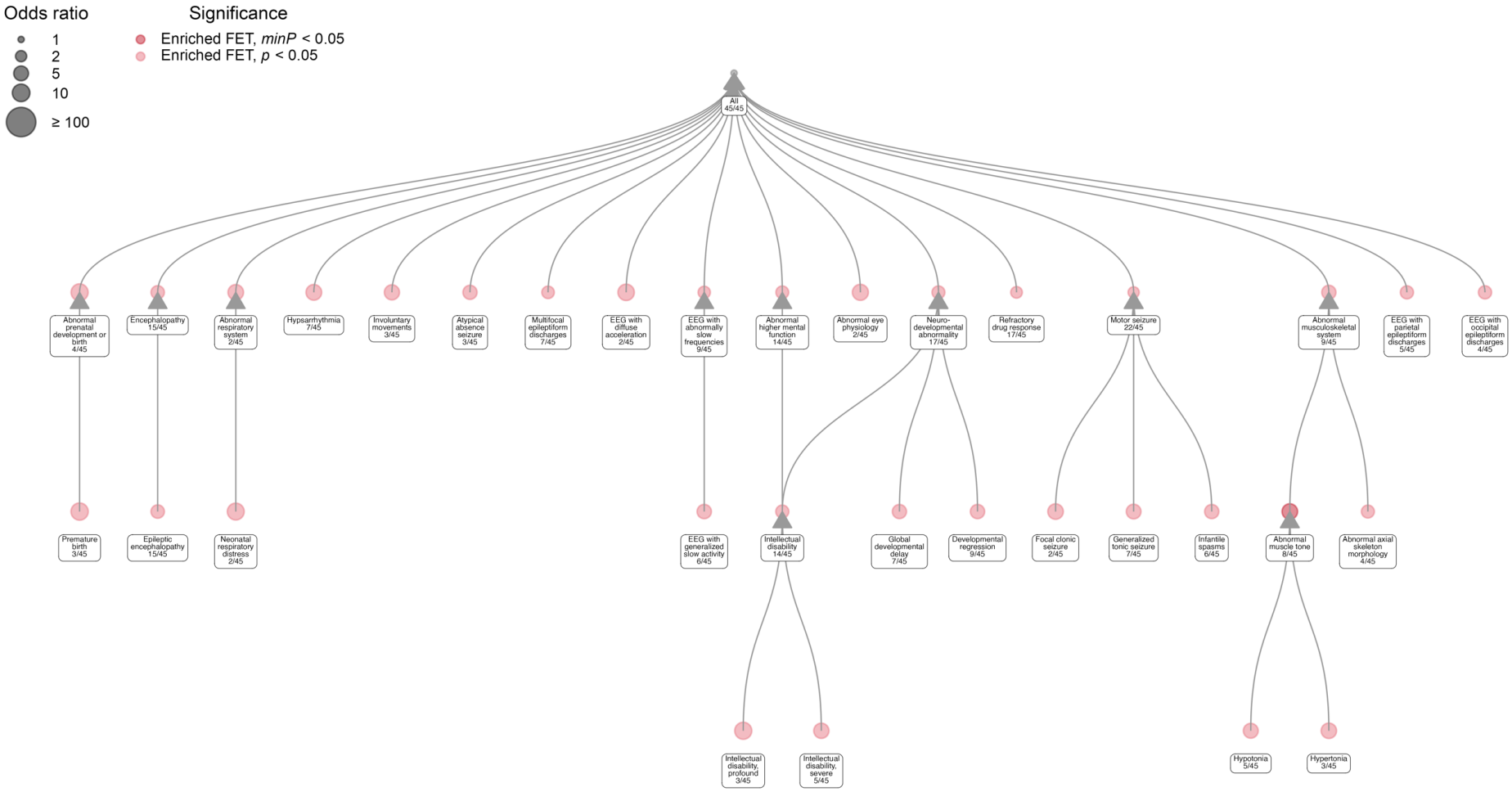


Figure 5-39. The associations of carriers of PAVs in *KANSL1L* shown conceptually.

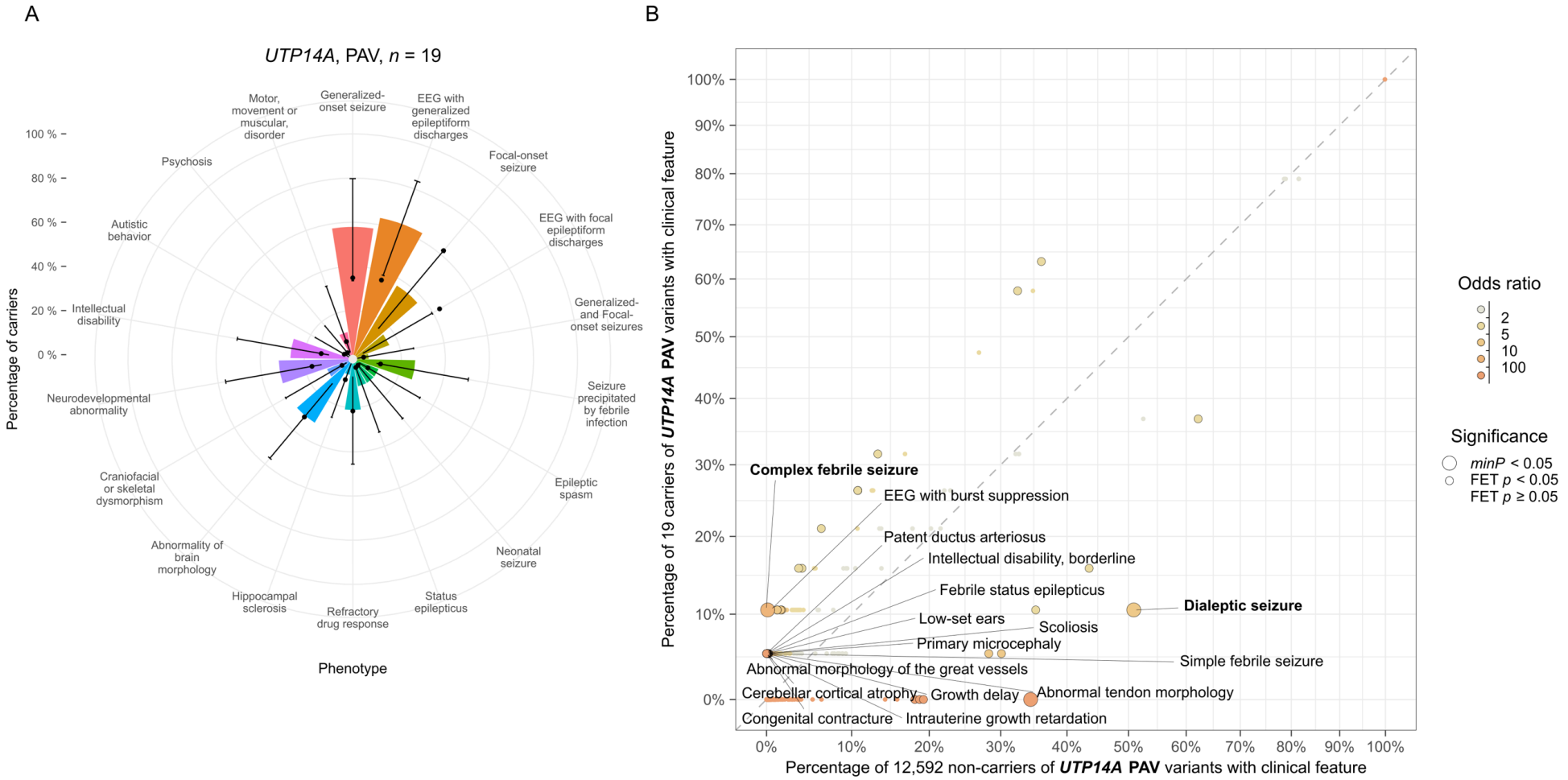


Figure 5-40. The associations of carriers of PAVs in *UTP14A*.

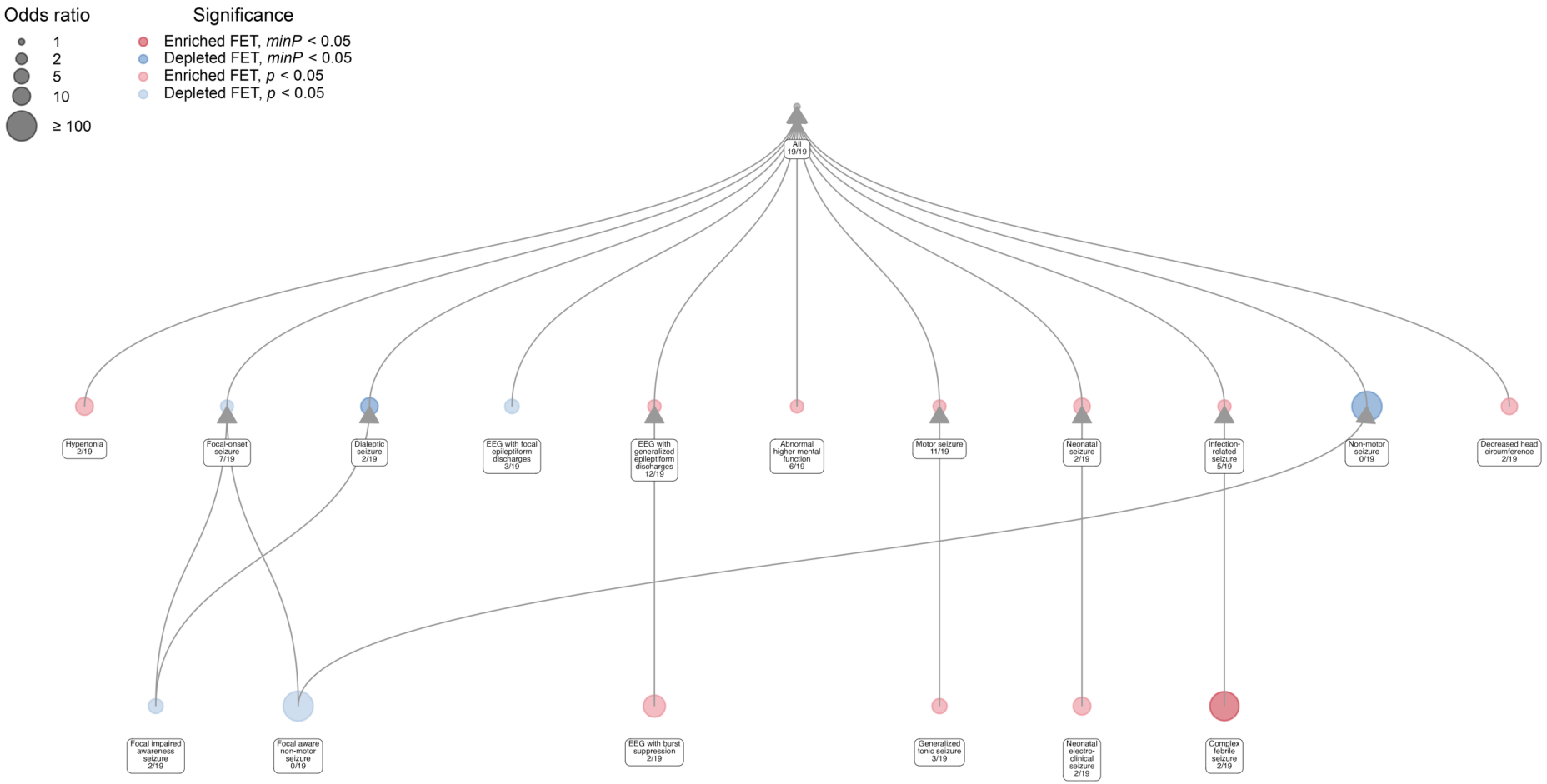


Figure 5-41. The associations of carriers of PAVs in *UTP14A* shown conceptually.

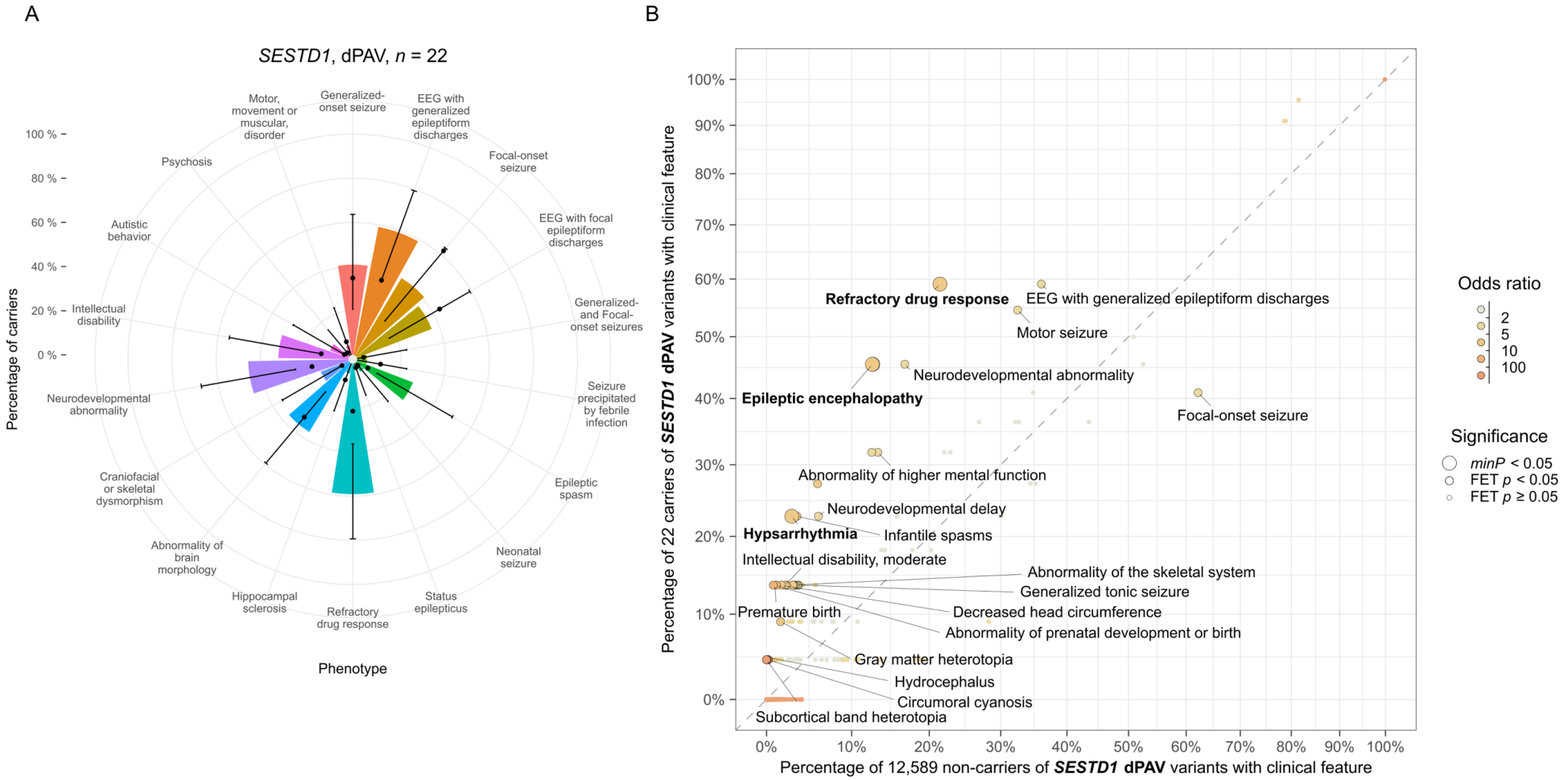


Figure 5-42. The associations of carriers of dPAVs in *SESTD1*.

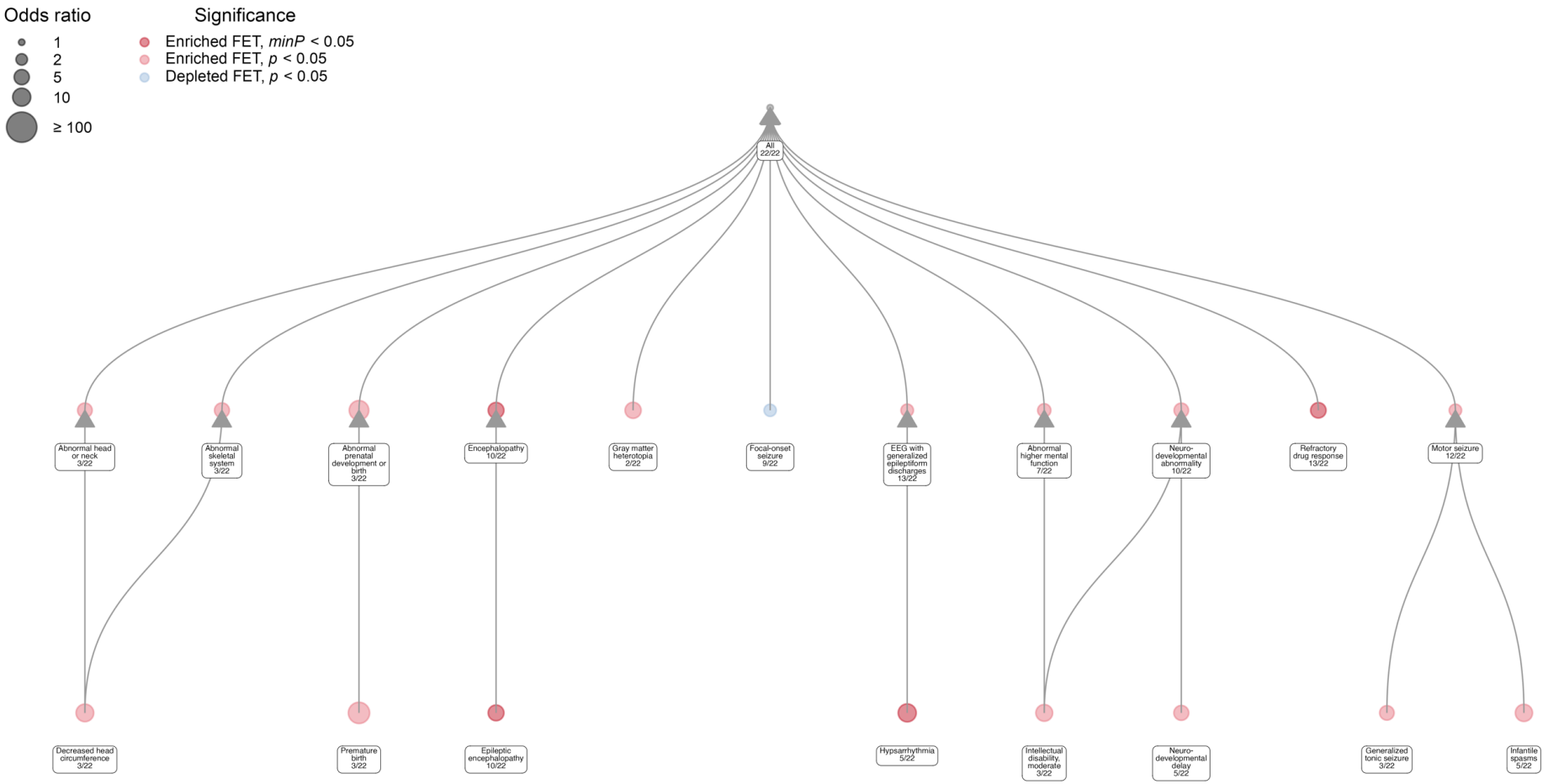


Figure 5-43. The associations of carriers of dPAVs in *SESTD1* shown conceptually.

PAK1IP1 qualified for inclusion because if has a high probability of being haploinsufficient. It encodes PAK1 (p21-activated kinase) interacting protein 1 and thought to negatively regulate PAK1 (Xia *et al.*, 2001), a serine-threonine protein kinase that has been found to have disturbed expression in Huntington's disease (Seefelder and Kochanek, 2021) and Fragile X syndrome (reviewed in (Ma *et al.*, 2012)), found to harbor gain-of-function variants in human malignancies (Radu *et al.*, 2014), and modulates early brain and craniofacial development (Asmar *et al.*, 2023 4599). De novo variants in *PAK1* have been discovered in seven people with developmental delay, macrocephaly, and seizures (Harms *et al.*, 2018; Horn *et al.*, 2019; Zhuang *et al.*, 2023), and PAK1 phosphorylation occurs downstream of *TRIO*, variants in which can cause autosomal dominant intellectual development disorder with microcephaly (type 44) [OMIM:617061] and with macrocephaly (type 63) [OMIM:618825] (Bonnet *et al.*). The *Pak1ip1*^{mray}/*Pak1ip1*^{mray} mouse model [MGI:5548169] exhibits abnormal craniofacial development with a median orofacial cleft due to maxillary insufficiency, with decreased forebrain (with distorted ventricular system) and hindbrain size, and relatively increased midbrain size (Ross *et al.*, 2013). *Pak1ip1* is preferentially expressed in brain, facial primordia, limb buds, and liver in embryonic mice (Ross *et al.*, 2013), to the extent that it has been proposed as a reference for normalization of gene expression (Hildyard;Wells and Piercy, 2021). Median orofacial cleft is rare in humans relative to the paramedian location and can be associated with morphologic ophthalmic and neurodevelopmental abnormalities as part of the holoprosencephaly spectrum (Mishra *et al.*, 2015). While skull bone growth in *Pak1ip1*^{mray}/*Pak1ip1*^{mray} mice is delayed, the craniofacial changes are thought insufficient to explain the embryonic lethality, hypothesized to be a consequence of the depletion of neural crest cells in the anterior head due to cell cycle arrest at G_0/G_1 (Panoutsopoulos *et al.*, 2020 4588). PAK1IP1 is found in the nucleolus and both knockdown and overexpression of *PAK1IP1* in human cell cultures inhibited cell proliferation at G_1 arrest in a p53-dependent manner (Yu *et al.*, 2011). Were *PAK1IP1* variants to be confirmed as modulating neurodevelopmental phenotypes, therapeutic PAK1 modulators developed for oncological therapy might be found to rescue some consequences (Semenova and Chernoff, 2017).

OPCML qualified for inclusion because it has a high probability of being haploinsufficient and triplosensitive, encoding opioid-binding protein/cell adhesion molecule-like. This is one of the GPI-anchored cell adhesion molecules with immunoglobulin domains known as IgLON. OPCML is necessary for normal neurite outgrowth (Faivre-Sarrailh and Rougon, 1997) and synaptogenesis (Hashimoto;Maekawa and Miyata, 2009). *Opcml*$^{-/-}$ mice [MGI:6469751] demonstrate abnormal behavior, difficulties with maze tasks, and abnormal dendritic spine

maturation, ameliorated by aripiprazole, but no primary motor or seizure phenotype has been reported (Zhang *et al.*, 2019). The authors identified an association with *OPCML* expression quantitative trait loci in a genome-wide association meta-analysis of schizophrenia in Han Chinese cohorts (Zhang *et al.*, 2019) and it had previously been identified as a susceptibility locus for schizophrenia in a Dutch cohort (Schol-Gelok *et al.*, 2010). Two people have been described with Jacobsen syndrome [OMIM:147791], consisting of neurodevelopmental and physical growth delay, trigonocephaly and facial dysmorphism, and thrombocytopenia, and found to have deletions of the region of 11q encompassing *OPCML* (Ji *et al.*, 2010). OPCML is thought to be a tumor suppressor and thus explored as an oncological therapeutic target (Antony *et al.*, 2021).

KANSL1L qualified for inclusion because it has a high probability of being haploinsufficient and encodes KAT8 regulatory NSL complex subunit 1-like. It has been identified as a candidate risk gene for neuroblastoma due to being near associated single nucleotide polymorphisms (SNPs) (Bae *et al.*, 2020) and found to be involved in somatic gene rearrangements in lipoblastomata (Lopez-Nunez *et al.*, 2020), but not directly implicated in epilepsy or mendelian disorders. However, *KANSL1L* can be confused with *KANSL1* (da Silva-Buttkus *et al.*, 2023), the cause of autosomal dominant Koolen-de Vries syndrome [OMIM:610443]. As epilepsy can develop as part of Koolen-de Vries syndrome, *KANSL1* (not *KANSL1L*) has been explored but not found to be a common cause of self-limited focal epilepsies of childhood (Myers *et al.*, 2018). The transgenic mouse crossed with the *Kansl1l*$^{tm1b(EUCOMM)Wtsi}$ [MGI:6379553], is reported to have largely ophthalmological and testicular, but not neurological, phenotypes (International Knockout Mouse Consortium).

KCNJ1 qualified for inclusion because it has a high probability of triplosensitivity, and encodes the potassium inwardly-rectifying channel, subfamily J, member 1. It can cause autosomal recessive Bartter's syndrome, type 2 [OMIM:241200], an antenatal-onset renal tubulopathy with severe salt-wasting, sometimes associated with seizures (due to hypocalcemia from hypercalciuria), and failure to thrive (Károlyi *et al.*, 1997). Epilepsy has not been reported and *Kcnj1*^{cnj1tm1Ges}/*Kcnj1*^{tm1Ges} mice [MGI:1927248] have homeostatic and renal phenotypes with postnatal growth restriction and lethality.

NECTIN2 is a ubiquitously expressed gene with a high probability of being loss-of-function intolerant. It encodes the nectin cell adhesion molecule 2, which like OPCML, has an immunoglobulin-like structure and is implicated in synapse and formation and homeostasis of neurons and astrocytes (reviewed in Mizutani *et al.*, 2022). It has been studied a target for neurooncological malignancy (Wienke *et al.*, 2024) and has been proposed as candidate

systemic biomarker of Alzheimer's disease risk (Tin *et al.*, 2023), but has not been associated with mendelian disease. Two homozygous mouse strains have been developed: *Nectin2*^{tm1Smu}/*Nectin2*^{tm1Smu} [MGI:2183406] and *Nectin*^{2tm1Vrr}/*Nectin2*^{tm1Vrr} [MGI:2183405] (Bouchard *et al.*, 2000). Both were reported to have reproductive phenotypes and the *Nectin2*^{tm1Smu}/*Nectin2*^{tm1Smu} strain was found to have reduced synaptogenesis (Shiotani *et al.*, 2021).

NOP14 qualified for inclusion because it has a high probability of triplosensitivity. It encodes the nucleolar protein 14. It has not been proven to cause mendelian disease but sits within a candidate region of chromosome 4p proposed to cause a mild form Wolf-Hirschhorn syndrome [OMIM:194190]. Classical Wolf-Hirschhorn syndrome presents with a craniofacial cleft and sometimes limb and cardiac morphological abnormalities, severe developmental delay and eventual intellectual disability, and occasionally seizures (Hannes *et al.*, 2012). No mouse model with isolated *Nop14* manipulation has been developed, but a candidate pathogenic variant in the homozygous state was identified in two unrelated Iranian families with recurrent pregnancy loss and knockout resulted in zebrafish lethality at 5 days (Suzuki *et al.*, 2018).

UTP14A is an X-chromosome gene that qualified for inclusion because it has a high probability of being loss-of-function intolerant. It encodes a small subunit processome component, UTP14A found to form a complex with NOP14, and implicated in the nucleolar transport of EMG1, a gene associated with Bowen-Conradi syndrome [OMIM:211180], an autosomal recessive disorder with neurodevelopmental delay, microcephaly, rocker bottom feet, and death in infancy (Warda *et al.*, 2016). No mouse model has been developed beyond cell lines and published studies of both *UTP14A* and *NOP14* have focused on their roles and possible prognostic value in various malignancies (Du *et al.*, 2017; Li *et al.*, 2019; Ren *et al.*, 2019; Lu *et al.*, 2022)

STX12 qualified for inclusion because it has a high probability of being loss-of-function intolerant. It encodes syntaxin 12, a sensitive factor attachment protein receptor (SNARE) that has role in recycling of glutamate receptors rather than typical SNARE neurotransmitter exocytosis (reviewed in Wang and Tang, 2006). It has not been definitively associated a mendelian disorder but was found to be deleted with *AHDC1* (the usual Xia-Gibbs syndrome gene) in a single person with Xia-Gibbs syndrome [OMIM:615829], which presents with facial dysmorphology, laryngomalacia with obstructive sleep apnea, and neurodevelopmental delay and rarely seizures (Chander *et al.*, 2022). The relevance of this to the carrier's syndrome remains unclear. Beyond the nervous system, syntaxin 12 interacts with VPS16B and VPS33B in megakaryocytes (Ambrosio;Febvre and Di Pietro, 2022). Biallelic variants in *VPS33B* can

cause Arthrogryposis, renal dysfunction, and cholestasis 1 [OMIM:208085] and Keratoderma-ichthyosis-deafness syndrome [OMIM:620009], which can be associated with neurodevelopmental delay, morphological changes of the brain (particularly corpus callosum dysgenesis), ataxia, and hypotonia, as well as bleeding. No monogenic mouse model has been developed for *STX12*.

PPFIA1 qualified for inclusion because it has a high probability of being haplosensitive, loss-of-function intolerant, and may have a role in the biology of autism. It encodes the protein-tyrosine phosphatase receptor-type F polypeptide-interacting protein alpha-1, an alpha-liprin. PPFIA1 is thought to regulate synapse maturation in response to synaptic activity via Ca^{2+}/calmodulin-dependent protein kinase II, and rat hippocampal neurons lacking the orthologous liprin-alpha-1 have reduced complexity of dendritic spines and number of synapses (Hoogenraad *et al.*, 2007). Mice carrying the *Ppfia1*$^{em1(IMPC)J}$ allele [MGI:6188059] have been described as having abnormal vocalization and hair pigmentation in the heterozygous state and abnormalities of embryo size and pre-weaning lethality in the homozygous state (International Knockout Mouse Consortium, 2014).

HCFC2 encodes host cell factor 2, and qualified for inclusion because it has a high probability of being missense intolerant, haplosensitive, and loss-of-function intolerant. Little is known about its role thus far, but its overexpression in HEK-293 cells inhibited mitosis and cellular proliferation (Gudkova *et al.*, 2019). Mouse models including crosses of *Hcfc*^{em1Btlr} [MGI:6108107], *Hcfc2*^{fkls} [MGI:5445585], *Hcfc2*^{min} [MGI:6108103], and *Hcfc2*^{sca} [MGI:6108106] have been reported only to have immunological phenotypes (Sun *et al.*, 2017).

SESTD1 qualified for inclusion because it has a high probability of being loss-of-function intolerant and triplosensitive. It encodes SEC14 and spectrin domain containing 1, a phospholipid-binding protein that constrains dendritic spine density and excitatory neurotransmission in cultured hippocampal neurons. Overexpression of SESTD1 resulted in reduced spine density and frequency of miniature post-synaptic currents and reciprocal changes were seen after knockdown (Lee;Huang and Hsu, 2015). While not associated with any mendelian disorder, it was found to contain an intronic SNP associated with lithium-responsive bipolar disorder (Song *et al.*, 2016). It has an important role in embryogenesis as part of the planar cell polarity pathway. *Sestd1*$^{tm1.1Bnrc}$/*Sestd1*$^{tm1.1Bnrc}$ mice [MGI:5520962] develop an abnormal neural tube resulting in spina bifida, insufficient colonic and urological tracts, and a short curly tail (Yang and Cheyette, 2013). SESTD1 is also believed to contribute to growth cone guidance via Ca^{2+}-sensitive interactions with the transient receptor potential cation channels (Miehe *et al.*, 2010).

VPS26B qualified for inclusion because it has a high probability of being haploinsufficient and loss-of-function intolerant. It encodes vacuolar protein sorting 26 and is not associated with mendelian disease. VPS26B forms part of the retromer complex, a scaffold for intracellular endosomal processing, implicated in neurodegenerative disorders (Gustavsson *et al.*, 2015) and reduced in people with trisomy 21 with or without Alzheimer's disease (Chen *et al.*, 2023). It is thought to be contribute to endosome recycling and thus synaptic function, potentially contributing to the regional propensity of Alzheimer's pathology for the trans-entorhinal cortex (Simoes *et al.*, 2021). VPS26B is a target of NEK1, a serine-threonine kinase implicated as susceptibility locus 24 to amyotrophic lateral sclerosis [OMIM:617892] and Short-rib thoracic dysplasia 6 with or without polydactyly [OMIM:263520], which includes neurodevelopmental and neuroanatomical abnormalities, and thought to contribute to pathology through blood-brain-barrier insufficiency (Wang *et al.*, 2021). No behavioral or morphologic phenotype was detected in *Vps26b*^{tm1Ktch}/*Vps26b*^{tm1Ktch} [MGI:4867701] mice (Kim *et al.*, 2010).

POU2F2 qualified for inclusion because it has a high probability of being loss-of-function intolerant and triplosensitive. It is a transcription factor that regulates B cell proliferation and that can promote tumor progression. Increased expression of POU2F2 in glioblastoma multiforme indicates a poor progress, apparently through promotion of glycolysis and activation of the AKT/mTOR pathway (Yang *et al.*, 2021). It may have a role in hypertension through its actions on vascular smooth muscles cells (Yang *et al.*, 2023). In addition to B cell and humoral immunological deficits, mice with *Pou2f2*^{tm1Bal} [MGI:3028896] alleles have lethargic behavior and poor suckling, and those with *Pou2f2*^{tm1Lmc} [MGI:3037094] have neonatal lethality (Corcoran *et al.*, 2004).

5.6 Discussion

This large international effort has enabled me to demonstrate that people with epilepsy who carry an ultra-rare PAV in the same gene often share clinical features, particularly if the gene already has a validated role in epilepsies with mendelian dominant or de novo transmission. The clinical features of the people carrying these URVs often resemble those of people with monogenic epilepsy associated with the gene. It may be that the variants of many of these participants are the cause of their epilepsy and with sufficient supporting information, their variant could be declared diagnostic (Richards *et al.*, 2015). Thus, one interpretation of the results regarding the established dominant epilepsy genes is that formal analysis of phenotypic data collected blind to genotype confirms the phenotypic spectra that have come to characterize

monogenic epilepsies through traditional phenotypic studies. However, beyond the PTVs, one usually requires the support of segregation analysis or functional laboratory assays to establish that a missense variant reaches the diagnostic threshold. We do not have access to familial DNA samples and are left in the situation commonly encountered in the adult epilepsy clinic: the patient has clinical features that fit the gene and the variant is ultra-rare and may look suspicious bioinformatically. However, without parental samples it may not reach the ACMG threshold at which it can be classified as diagnostic. These results suggest that even if a variant cannot be proven to be the cause of a person's epilepsy, it may yet explain some of their clinical features. Consequently, we should consider whether patients found to have URVs and clinical features consistent with a monogenic epilepsy might benefit from precision treatments developed for that monogenic epilepsy, even if their variant does not reach current diagnostic thresholds.

Some surprising associations with established epilepsy genes emerged. For example, *KCNT1*, a gene classically associated with neonatal or early infantile severe Nocturnal Frontal Lobe Epilepsy or Epilepsy in Infancy with Migrating Focal Seizures (McTague *et al.*, 2018), was found to associate strongly with an oral cleft, a common morphological abnormality affecting approximately 15/10,000 live births in Europe (EUROCAT, 2024). This was because three participants with predicted damaging missense variants (not inconsistent with the proposed gain-of-function pathogenic mechanism of variants in monogenic *KCNT1*-related epilepsies) had been annotated with forms of oral cleft. These were the only three members of the cohort of 12,611 to have been annotated with *Cleft lip [HP:0410030]*, suggesting that the phenotype was under-annotated (it seems unlikely that people with epilepsy are less likely to have oral cleft deformities than the general population). As with other associations tested, this may have arisen by statistical chance ($minP < 1.00 \times 10^{-5}$). Alternatively, this may be a previously unrecognized true association, and potential confounders should be considered. First, consider the coincidental carrying of another genetic variant responsible for oral cleft abnormalities in these participants. Two of the three participants annotated with *Cleft lip [HP:0410030]* and dMissense variants in *KCNT1* also carried dMissense variants in *KCNJ1*, *IRF2BPL*, *DMPK*, and variants in other genes with less well-established roles in mendelian disorders. While the syndromes associated with variants in these genes are not known to include oral cleft abnormalities, I cannot exclude the possibility of an alternative monogenic cause or copy number variant being responsible for these participants' oral clefts as my analysis has focused on validated and candidate neurodevelopmental genes rather than oral cleft genes (Wilson;Newbury and Kini, 2023). Such latent variants responsible for the oral cleft abnormalities could be in linkage disequilibrium with the variants assessed in *KCNT1*. Alternatively, systematic biases of data

collection in some centers could combine with genetic stratification to give the appearance of a biological association. However, the three participants annotated with *Cleft lip [HP:0410030]* were recruited from three independent sites (in Italy, Lithuania, and the United States of America) making it unlikely that the apparent association may have resulted from the coincidence of a localized ancestral tendency to carry ultra-rare missense variants in *KCNT1* with recruitment from an Epi25 center with a particular interest in oral clefts. On the other hand, it is possible that enriched annotation of oral cleft abnormalities reflects the early onset of an epilepsy due to variants in *KCNT1*, *KCNJ1*, *IRF2BPL*, *DMPK*, or some other cause in these three participants rather than an increased risk of actually having such morphological abnormalities. Its association could have appeared primarily through confounding due to age at recruitment rather than a true enrichment of oral cleft abnormalities in children with early onset epilepsy. This is a phenotype that adult neurologists may not recognize, particularly if it has resolved or been treated with successful surgery years before. However, a pediatrician would pay greater attention to an oral cleft because of its consequences for feeding or the suspicion that it may form part of a syndromic diagnosis that explains the epilepsy (for example, Cornelia de Lange syndrome which can present with oral cleft and seizures (Kline *et al.*, 2018)) and therefore be more likely to report it to a research study. This hypothesis is consistent with the ages of the three participants: two of the three participants were infants when recruited and considered to have a DEE. The third, recruited from Italy, had data collected at age 15 years (when they were likely to be under the care of a pediatrician) but was annotated as having onset of a DEE with tonic and focal-onset seizures and regression in infancy. Finally, cleft palate is a recognized teratogenic effect of various ASMs, raising the (less probable) possibility of a maternal history of epilepsy sufficiently severe to continue ASM in pregnancy and resulting in utero exposure (Hernandez-Diaz *et al.*, 2018).

Some genes with established contributions to neurodevelopmental disorders that are not primarily considered epilepsies had high similarities and significant associations, even with core epilepsy phenotypes. For example, people with PAVs in *LZTR1* were distinctive because of their shared features suggestive of Lennox Gastaut Syndrome, and with PAVs in *PCGF2*, who had focal epilepsy. This reinforces the pleiotropism of neurodevelopmental genes and consequently the importance of screening for genetic causes beyond the realms of traditional epilepsy genes, as many people with other neurodevelopmental disorders may have recurrent unprovoked seizures (i.e., epilepsy) as part of their syndrome.

This analysis concluded with an exploration of genes that might be hypothesized to harbor URVs that affect neurological phenotypes. The associations and similarities presented propose

hypotheses that can be tested in subsequent clinical and mechanistic studies, potentially confirming a role, be it a novel cause of epilepsy or a gene in which variants might influence the presentation of epilepsy. *KANSL1L* is of particular interest because it harbored URVs relatively commonly (0.36 % of the cohort), appeared to present as a DEE, and would conform to the emerging paradigm of finding multiple members of the same gene family to be implicated in neurodevelopmental disorders (Lal *et al.*, 2020). In this case it would follow the known association of its paralogue *KANSL1* with the neurodevelopmental disorder Koolen de Vries syndrome [OMIM:610443] (Koolen *et al.*, 2016).

Beyond the limitations to the power of these analyses resulting from the sparsity of phenotypic data, there are some important limitations to this study. The major one is that I have been unable to take much account of age, an important dimension of epilepsy syndromes (Wirrell *et al.*, 2022a). The work presented here uses a few age-specific phenotypes (for example, *Infantile spasms [HP:0012469]*, *Febrile seizure (within the age range of 3 months to 6 years) [HP:0002373]*, and the subontology of *Neonatal seizures [HP:0032807]*). I have undertaken preliminary analyses after extending the HPO seizure subontology in a chronological dimension, binning seizure annotations by age of onset. However, the age data available for years 1–3 of Epi25 are limited and this is an approach that I look forward to exploring further with data from subsequent years of Epi25, which are expected to be more complete. I would have liked to identify associations with particularly important phenotypes such as sudden unexpected death in epilepsy and outcomes such as the response to particular ASMs. However, *Sudden unexpected death in epilepsy [HP:0033258]* was annotated to only four participants, leaving insufficient power to detect a robust association. Future studies may prioritize the prospective reporting of sudden unexpected death in epilepsy so that genomic and clinical data collected for cross-sectional studies such as this can be reused to address this rare and important phenotype that is so notoriously difficult to study. With regard to medication, the relevant data in the Epi25 study were limited to the indication of whether or not the participant had drug-resistant epilepsy (Kwan *et al.*, 2010), without details of their medication history. In future, integrated healthcare record systems may allow data including time-stamped prescriptions and CDEs regarding seizure frequency to be linked to research records, enabling pharmacogenomic dimensions to these studies.

Finally, this analysis focused on stratification according to single URVs, defined by their absence from ever-growing population databases. Future studies could explore rare variants with higher minor allele frequencies as well as polygenic contributions to see whether these have a detectable effect on the precise manifestations of a person's epilepsy.

In conclusion, this study has demonstrated the potential of large-scale phenotypic analysis to complement genomic analyses in explaining the genetic basis of the epilepsies. The results indicate that an URV may partially explain a person's detailed clinical features, even without being able to prove that it is the cause of their epilepsy. I look forward to repeating this analysis on cohorts from subsequent years of Epi25 to see which genetic associations can be validated in independent cases and whether further improvements in data capture and modeling might yield greater discoveries.

Chapter 6. Genome-wide identification and phenotypic characterization of seizure-associated copy number variations in 741,075 individuals

6.1 Preface

This chapter presents my work applying phenotypic analysis to 10,880 people from years 1–3 of Epi25 according to the CNVs that they were found to carry. These are a subset of the 12,611 referred to in Chapter 5, for whom CNVs were called and passed genomic quality control. My work in this chapter was conducted in collaboration with Prof. Dennis Lal's and Prof. Michael Talkowski's research groups, with their contributions led by Dr Ludovica Montanucci and Dr Ryan Collins, respectively. They called the CNVs, performed the fine mapping of credible intervals for the epilepsy liability associated with each CNV using a Bayesian model, and drafted the text and figures relating to these. This collaboration culminated in the following publication, the author accepted manuscript of which is reproduced and modified here under CC-BY, with formatting and stylistic modifications, and the addition of a phenotypic similarity analysis (Methods 6.4.10, Results 6.5.5).

Montanucci, L.,† Lewis-Smith, D.,† Collins, R.L.,† Niestroj, L.M.,† Parthasarathy, S., Xian, J., Ganesan, S., Macnee, M., Brunger, T., Thomas, R.H., Talkowski, M., Epi25 Collaborative, Helbig, I., Leu, C.‡ and Lal, D.‡ (2023) 'Genome-wide identification and phenotypic characterization of seizure-associated copy number variations in 741,075 individuals', Nat Commun*, 14(1), p. 4392.*[29]

I cleaned the phenotypic data, designed, and conducted the phenotypic analysis, produced the phenotypic illustrations included in this chapter (including modifying the genetic figures stylistically), and drafted the clinical and phenotypic sections of the introduction, methods, results, and discussion, as well as contributing to the prioritization of candidate genes within each credible interval having reviewed the literature. With regards to the phenotypic analyses that constitute my work, co-authors assisted with the generation of statistical association plots for the publication but not included in this chapter (S.P., J.X., S.G.), and supervised my work (R.H.T., I.H.). I thank *Nature Communication*'s reviewers and editor (Dr. Margot Brandt) for their comments.

[29] † These authors contributed equally. ‡ These authors jointly supervised this work.

6.2 Abstract

CNV are established risk factors for neurodevelopmental disorders with seizures or epilepsy. With the hypothesis that seizure disorders share genetic risk factors, we pooled CNV data from 10,590 individuals with seizure disorders, 16,109 individuals with clinically validated epilepsy, and 492,324 population controls and identified 25 genome-wide significant loci, 22 of which are novel for seizure disorders, such as deletions at 1p36.33, 1q44, 2p21-p16.3, 3q29, 8p23.3-p23.2, 9p24.3, 10q26.3, 15q11.2, 15q12-q13.1, 16p12.2, 17q21.31, duplications at 2q13, 9q34.3, 16p13.3, 17q12, 19p13.3, 20q13.33, and reciprocal CNVs at 16p11.2, and 22q11.21. Using genetic data from additional 248,751 individuals with 23 neuropsychiatric phenotypes, we explored the pleiotropy of these 25 loci. Finally, in a subset of individuals with epilepsy and detailed clinical data available, we performed phenome-wide association analyses between individual CNVs and clinical annotations categorized through the HPO. For six CNVs, we identified 19 significant associations with specific HPO terms and generated, for all CNVs, phenotype signatures across 17 clinical categories relevant for epileptologists. This is the most comprehensive investigation of CNVs in epilepsy and related seizure disorders, with potential implications for clinical practice.

6.3 Introduction

An epileptic seizure is a paroxysm of symptoms and signs due to abnormally excessive or synchronous neuronal activity (Fisher *et al.*, 2014). Seizures are classified based on their characteristics and electroencephalogram (EEG) as focal-onset seizures (which start in a specific brain region) and generalized-onset seizures (which are rapidly seen across bihemispheric networks) (Fisher *et al.*, 2014; Fisher *et al.*, 2017b). The utility of this seizure classification is that it categorizes epilepsy into syndromes and allows clinicians to make implications about disease etiology, trajectory, and response to medication. Clinical manifestations vary from whole-body convulsions with loss of consciousness (tonic-clonic seizures), to movements involving only part of the body with variable levels of consciousness (focal motor seizure), to a brief loss of awareness (absence seizure) (Fisher *et al.*, 2014; Fisher *et al.*, 2017b). Seizures can be provoked by head trauma, infection, or acute toxic-metabolic imbalance, or they can be spontaneous and unprovoked. Individuals who exhibit at least one unprovoked seizure with an enduring elevated risk of further seizures or who have the electroclinical features of one of a few specific epilepsy syndromes that can be diagnosed without recurrent seizures fulfill the criteria for a diagnosis of epilepsy (Fisher *et al.*, 2014). Seizures and epilepsy are common in the general population. Neonatal seizures occur in 1.5 %

of neonates, febrile seizures in 2–4 % of young children, and epilepsy in up to 1 % of children and adolescents (Berg;Jallon and Preux, 2013). Seizures are common among individuals with neurodevelopmental disorders, affecting 21.5 % of those with autism and intellectual disability and 8 % with autism without intellectual disability (Amiet *et al.*, 2008).

CNVs such as deletions and duplications, change the dosage of genomic segments and are established risk factors for various types of epilepsy (Dibbens *et al.*, 2009; Helbig *et al.*, 2009; de Kovel *et al.*, 2010; Heinzen *et al.*, 2010; Sisodiya and Mefford, 2011; Mefford, 2014; Olson *et al.*, 2014; Lal *et al.*, 2015; Addis *et al.*, 2016; Pérez-Palma *et al.*, 2017), seizures (Fortin *et al.*, 2020), and neuropsychiatric disorders (Pinto *et al.*, 2014; Sanders *et al.*, 2015; Leppa *et al.*, 2016; Takumi and Tamada, 2018). Large CNVs can affect multiple dosage-sensitive genes, leading to complex clinical presentations. To date, only one hypothesis-free genome-wide CNV association study has been reported for epilepsy (Niestroj *et al.*, 2020). This genome-wide CNV association study in 10,712 individuals with epilepsy and 6,746 controls identified three genome-wide significant CNVs (Niestroj *et al.*, 2020). High-resolution CNV screening has become routine in clinical molecular diagnostics, leading to greater detection of chromosomal abnormalities in patients (Coppola *et al.*, 2019). Diagnostic CNVs can be identified in 1–4 % of individuals with epilepsy and > 10 % of those with seizures and neurodevelopmental disorders (Pérez-Palma *et al.*, 2017; Coppola *et al.*, 2019; Niestroj *et al.*, 2020; Sheidley *et al.*, 2022). However, the pleiotropy of pathogenic CNVs, partially driven by structural properties (size, fixed versus variable breakpoints, number of affected genes), represents a significant challenge in the clinical interpretation of CNVs, limiting their utility for disorder classification, prognostication, and the development of precision medicine treatments that specifically target the critical pathogenic gene(s) altered by the CNV. The majority of pathogenic and likely pathogenic CNVs are greater than 1 megabase (Mb) in size, and it is often unclear which gene(s) or genomic element(s) affected by the CNV contribute to one or more disorders (Riggs *et al.*, 2020; Okur *et al.*, 2021; Riggs *et al.*, 2021). A well-powered seizure CNV discovery screen combined with detailed genotype-phenotype analyses could identify genomic segments that confer risk for seizures, identify clinical characteristics in affected patients and consequently guide genetic test interpretation.

Although many individuals with neuropsychiatric and developmental disorders have comorbid seizures, genome-wide CNV association analyses across epilepsy and seizure have yet to be reported. We hypothesized that genetic risk for seizures is shared in individuals with epilepsy diagnosed according to ILAE criteria (Fisher *et al.*, 2014) and related neurological and neurodevelopmental disorders who also have seizures. Therefore, a joint analysis could add to

the three epilepsy-associated CNV loci reported previously (Niestroj *et al.*, 2020). To explore this hypothesis, we performed a meta-analysis of genome-wide association studies (GWAS) comprising 26,699 individuals with diagnosed epilepsy or seizures and 492,324 controls. Since both definitions are based on the presence of seizures, we refer to individuals affected by either condition as individuals with seizures from here on forward. The effective sample size of this study (n_{eff} = 101,302) provides adequate power to identify significant associations of risk CNVs that are present in the general healthy population, therefore, do not exhibit complete penetrance. However, the analytic setup restricts the frequency in the general population to up to 1 % for quality purposes. We assessed the pleiotropy of any identified seizure-associated CNV in subsequent meta-analyses of epilepsy and 238,161 independent individuals affected by a range of 23 neuropsychiatric disorders. Finally, using a subset of the seizure cohort comprising 10,880 individuals with epilepsy detailed using 214,203 HPO annotations (Kohler *et al.*, 2021), we evaluated the clinical features characterizing carriers of each seizure-associated CNV.

6.4 Methods

6.4.1 Study cohorts

Each center's ethics committees/institutional review boards approved data collection and use. For the Epi25 cohort, patients or their legal guardians provided signed informed consent/assent according to local IRB requirements; as samples had been collected over 20 years in some centers, forms reflected standards at the time of collection. For Epi25 samples collected after 2015-1-25, forms required specific language according to the NIH Genomic Data Sharing Policy.

6.4.2 Individuals with clinically defined epilepsy - Epi25 Collaborative

Individuals with ILAE-defined epilepsy (n = 16,109) were collected through Epi25 . The epilepsy diagnosis was performed according to clinical criteria (clinical interview, neurological examination, EEG, imaging data), following ILAE classifications (Berg *et al.*, 2010). All cohorts are detailed in **Table 6-5**. All individuals of the Epi25 cohort were selected to be of PCA-defined European ancestry. Ancestry-matched population controls (n = 8,545) for the Epi25 arm of the study were recruited through (i) Epi25, (ii) a Broad Institute project on inflammatory bowel disease without reported epilepsy (part of the Inflammatory Bowel Disease Genetics Collaborative), (iii) healthy individuals from the Genetics of Personality Collaborative, and (iv) the THL Institute for Health and Welfare, a subsample of the FINRISK

study (Borodulin *et al.*, 2018). Genotyping for all cases and controls was performed on the same genotyping array (Illumina Infinium Global Screening Array, GSA-MD version 1.0) and at the same center (Broad Institute) as the epilepsy cases. For a detailed description, see Niestroj *et al.* (2020).

6.4.3 CNV calling and quality control - Epi25 Collaborative

We restricted our analysis to only autosomal CNVs due to a higher quality of calls and followed the quality control pipeline developed in our previous study (Niestroj *et al.*, 2020). In detail, quality control was performed in two major steps: (i) pre-CNV calling quality control and (ii) post-CNV calling quality control. For pre-CNV calling quality control, we excluded samples with a call rate < 0.96 or discordant sex status. To select individuals of European ancestry, we filtered autosomal SNPs for low genotyping rate (< 0.98), a high difference in the SNP minor allele frequency between cases and controls (> 0.05), deviation from Hardy-Weinberg equilibrium with $p \leq 0.001$), and pruned the remaining SNPs for linkage disequilibrium (-indep-pairwise 200 100 0.2) using PLINK v1.9 (Chang *et al.*, 2015). We then performed a PCA of the Epi25 cases and controls using PLINK and GCTA (Yang *et al.*, 2011). European individuals were defined as individuals clustering with the 1000 Genomes Project (1000 Genomes Project Consortium *et al.*, 2015) European samples. We created GC wave-adjusted Log-R ratio intensity files for all samples using PennCNV, generated a custom population B-allele frequency file, and employed PennCNV's CNV calling algorithms (Wang *et al.*, 2007b) to detect CNVs in our dataset. The post-CNV calling quality control included the following steps: (i) CNV calls of the same type (deletion or duplication) were merged if the number of SNP/intensity markers between them was < 20 % of the total number when both segments were combined; (ii) CNVs supported by < 20 markers, < 20 kb long, and with a SNP density < 0.0001 were excluded from subsequent analyses; (iii) CNVs that overlapped other CNVs in ≥ 1 % of all samples within the Epi25 dataset were excluded to remove potential platform-specific artifacts, (iv) CNVs with > 50 % overlap with telomeric, centromeric, and immunoglobulin regions of the hg19 reference assembly were excluded; (v) CNVs with ≥ 50 % overlap with reported common CNVs (allele frequency > 1 %) in two independent CNV reference catalogs: DGV Gold Standard Dataset (Lee *et al.*, 2020) and DECIPHER Population Copy-Number Variation Frequencies (Bragin *et al.*, 2014) were excluded. Finally, the probe-level intensity plots of all CNVs supporting the seizure-associated regions (**Table 6-2**) were visually inspected to exclude any remaining artifacts. The DGV Gold Standard and DECIPHER Population frequencies of the remaining CNVs are given in **Table 6-7**.

6.4.4 Individuals with seizures or neuropsychiatric phenotypes

A large CNV dataset from individuals with a range of neuropsychiatric disorders (including seizure disorders) was aggregated from 17 different sources by Collins *et al.* (2022). The contributors of each cohort provided the specific clinical phenotypes. The aggregated individuals were grouped into 54 partially overlapping disease phenotypes standardized through the Human Phenome Ontology (Kohler *et al.*, 2019). The 54 different phenotypes of Collins *et al.* (2022) were obtained through a recursive hierarchical clustering that defined a minimal set of nonredundant primary phenotypes, each including a minimum of > 300 samples in at least three independent cohorts, > 3,000 samples in total across all cohorts, and had less than 80 % sample overlap with any other phenotype. Of the 54 phenotypes, we only selected neurological and psychiatric HPO-based phenotypes (n = 23, excluding Seizures, **Table 6-6**). The architecture of these HPO-based phenotypes allows the identification of associations at different levels, from broad to narrow phenotypes, providing the opportunity to distill between pleiotropic and specific associations. This data set also included the Epi25 cohort from our previous genome-wide CNV study (Niestroj *et al.*, 2020). This previous (outdated) Epi25 cohort was excluded from the neuropsychiatric cohort for cross-disorder meta-analyses in the present work. All the considered cohorts are listed in **Table 6-5**. This aggregated CNV dataset comprised 248,751 individuals affected by at least one of 24 neuropsychiatric disorders, including 10,590 individuals with seizures and 483,779 population controls.

6.4.5 Quality control - neuropsychiatric disorders cohort

The CNV harmonization procedure for the Neuropsychiatric cohort is described in the Supplementary Materials of Collins *et al.* (2022) and included following steps: (i) CNV calls of the same type (deletion or duplication) were merged if their breakpoints were within ±25 % of the size of their corresponding original CNV calls to avoid over-segmentation of large CNV calls; (ii) CNVs not mapped to autosomes from the primary hg19 assembly were excluded; (iii) only CNVs between ≥ 100 kb and ≤ 20 Mb in size were considered; (iv) CNVs that matched reported common CNVs (allele frequency > 1 %) in three independent CNV reference catalogs derived from genome sequencing Sudmant *et al.* (2015), Abel *et al.* (2020), and Collins *et al.* (2022) were excluded; (v) CNVs that overlapped other CNVs in ≥ 1 % of samples within the same dataset or in any of the other array CNV datasets were excluded to remove potential platform specific artifacts; (vi) we excluded all CNVs with ≥ 30 % overlap with somatic

hypermutable sites, segmental duplications, simple/low-complexity/satellite repeats, or *n*-masked bases of the hg19 reference assembly.

6.4.6 Genome-wide association analysis

We performed segment-based CNV burden analyses to identify genomic regions with a significant increase of CNVs in epilepsy cases compared to controls, separated by CNV type (deletion or duplication). We adopted a sliding window approach as introduced by Collins *et al.* (2022). The sliding windows model allowed association testing of all autosomes through 267,237 sliding windows characterized by a window size of 200 kb and a step size of 10 kb, corresponding to 13,339.6 non-overlapping windows. Each of these windows was required to have a low overlap with hypermutable sites, segmental duplications, simple/low-complexity/satellite repeats, and N-masked regions (> 30 %). For each of the genomic regions, we counted the number of overlapping CNVs separately for cases and controls for each CNV type (deletion or duplication). We required an overlap between the CNV and the genomic window of ≥ 10 % to reveal the potential burden of small deletions or duplications (size ≥ 20 kb). We used the one-sided Fisher test as the test statistic for the CNVs collapsed for each segment. Cases/control CNV counts and the Fisher tests were performed using the CNV docker available at https://hub.docker.com/r/talkowski/rcnv and custom python (version 3.7.9) and R (version 3.6.1) scripts. The same procedure was applied to the cohorts of the neuropsychiatric disorder dataset, as detailed in Collins *et al.* (2022).

6.4.7 Meta-analysis and fine-mapping

Fixed-effects meta-analyses were performed using the *metafor* (version 3.6.1) package with an empirical continuity correction (Sweeting;Sutton and Lambert, 2004) and a saddlepoint re-approximation of the null distribution used for inference. The meta-analysis procedure is detailed in Collins *et al.* (2022). We meta-analyzed the effect sizes from 7 GWAS derived from the 17 cohorts of the neuropsychiatric disorder dataset with each segment-based *p*-value of the Epi25 dataset. The threshold for genome-wide significance was set to $\alpha = 3.74 \times 10^{-6}$ after Bonferroni correction for multiples testing corresponding to the number of independent, non-overlapping 200 kb windows, calculated by merging all overlapping windows and dividing the sum of their sizes by 200 kb (effective n = 13,339.6 independent windows; $p = 3.74 \times 10^{-6}$). To account for possible cohort-specific biases, we expected each segment to fulfill the following additional criteria: (i) at least two cohorts featuring nominal significant *p*-values ($p < 0.05$) for the given segment, and (ii) a meta-analysis $p < 0.05$ after excluding the single most significant

cohort. We then used a Bayesian algorithm (Wakefield, 2009) to identify the minimal credible interval(s) that contained the causal element(s) or genes with 95 % confidence, as in Collins *et al.* (2022). Finally, we explored the known biological function of all genes within the credible intervals and performed pathway analyses using Enrichr (https://maayanlab.cloud/Enrichr/, Chen *et al.*; Kuleshov *et al.*, 2016). All resources used to investigate the knowledge basis of all seizure-associated CNV regions are described in **Table 6-8**.

6.4.8 Detailed HPO characterization of Epi25 participants

To identify phenotypic associations with each of the CNVs within a cohort of individuals with epilepsy, we translated clinical data from years 1–3 of the deeply phenotyped Epi25 Collaborative international cohort into HPO (version released 2022-2-14) concepts, following our optimization of the HPO for epilepsy phenotypes (Epi25 Collaborative, 2019; Lewis-Smith *et al.*, 2021a). We selected only individuals with CNV data and sufficiently detailed clinical data (as of 2022-01-25) to confirm the presence of seizures or epileptic encephalopathy with spike-and-wave activation in sleep (EE-SWAS, an epilepsy syndrome in which overt clinical seizures may not always be observed). Categorical clinical data were mapped to HPO concepts using a data dictionary. Free text data were annotated with HPO terms manually by D.L.-S. under the supervision of I.H. and R.H.T. (Kohler *et al.*, 2021). Quantitative data related to the gestational age, weight, and head circumference at birth were categorized to match HPO definitions using sex-stratified distributions from the INTERGROWTH-21th Project using the *R growthstandards* package (version 0.1.5, Villar *et al.*, 2014).

We inferred all HPO concepts applicable to each individual from those translated from the clinical data by propagation, following the *is_a* relationships between HPO concepts as previously described (Lewis-Smith *et al.*, 2021a), using the R *ontologyIndex* package (version 2.7, Greene;Richardson and Turro, 2017). We excluded HPO terms that carried no information in the context of this cohort (those that were annotated ubiquitously) and modified the relationships of others, tailoring them to this analysis (**Table 5-1**). Phenotypes were annotated as being explicitly present or not, without annotating any phenotypes as being explicitly absent. Taking this open-world perspective is conservative, meaning that the proportion of individuals in a group annotated with a particular phenotype should be considered a lower limit while still allowing statistical testing of phenotypic associations and mitigating the risk of explicitly annotating a phenotype as absent when it was present but not recorded or the individual will manifest the phenotype at some point in the future (Lewis-Smith *et al.*, 2022).

After excluding individuals with markers of acquired epilepsy that are unlikely to be part of the phenotype, such as significant brain trauma, encephalitis, or meningitis, 10,880 individuals from the genomic analysis had adequate phenotypic data available for analysis. Of these, 10,106 individuals are of European ancestry, 602 of East Asian ancestry, and 172 of African ancestry, according to PCA analysis. After propagation to infer generic phenotypic descriptors from specific ones, this cohort had 214,203 informative annotations (median = 17 per individual, range = 1–128), spanning a repertoire of 1,667 phenotypic concepts. The proportion of the phenomic cohort annotated with clinically selected phenotypes is provided in **Table 6-1** (data for all 1,667 phenotypes are available in **Supplementary Table 6-3)**.

HPO CUI	HPO name	Number of annotated individuals	Percentage of annotated individuals (%)
HP:0001250	Seizure	10,869	99.9
HP:0002353	EEG abnormality	9,127	83.9
HP:0007359	Focal-onset seizure	6,991	64.3
HP:0002069	Bilateral tonic-clonic seizure	5,772	53.1
HP:0011146	Dialeptic seizure	5,608	51.5
HP:0011185	EEG with focal epileptiform discharges	4,987	45.8
HP:0033259	Non-motor seizure	3,959	36.4
HP:0011198	EEG with generalized epileptiform discharges	3,924	36.1
HP:0002384	Focal impaired awareness seizure	3,866	35.5
HP:0002197	Generalized-onset seizure	3,786	34.8
HP:0012443	Abnormality of brain morphology	3,695	34.0
HP:0033717	EEG with temporal epileptiform discharges	3,489	32.1
HP:0007334	Bilateral tonic-clonic seizure with focal onset	3,168	29.1
HP:0032677	Generalized-onset motor seizure	2,931	26.9
HP:0020174	Refractory drug response	2,658	24.4
HP:0012759	Neurodevelopmental abnormality	1,895	17.4

Table 6-1. The percentage of 10,880 Epi25 participants included in the phenomic analysis cohort annotated with a clinically-prioritized selection of common epilepsy phenotypes.

6.4.9 Phenome-wide association analysis of CNVs

All association analyses and phenomic visualizations were performed in R. Associations between CNVs, and HPO concepts were calculated using the two-sided Fisher's exact test (*stats::fisher.test()*). The tested phenotypes were all those 1,667 HPO terms translated from clinical data that were informative (not ubiquitous) and are detailed in **Supplementary Table 6-4**. While this was a descriptive analysis, given a large number of tests performed,[30] we sought to aid identification of the most robust associations. Bonferroni's single step and Holm's step-down adjustments are overly conservative given the dependence structure of propagated HPO annotations. For example, after full harmonization, annotations of *Typical absence seizure*

[30] 1,667 informative HPO concepts x (32 groups of multiple individuals + 3 groups of a single individual) = 58,345

[HP:0011147], *Generalized non-motor (absence) seizure [HP:0002121]*, and *Generalized-onset seizure [HP:0002197]* will be highly correlated because an individual cannot have the first without the second or the second without the last as a result of there *is_a* relationships in the HPO. Therefore we applied the *minP step-down* procedure, which uses a permutation-based approach to control the family-wise error rate (Westfall and Wolfinger, 1997). We selected 100,000 randomly generated groups of individuals from the Epi25 phenomic analysis cohort of size *k,* where *k* is the number of carriers of each CNV. Then for each of these groups, we calculated the two-sided Fisher's exact test *p*-values for every one of the 1,667 HPO concepts. We used the *NRejections::adj_Wstep* function (version 1.2.0) in R to perform the step-down procedure. This generated *p*-values corrected for the correlation-adjusted number of tested HPO annotations. We did not adjust *p*-values across CNVs because we were interested only in identifying those associations that were most robust in this descriptive analysis.

6.4.10 Phenotypic similarity analysis

We compared the Epi25 carriers of each of the 32 CNV credible intervals with multiple carriers across all of the phenotypic annotations according to $\text{Sim}_{\text{MinTO}}$ and five supplementary phenotypic similarity measures: $\text{Sim}_{\text{Phrank}}$, $\text{Sim}_{\text{Resnik-BMS}}$, $\text{Sim}_{\text{Resnik-BMA}}$, $\text{Sim}_{\text{Rel-BMS}}$, and $\text{Sim}_{\text{Rel-BMS}}$. Details of these can be found in Chapter 4, but a brief summary follows for convenience.

$\text{Sim}_{\text{MinTO}}$ and $\text{Sim}_{\text{Phrank}}$ are weighted measures calculated from the intersection of a pair of peoples propagated HPO annotations. $\text{Sim}_{\text{MinTO}}$ defines the similarity of a pair of people as the sum of the monotonically-adjusted IC of the set of HPO terms necessary and sufficient to summarize all their shared clinical features. This minimal set can reconcile the intersection of the pair's propagated HPO annotations by propagation. $\text{Sim}_{\text{Phrank}}$ defines the pairwise similarity of two people as the sum of the IC of each term in the intersection of their propagated sets, conditional on the IC of their parent terms. This mitigates variability in nodal density across the HPO. The remaining four measures are semantic similarity approaches in which the minimal set of annotations of each person in the pair are compared exhaustively to produce a matrix of term-term similarities. Term-term similarity measures in this analysis are based on Resnik's shared information ($\text{Sim}_{\text{Resnik}}$, the IC of each pair's MICA), or Relevance similarity (Sim_{Rel}), which is a form of Lin's similarity (Sim_{Lin}) weighted by the specificity of the pair's MICA. This matrix is then summarized to give a single symmetric similarity of the two people based on the mean of the sum (Sim_{BMS}) or mean (Sim_{BMA}) of the best matches of each person's

set of annotations. The distribution of each of these similarity measures is discrete and varies by the size of group, n. Consequently, we interpret the raw similarity values based on Monte Carlo simulation. An unadjusted one-sided p-value is calculated based on the proportion of 10^6 randomly selected groups of size n achieving a raw similarity at least as great as that observed for the group of interest.

6.4.11 Data availability

All genome-wide CNV association summary statistics are available at Zenodo (zenodo.org/record/7939126#.ZGK7yi-B29Y with DOI 10.5281/zenodo.7939126). Individual-level CNV data for epilepsy patients are available from the Epi25 Consortium (http://epi-25.org/) upon reasonable request and approved ethics protocol. Furthermore, raw data is deposited at dbGAP (www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001551.v1.p1). All HPO-based phenome-wide summary statistics with raw $p < 0.05$ are available in **Supplementary Table 6-4** of this manuscript. Fine-mapping results are available in **Table 6-3** and **Supplementary Tables 6-1** and **6-2** of this manuscript. The CNV data of the Neuropsychiatric cohort are described in the Supplementary Materials of Collins *et al.* (2022). They can be accessed from existing publications, public resources, or, upon request, from the authors of Collins *et al.* (2022) (see 'Key resources table' and Table S2 in Collins *et al.* (2022)). The CNV data reported by *GeneDx* and Indiana University clinical testing sites were not consented for public release. All datasets used in this study are detailed in **Table 6-5** of our manuscript.

6.4.12 Code availability

The code for the association and meta-analysis is available and have been deposited at Zenodo (github.com/talkowski-lab/rCNV2/tree/v1.0, with DOI 10.5281/zenodo.6647918). Also, we provided a Docker image hosted on DockerHub (hub.docker.com/r/talkowski/rcnv) and Google Container Registry (gcr.io/gnomad-wgs-v2-sv/rcnv), which provides a controlled container environment containing all dependencies necessary to execute the code identically as presented in this study.

6.5 Results

6.5.1 Discovery of 25 genome-wide significant seizure-associated CNVs regions

We performed a meta-analysis of 16,109 individuals with epilepsy and 8,545 population controls (Epi25 cohort) with 10,590 individuals with seizures (not explicitly meeting diagnostic criteria for epilepsy) and 483,779 population controls, derived from an aggregated CNV dataset of 17 cohorts (neuropsychiatric disorders cohort) (see all cohorts of this study in **Table 6-5**). The genome was scanned using 267,237 genomic segments of 200 kb size in a 10 kb sliding window approach (Collins *et al.*, 2022). After applying Bonferroni correction of the threshold for a significant association in the meta-analysis and fine-mapping, we identified 25 loci associated with seizures at genome-wide significance ($p \leq 3.74 \times 10^{-6}$). All 25 loci are shown in **Figure 6-1** and detailed in **Table 6-2**. The 25 identified loci included 15 deletion CNVs (size range: 230 kb to 5 Mb) and ten duplication CNVs (size range: 290 kb to 8.9 Mb). All the genome-wide associated deletions found in this study consisted of the loss of one copy, while all duplications consisted of the gain of one copy. Three of the 25 seizure-associated loci (15q11.2-q13.3 dup, 15q13.2-q13.3 del, 16p13.11 del) had previous genome-wide statistical support for an association with epilepsy from our previous study (Niestroj *et al.*, 2020) that included 40 % of the individuals with seizures of this study. All other identified CNVs (22/25, 88 %) represent new genome-wide significant loci for seizures, with 10/22 (59 %) loci previously implicated in neurological and psychiatric disorders, 6/22 (23 %) specifically in epilepsy by studies without genome-wide statistical support, 2/22 (9 %) reported in individuals without neurological or psychiatric disorders, and 4/22 (18 %) not previously reported regions. We detailed in **Table 6-3** all commonly reported diseases for the 25 identified seizure-associated loci. Our meta-analysis in seizure disorders was likely not powered enough to identify some of the known CNVs implicated in epilepsy (without genome-wide statistical support) associated with seizures (for example, 1q21.1 del/dup). Reciprocal CNVs, defined by deletions and duplications associated with seizures involving overlapping genomic segments, were found at 15q11.2, 16p11.2, and 22q11.21. No overlap existed between the seizure-associated CNV regions identified in this study and the most recent SNP-based GWAS study in epilepsy (ILAE Consortium on Complex Epilepsies, 2023).

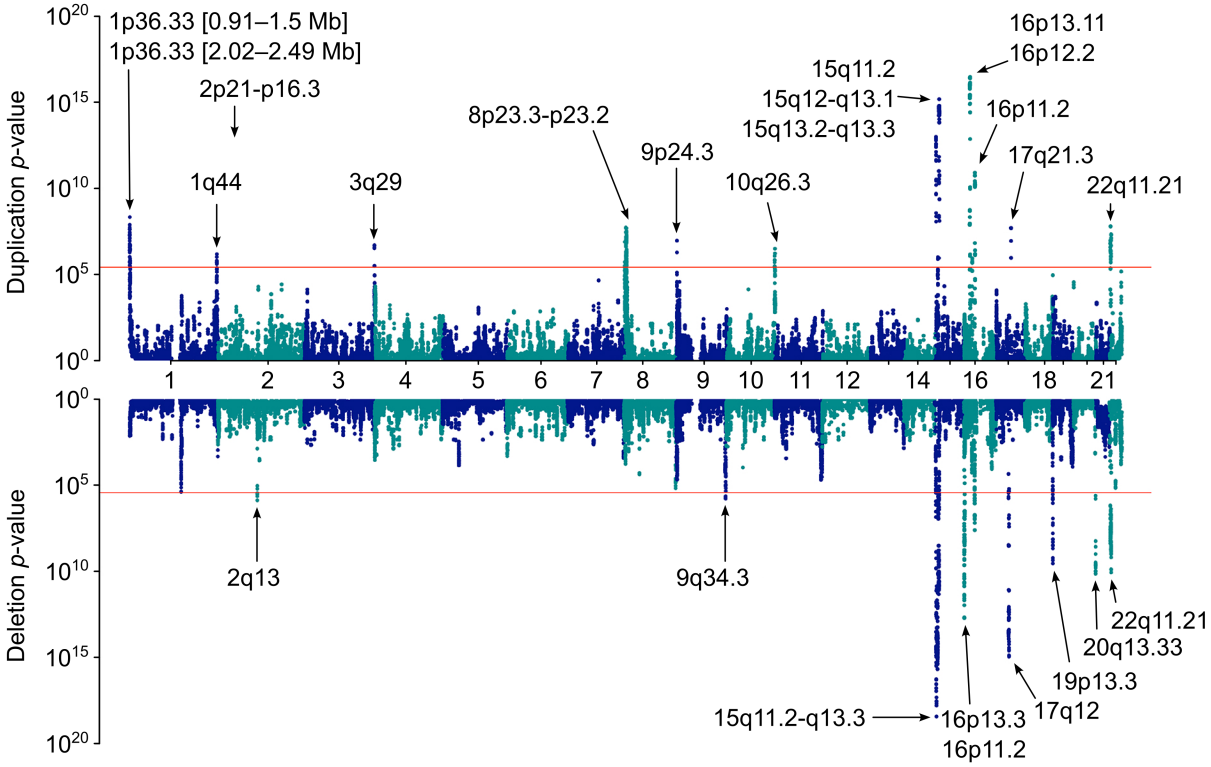


Figure 6-1. Genome-wide meta-analysis identifies 25 CNVs associated with seizure disorders. Miami plot (deletions above with duplications beneath) of the meta-analysis of the CNV genome-wide association analyses of (i) 16,109 individuals with clinically validated epilepsy versus 8,545 controls and (ii) 10,590 individuals with seizure disorders versus 483,779 controls. Dots represent -log10 of the meta-analysis *p*-values of the cohort-specific Fisher exact tests for the enrichment of CNVs in cases versus controls for each a 200 kb sliding window. Genomic regions surpassing the Bonferroni-corrected threshold for significance (red line, α = 3.74x10-6) were annotated.

6.5.2 Fine-mapping and candidate genes

Out of the three CNV regions with previous genome-wide statistical support, our fine-mapping approach narrowed down the critical seizure-relevant region for the known 15q11-q13 duplication to the imprinted promoter/exon 1 region of *SNPRN* (**Table 6-3** and **Figure 6-2**). The *SNRPN* promoter/exon 1 region was suggested to regulate the imprinting of the critical region for *Prader-Willi syndrome [OMIM:176270]* (Ohta *et al.*, 1999; Bielinska *et al.*, 2000). Overexpression of *SNRPN*, corresponding to the seizure-associated duplication of the region, was found to cause abnormal neural development in cultured primary cortical neurons (Li *et al.*, 2016). Conversely, *SNRPN* knockdown was found in the same study to also cause subtle neuronal abnormalities, in line with reports of short *SNRPN* deletions in Prader-Willi syndrome (Grootjen *et al.*, 2022). For the other two CNV regions with previous genome-wide statistical support, we identified several genes with a brain phenotype in the minimal credible intervals. The 15q13.2-q13.3 deletion credible interval includes the haploinsufficient gene

OTUD7A, shown to cause abnormal development of cortical dendritic spines and dendrite outgrowth in *Otud7a*$^{DEL/+}$ mice (Uddin *et al.*, 2018), and *KLF13*, shown to cause a layer-specific decrease of cortical interneurons in *Klf13*$^{DEL/+}$ mice (Malwade *et al.*, 2022). The 16p13.11 deletion credible interval includes two haploinsufficient genes: *MYH11*, implicated in cerebrovascular disorders (Aortic aneurysm, familial thoracic 4 [OMIM: 132900] Ravindra *et al.*, 2016; Keylock *et al.*, 2018) that are a risk factor for seizures (Steinlin, 2013), and *MARF1*, involved in cortical neurogenesis (Kanemitsu *et al.*, 2017).

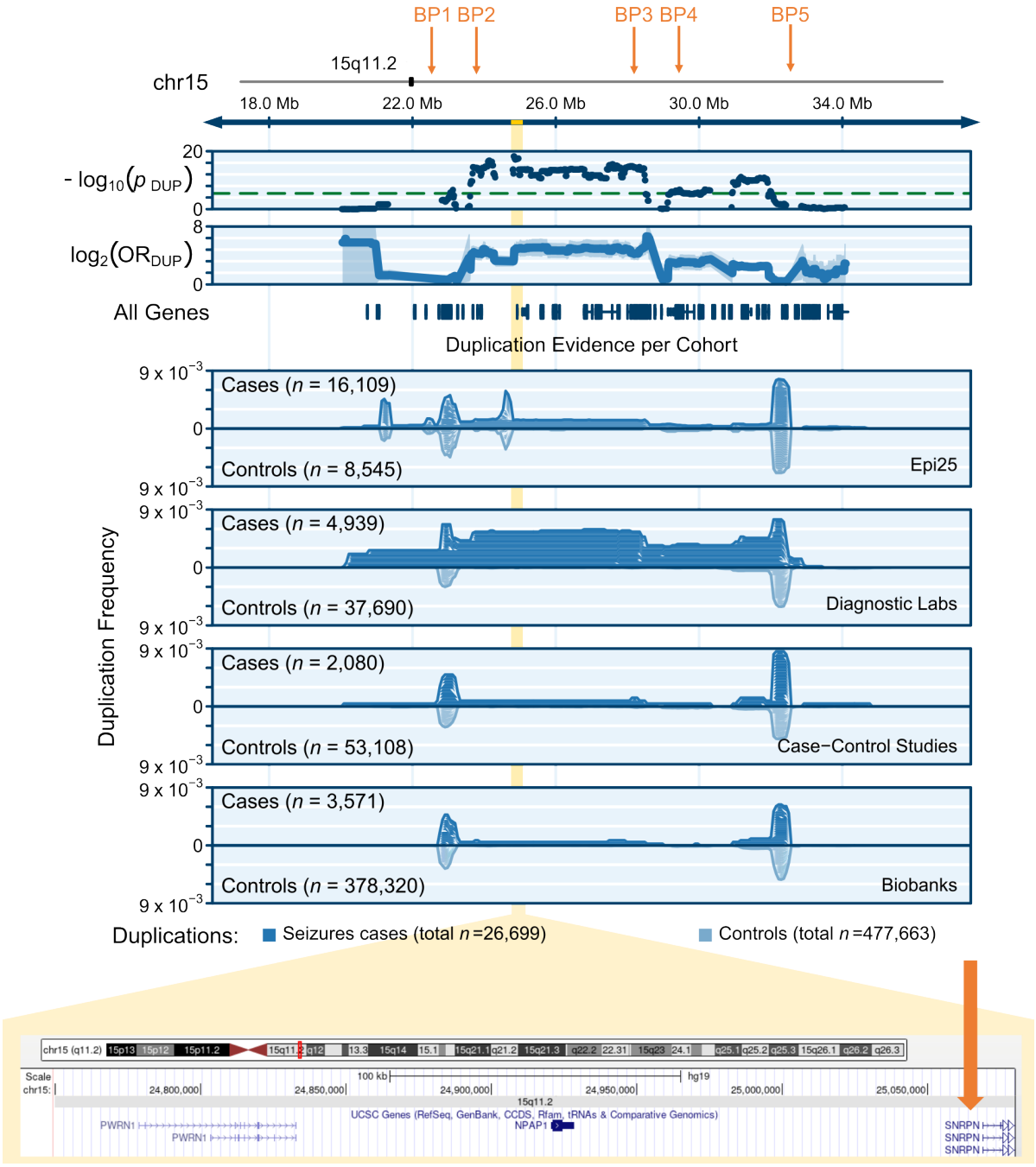


Figure 6-2. Regional CNV distribution plot and fine-mapping of the 15q11-q13 duplication. Known breakpoints are labeled BP1 to BP5. The dashed line in the first plot represents the Bonferroni-corrected threshold for genome-wide significance, $\alpha = 3.74 \times 10^{-6}$. The credible interval containing the causal element/gene with 95 % confidence is highlighted in yellow.

Cytoband	CNV type	Hg19 Start (Mb)	Hg19 End (Mb)	Lowest *p*-value in region	OR (95 % CI)	Credible interval containing the causal element/gene with 95 % confidence	Number of associations with neuropsychiatric phenotypes	Highest odds ratio in neuropsychiatric disorder/seizure meta-analyses (95 % CI)
1p36.33	DEL	0.91	1.51	4.65×10^{-9}	23 (10-54)	910000-1510000	5	11 (5-24) CNS abnormality
1p36.33	DEL	2.02	2.49	1.95×10^{-7}	44 (14-141)	2020000-2490000	3	48 (14-170) Abnormalities of Cognition
1q44	DEL	245.29	245.86	6.59×10^{-7}	41 (13-133)	245290000-245860000	6	62 (19-207) Abnormal brain morphology
2p21-p16.3	DEL	47.5	47.85	3.00×10^{-13}	12 (7-22)	47500000-47850000	23	13 (8-24) Behavioral abnormalities
2q13	DUP	110.77	111.06	1.33×10^{-6}	3 (2-5)	110770000-111060000	6	24 (9-63) Hyperactivity
3q29	DEL	195.76	196.24	2.01×10^{-7}	40 (13-122)	195760000-196240000	0	no significant association signal
8p23.3-p23.2	DEL	0.4	5.47	1.83×10^{-8}	12 (6-25)	400000-610000 3040000-3780000 4810000-5470000	4	16 (7-39) Intellectual disability
9p24.3	DEL	0.33	0.56	1.08×10^{-7}	13 (6-29)	330000-560000	1	7 (4-12) Neurodevelopmental abnormality
9q34.3	DUP	139.21	140.12	1.67×10^{-6}	12 (5-27)	139210000-139590000 139890000-140120000	21	12 (6-25) Schizophrenia
10q26.3	DEL	133.41	134.68	3.16×10^{-7}	40 (13-125)	133410000-133740000 134370000-134680000	2	32 (10-107) Sleep disorder
15q11.2	DEL	22.74	23.28	1.02×10^{-13}	3 (2-3)	22740000-23280000	12	3 (2-4) Abnormalities of Cognition
15q11.2-q13.3	DUP	22.98	32.15	3.68×10^{-19}	27 (14-52)	24750000-25080000	23	33 (16-67) Intellectual disability
15q12-q13.1	DEL	27.93	28.23	9.85×10^{-7}	14 (6-34)	27930000-28230000	3	24 (11-52) Intellectual disability
15q13.2-q13.3	DEL	31.06	32.51	6.71×10^{-16}	14 (8-24)	31060000-32510000	17	16 (6-44) Sleep disorder
16p13.3	DUP	0.6	0.89	1.98×10^{-13}	9 (5-14)	600000-890000	23	12 (6-26) Schizophrenia
16p13.11	DEL	15.42	16.35	3.53×10^{-17}	9 (6-14)	15420000-16350000	18	13 (4-44) CNS atrophy
16p12.2	DEL	21.88	22.5	1.14×10^{-6}	4 (2-5)	21880000-22500000	6	4 (2-6) Hyperactivity
16p11.2	DEL	29.56	30.19	1.25×10^{-11}	9 (5-15)	29560000-30190000	11	13 (8-21) Intellectual disability
16p11.2	DUP	29.87	30.19	2.45×10^{-8}	6 (3-10)	29870000-30190000	12	11 (5-24) Sleep disorder
17q12	DUP	34.76	36.25	1.12×10^{-15}	15 (9-27)	34760000-35510000 35960000-36250000	14	10 (5-20) Abnormal brain morphology
17q21.31	DEL	41.08	41.45	1.98×10^{-8}	5 (3-9)	41080000-41450000	20	6 (4-10) Abnormality of the nervous system
19p13.3	DUP	1.04	1.34	2.85×10^{-10}	7 (4-12)	1040000-1340000	23	7 (4-14) Schizophrenia
20q13.33	DUP	62	62.35	7.12×10^{-11}	6 (4-10)	62000000-62350000	23	8 (5-13) CNS abnormality
22q11.21	DUP	18.99	21.54	8.73×10^{-11}	5 (3-7)	18990000-1937000 20200000-21540000	9	5 (3-7) Hyperactivity
22q11.21	DEL	18.99	21.54	1.57×10^{-8}	26 (10-65)	18990000-19400000 19670000-19960000 20650000-21540000	12	26 (9-70) Central motor dysfunction

Table 6-2. Genome-wide significantly associated CNV regions and credible intervals.

Cytoband	CNV type	Best overlapping syndrome	Credible interval containing the causal element/gene with 95 % confidence	Brain-related disease genes (high confidence)	PubMedID
15q11.2-q13.3 [a]	DUP	15q11-q13 duplication syndrome (Prader-Willi/Angelman critical region)	15:24750000-25080000	*SNRPN* overexpression (Neurodevelopmental phenotype)	27430727
				SNRPN deletion (Prader-Willi) if the CNV is gene disrupting	35956251
15q13.2-q13.3 [a]	DEL	15q13.3 deletion syndrome	15:31060000-32510000	-	-
16p13.11 [a]	DEL	16p13.11 deletion syndrome	16:15420000-16350000	*MYH11* (Moyamoya-like cerebrovascular disease, cerebral artery aneurysm)	29263223, 27367753
1p36.33 [b]	DEL	1p36 deletion syndrome (Seizures critical region 1)	1:910000-1510000	*DVL1* (Robinow syndrome)	25817016
				TMEM240 (Spinocerebellar ataxia 21)	25070513
1p36.33 [b]	DEL	1p36 deletion syndrome (Seizures critical region 3)	1:2020000-2490000	*SKI* (Shprintzen-Goldberg syndrome)	23023332
1q44 [b]	DEL	KIF26B deletion	1:245290000-245860000	*KIF26B* (Pontocerebellar hypoplasia)	30151950
16p12.2 [b]	DEL	16p12.1 deletion syndrome	16:21880000-22500000	-	-
16p11.2 [b]	DEL	16p11.2 deletion syndrome (BP4-BP5)	16:29560000-30190000	*PRRT2* (Benign familial infantile seizures)	33746883
				TAOK2 (Autism spectrum disorder)	29467497
17q12 [b]	DUP	17q12 duplication syndrome	17:34760000-35510000	-	-
			17:35960000-36250000	-	-
2q13 [c]	DUP	NPHP1 duplication	2:110770000-111060000	*NPHP1* duplication (Autism spectrum disorder, global developmental delay)	25126106, 16892302
3q29 [c]	DEL	3q29 deletion syndrome	3:195760000-196240000	-	-
8p23.3-p23.2 [c]	DEL	8p23.2-pter deletion syndrome	8:400000-610000	-	-
			8:3040000-3780000	-	-
			8:4810000-5470000	-	-
9p24.3 [c]	DEL	9p24.3 DOCK8 / KANK1 deletion	9:330000-560000	*KANK1* (Cerebral palsy spastic quadriplegic 2)	16301218
9q34.3 [c]	DUP	interstitial 9q34.3 duplication (not encompassing EHMT1)	9:139210000-139590000	-	-
			9:139890000-140120000	*GRIN1* gain of function (Polymicrogyria)	29365063
10q26.3 [c]	DEL	10q26 deletion syndrome	10:133410000-133740000	-	-
			10:134370000-134680000	-	-
15q11.2 [c]	DEL	15q11.2 deletion syndrome (BP1-BP2)	15:22740000-23280000	*NIPA1* (Spastic paraplegia 6, autosomal dominant)	23897027
16p11.2 [c]	DUP	16p11.2 duplication syndrome (BP4-BP5)	16:29870000-30190000	-	-
22q11.21 [c]	DUP	22q11.21 deletion syndrome (LCRA-LCRD)	22:18990000-19370000	-	-
			22:20200000-21540000	-	-

Table 6-3. Known disease genes in the credible intervals of the seizure-associated CNV regions. [a] Three CNV regions with previous genome-wide statistical support for epilepsy (PMID: 32568404) (Niestroj *et al.*, 2020), [b] six CNV regions previously implicated in epilepsy without genome-wide statistical support, [c] ten CNV regions previously reported in other neurological and psychiatric disorders, and [d] four novel CNV regions never reported in neurological or psychiatric disorders. BP, breakpoint (continued on next page).

Cytoband	CNV type	Best overlapping syndrome	Credible interval containing the causal element/gene with 95 % confidence	Brain-related disease genes (high confidence)	PubMedID
22q11.21 [c]	DEL	22q11.21 deletion syndrome (LCRA-LCRD)	22:18990000-19400000	-	-
			22:19670000-19960000	*TBX1* (22q11.21 deletion syndrome)	14585638
			22:20650000-21540000	*LZTR1* (Noonan syndrome)	30368668
2p21-p16.3	DEL	Lynch syndrome locus	2:47500000-47850000	-	-
19p13.3	DUP	non-canonical 19p13.3 duplication	19:1040000-1340000	-	-
15q12-q13.1 [d]	DEL	OCA2 deletion	15:27930000-28230000	-	-
16p13.3 [d]	DUP	non-canonical 16p13.3 duplication	16:600000-890000	*STUB1* gain of function (early onset dementia syndrome, autosomal dominant ataxia with cognitive decline and autism)	35493319, 32211513
17q21.31 [d]	DEL	non-canonical 17q21.31 deletion	17:41080000-41450000	*BRCA1* (Cancer)	35393462
20q13.33 [d]	DUP	novel 20q13.33 duplication	20:62000000-62350000	*KCNQ2* gain of function (Neurodevelopmental disability, neonatal encephalopathy)	35780567, 28139826
				EEF1A2 gain of function (Neurodevelopmental disorders)	32160274

Out of the six seizure-associated CNV regions previously implicated in epilepsy without genome-wide statistical support, we mapped the credible intervals of the two seizure-associated deletions at 1p36 to the first and third known critical regions for seizures within the phenotype spectrum of the 1p36 deletion syndrome (Jordan;Zaveri and Scott, 2015). Known disease genes in the credible intervals at 1p36 are *DVL1* (Robinow syndrome, autosomal dominant 2 [OMIM: 616331], White *et al.*, 2015), *TMEM240* (Spinocerebellar ataxia 21 [OMIM: 607454], Delplanque *et al.*, 2014), and *SKI* (Shprintzen-Goldberg syndrome (Doyle *et al.*, 2012)). In the credible intervals of the remaining CNV regions, we identified the following known disease genes: (i) the haploinsufficient *KIF26B* gene (Pontocerebellar hypoplasia (Wojcik *et al.*, 2018)) as the only gene affected by the 1q44 deletion, and (ii) *PRRT2* (self-limited familial infantile epilepsy, paroxysmal dyskinesia (Landolfi;Barone and Erro, 2021)) and the haploinsufficient *TAOK2* gene (Autism (Richter *et al.*, 2019)) at the 16p11.2 BP4-BP5 deletion syndrome locus. Of note, single nucleotide variants in *PRRT2* are among the most frequent findings in clinical genetic testing of epilepsy (Lindy *et al.*, 2018; Symonds *et al.*, 2019).

Among the ten seizure-associated CNV regions previously reported in other neurological and psychiatric disorders, we identified one credible interval suggesting a different causal gene than previously reported: an interstitial 9q34.3 duplication not encompassing *EHMT1* that is considered as the causal gene based on one out of 22 reported 9q34.3 duplication carrier (Bonati *et al.*, 2019). The top candidate gene within the credible interval identified by our meta-analysis is *GRIN1*, affected by 9q34.3 duplications in 21 of all reported carriers (Bonati *et al.*, 2019). *GRIN1* gain of function variants are known to cause a developmental epileptic encephalopathy, often with polymicrogyria (Neurodevelopmental disorder with or without hyperkinetic movements and seizures, autosomal dominant [OMIM: 614254], Fry *et al.*, 2018). In contrast, our fine-mapping analysis confirms *TBX1* as the (known) causal gene for the 22q11.21 deletion/DiGeorge syndrome [OMIM:188400] (Yagi *et al.*, 2003). We also found *LZTR1* (Noonan syndrome 2 [OMIM: 605275], Noonan syndrome 10 [OMIM: 616564], Umeki *et al.*, 2019) within the credible 22q11.21 deletion intervals. Other known disease genes in the credible intervals of the remaining CNV regions implicated in neurological and psychiatric disorders were: *NPHP1* inside a 2q13 duplication (Autism and global developmental delay, Baris *et al.*, 2006; Yasuda *et al.*, 2014), *KANK1* (Cerebral palsy, spastic quadriplegic 2 [OMIM:612900] (Lerer *et al.*, 2005) inside a small 9p24.3 *DOCK8*/*KANK1* deletion, and *NIPA1* (Autosomal dominant spastic paraplegia 6 [OMIM:600363] (Fink, 2013)) inside the 15q11.2 BP1-BP2 deletion syndrome region.

Finally, we identified four novel CNV regions associated with seizures. Three out of four harbored known disease genes. The credible region of a non-canonical 16p13.3 duplication included *STUB1*. *STUB1* gain of function was reported to cause early onset dementia syndrome (Reis *et al.*, 2022) and autosomal dominant ataxia with cognitive decline and autism, Spinocerebellar ataxia 48 [OMIM:618093] (Chen *et al.*, 2020). The credible region of a non-canonical 17q21.31 deletion included *BRCA1*. *BRCA1* mutations are well-known in cancer (Breast-ovarian cancer, familial, 1 [OMIM: 604370], Garber *et al.*, 2022), with *BRCA1* as a possible mediator of glioma cell proliferation, migration, and glioma stem cell self-renewal (Wang *et al.*, 2020). The credible region of a novel 20q13.33 duplication included *KCNQ2* and *EEF1A2*. *KCNQ2* gain of function is known to cause a neurodevelopmental disability and neonatal encephalopathy, DEE 7 [OMIM:613720] (Mulkey *et al.*, 2017; Miceli *et al.*, 2022). *EEF1A2* gain of function was shown to cause neurodevelopmental disorders, including epilepsy (DEE 33 [OMIM:616409]), and intellectual disability (Intellectual developmental disorder, autosomal dominant 38 [OMIM: 616393], Davies *et al.*, 2020).

Significantly enriched GO Biological Processes among all known brain-related disease genes in the credible intervals were: Chordate embryonic development [GO:0043009], Sensory organ morphogenesis [GO:0090596], Mitotic G2 DNA damage checkpoint signaling [GO:0007095], Neural tube closure [GO:0001843], Negative regulation of Ras protein signal transduction [GO:0046580], Dendrite morphogenesis [GO:0048813], and Mitotic G2/M transition checkpoint [GO:0044818]. No GO Biological Process was significantly enriched when considering all genes inside all credible intervals, pointing to likely heterogeneous disease mechanisms of the 25 seizure-associated CNV regions. All credible intervals and known brain-related disease genes are detailed in , additional candidate genes of lower confidence are detailed in **Supplementary Table 6-1**, and all genes inside the credible intervals are detailed in **Supplementary Table 6-2**.

6.5.3 Most of the 25 identified risk CNVs are pleiotropic

We performed 23 meta-analyses of epilepsy with 23 other neuropsychiatric disorders (listed in **Table 6-6**) in an additional 238,161 individuals with neuropsychiatric disorders and 492,324 controls to explore pleiotropy of the 25 identified CNVs. 24 out of 25 seizure-associated CNVs were significantly associated in at least one of the 23 meta-analyses with a neuropsychiatric disorder. The number of neuropsychiatric disorders with which a significant association was found and their greatest ORs is reported in **Table 6-2**. About two thirds (60 %) of all CNVs were highly pleiotropic and showed significant associations with

> 10 epilepsy/neuropsychiatric disorder meta-analyses. The most frequently co-associated phenotype was *Neurodevelopmental abnormality [HP:0012759]*, associated with 36 % of all seizure-associated CNVs.

6.5.4 Characterization of the clinical phenotypes enriched in the carriers of each seizure-associated CNV in epilepsy patients with deep phenotypes

We performed phenome-wide association analyses for each of 35 distinct groups of carriers of CNVs at the credible intervals identified across the 25 CNV regions to characterize the high-resolution clinical manifestations associated with each CNV.[31] This analysis was performed on a subset of the Epi25 cohort (Phenomic cohort, **Table 6-5**) comprising 10,880 individuals with non-acquired epilepsy and deep phenotypic data. In the Phenomic cohort, 562 individuals (5.2 %) carried at least one seizure-associated credible interval (n = 498, 4.6 % carried one credible interval, n = 64, 0.6 % carried 2–5 credible intervals). The most common credible interval (deletion at 2p21-p16.3) was carried by 114 (1.0 %) individuals, and 18 credible intervals were found in at least 0.1 % of the cohort (≥ 11 carriers). One CNV was not found (deletion at 9p24.3, containing a single credible interval). Across the 35 detected credible intervals and 1,667 annotated HPO concepts, we identified 624 nominally significant associations (two-sided Fisher's exact test, **Supplementary Table 6-4).** The minimal sets of these per CNV (those terms that summarize the characteristic phenotypic features of carriers succinctly while remaining precise) are provided in **Supplementary Table 6-5**. Given the large number of associations tested and that HPO annotations describing the same clinical feature at different levels of precision are highly correlated, we applied the *minP* step-down procedure to aid interpretation (Westfall and Wolfinger, 1997), yielding 19 associations robust to multiple testing within each genetically defined group ($minPsd < 0.05$, **Table 6-4**).

[31] These 35 groups spanned 38 credible intervals: the same nine people carried duplications at 17q12 [34.75-35.51 Mb] and 17q12 [34.75-36.25 Mb], and one person had deletion at 10q26.3 [133.41-134.68 Mb], 10q26.3 [133.41-133.74 Mb], and 10q26.3 [134.37-134.68 Mb].

Locus	CNV type	HPO term	Odds ratio (95 % CI)	Relative risk	*p*-value		CNV carriers			CNV non-carriers	
					Raw	Adjusted	Prop	n_{pheno}	n_{tot}	n_{pheno}	n_{tot}
15q13.2-q13.3 [31.06-32.51Mb]	DEL	Generalized non-motor (absence) seizure [HP:0002121]	10.5 (4.25-28.5)	4.18	3.70×10^{-8}	1.00×10^{-5}	0.667	16	24	1731	10856
15q13.2-q13.3 [31.06-32.51Mb]	DEL	Typical absence seizure [HP:0011147]	8.43 (3.48-21.3)	4.10	6.94×10^{-7}	1.10×10^{-4}	0.583	14	24	1545	10856
15q13.2-q13.3 [31.06-32.51Mb]	DEL	EEG with spike-wave complexes [HP:0010850]	7.84 (3.16-21.2)	3.28	1.18×10^{-6}	2.00×10^{-4}	0.667	16	24	2205	10856
15q13.2-q13.3 [31.06-32.51Mb]	DEL	Generalized-onset seizure [HP:0002197]	9.41 (3.15-37.9)	2.40	1.41×10^{-6}	2.20×10^{-4}	0.833	20	24	3766	10856
15q13.2-q13.3 [31.06-32.51Mb]	DEL	EEG with generalized epileptiform discharges [HP:0011198]	6.76 (2.44-23.2)	2.20	1.98×10^{-5}	0.00379	0.792	19	24	3905	10856
15q13.2-q13.3 [31.06-32.51Mb]	DEL	Bilateral tonic-clonic seizure with focal onset [HP:0007334]	0 (0-0.404)	0.00	4.07×10^{-4}	0.0484	0.00	0	24	3168	10856
1p36.33 [0.91-1.51Mb]	DEL	Hypotonia [HP:0001252]	12.2 (3.95-32)	9.51	3.23×10^{-5}	0.00674	0.240	6	25	274	10855
1p36.33 [0.91-1.51Mb]	DEL	Epileptic spasm [HP:0011097]	7.47 (2.78-18.4)	5.40	6.85×10^{-5}	0.0108	0.320	8	25	643	10855
1p36.33 [0.91-1.51Mb]	DEL	Abnormal muscle tone [HP:0003808]	8.65 (2.81-22.7)	6.82	1.97×10^{-4}	0.0287	0.240	6	25	382	10855
1p36.33 [0.91-1.51Mb]	DEL	Infantile spasms [HP:0012469]	8.34 (2.71-21.9)	6.58	2.39×10^{-4}	0.0324	0.240	6	25	396	10855
1p36.33 [0.91-1.51Mb]	DEL	Abnormal muscle physiology [HP:0011804]	8.21 (2.67-21.5)	6.48	2.59×10^{-4}	0.0339	0.240	6	25	402	10855
1p36.33 [0.91-1.51Mb]	DEL	Abnormality of the musculature [HP:0003011]	8.04 (2.61-21.1)	6.35	2.87×10^{-4}	0.038	0.240	6	25	410	10855
1p36.33 [0.91-1.51Mb]	DEL	Plagiocephaly [HP:0001357]	93.8 (9.48-482)	86.8	3.30×10^{-4}	0.045	0.080	2	25	10	10855
2p21-p16.3 [47.50-47.85Mb]	DEL	Focal-onset seizure [HP:0007359]	0.463 (0.313-0.681)	0.708	4.79×10^{-5}	0.0086	0.456	52	114	6939	10766
2p21-p16.3 [47.50-47.85Mb]	DEL	Bilateral tonic-clonic seizure with generalized onset [HP:0025190]	2.3 (1.5-3.46)	1.88	9.09×10^{-5}	0.0157	0.325	37	114	1861	10766
15q12-q13.1 [27.93-28.23Mb]	DEL	Global developmental delay [HP:0001263]	69.1 (5.55-3540)	18.1	2.80×10^{-4}	0.0127	0.750	3	4	451	10876
15q12-q13.1 [27.93-28.23Mb]	DEL	Epileptic encephalopathy [HP:0200134]	Inf (4.43-Inf)	7.72	2.83×10^{-4}	0.0127	1.00	4	4	1408	10876
15q12-q13.1 [27.93-28.23Mb]	DEL	Encephalopathy [HP:0001298]	Inf (4.41-Inf)	7.69	2.87×10^{-4}	0.0129	1.00	4	4	1414	10876
16p11.2 [29.87-30.19Mb]	DUP	Psychogenic non-epileptic seizure [HP:0033052]	81.5 (7.85-471)	61.8	4.82×10^{-4}	0.0297	0.250	2	8	44	10872

Table 6-4. Significant individual CNV-HPO associations in the Epi25 phenomic cohort. n_{pheno}, number of people in the group who were annotated with the phenotype; n_{tot}, number of people in the group.

Carriers of deletions at 1p36.33 [0.91-1.51 Mb] (n = 25, 0.23 % of the Phenomic cohort), 1p36.33 [2.02–2.49 Mb] (n = 17, 0.16 %, **Figure 6-3**), or 15q12–q13.1 (n = 4, 0.037 %), and carriers of duplications at 15q11.2-q13.3 (n = 46, 0.42 %) were enriched with clinical features suggestive of developmental and epileptic encephalopathies (**Figure 6-4**), such as epileptic spasms and tonic seizures, epileptic encephalopathy, and other neurodevelopmental disorders, sudden unexpected death in epilepsy, and morphological abnormalities (Scheffer *et al.*, 2017). Features characterizing GGE were associated with deletions at 2p21-p16.3 (n = 114, 1.05 %, generalized tonic-clonic and absence seizures), 15q11.2 (n = 56, 0.52 %, eyelid myoclonia and absence seizures, **Figure 6-5**), 16p13.11 (n = 42, 0.39 %, generalized tonic-clonic seizures, **Figure 6-6**), 15q13.2-q13.3 (n = 24, 0.22 %, absence seizures) or 22q11.21 [20.65-21.54 Mb] (n = 6, 0.055 %, juvenile myoclonic epilepsy-like features, **Figure 6-7**). Duplications at 16p11.2 (n = 8, 0.074 %) were associated with non-epileptic seizures comorbid with epilepsy (OR = 81.5, unadjusted $p = 4.82 \times 10^{-4}$, *minPsd* = 0.0297), and showed a nonsignificant greater frequency of microcephaly (OR = 31.5, unadjusted $p = 3.62 \times 10^{-2}$, *minPsd* = 0.92) that replicates the mirror microcephaly/macrocephaly phenotype of the reciprocal 16p11.2 CNVs (Golzio *et al.*, 2012).

We interrogated the phenotypic annotations of CNV carriers regarding the candidate genes prioritized in our fine-mapping analysis. *MSH2* was prioritized as the candidate gene for the most common deletion in the Phenomic cohort (2p21-p16.3). Heterozygous loss of function variants of the haploinsufficient gene *MSH2* cause *Lynch syndrome 1 [OMIM:120435]* (Barrow *et al.*, 2008), and complete knockout of paralog *Msh2* in $Ccm1^{+/-}$ mice causes multiple cavernoma through a presumed second hit (McDonald *et al.*, 2011). We found that carriers had a nonsignificant greater frequency of neoplasms (OR = 2.35, unadjusted $p = 2.49 \times 10^{-2}$, *minPsd* = 1.00) and cerebral cavernomata (OR = 5.23, unadjusted $p = 6.58 \times 10^{-4}$, *minPsd* = 0.157) than non-carriers. Carriers of the 1p36.33 [2.02-2.49 Mb] deletion overlapping the gene *SKI* had features (hypotonia, talipes equinovarus, abnormalities of the globe and nose, osteoporosis, global developmental delay, and Chiari malformation) concordant to the Shprintzen-Goldberg craniosynostosis syndrome [OMIM:182212] caused by *SKI* (Doyle *et al.*, 2012). All 15 individuals with duplication of 9q34.3 had focal-onset seizures that were rarely drug-resistant, without any individual annotated with a neurodevelopmental disorder or polymicrogyria despite the presence of the *GRIN1*, which can cause polymicrogyria when affected by gain-of-function variants (**Figure 6-8**; Fry *et al.*, 2018). Sixteen of 24 individuals carrying deletions at 15q13.3 [31.06-32.51 Mb] had generalized absence seizures (OR = 10.5, unadjusted $p = 3.70 \times 10^{-8}$, *minPsd* = 1×10^{-5}), in line with the primary

seizure type reported in carriers of the 15q13.3 deletion (Whitney *et al.*, 2021). We found generalized myoclonic seizures annotated to half of the carriers of the 22q11.21 [19.67-19.96 Mb] deletion (**Figure 6-9**), confirming the relationship between this CNV and these seizures (Strehlow *et al.*, 2016), implicating *TBX167*, the known causal gene for the 22q11.21 deletion/DiGeorge syndrome in myoclonic epilepsies (Yagi *et al.*, 2003). Features suggestive of juvenile myoclonic epilepsy were also found among six people carrying deletions overlapping with the second credible interval at 22q11.2 [20.65-21.54 Mb] spanning the Noonan syndrome 10 locus containing in which a single individual was reported with seizures (Umeki *et al.*, 2019). However, none of these six individuals had annotations beyond seizures and electroencephalography phenotypes that would support a multisystemic syndrome. Duplication at 22q11.21 appeared to have a reciprocal effect on broad seizure type to the deletions, with 15 of 16 carriers of duplications at this locus having *Focal-onset seizure [HP:0007359]* (**Figure 6-10**). A few CNVs such as duplication at 17q12 appeared to have little evidence of distinctive phenotypic features.

Finally, clinicians may want to know the frequency of broad clinical features among carriers of the CNV identified in their patients to improve the interpretation of its clinical relevance and to facilitate genetically stratified prognostication. Therefore, we prioritized 17 common, conceptually broad, and important epilepsy manifestations and comorbidities for visualization, including the co-occurrence of generalized-onset and focal-onset seizures that characterizes the combined generalized and focal epilepsy type (Scheffer *et al.*, 2017) (**Figure 6-11**). The most common CNV, deletion at 2p21-p16.3, appeared to modestly increase the likelihood of a carrier having generalized epilepsy. However, a few CNVs had a profile dominated by core electroclinical features of generalized (for example, deletions at 15q13.2-15q13.3) or focal epilepsy (duplications at 9q34.3 [139.89-140.12 Mb]), with comorbid features being rare. Conversely, carriers of other CNVs had relatively high frequencies of neurodevelopmental disorders, epileptic spasms, and drug resistance suggestive of DEE (deletions at 1p36.33). However, no CNV was found exclusively in people with a particular seizure type, and carriers of some CNVs appeared to have broad clinical features at frequencies indistinguishable from the cohort's baseline (duplications at 19p13.3), suggesting some generic contribution to epilepsy risk across epilepsy types.

6.5.5 Syndromic similarity among carriers of the same CNV

We assessed the overall similarity across all phenotypic annotations for each of the 32 groups of multiple people carrying a CNV spanning the same credible. We found four of 31 groups

were sufficiently similar according to our primary measure, $\mathrm{Sim}_{\mathrm{MinTO}}$ to reach $p \leq 0.05$, meaning that the median pairwise similarity of carriers was in the top 5 % of those expected to be observed by chance for groups of the same size (**Supplementary Table 6-6**). The 24 carriers of deletions at 15q13.2-q13.3 [31.06-32.51 Mb] (one-sided raw $p = 2.1 \times 10^{-5}$, Holm's $p = 6.72 \times 10^{-4}$), and those four with deletions of 15q12-q13.1 [27.93-28.23 Mb] (raw $p = 2.11 \times 10^{-4}$, Holm's $p = 0.00654$) were particularly similar. The former was significant across all six similarity measures assessed with Holm's $p < 0.05$. The latter reached Holm's $p < 0.05$ by all except the best match average measures, suggesting that the similarity of deletion at 15q12-q13.1 is more dependent on its carriers having broader phenotypes than sharing relatively more specific features. This was supported by review of the minimal sets of annotations of the carriers (**Supplementary Table 6-5**). Compared with those carrying deletions at 15q13.2-q13.3, those carrying deletions at 15q12-q13.1 had more annotations (median 10.5 vs 3, Wilcoxon rank sum test with continuity correction, WRST, $W = 93$, two-sided $p = 0.00302$, pseudomedian = 6.81) but only of modestly greater information (median 4.62 bits vs 3.26 bits, WRST $W = 81$, two-sided $p = 0.0328$, pseudomedian = 1.22). The relative breadth of the presentation of those carrying deletions at 15q12-q13.1 compared to those carrying deletions at 15q13.2-q13.3 is evident on reviewing their positively associated phenotypes reaching two-sided raw $p < 0.5$ (**Supplementary Table 6-4**). The minimal set of those associated with 15q12-q13.1 deletions includes 12 conceptually independent terms suggestive of a DEE: *Global developmental delay [HP:0001263]*, *Epileptic encephalopathy [HP:0200134]*, *Severe intellectual disability [HP:0010864]*, *Delayed CNS myelination [HP:0002188]*, *Periventricular leukomalacia [HP:0006970]*, *Post-term pregnancy [HP:0031169]*, *Hypotonia [HP:0001252]*, *Ataxia [HP:0001251]*, *Convulsive status epilepticus [HP:0032660]*, *Status epilepticus without prominent motor symptoms [HP:0031475]*, *Seizure precipitated by febrile infection [HP:0032894]*, and *Refractory drug response [HP:0020174]*. The minimal set of phenotypes positively associated with deletions at 15q13.2-q13.3 includes 8 terms suggesting a milder neurodevelopmental, cerebellar, and generalized epilepsy syndrome with *Typical absence seizure [HP:0011147]*, *EEG with spike-wave complexes (> 3 Hz) [HP:0010849]*, *Mild global developmental delay [HP:0011342]*, *Delayed gross motor development [HP:0002194]*, *Mild intellectual disability [HP:0001256]*, *Dysdiadochokinesis [HP:0002075]*, *Clumsiness [HP:0002312]*, and *Abnormal cerebellar cortex morphology [HP:0031422]*. Furthermore, 15q13.2-q13.3 was negatively associated with five relatively common seizure and EEG phenotypes, reinforcing the narrower clinical presentation of this CNV.

Other CNVs reaching raw one-sided $p < 0.05$ according to $\mathrm{Sim}_{\mathrm{MinTO}}$ were the deletions at 1p36.33 [2.02-2.49 Mb] (phenotypically broad and suggestive an epileptic encephalopathy with minimal sets of 30 positively associated phenotypes and negatively associated with *Bilateral tonic-clonic seizure [HP:0007334]*) and 3q29 [195.76-196.24 Mb] (minimal set of positively associated phenotypes of *Febrile seizure (within the age range of 3 months to 6 years) [HP:0002373]*, *Severe intellectual disability [HP:0010864]*, *Facial asymmetry [HP:0000324]*, *Limb tremor [HP:0200085]*, *Decreased head circumference [HP:0040195]*, *Psychosis [HP:0000709]*, and *Joint stiffness [HP:0001387]*). These also had moderately high similarities according by the other measures. Two loci had high similarity by the $\mathrm{Sim}_{\mathrm{BMA}}$ measures only: deletions at 22q11.21 [20.65-21.54 Mb] and 22q11.21 [18.99-21.54 Mb], with the carriers of the latter consisting of five of the six carriers of the former more limited deletion. These both had a median of 4 base annotations per carrier, with relatively high information (median of the median information of each base annotation per carrier 3.33 bits and 2.89 bits). The minimal set of phenotypes associated with 22q11.21 [20.65-21.54 Mb] included *Bilateral tonic-clonic seizure with generalized onset [HP:0025190]*, *Focal autonomic seizure [HP:0011154]*, *Generalized myoclonic seizure [HP:0002123]*, *EEG with polyspike wave complexes [HP:0002392]*, and *EEG with focal slow activity [HP:0010843]*, and there was a negative association with *Focal-onset seizure [HP:0007359]*. For 22q11.21 [18.99-21.54 Mb] there were positive associations with *Generalized myoclonic seizure [HP:0002123]*, *Bilateral tonic-clonic seizure with generalized onset [HP:0025190]*, *Focal autonomic seizure [HP:0011154]*, and *EEG with polyspike wave complexes [HP:0002392].*

6.6 Discussion

In this study, we leveraged a substantial increase in sample size to identify novel seizure-associated CNVs when jointly analyzing 26,699 individuals with various types of seizure disorders against 492,324 population controls. We identified 25 novel loci with genome-wide significance for seizure disorders. In addition, all three previously reported epilepsy-associated loci at genome-wide level maintained genome-wide significance for seizure disorders in our meta-analysis that included the epilepsy cohort from the previous study (Collins *et al.*, 2022). Of the 25 seizure-associated loci, 16 were previously implicated in neurological and psychiatric disorders, including epilepsy. Five were flanked by known segmental duplications or low copy number repeats. Of note, our fine-mapping analysis confirmed the first and third known critical regions for seizures within the phenotype spectrum of the 1p36 deletion syndrome (Jordan;Zaveri and Scott, 2015), *TBX1* as the (known) causal gene for the 22q11.21 deletion/DiGeorge syndrome (Yagi *et al.*, 2003), and suggested the *SNRPN* promoter/exon 1

region as the causal element for seizures within the larger BP2-BP3 15q11.2-q13 duplication region. However, our study design did not support the assessment of whether the imprinting status of the duplicated region itself plays an additional role besides the previously suggested role of *SNRPN* promoter/exon 1 region in regulating the imprinting of the Prader-Willi critical region. Future studies that also include genomic screens of parents will shed light on this open question.

In a high-resolution phenomic analysis in a subset of 10,880 individuals from our cohort with epilepsy (from the Epi25 cohort), we identified 605 suggestive and 19 significant clinical associations informative for epileptologists among CNV carriers. This observation indicates that beyond contributing to the generic risk of seizures, several CNVs contribute to specific epilepsy types. Carriers of some CNVs tended to have features typical of developmental and epileptic encephalopathies with neurodevelopmental and non-seizure phenotypes. Conversely, carriers of others had phenotypes restricted to the core epileptic features of seizures and electroencephalographic abnormalities (both generalized and focal). Interestingly, reciprocal CNVs involving 22q11.21 seemed to produce opposite epilepsy types, with deletion and duplication carriers tending to have generalized and focal epilepsies, respectively. Dose-dependent effects of KLHL22 on DEPDC5 degradation are a possible explanation (Chen *et al.*, 2018c). Overall, the high degree of pleiotropy among seizure-associated CNVs implies that these CNVs likely impair neurodevelopmental processes rather generically and contribute to the broad spectrum of neurodevelopmental disorders. According to the oligo-/polygenic inheritance model, CNVs may interact with the genetic background or environmental factors to generate the final disease phenotype. Interaction between CNVs and the polygenic background was recently demonstrated in carriers of the schizophrenia-associated 22q11.2 deletion (Cleynen *et al.*, 2021). Support for an oligogenic-CNV disorder model was also recently published (Stefanski *et al.*, 2022).

Genome-wide genetic screening for pathogenic CNVs is recommended as a first-tier approach for the postnatal evaluation of individuals with intellectual disability, developmental delay, autism spectrum disorder, multiple congenital anomalies, and prenatal evaluation of fetuses with structural anomalies observed by ultrasound (Manning and Hudgins, 2010; Miller *et al.*, 2010; Stosic;Levy and Wapner, 2018; Manning and Hudgins, 2020). It has previously been shown that CNVs confer significant risk towards epilepsy (Amiet *et al.*, 2008; Heinzen *et al.*, 2010; Mefford *et al.*, 2011; Sisodiya and Mefford, 2011; Fisher *et al.*, 2014; Olson *et al.*, 2014; Lal *et al.*, 2015; Addis *et al.*, 2016; Fisher *et al.*, 2017b; Pérez-Palma *et al.*, 2017), particularly for individuals with comorbid neurodevelopmental disorders such as intellectual disability

(Mefford *et al.*, 2011; Mullen *et al.*, 2013; Reinthaler *et al.*, 2014; Coppola *et al.*, 2019). In contrast to single nucleotide polymorphism SNP GWASs for epilepsy or seizures, where the risk of identified variants is small (OR < 2, ILAE Consortium on Complex Epilepsies, 2018; Skotte *et al.*, 2022; ILAE Consortium on Complex Epilepsies, 2023), the effect sizes of the 25 CNVs identified in this study are large (median OR = 11, range 2–53). Our high-resolution phenomic analysis of 10,880 individuals with epilepsy grouped by CNV carrier status illustrates the seizures, EEG and brain imaging findings, and neurodevelopmental and other comorbidities associated with each CNV. This genotype-first approach complements the traditional single-phenotype, case-control paradigm by taking a simultaneous phenome-wide perspective in individuals deeply phenotyped according to standardized protocols before CNV discovery or genetic association tests. We found phenotypic evidence supporting associations between CNVs, broad markers of epilepsy types, and fine-grained phenotypes. The high-resolution phenotype associations that an epileptologist can recognize derived from the HPO phenotype association analysis and disease risk estimates from the meta-analysis for each CNV can enhance the interpretation of clinical relevance and pathogenicity following the American College for Genetics and Genomics CNV interpretation guidelines (Riggs *et al.*, 2020; Riggs *et al.*, 2021).

Our study has several limitations. First, many of the patients with seizures included in this study have comorbid neurological and psychiatric disorders. Therefore, some of the identified CNV loci may be associated with other clinical phenotypes present in a high percentage of all cases. Second, we did not detect robust associations with two important outcomes in our HPO analysis, refractory drug response and sudden unexpected death. Sudden unexpected death in epilepsy is poorly suited to cross-sectional studies: it was annotated to only 4 of 10,880 individuals, far fewer cases than expected to occur with follow-up of this cohort of individuals requiring tertiary center care (Harden *et al.*, 2017). This emphasizes the open-world interpretation required for our results: in any study that is cross-sectional and of a disorder that has inherently variable phenotyping depth (epilepsy presentations can often be classified only incompletely) (Fisher *et al.*, 2014; Scheffer *et al.*, 2017), and which is characterized by some phenotypes that are age-dependent (such as some seizure types, autism, and intellectual disability), one should rarely assume that the absence of an annotation can be interpreted as the absence of that phenotype over the lifetime of the carrier. Thus, the proportion of individuals annotated with a phenotype is likely lower than the actual proportion manifesting it over their lifetimes (Lewis-Smith *et al.*, 2022). Third, in contrast to conventional SNP-based GWASs, genome-wide CNV association studies have major challenges in identifying the causal gene(s)

impacted by the CNV. Among the 25 identified CNVs, deletions ranged from 230 kb to 5 Mb and duplications from 290 kb to 9 Mb, affecting 14.2 genes on average. CNV breakpoints in the current study are estimated from genotyped SNPs around the actual breakpoint. These breakpoint estimates are limited by the resolution of the genotyping platform used to call the CNVs. In fact, microarrays have many technical limitations, such as poor breakpoint resolution and limited sensitivity for small CNVs (Sudmant *et al.*, 2015). Newer technologies like whole-genome sequencing will enable the assessment of a more comprehensive array of rare variants, including balanced rearrangements, small (exonic) CNVs (Brandler *et al.*, 2016), short tandem repeats, and other structural variants (Gymrek *et al.*, 2016). However, some genomic regions harbor complex deletion/duplication/inversion rearrangements (for example, 22q11.21 (Vervoort and Vermeesch, 2022), 15q11.2 (Gimelli *et al.*, 2003)) that can even show population stratification (for example, 16p11.2 (González *et al.*, 2014)). More accurate and complete (pangenome) references will be needed to determine the exact breakpoints of such complex rearrangements (Liao *et al.*, 2023; Porubsky *et al.*, 2023), even in the case of sequencing-based CNVs discovery. Lastly, we performed joint epilepsy/seizures and cross-disorder meta-analyses in individuals with minimal clinical information. Future studies with access to rich clinical metadata, such as electronic health records, will likely identify additional seizure-associated CNVs. It is important to consider the inclusion criteria for this cohort and the definition of cases and controls when interpreting associations and their relevance to a patient. Our phenomic analysis cohort was performed using the data of years 1–3 of Epi25, predominantly recruited from academic epilepsy centers and of European ancestry (92.9 %). Additionally, we screened cases to exclude those with brain trauma, meningitis, or encephalitis. Thus, our clinical associations should be considered most valid in individuals of European ancestry with likely genetic or unexplained epilepsies attending specialist epilepsy centers. Future data analyses from subsequent years of Epi25 will provide data more applicable to other populations.

Large-scale collaborations that enable the aggregation of massive datasets have greatly advanced epilepsy and the discovery of genetic factors through GWASs. Here, we have extended this framework to CNV discovery by meta-analyzing epilepsy and seizure disorders, followed by additional meta-analyses in neuropsychiatric disorders and traits to explore pleiotropy. We also identified fine-grained genotype-phenotype associations and clinical profiles for each CNV. Our results will help refine promising candidate CNVs associated with specific epilepsy types and extend their clinical value. We are confident that applying this framework to even larger datasets has the potential to advance the discovery of all clinically

relevant risk loci, ultra-rare high-risk CNVs missed by this study, and the underlying genes or functional elements.

Cohort	Data Source	Short Name	Citation	Original Ref. Build	Case Phenotypes	Cases	Controls	Meta-analysis cohort assignment
Epi25 Collaborative	Epi25 Collaborative	Epi25	Epi25 Collaborative, Am J Hum Genet (2021)	hg19	Epilepsy and controls	16,109	8,545	NA
Neuropsychiatric cohorts	BioVU	BioVU	Morley et al., Nat. Med. (2021)	hg19	Mixed	32,306	14,661	Cohort 7
	Boston Children's Hospital	BCH	Talkowski et al., Cell (2012)	hg18	Mixed	3,591	0	Cohort 1
	Children's Hospital of Philadelphia	CHOP	Li et al., Nat. Commun. (2020)	hg19	Mixed	153,870	24,161	Cohort 4
	Eichler Lab	Coe	Coe et al., Nat. Genet. (2014)	hg19	Developmental disorders	29,104	11,256	Cohort 1
	Eichler Lab	Cooper	Cooper et al., Nat. Genet. (2011)	hg19	N/A	0	8,329	Cohort 1
	Estonian Biobank	EstBB	Leitsalu et al., Int. J. Epidemiol. (2014)	hg19	Mixed	63,183	15,659	Cohort 6
	GeneDX	GDX	-	hg18 &19	Mixed	74,028	0	Cohort 2
	Indiana University	IU	-	hg19	Mixed	1,577	0	Cohort 1
	Ontario Population Genomics Platform	Ontario	Uddin et al., Genet. Med. (2015)	hg19	N/A	0	873	Cohort 2
	Psychiatric Genetics Consortium	PGC	Marshall et al., Nat. Genet. (2017)	hg18	Schizophrenia	21,094	20,277	Cohort 3
	Radboud University Medical Center	RUMC	Vulto-van Silfhout et al., Hum. Mutat. (2013)	hg17	Intellectual disability	5,531	0	Cohort 3
	SickKids Hospital	SickKids	Zarrei et al., NPJ Genomic Medicine (2019)	hg19	Developmental disorders	2,689	0	Cohort 3
	Simons Simplex Collection	SSC	Sanders et al., Neuron (2015)	hg18	Autism	2,795	0	Cohort 3
	The Cancer Genome Atlas	TCGA	Zack et al., Nat. Genet. (2013)	hg19	N/A	0	8,670	Cohort 3
	The Genetic Etiology of Tourette Syndrome Consortium	TSAICG	Huang et al., Neuron (2017)	hg19	Tourette Syndrome	2,434	4,093	Cohort 2
	UK Biobank	UKBB	Macé et al., Nat. Commun. (2017)	hg19	Mixed	54,071	375,800	Cohort 2 (3,139 randomly selected controls); Cohort 5 (all remaining samples)
Phenomic cohort	Epi25 Collaborative	Pheno-Epi25	Epi25 Collaborative, Am J Hum Genet (2021)	hg19	Non-acquired epilepsy	10,880	0	-

Table 6-5. List of cohorts for which CNV calls were made. The phenomic cohort is a subset of the Epi25 cohort.

Disorder Full Name	HPO	Number of participants
Seizures	HP:0001250	10,590
23 neuropsychiatric disorders		
Abnormality of the nervous system	HP:0000707	248,751
Abnormality of higher mental function	HP:0011446	15,416
Neurodevelopmental abnormality	HP:0012759	55,760
Autistic behavior	HP:0000729	26,199
Intellectual disability	HP:0001249	10,371
Behavioral abnormality	HP;0000708	135,483
Impairment in personality functioning	HP:0031466	58,323
Hyperactivity	HP:0000752	14,955
Bipolar affective disorder	HP:0007302	34,915
Schizophrenia	HP:0100753	25,108
Abnormal fear/anxiety-related behavior	HP:0100852	25,856
Sleep disturbance	HP:0002360	6,091
Abnormality of nervous system morphology	HP:0012639	68,594
Morphological abnormality of the central nervous system	HP:0002011	37,800
Abnormality of brain morphology	HP:0012443	7,688
Atrophy/Degeneration affecting the central nervous system	HP:0007367	26,163
Abnormality of central motor function	HP:0011442	5,341
Abnormality of movement	HP:0100022	12,628
CNS hypermyelination	HP:0012447	3,204
Involuntary movements	HP:0004305	4,928
Abnormality of the peripheral nervous system	HP:0410008	6,005
Abnormal peripheral nervous system morphology	HP:0000759	32,665
Peripheral neuropathy	HP:0009830	22,571

Table 6-6. Phenotypes of neuropsychiatric disorders cohorts used in this study.

Cytoband	CNV type	Hg19 Start (Mb)	Hg19 End (Mb)	Cases carrier frequency (%)	Control carrier frequency (%)	Frequency in DGV Gold Standard (%)	Frequency in DECIPHER (%)
1p36.33	DEL	0.91	1.51	0.086	0.015	0	0
1p36.33	DEL	2.02	2.49	0.146	0.003	0	0
1q44	DEL	245.29	245.86	0.041	0.002	0	0
2p21-p16.3	DEL	47.5	47.85	0.678	0.002	0	0
2q13	DUP	110.77	111.06	0.139	0.108	0	0.27
3q29	DEL	195.76	196.24	0.034	0.002	0	0.12
8p23.3-p23.2	DEL	0.4	5.47	0.067	0.010	0	0
9p24.3	DEL	0.33	0.56	0.049	0.007	0	0.02
9q34.3	DUP	139.21	140.12	0.315	0.003	0	0
10q26.3	DEL	133.41	134.68	0.030	0.002	0	0
15q11.2	DEL	22.74	23.28	0.689	0.284	0.41	0.83
15q11.2-q13.3	DUP	22.98	32.15	0.258	0.005	0	0.12
15q12-q13.1	DEL	27.93	28.23	0.097	0.008	0	0
15q13.2-q13.3	DEL	31.06	32.51	0.243	0.012	0	0.02
16p13.3	DUP	0.6	0.89	0.573	0.009	0	0
16p13.11	DEL	15.42	16.35	0.363	0.031	0.03	0.07
16p12.2	DEL	21.88	22.5	0.191	0.054	0.09	0.08
16p11.2	DEL	29.56	30.19	0.165	0.024	0.05	0.12
16p11.2	DUP	29.87	30.19	0.127	0.026	0	0.03
17q12	DUP	34.76	36.25	0.187	0.014	0.02	0.12
17q21.31	DEL	41.08	41.45	0.461	0.004	0	0
19p13.3	DUP	1.04	1.34	0.427	0.003	0	0.08
20q13.33	DUP	62	62.35	0.479	0.006	0	0
22q11.21	DUP	18.99	21.54	0.199	0.067	0	0
22q11.21	DEL	18.99	21.54	0.120	0.009	0	0.19

Table 6-7. CNV frequencies in the cases & controls of the meta-analysis, DGV-Gold Standard, and DECIPHER databases. The frequencies in the DGV Gold Standard and DECIPHER Population databases are given for CNVs with ≥ 50 % overlap with the seizure-associated CNV regions.

Name	Web address	Used for
DGV	http://dgv.tcag.ca/dgv/app/downloads?ref=	Annotation of CNV frequencies from external data
DECIPHER	https://www.deciphergenomics.org/about/downloads/data https://www.deciphergenomics.org	Localization of reported phenotypes, curated genes, dosage sensitivity Reported CNV Syndromes
ClinGen	https://www.clinicalgenome.org	Localization of reported phenotypes, curated genes, dosage sensitivity
GARD	https://rarediseases.info.nih.gov	Detailed phenotypes of disorders caused by CNVs or genes overlapping with the seizure-associated regions
OMIM	https://www.omim.org	Molecular Genetics and reported patients for any gene
PubMed	https://pubmed.ncbi.nlm.nih.gov	Screening of published literature
UCSC Genome Browser on Human GRCh37/hg19	https://genome.ucsc.edu/	Visualization of the credible intervals
GWAS Catalog	https://www.ebi.ac.uk/gwas/home	Reported GWAS hits in credible intervals
Enrichr	https://maayanlab.cloud/Enrichr/	Pathway analyses

Table 6-8. Resources used to collect information about CNVs and the affected genes.

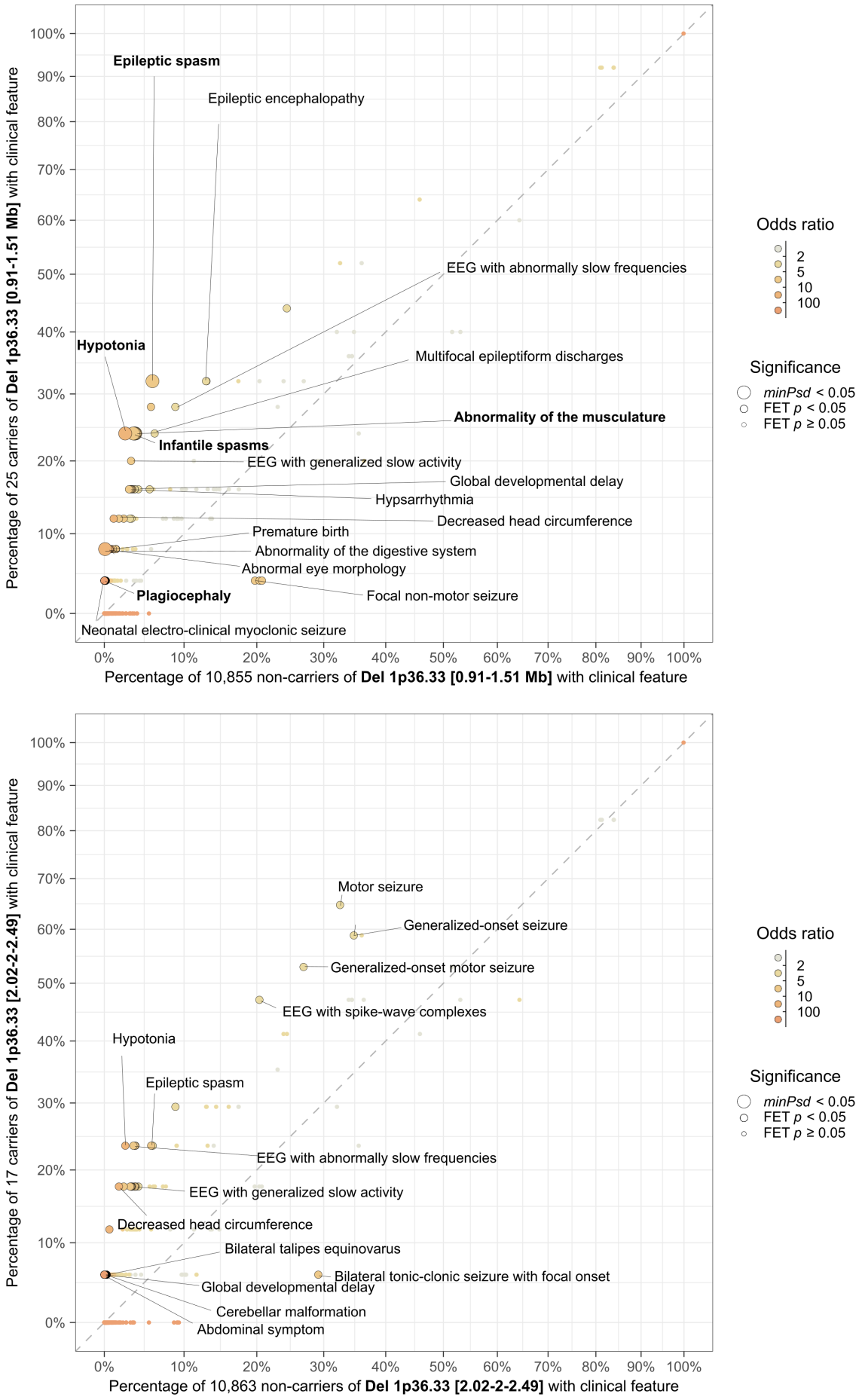


Figure 6-3. Associations of deletions at 1p36.33 [0.91-1.51 Mb] and [2.02–2.49 Mb].

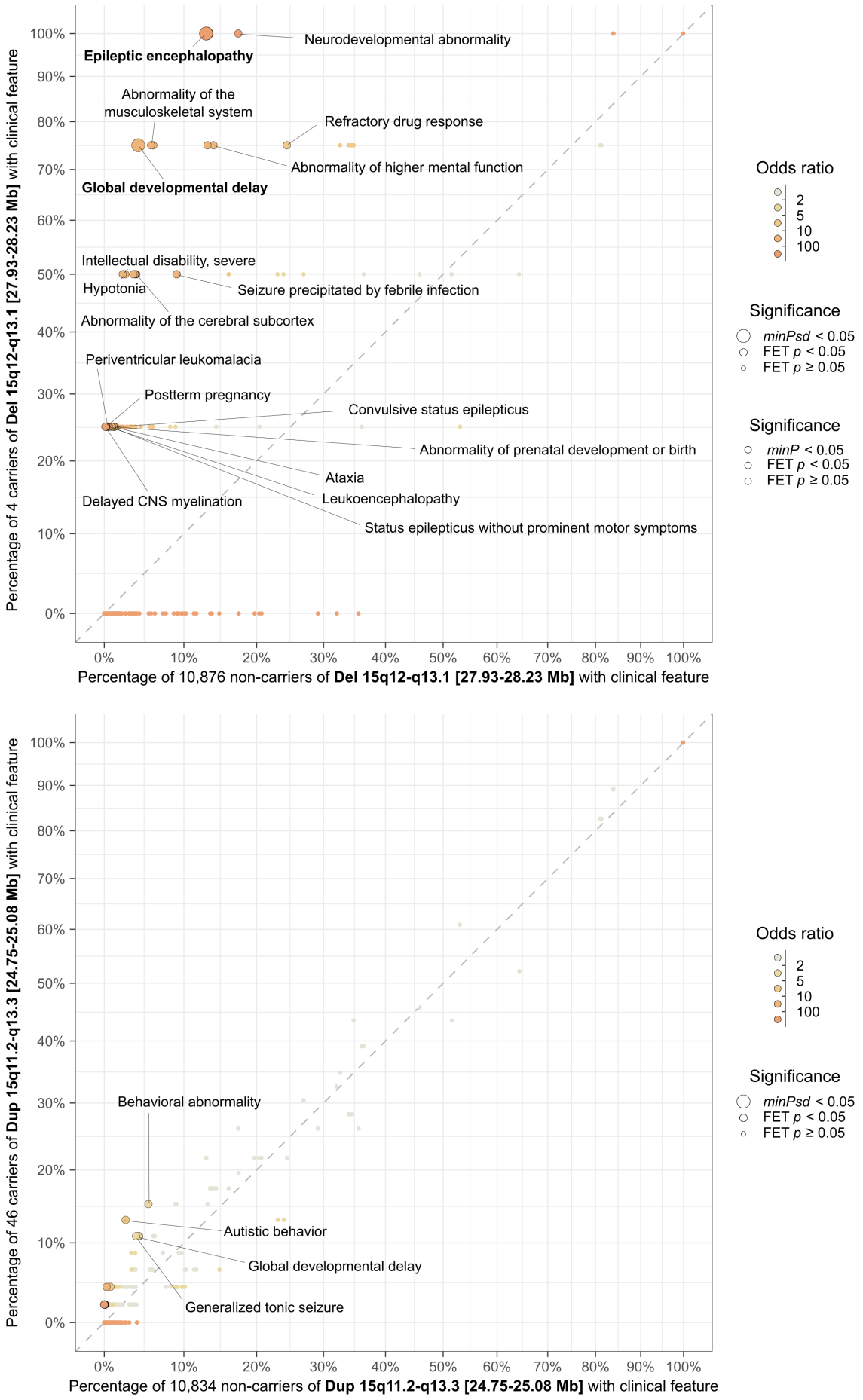


Figure 6-4. Associations of deletions at 15q12–q13.1 and duplications at 15q11.2-q13.3.

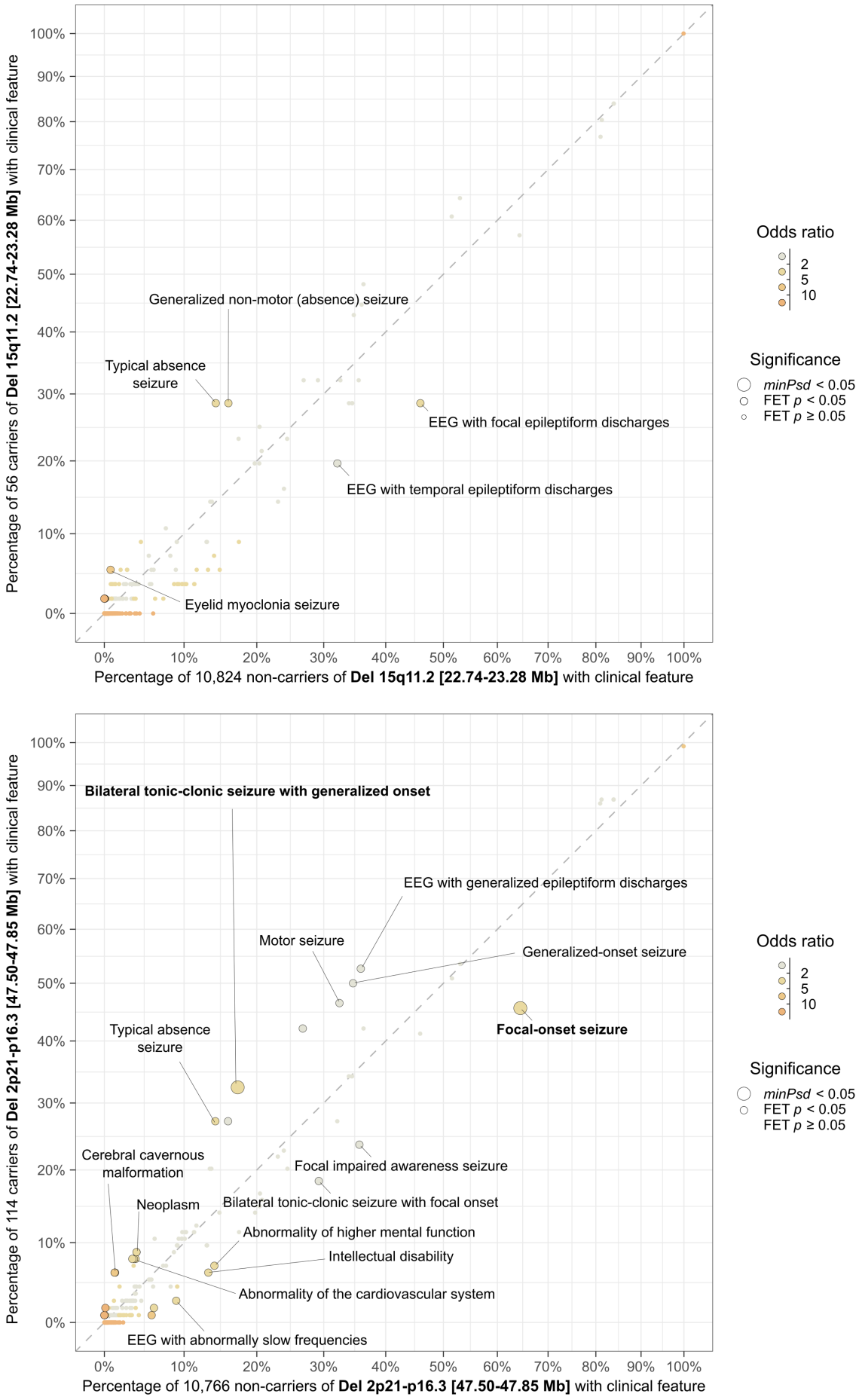


Figure 6-5. Associations of deletions at 2p21-p16.3 and 15q11.2.

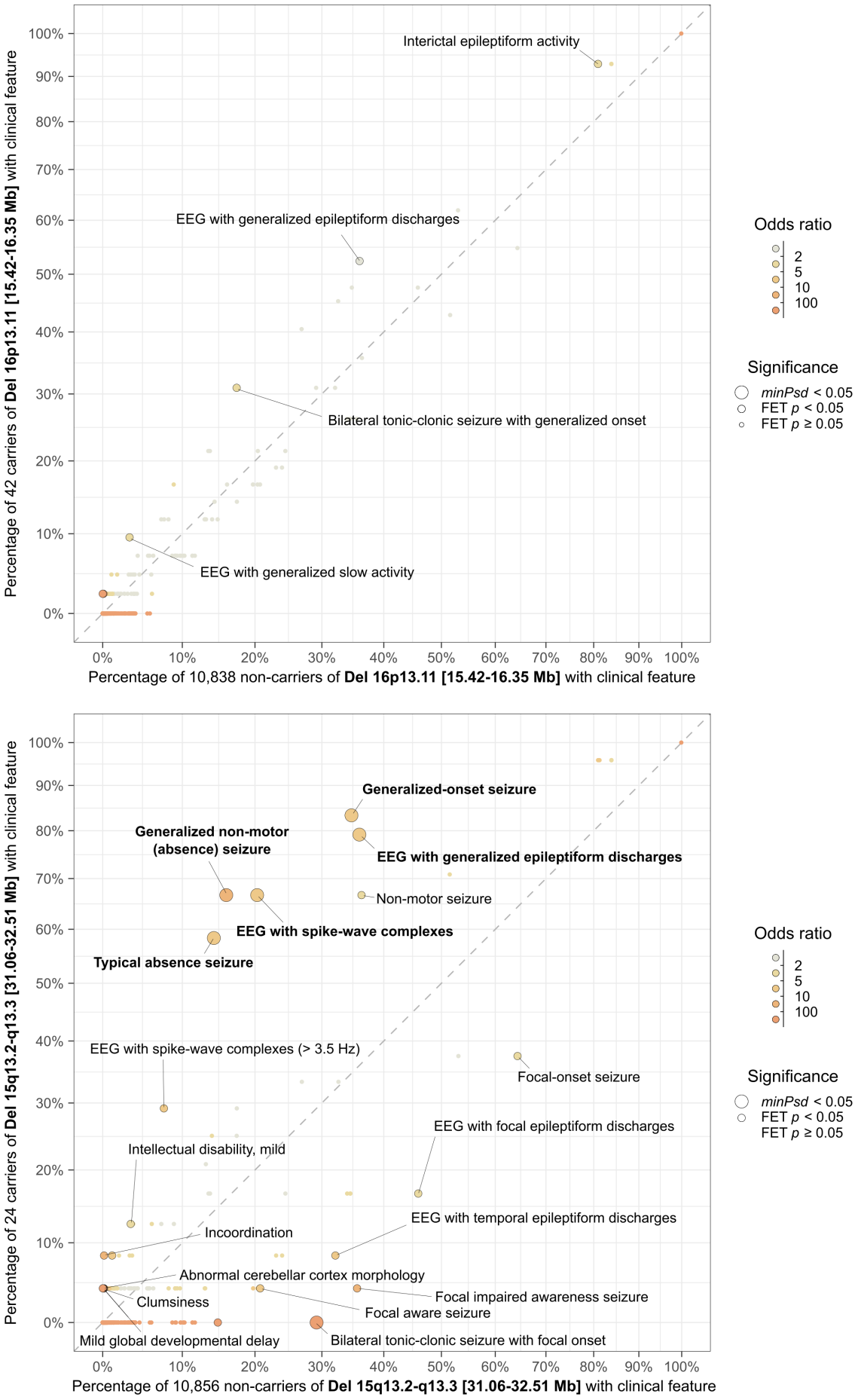


Figure 6-6. Associations of deletions at 16p13.11 and 15q13.2-q13.3.

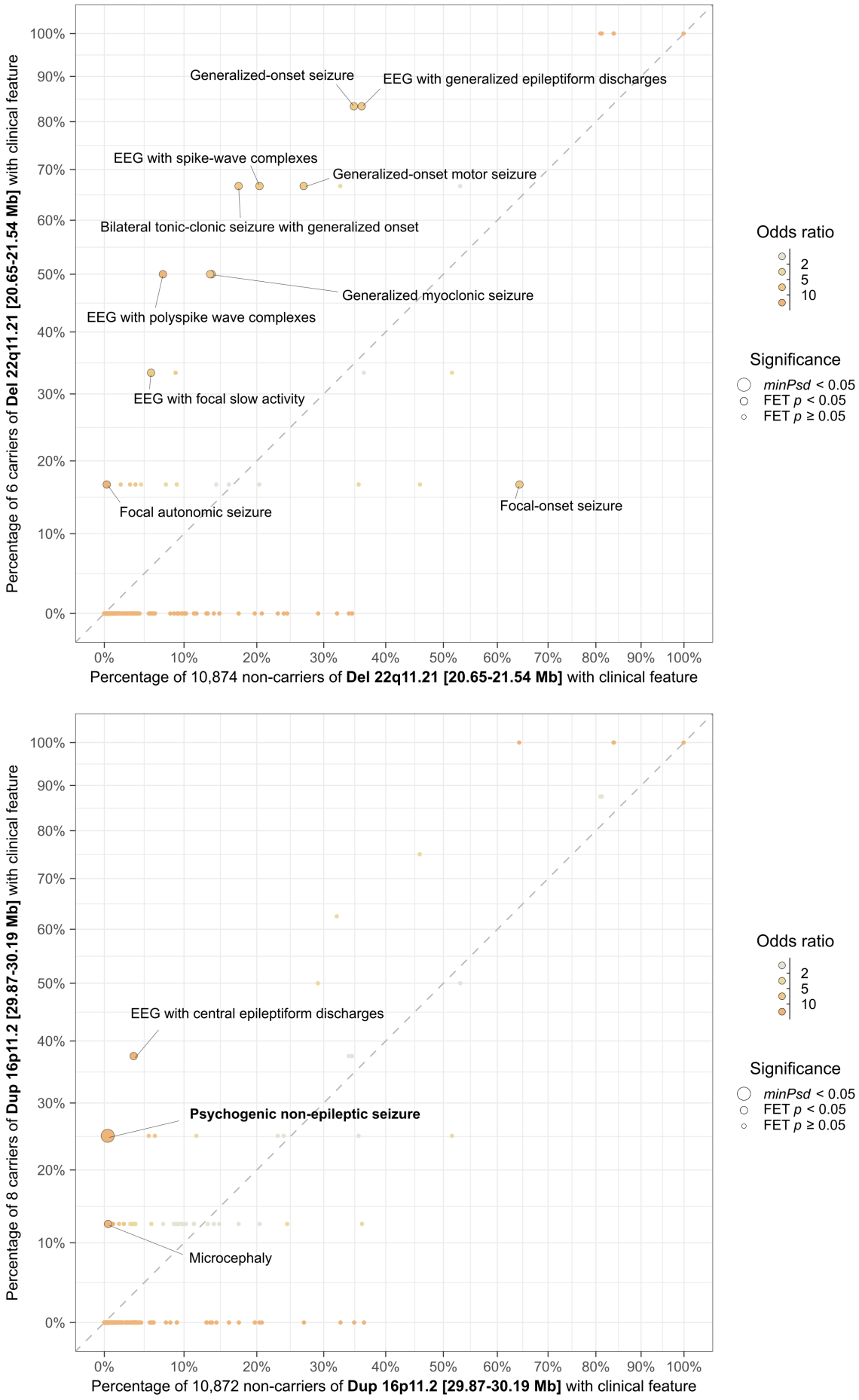


Figure 6-7. Associations of deletions at 22q11.21 and duplications at 16p11.2.

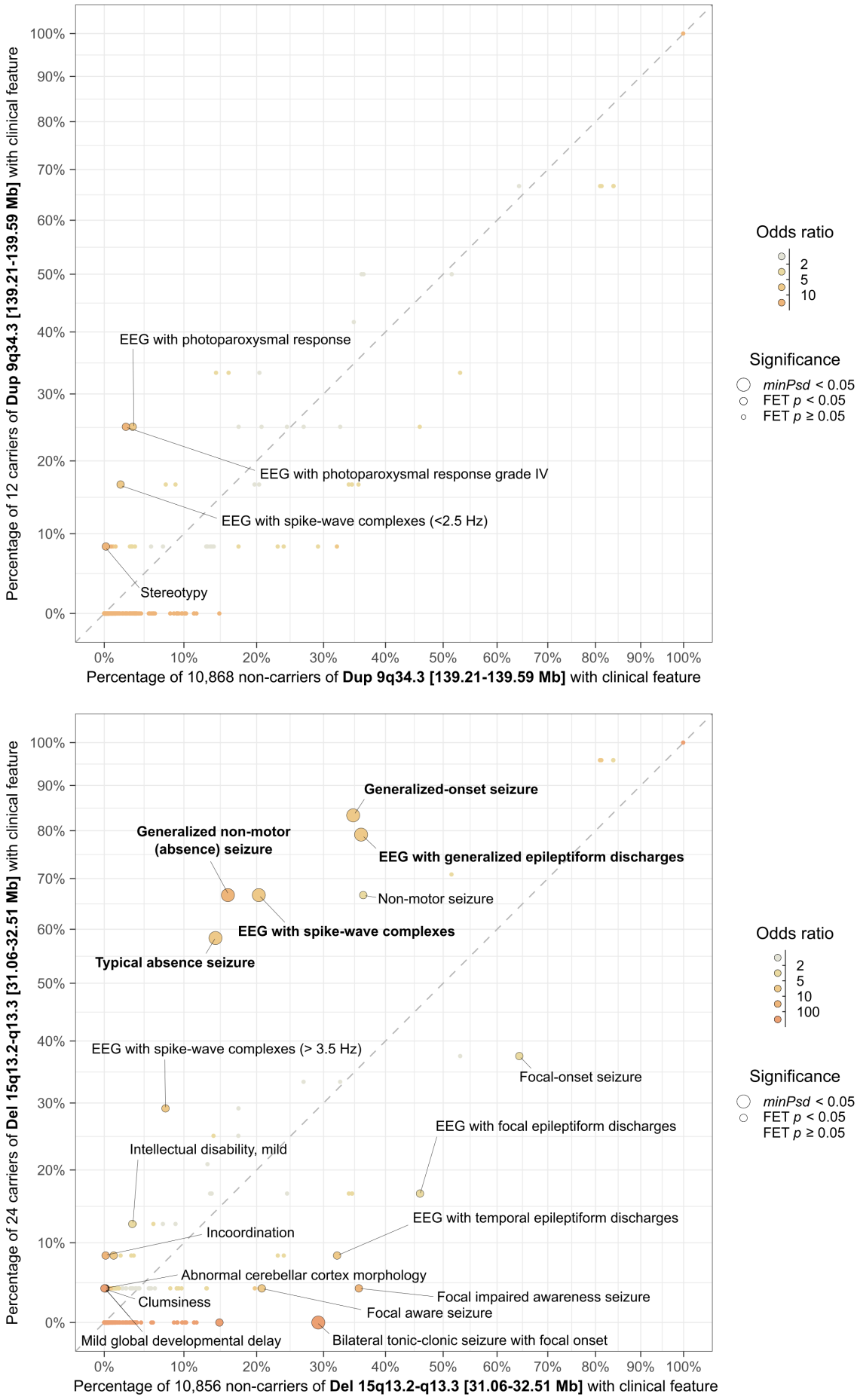


Figure 6-8. Associations of duplications at 9q34.3 and deletions at 15q13.3.

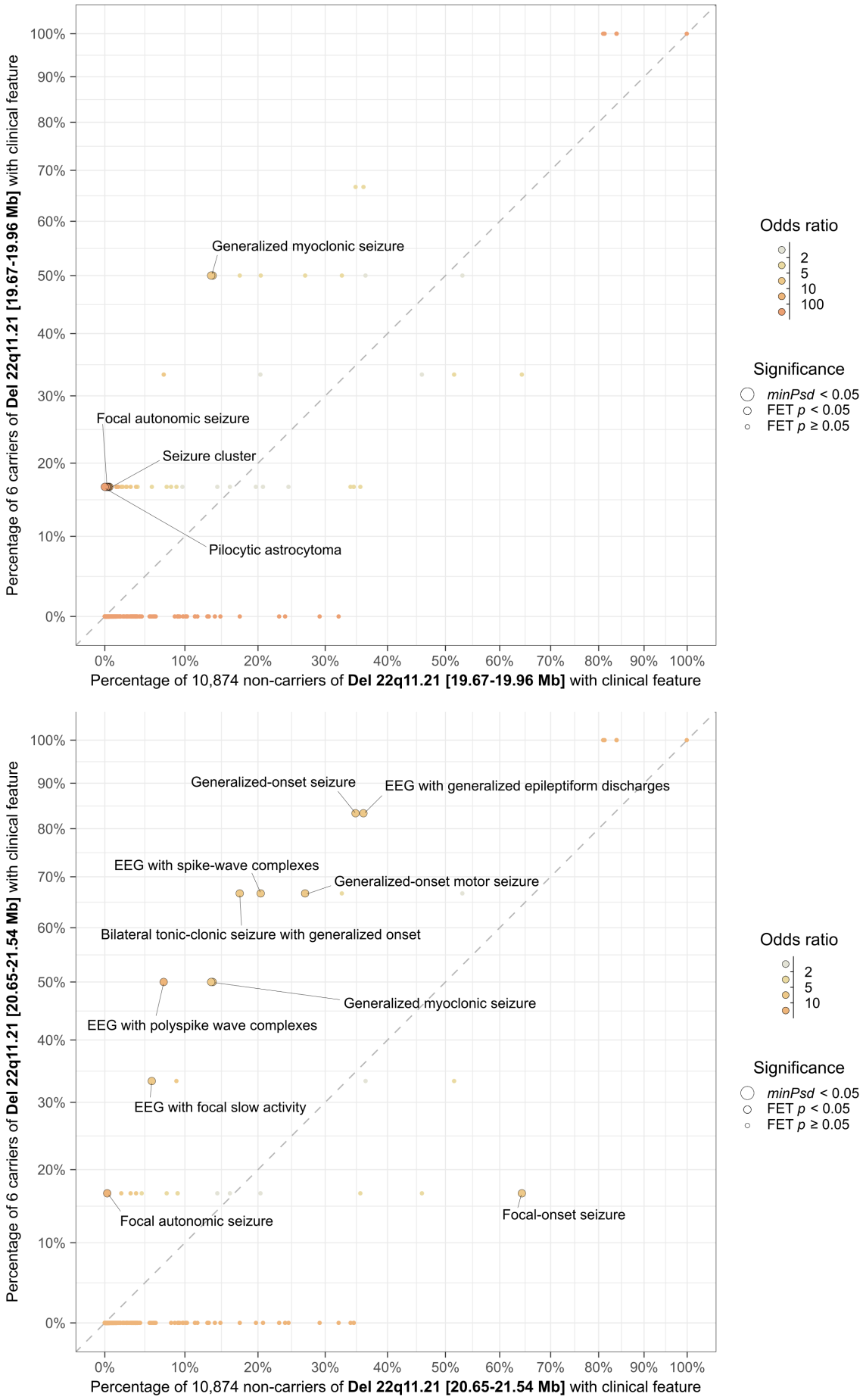


Figure 6-9. Associations of deletions at 22q11.21 [19.67-19.96 Mb] and [20.65-21.54 Mb].

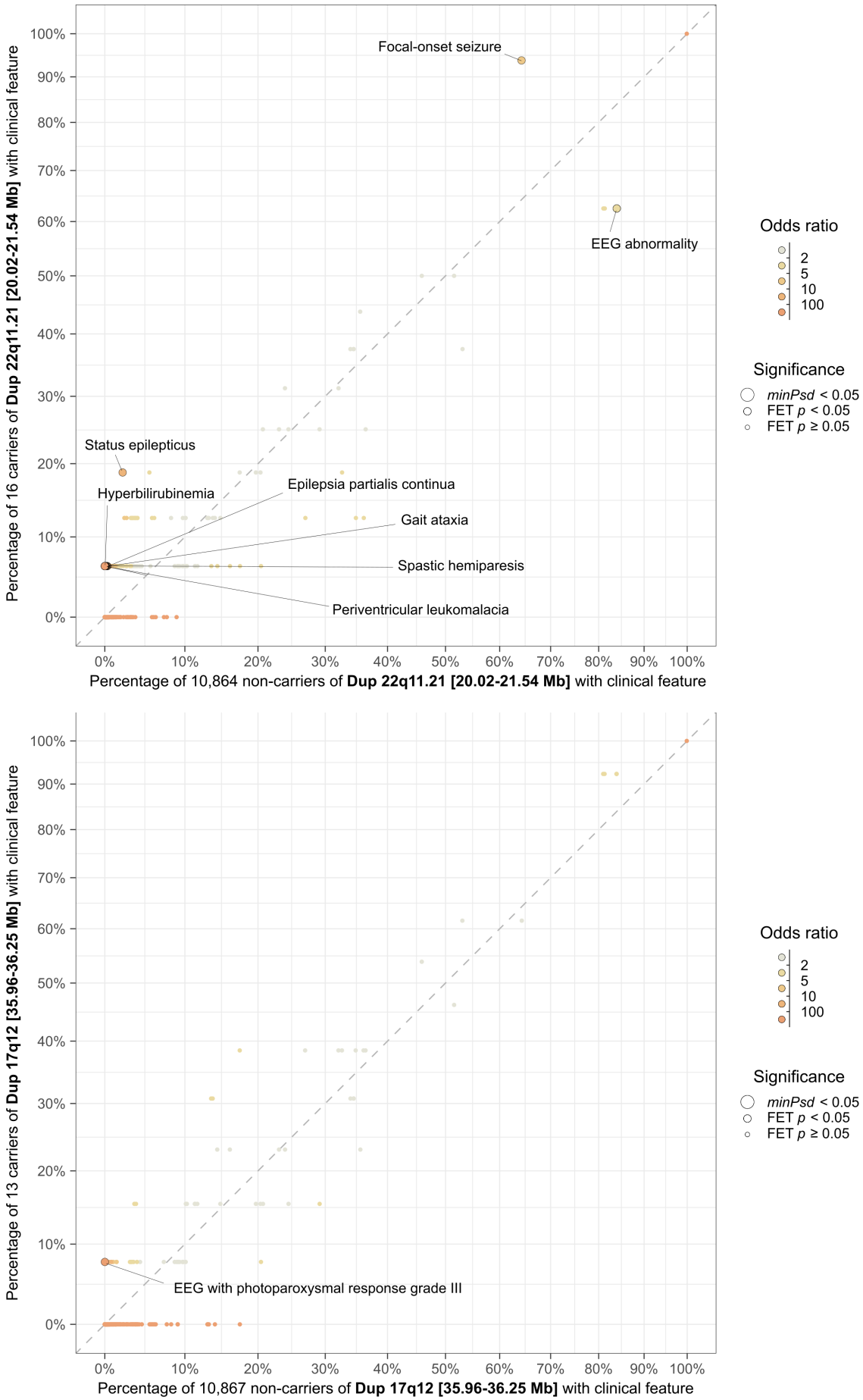


Figure 6-10. Associations of duplication at 22q11.21 [20.02-21.54 Mb] and 17q12.

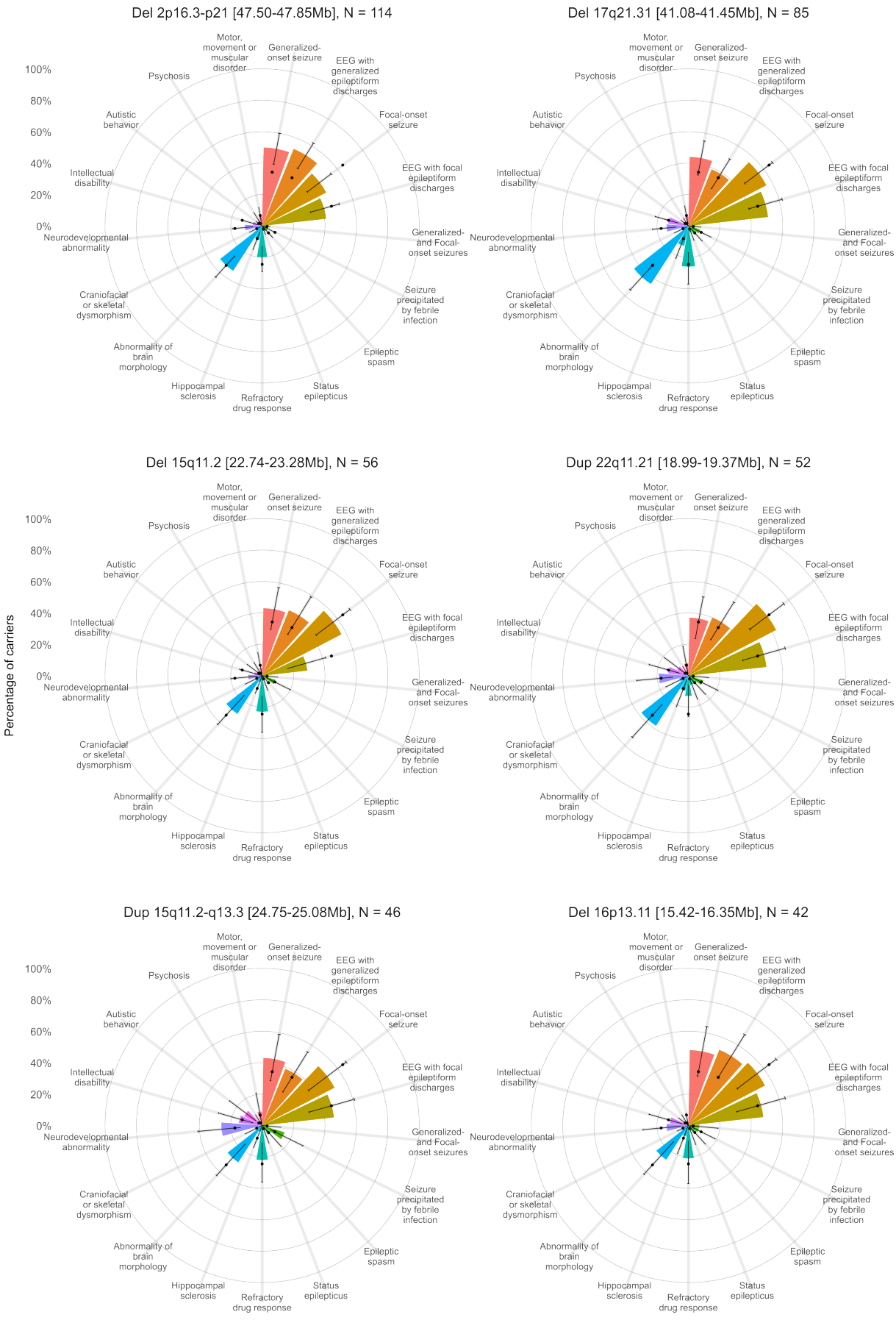
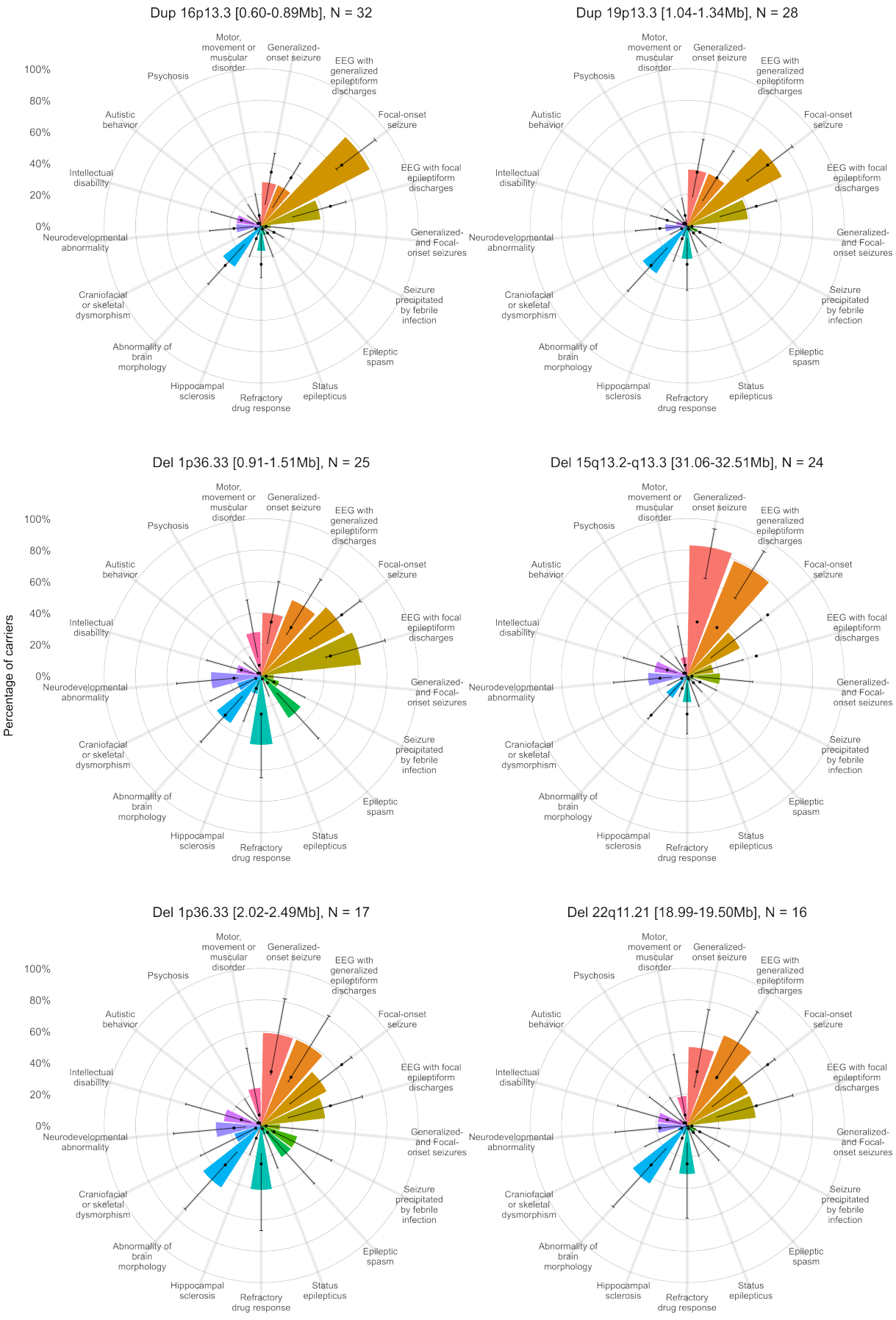
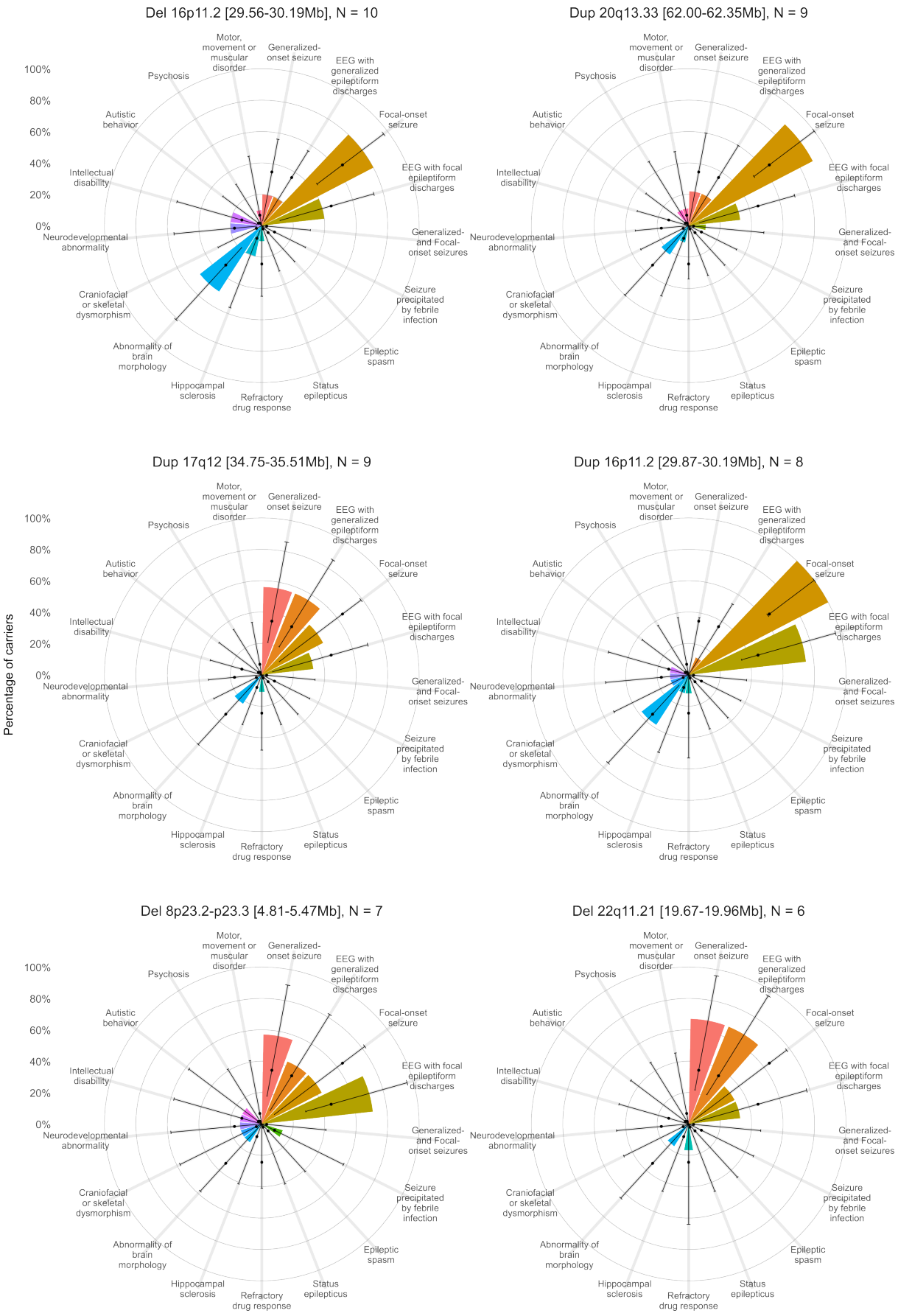
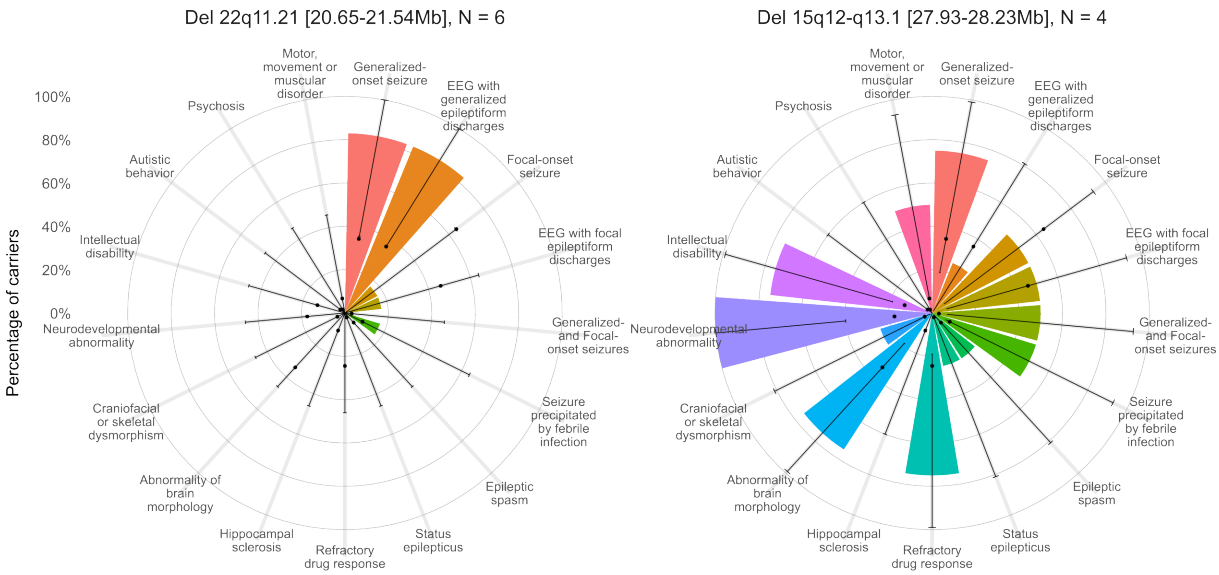


Figure 6-11. Broad clinical profiles of carriers of carriers of CNVs. (Continues for four further pages with guidance as to the interpretation of the figures on last of these.)


Dup 16p13.3 [0.60-0.89Mb], N = 32
Dup 19p13.3 [1.04-1.34Mb], N = 28
Del 1p36.33 [0.91-1.51Mb], N = 25
Del 15q13.2-q13.3 [31.06-32.51Mb], N = 24
Del 1p36.33 [2.02-2.49Mb], N = 17
Del 22q11.21 [18.99-19.50Mb], N = 16
Percentage of carriers
100%
80%
60%
40%
20%
0%
Motor, movement or muscular disorder
Generalized-onset seizure
EEG with generalized epileptiform discharges
Focal-onset seizure
EEG with focal epileptiform discharges
Generalized- and Focal-onset seizures
Seizure precipitated by febrile infection
Epileptic spasm
Status epilepticus
Refractory drug response
Hippocampal sclerosis
Abnormality of brain morphology
Craniofacial or skeletal dysmorphism
Neurodevelopmental abnormality
Intellectual disability
Autistic behavior
Psychosis


Dup 22q11.21 [20.02-21.54Mb], N = 16
Dup 9q34.3 [139.89-140.12Mb], N = 15
Dup 17q12 [35.96-36.25Mb], N = 13
Del 16p12.2 [21.88-22.50Mb], N = 13
Dup 9q34.3 [139.21-139.59Mb], N = 12
Del 8p23.2-p23.3 [3.04-3.78Mb], N = 11
Percentage of carriers
100%
80%
60%
40%
20%
0%
Motor, movement or muscular disorder
Generalized-onset seizure
EEG with generalized epileptiform discharges
Focal-onset seizure
EEG with focal epileptiform discharges
Generalized- and Focal-onset seizures
Seizure precipitated by febrile infection
Epileptic spasm
Status epilepticus
Refractory drug response
Hippocampal sclerosis
Abnormality of brain morphology
Craniofacial or skeletal dysmorphism
Neurodevelopmental abnormality
Intellectual disability
Autistic behavior
Psychosis


Del 16p11.2 [29.56-30.19Mb], N = 10
Dup 20q13.33 [62.00-62.35Mb], N = 9
Dup 17q12 [34.75-35.51Mb], N = 9
Dup 16p11.2 [29.87-30.19Mb], N = 8
Del 8p23.2-p23.3 [4.81-5.47Mb], N = 7
Del 22q11.21 [19.67-19.96Mb], N = 6
Percentage of carriers
100%
80%
60%
40%
20%
0%
Motor, movement or muscular disorder
Generalized-onset seizure
EEG with generalized epileptiform discharges
Focal-onset seizure
EEG with focal epileptiform discharges
Generalized- and Focal-onset seizures
Seizure precipitated by febrile infection
Epileptic spasm
Status epilepticus
Refractory drug response
Hippocampal sclerosis
Abnormality of brain morphology
Craniofacial or skeletal dysmorphism
Neurodevelopmental abnormality
Intellectual disability
Autistic behavior
Psychosis



Summary clinical signatures of CNVs in a deeply phenotyped epilepsy cohort. The percentage of carriers of the CNV with each broad phenotype is shown by the height of bars arranged on a polar axis, with two-sided 95 % confidence interval error bars for these percentages derived from the binomial distribution. The dots indicate the percentage of the entire Phenomic cohort of 10,880 people with each broad phenotype and are shown for reference only (representing the prior probability of a person having the phenotype without genetic stratification). The corresponding 95 % confidence intervals of the complete phenomic cohort are no wider than 1.9 % (and would not extend beyond the dot so are omitted for clarity). 'Craniofacial or skeletal dysmorphism' includes individuals with either *Abnormality of the head [HP:0000234]* (which excludes isolated brain structural abnormalities) or *Abnormal skeletal morphology [HP:0011842]*. 'Motor, movement or muscular disorder' includes individuals with any of *Abnormal central motor function [HP:0011442]*, *Abnormality of movement [HP:0100022]*, or *Abnormality of the musculature [HP:0003011]*, but not *Motor delay [HP:0001270]*, which is included in 'Neurodevelopmental abnormality'. While 'Neurodevelopmental abnormality' includes those with 'Intellectual disability', the latter is shown additionally as it is a neurodevelopmental outcome with particularly important socioeconomically important consequences.

Chapter 7. The natural history of CDKL5 Deficiency Disorder into adulthood

7.1 Preface

In this chapter I present a study that I undertook in collaboration with Dr Ángel Aledo-Serrano (Synaptia Madrid Epilepsy Center) and Prof. Rikke S. Møller (University of Southern Denmark and Danish Epilepsy Centre) using data from a team of international collaborators. I designed the data entry tool with Dr Aledo-Serrano, and Prof. Møller's team re-evaluated the diagnostic validity of variants. I decided on the analysis strategy and hypotheses to test, performed the analysis, and drafted this chapter and its figures independently.

I am grateful to the people with CDKL5 Deficiency Disorder (CDD) and their caregivers who consented to the use of their data, the clinicians who provided them (listed in Appendix B), and the assistance of the Loulou Foundation, the groups forming the CDKL5 Alliance, the International CDKL5 Disorder Database, and the EpiCARE platform helping to recruit participants.

7.2 Abstract

Knowledge of the epilepsy, neurodevelopmental, and broader spectrum of phenotypes and outcomes of CDD is limited to the results of cross-sectional analysis of largely pediatric cohorts. Assessment of outcomes in adulthood is challenging because of the diagnostic gap and duration of follow-up that would be required for prospective studies. Consequently, we aimed to delineate the natural history retrospectively from adulthood, describing the outcomes and assessing the factors influencing these.

We collected clinical data about an international cohort of 67 adults. I harmonized and analyzed demographic, phenotypic, CDKL5 Developmental Score (CDS), and treatment data, and tested associations with genetic factors, sex, and a positive or negative history of neonatal seizures, an early phenotype that I hypothesized might indicate a more severe prognosis.

The 67 adults had a median age of 24 years at last follow-up (55 females). Three had died. All but one had epilepsy, typically beginning with epileptic spasms or tonic seizures before 4 months of age, and focal-onset and non-motor seizures emerging later. Fewer than a third had been documented as having bilateral tonic-clonic seizures or status epilepticus by adulthood and seizures often improved with age. However, 73 % had never experienced 6 months of seizure-freedom. No specific antiseizure treatment demonstrated superiority, but by adulthood

the most common were clobazam, sodium valproate, and lamotrigine. Common comorbidities included movement disorders, visual impairment, sleep disorders, constipation, and scoliosis.

All participants had intellectual disability: 75 % had not acquired speech and 45 % had regressed developmentally. 16 % never achieved any CDS skill, but most attained at least three, and 28 % attained six or all seven. By adulthood, half of those who had achieved any CDS skill continued to retain all their acquired CDS skills. The skills most frequently lost were independent walking and standing.

Those with a history of neonatal seizures tended to attain fewer CDS skills and were more likely to have been annotated with abnormal muscle tone in adulthood, atrioventricular conduction delay, and potential complications of their illness and treatment.

People carrying missense variants tended to attain more CDS skills than those with other variants and were more likely to lose skills in adulthood and develop anxiety, possibly reflecting the limited neurodevelopment of those with non-missense variants, who manifested a more multisystemic disorder.

Retrospective data from adulthood elucidates the evolution of symptoms, variation in developmental outcomes, and which antiseizure treatments clinicians gravitate towards in CDD. Carriage of a non-missense variant or a history of neonatal seizures indicates a tendency towards a more complex disorder and lower developmental trajectory. These findings should inform management decisions, prognostication, and the design of clinical trials.

7.3 Introduction

Cyclin-dependent kinase-like 5 (CDKL5), encoded by *CDKL5* (Xp22.13), is a serine/threonine kinase with multiple phosphorylation targets and its deficiency results in abnormalities of axonal and dendritic development and morphology, with impaired synaptic vesicle recycling, (reviewed by Van Bergen *et al.*, 2022). Pathogenic variants in *CDKL5* cause one of the more *common* of the rare genetic DEEs: CDD, which presents with seizures before 36 months of age in 2.36 per 100,000 livebirths (Weaving *et al.*, 2004; Symonds *et al.*, 2019).

CDD is an early-onset severe neurodevelopmental disorder with drug-resistant seizures, motor and cognitive developmental stagnation or regression, behavioral disturbances, cerebral visual impairment, sleep dysregulation, and gastrointestinal dysfunction (MacKay *et al.*, 2021; Leonard *et al.*, 2022a). Knowing the natural history according to etiology complements knowledge of clinical syndromes to inform care for an individual with epilepsy (Scheffer *et al.*,

2017). Across other rare genetic DEEs, older individuals are receiving molecular diagnoses and those affected may be living longer with advances in pediatric care (Borlot *et al.*, 2019; Johannesen *et al.*, 2020; Lewis-Smith *et al.*, 2020b; Minardi *et al.*, 2020; Selvarajah *et al.*, 2021; Zacher *et al.*, 2021; de Maria *et al.*, 2022; Fowkes *et al.*, 2022; Stamberger *et al.*, 2022; Nascimento *et al.*, 2023; Rong *et al.*, 2023; Postma *et al.*, 2024). When caring for a particular patient, accurate and precise prognostication in clinic requires comprehensive knowledge of the clinical landscape and range of trajectories of a disorder, which could be enhanced by known associations between clinical features, or genotype-phenotype associations. Importantly, in other DEEs some comorbidities become more apparent in adulthood (Vlaskamp *et al.*, 2019; Selvarajah *et al.*, 2022; Stamberger *et al.*, 2022; Rong *et al.*, 2023). Consequently, knowledge of the natural history of disorders is necessary for clinicians to prioritize potential comorbidities for screening during follow-up, and economists to model disease outcomes for service development and the economic consequences of a disease. However, no studies have focused on the clinical evolution of CDD through to, and within, adulthood, which limits the evidence available to inform management of adolescents and adults and to prognosticate. Full evaluation of the efficacy of candidate disease-modifying treatments will require greater elucidation of the natural history.

Previously, a cross-sectional study of two composite outcome measures in 285 people with CDD found associations with some recurrent genotypes (MacKay *et al.*, 2021). However, the median age was 4.3 years, and the number of adults was unclear: only 38 participants had been followed up beyond 13 years of age. Here, I report the clinical features and histories of 67 adults with disease-causing *CDKL5* variants, describing how the clinical manifestations of CDD evolve into adulthood, and assessing the effect of sex, genotype, and onset with neonatal seizures.

7.4 Methods

7.4.1 Participant recruitment and data collection

Participants were recruited for this retrospective study through an international network of pediatric and adult neurologists and medical geneticists, and the International CDKL5 Disorder Database (Fehr *et al.*, 2015). For inclusion, participants had to be at least 18 years old at last clinical review and have a pathogenic or likely pathogenic variant in *CDKL5* MANE select transcript (Morales *et al.*, 2022) (NM_001323289.2, ENST00000623535.2) according to the ACMG/AMP guidelines (Richards *et al.*, 2015), assessed using *InterVar* (Li and Wang, 2017). These could include deletion of *CDKL5* exons or the entire gene, noting if other genes were

affected. We obtained updated information about the 34 participants in the International CDKL5 Disorder Database (Fehr *et al.*, 2015; Fehr *et al.*, 2016a; Mangatt *et al.*, 2016; Lim *et al.*, 2017; Mori *et al.*, 2017; Lim *et al.*, 2018; Tangarorang *et al.*, 2019; Dale *et al.*, 2021; Leonard *et al.*, 2021; Demarest *et al.*, 2022; Leonard *et al.*, 2022b) and individuals who have been reported previously (Archer *et al.*, 2006; Hagebeuk;van den Bossche and de Weerd, 2013; Ortega-Moreno *et al.*, 2017; Demarest *et al.*, 2019b; Musante *et al.*, 2022) (**Supplementary Table 7-1**). Data were entered into the International CDKL5 Disorder Database by caregivers rather than generated by the current investigators or extracted from peer-reviewed publications. Dr Aledo-Serrano and collaborators contacted caregivers of participants where necessary to clarify details.

7.4.2 Protocol approval and consent

Written informed consent for participation and publication was obtained from all parents or legal guardians. The study protocol and design were approved by the ethics committee of Madrid Ruber International Hospital and the New York University Langone Health IRB.

7.4.3 Mapping of phenotypic data

Clinical data were entered into a study-specific REDCap resource (Harris *et al.*, 2009) we adapted from the Epi25 REDCap template (Epi25 Collaborative, 2020). I mapped clinical features from REDCap fields to HPO concepts (version 1.7.14, release date 2022-02-14) with binning of quantitative gestational age, birth weight, and head circumference data and modification of relationships between concepts as described previously (5.4.5, 5.4.6, and 5.4.7). I added 19 study-specific concepts to the HPO to maximize the capture of clinical features that have been prioritized in CDD. These include multiphasic seizures, more granular subtypes of *Stereotypy [HP:0000733]*, age-specific forms of phenotypes such as *Hypertonia [HP:0001276]*, and specification of how active seizures were at the time of any *Regression [HP:0002376]*. I used the in-house identifier prefix *mHP:* for these to avoid confusion with official HPO concepts (**Table 7-1**).

Term	Parents	Children
Modified existing terms		
Bruxism [HP:0003763] [a]	Behavioral abnormality [HP:0000708]	None
Study-specific terms		
Neonatal hypertonia [mHP:0000001]	Hypertonia [HP:0001276]	None
Adult hypertonia [mHP:0000002]	Hypertonia [HP:0001276]	Adult dystonia [mHP:0000004] Adult spasticity [mHP:0000005]
Adult hypotonia [mHP:0000003]	Hypotonia [HP:0001252]	None
Adult dystonia [mHP:0000004]	Dystonia [HP:0001332] Adult hypertonia [mHP:0000002]	None
Adult spasticity [mHP:0000005]	Spasticity [HP:0001257] Adult hypertonia [mHP:0000002]	None
Multiphasic seizure [mHP:0000006]	Seizure [HP:0001250]	None
Neurodevelopmental delay prior to seizure-onset [mHP:0000007]	Neurodevelopmental delay [HP:0012758]	None
Regression prior to seizures [mHP:0000008]	Developmental regression [HP:0002376]	None
Regression during stable or improving seizures [mHP:0000009]	Developmental regression [HP:0002376] Seizure [HP:0001250]	None
Regression in adulthood [mHP:0000010]	Developmental regression [HP:0002376]	None
Regression while seizures active [mHP:0000011]	Developmental regression [HP:0002376] Seizure [HP:0001250]	None
Regression while seizure-free on treatment [mHP:0000012]	Developmental regression [HP:0002376]	None
Regression while seizure-free off treatment [mHP:0000013]	Developmental regression [HP:0002376]	None
Stereotypical humming [mHP:0000014]	Stereotypy [HP:0000733]	None
Stereotypical touching of face [mHP:0000015]	Stereotypy [HP:0000733]	None
Stereotypical insertion of hand in mouth [mHP:0000016]	Stereotypy [HP:0000733]	None
Stereotypical leg crossing [mHP:0000017]	Stereotypy [HP:0000733]	None
Stereotypical oromotor behavior [mHP:0000018]	Stereotypy [HP:0000733]	None
Stereotypical vocalization [mHP:0000019]	Stereotypy [HP:0000733]	None

Table 7-1. Modification of HPO version 1.7.14 (release 2022-02-14) for this study. Terms modified as in Table 5-1 not shown. Parentage of *Bruxism [HP:0003763]* was changed in recognition that bruxism can occur while awake. This is how *Bruxism [HP:0003763]* has been represented in the more HPO version 1.8.12 (released 2023-10-09).

7.4.4 Negated phenotypes

While other investigators have made close-world assumptions, particularly regarding age-dependent phenotypes in cohorts of varied ages (Andrews *et al.*, 2015), I decided to interpret phenotypes conservatively if they were not documented as present. I only annotated their absence if a clinician had indicated their absence, and, for those phenotypes that require an investigation such as cross-sectional imaging, only if this investigation had been undertaken.

Some phenotypes were excluded explicitly up until the age of the most recent clinical assessment. From these phenotypes one can infer other phenotypic terms that can be negated by reverse propagation towards the terminal leaf branches of the HPO rather than towards its root (see 3.8; Crawford *et al.*, 2021). This generates many negated annotations (for example,

no Seizure [NP:0001250] implies the absence of all subclasses of *Seizure [HP:0001250]*, over 300 concepts), many of which are redundant for our analyses because they are globally absent in the cohort or are annotated to only the identical set of people as their more generic parent HPO concept. Accordingly, I pruned the cohort's negated annotations to retain only those concepts that after reverse propagation were annotated negatively to more people than any of their parent concepts and those that were annotated positively to at least one person in the study.

In some cases, phenotypes were excluded with the caveat that they were dynamic phenotypes that could only be detected using a clinical investigation. This leaves some uncertainty as to whether they might have been detected at some other time, particularly if continuous thorough screening were feasible. This is pertinent to electroencephalographic features that conceivably may have occurred only at times when electroencephalography was not being performed. However, uncertainty due to sampling applies to some extent for many phenotypes that are not obvious congenital and permanent malformations. I did accept negation of electroencephalographic phenotypes if electroencephalography had been performed and the phenotype had not been reported, accepting some uncertainty as this reflects reasoning in clinical practice. The interpretation of such an annotation is that the phenotype was absent within the limitations of active screening for it according to routine clinical standards, rather than an absolute assertion that it could not have occurred.

I assessed each person's sets of positive and negative propagated annotations to identify contradictions: those phenotypes that were explicitly or implicitly said to be both present at some age and absent in the same individual. Conflicting annotations were resolved by expert review of REDCap data (D.L.-S.) and confirmation in raw data when necessary (A.A.S.).

After propagation of positive and negative annotations, I was able to quantify the *phenotyping gap* for each HPO concept: the proportion of people whose data were insufficient to establish the presence or absence of the respective phenotype (Lewis-Smith *et al.*, 2022).

7.4.5 Clinical outcomes

I explored the onset of each seizure type in each person by propagating their positive seizure annotations, and then taking the minimum age of onset of all the subtypes of the inferred HPO seizure phenotype. For example, individual CDKL5_09 had two types of *Motor seizure [HP:0020219]*: *Epileptic spasm [HP:0011097]* starting on day 3 of life and *Tonic seizure [HP:0032792]* starting aged 2 years. Accordingly, her predisposition to *Motor seizure*

[HP:0020219] first manifested at 2 days old. The maximum age at each seizure type was calculated similarly from the maxima.

For each seizure type, individuals with any missing ages of onset were excluded from analysis of onset and longitudinal duration, and similarly, where the age at the last occurrence of the seizure phenotype or one of its subtypes was missing, this individual was censored from the analysis of activity. Consequently, the proportion of people with more common, generic seizure phenotypes, such as *Seizure [HP:0001250]* and *Motor seizure [HP:0020219]*, who had sufficient data to be included tended to be lower than for some rarer, more precise seizure phenotypes such as *Tonic seizure [HP:0032792]* and *Focal autonomic seizure [HP:0011154]*.

Epilepsy outcomes and the consequences of ASM, vagus nerve simulation (VNS), and the ketogenic diet were annotated based on the impressions of caregivers and treating physicians. Current seizure freedom was defined as the period without seizures being at least three-fold longer than the longest inter-seizure interval in the preceding year (Kwan *et al.*, 2010), with a minimum of 6 months to minimize the risk of bias due to the transient honey-moon periods on starting a new ASM in CDD (Fehr *et al.*, 2016b; Muller *et al.*, 2016).

CDS was calculated on a scale of 0–7 by summation of developmental skills obtained (Demarest *et al.*, 2019b). We aimed to identify the developmental milestones eventually acquired by adulthood and how many people had lost skills by their most recent follow-up. Precise age annotations of CDS were rarely available. However, we were able to calculate the person's CDS total from item scores at most recent follow-up in adulthood (when clinically stable) and compare this to their historic maximum to allow identification of skill loss. The maximum CDS was age-agnostic. Consequently, it is possible that a person never held all the skills concurrently.

7.4.6 Genetic stratification

I grouped participants according to their *CDKL5* variant in MANE select transcript (Morales *et al.*, 2022) NM_001323289.2/ENST00000623535.2 (**Supplementary Table 7-1**). Primary genetic analyses focused on the coarse dichotomy between missense variants and non-missense variants because of the uncertainty as to the optimal biological and clinical stratification of variants according to more complex features (Fehr *et al.*, 2015; MacKay *et al.*, 2021). However, I undertook additional analyses according to each:

1. Unique sequence variant using Human Genomic Variation Society nomenclature.

2. Variant Effect Predictor consequence, encompassing missense, frameshift, stop gain, and splicing (McLaren *et al.*, 2016). Splicing variants were further validated using SpliceAI predictions, accepting variants as affecting splicing if any SpliceAI donor or acceptor prediction (delta score) ≥ 0.5 as recommended by Jaganathan *et al.* (2019).[32]

3. Location within the protein. While doubts remain about the optimal location-based groupings (Fehr *et al.*, 2015; MacKay *et al.*, 2021), I stratified variants into three groups: those residing in the kinase domain (amino acids 13–297), those in middle of the protein (amino acids 298–781), and those in C-terminus downstream of amino acid 781, where a nuclear import motif is suspected (Kilstrup-Nielsen *et al.*, 2012; UniProt, 2023).

4. Large deletions, including at least a complete exon, but often the whole gene. We identified two people whose deletions had been shown to include *NHS*. We also considered these separately because of this gene's known clinical significance as the cause of X-linked dominant Nance-Horan Syndrome, which includes neurodevelopmental, visual (often cataract), and dental disorders with dysmorphia [OMIM:302350] (McKusick-Nathans Institute of Genetic Medicine, 2020). Additional loss of *NHS* might be expected to contribute or exacerbate such features in people with CDD. *CDKL5* deletions involving *NHS* also span *SCML1*, *RAI2*, *BEND2*, and *SCML2*, genes of unknown clinical relevance.

5. Demonstrated de novo or inherited transmission.

6. Demonstrated mosaicism, noting that for many participants, we did not know if sufficient investigation had been undertaken to identify or exclude mosaicism.

7.4.7 Statistical analysis

I undertook analyses after stratifying people by demographic and genetic factors, and the presence or absence of neonatal seizures (*Neonatal electro-clinical seizure [HP:0032809]*), hypothesizing that these might indicate clinically relevant strata in CDD. Neonatal seizures are an early onset phenotype that may be the first manifestation of an underlying severe neurodevelopmental disorder, and/or may contribute to the underlying disorder as an epileptic encephalopathy (Scheffer *et al.*, 2017).

[32] The lowest SpliceAI delta score observed across all donor and acceptor predictions for variants meeting this criterion was 0.75.

I tested the association between the presence and absence of a phenotypic association with membership of these strata. I tested positive and negated phenotypes separately (for example, one contingency table for each of *Neonatal electro-clinical seizure [HP:0032809]* and *no Neonatal electro-clinical seizure [NP:0032809]*) in recognition of phenotyping gaps (Lewis-Smith *et al.*, 2022). I calculated the conditional maximum likelihood estimate derived OR, RR, and *p*-values using FET (*stats::fisher.text()*), reporting unadjusted two-sided *p*-values and applying the *minP stepdown* (*minPsd*, using the *NRejections::adj_Wstep* function, version 1.2.0) procedure to each stratum (Westfall and Wolfinger, 1997).

I used a $\mathrm{Sim}_{\mathrm{MinTO}}$ as the primary similarity measure, with 25 secondary measures that might provide validation (in the case of other measures prioritizing shared information and an open-world interpretation) or an alternative perspective, in particular, a closed-world perspective using $\mathrm{Sim}_{\mathrm{Goodall3}}$. Only positive annotations were used in similarity analyses, with all those phenotypes not inferred to have been present presumed to be absent when applying $\mathrm{Sim}_{\mathrm{Goodall3}}$. I interpreted the median pairwise similarities of each group using Monte Carlo simulation with 1,000,000 randomly selected groups of each group size.

For similarity analysis of those with and without neonatal seizures, I created a new similarity array using the same similarity measures but having censored any of *Neonatal seizure [HP:0032807]* and its descendants. I performed the analysis on this similarity array for two groups: those annotated as having had neonatal seizures (annotated with *Neonatal electro-clinical seizure [HP:0032809]*) and those asserted to have not had neonatal seizures (annotated with *no Neonatal electro-clinical seizure [NP:0032809]),* generating new Monte Carlo simulations for these groups from this different array.

7.5 Results

7.5.1 Description of the study cohort

Sixty-seven adults from 66 families (one pair of monozygotic twins) were included (**Figure 7-1A** and **Table 7-2**). Most were in their twenties at last follow-up, with the eldest aged 57 years The specialty of the most recent responsible clinician was known for 46 people. The 26 (57 %) remaining under the care of a pediatric neurologist tended to be younger (median age at last follow-up = 21.5 versus 28.0 years old, range = 18–36 versus 18-57 years old, Wilcoxon rank sum test with continuity correction, WRST $W = 148$, two-sided $p = 0.0130$) than the 20 (42 %) under an adult neurologist. Excluding large deletions (in ten people), 49 different variants were detected, seven recurrently with c.183del (p.Met63Cysfs*13) carried by monozygotic twins

(**Figure 7-1B** and **Supplementary Table 7-1**). Participants were recruited across 13 countries (**Figure 7-2**), 20 from the International CDKL5 Disorder Database. Ancestry was predominantly reported as Caucasian (*n* = 58, 87 %). Three were Hispanic and two Asian.

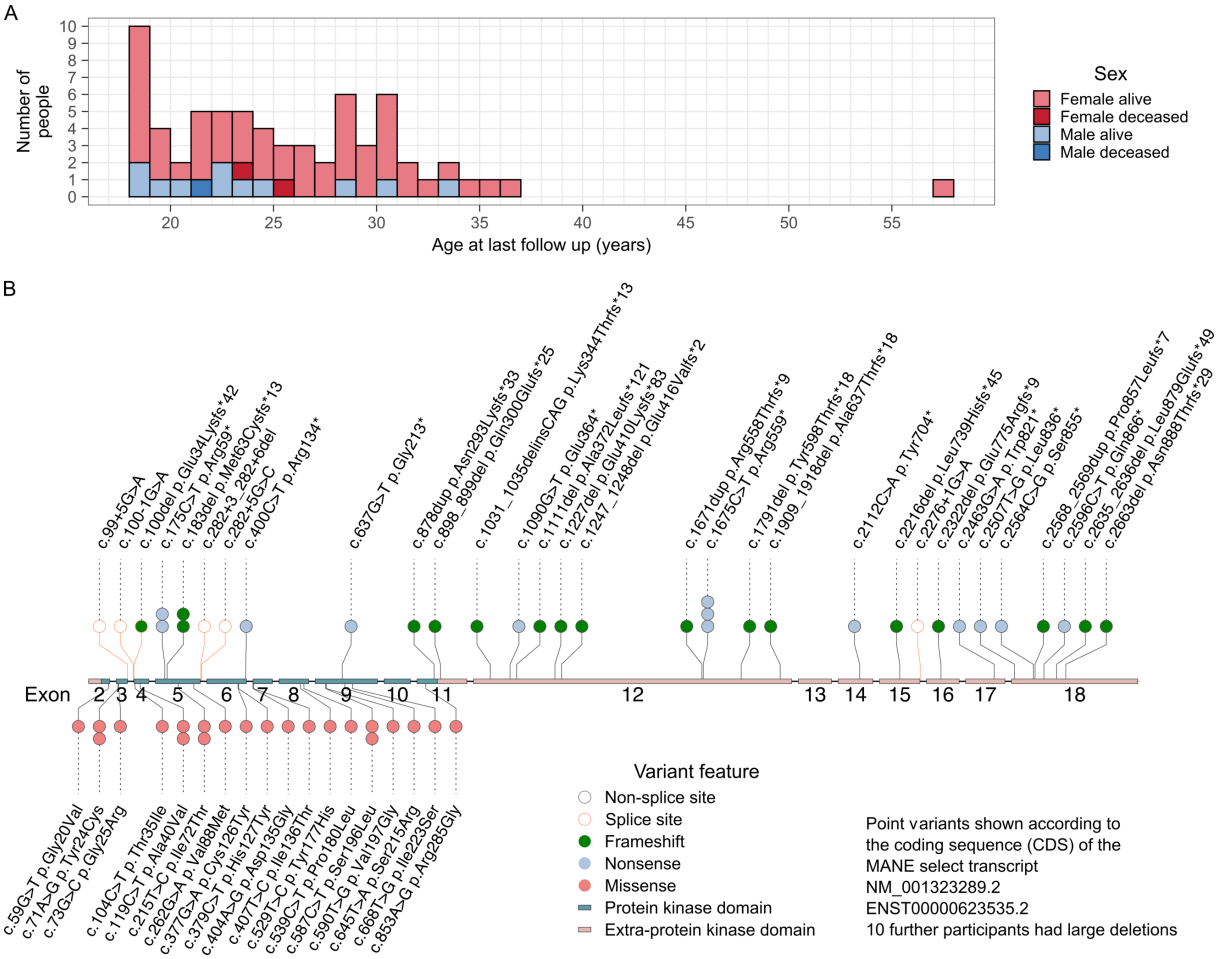


Figure 7-1. Age at last follow up (A) and *CDKL5* variants (B) in the cohort.

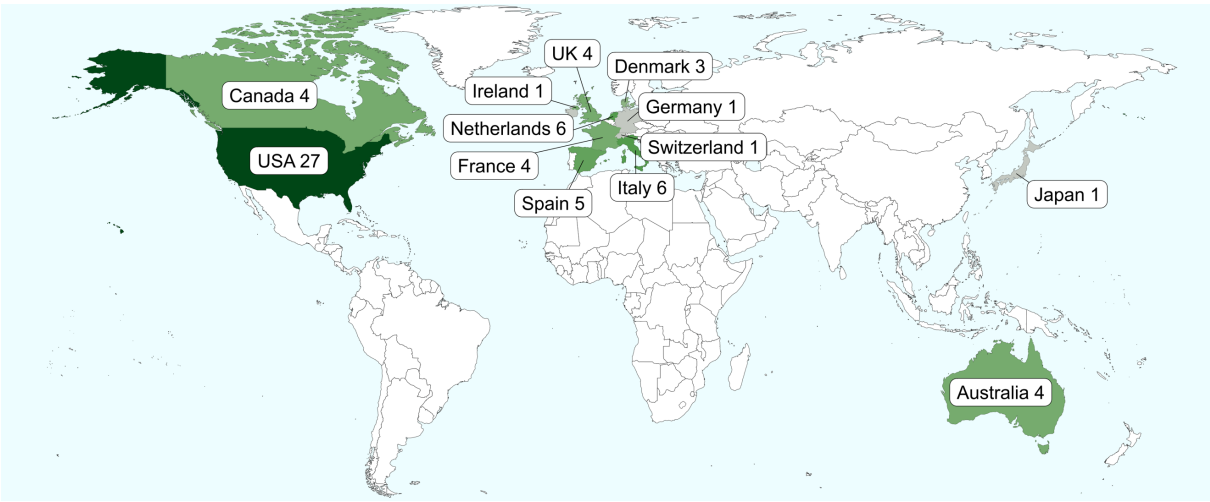


Figure 7-2. The geographic distribution of the cohort.

Feature	Number or median (Number with data)	Percentage of those with data or minimum, interquartile range, maximum
Sex		
Female	55 (67)	82 %
Male	12 (67)	18 %
Vital status		
Alive at last data collection	64 (67)	96 %
Deceased	3 (67)	4.5 %
Accommodation at last follow-up		
Most nights spent with family	45 (66)	68 %
Completely institutionalized	15 (66)	23 %
Lead neurologist at last follow-up		
Pediatric	26 (46)	57 %
Adult	20 (46)	43 %
Variant summary		
Missense kinase domain	22 (67)	33 %
Non-missense kinase domain	12 (67)	18 %
Non-missense non-kinase domain	23 (67)	34 %
Large deletion	10 (67)	15 %
Confirmed de novo	40 (41)	98 %
Confirmed inherited	1 (41)	2.4 %
Confirmed mosaicism	5 (unknown)	7.5 % (of all 67)
CDKL5 Developmental Score		
Total maximum CDKL5 Developmental Score [a]	3 (67)	0, 1–6, 7
Ever able to sit independently	42 (67)	63 %
Ever able to stand independently	32 (67)	48 %
Ever able to walk independently	28 (67)	42 %
Ever able to babble	42 (67)	63 %
Ever able to use a word	17 (67)	25 %
Ever able to use a raking grasp	47 (67)	70 %
Ever able to use a pincer grasp	18 (67)	27 %
Recent Total CDKL5 Developmental Score [a]	2 (67)	0, 1–4, 7
Able to sit independently at last data collection	38 (67)	57 % (90 % of those attaining skill)
Able to stand independently at last data collection	19 (67)	28 % (59 % of those attaining skill)
Able to walk independently at last data collection	14 (67)	21 % (50 % of those attaining skill)
Able to babble at last data collection	36 (67)	54 % (86 % of those attaining skill)
Able to use a word at last data collection	13 (67)	19 % (76 % of those attaining skill)
Able to use a raking grasp at last data collection	42 (67)	63 % (89 % of those attaining skill)
Able to use a pincer grasp at last data collection	14 (67)	21 % (78 % of those attaining skill)
Age		
Age at last data collection (years)	24 (67)	18, 21–29, 57
Age at diagnosis (years)	16 (55)	2, 14–22, 51
Attained independent sitting (months) [b]	12 (25)	3, 8–15, 26
Attained use of a word (months) [b]	19 (8)	15, 17.5–31.5, 72
Attained independent walking (months) [b]	30 (17)	12, 17–60, 108
Regression thought have begun in those with any documented age of regression (months)	132 (27)	8, 27–228, 624
Regression thought have begun in those with documented age of regression prior to adulthood (months)	48 (18)	8, 18–129, 168

Table 7-2. Description of the demographics, CDKL5 Developmental Scores, and variants of the cohort. [a] Minimum total CDKL5 Developmental Score = 0, maximum = 7. [b] Among those attained the skill.

7.5.2 Phenotypic annotation

3,441 data points relating to 67 participants were mapped directly to HPO concepts: 2,506 asserting a phenotype to have been present and 935 indicating that a phenotype had been explicitly absent thus far. Values from some REDCap data fields mapped to recurrent HPO concepts. Removal of duplicate phenotypic annotations per person reduced this to 3,236.

Propagation of positive and negative annotations of each person allowed identification of contradictions where a phenotype was inferred to have been both present at some time before data capture and never present in the same person. Positive annotations usually took preference following manual review. Taking a more conservative interpretation of the negation of phenotypes than phenotypes believed to be present is in keeping with the greater confidence that clinicians can often take in confirming a clinical finding than in excluding one that might be subtle or paroxysmal and consequently not detected at the time of assessment. Contradictions were typically explained by misalignment of broad clinical concept labels between the HPO and how these terminologies are used by the clinical investigators. For example, in this version of the HPO (released 2022-02-14), *Stereotypy [HP:0000733]* is a type of *Involuntary movements [HP:0004305]*, which is a type of *Abnormality of Movement [HP:0100022]*. However, many clinicians would consider a stereotypy to be an abnormal behavior rather than a movement disorder, and thus would consider the presence of stereotypy to not be inconsistent with a patient having no abnormality of movement in the traditional neurological sense at the time of data entry.[33] In one case, CDKL5_60, I decided to censor a positive annotation in addition to two negative annotations. CDKL5_60 was a boy born at 31 weeks weighing 1826 g, who had manually been coded as being *Small for gestational age [HP:0001518]* despite 1826 g falling between the 50–75th centiles for sex and a gestational age of 31–32 weeks according to the INTERGROWTH-21st Project (Villar *et al.*, 2014), and UK-WHO growth charts for neonates (Royal College of Paediatrics and Child Health;Department of Health and Scottish Government, 2009). I censored the explicit *Small for gestational age [HP:0001518]* annotation as

[33] The recent HPO version 1.8.12 (released 2023-10-09) recognizes this, renaming the stereotypy concept *Abnormal repetitive mannerisms [HP:0000733]* and making it a type of *Restricted or repetitive behaviors or interests [HP:0031432]* and a descendent of *Abnormality of mental function [HP:0011446]* without being a descendant of *Abnormality of movement [HP:0100022]*. This emphasizes the importance of those performing data entry and interpreting HPO annotations being aware of the phenotypic concepts subsumed by hypernyms of the HPO in the version they are working with, the need for consistency checks when using both positive and negative annotations, and the importance of clinicians engaging in HPO development to ensure it reflects clinical practice.

well as the annotations based on the *z*-scores calculated from the weight and gestational age provided: *no Large for gestational age [NP:0001520]* and *no Small for gestational age [NP:0001518]*.

This quality control to resolve apparently contradictory annotations refined the number of annotations before propagation to 3,217 (median = 48, range = 23–67 per person), comprising 2,303 positive (median = 34, range = 9–54) and 914 negative annotations (median = 14, range = 6–21).

Propagation resulted in 5,894 positive annotations to the cohort (median = 88, range = 29–152 per person), with a repertoire of 503 distinct concepts. Refinement of each participant's positive annotations to a minimal set sufficient to infer their propagated set completely, assuming knowledge represented by the HPO, showed that all phenotypes reported or inferred as present in the cohort could be captured using 1,669 annotations (median = 25, range = 7–41 per person). Clinical data entry is burdensome and remains a major bottleneck to clinical research. Considering the efficiency of our REDCap-HPO pipeline, this indicates redundancy of 33 % (1-(1,669/2,506)) for the capture of concepts asserting that a phenotype was present.

Reverse propagation of negated phenotypes followed by pruning to only those phenotypes that were either annotated as being present in at least one individual, or that were annotated differently from any of their parent concepts in this cohort, resulted in 1,845 annotations representing 98 negated phenotypic concepts. Of these, 1,729 annotations indicated the absence of one of 94 clinical phenotypes that were annotated as present in at least one person in the cohort, and 116 annotations indicated the absence of one of four clinical phenotypes that were not annotated as present in any member of the cohort.

In total, we obtained 7,739 propagated annotations for the cohort relating to the presence or negation of 507 HPO concepts that describe the cohort's phenotypic repertoire (**Figure 7-3**, **Figure 7-4**, and **Supplementary Table 7-2**).

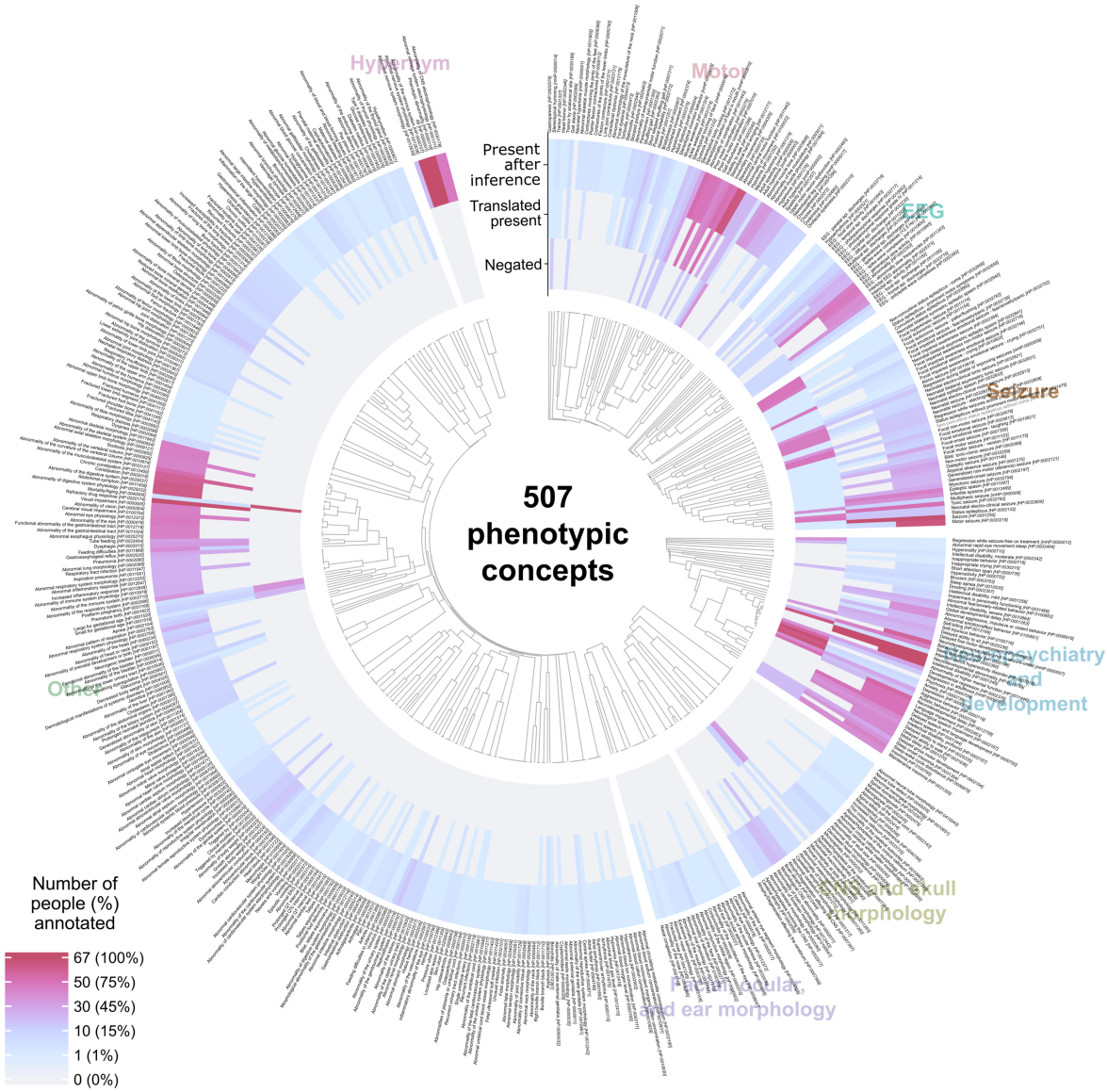


Figure 7-3. Visualization of phenotypic data harmonization. The difference in intensity between the middle and outer track indicates the effect of propagation.

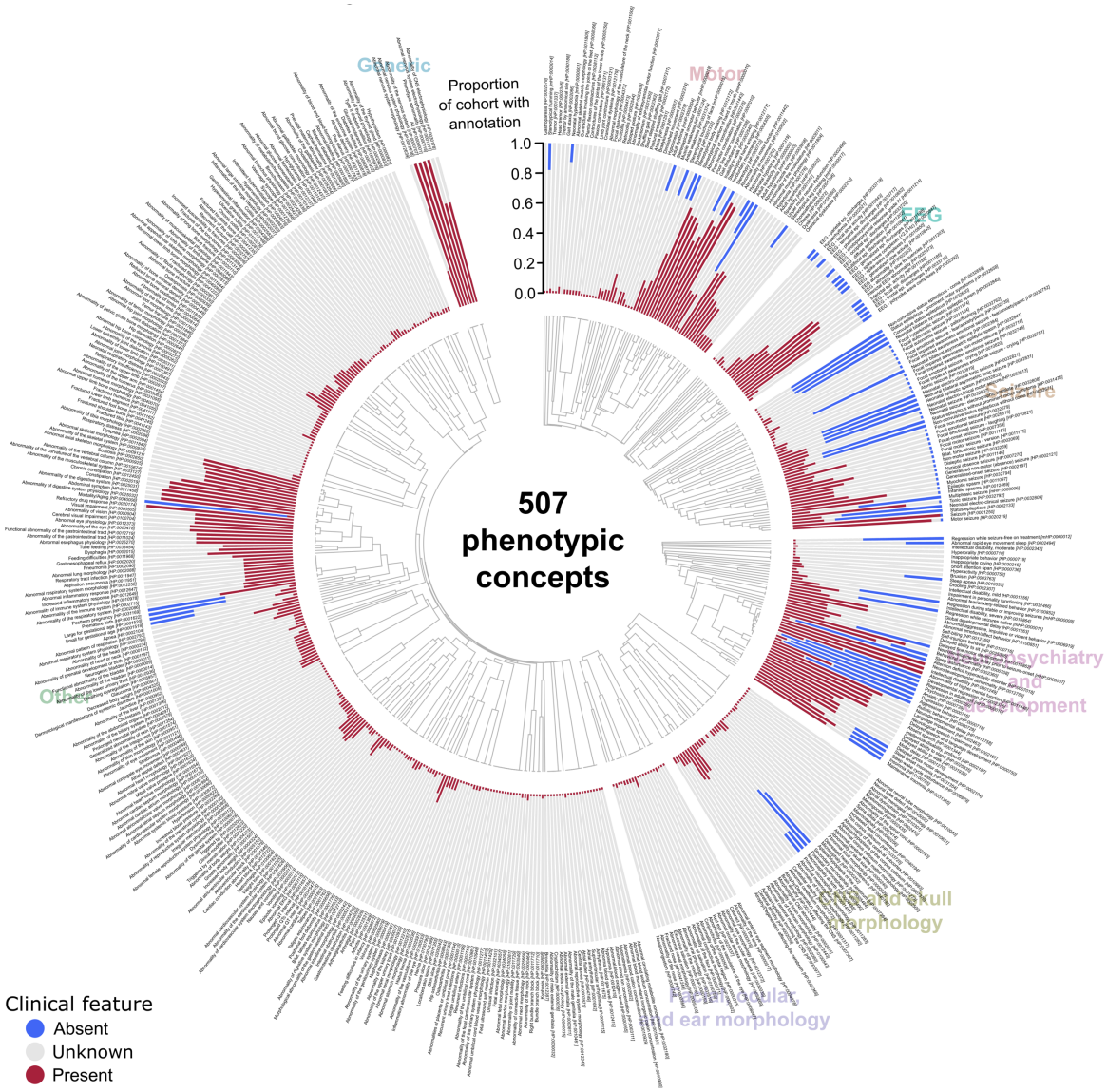


Figure 7-4. The phenotypic landscape of the cohort, including the phenotyping gap. A high-resolution version to enable zooming is available in the supplementary files.

Association analysis need only consider informative annotations: those carrying information that is not redundant. In the case of those annotations that are annotated as present, these are uninformative if annotated ubiquitously. For example, there is no value in testing the association of *Intellectual disability [HP:0001249]* as this was annotated to all 67 people in the cohort. Regarding those that are negated, a phenotype must be negated in more people than at least one of its parent concepts to necessitate testing independently from its parents. Consequently, for association analysis I further pruned each person's annotations to include only those that are informative. This resulted in 5,425 positive annotations (median = 81, range = 22–145), representing 496 concepts that were present in at least one but not all members of the cohort and 951 negated annotations (median = 14, range = 6–22 per individual), with a

repertoire of 29 HPO concepts. Thus, the total number of phenotypic annotations used for the association analysis was 6,376 (median = 96, range = 28–160 per individual), comprising 525 concepts representing the presence or absence of 500 phenotypes. The associations of those negated phenotypes that were not tested directly can be inferred from the association results of their parents (**Supplementary Table 7-3**).

I excluded one uninformative annotation from the minimal set of a single person (*Intellectual disability [HP:0001249]* in CDKL5_51) for phenotypic similarity analysis by semantic similarity measures, because the comparison of uninformative annotations precludes $\mathrm{Sim}_{\mathrm{Rel}}$ (relevance, Schlicker *et al.*, 2006) and $\mathrm{Sim}_{\mathrm{Lin}}$ (Lin, 1998) similarity measures (**Equation 4-14** and **Equation 4-16**).

7.5.3 Seizures

All but one individual had seizures (99 %, hereafter, percentages in the text to the entire cohort unless an alternative denominator is specified). Four people (6.0 %) had seizures in the first week of life. Thirteen (19 %) had been documented as having had, and 45 (67 %) as having not had seizures in the neonatal period (phenotyping gap 13 %).

Most people had seizures beginning with motor manifestations (n = 62, 93 %), usually tonic seizures (irrespective of electrographic onset, n = 51, 76 %) or epileptic spasms (n = 43, 64 %). The latter usually started in infancy (n = 37/43, 86 %). Among the 13 people with neonatal seizures, these were predominantly epileptic spasms (n = 7/13, 54 %) and tonic seizures (n = 5/13, 38 %). Myoclonic (n = 26, 39 %) and non-motor (n = 30, 45 %) seizures were common. The latter comprised atypical absences in 20/30 (67 %) and focal non-motor seizures in 15/30 (50 %). Five of 30 (17 %) had been diagnosed with both. Broadly, seizures could be classified as focal-onset in 28 (42 %) and generalized-onset in 20 (30 %), with eight (12 %) having both. In 26 people (39 %) no electroclinical onset was documented. Bilateral tonic-clonic seizures (regardless of onset) were reported in 21 (31 %) and seizures with a dialeptic component (incorporating absence and focal impaired awareness seizures) in 25 (37 %). Status epilepticus had occurred in 14 (21 %), more commonly specified without (n = 11, 16 %) rather than with (n = 1, 1.5 %) prominent motor symptoms. Forty-eight (72 %) had not had recognized status epilepticus. Multiphasic seizures with evolving motor components have been reported in CDD and were documented in eight (12 %) and described in six, typically involving a tonic phase (n = 5/6) that evolved into spasms (n = 4/5).

Onset of seizures was before 2 months of age in half, and before 11 months in all 38 with complete ages of onset for all documented seizure types (**Figure 7-5**). Repeating the analysis in all 66 with seizures, censoring seizure types with missing age of onset gave an upper limit for the median age of seizure onset in this cohort within the fourth month (cumulative risk of 62 % by 4 months old). Fewer than 20 % were yet to manifest a seizure at 9 months old.

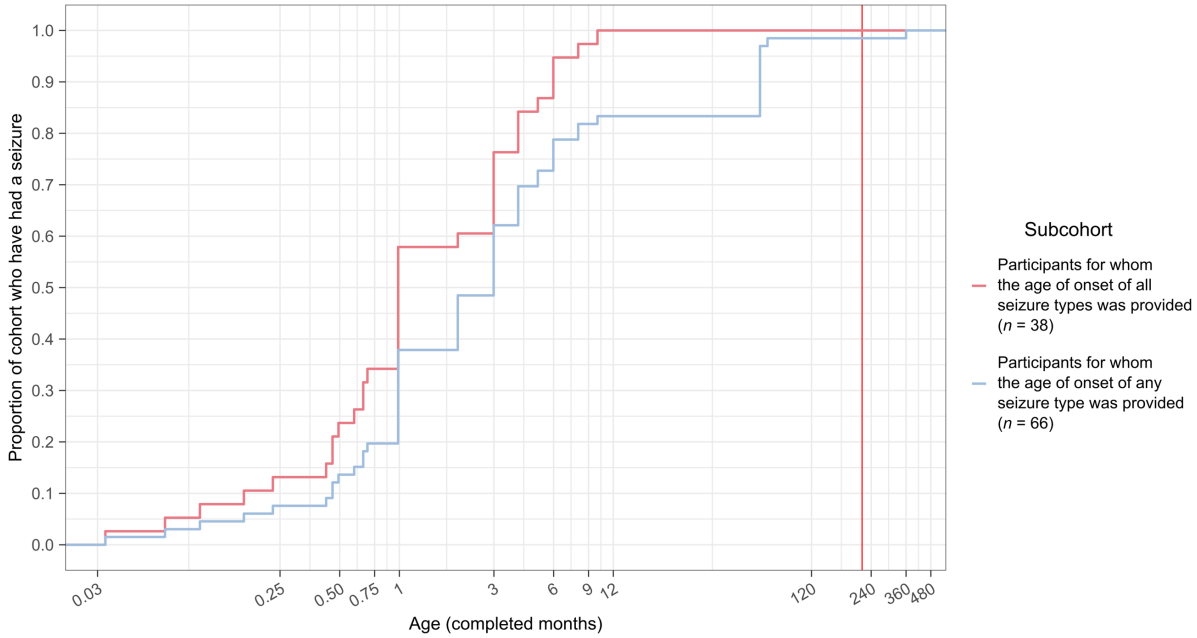


Figure 7-5. The distribution of age of onset of epileptic seizures of any type. The vertical red line indicates 18 years of age, beyond this age, follow-up is incomplete.

So far, five of 66 people with epilepsy (7.6 %) had manifested new types of seizure in adulthood, aged 18–23 years (**Figure 7-6**). These included myoclonic, focal impaired awareness emotional, and atypical absence seizures. To evaluate the chronological pattern of seizure types in CDD, I summarized the age of onset of each seizure type in the people with sufficient data (**Figure 7-7A**). After tonic seizures and epileptic spasms in infancy, tonic-clonic, focal-onset, myoclonic, and non-motor seizures began in early childhood, with seizures that have initial motor semiology tending to present earlier than non-motor seizures. Atonic seizures and status epilepticus, usually non-convulsive, presented in later childhood.

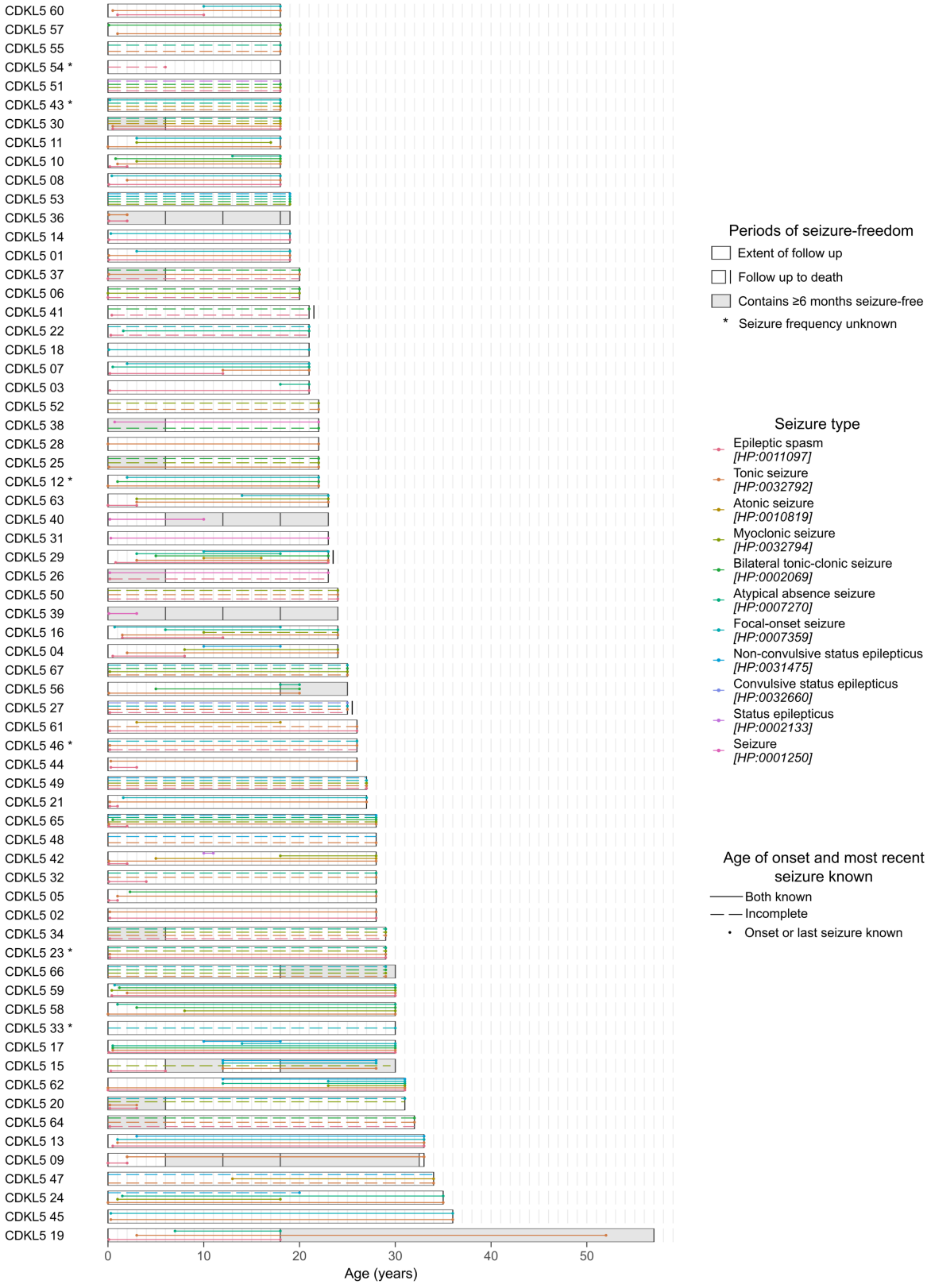


Figure 7-6. The evolution of seizures by participant with CDD. The minimal set of HPO phenotypes to capture the cohort's repertoire of seizures is shown with reduction in the specific descriptors of focal-onset seizures to enable visualization without further color dimensions.

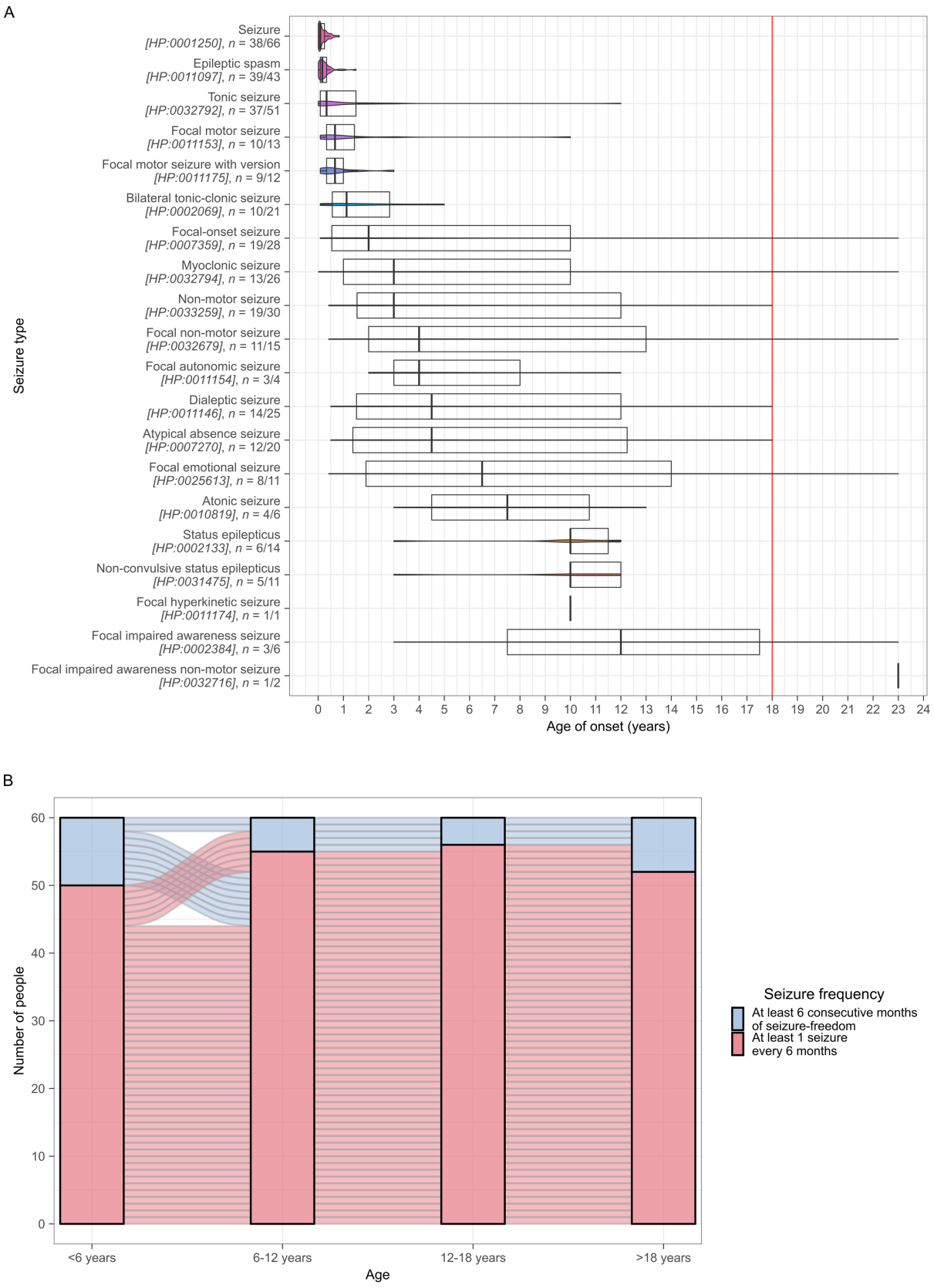


Figure 7-7. (A) The evolution of seizure types by onset, and (B) 6-year intervals containing 6 months of seizure freedom. In (A), the numerator of *n* indicates the number of people for whom the age of onset of the specified seizure type is known. The denominator of *n* indicates the total number of people with the seizure phenotype in the cohort, regardless of whether the chronology of their seizure phenotype is known, to give an indication of the extent of missing data. The vertical red line indicates 18 years of age, beyond this age follow-up is incomplete.

For 60 people with epilepsy, I was able to assess the occurrence of six consecutive months of seizure-freedom in each consecutive 6-year period from birth to adulthood (**Figure 7-7B**). Fewer than 20 % attained 6 months of seizure-freedom in any 6-year period, the majority (n = 44, 73 %) never did, and only two (3 %) attained it consistently.

Twenty-five people had been annotated with at least one seizure type complete with not only an age of onset but also the age of most recent occurrence. These limited data suggest that epileptic spasms entered long-term remission in most, but not all, by adolescence. Tonic, other motor seizures, focal-onset, and seizures with loss of awareness tended to persist (**Figure 7-8**).

At last data collection, of 62 adults who remained alive with a history of seizures (data missing for one), 35 (56 %) had seizures at least daily, a further ten (16 %) at least weekly, and nine (15 %) at least monthly. Eight (13 %) had been seizure-free for at least 6 months: 5/8 (62.5 %) continuing and 3/8 (37.5 %) having stopped ASM. Catamenial worsening of seizures was reported for 11/55 women (20 %), denied for two (3.6 %), and unknown in 42 (76 %).

A caregiver's or physician's subjective description of the evolution of childhood-onset seizures in adulthood was available for 63. Twenty-six (41 %) had an improvement, including 8 (13 %) who were seizure-free. Seizures remained stable in 31 (49 %) and worsened in seven (11 %, exact binomial test for 26 reporting improvement among 33 reporting a change in seizures, 79 %, two-sided p = 00132). The oldest individual (CDKL5_19, 57 years old) had periods of improvement and worsening (included in both figures). Periods of improvement or worsening were thought to have occurred in response to treatment changes in 13 and two, and spontaneously in eight and four, respectively. While the proportions favored changing treatment, this did not reach significance (13/8 versus 2/4, RR = 3.3, OR = 3.1, 95 % CI = 0.35–42, FET unadjusted two-sided p = 0.358).

7.5.4 Electroencephalography

Electroencephalographic abnormalities had been documented in 46 people (69 %). Most had generalized slow activity. Thirty-eight (57 %) had epileptiform discharges: usually focal (n = 35/38, 92 %), frequently multifocal (n = 15/35, 43 %). Generalized discharges were reported in 22/38 (58 %), most of whom also had focal epileptiform discharges on the same or another EEG (n = 19/22, 86 %). Hypsarrhythmia and a photoparoxysmal response were documented in only one individual each. Electroencephalographic reports in adulthood and those in childhood differed only in the provision of spatial localization of focal epileptiform discharges in adulthood (**Table 7-3**).

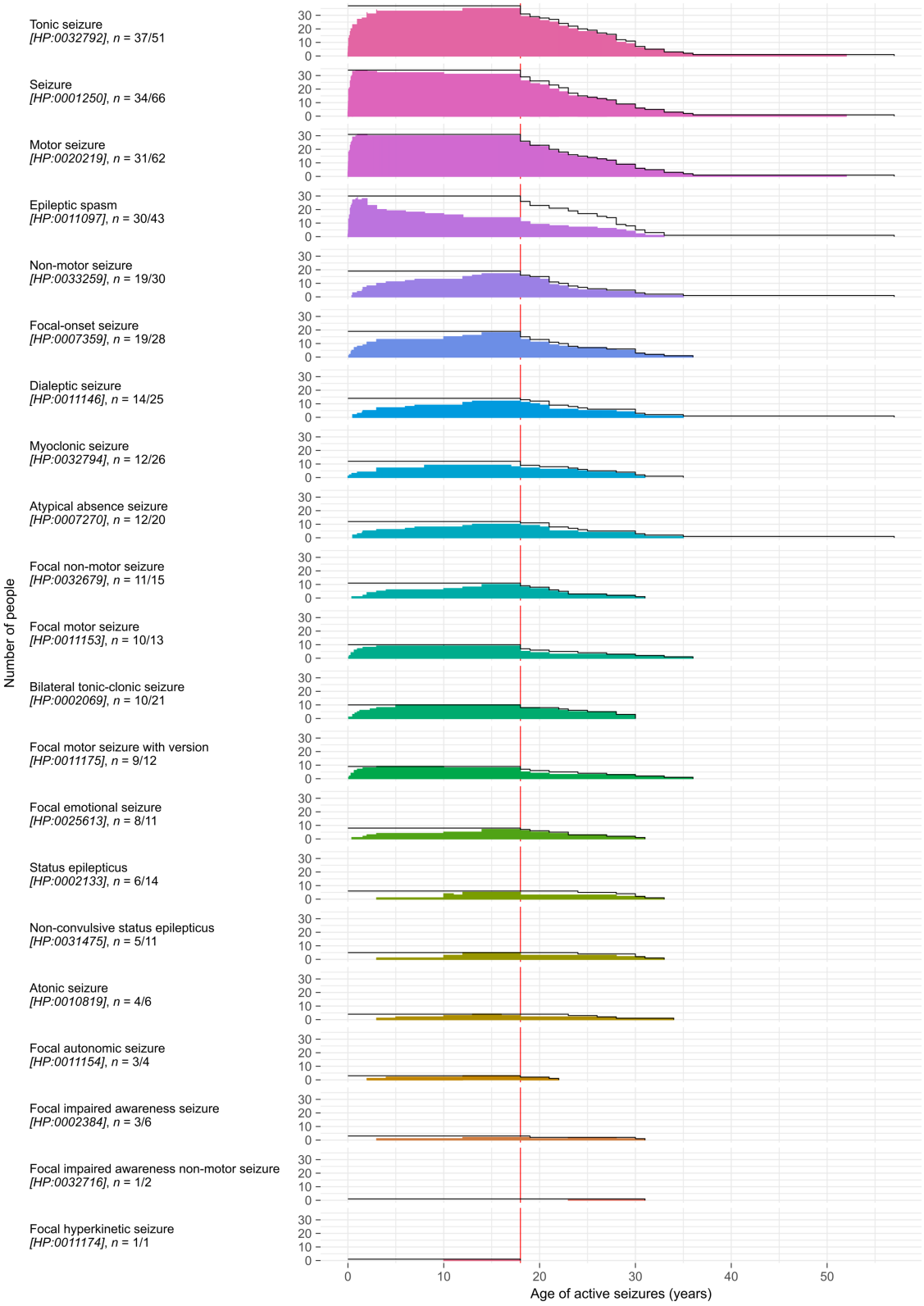


Figure 7-8. The evolution of seizures in the cohort by seizure type. Only people with a known age at onset and at the most recent occurrence of the seizure are included. The numerator of *n* indicates the number of people for whom the age of onset and at last occurrence of the seizure type is known. The stepped black line indicates how many of these people remain under follow up at the age indicated on the x-axis. The denominator of *n* indicates the total number of people with the seizure phenotype in the cohort, regardless of whether the chronology of the seizure phenotype is known for them, to give an indication of the extent of missing data. The vertical red line indicates 18 years of age, beyond this age follow-up is incomplete.

HPO term		Phenotype reported		Phenotype not reported		Odds ratio	Fisher exact test two-sided *p*-value
CUI	Term name	Child EEG	Adult EEG	Child EEG	Adult EEG		
HP:0010841	Multifocal epileptiform discharges	15	0	31	46	Infinite	1.56×10^{-5}
HP:0033720	EEG with occipital epileptiform discharges	9	0	37	46	Infinite	0.00254
HP:0033716	EEG with frontal epileptiform discharges	6	0	40	46	Infinite	0.0263
HP:0011202	EEG with diffuse acceleration	7	2	39	44	3.89	0.158
HP:0033717	EEG with temporal epileptiform discharges	3	0	43	46	Infinite	0.242
HP:0011198	EEG with generalized epileptiform discharges	15	11	31	35	1.53	0.488
HP:0002392	EEG with polyspike wave complexes	2	0	44	46	Infinite	0.495
HP:0010847	EEG with spike-wave complexes (< 2.5 Hz)	13	10	33	36	1.41	0.631
HP:0010850	EEG with spike-wave complexes	13	10	33	36	1.41	0.631
HP:0010845	EEG with generalized slow activity	30	28	16	18	1.20	0.829
HP:0011185	EEG with focal epileptiform discharges	21	23	25	23	0.842	0.835
HP:0010843	EEG with focal slow activity	2	3	44	43	0.655	1
HP:0011182	Interictal epileptiform activity	26	27	20	19	0.916	1
HP:0011203	EEG with abnormally slow frequencies	30	31	16	15	0.908	1
HP:0025373	Interictal EEG abnormality	26	27	20	19	0.916	1
HP:0033719	EEG with parietal epileptiform discharges	1	0	45	46	Infinite	1
HP:0002521	Hypsarrhythmia	0	1	46	45	0	1

Table 7-3. Comparison of electroencephalographic abnormalities in childhood and adulthood.

7.5.5 Antiseizure treatment

The 66 people with a history of seizures had been exposed to 36 different antiseizure treatments (**Table 7-4**). The median number of antiseizure treatments to which a participant had been exposed was seven (mean = 7.7, range = 1–18). Of the total 513 treatment exposures, 40 were considered to have benefitted (7.8 %) and 25 to have aggravated seizure control (4.9 %) by the treating clinician. At most recent follow-up, three of 66 (4.5 %) were seizure-free without continuing ASM. In the remaining 63, the median number of regular antiseizure treatments was three (mean = 2.95, range = 1–6), with four people (6.3 %) additionally prescribed a rescue medication.

The ASM to which participants had most frequently been exposed were lamotrigine (59 %), clobazam (58 %), sodium valproate (56 %), levetiracetam (55 %), and topiramate (52 %). The most commonly prescribed currently (among those 63 receiving ASM at last follow-up) were clobazam (40 %), sodium valproate (32 %), and lamotrigine (25 %). This was reflected in the most common combinations: 41 of 63 (65 %) were taking at least one, 17 (27 %) were taking at least two of these, and three (4.8 %) were taking all three.

After adjusting for multiple testing, insufficient evidence was found to demonstrate that any antiseizure treatment was reported to be particularly beneficial or detrimental to seizure control (seizure type not specified). However, those with the most promising by OR were ganaxolone, cenobamate, cannabidiol, adrenocorticotropin, and lacosamide. None was convincingly associated with seizure aggravation. Ketogenic diets had been attempted by 21 people (commenced in adulthood in two) but were considered beneficial in only one and thought to have aggravated seizures in another. None remained on this treatment at last follow-up. Vagus nerve stimulation (VNS) was started in 19 people (five in adulthood) and continued in ten to last follow-up, one of whom was seizure-free. It was considered beneficial in three and aggravating in none.

Fifty-three new courses of treatment were commenced in adulthood, involving 21 treatments in 22 adults, and 32 of these treatments were continued until last follow-up. The median number of new treatments commenced per individual among these 22 was 2 (range = 1–7). Sixteen (30 %) courses of treatment commenced in adulthood were considered beneficial while two (3.8 %, trials of topiramate and primidone in different people) were considered detrimental to seizure control (**Table 7-5**).

	Number of people							Considered beneficial				Considered aggravating	
Antiseizure treatment	Exposed	Current regular treatment	Previous regular treatment	Current rescue therapy	Commenced in adulthood	Current regular treatment while seizures active	Current regular treatment while seizures in remission	Number of people	Odds ratio (95 % CI)	Unadjusted FET two-sided *p*-value	Number of people	Odds ratio (95 % CI)	Unadjusted FET two-sided *p*-value
Ganaxolone	2	2	0	0	1	2	0	2	Infinite (2.25-Infinite)	0.00594	0	0 (0-106)	1.00
Cannabidiol	9	8	1	0	1	8	0	3	6.27 (0.975-30.8)	0.0266	0	0 (0-10.3)	1.00
Lacosamide	17	9	8	0	4	9	0	4	3.91 (0.884-13.6)	0.0359	0	0 (0-4.89)	1.00
ACTH	12	0	11	0	2	0	0	3	4.16 (0.695-17.6)	0.059	0	0 (0-7.31)	1.00
Cenobamate	1	1	0	0	1	0	1	1	Infinite (0.303-Infinite)	0.078	0	0 (0-752)	1.00
Phenobarbital	32	11	21	0	2	10	1	5	2.35 (0.667-6.75)	0.0932	1	0.615 (0.0145-4.04)	1.00
Perampanel	17	9	8	0	7	9	0	3	2.65 (0.468-10.1)	0.139	1	1.23 (0.0282-8.59)	0.578
VNS	19	10	8	0	5	9	1	3	2.31 (0.413-8.62)	0.178	0	0 (0-4.31)	0.616
Clonazepam	26	10	16	0	2	8	2	0	0 (0-1.79)	0.250	3	2.75 (0.492-10.2)	0.126
Levetiracetam	36	12	24	0	2	12	0	1	0.321 (0.00771-2.03)	0.345	3	1.88 (0.342-6.75)	0.407
Phenytoin	18	2	16	0	0	2	0	0	0 (0-2.71)	0.384	2	2.56 (0.27-12)	0.217
Topiramate	34	12	22	0	3	12	0	1	0.342 (0.0082-2.17)	0.503	2	1.24 (0.136-5.4)	0.678
Clobazam	38	25	13	0	5	23	2	4	1.43 (0.35-4.36)	0.524	3	1.76 (0.322-6.31)	0.419
Oxcarbazepine	10	4	6	0	0	4	0	1	1.32 (0.0294-9.98)	0.559	0	0 (0-9.07)	1.00
Diazepam	12	4	6	4	0	4	0	0	0 (0-4.33)	0.612	0	0 (0-7.31)	1.00
Ethosuximide	12	0	12	0	0	0	0	0	0 (0-4.33)	0.612	0	0 (0-7.31)	1.00
Zonisamide	15	6	9	0	1	6	0	0	0 (0-3.34)	0.620	0	0 (0-5.64)	1.00
Vigabatrin	26	3	23	0	0	3	0	1	0.46 (0.0109-2.97)	0.711	2	1.68 (0.182-7.49)	0.366
Lamotrigine	39	16	23	0	3	14	2	2	0.621 (0.0698-2.57)	0.758	3	1.71 (0.313-6.11)	0.426
Valproate	37	20	17	0	6	16	4	2	0.659 (0.074-2.74)	0.758	1	0.524 (0.0124-3.41)	1.00
Carbamazepine	26	5	21	0	2	4	1	2	0.985 (0.109-4.23)	1.00	0	0 (0-3.03)	0.630
Ketogenic diet	21	0	21	0	2	0	0	1	0.581 (0.0137-3.84)	1.00	1	0.975 (0.0226-6.64)	1.00
Rufinamide	16	7	9	0	1	6	1	1	0.783 (0.0181-5.36)	1.00	1	1.31 (0.03-9.26)	0.556
Primidone	6	1	5	0	1	1	0	0	0 (0-10.3)	1.00	1	4 (0.0817-37.9)	0.260
Brivaracetam	5	3	2	0	0	3	0	0	0 (0-13.2)	1.00	0	0 (0-22.1)	1.00
Gabapentin	5	1	4	0	0	1	0	0	0 (0-13.2)	1.00	0	0 (0-22.1)	1.00
Nitrazepam	5	1	4	0	0	1	0	0	0 (0-13.2)	1.00	0	0 (0-22.1)	1.00
Felbamate	5	0	5	0	1	0	0	0	0 (0-13.2)	1.00	0	0 (0-22.1)	1.00

Antiseizure treatment	Number of people							Considered beneficial			Considered aggravating		
	Exposed	Current regular treatment	Previous regular treatment	Current rescue therapy	Commenced in adulthood	Current regular treatment while seizures active	Current regular treatment while seizures in remission	Number of people	Odds ratio (95 % CI)	Unadjusted FET two-sided *p*-value	Number of people	Odds ratio (95 % CI)	Unadjusted FET two-sided *p*-value
Prednisolone	3	0	3	0	0	0	0	0	0 (0-29.1)	1.00	0	0 (0-48.5)	1.00
Chlorazepate	2	2	0	0	0	2	0	0	0 (0-63.5)	1.00	0	0 (0-106)	1.00
Acetazolamide	2	0	2	0	0	0	0	0	0 (0-63.5)	1.00	1	19.9 (0.248-1570)	0.0952
Stiripentol	1	1	0	0	0	1	0	0	0 (0-458)	1.00	0	0 (0-752)	1.00
Soticlestat	1	1	0	0	0	1	0	0	0 (0-458)	1.00	0	0 (0-752)	1.00
Bromide	1	0	1	0	0	0	0	0	0 (0-458)	1.00	0	0 (0-752)	1.00
Fenfluramine	1	0	1	0	1	0	0	0	0 (0-458)	1.00	0	0 (0-752)	1.00
Sulthiame	1	0	1	0	0	0	0	0	0 (0-458)	1.00	0	0 (0-752)	1.00

Table 7-4. Antiseizure treatment across all ages. ACTH, adrenocorticotropin; VNS, vagus nerve stimulator. Commenced on the previous page.

Antiseizure treatment	Number of people					Considered beneficial			Considered aggravating		
	Commenced in adulthood	Current regular treatment	Previous regular treatment	Current regular treatment while seizures currently active	Current regular treatment while seizures in remission	Number of people	Odds ratio (95 % CI)	Unadjusted FET two-sided *p*-value	Number of people	Odds ratio (95 % CI)	Unadjusted FET two-sided *p*-value
Perampanel	7	5	2	9	0	2	0.916 (0.0782-6.47)	1.00	0	0 (0-36.8)	1.00
Valproate	6	5	1	16	4	2	1.17 (0.0959-9.34)	1.00	0	0 (0-44.4)	1.00
Clobazam	5	4	1	23	2	2	1.6 (0.122-15.6)	0.632	0	0 (0-55.4)	1.00
VNS	5	3	1	9	1	1	0.556 (0.0105-6.27)	1.00	0	0 (0-55.4)	1.00
Lacosamide	4	3	1	9	0	2	2.45 (0.163-36.9)	0.575	0	0 (0-72.5)	1.00
Lamotrigine	3	2	1	14	2	0	0 (0-5.65)	0.545	0	0 (0-104)	1.00
Topiramate	3	2	1	12	0	0	0 (0-5.65)	0.545	1	20.2 (0.209-1920)	0.111
Carbamazepine	2	2	0	4	1	1	2.36 (0.0287-193)	0.517	0	0 (0-179)	1.00
Levetiracetam	2	2	0	12	0	0	0 (0-12.5)	1.00	0	0 (0-179)	1.00
Phenobarbital	2	1	1	10	1	1	2.36 (0.0287-193)	0.517	0	0 (0-179)	1.00
Cannabidiol	1	1	0	8	0	1	Infinite (0.0593-Infinite)	0.302	0	0 (0-979)	1.00
Cenobamate	1	1	0	0	1	1	Infinite (0.0593-Infinite)	0.302	0	0 (0-979)	1.00
Ganaxolone	1	1	0	2	0	1	Infinite (0.0593-Infinite)	0.302	0	0 (0-979)	1.00
Clonazepam	2	0	2	8	2	0	0 (0-12.5)	1.00	0	0 (0-179)	1.00
Ketogenic diet	2	0	2	0	0	0	0 (0-12.5)	1.00	0	0 (0-179)	1.00
ACTH	2	0	1	0	0	2	Infinite (0.446-Infinite)	0.0871	0	0 (0-179)	1.00
Felbamate	1	0	1	0	0	0	0 (0-90.1)	1.00	0	0 (0-979)	1.00
Fenfluramine	1	0	1	0	0	0	0 (0-90.1)	1.00	0	0 (0-979)	1.00
Primidone	1	0	1	1	0	0	0 (0-90.1)	1.00	1	Infinite (0.654-Infinite)	0.0377
Rufinamide	1	0	1	6	1	0	0 (0-90.1)	1.00	0	0 (0-979)	1.00
Zonisamide	1	0	1	6	0	0	0 (0-90.1)	1.00	0	0 (0-979)	1.00

Table 7-5. Antiseizure treatment in adulthood. ACTH, adrenocorticotropin; VNS, vagus nerve stimulator.

7.5.6 Neurodevelopmental, psychiatric, and other neurological features

All had intellectual disability, usually profound or severe (*n* = 61, 91 %). Fifty (75 %) had not acquired speech. Neurodevelopmental delay was reported in 54 (81 %), noted prior to the onset of seizures in 18 of these (33 %), and usually specified as gross motor (*n* = 39/54, 72 %) or global (*n* = 18/54, 33 %). Fine motor delay was annotated to only a subset of those with gross motor delay (*n* = 20/39, 51 %). Developmental regression was documented thoroughly, occurring in 30 (45 %) and not in 36 (54 %). Regression occurred during periods of subjectively higher (*n* = 14/30, 47 %) and lower (*n* = 14/30, 47 %) seizure burden, including in one person when seizure-free. Twelve of 30 (40 %) lost skills in adulthood. The person without seizures had not regressed.

Autistic behavior was described in 15 (22 %). Attention deficit hyperactivity disorder was reported in three (4.5 %), improving during adulthood in all, and affective disorders in seven (10 %): five with anxiety and three with depression, including one with both. Psychosis was not reported. Self-injurious behaviors (all self-biting) were documented in 18 (27 %), and other aggressive behaviors in 28 (42 %). Aggressive behaviors began in adulthood in seven of 28 (25 %). In those 21 with aggressive behaviors beginning before adulthood, behavior improved in five (24 %) in adulthood, remained stable in eight (38 %), and worsened in eight (38 %).

Stereotypies were reported in 49 people (73 %), most frequently hand wringing (*n* = 42/49, 86 %), a tendency to insert their hand in their mouth (*n* = 29/49, 59 %), or body rocking (*n* = 15/49, 31 %). All 18 people with self-injurious behavior had stereotypies.

Sixty-two people (93 %) were documented as having either *Abnormality of movement [HP:0100022]* or *Abnormality of central motor function [HP:0011442]*. Both abnormalities of coordination (*n* = 52, 78 %) and involuntary movements including stereotypies (*n* = 51, 76 %) were common. Beyond stereotypies, specific hyperkinetic disorders occurred in 18 (27 %), including dystonia in 11 (16 %); or the chorea dyskinesia spectrum in nine (13 %); and tremor in two (3 %). Hypokinesia (*n* = 6, 9.0 %) and bruxism (*n* = 4, 6.0 %) were reported more rarely. Mobility problems were reported in 51 (76 %) and included gait disturbance (*n* = 47, 70 %), often described as shuffling (*n* = 5/47, 11 %) or ataxic (*n* = 3/47, 6.4 %), and postural instability (*n* = 13, 19 %). Thirty-nine never walked (58 % of the cohort, 76 % of those 51 with gait disturbance or postural instability).

Sixteen (24 %) had a parkinsonian feature (bradykinesia, hypokinesia, postural instability, retrocollis, or short-stepped shuffling gait), usually beginning in adolescence. I did not find an

association between greater age at last assessment and either the presence of a parkinsonian feature (WRST $W = 361.5$, one-sided $p = 0.756$) or the number of parkinsonian features (Kendall's $\tau = -0.051$, one-sided $p = 0.693$). Despite the absence of evidence of chronological causation, four people with parkinsonian features were exposed to valproic acid ($n = 2$) or antipsychotics ($n = 2$). One person had self-limited focal dystonia following exposure to metoclopramide during adulthood. Medications prescribed for movement disorders included trihexyphenidyl ($n = 2$), levodopa ($n = 1$), and amantadine ($n = 1$), without data on response. The evolution of movement disorders was described in 26. In 16/26 (62 %) these began in adulthood. Of the 10 with movement disorders beginning before, eight became more severe and two remained stable in adulthood.

Abnormal muscle tone at some age was reported in 32 (48 %). Neonatally, hypotonia was reported in 18 (27 %) and excluded in 8 (12 %). Neonatal hypertonia was not recorded. In adulthood, 21 (31 %) had hypertonia, usually spasticity with one treated with botulinum toxin and another with baclofen. Hypotonia was documented in two adults (3 %).

Sleep disorders were reported in 43 (64 %), with difficulty with sleep maintenance, sleep-wake circadian rhythm disturbance, and sleep onset insomnia common among these. Sleep apnea was reported in five of 43 (12 %). One person had abnormal rapid eye movement sleep, in addition to postural instability, stooped posture, and short stepped shuffling gait. Sleep disorders were adult-onset in ten of 43 (23 %). In those with earlier onset sleep disorders, these improved in seven (21 %), worsened in two (6.1 %), or remained stable in 24 (73 %) in adulthood. Medications for sleep symptoms included melatonin ($n = 8$), trazodone ($n = 5$), benzodiazepines ($n = 3$), clonidine ($n = 1$), and zopiclone ($n = 1$), as well as antidepressants and antipsychotics for which sleep may have contributed to the indication.

Visual impairments, always interpreted as cerebral, were recorded in 44 (66 %).

7.5.7 Neurodevelopmental evolution

We determined the total maximum CDS (defined as the sum of a person's peak abilities across the seven CDS milestones regardless of whether these skills were present concurrently) and the most recent total CDS, those skills present at the time of last assessment in adulthood (**Table 7-2** and **Figure 7-9**).

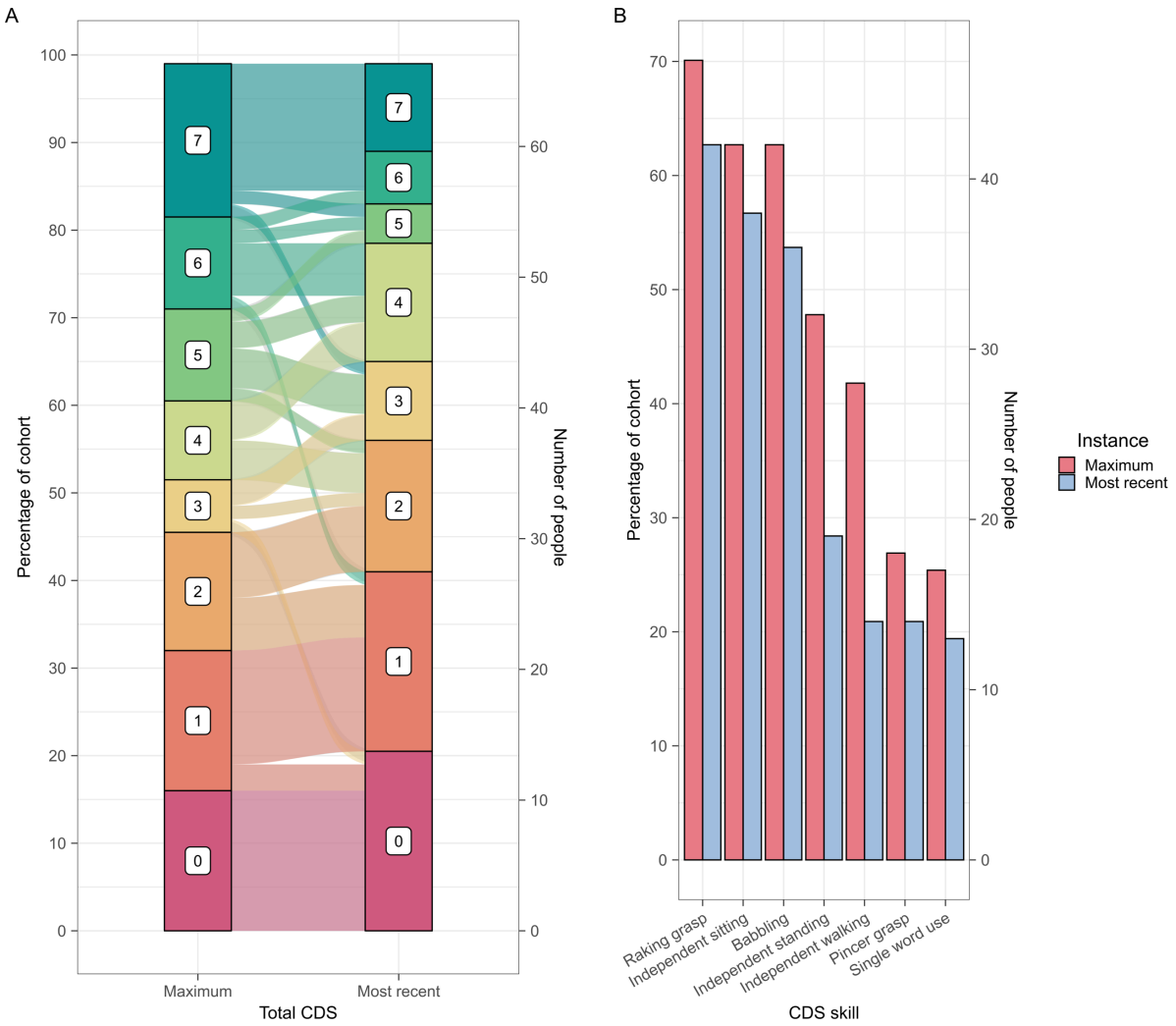


Figure 7-9. (A) The distribution of total maximum CDS across the entire cohort and how this compares to the most recent total CDS per person and (B) the percentage of the cohort attaining and retaining each constituent skill.

The median age at which children were able to sit was 12 months, to walk was 30 months, and to use a word was 19 months (**Figure 7-10**). The median age at which regression was reported to have begun was 48 months among those 18 demonstrating regression prior to adulthood and 132 months including those with loss of skills during adulthood (**Table 7-2** and **Table 7-6**).

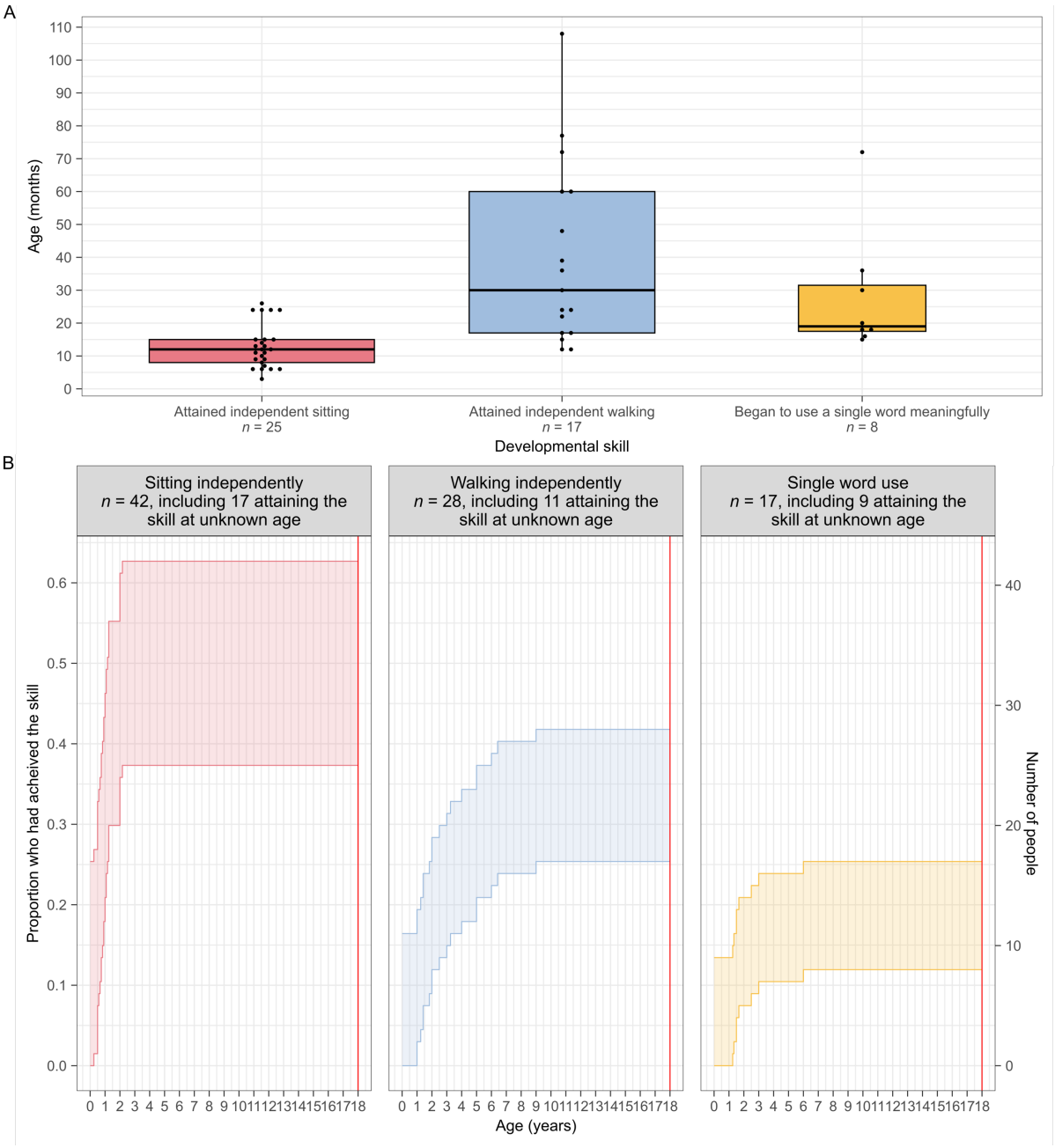


Figure 7-10. The age of attainment of developmental milestones. In (A) and subsequent boxplots, the plots summarize the raw data (black points), showing the median (central line), lower and upper quartiles (hinges), and minimum of 1.5-fold the interquartile range and full range (whiskers). Box width is proportional to the square-root of the number of people in the group. In (B), the vertical red lines in indicate 18 years of age, beyond this age follow-up is incomplete.

Total maximum CDS	Reaching total maximum CDS		Maintained maximum ability across all CDS skills to most recent review		Loss of at least one CDS skill	
	Number of people	Percentage of cohort	Number of people	Percentage of those with total maximum CDS	Number of people	Percentage of those with total maximum CDS
0	11	16 %	11	100 %	0	0 %
1	11	16 %	9	82 %	2	18 %
2	9	13 %	5	56 %	4	44 %
3	4	6 %	2	50 %	2	50 %
4	6	9 %	3	50 %	3	50 %
5	7	10 %	1	14 %	6	86 %
6	7	10 %	1	14 %	6	86 %
7	12	18 %	7	58 %	5	42 %

Table 7-6. The evolution of CDS into adulthood.

Eleven people (16 %) never reached any CDS milestone, most attained at least three, and 19 (28 %) attained at least six, of whom 12 (18 % of 67) attained all seven skills (**Figure 7-9A** and **Table 7-6**). Under half were ever able to stand unaided, but most of those who could also attained independent walking (n = 28/32, 88 %). Fewer than half of those able to babble were ever able to use one word (n = 17/42, 40 %). Similarly, nearly half of those who could use a rake grasp were ever able to use a pincer grasp (n = 18/47, 38 %, (**Figure 7-9B** and **Table 7-2)**. At last assessment as an adult, the median total CDS was 2, with 14 people (21 %) not being able to perform any of the skills, 4 (6.0 %) performing six, and seven (10 %) performing all seven. Standing and walking independently were the skills most frequently lost. Half of people who had attained any CDS skill retained all their skills into adulthood (n = 28/56). The number of skills lost increased with the number of skills attained, except in those attaining all seven.

I hypothesized that lower maximum CDS may be associated with early-onset seizures, as a marker of more severe underlying neurodevelopmental encephalopathy, or through the consequences of an early epileptic encephalopathy. The data supported this (WRST W = 430.5, p = 0.00481, effect size = 0.341, **Figure 7-11** and **Supplementary Table 7-4**). The difference persisted with a closed-world assumption that those nine not annotated with *Neonatal electro-clinical seizure [HP:0032809]* had not had neonatal seizures (medians unchanged, WRST W = 511.5, one-sided p = 0.00520, effect size = 0.314) and using the most recent total CDS (median = 2 versus 0, WRST W = 442.5, one-sided p = 0.00236, effect size = 0.372).

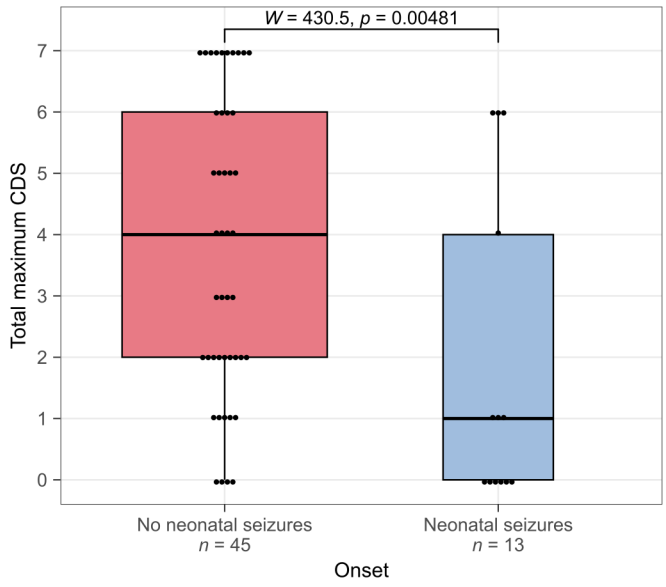


Figure 7-11. The association between the total number of CDS skills attained and the presence or absence of neonatal onset seizures.

After finding an association between a positive history of neonatal seizures (onset in the first month of life) and total maximum CDS, I explored whether the relationship between the age of epilepsy onset and total maximum CDS extended beyond this early threshold. Among those 38 people with a known age of seizure-onset, I did not find a correlation between the age of seizure onset and total maximum CDS (Kendall's $\tau = 0.0652$, one-sided $p = 0.299$, **Figure 7-12**).

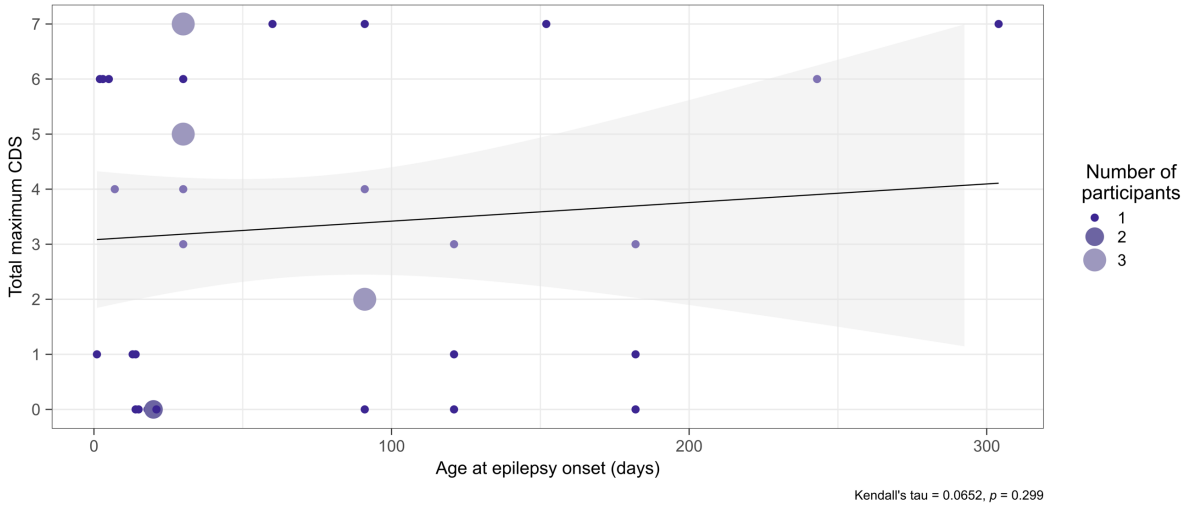


Figure 7-12. The relationship between age of epilepsy onset and total maximum CDS.

I also investigated the effect of moving the seizure-onset threshold from 1 to 2 months of age. When including only the 38 people with known age of onset, the median total maximum CDS of those 22 with seizure-onset aged < 2 months was 4 versus 2–3 for the remaining 16 (WRST

W = 180, one-sided p = 0.458), a non-significant trend in the opposite direction to that expected (**Figure 7-13A**). Considering all 67 people, the median total maximum CDS of those 25 with seizure-onset aged < 2 months was 4 versus 2–3 for the remaining 42, demonstrating a similar non-significant trend (WRST W = 549.5, one-sided p = 0.377, **Figure 7-13B**).

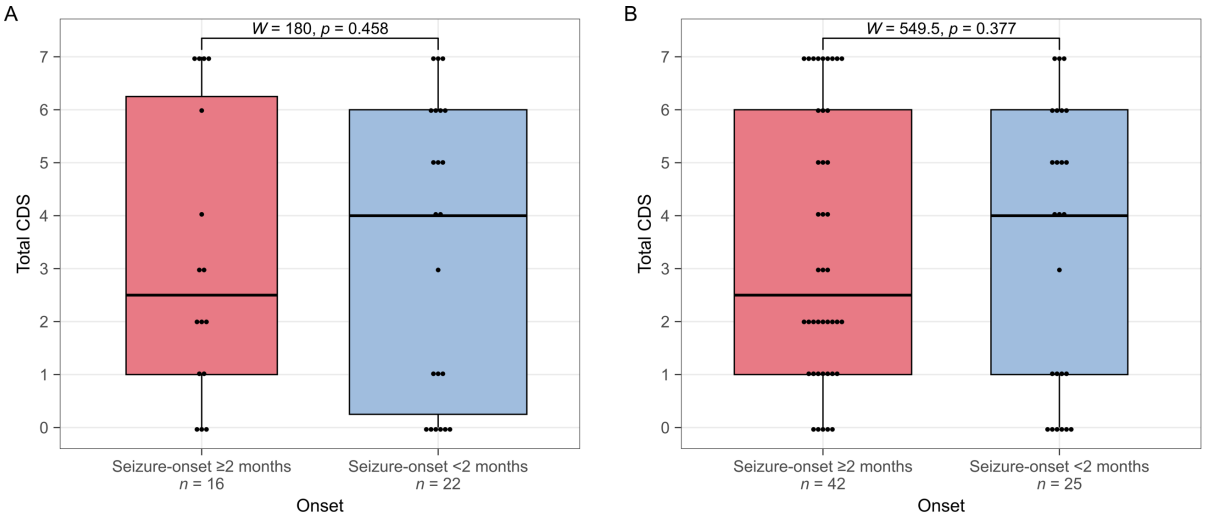


Figure 7-13. Total maximum CDS according to the presence or absence of seizures in first 2 months of life. (A) Considering only those with known age of seizure onset for all seizure types. (B) Considering the entire cohort by earliest known age of any seizure type.

There was no significant evidence of a relationship between either total maximum CDS (**Figure 7-14**) or the most recent total CDS (**Figure 7-15**) and age at last data collection. I also found no relationship between age at last assessment and the loss of attained CDS skills by logistic regression (*stats::glm(family = binomial(link = 'logit'))*) or WRST, focusing on only people with total maximum CDS > 0 (**Table 7-7**).

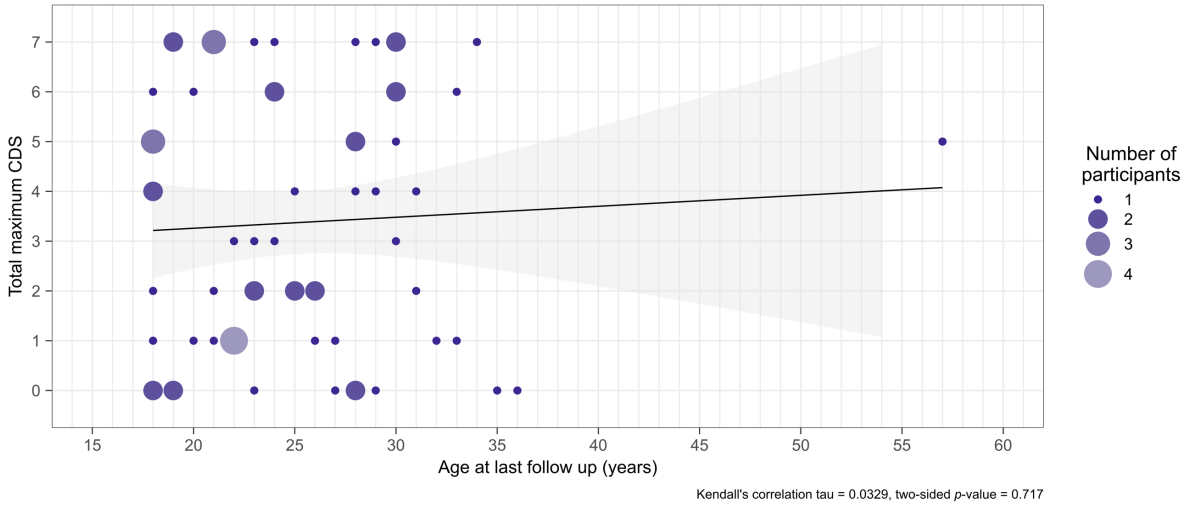


Figure 7-14. The relationship between age at last collection and total maximum CDS.

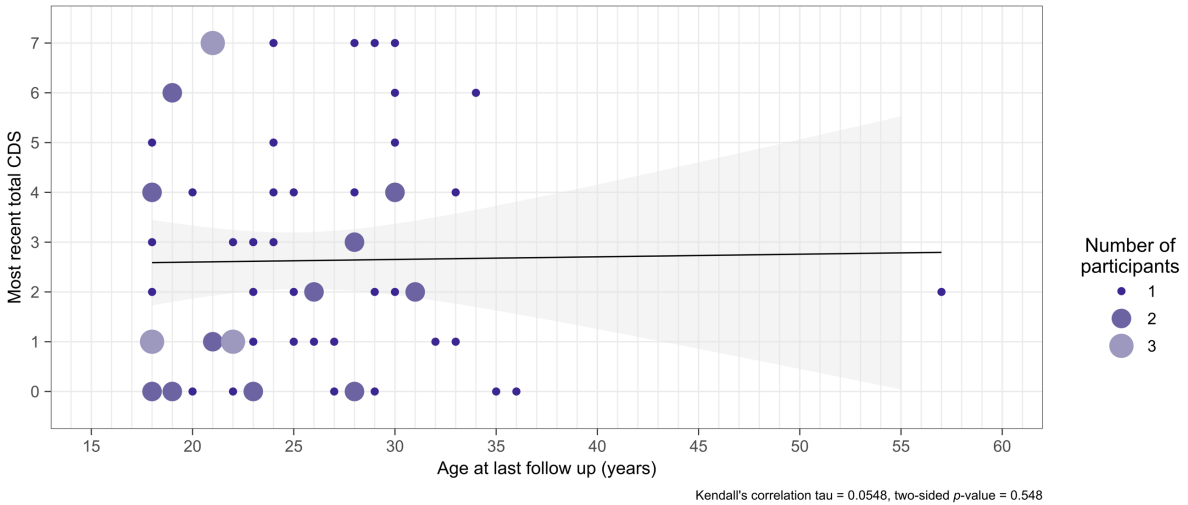


Figure 7-15. The relationship between age at last collection and total recent CDS.

Skill	Number of people			Median age at last follow-up (years)		Logistic regression				WRST	
	Acquiring the skill	Retaining the skill	Losing the skill having acquired it	Among those retaining the skill	Among those who had lost the skill	Coefficient	Standard error	z-value	Two-sided *p*-value	W	Two-sided *p*-value
Fall in the total maximum CDS for those acquiring any CDS skill	56	28	28	24.5	23.5	0.01811	0.0431	0.420	0.674	378.5	0.831
Walking independently	28	14	14	26.5	23.5	-0.009404	0.0487	-0.193	0.847	79.5	0.405
Standing independently	32	19	13	24	24	-0.04	0.04983	-0.803	0.422	128	0.877
Sitting independently	42	38	4	24	27.5	-0.035	0.06298	-0.556	0.578	103.5	0.245
Using a word	17	13	4	24	26.5	0.02225	0.121	0.184	0.854	24.5	0.909
Babbling	42	36	6	24	26	-0.0105	0.0957	-0.110	0.913	110	0.957
Pincer grasp	18	14	4	22.5	26	-0.07805	0.06419	-1.22	0.224	31	0.790
Raking grasp	47	42	5	25.5	22	0.2311	0.1342	1.72	0.0852	49	0.0547

Table 7-7. The relationship between age and loss of skills.

7.5.8 Comorbidities

Gastroenterological abnormalities were common (*n* = 60, 90 %), with most having suffered constipation (*n* = 57, 85 %). Twenty-three (34 %) received long-term treatment with laxatives (mostly polyethylene glycols) or recurrent enemata. Difficulty feeding or dysphagia was common (*n* = 40, 59 %). Thirty-one (46 %) required gastrostomy feeding, usually before adolescence (*n* = 26, 39 %). Five of 41 (12 %) not requiring a gastrostomy by adolescence had the procedure subsequently. Thirty (45 %) had gastroesophageal reflux (*n* = 27, 40 %) or episodic vomiting (*n* = 3, 4.5 %). Six were treated chronically with anti-reflux medications and four with antiemetics. Two people underwent Nissen fundoplication, and another required five colonoscopies and a laparotomy for recurrent volvulus, all three undergoing these procedures during adulthood. Three people received antimuscarinics to mitigate for difficulty swallowing saliva. A third of the cohort had experienced at least one episode of aspiration pneumonia.

Fourteen of 42 people with scoliosis (21 %, 32 % of 63 % of the cohort) underwent corrective surgery (seven during adulthood). Osteoporosis was diagnosed in eight (12 %), and fractures were reported in five (three fibular, one tibial, and one humeral), four of whom had osteoporosis. Hip dislocation (*n* = 3, 4.5 %) and contractures (*n* = 2, 3.0 %) were reported rarely.

Significant cardiovascular abnormalities were reported in eight people (12 %), each with a single anomaly. Five (7.5 %) had electrocardiographic abnormalities, all detected in adulthood: atrioventricular block (*n* = 2, 3.0 %), as well as atrial flutter, prolonged QTc, and right bundle branch block (each in one person). Syncope was reported in one individual (1.5 %) and was vasovagal. Systemic arterial hypertension was diagnosed in two (3.0 %) and atrial septal defect with mitral valve prolapse detected in one.

Hypothyroidism was diagnosed in two people (3.0 %), one of whom also had type 2 diabetes mellitus. Five (7.5 %) developed nephrolithiasis, all in adulthood following topiramate or ketogenic diet treatment, with one requiring surgical treatment.

7.5.9 Neuroimaging

Seventeen people (25 %) had a documented abnormality of central nervous system morphology. Three (4.5 %) had microcephaly (developing after birth) and two (3.0 %) had macrocephaly (since birth). Brain atrophy was detected in 14 (21 %), specified as cerebral in three (4.5 %) and cerebellar in another three (4.5 %). One person had spina bifida with a myelomeningocele and another had a thin corpus callosum.

7.5.10 Mortality

All three who died carried large deletions (one involving the whole *CDKL5* gene). The cause of death was known for two, both dying after hospitalization for infections with seizures worsened sufficiently to have been judged to contribute to death. There was no significant difference between the total maximum CDS of those three who were known to have died (median = 2) and those 64 still alive (median = 3) at last data collection (WRST $W = 92.5$, two-sided $p = 0.927$, **Figure 7-16**).

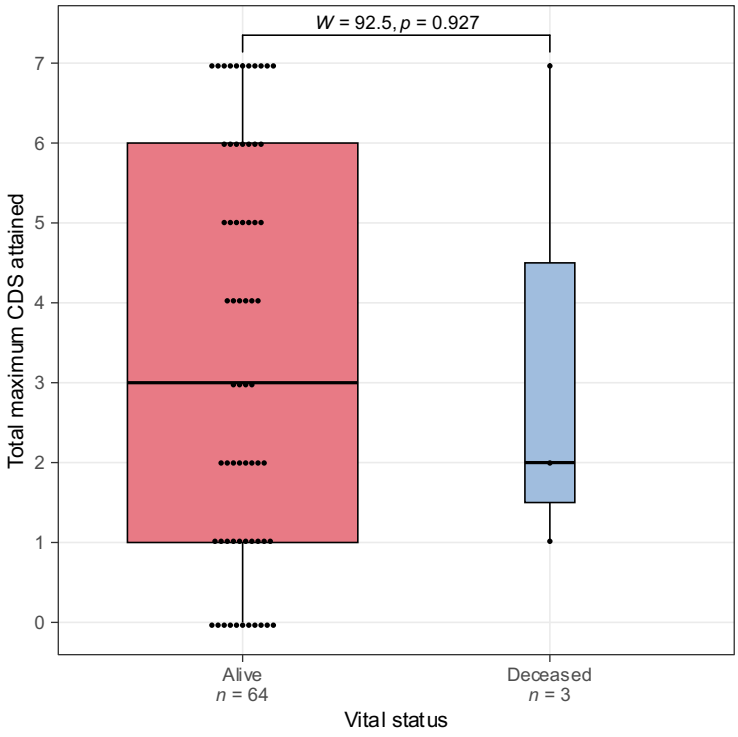


Figure 7-16. The total maximum CDS by vital status.

7.5.11 Molecular diagnosis

The year of molecular diagnosis was available for 55 people (82 %) with a median age at diagnosis of 16 years (range = 2–51 years) and a trend towards more recent diagnoses being at more advanced ages (**Figure 7-17**). Clinicians reported that molecular diagnosis had altered clinical management for 25 participants (37 %) and had not for 26 (39 %).

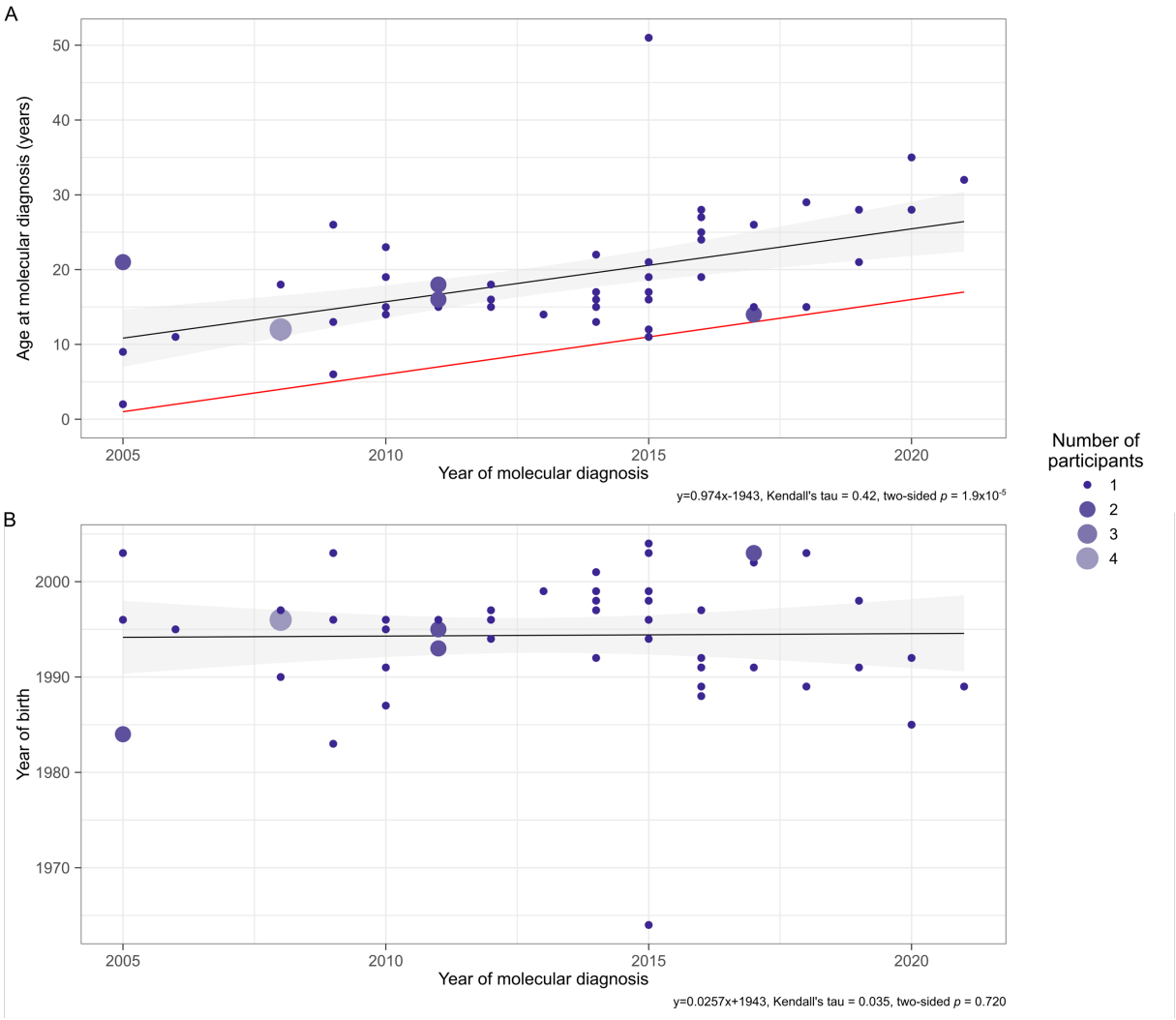


Figure 7-17. The relationship between age at diagnosis and year of diagnosis. (A) The red line with slope = 1 indicates the earliest possible age of diagnosis for this cohort. (B) The same data presented relative to year of birth.

I hypothesized that clinician perception of the benefit of obtaining a molecular diagnosis might be explained by seizure burden and the availability of precision medicines. I found those for whom molecular diagnosis was not felt to have altered management were more likely have had 6-month periods of seizure freedom (OR = 4.4, FET two-sided $p = 0.00604$). Molecular diagnoses perceived as beneficial to management tended to be more recent (median year = 2015 versus 2011, WRST $W = 368.5$, two-sided $p = 0.0225$). There was no difference in age at diagnosis between those in whom molecular diagnosis affected management and those in whom it had not (medians 16 years and 17 years respectively, WRST $W = 262.5$, two-sided $p = 0.974$).

7.5.12 Genotype-phenotype associations

I was primarily interested in any clinical distinctions between two broad groups: those 22 participants with missense variants and those 45 with other variants (**Figure 7-18**, **Figure 7-19**, and **Supplementary Table 7-5**). While no missense versus non-missense association with phenotype reached *minPsd* < 0.05, according to raw *p*-values *Regression in adulthood [mHP:0000010]* (OR = 5.7, $p = 0.0140$, *minPsd* = 0.634) and *Anxiety [HP:0000739]* (OR = 9.4, $p = 0.0368$, *minPsd* = 0.953) were more common among the 22 people with missense variants. Conversely, those 45 with non-missense variants were more frequently annotated with morphological (*Abnormal skeletal morphology [HP:0011842]*, OR = 3.63, $p = 0.0270$, *minPsd* = 0.814), motor (*Poor motor coordination [HP:0002275]*, OR = 5.27, $p = 0.00683$, *minPsd* = 0.357), gastrointestinal (*Gastroesophageal reflux [HP:0002020]*, OR = 7.03, $p = 0.00301$, *minPsd* = 0.167), infective (*Aspiration pneumonia [HP:0011951]*, OR = 4.53, $p = 0.0264$, *minPsd* = 0.814), and neurodevelopmental abnormalities (*Motor delay [HP:0001270]*, OR = 3.79, $p = 0.0173$, *minPsd* = 0.672). Among those associations falling short of $p < 0.05$ but of clinical interest were that people with missense variants might be more likely to have *no Neurodevelopmental delay prior to seizure-onset [mNP:0000007]* and *Atonic seizure [HP:0010819]*.

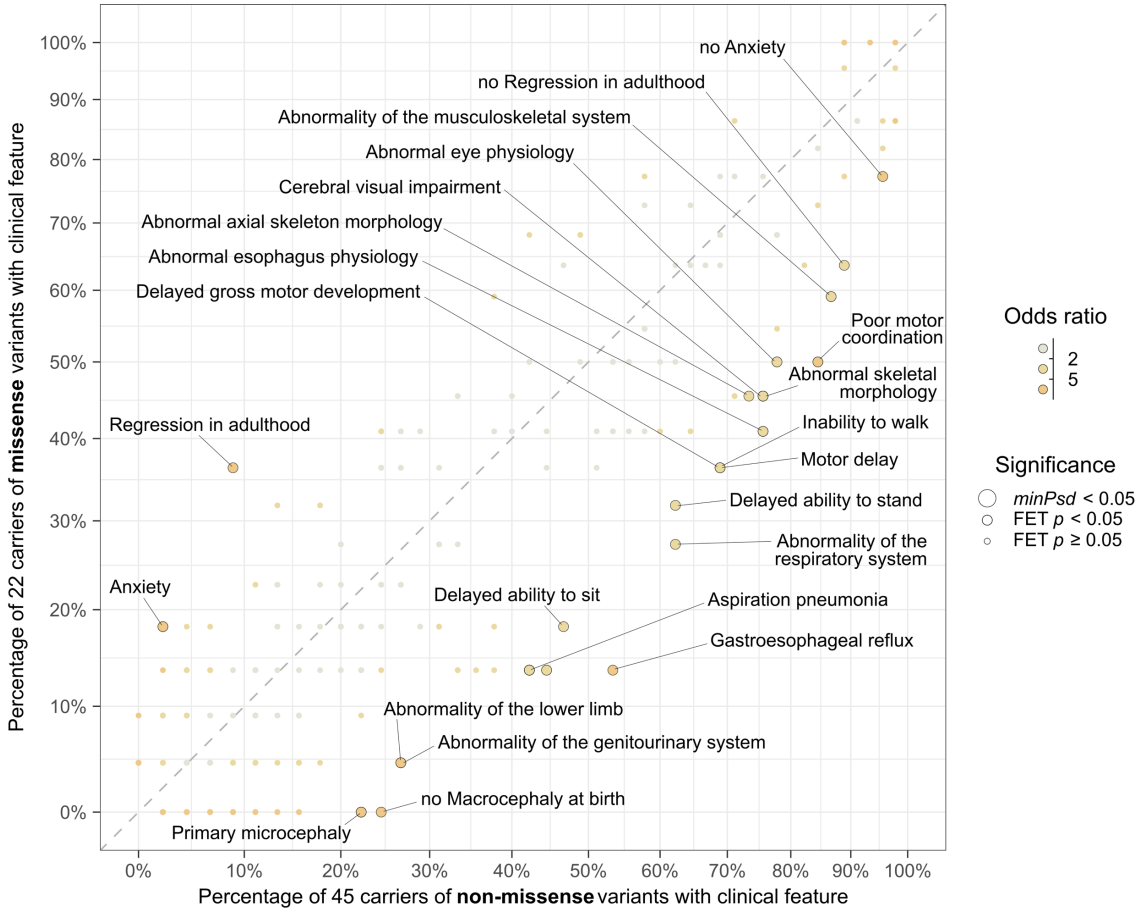


Figure 7-18. The phenotypic associations of carriers of missense *CDKL5* variants.

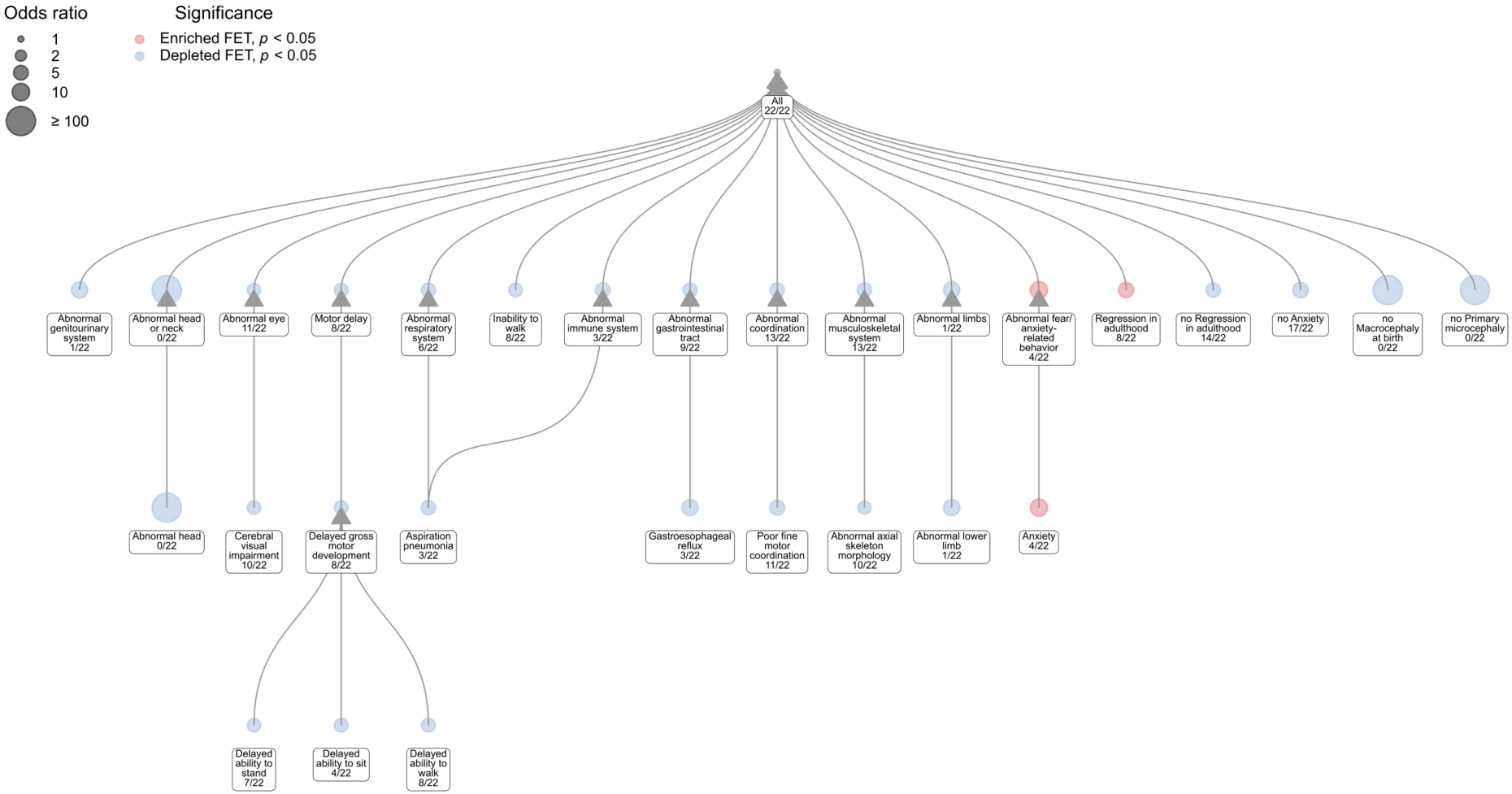


Figure 7-19. The phenotypic associations of carriers of missense *CDKL5* variants viewed conceptually. Associations for non-missense variants are identical with reversal of the direction (enrichment versus depletion).

People with missense variants tended to attain more CDS skills than those with non-missense variants (**Figure 7-20**, **Supplementary Table 7-4**). They may also be more likely to learn to use a pincer grasp, sit, stand, and walk (**Supplementary Table 7-6**).

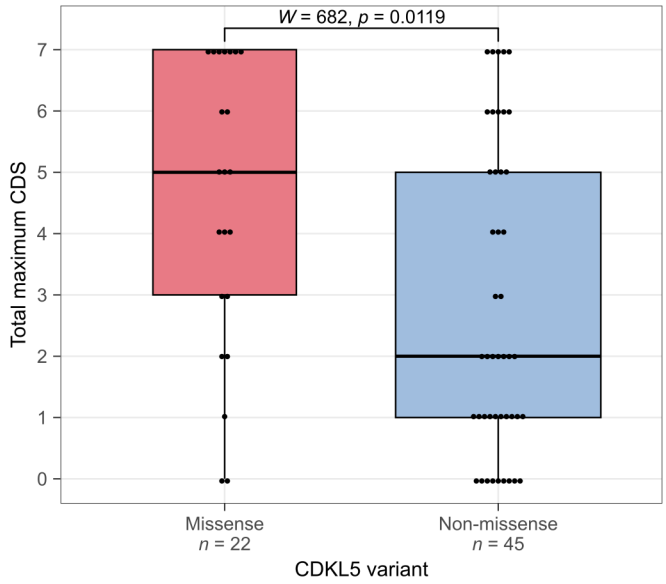


Figure 7-20. The association between the total number of CDS skills attained and the broad class of *CDKL5* variant.

I tested the associations of individual phenotypes with the membership of 61 genetically defined dependent groups, including specific genotypes, predicted variant consequences, variant locations, and the confirmation of de novo inheritance and mosaicism (**Supplementary Table 7-5**). The only phenotypic associations of genetic groups reaching *minPsd* < 0.05 were the increased odds of those 40 people with demonstrated de novo variants being annotated with electroencephalographic abnormalities, particularly *EEG with abnormally slow frequencies [HP:0011203]* (OR = 11.4, $p = 3.56 \times 10^{-5}$, *minPsd* = 0.00151, **Figure 7-21**). Despite all three people who had died being among those ten carrying large deletions, this reached significance only when not adjusting for multiple tests (surrogate term *Mortality/Aging [HP:0040006]*, OR = infinite, $p = 0.00250$, *minPsd* = 0.117, **Figure 7-22**). The two with deletion of *NHS* in addition to *CDKL5* had an enrichment of *EEG with diffuse acceleration [HP:0011202]* and *Multifocal epileptiform discharges [HP:0010841]*, but not of morphological features of Nance-Horan Syndrome. Other than those with stop gain variants appearing less likely to use a raking grasp (but showing no difference with regard to the more advanced pincer grasp), and the two carriers of c.587C>T reaching greater maximum total CDS, secondary analysis according to other genetic groups yielded no confident associations with phenotypes or CDS (**Supplementary Tables 7-4**, **7-5**, and **7-6**).

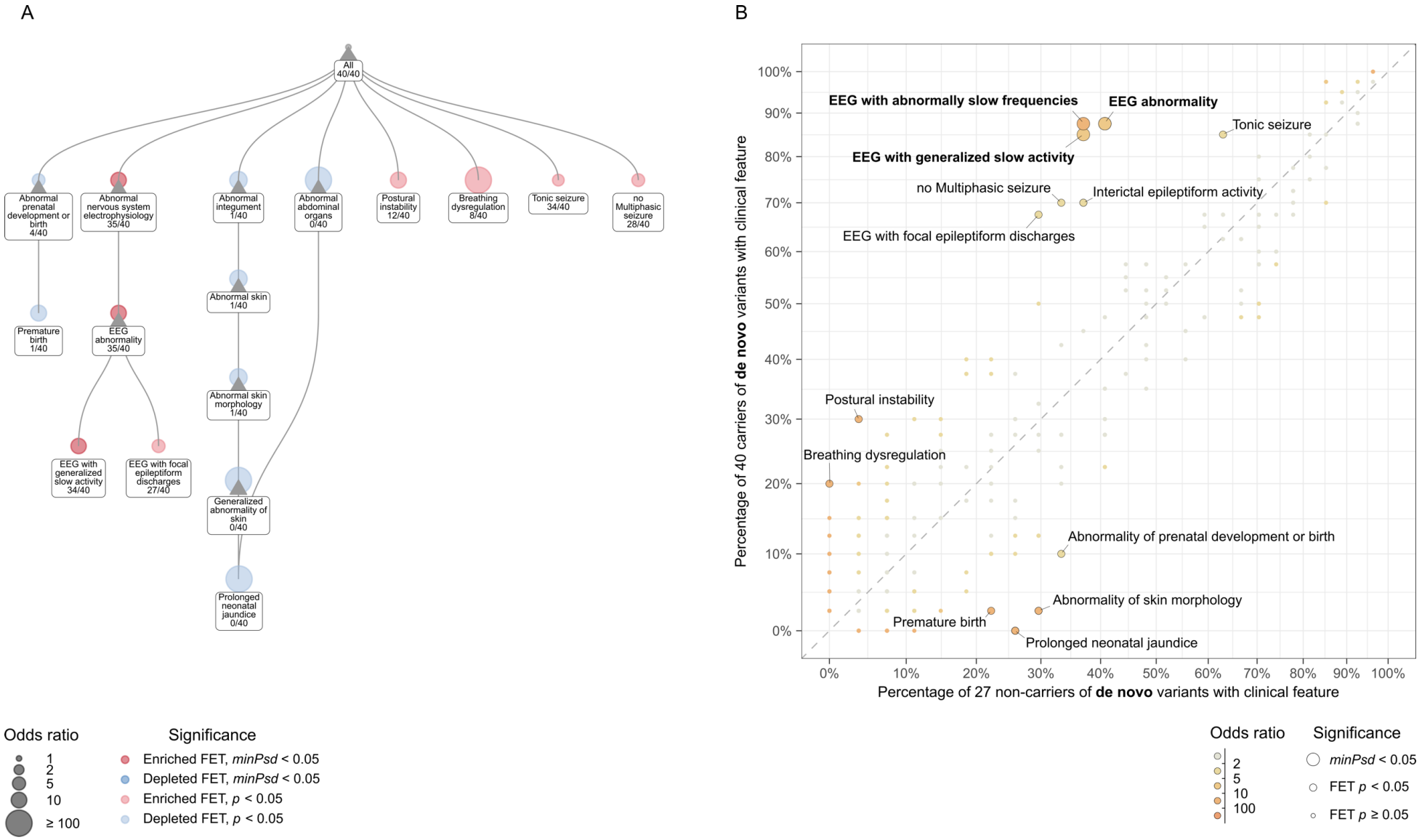


Figure 7-21. The phenotypic associations of 40 carriers of de novo *CDKL5* variants.

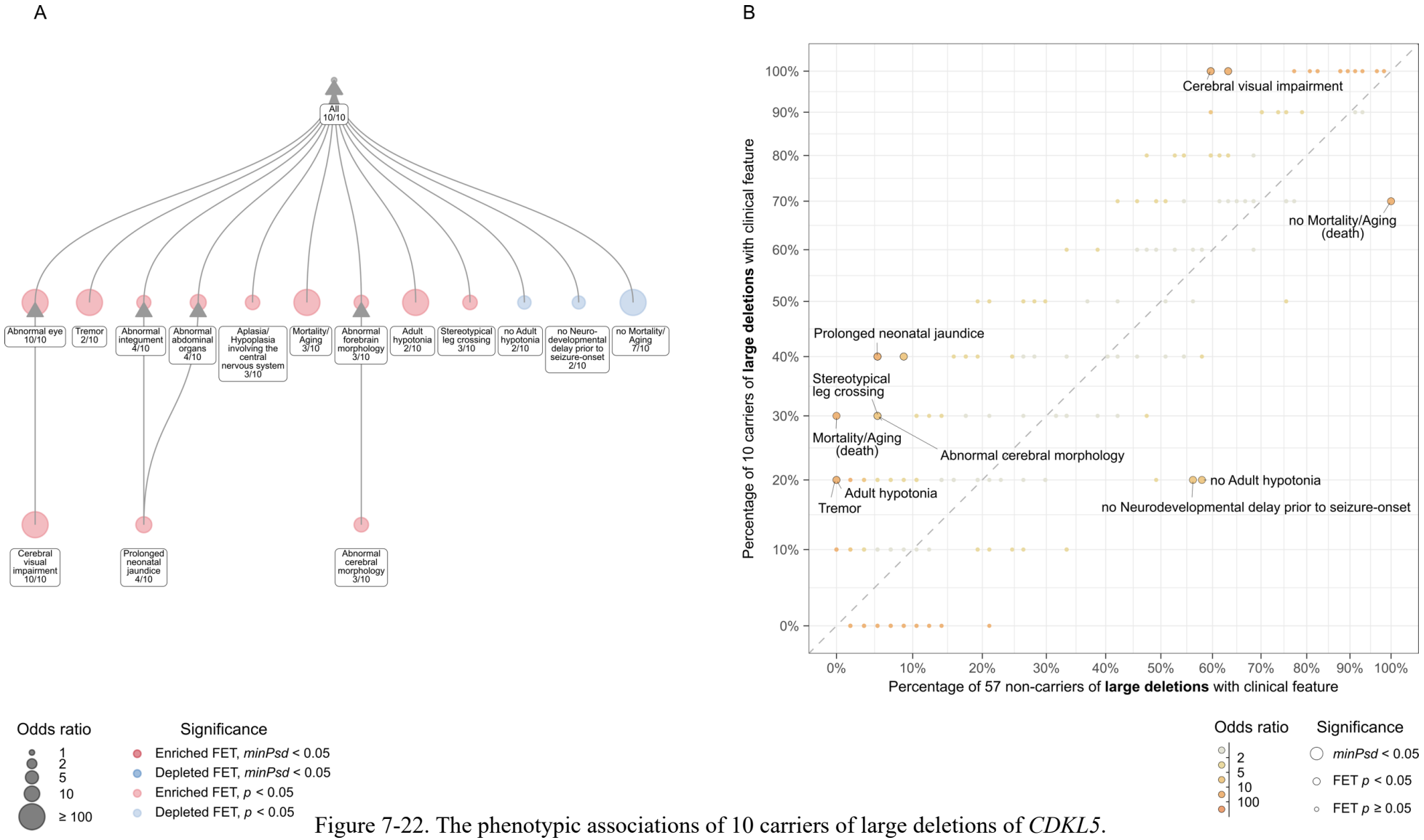


Figure 7-22. The phenotypic associations of 10 carriers of large deletions of *CDKL5*.

7.5.13 Genotype-based phenotypic similarity analysis

Those 45 with non-missense variants (one-sided Monte Carlo-derived one-sided $p = 0.0144$) and those ten with large deletions ($p = 0.00389$) were particularly similar according to $\mathrm{Sim}_{\mathrm{MinTO}}$ (**Supplementary Table 7-7**). Other measures that focus on the clinical features that are *present in both* of a pair of people (suited to an open-world interpretation), tended to agree (in particular $\mathrm{Sim}_{\mathrm{Resnik\text{-}BMS}}$, $\mathrm{Sim}_{\mathrm{TO}}$, $\mathrm{Sim}_{\mathrm{Resnik\text{-}BMA}}$, $\mathrm{Sim}_{\mathrm{Rel\text{-}BMS}}$, $\mathrm{Sim}_{\mathrm{Rel\text{-}BMA}}$, $\mathrm{Sim}_{\mathrm{Phrank}}$, and $\mathrm{Sim}_{\mathrm{RBP}}$). Conversely, $\mathrm{Sim}_{\mathrm{Goodall3}}$ is a function of those phenotypes that are *concordantly present or absent* in a pair. In this $\mathrm{Sim}_{\mathrm{Goodall3}}$ analysis I assumed that if a phenotype is not annotated as present, then it is absent. I found that those 45 with non-missense ($p = 0.00577$), those 22 with missense ($p = 0.0295$), and those two with c.587C>T ($p = 0.0342$) were similar according to $\mathrm{Sim}_{\mathrm{Goodall3}}$. Together these similarity and association results suggest that while those with missense variants might not share highly informative clinical phenotype, they can be considered similar in the context of people with CDD *because of the phenotypes that they lack*. The two people carrying deletions involving *NHS* were not particularly similar by any measure. Other similarity measures with published applications in biomedical, molecular, or linguistic fields that are less well suited to phenotypic analysis on theoretical grounds (4.6) are included for reference.

7.5.14 Sex-phenotype associations

The associations with sex reaching $p < 0.05$ were limited to indicating that the 12 male participants were more frequently annotated with *Focal motor seizure [HP:0011153]* (RR = 3.9, OR = 6.6, unadjusted FET two-sided $p = 0.00837$) and *Excessive salivation [HP:0003781]* (RR = infinite, OR = infinite, $p = 0.0299$) than the 55 females (**Figure 7-23** and **Supplementary Table 7-5**). Associations of clinical interest nearing this threshold suggested that females may be more likely to be noted to have *Abnormal aggressive, impulsive, or violent behavior [HP:0006919]* (RR = 2.4, OR = 4.4, $p = 0.0522$) and *Adult hypertonia [mHP:0000002]* (RR = 4.4, OR = 6.2, $p = 0.0861$). No sex-phenotype association reached $minPsd < 0.05$ and neither sex had increased phenotypic similarity according to any measure (**Supplementary Table 7-7**).

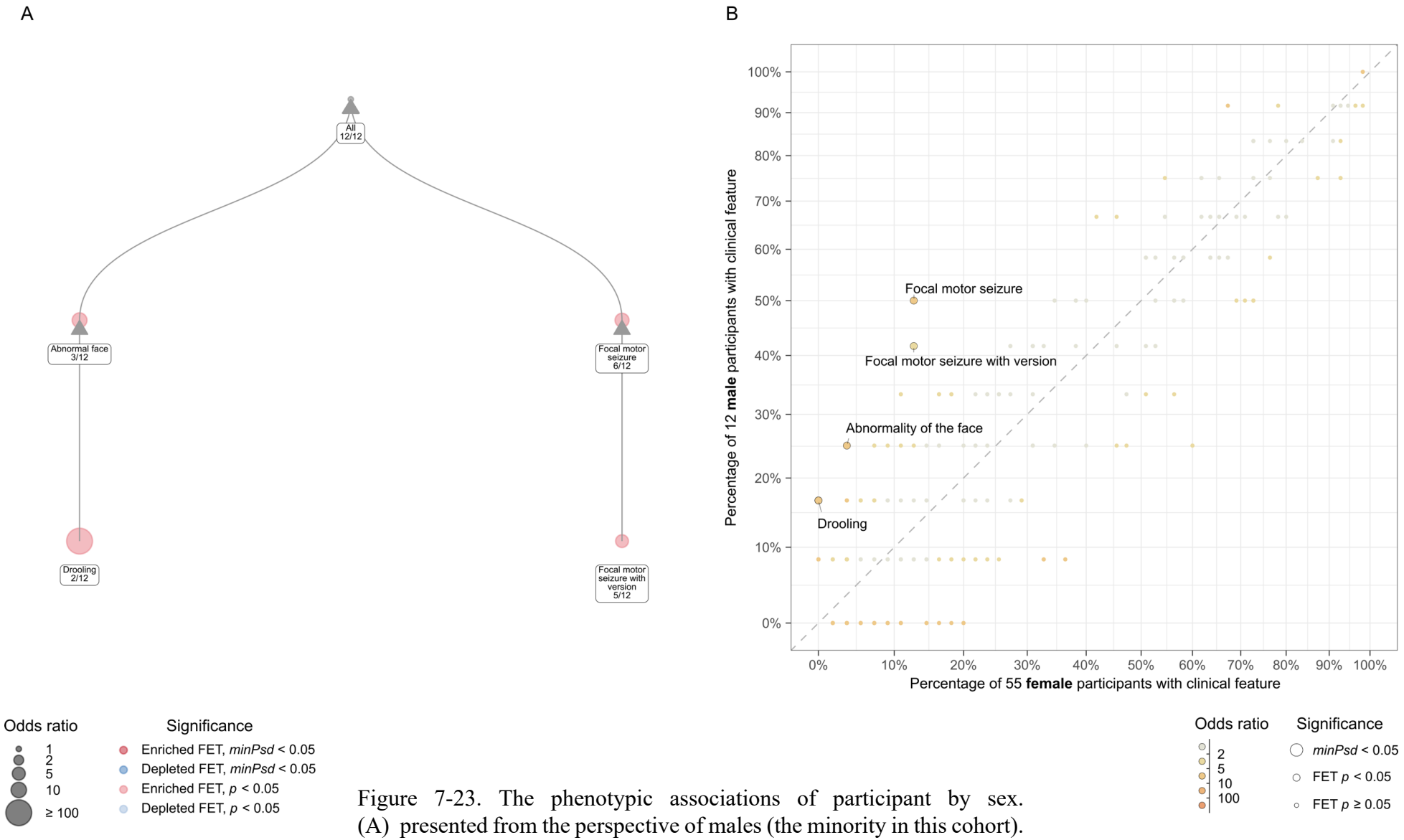


Figure 7-23. The phenotypic associations of participant by sex. (A) presented from the perspective of males (the minority in this cohort).

Previous studies have described lower rates of CDS skill attainment in males than females (Fehr *et al.*, 2015). I found insufficient evidence to support a relationship between sex and total maximum CDS (median among 12 males = 2 and among 55 females = 3, WRST $W = 368.5$, two-sided $p = 0.530$, **Figure 7-24A** and **Supplementary Table 7-4**) or most recent total CDS (median for males = 1 and females = 2, WRST $W = 400$, two-sided $p = 0.250$, **Figure 7-24B**). While I found that a smaller proportion of males attained some skills, these differences were not statistically significant (**Supplementary Table 7-6**).

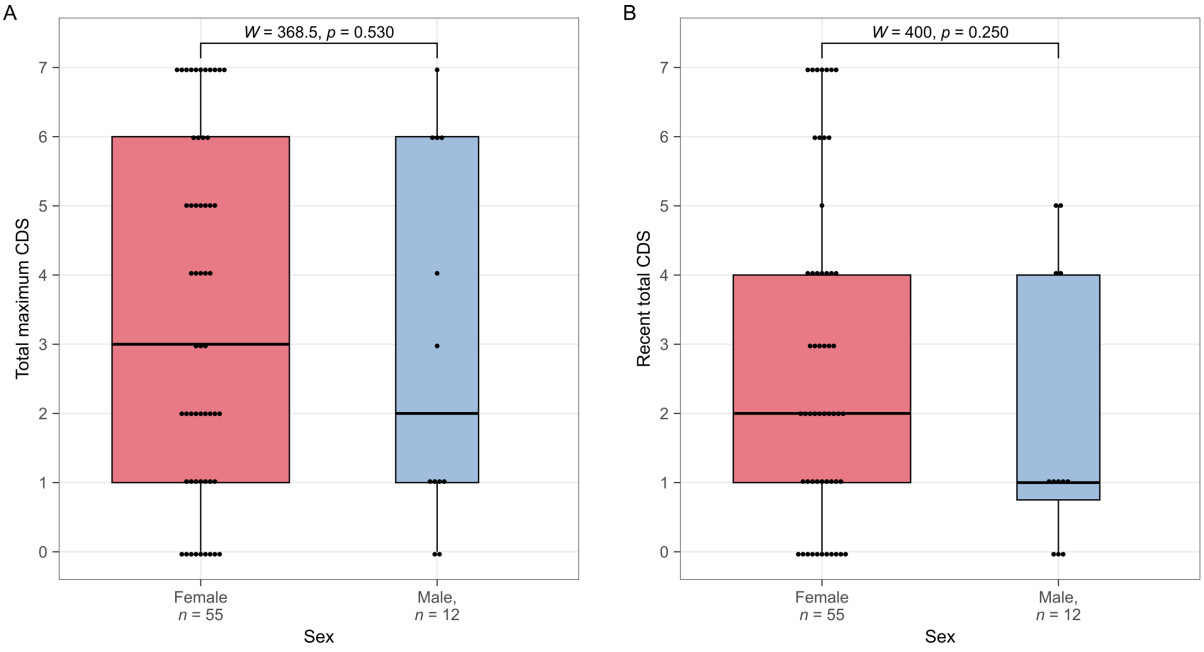


Figure 7-24. Total maximum and total recent CDS by sex.

7.5.15 Phenotypic associations of neonatal seizures

I assessed the phenotype-phenotype associations of neonatal seizures (*Neonatal electro-clinical seizure [HP:0032809]*) as this is an early phenotype that may indicate a more severe trajectory (**Supplementary Table 7-5**). Those 13 with a history of neonatal seizure were more likely to have been annotated with phenotypes including nephrolithiasis and pressure ulcers, which are likely to be complications of treatment and a severe motor disorder (**Figure 7-25**). Those 45 asserted to have not had neonatal seizures were more likely to present with global developmental delay and infantile (rather than neonatal) onset seizures (**Figure 7-26** and **Figure 7-27**),

Those with a history of neonatal seizures, but not those asserted not to have had neonatal seizures, were similar according to $\mathrm{Sim}_{\mathrm{MinTO}}$ and other open-world measures. Neither was particularly similar according to $\mathrm{Sim}_{\mathrm{Goodall3}}$.

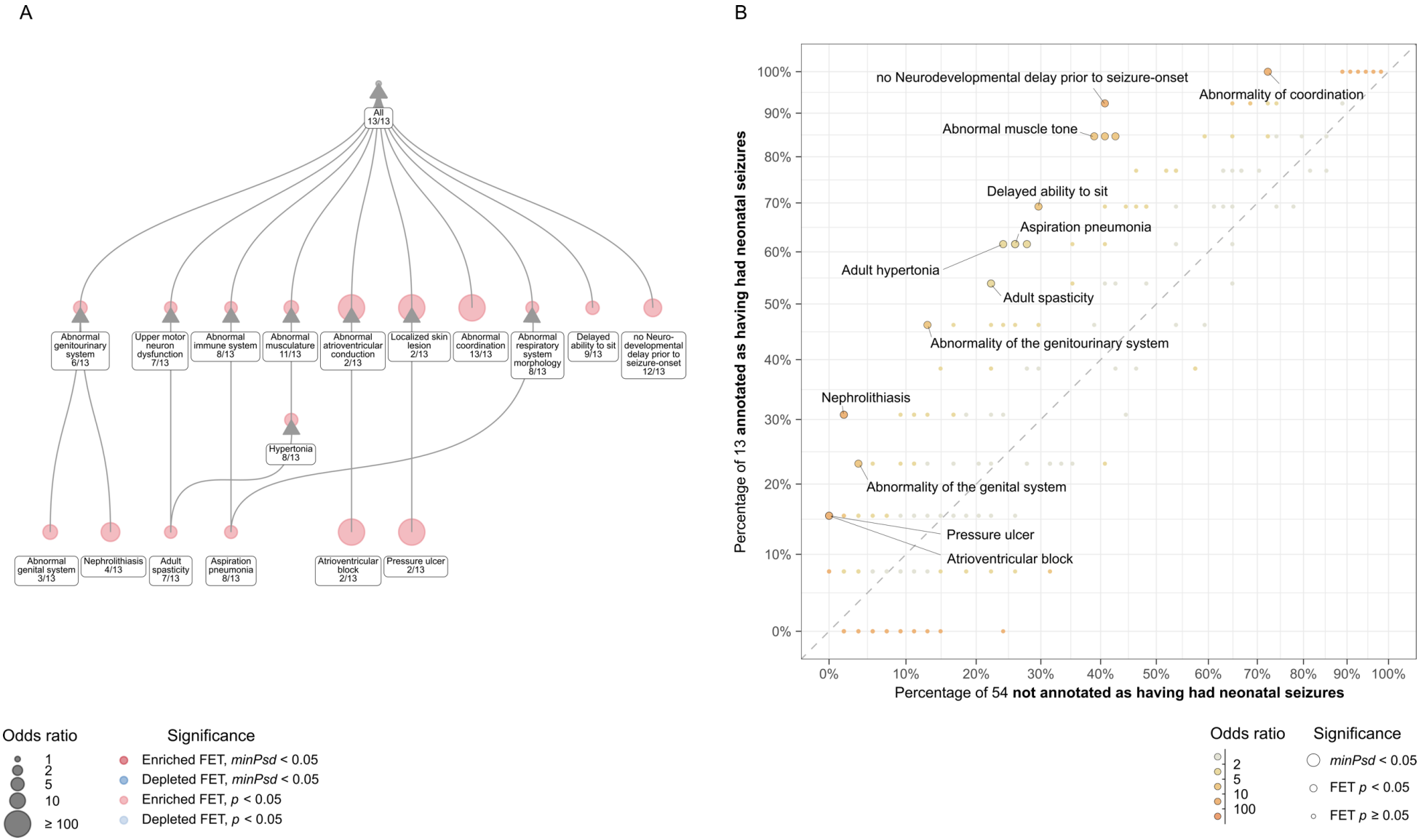


Figure 7-25. The phenotypic associations of those believed to have had neonatal seizures.

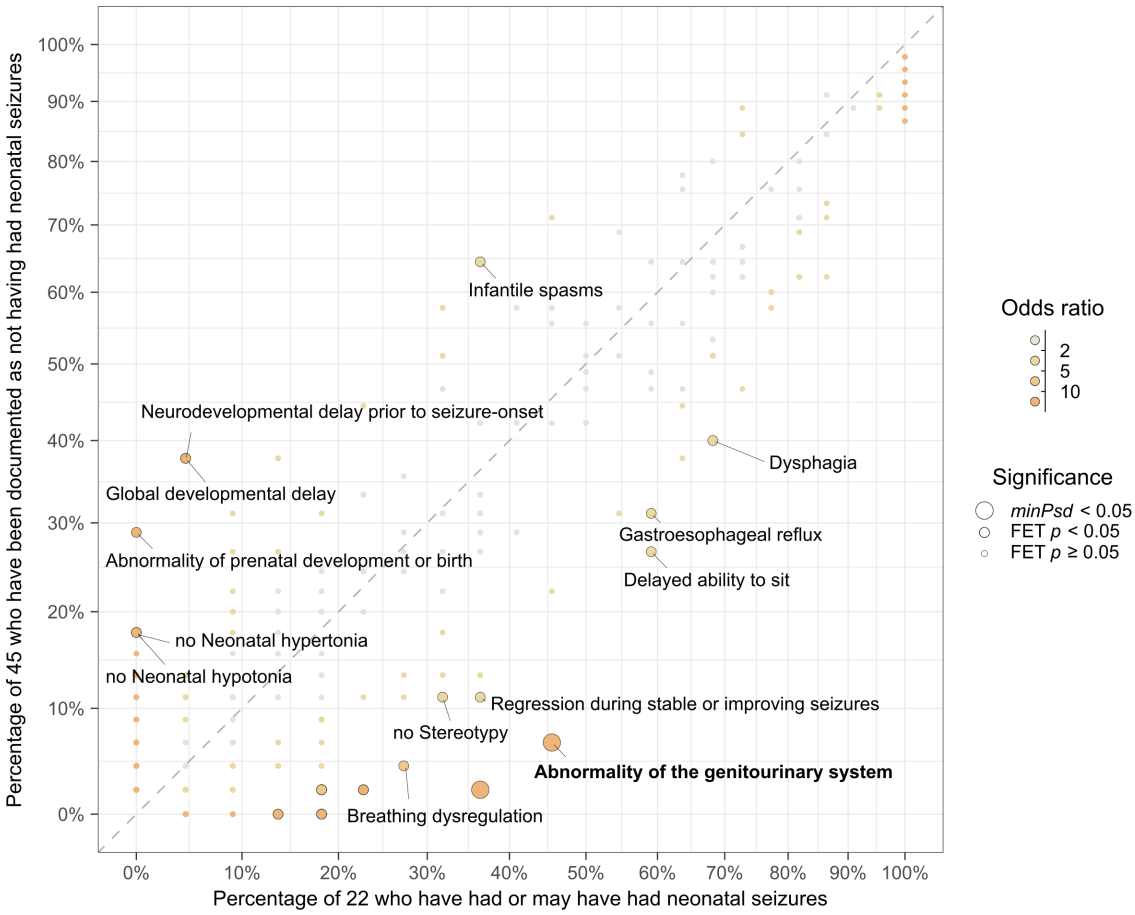


Figure 7-26. The phenotypic associations of those believed not to have had neonatal seizures, viewed statistically.

The 13 known to have had neonatal seizures were less likely to have be able to babble or sit independently. Those who were asserted to have not had neonatal seizures were more likely to have developed independent sitting (**Supplementary Table 7-6**).

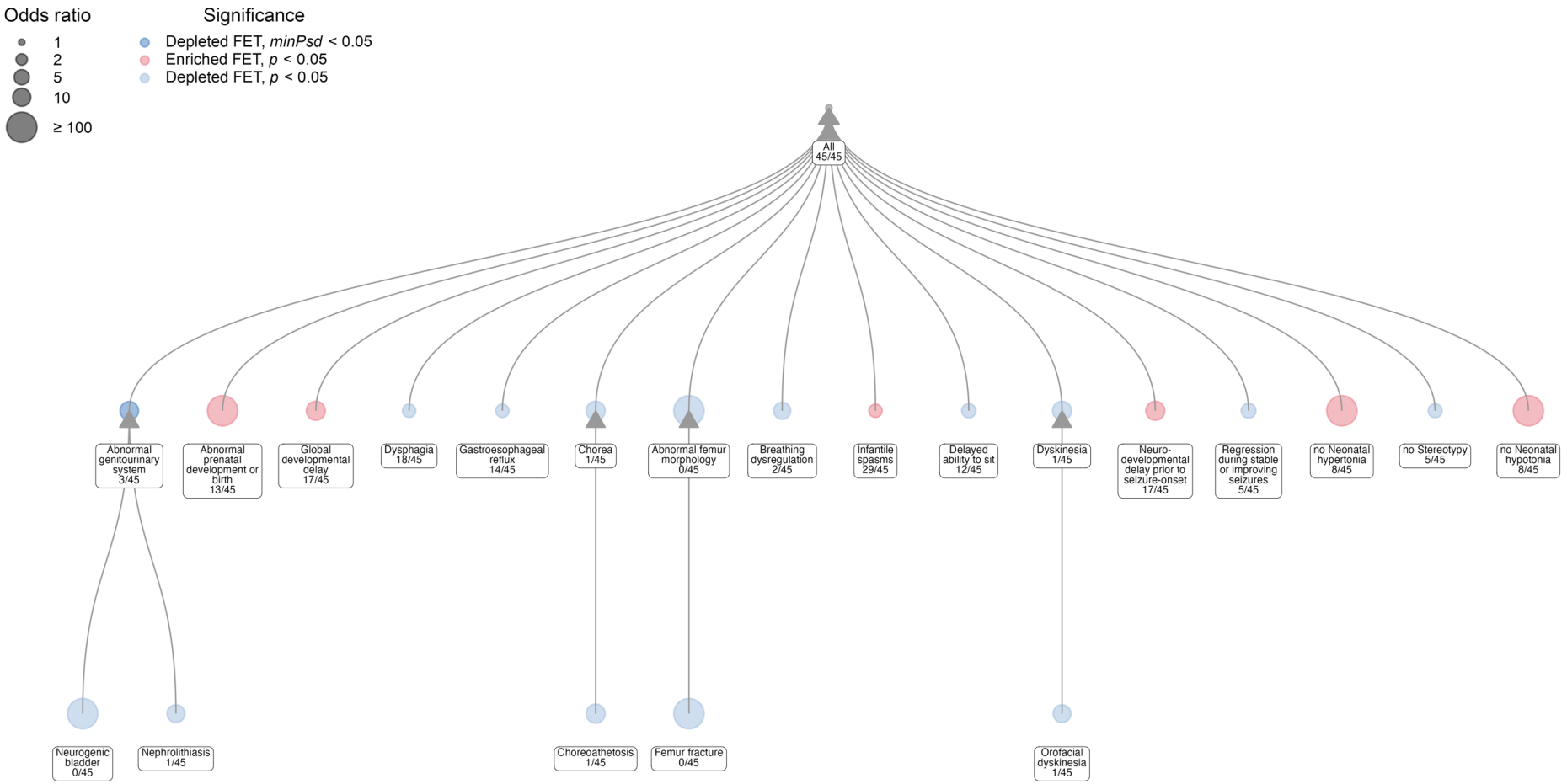


Figure 7-27. The phenotypic associations of those believed not to have had neonatal seizures, viewed conceptually.

7.5.16 Predicting total maximum CDS

A history of neonatal seizures and whether the *CDKL5* variant was missense or not were independently associated with total maximum CDS (**Supplementary Table 7-4**). Consequently, I explored the effect of stratification according to the combination of these two on total maximum CDS, focusing on the 58 people for whom a history of neonatal seizures had been annotated or asserted to have been excluded.

I found significant group-wise differences in total maximum CDS (Kruskal-Wallis rank sum test, *stats::kruskal.test()*, chi-squared = 9.85 with 3 degrees of freedom, $p = 0.0199$). Those who had both a history of neonatal seizures and a non-missense variant tended to have more limited skill acquisition (median = 0, mode = 0, **Figure 7-28**). More skills tended to be acquired by those with missense variants and no neonatal seizures (median = 4–5, mode = 7) and by those with non-missense variants and no neonatal seizures (median = 3, mode = 2). Only three people were known to have a missense variant and history of neonatal seizures (median = 4 range = 1–6), which limits the power to interpret comparisons with this combination of risk factors. Importantly for advising patients and caregivers at the individual clinical level, all groups spanned a wide range of total maximum CDS, indicating that regardless of variant type and the history of neonatal seizures, some people with CDD acquire multiple skills, and outcomes must be influenced by other factors. Two with a history of neonatal seizures and carrying non-missense variants attained all skills except a pincer grasp.

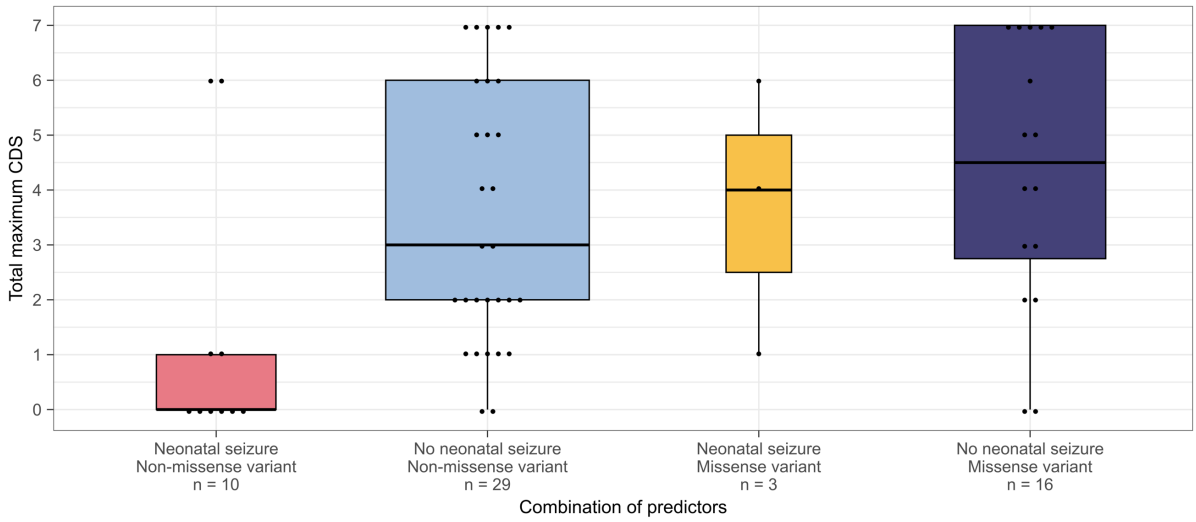


Figure 7-28. The total maximum CDS by the occurrence of neonatal seizures and whether the diagnostic *CDKL5* variant is missense.

Consider the situation where a genetic diagnosis is obtained after knowing that the person has had neonatal seizures. Without knowing the variant type, we would predict the median total

maximum CDS achieved by adulthood in this scenario to be 1 (range 0–6, $n = 13$, **Figure 7-11A**). Finding a missense diagnostic variant would allow revision of this prediction to a median of 4 (range 1–6, $n = 3$), and finding a non-missense variant to 0 (range 0–6, $n = 10$, **Figure 7-28**). The variant classification-agnostic median total maximum CDS was 4 (range 0-7, $n = 45$) for those without neonatal seizures, with little change on finding a missense (median = 4–5, range 0–7, $n = 16$) or non-missense (median = 3, range 0–7, $n = 29$) variant.

For reference, I provide the proportion of people acquiring each CDS skill across the entire cohort according to sex, history of neonatal seizures, broad variant class, and the combination of the last two of these factors (**Figure 7-29**, **Table 7-8**, and **Table 7-9**).

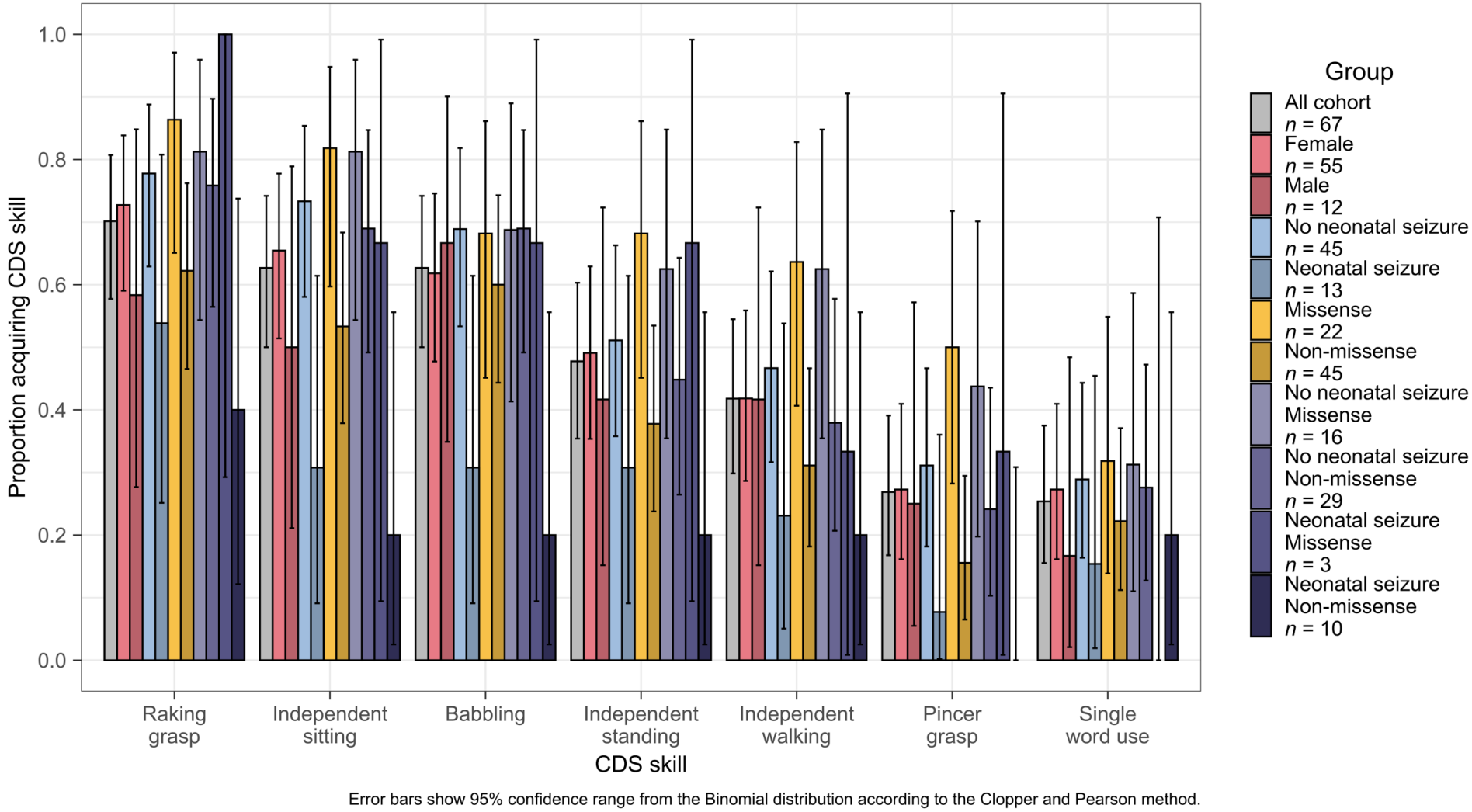


Figure 7-29. Attainment of CDS skills by sex, the occurrence of neonatal seizures, and whether the diagnostic *CDKL5* variant is missense.

Skill	Entire cohort (*n* = 67)		Neonatal seizures (*n* = 13)		Neonatal seizures and missense variant (*n* = 3)		Neonatal seizures and non-missense variant (*n* = 10)	
	Number acquiring skill	Proportion	Number acquiring skill	Proportion	Number acquiring skill	Proportion	Number acquiring skill	Proportion
Raking grasp	47	0.70	7	0.54	3	1.00	4	0.4
Independent sitting	42	0.63	4	0.31	2	0.67	2	0.2
Babbling	42	0.63	4	0.31	2	0.67	2	0.2
Independent standing	32	0.48	4	0.31	2	0.67	2	0.2
Independent walking	28	0.42	3	0.23	1	0.33	2	0.2
Pincer grasp	18	0.27	1	0.077	1	0.33	0	0.0
Single word use	17	0.25	2	0.15	0	0.00	2	0.2

Table 7-8. Attainment of CDS skills by whether the diagnostic *CDKL5* variant is missense or not, among those with neonatal seizures.

Skill	Entire cohort (*n* = 67)		No neonatal seizures (*n* = 45)		No neonatal seizures and missense variant (*n* = 16)		No neonatal seizures and non-missense variant (*n* = 29)	
	Number acquiring skill	Proportion	Number acquiring skill	Proportion	Number acquiring skill	Proportion	Number acquiring skill	Proportion
Raking grasp	47	0.70	35	0.78	13	0.81	22	0.76
Independent sitting	42	0.63	33	0.73	13	0.81	20	0.69
Babbling	42	0.63	31	0.69	11	0.69	20	0.69
Independent standing	32	0.48	23	0.51	10	0.63	13	0.45
Independent walking	28	0.42	21	0.47	10	0.63	11	0.38
Pincer grasp	18	0.27	14	0.31	7	0.44	7	0.24
Single word use	17	0.25	13	0.29	5	0.31	8	0.28

Table 7-9. Attainment of CDS skills by whether the diagnostic *CDKL5* variant is missense or not, among those without neonatal seizures.

7.6 Discussion

This is the first study to focus on the clinical features and experiences of CDD in adults, describing the developmental outcomes and providing a perspective of the range trajectories through to adulthood. It remains possible that this cohort overrepresents early-onset and severe presentations as these features may increase the likelihood of being offered testing and referral to academic centers (Aledo-Serrano *et al.*, 2020; Lewis-Smith *et al.*, 2020b). However, this retrospective international perspective confirms that epilepsy is common, usually presenting in infancy, and usually remains drug-resistant (Bahi-Buisson *et al.*, 2008a; Bahi-Buisson *et al.*, 2008b; Fehr *et al.*, 2013; Fehr *et al.*, 2016b). I can confirm that neurodevelopment is delayed with half manifesting regression (Fehr *et al.*, 2015; Fehr *et al.*, 2016a), and that skill acquisition remains limited despite prolonged follow-up. A history of neonatal seizures and classification of the variant may refine early prognostication.

I have extended the characterization of seizure trajectory into adulthood (Bahi-Buisson *et al.*, 2008a; Demarest *et al.*, 2019b). Tonic seizures and spasms, predominating in infancy, can persist into adulthood. Myoclonic, tonic-clonic, and focal-onset seizures may emerge from early childhood, with non-motor seizures later. Status epilepticus is rare compared to other genetic DEEs and typically non-convulsive (Donnan *et al.*, 2023). The majority of people never obtain 6 months of seizure freedom despite follow-up to adulthood, and most continue to have seizures daily, as in childhood (Fehr *et al.*, 2016b). However, many have subjective improvement in seizure burden in adulthood and worsening appears rare. EEG changes remain focal or multifocal (Bahi-Buisson *et al.*, 2008a) and hypsarrhythmia may be rare despite epileptic spasms (Demarest *et al.*, 2019b).

Few adults were seizure-free without treatment in adulthood and continuing polytherapy was common. We had only pragmatic subjective measures of efficacy, covering only a small proportion of treatment trials. Many people settle on clobazam, sodium valproate, or lamotrigine in adulthood, ASM found beneficial in a younger cohort (Muller *et al.*, 2016). Yet newer treatments appear promising. I found little to suggest that VNS or ketogenic diets offer sufficient long-term benefit to justify their burden on adult patients. Some people seem to benefit from therapeutic changes in adulthood, although recency and honeymoon periods could have biased evaluation of these most recent changes (Fehr *et al.*, 2016b; Muller *et al.*, 2016). Future studies of real-world treatment response may benefit from contemporaneous epilepsy-tailored electronic health records within epilepsy learning healthcare systems (Xian *et al.*, 2023).

Developmental delay can present before seizures, despite their infantile-onset. Nearly half lose skills, sometimes in early adulthood and despite improving seizures. I have referred to loss of developmental skills as *regression* pragmatically including when referring to adolescents and young adults, because we use it in clinic for want of a term specific to describe the loss of skills in young adults. It is unclear where the limits of *neurodevelopmental regression* and *dementia* lie, particularly where there is a subsequent stabilization without further skill loss. However, I accept that the loss of skills in early adulthood may be different from that during childhood (Kirsten *et al.*, 2023).

Intellectual disability without speech, with visual impairment and inability to stand or walk independently is typical of CDD in adulthood. One in six adults has never attained any CDS skill, but of those who do, half retain all their CDS skills with 16 % performing all or all but one of the skills, and our data do not suggest that skills are gradually lost in early adulthood. Extended follow-up shows that the proportion eventually able to stand and walk independently may be greater than previously found (Fehr *et al.*, 2015), but half of those learning to stand or walk will lose this ability. Furthermore, a history of neonatal seizures, present in 19 %, predicts more limited developmental outcomes. Whether these contribute to the developmental disorder or indicate manifestation of more severe disease remains unclear, although in pediatric populations low seizure burden may be associated with improvements in CDS (Leonard *et al.*, 2022b).

Movement, gastrointestinal, and sleep disorders remain common in early adulthood, and are occasionally parkinsonian. Many adults use laxatives and may have had gastrostomy or scoliosis surgery. I suspect that our data may underestimate osteoporosis as half of these 12 % had suffered fractures. The proportion who had undergone bone dual-energy x-ray absorptiometry is unknown, leading us to suspect that latent osteoporosis may be more common. Finding electrocardiographic abnormalities in 7.5 %, is in keeping with findings of prolonged QT_c (Stansauk *et al.*, 2023) and supraventricular tachycardia (Ivaniuk *et al.*, 2024) in pediatric CDD. These, the osteoporosis, and nephrolithiasis, emphasize the importance of considering the risks of ASMs and screening for subclinical comorbidities (Jefferson *et al.*, 2016; Winterhalder *et al.*, 2022).

People carrying missense variants were more likely to have anxiety or loss of skills in adulthood, potentially because those with non-missense variants had more limited neurodevelopment and consequently less potential to lose skills or ability to express anxiety. Carriers of non-missense variants were enriched with a broad range of phenotypes. Our reported

associations may heighten clinicians' attention regarding clinical features to remain vigilant for.

Our retrospective perspective from adulthood does not capture the experience of those dying before adulthood. While infections were the principal cause of death, the report of active seizures in the final days of all three emphasizes the importance of seizure control in people with infections and during palliative care.

A history of neonatal seizures appears independent of whether a variant is missense or not, yet both appear to predict lower peak neurodevelopment by CDS and a broader range of comorbidities. I propose recognition of these features as markers of a general prognostic trend but advise against any deterministic interpretation of a formal predictive model based solely on these for any particular person.

Despite five people receiving a molecular diagnosis within 3 years of *CDKL5* being established as a cause of DEE (Weaving *et al.*, 2004), most were diagnosed in adolescence. Consequently, if early molecular diagnosis becomes the norm, and if genetically stratified care leads to better management and eventually treatments that mitigate pathogenesis, future cohorts may have a different trajectory. People with more frequent seizures were said to have benefited the most from diagnosis, potentially because of previous reports of treatment outcomes and the emergence of ganaxolone (Muller *et al.*, 2016; Aledo-Serrano *et al.*, 2021; Amin *et al.*, 2022; Knight *et al.*, 2022).

Manually coding of highly dimensional clinical data into a formal framework that allows scalable inference has allowed us to maximize the information included in our analyses, despite the challenges of retrospective real-world data, variability in phenotypic screening, and documentation across centers and healthcare systems. I was able to capture a median of 25 independent phenotypes as present and 14 as absent per person, representing 507 harmonized concepts. In contrast, retrospective case series often report only the proportion of individuals known to have had each phenotype. This risks the closed-world interpretation that the phenotype was absent in the remainder. As in Crawford *et al.* (2021), here I differentiated between phenotypes that were excluded and those that had not been ascertained in a particular participant. This acknowledges the variability in phenotypic capture from real-world data, particularly when collating data from multiple clinical settings and spanning multiple ages, and the inherent uncertainties and inaccuracies that clinicians are used to accommodating. Some phenotypes may not have been reliably documented due to limited relevance to neurological care (for example, irregular menstruation). Others may be difficult to exclude without investigations that may not have been performed universally, such as osteoporosis. In

describing the phenotypic repertoire, I focus on the proportion asserted to have manifest a particular phenotype, as the true value is likely to lie near this lower bound, particularly for neurological phenotypes in the study of neurological disorders. Quantifying the proportion found not have it despite observation and the resulting phenotyping gap allows one to refine the upper limit (Lewis-Smith *et al.*, 2022). Furthermore, the definite absence of a phenotype can be important because of its functional consequences (for example, the absence of profound intellectual disability or inability to speak), its diagnostic value (the absence of hypsarrhythmia during epileptic spasms is suggestive of CDD (Demarest *et al.*, 2019b)), or its prognostic value (the absence of neonatal seizures (this study) and the absence of cortical visual impairment (Demarest *et al.*, 2019b) are associated with better neurodevelopmental outcomes).

While increasing use of CDEs (Grinspan *et al.*, 2021; Lewis-Smith *et al.*, 2022) and CDD-tailored assessments (Demarest *et al.*, 2019a) will strengthen EMR-based analyses in the future, this study demonstrates the value of critical harmonization and pragmatic retrospective analysis of international real-world data in rare diseases, adding a longitudinal dimension that informs clinical prognosis and management and the design of future clinical trials. The presence of neonatal seizures and classification of the variant may indicate different trajectories, refining prognostication at the time of diagnosis.

Chapter 8. Epilogue

I have shown that clinical reasoning can be modeled to harmonize and analyze large amounts of clinical data at the detailed level of both single phenotypes and each person's constellation of phenotypes, their personal syndrome. The methods I have presented can accommodate data provided at different degrees of granularity, an intrinsic feature of clinical epilepsy, and can automate reasoning at a scale that precludes manual analysis yet is explainable. As with any model, there are assumptions that need consideration when interpreting the data and deciding upon the methods to employ and interpreting their results. Most importantly, the clinical data usually relate to the presence or absence of documentation of a phenotype rather than the presence or absence of the phenotype itself and are commonly incomplete, reflecting the perspectives, motivations, and limitations of those undertaking clinical assessments and recording the outcomes of these. Furthermore, the means by which data have been collected and how they have been structured to facilitate analysis should influence interpretation. For example, *Homonymous hemianopia [HP:0030516]* classically indicates a retrochiasmic lesion to a neurologist, most frequently involving the contralateral occipital lobe. However, at present the structure of the HPO represents this as a subtype of *Abnormality of the eye [HP:0000478]* without any relationship to *Abnormality of the nervous system [HP:0000707]* because it aims to represent phenotypes by their anatomic and physiological systems rather than interpretation of their mechanisms.[34] Consequently, while such extreme examples are rare, one must be aware of the peculiarities of the HPO when interpreting the results of analyses based on its representations.

My work as reported here shows that formal representations of clinical features can be improved to integrate features of multiple classifications to allow automated analysis of data from international sources within a harmonized framework that is interpretable as a model of clinical reasoning. Phenome-wide association analyses can be performed and interpreted while allowing for the large number of hypotheses tested while recognizing their dependency structure. Furthermore, a variety of phenotypic similarity measures can provide complementary interpretations of the data, on the basis of different assumptions and capturing different concepts of *similarity*. The scale at which we can implement these approaches now is an order of magnitude greater than five years ago: we analyzed 31,742 annotations of 846 participants in our study of DEEs (Galer *et al.*, 2020) compared with 291,432 annotations of 12,611 participants in **Table 5-6**.

[34] A summary of my discussion of this with Prof. Peter N. Robinson is available on https://github.com/obophenotype/human-phenotype-ontology/issues/7291.

The primary aim of my work on Epi25 data was to identify the genetic contribution to phenotypic variability within populations with epilepsy rather than to discover new causes of epilepsy. This aim addresses the scenario in which a person undergoing genetic testing already has epilepsy and the test aims to provide information about how it is likely to affect them, rather than solely explaining why they have epilepsy and quantifying the familial recurrence risk. I have shown that stratifying people with epilepsy by the CNVs and URVs that they carry can indicate their risk of a wide range of clinical features, even on the basis of research data provided primarily for quality control of broad epilepsy type in a case-control genetic association study. These data are sparse when compared with the information a clinician considers during a clinical encounter, and the proportion of participants annotated with some phenotypes is likely to underestimate their prevalence. This suggests that more complete data may increase the power of these analyses. This may explain why many phenotypes that might be expected to be associated with established epilepsy genes in 5.5.7 fell short of the $minP < 0.05$ threshold. For example, only one person was annotated with *Cortical tubers [HP:0009717]* and consequently the study was underpowered to demonstrate a robust association with these despite the person affected carrying a dMissense variant in the appropriate gene, *TSC2* (predictive value = 1, $p = 0.00349$, $minP = 0.708$). I hope that were these findings to be replicated, consolidated, and even extended with richer data sets, they might contribute to genetic stratification, improving the precision and accuracy of prognostication, and informing clinical decisions relating to treatment and screening for comorbidities. Tests for these genetic factors are readily available to clinicians and patients fortunate enough to benefit from developed healthcare systems, and both the NHS Genomic Medicine Service (NHS England, 2024) and US healthcare organizations (Verma *et al.*, 2022; Wiley *et al.*, 2024) are eager to deliver on the promises of genomic precision medicine. Additionally, similar association and similar phenotypic analyses could be applied to groups defined by hypothesized explanatory factors beyond genetics. For example, one could stratify participants by candidate biomarkers on MRI or EEG that might indicate different system disorders nearer to clinical manifestations than a genetic variant. Furthermore, one could identify the clinical features of super-responders to a particular ASM and use these to stratify first-line treatments on phenotypic grounds.

I have shown that one can generate phenotype-derived hypotheses implicating particular novel genetic factors may have a role in epilepsy. I shall be interested to see if independent evidence for some of the novel candidate genes identified in Chapters 5 and 6 supports their role as causes, risk factors, or phenotypic modifiers.

I have not addressed the phenotypic consequences of common variants (SNPs) or combinations of variants in the work recorded here. The phenotypic consequences of individual common variants on precise phenotypes will be modest. However, as polygenic risk scores are refined and evidence builds of a role for SNPs in even presumed monogenic disorders (Campbell *et al.*, 2022; Oliver *et al.*, 2022; Martins Custodio *et al.*, 2023) it will be interesting to see if these approaches can be adapted for analyzing multidimensional genetic data, integrating the genomic architecture of URVs, CNVs, and SNPs.

A major remaining challenge is the incorporation of the chronological dimension. In addition to the broad epilepsy type (Scheffer *et al.*, 2017), traditional epilepsy syndromes are primarily conceptualized by the developmental stage at which the epilepsy manifests (Wirrell *et al.*, 2022a). A myoclonic seizure is likely to be a symptom of different pathophysiological processes in a neonate, toddler, adolescent, and pensioner. Furthermore, the age of onset of some phenotypes is a strong prognostic indicator, for example in differentiating between those *SCN1A*-related epilepsies that will progress to Dravet Syndrome and those that progress to febrile seizures-plus with little, if any, developmental consequences (Brunklaus *et al.*, 2022). The results presented in Chapter 7 suggest that age of seizure onset may predict developmental outcomes in CDD, a condition in which all adults had some degree of intellectual disability. We and others (Boßelmann *et al.*, 2024; Ivaniuk *et al.*, 2024) have started developing a framework for identifying clinical associations, and performing digital natural history studies in epilepsy based on analysis of routine EMR data to see how phenotypic associations (Ganesan *et al.*, 2020) and similarity (Lewis-Smith *et al.*, 2021b), seizure burden (Fitzgerald *et al.*, 2021), developmental outcomes, and prescription patterns (Xian *et al.*, 2023) evolve over time from EMRs, reviewed in Chapter 3 and Lewis-Smith (2022).[35]

Monogenic DEEs are one of the groups of disorders for which precision medicine makes big promises which it has yet to fulfill (Balestrini *et al.*, 2021; Knowles *et al.*, 2022). However, while candidate treatments have emerged for clinical trials, the natural history against which we should assess the effectiveness of treatments is not known. Prospective natural history studies are the gold standard for fully characterizing the evolution of a disease. Yet, these will take many years to complete if they are to chart the trajectory of a disease into adulthood, particularly as many of these disorders have only recently been attributed to a particular genetic cause. Furthermore, prolonged natural history studies and clinical trials are expensive and place

[35] Additionally, I have explored the mapping and harmonization of seizure phenotypes into age-defined bins for association and similarity analysis of Epi25 data.

a significant burden upon the relatively small population of patients with each rare disease and their caregivers in addition to the direct consequences of the disease. The increasing availability of clinical data from EMRs, and their collation by organizations such as Ciitizen (Ciitizen corporation, 2024), Rare-X (Rare-X, 2024), Simon's Searchlight (Simons Searchlight, 2024), and Komodo Health (Komodo Health, 2024), often in partnership with rare disease charities, makes retrospective natural history studies feasible (Xian *et al.*, 2023). As demonstrated in the study of CDD in Chapter 7, routinely collected data can make discoveries that guide care, service development, and further research more rapidly and cost-effectively.

Compared to the high-throughput of modern genomics, MRI, and EEG, from which vast quantities of data can be collected and processed by reproducible pipelines, the burden of collecting, curating, and validating clinical phenotypes remains heavy. Yet, it is the symptoms of epilepsy, its outcomes, and the effectiveness and side-effects of its treatment that people with epilepsy prioritize (Epilepsy Research Institute UK, 2021). While there will be cultural and political challenges, the standardization of epilepsy records promises to bring about a data revolution in which CDEs for epilepsy outcomes are routinely collected for more precise care of individual patients, and exploited at scale for the appraisal of the quality of care and research (Fitzgerald *et al.*, 2021; Lewis-Smith, 2022; Mitchell *et al.*, 2022). Such core data structures might be enhanced by improved natural language processing coupled to advances in large language models, yet for the care of individual patients these have some way to go without significant manual supervision (Fonferko-Shadrach *et al.*, 2019; Xie *et al.*, 2023; Landais;Sultan and Thomas, 2024; van Diessen *et al.*, 2024).

The future promises learning healthcare systems, integrating personalized care, research, and quality improvement, to produce longitudinal, multimodal datasets that include clinical phenotypes, patient-centered outcomes, and interventions (Fitzsimons *et al.*, 2013; Health and Board on Health Sciences Policy Institute of Medicine, 2015; Delanty *et al.*, 2019; Donahue *et al.*, 2021; Grinspan *et al.*, 2021). I hope that the methods described in this work will help to maximize the value of these by enabling near real-time continuous and large-scale evaluation of how epilepsy presents, evolves, and is modified by interventions for people stratified by their genetics and other factors, with the ultimate aim of enabling people with epilepsy to live their lives as freely and fully as possible.

Bibliography

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A. and Abecasis, G.R. (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68-74.

Abel, H.J., Larson, D.E., Regier, A.A., Chiang, C., Das, I., Kanchi, K.L., Layer, R.M., Neale, B.M., Salerno, W.J., Reeves, C., Buyske, S., Matise, T.C., Muzny, D.M., Zody, M.C., Lander, E.S., Dutcher, S.K., Stitziel, N.O. and Hall, I.M. (2020) 'Mapping and characterization of structural variation in 17,795 human genomes', *Nature*, 583(7814), pp. 83-89.

Abul-Husn, N.S., Cheng, X., Li, A.H., Xin, Y., Schurmann, C., Stevis, P., Liu, Y., Kozlitina, J., Stender, S., Wood, G.C., Stepanchick, A.N., Still, M.D., McCarthy, S., O'Dushlaine, C., Packer, J.S., Balasubramanian, S., Gosalia, N., Esopi, D., Kim, S.Y., Mukherjee, S., Lopez, A.E., Fuller, E.D., Penn, J., Chu, X., Luo, J.Z., Mirshahi, U.L., Carey, D.J., Still, C.D., Feldman, M.D., Small, A., Damrauer, S.M., Rader, D.J., Zambrowicz, B., Olson, W., Murphy, A.J., Borecki, I.B., Shuldiner, A.R., Reid, J.G., Overton, J.D., Yancopoulos, G.D., Hobbs, H.H., Cohen, J.C., Gottesman, O., Teslovich, T.M., Baras, A., Mirshahi, T., Gromada, J. and Dewey, F.E. (2018) 'A Protein-Truncating HSD17B13 Variant and Protection from Chronic Liver Disease', *N Engl J Med*, 378(12), pp. 1096-1106.

Addis, L., Rosch, R.E., Valentin, A., Makoff, A., Robinson, R., Everett, K.V., Nashef, L. and Pal, D.K. (2016) 'Analysis of rare copy number variation in absence epilepsies', *Neurol Genet*, 2(2), p. e56.

Addis, L., Sproviero, W., Thomas, S.V., Caraballo, R.H., Newhouse, S.J., Gomez, K., Hughes, E., Kinali, M., McCormick, D., Hannan, S., Cossu, S., Taylor, J., Akman, C.I., Wolf, S.M., Mandelbaum, D.E., Gupta, R., van der Spek, R.A., Pruna, D. and Pal, D.K. (2018) 'Identification of new risk factors for rolandic epilepsy: CNV at Xp22.31 and alterations at cholinergic synapses', *J Med Genet*, 55(9), pp. 607-616.

Adult Genotype Tissue Expression (GTEx) Project (2017) 'GTEx Portal'. 2022-10-21. pp. GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz. Available at: https://gtexportal.org/home/downloads/adult-gtex/overview.

Afawi, Z., Oliver, K.L., Kivity, S., Mazarib, A., Blatt, I., Neufeld, M.Y., Helbig, K.L., Goldberg-Stern, H., Misk, A.J., Straussberg, R., Walid, S., Mahajnah, M., Lerman-Sagie, T., Ben-Zeev, B., Kahana, E., Masalha, R., Kramer, U., Ekstein, D., Shorer, Z., Wallace, R.H., Mangelsdorf, M., MacPherson, J.N., Carvill, G.L., Mefford, H.C., Jackson, G.D., Scheffer, I.E., Bahlo, M., Gecz, J., Heron, S.E., Corbett, M., Mulley, J.C., Dibbens, L.M., Korczyn, A.D. and Berkovic, S.F. (2016) 'Multiplex families with epilepsy: Success of clinical and molecular genetic characterization', *Neurology*, 86(8), pp. 713-22.

Akawi, N., McRae, J., Ansari, M., Balasubramanian, M., Blyth, M., Brady, A.F., Clayton, S., Cole, T., Deshpande, C., Fitzgerald, T.W., Foulds, N., Francis, R., Gabriel, G., Gerety, S.S., Goodship, J., Hobson, E., Jones, W.D., Joss, S., King, D., Klena, N., Kumar, A., Lees, M., Lelliott, C., Lord, J., McMullan, D., O'Regan, M., Osio, D., Piombo, V., Prigmore, E., Rajan, D., Rosser, E., Sifrim, A., Smith, A., Swaminathan, G.J., Turnpenny, P., Whitworth, J., Wright, C.F., Firth, H.V., Barrett, J.C., Lo, C.W., FitzPatrick, D.R., Hurles, M.E. and study, D.D.D. (2015) 'Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families', *Nat Genet*, 47(11), pp. 1363-9.

Aledo-Serrano, A., García-Morales, I., Toledano, R., Jiménez-Huete, A., Parejo, B., Anciones, C., Mingorance, A., Ramos, P. and Gil-Nagel, A. (2020) 'Diagnostic gap in genetic epilepsies: A matter of age', *Epilepsy Behav*, 111, p. 107266.

Aledo-Serrano, Á., Gómez-Iglesias, P., Toledano, R., Garcia-Peñas, J.J., Garcia-Morales, I., Anciones, C., Soto-Insuga, V., Benke, T.A., Del Pino, I. and Gil-Nagel, A. (2021) 'Sodium channel blockers for the treatment of epilepsy in CDKL5 deficiency disorder: Findings from a multicenter cohort', *Epilepsy Behav*, 118, p. 107946.

Aledo-Serrano, Á., Valls-Carbó, A., Fenger, C.D., Groeppel, G., Hartlieb, T., Pascual, I., Herraez, E., Cabal, B., García-Morales, I., Toledano, R., Budke, M., Beltran-Corbellini, Á., Baldassari, S., Coras, R., Kobow, K., Herrera, D.M., Del Barrio, A., Dahl, H.A., Del Pino, I., Baulac, S., Blumcke, I., Møller, R.S. and Gil-Nagel, A. (2023) 'D-galactose Supplementation for the Treatment of Mild Malformation of Cortical Development with Oligodendroglial Hyperplasia in Epilepsy (MOGHE): A Pilot Trial of Precision Medicine After Epilepsy Surgery', *Neurotherapeutics*, 20(5), pp. 1294-1304.

Allers, K., Essue, B.M., Hackett, M.L., Muhunthan, J., Anderson, C.S., Pickles, K., Scheibe, F. and Jan, S. (2015) 'The economic impact of epilepsy: a systematic review', *BMC neurology*, 15, pp. 245–245.

Alsfouk, B.A.A., Brodie, M.J., Walters, M., Kwan, P. and Chen, Z. (2020) 'Tolerability of Antiseizure Medications in Individuals With Newly Diagnosed Epilepsy', *JAMA Neurol*.

Ambrosio, A.L., Febvre, H.P. and Di Pietro, S.M. (2022) 'Syntaxin 12 and COMMD3 are new factors that function with VPS33B in the biogenesis of platelet α-granules', *Blood*, 139(6), pp. 922-935.

American Academy of Pediatrics, Steering Committee on Quality, Improvement, Management, Subcommittee on Febrile Seizures, (2008) 'Febrile seizures: clinical practice guideline for the long-term management of the child with simple febrile seizures', *Pediatrics*, 121(6), pp. 1281-6.

Amiet, C., Gourfinkel-An, I., Bouzamondo, A., Tordjman, S., Baulac, M., Lechat, P., Mottron, L. and Cohen, D. (2008) 'Epilepsy in autism is associated with intellectual disability and gender: evidence from a meta-analysis', *Biol Psychiatry*, 64(7), pp. 577-82.

Amin, S., Monaghan, M., Aledo-Serrano, A., Bahi-Buisson, N., Chin, R.F., Clarke, A.J., Cross, J.H., Demarest, S., Devinsky, O., Downs, J., Pestana Knight, E.M., Olson, H., Partridge, C.A., Stuart, G., Trivisano, M., Zuberi, S. and Benke, T.A. (2022) 'International Consensus Recommendations for the Assessment and Management of Individuals With CDKL5 Deficiency Disorder', *Front Neurol*, 13, p. 874695.

Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U. and Zoghbi, H.Y. (1999) 'Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2', *Nat Genet*, 23(2), pp. 185-8.

Andrews, T., Meader, S., Vulto-van Silfhout, A., Taylor, A., Steinberg, J., Hehir-Kwa, J., Pfundt, R., de Leeuw, N., de Vries, B.B. and Webber, C. (2015) 'Gene networks underlying convergent and pleiotropic phenotypes in a large and systematically-phenotyped cohort with heterogeneous developmental disorders', *PLoS Genet*, 11(3), p. e1005012.

Androsova, G., Krause, R., Borghei, M., Wassenaar, M., Auce, P., Avbersek, A., Becker, F., Berghuis, B., Campbell, E., Coppola, A., Francis, B., Wolking, S., Cavalleri, G.L., Craig, J., Delanty, N., Koeleman, B.P.C., Kunz, W.S., Lerche, H., Marson, A.G., Sander, J.W., Sills, G.J., Striano, P., Zara, F., Sisodiya, S.M., Depondt, C. and Epi, P.G.X.C. (2017) 'Comparative effectiveness of antiepileptic drugs in patients with mesial temporal lobe epilepsy with hippocampal sclerosis', *Epilepsia*, 58(10), pp. 1734-1741.

Antony, J., Zanini, E., Birtley, J.R., Gabra, H. and Recchi, C. (2021) 'Emerging roles for the GPI-anchored tumor suppressor OPCML in cancers', *Cancer Gene Ther*, 28(1-2), pp. 18-26.

Archer, H.L., Evans, J., Edwards, S., Colley, J., Newbury-Ecob, R., O'Callaghan, F., Huyton, M., O'Regan, M., Tolmie, J., Sampson, J., Clarke, A. and Osborne, J. (2006) 'CDKL5 mutations cause infantile spasms, early onset seizures, and severe mental retardation in female patients', *J Med Genet*, 43(9), pp. 729-34.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) 'Gene ontology: tool for the unification of biology. The Gene Ontology Consortium', *Nat Genet*, 25(1), pp. 25-9.

Asmar, A.J., Abrams, S.R., Hsin, J., Collins, J.C., Yazejian, R.M., Wu, Y., Cho, J., Doyle, A.D., Cinthala, S., Simon, M., van Jaarsveld, R.H., Beck, D.B., Kerosuo, L. and Werner, A. (2023) 'A ubiquitin-based effector-to-inhibitor switch coordinates early brain, craniofacial, and skin development', *Nat Commun*, 14(1), p. 4499.

Avanzini, G. (2003) 'Of Cabbages and Kings: Do We Really Need a Systematic Classification of Epilepsies?', *Epilepsia*, 44(1), pp. 12–13.

Bae, J.S., Lee, J.W., Yoo, J.E., Joung, J.G., Yoo, K.H., Koo, H.H., Song, Y.M. and Sung, K.W. (2020) 'Genome-Wide Association Study for the Identification of Novel Genetic Variants Associated with the Risk of Neuroblastoma in Korean Children', *Cancer Res Treat*, 52(4), pp. 1251-1261.

Bahi-Buisson, N., Kaminska, A., Boddaert, N., Rio, M., Afenjar, A., Gerard, M., Giuliano, F., Motte, J., Heron, D., Morel, M.A., Plouin, P., Richelme, C., des Portes, V., Dulac, O., Philippe, C., Chiron, C., Nabbout, R. and Bienvenu, T. (2008a) 'The three stages of epilepsy in patients with CDKL5 mutations', *Epilepsia*, 49(6), pp. 1027-37.

Bahi-Buisson, N., Nectoux, J., Rosas-Vargas, H., Milh, M., Boddaert, N., Girard, B., Cances, C., Ville, D., Afenjar, A., Rio, M., Héron, D., N'Guyen Morel, M.A., Arzimanoglou, A., Philippe, C., Jonveaux, P., Chelly, J. and Bienvenu, T. (2008b) 'Key clinical features to identify girls with CDKL5 mutations', *Brain*, 131(Pt 10), pp. 2647-61.

Bailey, J.N., de Nijs, L., Bai, D., Suzuki, T., Miyamoto, H., Tanaka, M., Patterson, C., Lin, Y.C., Medina, M.T., Alonso, M.E., Serratosa, J.M., Duron, R.M., Nguyen, V.H., Wight, J.E., Martinez-Juarez, I.E., Ochoa, A., Jara-Prado, A., Guilhoto, L., Molina, Y., Yacubian, E.M., Lopez-Ruiz, M., Inoue, Y., Kaneko, S., Hirose, S., Osawa, M., Oguni, H., Fujimoto, S., Grisar, T.M., Stern, J.M., Yamakawa, K., Lakaye, B. and Delgado-Escueta, A.V. (2018) 'Variant Intestinal-Cell Kinase in Juvenile Myoclonic Epilepsy', *N Engl J Med*, 378(11), pp. 1018–1028.

Bailey, J.N., Patterson, C., de Nijs, L., Duron, R.M., Nguyen, V.H., Tanaka, M., Medina, M.T., Jara-Prado, A., Martinez-Juarez, I.E., Ochoa, A., Molina, Y., Suzuki, T., Alonso, M.E., Wight,

J.E., Lin, Y.C., Guilhoto, L., Targas Yacubian, E.M., Machado-Salas, J., Daga, A., Yamakawa, K., Grisar, T.M., Lakaye, B. and Delgado-Escueta, A.V. (2017) 'EFHC1 variants in juvenile myoclonic epilepsy: reanalysis according to NHGRI and ACMG guidelines for assigning disease causality', *Genet Med*, 19(2), pp. 144–156.

Balestrini, S., Chiarello, D., Gogou, M., Silvennoinen, K., Puvirajasinghe, C., Jones, W.D., Reif, P., Klein, K.M., Rosenow, F., Weber, Y.G., Lerche, H., Schubert-Bast, S., Borggraefe, I., Coppola, A., Troisi, S., Møller, R.S., Riva, A., Striano, P., Zara, F., Hemingway, C., Marini, C., Rosati, A., Mei, D., Montomoli, M., Guerrini, R., Cross, J.H. and Sisodiya, S.M. (2021) 'Real-life survey of pitfalls and successes of precision medicine in genetic epilepsies', *J Neurol Neurosurg Psychiatry*, 92(10), pp. 1044-1052.

Banerjee, P.N., Filippi, D. and Hauser, W.A. (2009) 'The descriptive epidemiology of epilepsy—A review', *Epilepsy Res*, 85(1), pp. 31–45.

Baris, H., Bejjani, B.A., Tan, W.H., Coulter, D.L., Martin, J.A., Storm, A.L., Burton, B.K., Saitta, S.C., Gajecka, M., Ballif, B.C., Irons, M.B., Shaffer, L.G. and Kimonis, V.E. (2006) 'Identification of a novel polymorphism--the duplication of the NPHP1 (nephronophthisis 1) gene', *Am J Med Genet A*, 140a(17), pp. 1876-9.

Barrow, E., Alduaij, W., Robinson, L., Shenton, A., Clancy, T., Lalloo, F., Hill, J. and Evans, D.G. (2008) 'Colorectal cancer in HNPCC: cumulative lifetime incidence, survival and tumour distribution. A report of 121 families with proven mutations', *Clin Genet*, 74(3), pp. 233-42.

Bastarache, L. (2021) 'Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS', *Annu Rev Biomed Data Sci*, 4(1), pp. 1-19.

Basu, S.N., Kollu, R. and Banerjee-Basu, S. (2009) 'AutDB: a gene reference resource for autism research', *Nucleic Acids Res*, 37(Database issue), pp. D832-6.

Bayat, A., Bayat, M., Rubboli, G. and Møller, R.S. (2021) 'Epilepsy Syndromes in the First Year of Life and Usefulness of Genetic Testing for Precision Therapy', *Genes (Basel)*, 12(7).

Bayat, A., Fenger, C.D., Techlo, T.R., Højte, A.F., Nørgaard, I., Hansen, T.F., Rubboli, G., Møller, R.S. and Group, D. (2022) 'Impact of Genetic Testing on Therapeutic Decision-Making in Childhood-Onset Epilepsies-a Study in a Tertiary Epilepsy Center', *Neurotherapeutics*, 19(4), pp. 1353-1367.

Beghi, E. and Giussani, G. (2018) 'Aging and the Epidemiology of Epilepsy', *Neuroepidemiology*, 51(3-4), pp. 216-223.

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *J R Stat Soc*, 57(1), pp. 289-300.

Berg, A.T., Berkovic, S.F., Brodie, M.J., Buchhalter, J., Cross, J.H., van Emde Boas, W., Engel, J., French, J., Glauser, T.A., Mathern, G.W., Moshe, S.L., Nordli, D., Plouin, P. and Scheffer, I.E. (2010) 'Revised terminology and concepts for organization of seizures and epilepsies: report of the ILAE Commission on Classification and Terminology, 2005-2009', *Epilepsia*, 51(4), pp. 676-85.

Berg, A.T. and Blackstone, N.W. (2003) 'Of Cabbages and Kings: Perspectives on Classification from the Field of Systematics', *Epilepsia*, 44(1), pp. 8–11.

Berg, A.T., Jallon, P. and Preux, P.M. (2013) 'The epidemiology of seizure disorders in infancy and childhood: definitions and classifications', *Handb Clin Neurol*, 111, pp. 391-8.

Berkovic, S.F., Howell, R.A., Hay, D.A. and Hopper, J.L. (1998) 'Epilepsies in twins: genetics of the major epilepsy syndromes', *Ann Neurol*, 43(4), pp. 435–45.

Beveridge, W.I.B. (1957) *The art of scientific investigation*. [Rev. ed.] edn. New York: Norton.

Beyenburg, S., Stavem, K. and Schmidt, D. (2010) 'Placebo-corrected efficacy of modern antiepileptic drugs for refractory epilepsy: Systematic review and meta-analysis', *Epilepsia*, 51(1), pp. 7-26.

Bielinska, B., Blaydes, S.M., Buiting, K., Yang, T., Krajewska-Walasek, M., Horsthemke, B. and Brannan, C.I. (2000) 'De novo deletions of SNRPN exon 1 in early human and mouse embryos result in a paternal to maternal imprint switch', *Nat Genet*, 25(1), pp. 74-8.

Blake, J.A., Baldarelli, R., Kadin, J.A., Richardson, J.E., Smith, C.L. and Bult, C.J. (2021) 'Mouse Genome Database (MGD): Knowledgebase for mouse-human comparative biology', *Nucleic Acids Res*, 49(D1), pp. D981-d987.

Blume, W.T., Luders, H.O., Mizrahi, E., Tassinari, C., van Emde Boas, W. and Engel, J., Jr. (2001) 'Glossary of descriptive terminology for ictal semiology: report of the ILAE task force on classification and terminology', *Epilepsia*, 42(9), pp. 1212-8.

Bobbili, D.R., Lal, D., May, P., Reinthaler, E.M., Jabbari, K., Thiele, H., Nothnagel, M., Jurkowski, W., Feucht, M., Nurnberg, P., Lerche, H., Zimprich, F., Krause, R., Neubauer, B.A., Reinthaler, E.M., Zimprich, F., Feucht, M., Steinbock, H., Neophytou, B., Geldner, J., Gruber-Sedlmayr, U., Haberlandt, E., Ronen, G.M., Altmuller, J., Lal, D., Nurnberg, P., Sander, T., Thiele, H., Krause, R., May, P., Balling, R., Lerche, H., Neubauer, B.A. and Consortium, E.C. (2018) 'Exome-wide analysis of mutational burden in patients with typical and atypical Rolandic epilepsy', *Eur J Hum Genet*, 26(2), pp. 258–264.

Bodenreider, O. (2004) 'The Unified Medical Language System (UMLS): integrating biomedical terminology', *Nucleic Acids Res*, 32(Database issue), pp. D267-70.

Boerma, R.S., Braun, K.P., van den Broek, M.P., van Berkestijn, F.M., Swinkels, M.E., Hagebeuk, E.O., Lindhout, D., van Kempen, M., Boon, M., Nicolai, J., de Kovel, C.G., Brilstra, E.H. and Koeleman, B.P. (2016) 'Remarkable Phenytoin Sensitivity in 4 Children with SCN8A-related Epilepsy: A Molecular Neuropharmacological Approach', *Neurotherapeutics*, 13(1), pp. 192–7.

Bomba, L., Walter, K. and Soranzo, N. (2017) 'The impact of rare and low-frequency genetic variants in common disease', *Genome Biol*, 18(1), p. 77.

Bonati, M.T., Castronovo, C., Sironi, A., Zimbalatti, D., Bestetti, I., Crippa, M., Novelli, A., Loddo, S., Dentici, M.L., Taylor, J., Devillard, F., Larizza, L. and Finelli, P. (2019) '9q34.3 microduplications lead to neurodevelopmental disorders through EHMT1 overexpression', *Neurogenetics*, 20(3), pp. 145-154.

Bonnet, M., Roche, F., Fagotto-Kaufmann, C., Gazdagh, G., Truong, I., Comunale, F., Barbosa, S., Bonhomme, M., Nafati, N., Hunt, D., Rodriguez, M.P., Chaudhry, A., Shears, D., Madruga, M., Vansenne, F., Curie, A., Kajava, A.V., Baralle, D., Fassier, C., Debant, A. and Schmidt, S. (2023) 'Pathogenic TRIO variants associated with neurodevelopmental disorders perturb the

molecular regulation of TRIO and axon pathfinding in vivo', *Molecular Psychiatry*, 28(4), pp. 1527-1544.

Boriah, S., Chandola, V. and Kumar, V. (2008) 'Similarity Measures for Categorical Data: A Comparative Evaluation', in *Proc 2008 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, pp. 243-254.

Borlot, F., de Almeida, B.I., Combe, S.L., Andrade, D.M., Filloux, F.M. and Myers, K.A. (2019) 'Clinical utility of multigene panel testing in adults with epilepsy and intellectual disability', *Epilepsia*, 60(8), pp. 1661-1669.

Borodulin, K., Tolonen, H., Jousilahti, P., Jula, A., Juolevi, A., Koskinen, S., Kuulasmaa, K., Laatikainen, T., Männistö, S., Peltonen, M., Perola, M., Puska, P., Salomaa, V., Sundvall, J., Virtanen, S.M. and Vartiainen, E. (2018) 'Cohort Profile: The National FINRISK Study', *Int J Epidemiol*, 47(3), pp. 696-696i.

Boßelmann, C.M., Ivaniuk, A., St John, M., Taylor, S.C., Krishnaswamy, G., Milinovich, A., Leu, C., Gupta, A., Pestana-Knight, E.M., Najm, I. and Lal, D. (2024) 'Healthcare utilization and clinical characteristics of genetic epilepsy in electronic health records', *Brain Commun*, 6(2), p. fcae090.

Bouchard, M.J., Dong, Y., McDermott, B.M., Jr., Lam, D.H., Brown, K.R., Shelanski, M., Bellvé, A.R. and Racaniello, V.R. (2000) 'Defects in nuclear and cytoskeletal morphology and mitochondrial localization in spermatozoa of mice lacking nectin-2, a component of cell-cell adherens junctions', *Mol Cell Biol*, 20(8), pp. 2865-73.

Bragin, E., Chatzimichali, E.A., Wright, C.F., Hurles, M.E., Firth, H.V., Bevan, A.P. and Swaminathan, G.J. (2014) 'DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation', *Nucleic Acids Res*, 42(Database issue), pp. D993-d1000.

Brainstorm, C., Anttila, V., Bulik-Sullivan, B., Finucane, H.K., Walters, R.K., Bras, J., Duncan, L., Escott-Price, V., Falcone, G.J., Gormley, P., Malik, R., Patsopoulos, N.A., Ripke, S., Wei, Z., Yu, D., Lee, P.H., Turley, P., Grenier-Boley, B., Chouraki, V., Kamatani, Y., Berr, C., Letenneur, L., Hannequin, D., Amouyel, P., Boland, A., Deleuze, J.F., Duron, E., Vardarajan, B.N., Reitz, C., Goate, A.M., Huentelman, M.J., Kamboh, M.I., Larson, E.B., Rogaeva, E., St George-Hyslop, P., Hakonarson, H., Kukull, W.A., Farrer, L.A., Barnes, L.L., Beach, T.G., Demirci, F.Y., Head, E., Hulette, C.M., Jicha, G.A., Kauwe, J.S.K., Kaye, J.A., Leverenz, J.B., Levey, A.I., Lieberman, A.P., Pankratz, V.S., Poon, W.W., Quinn, J.F., Saykin, A.J., Schneider, L.S., Smith, A.G., Sonnen, J.A., Stern, R.A., Van Deerlin, V.M., Van Eldik, L.J., Harold, D., Russo, G., Rubinsztein, D.C., Bayer, A., Tsolaki, M., Proitsi, P., Fox, N.C., Hampel, H., Owen, M.J., Mead, S., Passmore, P., Morgan, K., Nothen, M.M., Rossor, M., Lupton, M.K., Hoffmann, P., Kornhuber, J., Lawlor, B., McQuillin, A., Al-Chalabi, A., Bis, J.C., Ruiz, A., Boada, M., Seshadri, S., Beiser, A., Rice, K., van der Lee, S.J., De Jager, P.L., Geschwind, D.H., Riemenschneider, M., Riedel-Heller, S., Rotter, J.I., Ransmayr, G., Hyman, B.T., Cruchaga, C., Alegret, M., Winsvold, B., Palta, P., Farh, K.H., Cuenca-Leon, E., Furlotte, N., et al. (2018) 'Analysis of shared heritability in common disorders of the brain', *Science*, 360(6395).

Brandler, W.M., Antaki, D., Gujral, M., Noor, A., Rosanio, G., Chapman, T.R., Barrera, D.J., Lin, G.N., Malhotra, D., Watts, A.C., Wong, L.C., Estabillo, J.A., Gadomski, T.E., Hong, O., Fajardo, K.V., Bhandari, A., Owen, R., Baughn, M., Yuan, J., Solomon, T., Moyzis, A.G., Maile, M.S., Sanders, S.J., Reiner, G.E., Vaux, K.K., Strom, C.M., Zhang, K., Muotri, A.R.,

Akshoomoff, N., Leal, S.M., Pierce, K., Courchesne, E., Iakoucheva, L.M., Corsello, C. and Sebat, J. (2016) 'Frequency and Complexity of De Novo Structural Mutation in Autism', *Am J Hum Genet*, 98(4), pp. 667-79.

Broad Institute (2023) 'Genome Aggregation Database'. 2023-10-12. Available at: https://gnomad.broadinstitute.org/downloads.

Brunklaus, A., Ellis, R., Reavey, E., Forbes, G.H. and Zuberi, S.M. (2012) 'Prognostic, clinical and demographic features in SCN1A mutation-positive Dravet syndrome', *Brain*, 135(Pt 8), pp. 2329–36.

Brunklaus, A., Pérez-Palma, E., Ghanty, I., Xinge, J., Brilstra, E., Ceulemans, B., Chemaly, N., de Lange, I., Depienne, C., Guerrini, R., Mei, D., Møller, R.S., Nabbout, R., Regan, B.M., Schneider, A.L., Scheffer, I.E., Schoonjans, A.-S., Symonds, J.D., Weckhuysen, S., Kattan, M.W., Zuberi, S.M. and Lal, D. (2022) 'Development and Validation of a Prediction Model for Early Diagnosis of SCN1A-Related Epilepsies', *Neurology*, 98(11), pp. e1163-e1174.

Campbell, C., Leu, C., Feng, Y.A., Wolking, S., Moreau, C., Ellis, C., Ganesan, S., Martins, H., Oliver, K., Boothman, I., Benson, K., Molloy, A., Brody, L., Michaud, J.L., Hamdan, F.F., Minassian, B.A., Lerche, H., Scheffer, I.E., Sisodiya, S., Girard, S., Cosette, P., Delanty, N., Lal, D. and Cavalleri, G.L. (2022) 'The role of common genetic variation in presumed monogenic epilepsies', *EBioMedicine*, 81, p. 104098.

Caraballo, R.H., Ganez, L., Santos Cde, L., Espeche, A., Cersosimo, R. and Fejerman, N. (2009) 'Benign infantile seizures with mild gastroenteritis: study of 22 patients', *Seizure*, 18(10), pp. 686-9.

Carpio, A., Salgado, C., DiCapua, D., Fleury, A., Suastegui, R., Giagante, B., Aguilera, L., Quijada, A., Nuñez, L., Villanueva, M., Plascencia-Alvarez, N., Hamamoto Filho, P.T., Piedra, L.M., Silva-Rosas, C. and Kelvin, E.A. (2024) 'Causes and prognosis of adults experiencing a first seizure in adulthood: A pilot cohort study conducted in five countries in Latin America', *Epilepsia Open*.

Cavalleri, G.L., Walley, N.M., Soranzo, N., Mulley, J., Doherty, C.P., Kapoor, A., Depondt, C., Lynch, J.M., Scheffer, I.E., Heils, A., Gehrmann, A., Kinirons, P., Gandhi, S., Satishchandra, P., Wood, N.W., Anand, A., Sander, T., Berkovic, S.F., Delanty, N., Goldstein, D.B. and Sisodiya, S.M. (2007) 'A multicenter study of BRD2 as a risk factor for juvenile myoclonic epilepsy', *Epilepsia*, 48(4), pp. 706–12.

Chabalier, J., Mosser, J. and Burgun, A. (2007) 'A transversal approach to predict gene product networks from ontology-based similarity', *BMC Bioinformatics*, 8(1), p. 235.

Chander, V., Mahmoud, M., Hu, J., Dardas, Z., Grochowski, C.M., Dawood, M., Khayat, M.M., Li, H., Li, S., Jhangiani, S., Korchina, V., Shen, H., Weissenberger, G., Meng, Q., Gingras, M.C., Muzny, D.M., Doddapaneni, H., Posey, J.E., Lupski, J.R., Sabo, A., Murdock, D.R., Sedlazeck, F.J. and Gibbs, R.A. (2022) 'Long read sequencing and expression studies of AHDC1 deletions in Xia-Gibbs syndrome reveal a novel genetic regulatory mechanism', *Hum Mutat*, 43(12), pp. 2033-2053.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M. and Lee, J.J. (2015) 'Second-generation PLINK: rising to the challenge of larger and richer datasets', *Gigascience*, 4, p. 7.

Chen, B., Cheng, M., Hong, S., Liao, S., Ma, J., Li, T. and Jiang, L. (2018a) 'Clinical outcome of recurrent afebrile seizures in children with benign convulsions associated with mild gastroenteritis', *Seizure*, 60, pp. 110-114.

Chen, C.K., Mungall, C.J., Gkoutos, G.V., Doelken, S.C., Kohler, S., Ruef, B.J., Smith, C., Westerfield, M., Robinson, P.N., Lewis, S.E., Schofield, P.N. and Smedley, D. (2012) 'MouseFinder: Candidate disease genes from mouse phenotype data', *Hum Mutat*, 33(5), pp. 858-66.

Chen, D.H., Latimer, C., Yagi, M., Ndugga-Kabuye, M.K., Heigham, E., Jayadev, S., Meabon, J.S., Gomez, C.M., Keene, C.D., Cook, D.G., Raskind, W.H. and Bird, T.D. (2020) 'Heterozygous STUB1 missense variants cause ataxia, cognitive decline, and STUB1 mislocalization', *Neurol Genet*, 6(2), pp. 1-13.

Chen, D.Y., Chen, C.C., Crawford, J.R. and Wang, S.G. (2018b) 'Tumor-related epilepsy: epidemiology, pathogenesis and management', *J Neurooncol*, 139(1), pp. 13-21.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R. and Ma'ayan, A. (2013) 'Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool', *BMC Bioinformatics*, 14, p. 128.

Chen, J., Ou, Y., Yang, Y., Li, W., Xu, Y., Xie, Y. and Liu, Y. (2018c) 'KLHL22 activates amino-acid-dependent mTORC1 signalling to promote tumorigenesis and ageing', *Nature*, 557(7706), pp. 585-589.

Chen, X.Q., Sawa, M., Becker, A., Karachentsev, D., Zuo, X., Rynearson, K.D., Tanzi, R.E. and Mobley, W.C. (2023) 'Retromer Proteins Reduced in Down Syndrome and the Dp16 Model: Impact of APP Dose and Preclinical Studies of a γ-Secretase Modulator', *Ann Neurol*, 94(2), pp. 245-258.

Chen, Z., Brodie, M.J., Liew, D. and Kwan, P. (2018d) 'Treatment Outcomes in Patients With Newly Diagnosed Epilepsy Treated With Established and New Antiepileptic Drugs: A 30-Year Longitudinal Cohort Study', *JAMA Neurol*, 75(3), pp. 279-286.

Cheng, J., Cline, M., Martin, J., Finkelstein, D., Awad, T., Kulp, D. and Siani-Rose, M.A. (2004) 'A knowledge-based clustering algorithm driven by Gene Ontology', *J Biopharm Stat*, 14(3), pp. 687-700.

Chiang, K.L., Huang, C.Y., Hsieh, L.P. and Chang, K.P. (2020) 'A propositional AI system for supporting epilepsy diagnosis based on the 2017 epilepsy classification: Illustrated by Dravet syndrome', *Epilepsy Behav*, 106, p. 107021.

Chiron, C., Marchand, M.C., Tran, A., Rey, E., d'Athis, P., Vincent, J., Dulac, O. and Pons, G. (2000) 'Stiripentol in severe myoclonic epilepsy in infancy: a randomised placebo-controlled syndrome-dedicated trial. STICLO study group', *Lancet*, 356(9242), pp. 1638-42.

Cho, Y.R., Hwang, W., Ramanathan, M. and Zhang, A. (2007) 'Semantic integration to identify overlapping functional modules in protein interaction networks', *BMC Bioinformatics*, 8(1), p. 265.

Ciitizen corporation (2024). Available at: https://www.ciitizen.com/about/ (Accessed: 2024-03-10).

Claes, L., Del-Favero, J., Ceulemans, B., Lagae, L., Van Broeckhoven, C. and De Jonghe, P. (2001) 'De novo mutations in the sodium-channel gene SCN1A cause severe myoclonic epilepsy of infancy', *Am J Hum Genet*, 68(6), pp. 1327-32.

Cleynen, I., Engchuan, W., Hestand, M.S., Heung, T., Holleman, A.M., Johnston, H.R., Monfeuga, T., McDonald-McGinn, D.M., Gur, R.E., Morrow, B.E., Swillen, A., Vorstman, J.A.S., Bearden, C.E., Chow, E.W.C., van den Bree, M., Emanuel, B.S., Vermeesch, J.R., Warren, S.T., Owen, M.J., Chopra, P., Cutler, D.J., Duncan, R., Kotlar, A.V., Mulle, J.G., Voss, A.J., Zwick, M.E., Diacou, A., Golden, A., Guo, T., Lin, J.R., Wang, T., Zhang, Z., Zhao, Y., Marshall, C., Merico, D., Jin, A., Lilley, B., Salmons, H.I., Tran, O., Holmans, P., Pardinas, A., Walters, J.T.R., Demaerel, W., Boot, E., Butcher, N.J., Costain, G.A., Lowther, C., Evers, R., van Amelsvoort, T., van Duin, E., Vingerhoets, C., Breckpot, J., Devriendt, K., Vergaelen, E., Vogels, A., Crowley, T.B., McGinn, D.E., Moss, E.M., Sharkus, R.J., Unolt, M., Zackai, E.H., Calkins, M.E., Gallagher, R.S., Gur, R.C., Tang, S.X., Fritsch, R., Ornstein, C., Repetto, G.M., Breetvelt, E., Duijff, S.N., Fiksinski, A., Moss, H., Niarchou, M., Murphy, K.C., Prasad, S.E., Daly, E.M., Gudbrandsen, M., Murphy, C.M., Murphy, D.G., Buzzanca, A., Fabio, F.D., Digilio, M.C., Pontillo, M., Marino, B., Vicari, S., Coleman, K., Cubells, J.F., Ousley, O.Y., Carmel, M., Gothelf, D., Mekori-Domachevsky, E., Michaelovsky, E., Weinberger, R., Weizman, A., Kushan, L., Jalbrzikowski, M., Armando, M., Eliez, S., Sandini, C., Schneider, M., et al. (2021) 'Genetic contributors to risk of schizophrenia in the presence of a 22q11.2 deletion', *Mol Psychiatry*, 26(8), pp. 4496-4510.

Clinical Genome Resource (2024) 'Summary report'. 06 March 2024. Available at: https://search.clinicalgenome.org/kb/downloads.

Collins, R.L., Glessner, J.T., Porcu, E., Lepamets, M., Brandon, R., Lauricella, C., Han, L., Morley, T., Niestroj, L.M., Ulirsch, J., Everett, S., Howrigan, D.P., Boone, P.M., Fu, J., Karczewski, K.J., Kellaris, G., Lowther, C., Lucente, D., Mohajeri, K., Noukas, M., Nuttle, X., Samocha, K.E., Trinh, M., Ullah, F., Vosa, U., Epi, C., Estonian Biobank Research, T., Hurles, M.E., Aradhya, S., Davis, E.E., Finucane, H., Gusella, J.F., Janze, A., Katsanis, N., Matyakhina, L., Neale, B.M., Sanders, D., Warren, S., Hodge, J.C., Lal, D., Ruderfer, D.M., Meck, J., Magi, R., Esko, T., Reymond, A., Kutalik, Z., Hakonarson, H., Sunyaev, S., Brand, H. and Talkowski, M.E. (2022) 'A cross-disorder dosage sensitivity map of the human genome', *Cell*, 185(16), pp. 3041-3055 e25.

Commission on Classification and Terminology of the International League Against Epilepsy (1981) 'Proposal for revised clinical and electroencephalographic classification of epileptic seizures.', *Epilepsia*, 22(4), pp. 489-501.

Commission on Classification and Terminology of the International League Against Epilepsy (1989) 'Proposal for revised classification of epilepsies and epileptic syndromes.', *Epilepsia*, 30(4), pp. 389–99.

Commission on Terminology of the International League Against Epilepsy (1964) 'A Proposed International Classification of Epileptic Seizures', *Epilepsia*, 5(4), pp. 297—306.

Coppola, A., Cellini, E., Stamberger, H., Saarentaus, E., Cetica, V., Lal, D., Djémié, T., Bartnik-Glaska, M., Ceulemans, B., Helen Cross, J., Deconinck, T., Masi, S., Dorn, T., Guerrini, R., Hoffman-Zacharska, D., Kooy, F., Lagae, L., Lench, N., Lemke, J.R., Lucenteforte, E., Madia, F., Mefford, H.C., Morrogh, D., Nuernberg, P., Palotie, A., Schoonjans, A.S., Striano, P., Szczepanik, E., Tostevin, A., Vermeesch, J.R., Van Esch, H., Van Paesschen, W., Waters, J.J., Weckhuysen, S., Zara, F., De Jonghe, P., Sisodiya, S.M. and

Marini, C. (2019) 'Diagnostic implications of genetic copy number variation in epilepsy plus', *Epilepsia*, 60(4), pp. 689-706.

Coppola, G. (2009) 'Malignant migrating partial seizures in infancy: an epilepsy syndrome of unknown etiology', *Epilepsia*, 50 Suppl 5, pp. 49-51.

Corcoran, L.M., Koentgen, F., Dietrich, W., Veale, M. and Humbert, P.O. (2004) 'All known in vivo functions of the Oct-2 transcription factor require the C-terminal protein domain', *J Immunol*, 172(5), pp. 2962-9.

Core Human Phenotype Ontology/Monarch Initiative Team *The Human Phenotype Ontology*. Available at: https://hpo.jax.org/app/ (Accessed: 2021-01-06).

Couto, F.M., Silva, M.J. and Coutinho, P.M. (2007) 'Measuring semantic similarity between Gene Ontology terms', *Data Knowl Eng*, 61(1), pp. 137-152.

Crawford, K., Xian, J., Helbig, K.L., Galer, P.D., Parthasarathy, S., Lewis-Smith, D., Kaufman, M.C., Fitch, E., Ganesan, S., O'Brien, M., Codoni, V., Ellis, C.A., Conway, L.J., Taylor, D., Krause, R. and Helbig, I. (2021) 'Computational analysis of 10,860 phenotypic annotations in individuals with SCN2A-related disorders', *Genet Med*, 23(7), pp. 1263-1272.

da Silva-Buttkus, P., Spielmann, N., Klein-Rodewald, T., Schütt, C., Aguilar-Pimentel, A., Amarie, O.V., Becker, L., Calzada-Wack, J., Garrett, L., Gerlini, R., Kraiger, M., Leuchtenberger, S., Östereicher, M.A., Rathkolb, B., Sanz-Moreno, A., Stöger, C., Hölter, S.M., Seisenberger, C., Marschall, S., Fuchs, H., Gailus-Durner, V. and Hrabě de Angelis, M. (2023) 'Knockout mouse models as a resource for the study of rare diseases', *Mamm Genome*, 34(2), pp. 244-261.

Dale, T., Downs, J., Wong, K. and Leonard, H. (2021) 'The perceived effects of cannabis products in the management of seizures in CDKL5 Deficiency Disorder', *Epilepsy Behav*, 122, p. 108152.

Danesi, M.A. (1988) 'Seasonal variations in the incidence of photoparoxysmal response to stimulation among photosensitive epileptic patients: evidence from repeated EEG recordings', *J Neurol Neurosurg Psychiatry*, 51(6), pp. 875-7.

Davies, F.C.J., Hope, J.E., McLachlan, F., Marshall, G.F., Kaminioti-Dumont, L., Qarkaxhija, V., Nunez, F., Dando, O., Smith, C., Wood, E., MacDonald, J., Hardt, O. and Abbott, C.M. (2020) 'Recapitulation of the EEF1A2 D252H neurodevelopmental disorder-causing missense mutation in mice reveals a toxic gain of function', *Hum Mol Genet*, 29(10), pp. 1592-1606.

de Kovel, C.G., Pinto, D., de Haan, G.J., Kasteleijn-Nolst Trenite, D.G., Lindhout, D. and Koeleman, B.P. (2007) 'Association analysis of BRD2 (RING3) and epilepsy in a Dutch population', *Epilepsia*, 48(11), pp. 2191–2.

de Kovel, C.G., Trucks, H., Helbig, I., Mefford, H.C., Baker, C., Leu, C., Kluck, C., Muhle, H., von Spiczak, S., Ostertag, P., Obermeier, T., Kleefuss-Lie, A.A., Hallmann, K., Steffens, M., Gaus, V., Klein, K.M., Hamer, H.M., Rosenow, F., Brilstra, E.H., Trenite, D.K., Swinkels, M.E., Weber, Y.G., Unterberger, I., Zimprich, F., Urak, L., Feucht, M., Fuchs, K., Moller, R.S., Hjalgrim, H., De Jonghe, P., Suls, A., Ruckert, I.M., Wichmann, H.E., Franke, A., Schreiber, S., Nurnberg, P., Elger, C.E., Lerche, H., Stephani, U., Koeleman, B.P., Lindhout, D., Eichler, E.E. and Sander, T. (2010) 'Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies', *Brain*, 133(Pt 1), pp. 23-32.

de Lange, I.M., Mulder, F., van 't Slot, R., Sonsma, A.C.M., van Kempen, M.J.A., Nijman, I.J., Ernst, R.F., Knoers, N., Brilstra, E.H. and Koeleman, B.P.C. (2020) 'Modifier genes in SCN1A-related epilepsy syndromes', *Mol Genet Genomic Med*, 8(4), p. e1103.

de Maria, B., Balestrini, S., Mei, D., Melani, F., Pellacani, S., Pisano, T., Rosati, A., Scaturro, G.M., Giordano, L., Cantalupo, G., Fontana, E., Zammarchi, C., Said, E., Leuzzi, V., Mastrangelo, M., Galosi, S., Parrini, E. and Guerrini, R. (2022) 'Expanding the genetic and phenotypic spectrum of CHD2-related disease: From early neurodevelopmental disorders to adult-onset epilepsy', *Am J Med Genet A*, 188(2), pp. 522-533.

de Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., Singh, T., Klei, L., Kosmicki, J., Shih-Chen, F., Aleksic, B., Biscaldi, M., Bolton, P.F., Brownfeld, J.M., Cai, J., Campbell, N.G., Carracedo, A., Chahrour, M.H., Chiocchetti, A.G., Coon, H., Crawford, E.L., Curran, S.R., Dawson, G., Duketis, E., Fernandez, B.A., Gallagher, L., Geller, E., Guter, S.J., Hill, R.S., Ionita-Laza, J., Jimenz Gonzalez, P., Kilpinen, H., Klauck, S.M., Kolevzon, A., Lee, I., Lei, I., Lei, J., Lehtimäki, T., Lin, C.F., Ma'ayan, A., Marshall, C.R., McInnes, A.L., Neale, B., Owen, M.J., Ozaki, N., Parellada, M., Parr, J.R., Purcell, S., Puura, K., Rajagopalan, D., Rehnström, K., Reichenberg, A., Sabo, A., Sachse, M., Sanders, S.J., Schafer, C., Schulte-Rüther, M., Skuse, D., Stevens, C., Szatmari, P., Tammimies, K., Valladares, O., Voran, A., Li-San, W., Weiss, L.A., Willsey, A.J., Yu, T.W., Yuen, R.K., Cook, E.H., Freitag, C.M., Gill, M., Hultman, C.M., Lehner, T., Palotie, A., Schellenberg, G.D., Sklar, P., State, M.W., Sutcliffe, J.S., Walsh, C.A., Scherer, S.W., Zwick, M.E., Barett, J.C., Cutler, D.J., Roeder, K., Devlin, B., Daly, M.J. and Buxbaum, J.D. (2014) 'Synaptic, transcriptional and chromatin genes disrupted in autism', *Nature*, 515(7526), pp. 209-15.

de Vivo, D.C., Trifiletti, R.R., Jacobson, R.I., Ronen, G.M., Behmand, R.A. and Harik, S.I. (1991) 'Defective glucose transport across the blood-brain barrier as a cause of persistent hypoglycorrhachia, seizures, and developmental delay', *N Engl J Med*, 325(10), pp. 703–9.

Delanty, N., White, M., Benson, K., McCormack, M., Heavin, S., Comerford, E., Gangadharan, N., Power, K., Dunleavey, B., El-Naggar, H., Doherty, C., Greally, M., Cavalleri, G. and Fitzsimons, M. (2019) 'Development of a genomics module within an epilepsy-specific electronic health record: Toward genomic medicine in epilepsy care', *Epilepsia*, 60(8), pp. 1670-1677.

Delplanque, J., Devos, D., Huin, V., Genet, A., Sand, O., Moreau, C., Goizet, C., Charles, P., Anheim, M., Monin, M.L., Buée, L., Destée, A., Grolez, G., Delmaire, C., Dujardin, K., Dellacherie, D., Brice, A., Stevanin, G., Strubi-Vuillaume, I., Dürr, A. and Sablonnière, B. (2014) 'TMEM240 mutations cause spinocerebellar ataxia 21 with mental retardation and severe cognitive impairment', *Brain*, 137(Pt 10), pp. 2657-63.

Demarest, S., Marsh, R., Treat, L., Fisher, M.P., Dempsey, A., Junaid, M., Downs, J., Leonard, H., Benke, T. and Morris, M.A. (2022) 'The Lived Experience of Parents' Receiving the Diagnosis of CDKL5 Deficiency Disorder for Their Child', *J Child Neurol*, 37(6), pp. 451-460.

Demarest, S., Pestana-Knight, E.M., Olson, H.E., Downs, J., Marsh, E.D., Kaufmann, W.E., Partridge, C.A., Leonard, H., Gwadry-Sridhar, F., Frame, K.E., Cross, J.H., Chin, R.F.M., Parikh, S., Panzer, A., Weisenberg, J., Utley, K., Jaksha, A., Amin, S., Khwaja, O., Devinsky, O., Neul, J.L., Percy, A.K. and Benke, T.A. (2019a) 'Severity Assessment in CDKL5 Deficiency Disorder', *Pediatr Neurol*, 97, pp. 38-42.

Demarest, S.T., Olson, H.E., Moss, A., Pestana-Knight, E., Zhang, X., Parikh, S., Swanson, L.C., Riley, K.D., Bazin, G.A., Angione, K., Niestroj, L.M., Lal, D., Juarez-Colunga, E. and Benke, T.A. (2019b) 'CDKL5 deficiency disorder: Relationship between genotype, epilepsy, cortical visual impairment, and development', *Epilepsia*, 60(8), pp. 1733-1742.

Deng, Y., Gao, L., Wang, B. and Guo, X. (2015) 'HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology', *PLoS One*, 10(2), p. e0115692.

Devinsky, O., Cross, J.H., Laux, L., Marsh, E., Miller, I., Nabbout, R., Scheffer, I.E., Thiele, E.A., Wright, S. and Cannabidiol in Dravet Syndrome Study, G. (2017) 'Trial of Cannabidiol for Drug-Resistant Seizures in the Dravet Syndrome', *N Engl J Med*, 376(21), pp. 2011–2020.

Devinsky, O., Patel, A.D., Cross, J.H., Villanueva, V., Wirrell, E.C., Privitera, M., Greenwood, S.M., Roberts, C., Checketts, D., VanLandingham, K.E., Zuberi, S.M. and Group, G.S. (2018) 'Effect of Cannabidiol on Drop Seizures in the Lennox-Gastaut Syndrome', *N Engl J Med*, 378(20), pp. 1888–1897.

Dewey, F.E., Gusarova, V., O'Dushlaine, C., Gottesman, O., Trejos, J., Hunt, C., Van Hout, C.V., Habegger, L., Buckler, D., Lai, K.M., Leader, J.B., Murray, M.F., Ritchie, M.D., Kirchner, H.L., Ledbetter, D.H., Penn, J., Lopez, A., Borecki, I.B., Overton, J.D., Reid, J.G., Carey, D.J., Murphy, A.J., Yancopoulos, G.D., Baras, A., Gromada, J. and Shuldiner, A.R. (2016) 'Inactivating Variants in ANGPTL4 and Risk of Coronary Artery Disease', *N Engl J Med*, 374(12), pp. 1123-33.

Dibbens, L.M., Mullen, S., Helbig, I., Mefford, H.C., Bayly, M.A., Bellows, S., Leu, C., Trucks, H., Obermeier, T., Wittig, M., Franke, A., Caglayan, H., Yapici, Z., Consortium, E., Sander, T., Eichler, E.E., Scheffer, I.E., Mulley, J.C. and Berkovic, S.F. (2009) 'Familial and sporadic 15q13.3 microdeletions in idiopathic generalized epilepsy: precedent for disorders with complex inheritance', *Hum Mol Genet*, 18(19), pp. 3626-31.

Dice, L.R. (1945) 'Measures of the Amount of Ecologic Association Between Species', *Ecology*, 26(3), pp. 297-302.

Dickson, J.M., Jacques, R., Reuber, M., Hick, J., Campbell, M.J., Morley, R. and Grünewald, R.A. (2018) 'Emergency hospital care for adults with suspected seizures in the NHS in England 2007–2013: a cross-sectional study', *BMJ Open*, 8(10), p. e023352.

Donahue, M.A., Herman, S.T., Dass, D., Farrell, K., Kukla, A., Abend, N.S., Moura, L., Buchhalter, J.R. and Fureman, B.E. (2021) 'Establishing a learning healthcare system to improve health outcomes for people with epilepsy', *Epilepsy Behav*, 117, p. 107805.

Donnan, A.M., Schneider, A.L., Russ-Hall, S., Churilov, L. and Scheffer, I.E. (2023) 'Rates of Status Epilepticus and Sudden Unexplained Death in Epilepsy in People With Genetic Developmental and Epileptic Encephalopathies', *Neurology*, 100(16), pp. e1712-e1722.

Doyle, A.J., Doyle, J.J., Bessling, S.L., Maragh, S., Lindsay, M.E., Schepers, D., Gillis, E., Mortier, G., Homfray, T., Sauls, K., Norris, R.A., Huso, N.D., Leahy, D., Mohr, D.W., Caulfield, M.J., Scott, A.F., Destrée, A., Hennekam, R.C., Arn, P.H., Curry, C.J., Van Laer, L., McCallion, A.S., Loeys, B.L. and Dietz, H.C. (2012) 'Mutations in the TGF-β repressor SKI cause Shprintzen-Goldberg syndrome with aortic aneurysm', *Nat Genet*, 44(11), pp. 1249-54.

Dravet, C. (2011) 'The core Dravet syndrome phenotype', *Epilepsia*, 52 Suppl 2, pp. 3-9.

Du, Y., Liu, Z., You, L., Hou, P., Ren, X., Jiao, T., Zhao, W., Li, Z., Shu, H., Liu, C. and Zhao, Y. (2017) 'Pancreatic Cancer Progression Relies upon Mutant p53-Induced Oncogenic Signaling Mediated by NOP14', *Cancer Res*, 77(10), pp. 2661-2673.

Durner, M., Keddache, M.A., Tomasini, L., Shinnar, S., Resor, S.R., Cohen, J., Harden, C., Moshe, S.L., Rosenbaum, D., Kang, H., Ballaban-Gil, K., Hertz, S., Labar, D.R., Luciano, D., Wallace, S., Yohai, D., Klotz, I., Dicker, E. and Greenberg, D.A. (2001) 'Genome scan of idiopathic generalized epilepsy: evidence for major susceptibility gene and modifying genes influencing the seizure type', *Ann Neurol*, 49(3), pp. 328-35.

Ekblom, A.G., Laurell, T. and Arner, M. (2010) 'Epidemiology of Congenital Upper Limb Anomalies in 562 Children Born in 1997 to 2007: A Total Population Study from Stockholm, Sweden', *The Journal of Hand Surgery*, 35(11), pp. 1742-1754.

Ellis, C.A., Churilov, L., Epstein, M.P., Xie, S.X., Bellows, S.T., Ottman, R., Berkovic, S.F. and Consortium, E.K. (2019) 'Epilepsy in families: Age at onset is a familial trait, independent of syndrome', *Ann Neurol*, 86(1), pp. 91-98.

eMerge (2007). Available at: https://emerge-network.org (Accessed: 2022-01-06).

Engel, J., Jr. (1998) 'Classifications of the International League Against Epilepsy: time for reappraisal', *Epilepsia*, 39(9), pp. 1014–7.

Engel, J., Jr. (2001) 'A proposed diagnostic scheme for people with epileptic seizures and with epilepsy: report of the ILAE Task Force on Classification and Terminology', *Epilepsia*, 42(6), pp. 796-803.

Engel, J., Jr. (2003) 'Reply to “Of Cabbages and Kings: Some Considerations on Classifications, Diagnostic Schemes, Semiology, and Concepts”', *Epilepsia*, 44(1), pp. 4–5.

Engel, J., Jr. (2006) 'Report of the ILAE classification core group', *Epilepsia*, 47(9), pp. 1558–68.

Epi4K Consortium (2017) 'Phenotypic analysis of 303 multiplex families with common epilepsies', *Brain*, 140(8), pp. 2144-2156.

Epi4K Consortium (2019) 'Quantitative analysis of phenotypic elements augments traditional electroclinical classification of common familial epilepsies', *Epilepsia*, 60(11), pp. 2194-2203.

Epi4K Consortium and Epilepsy Phenome Genome Project (2013) 'De novo mutations in epileptic encephalopathies', *Nature*, 501(7466), pp. 217-21.

Epi4K Consortium and Epilepsy Phenome Genome Project (2017) 'Ultra-rare genetic variation in common epilepsies: a case-control sequencing study', *Lancet Neurol*, 16(2), pp. 135-143.

Epi25 Collaborative (2019) 'Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals', *Am J Hum Genet*, 105(2), pp. 267-282.

Epi25 Collaborative (2020). Available at: http://epi-25.org/epi25-forms (Accessed: 2022).

Epi25 Collaborative (2021) 'Sub-genic intolerance, ClinVar, and the epilepsies: A whole-exome sequencing study of 29,165 individuals', *Am J Hum Genet*, 108(6), pp. 965-982.

Epi25 Collaborative (2023) 'Shared and distinct ultra-rare genetic risk for diverse epilepsies: A whole-exome sequencing study of 54,423 individuals across multiple genetic ancestries', *medRxiv*.

Epilepsy Phenome Genome Project and Epi4K Consortium (2015) 'Copy number variant analysis from exome data in 349 patients with epileptic encephalopathy', *Ann Neurol*, 78(2), pp. 323-8.

Epilepsy Research Institute UK (2021) *UK Epilepsy Priority Setting Partnership (PSP)*. Available at: https://epilepsy-institute.org.uk/eri/about-epilepsy/uk-epilepsy-psp-top-ten/. .

Escayg, A., MacDonald, B.T., Meisler, M.H., Baulac, S., Huberfeld, G., An-Gourfinkel, I., Brice, A., LeGuern, E., Moulard, B., Chaigne, D., Buresi, C. and Malafosse, A. (2000) 'Mutations of SCN1A, encoding a neuronal sodium channel, in two families with GEFS+2', *Nat Genet*, 24(4), pp. 343-5.

EUROCAT (2024) *European Platform on Rare Disease Registration*. Available at: https://eu-rd-platform.jrc.ec.europa.eu/eurocat/eurocat-data/prevalence_en (Accessed: 2024-12-06).

Everitt, A.D. and Sander, J.W. (1999) 'Classification of the epilepsies: time for a change? A critical review of the International Classification of the Epilepsies and Epileptic Syndromes (ICEES) and its usefulness in clinical practice and epidemiological studies of epilepsy', *Eur Neurol*, 42(1), pp. 1–10.

Faivre-Sarrailh, C. and Rougon, G. (1997) 'Axonal molecules of the immunoglobulin superfamily bearing a GPI anchor: their role in controlling neurite outgrowth', *Mol Cell Neurosci*, 9(2), pp. 109-15.

Falcon, S. and Gentleman, R. (2006) 'Using GOstats to test gene lists for GO term association', *Bioinformatics*, 23(2), pp. 257-258.

Fehr, S., Downs, J., Ho, G., de Klerk, N., Forbes, D., Christodoulou, J., Williams, S. and Leonard, H. (2016a) 'Functional abilities in children and adults with the CDKL5 disorder', *Am J Med Genet A*, 170(11), pp. 2860-2869.

Fehr, S., Leonard, H., Ho, G., Williams, S., de Klerk, N., Forbes, D., Christodoulou, J. and Downs, J. (2015) 'There is variability in the attainment of developmental milestones in the CDKL5 disorder', *J Neurodev Disord*, 7(1), p. 2.

Fehr, S., Wilson, M., Downs, J., Williams, S., Murgia, A., Sartori, S., Vecchi, M., Ho, G., Polli, R., Psoni, S., Bao, X., de Klerk, N., Leonard, H. and Christodoulou, J. (2013) 'The CDKL5 disorder is an independent clinical entity associated with early-onset encephalopathy', *Eur J Hum Genet*, 21(3), pp. 266-73.

Fehr, S., Wong, K., Chin, R., Williams, S., de Klerk, N., Forbes, D., Krishnaraj, R., Christodoulou, J., Downs, J. and Leonard, H. (2016b) 'Seizure variables and their relationship to genotype and functional abilities in the CDKL5 disorder', *Neurology*, 87(21), pp. 2206-2213.

Fink, J.K. (2013) 'Hereditary spastic paraplegia: clinico-pathologic features and emerging molecular mechanisms', *Acta Neuropathol*, 126(3), pp. 307-28.

Firth, H.V., Wright, C.F. and Study, D.D.D. (2011) 'The Deciphering Developmental Disorders (DDD) study', *Dev Med Child Neurol*, 53(8), pp. 702-3.

Fisher, R.S. (2003) 'Editor's Introduction: Cabbages and Kings in the Classification of Seizures and the Epilepsies', *Epilepsia*, 44(1), pp. 1–1.

Fisher, R.S. (2017) 'An overview of the 2017 ILAE operational classification of seizure types', *Epilepsy Behav*, 70(Pt A), pp. 271-273.

Fisher, R.S., Acevedo, C., Arzimanoglou, A., Bogacz, A., Cross, J.H., Elger, C.E., Engel, J., Jr., Forsgren, L., French, J.A., Glynn, M., Hesdorffer, D.C., Lee, B.I., Mathern, G.W., Moshe, S.L., Perucca, E., Scheffer, I.E., Tomson, T., Watanabe, M. and Wiebe, S. (2014) 'ILAE official report: a practical clinical definition of epilepsy', *Epilepsia*, 55(4), pp. 475-82.

Fisher, R.S., Cross, H., D'Souza, C., French, J.A., Haut, S., Higurashi, N., Hirsch, E., Jansen, F.E., Peltola, J., Moshe, S.L., Perucca, E., Lagae, L., Roulet-Perez, E., Schulze-Bonhage, A., Scheffer, I.E., Somerville, E., Sperling, M.R., Wiebe, S., Yacubian, E.M. and Zuberi, S. (2019) '2017 International League Against Epilepsy classifications of seizures and epilepsy are steps in the right direction', *Epilepsia*, 60(6), pp. 1040-1044.

Fisher, R.S., Cross, J.H., D'Souza, C., French, J.A., Haut, S.R., Higurashi, N., Hirsch, E., Jansen, F.E., Lagae, L., Moshe, S.L., Peltola, J., Roulet Perez, E., Scheffer, I.E., Schulze-Bonhage, A., Somerville, E., Sperling, M., Yacubian, E.M. and Zuberi, S.M. (2017a) 'Instruction manual for the ILAE 2017 operational classification of seizure types', *Epilepsia*, 58(4), pp. 531-542.

Fisher, R.S., Cross, J.H., French, J.A., Higurashi, N., Hirsch, E., Jansen, F.E., Lagae, L., Moshe, S.L., Peltola, J., Roulet Perez, E., Scheffer, I.E. and Zuberi, S.M. (2017b) 'Operational classification of seizure types by the International League Against Epilepsy: Position Paper of the ILAE Commission for Classification and Terminology', *Epilepsia*, 58(4), pp. 522-530.

Fisher, R.S., van Emde Boas, W., Blume, W., Elger, C., Genton, P., Lee, P. and Engel, J., Jr. (2005) 'Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE)', *Epilepsia*, 46(4), pp. 470–2.

Fitzgerald, M.P., Kaufman, M.C., Massey, S.L., Fridinger, S., Prelack, M., Ellis, C., Ortiz-Gonzalez, X., Fried, L.E., DiGiovine, M.P., Collaborative, C.P.E.P., Melamed, S., Malcolm, M., Banwell, B., Stephenson, D., Witzman, S.M., Gonzalez, A., Dlugos, D., Kessler, S.K., Goldberg, E.M., Abend, N.S. and Helbig, I. (2021) 'Assessing seizure burden in pediatric epilepsy using an electronic medical record-based tool through a common data element approach', *Epilepsia*, 62(7), pp. 1617-1628.

Fitzsimons, M., Dunleavy, B., O'Byrne, P., Dunne, M., Grimson, J., Kalra, D., Normand, C. and Delanty, N. (2013) 'Assessing the quality of epilepsy care with an electronic patient record', *Seizure*, 22(8), pp. 604-10.

Fonferko-Shadrach, B., Lacey, A.S., Roberts, A., Akbari, A., Thompson, S., Ford, D.V., Lyons, R.A., Rees, M.I. and Pickrell, W.O. (2019) 'Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the ExECT (extraction of epilepsy clinical text) system', *BMJ Open*, 9(4), p. e023232.

Fortin, O., Vincelette, C., Chénier, S., Ghais, A., Shevell, M.I., Simard-Tremblay, E. and Myers, K.A. (2020) 'Copy number variation in genetic epilepsy with febrile seizures plus', *Eur J Paediatr Neurol*, 27, pp. 111-115.

Fowkes, R., Elwan, M., Akay, E., Mitchell, C.J., Thomas, R.H. and Lewis-Smith, D. (2022) 'A review of the clinical spectrum of BRAT1 disorders and case of developmental and epileptic encephalopathy surviving into adulthood', *Epilepsy Behav Rep*, 19, p. 100549.

Freiman, S., Hauser, W.A., Rider, F., Gulyaeva, N. and Guekht, A. (2024) 'Post-stroke epilepsy: From clinical predictors to possible mechanisms', *Epilepsy Res*, 199, p. 107282.

French, J.A. (2014) 'ILAE Classification Redux: Ready for Prime Time?', *Epilepsy Curr*, 14(2), pp. 84–5.

Fry, A.E., Fawcett, K.A., Zelnik, N., Yuan, H., Thompson, B.A.N., Shemer-Meiri, L., Cushion, T.D., Mugalaasi, H., Sims, D., Stoodley, N., Chung, S.K., Rees, M.I., Patel, C.V., Brueton, L.A., Layet, V., Giuliano, F., Kerr, M.P., Banne, E., Meiner, V., Lerman-Sagie, T., Helbig, K.L., Kofman, L.H., Knight, K.M., Chen, W., Kannan, V., Hu, C., Kusumoto, H., Zhang, J., Swanger, S.A., Shaulsky, G.H., Mirzaa, G.M., Muir, A.M., Mefford, H.C., Dobyns, W.B., Mackenzie, A.B., Mullins, J.G.L., Lemke, J.R., Bahi-Buisson, N., Traynelis, S.F., Iago, H.F. and Pilz, D.T. (2018) 'De novo mutations in GRIN1 cause extensive bilateral polymicrogyria', *Brain*, 141(3), pp. 698-712.

Gahl, W.A., Mulvihill, J.J., Toro, C., Markello, T.C., Wise, A.L., Ramoni, R.B., Adams, D.R., Tifft, C.J. and Udn (2016) 'The NIH Undiagnosed Diseases Program and Network: Applications to modern medicine', *Mol Genet Metab*, 117(4), pp. 393-400.

Galer, P.D., Ganesan, S., Lewis-Smith, D., McKeown, S.E., Pendziwiat, M., Helbig, K.L., Ellis, C.A., Rademacher, A., Smith, L., Poduri, A., Seiffert, S., von Spiczak, S., Muhle, H., van Baalen, A., Group, N.S., Investigators, E., Consortium, E.-R.E.S., Genomics, R., Innovation, N., Thomas, R.H., Krause, R., Weber, Y. and Helbig, I. (2020) 'Semantic Similarity Analysis Reveals Robust Gene-Disease Relationships in Developmental and Epileptic Encephalopathies', *Am J Hum Genet*, 107(4), pp. 683-697.

Ganesan, S., Galer, P.D., Helbig, K.L., McKeown, S.E., O'Brien, M., Gonzalez, A.K., Felmeister, A.S., Khankhanian, P., Ellis, C.A. and Helbig, I. (2020) 'A longitudinal footprint of genetic epilepsies using automated electronic medical record interpretation', *Genet Med*, 22(12), pp. 2060-2070.

Ganna, A., Satterstrom, F.K., Zekavat, S.M., Das, I., Kurki, M.I., Churchhouse, C., Alfoldi, J., Martin, A.R., Havulinna, A.S., Byrnes, A., Thompson, W.K., Nielsen, P.R., Karczewski, K.J., Saarentaus, E., Rivas, M.A., Gupta, N., Pietiläinen, O., Emdin, C.A., Lescai, F., Bybjerg-Grauholm, J., Flannick, J., Mercader, J.M., Udler, M., Laakso, M., Salomaa, V., Hultman, C., Ripatti, S., Hämäläinen, E., Moilanen, J.S., Körkkö, J., Kuismin, O., Nordentoft, M., Hougaard, D.M., Mors, O., Werge, T., Mortensen, P.B., MacArthur, D., Daly, M.J., Sullivan, P.F., Locke, A.E., Palotie, A., Børglum, A.D., Kathiresan, S. and Neale, B.M. (2018) 'Quantifying the Impact of Rare and Ultra-rare Coding Variation across the Phenotypic Spectrum', *Am J Hum Genet*, 102(6), pp. 1204-1211.

Gao, H., Sander, J.W., Xiao, Y., Zhang, Y. and Zhou, D. (2018) 'A comparison between the 1981 and 2017 International League Against Epilepsy classification of seizure types based on an outpatient setting', *Epileptic Disord*, 20(4), pp. 257-264.

Garber, H.R., Raghavendra, A.S., Lehner, M., Qiao, W., Gutierrez-Barrera, A.M., Tripathy, D., Arun, B. and Ibrahim, N.K. (2022) 'Incidence and impact of brain metastasis in patients with hereditary BRCA1 or BRCA2 mutated invasive breast cancer', *NPJ Breast Cancer*, 8(1), p. 46.

Gargano, M.A., Matentzoglu, N., Coleman, B., Addo-Lartey, E.B., Anagnostopoulos, A.V., Anderton, J., Avillach, P., Bagley, A.M., Bakštein, E., Balhoff, J.P., Baynam, G., Bello, S.M., Berk, M., Bertram, H., Bishop, S., Blau, H., Bodenstein, D.F., Botas, P., Boztug, K., Čady, J., Callahan, T.J., Cameron, R., Carbon, S.J., Castellanos, F., Caufield, J.H., Chan, L.E., Chute, C.G., Cruz-Rojo, J., Dahan-Oliel, N., Davids, J.R., de Dieuleveult, M., de Souza, V., de Vries, B.B.A., de Vries, E., DePaulo, J.R., Derfalvi, B., Dhombres, F., Diaz-Byrd, C., Dingemans, A.J.M., Donadille, B., Duyzend, M., Elfeky, R., Essaid, S., Fabrizzi, C., Fico, G., Firth, H.V., Freudenberg-Hua, Y., Fullerton, J.M., Gabriel, D.L., Gilmour, K., Giordano, J., Goes, F.S., Moses, R.G., Green, I., Griese, M., Groza, T., Gu, W., Guthrie, J., Gyori, B., Hamosh, A., Hanauer, M., Hanušová, K., He, Y.O., Hegde, H., Helbig, I., Holasová, K., Hoyt, C.T., Huang, S., Hurwitz, E., Jacobsen, J.O.B., Jiang, X., Joseph, L., Keramatian, K., King, B., Knoflach, K., Koolen, D.A., Kraus, M.L., Kroll, C., Kusters, M., Ladewig, M.S., Lagorce, D., Lai, M.C., Lapunzina, P., Laraway, B., Lewis-Smith, D., Li, X., Lucano, C., Majd, M., Marazita, M.L., Martinez-Glez, V., McHenry, T.H., McInnis, M.G., McMurry, J.A., Mihulová, M., Millett, C.E., Mitchell, P.B., Moslerová, V., Narutomi, K., Nematollahi, S., Nevado, J., et al. (2023) 'The Human Phenotype Ontology in 2024: phenotypes around the world', *Nucleic Acids Res*.

Gastaut, H. (1969) 'Classification of the epilepsies. Proposal for an international classification', *Epilepsia*, 10, pp. Suppl:14–21.

Geisinger Healthcare System (2021) *MyCode Community Healthcare Initiative*. Available at: https://www.geisinger.org/precision-health/mycode (Accessed: 2022-01-06).

Genomics England (2024a) 'Genomic Medicine Service Rare Disease Virtual Gene Panel: Genetic Epilepsy Syndromes R59 (version 4.178)'. 06 March 2024. Genomics England,. Available at: https://nhsgms-panelapp.genomicsengland.co.uk/panels/402/v4.0 (Accessed: 21 Apr 2021).

Genomics England (2024b) 'Genomic Medicine Service Rare Disease Virtual Gene Panel: Intellectual Disability R29 (version 5.493)'. 06 March 2024. Genomics England,. Available at: https://nhsgms-panelapp.genomicsengland.co.uk/panels/285/v5.0 (Accessed: 21 Apr 2021).

Gimelli, G., Pujana, M.A., Patricelli, M.G., Russo, S., Giardino, D., Larizza, L., Cheung, J., Armengol, L., Schinzel, A., Estivill, X. and Zuffardi, O. (2003) 'Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions', *Hum Mol Genet*, 12(8), pp. 849-58.

Girdea, M., Dumitriu, S., Fiume, M., Bowdin, S., Boycott, K.M., Chenier, S., Chitayat, D., Faghfoury, H., Meyn, M.S., Ray, P.N., So, J., Stavropoulos, D.J. and Brudno, M. (2013) 'PhenoTips: patient phenotyping software for clinical and research use', *Hum Mutat*, 34(8), pp. 1057-65.

Glauser, T.A., Cnaan, A., Shinnar, S., Hirtz, D.G., Dlugos, D., Masur, D., Clark, P.O., Adamson, P.C. and Childhood Absence Epilepsy Study, T. (2013) 'Ethosuximide, valproic acid, and lamotrigine in childhood absence epilepsy: initial monotherapy outcomes at 12 months', *Epilepsia*, 54(1), pp. 141-55.

Global Burden of Disease Epilepsy Collaborators (2019) 'Global, regional, and national burden of epilepsy, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016', *Lancet Neurol*, 18(4), pp. 357-375.

Golzio, C., Willer, J., Talkowski, M.E., Oh, E.C., Taniguchi, Y., Jacquemont, S., Reymond, A., Sun, M., Sawa, A., Gusella, J.F., Kamiya, A., Beckmann, J.S. and Katsanis, N. (2012) 'KCTD13

is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant', *Nature*, 485(7398), pp. 363-7.

Gong, X., Jiang, J., Duan, Z. and Lu, H. (2018) 'A new method to measure the semantic similarity from query phenotypic abnormalities to diseases based on the human phenotype ontology', *BMC Bioinformatics*, 19(Suppl 4), p. 162.

González, J.R., Cáceres, A., Esko, T., Cuscó, I., Puig, M., Esnaola, M., Reina, J., Siroux, V., Bouzigon, E., Nadif, R., Reinmaa, E., Milani, L., Bustamante, M., Jarvis, D., Antó, J.M., Sunyer, J., Demenais, F., Kogevinas, M., Metspalu, A., Cáceres, M. and Pérez-Jurado, L.A. (2014) 'A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity', *Am J Hum Genet*, 94(3), pp. 361-72.

Gorlov, I.P., Gorlova, O.Y., Frazier, M.L., Spitz, M.R. and Amos, C.I. (2011) 'Evolutionary evidence of the effect of rare variants on disease etiology', *Clin Genet*, 79(3), pp. 199-206.

Gramm, M., Leu, C., Perez-Palma, E., Ferguson, L., Jehi, L., Daly, M.J., Najm, I.M., Busch, R.M. and Lal, D. (2020) 'Polygenic risk heterogeneity among focal epilepsies', *Epilepsia*, 61(11), pp. e179-e185.

Greenberg, D.A., Cayanis, E., Strug, L., Marathe, S., Durner, M., Pal, D.K., Alvin, G.B., Klotz, I., Dicker, E., Shinnar, S., Bromfield, E.B., Resor, S., Cohen, J., Moshe, S.L., Harden, C. and Kang, H. (2005) 'Malic enzyme 2 may underlie susceptibility to adolescent-onset idiopathic generalized epilepsy', *Am J Hum Genet*, 76(1), pp. 139-46.

Greenberg, D.A., Delgado-Escueta, A.V., Maldonado, H.M. and Widelitz, H. (1988a) 'Segregation analysis of juvenile myoclonic epilepsy', *Genet Epidemiol*, 5(2), pp. 81-94.

Greenberg, D.A., Delgado-Escueta, A.V., Widelitz, H., Sparkes, R.S., Treiman, L., Maldonado, H.M., Park, M.S. and Terasaki, P.I. (1988b) 'Juvenile myoclonic epilepsy (JME) may be linked to the BF and HLA loci on human chromosome 6', *Am J Med Genet*, 31(1), pp. 185-92.

Greenberg, D.A. and Stewart, W.C. (2012) 'How should we be searching for genes for common epilepsy? A critique and a prescription', *Epilepsia*, 53 Suppl 4(0 4), pp. 72-80.

Greenberg, D.A. and Subaran, R. (2011) 'Blinders, phenotype, and fashionable genetic analysis: a critical examination of the current state of epilepsy genetic studies', *Epilepsia*, 52(1), pp. 1-9.

Greene, D., BioResource, N., Richardson, S. and Turro, E. (2016) 'Phenotype Similarity Regression for Identifying the Genetic Determinants of Rare Diseases', *Am J Hum Genet*, 98(3), pp. 490-499.

Greene, D., Richardson, S. and Turro, E. (2017) 'ontologyX: a suite of R packages for working with ontological data', *Bioinformatics*, 33(7), pp. 1104-1106.

Grinspan, Z.M., Patel, A.D., Shellhaas, R.A., Berg, A.T., Axeen, E.T., Bolton, J., Clarke, D.F., Coryell, J., Gaillard, W.D., Goodkin, H.P., Koh, S., Kukla, A., Mbwana, J.S., Morgan, L.A., Singhal, N.S., Storey, M.M., Yozawitz, E.G., Abend, N.S., Fitzgerald, M.P., Fridinger, S.E., Helbig, I., Massey, S.L., Prelack, M.S., Buchhalter, J. and Pediatric Epilepsy Learning Healthcare, S. (2021) 'Design and implementation of electronic health record common data elements for pediatric epilepsy: Foundations for a learning health care system', *Epilepsia*, 62(1), pp. 198-216.

Grootjen, L.N., Juriaans, A.F., Kerkhof, G.F. and Hokken-Koelega, A.C.S. (2022) 'Atypical 15q11.2-q13 Deletions and the Prader-Willi Phenotype', *J Clin Med*, 11(15).

Gruber, T.R. (1993) 'A translation approach to portable ontology specifications', *Knowledge Acquisition*, 5(2), pp. 199-220.

Gudkova, D., Dergai, O., Praz, V. and Herr, W. (2019) 'HCF-2 inhibits cell proliferation and activates differentiation-gene expression programs', *Nucleic Acids Res*, 47(11), pp. 5792-5808.

Guo, X., Liu, R., Shriver, C.D., Hu, H. and Liebman, M.N. (2006) 'Assessing semantic similarity measures for the characterization of human regulatory pathways', *Bioinformatics*, 22(8), pp. 967-73.

Gusarova, V., O'Dushlaine, C., Teslovich, T.M., Benotti, P.N., Mirshahi, T., Gottesman, O., Van Hout, C.V., Murray, M.F., Mahajan, A., Nielsen, J.B., Fritsche, L., Wulff, A.B., Gudbjartsson, D.F., Sjogren, M., Emdin, C.A., Scott, R.A., Lee, W.J., Small, A., Kwee, L.C., Dwivedi, O.P., Prasad, R.B., Bruse, S., Lopez, A.E., Penn, J., Marcketta, A., Leader, J.B., Still, C.D., Kirchner, H.L., Mirshahi, U.L., Wardeh, A.H., Hartle, C.M., Habegger, L., Fetterolf, S.N., Tusie-Luna, T., Morris, A.P., Holm, H., Steinthorsdottir, V., Sulem, P., Thorsteinsdottir, U., Rotter, J.I., Chuang, L.M., Damrauer, S., Birtwell, D., Brummett, C.M., Khera, A.V., Natarajan, P., Orho-Melander, M., Flannick, J., Lotta, L.A., Willer, C.J., Holmen, O.L., Ritchie, M.D., Ledbetter, D.H., Murphy, A.J., Borecki, I.B., Reid, J.G., Overton, J.D., Hansson, O., Groop, L., Shah, S.H., Kraus, W.E., Rader, D.J., Chen, Y.I., Hveem, K., Wareham, N.J., Kathiresan, S., Melander, O., Stefansson, K., Nordestgaard, B.G., Tybjaerg-Hansen, A., Abecasis, G.R., Altshuler, D., Florez, J.C., Boehnke, M., McCarthy, M.I., Yancopoulos, G.D., Carey, D.J., Shuldiner, A.R., Baras, A., Dewey, F.E. and Gromada, J. (2018) 'Genetic inactivation of ANGPTL4 improves glucose homeostasis and is associated with reduced risk of diabetes', *Nat Commun*, 9(1), p. 2252.

Gustavsson, E.K., Guella, I., Trinh, J., Szu-Tu, C., Rajput, A., Rajput, A.H., Steele, J.C., McKeown, M., Jeon, B.S., Aasly, J.O. and Farrer, M.J. (2015) 'Genetic variability of the retromer cargo recognition complex in parkinsonism', *Mov Disord*, 30(4), pp. 580-4.

Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., Daly, M.J., Price, A.L., Pritchard, J.K., Sharp, A.J. and Erlich, Y. (2016) 'Abundant contribution of short tandem repeats to gene expression variation in humans', *Nat Genet*, 48(1), pp. 22-9.

Hagberg, B., Aicardi, J., Dias, K. and Ramos, O. (1983) 'A progressive syndrome of autism, dementia, ataxia, and loss of purposeful hand use in girls: Rett's syndrome: report of 35 cases', *Ann Neurol*, 14(4), pp. 471-9.

Hagebeuk, E.E., van den Bossche, R.A. and de Weerd, A.W. (2013) 'Respiratory and sleep disorders in female children with atypical Rett syndrome caused by mutations in the CDKL5 gene', *Dev Med Child Neurol*, 55(5), pp. 480-4.

Hannes, F., Hammond, P., Quarrell, O., Fryns, J.P., Devriendt, K. and Vermeesch, J.R. (2012) 'A microdeletion proximal of the critical deletion region is associated with mild Wolf-Hirschhorn syndrome', *Am J Med Genet A*, 158a(5), pp. 996-1004.

Harden, C., Tomson, T., Gloss, D., Buchhalter, J., Cross, J.H., Donner, E., French, J.A., Gil-Nagel, A., Hesdorffer, D.C., Smithson, W.H., Spitz, M.C., Walczak, T.S., Sander, J.W. and Ryvlin, P. (2017) 'Practice Guideline Summary: Sudden Unexpected Death in Epilepsy Incidence Rates and Risk Factors: Report of the Guideline Development, Dissemination, and

Implementation Subcommittee of the American Academy of Neurology and the American Epilepsy Society', *Epilepsy Curr*, 17(3), pp. 180-187.

Harms, F.L., Kloth, K., Bley, A., Denecke, J., Santer, R., Lessel, D., Hempel, M. and Kutsche, K. (2018) 'Activating Mutations in PAK1, Encoding p21-Activated Kinase 1, Cause a Neurodevelopmental Disorder', *Am J Hum Genet*, 103(4), pp. 579-591.

Harris, P.A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N. and Conde, J.G. (2009) 'Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support', *J Biomed Inform*, 42(2), pp. 377-81.

Hasan, T.F. and Tatum, W.O. (2021) 'When should we obtain a routine EEG while managing people with epilepsy?', *Epilepsy Behav Rep*, 16, p. 100454.

Hashimoto, T., Maekawa, S. and Miyata, S. (2009) 'IgLON cell adhesion molecules regulate synaptogenesis in hippocampal neurons', *Cell Biochem Funct*, 27(7), pp. 496-8.

Hauser, W.A., Rich, S.S., Lee, J.R., Annegers, J.F. and Anderson, V.E. (1998) 'Risk of recurrent seizures after two unprovoked seizures', *N Engl J Med*, 338(7), pp. 429–34.

Health, R.o.T.G.-B.R.f. and Board on Health Sciences Policy Institute of Medicine (2015) 'The National Academies Collection: Reports funded by National Institutes of Health', in *Genomics-Enabled Learning Health Care Systems: Gathering and Using Genomic Information to Improve Patient Care and Research: Workshop Summary*. Washington (DC): National Academies Press (US)

Copyright 2015 by the National Academy of Sciences. All rights reserved.

Heinzen, E.L., Radtke, R.A., Urban, T.J., Cavalleri, G.L., Depondt, C., Need, A.C., Walley, N.M., Nicoletti, P., Ge, D., Catarino, C.B., Duncan, J.S., Kasperaviciute, D., Tate, S.K., Caboclo, L.O., Sander, J.W., Clayton, L., Linney, K.N., Shianna, K.V., Gumbs, C.E., Smith, J., Cronin, K.D., Maia, J.M., Doherty, C.P., Pandolfo, M., Leppert, D., Middleton, L.T., Gibson, R.A., Johnson, M.R., Matthews, P.M., Hosford, D., Kalviainen, R., Eriksson, K., Kantanen, A.M., Dorn, T., Hansen, J., Kramer, G., Steinhoff, B.J., Wieser, H.G., Zumsteg, D., Ortega, M., Wood, N.W., Huxley-Jones, J., Mikati, M., Gallentine, W.B., Husain, A.M., Buckley, P.G., Stallings, R.L., Podgoreanu, M.V., Delanty, N., Sisodiya, S.M. and Goldstein, D.B. (2010) 'Rare deletions at 16p13.11 predispose to a diverse spectrum of sporadic epilepsy syndromes', *Am J Hum Genet*, 86(5), pp. 707-18.

Helbig, I. and Ellis, C.A. (2020) 'Personalized medicine in genetic epilepsies - possibilities, challenges, and new frontiers', *Neuropharmacology*, 172, p. 107970.

Helbig, I., Lopez-Hernandez, T., Shor, O., Galer, P., Ganesan, S., Pendziwiat, M., Rademacher, A., Ellis, C.A., Humpfer, N., Schwarz, N., Seiffert, S., Peeden, J., Shen, J., Sterbova, K., Hammer, T.B., Moller, R.S., Shinde, D.N., Tang, S., Smith, L., Poduri, A., Krause, R., Benninger, F., Helbig, K.L., Haucke, V., Weber, Y.G., Consortium, E.-R. and Consortium, G. (2019) 'A Recurrent Missense Variant in AP2M1 Impairs Clathrin-Mediated Endocytosis and Causes Developmental and Epileptic Encephalopathy', *Am J Hum Genet*, 104(6), pp. 1060-1072.

Helbig, I., Mefford, H.C., Sharp, A.J., Guipponi, M., Fichera, M., Franke, A., Muhle, H., de Kovel, C., Baker, C., von Spiczak, S., Kron, K.L., Steinich, I., Kleefuss-Lie, A.A., Leu, C., Gaus, V., Schmitz, B., Klein, K.M., Reif, P.S., Rosenow, F., Weber, Y., Lerche, H., Zimprich,

F., Urak, L., Fuchs, K., Feucht, M., Genton, P., Thomas, P., Visscher, F., de Haan, G.J., Moller, R.S., Hjalgrim, H., Luciano, D., Wittig, M., Nothnagel, M., Elger, C.E., Nurnberg, P., Romano, C., Malafosse, A., Koeleman, B.P., Lindhout, D., Stephani, U., Schreiber, S., Eichler, E.E. and Sander, T. (2009) '15q13.3 microdeletions increase risk of idiopathic generalized epilepsy', *Nat Genet*, 41(2), pp. 160-2.

Helbig, I., Riggs, E.R., Barry, C.A., Klein, K.M., Dyment, D., Thaxton, C., Sadikovic, B., Sands, T.T., Wagnon, J.L., Liaquat, K., Cilio, M.R., Mirzaa, G., Park, K., Axeen, E., Butler, E., Bardakjian, T.M., Striano, P., Poduri, A., Siegert, R.K., Grant, A.R., Helbig, K.L. and Mefford, H.C. (2018) 'The ClinGen Epilepsy Gene Curation Expert Panel-Bridging the divide between clinical domain knowledge and formal gene curation criteria', *Hum Mutat*, 39(11), pp. 1476-1484.

Helbig, K.L., Farwell Hagman, K.D., Shinde, D.N., Mroske, C., Powis, Z., Li, S., Tang, S. and Helbig, I. (2016) 'Diagnostic exome sequencing provides a molecular diagnosis for a significant proportion of patients with epilepsy', *Genet Med*, 18(9), pp. 898-905.

Hernandez-Diaz, S., Huybrechts, K.F., Desai, R.J., Cohen, J.M., Mogun, H., Pennell, P.B., Bateman, B.T. and Patorno, E. (2018) 'Topiramate use early in pregnancy and the risk of oral clefts: A pregnancy cohort study', *Neurology*, 90(4), pp. e342-e351.

Hildyard, J.C.W., Wells, D.J. and Piercy, R.J. (2021) 'Identification of qPCR reference genes suitable for normalising gene expression in the developing mouse embryo', *Wellcome Open Res*, 6, p. 197.

Hirsch, E., French, J., Scheffer, I.E., Bogacz, A., Alsaadi, T., Sperling, M.R., Abdulla, F., Zuberi, S.M., Trinka, E., Specchio, N., Somerville, E., Samia, P., Riney, K., Nabbout, R., Jain, S., Wilmshurst, J.M., Auvin, S., Wiebe, S., Perucca, E., Moshe, S.L., Tinuper, P. and Wirrell, E.C. (2022) 'ILAE definition of the Idiopathic Generalized Epilepsy Syndromes: Position statement by the ILAE Task Force on Nosology and Definitions', *Epilepsia*, 63(6), pp. 1475-1499.

Hoehndorf, R., Schofield, P.N. and Gkoutos, G.V. (2011) 'PhenomeNET: a whole-phenome approach to disease gene discovery', *Nucleic Acids Res*, 39(18), p. e119.

Hoogenraad, C.C., Feliu-Mojer, M.I., Spangler, S.A., Milstein, A.D., Dunah, A.W., Hung, A.Y. and Sheng, M. (2007) 'Liprinalpha1 degradation by calcium/calmodulin-dependent protein kinase II regulates LAR receptor tyrosine phosphatase distribution and dendrite development', *Dev Cell*, 12(4), pp. 587-602.

Horn, S., Au, M., Basel-Salmon, L., Bayrak-Toydemir, P., Chapin, A., Cohen, L., Elting, M.W., Graham, J.M., Gonzaga-Jauregui, C., Konen, O., Holzer, M., Lemke, J., Miller, C.E., Rey, L.K., Wolf, N.I., Weiss, M.M., Waisfisz, Q., Mirzaa, G.M., Wieczorek, D., Sticht, H. and Abou Jamra, R. (2019) 'De novo variants in PAK1 lead to intellectual disability with macrocephaly and seizures', *Brain*, 142(11), pp. 3351-3359.

Iinuma, K., Morimoto, K., Akiyama, T., Ikeda, A. and Kurihara, M. (2006) 'Proposed diagnostic scheme for the classification of epileptic seizures and epilepsies (ILAE, 2001): proposal from Japan Epilepsy Society', *Epilepsia*, 47(9), pp. 1588–9; discussion 1590–1.

Institut National de la Santé Et de la Recherche Médicale (1997) 'Orphanet: an online database of rare diseases and orphan drugs'. 2023-05-16. Available at: http://www.orpha.net.

International Knockout Mouse Consortium (2014) 'MGI download of modified allele data from IKMC and creation of new knockout alleles,'. 2024-03-17. Available at: https://www.informatics.jax.org/reference/J:204739.

International League Against Epilepsy *EpilepsyDiagnosis.org*. Available at: https://www.epilepsydiagnosis.org (Accessed: 2020-12-10).

International League Against Epilepsy Consortium on Complex Epilepsies (2018) 'Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies', *Nat Commun*, 9(1), p. 5269.

International League Against Epilepsy Consortium on Complex Epilepsies (2023) 'GWAS meta-analysis of over 29,000 people with epilepsy identifies 26 risk loci and subtype-specific genetic architecture', *Nat Genet*, 55(9), pp. 1471-1482.

Ivaniuk, A., Boßelmann, C.M., Zhang, X., St. John, M., Taylor, S.C., Krishnaswamy, G., Milinovich, A., Aziz, P.F., Pestana-Knight, E. and Lal, D. (2024) 'Natural language processing and expert follow-up establishes tachycardia association with CDKL5 deficiency disorder', *Genet Med Open*, 2, p. 100842.

Jabbari, K., Bobbili, D.R., Lal, D., Reinthaler, E.M., Schubert, J., Wolking, S., Sinha, V., Motameny, S., Thiele, H., Kawalia, A., Altmuller, J., Toliat, M.R., Kraaij, R., van Rooij, J., Uitterlinden, A.G., Ikram, M.A., Euro, E.C.C., Zara, F., Lehesjoki, A.E., Krause, R., Zimprich, F., Sander, T., Neubauer, B.A., May, P., Lerche, H. and Nurnberg, P. (2018) 'Rare gene deletions in genetic generalized and Rolandic epilepsies', *PLoS One*, 13(8), p. e0202022.

Jaccard, P. (1901) 'Etude comparative de la distribution florale dans une portion des Alpes et des Jura', *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37, pp. 547--579.

Jackson, J.H. (1958) *Selected writings of John Hughlings Jackson*. Staples.

Jacobsen, J.O.B., Baudis, M., Baynam, G.S., Beckmann, J.S., Beltran, S., Buske, O.J., Callahan, T.J., Chute, C.G., Courtot, M., Danis, D., Elemento, O., Essenwanger, A., Freimuth, R.R., Gargano, M.A., Groza, T., Hamosh, A., Harris, N.L., Kaliyaperumal, R., Lloyd, K.C.K., Khalifa, A., Krawitz, P.M., Köhler, S., Laraway, B.J., Lehväslaiho, H., Matalonga, L., McMurry, J.A., Metke-Jimenez, A., Mungall, C.J., Munoz-Torres, M.C., Ogishima, S., Papakonstantinou, A., Piscia, D., Pontikos, N., Queralt-Rosinach, N., Roos, M., Sass, J., Schofield, P.N., Seelow, D., Siapos, A., Smedley, D., Smith, L.D., Steinhaus, R., Sundaramurthi, J.C., Swietlik, E.M., Thun, S., Vasilevsky, N.A., Wagner, A.H., Warner, J.L., Weiland, C., Axton, M., Babb, L., Boerkoel, C.F., Chaudhari, B.P., Chin, H.-L., Dumontier, M., Gazzaz, N., Hansen, D.P., Hochheiser, H., Kinsler, V.A., Lochmüller, H., Mankovich, A.R., Saunders, G.I., Sergouniotis, P.I., Thompson, R., Zankl, A., Haendel, M.A., Robinson, P.N. and The, G.P.M.C. (2022) 'The GA4GH Phenopacket schema defines a computable representation of clinical data', *Nat Biotechnol*, 40(6), pp. 817-820.

Jagadeesh, K.A., Birgmeier, J., Guturu, H., Deisseroth, C.A., Wenger, A.M., Bernstein, J.A. and Bejerano, G. (2019) 'Phrank measures phenotype sets similarity to greatly improve Mendelian diagnostic disease prioritization', *Genet Med*, 21(2), pp. 464-470.

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., Chow, E.D., Kanterakis, E., Gao, H., Kia, A., Batzoglou, S., Sanders, S.J. and Farh, K.K. (2019) 'Predicting Splicing from Primary Sequence with Deep Learning', *Cell*, 176(3), pp. 535-548.e24.

Jain, S. and Bader, G.D. (2010) 'An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology', *BMC Bioinformatics*, 11(1), p. 562.

Javed, A., Agrawal, S. and Ng, P.C. (2014) 'Phen-Gen: combining phenotype and genotype to analyze rare disorders', *Nat Methods*, 11(9), pp. 935-7.

Jefferson, A., Leonard, H., Siafarikas, A., Woodhead, H., Fyfe, S., Ward, L.M., Munns, C., Motil, K., Tarquinio, D., Shapiro, J.R., Brismar, T., Ben-Zeev, B., Bisgaard, A.M., Coppola, G., Ellaway, C., Freilinger, M., Geerts, S., Humphreys, P., Jones, M., Lane, J., Larsson, G., Lotan, M., Percy, A., Pineda, M., Skinner, S., Syhler, B., Thompson, S., Weiss, B., Witt Engerström, I. and Downs, J. (2016) 'Clinical Guidelines for Management of Bone Health in Rett Syndrome Based on Expert Consensus and Available Evidence', *PLoS One*, 11(2), p. e0146824.

Ji, T., Wu, Y., Wang, H., Wang, J. and Jiang, Y. (2010) 'Diagnosis and fine mapping of a deletion in distal 11q in two Chinese patients with developmental delay', *J Hum Genet*, 55(8), pp. 486-9.

Jiang, J.J.C., D. W. (1997) *Proc 10th Research on Computational Linguistics International Conference*. Taipei, Taiwan, August 1997. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP). Available at: https://aclanthology.org/O97-1002.

Johannesen, K.M., Nikanorova, N., Marjanovic, D., Pavbro, A., Larsen, L.H.G., Rubboli, G. and Moller, R.S. (2020) 'Utility of genetic testing for therapeutic decision-making in adults with epilepsy', *Epilepsia*, 61(6), pp. 1234-1239.

Jordan, V.K., Zaveri, H.P. and Scott, D.A. (2015) '1p36 deletion syndrome: an update', *Appl Clin Genet*, 8, pp. 189-200.

Kalachikov, S., Evgrafov, O., Ross, B., Winawer, M., Barker-Cummings, C., Martinelli Boneschi, F., Choi, C., Morozov, P., Das, K., Teplitskaya, E., Yu, A., Cayanis, E., Penchaszadeh, G., Kottmann, A.H., Pedley, T.A., Hauser, W.A., Ottman, R. and Gilliam, T.C. (2002) 'Mutations in LGI1 cause autosomal-dominant partial epilepsy with auditory features', *Nat Genet*, 30(3), pp. 335-41.

Kanemitsu, Y., Fujitani, M., Fujita, Y., Zhang, S., Su, Y.Q., Kawahara, Y. and Yamashita, T. (2017) 'The RNA-binding protein MARF1 promotes cortical neurogenesis through its RNase activity domain', *Sci Rep*, 7(1), p. 1155.

Károlyi, L., Konrad, M., Köckerling, A., Ziegler, A., Zimmermann, D.K., Roth, B., Wieg, C., Grzeschik, K.-H., Koch, M.C., Seyberth, H.W., Vargas, R., Forestier, L., Jean, G., Deschaux, M., Rizzoni, G.F., Niaudet, P., Antignac, C., Feldmann, D., Lorridon, F., Cougoureux, E., Laroze, F., Alessandri, J.-L., David, L., Saunier, P., Deschenes, G., Hildebrandt, F., Vollmer, M., Proesmans, W., Brandis, M., van den Heuvel, L.P.W.J., Lemmink, H.H., Nillesen, W., Monnens, L.A.H., Knoers, N.V.A.M., Guay-Woodford, L.M., Wright, C.J., Madrigal, G. and Hebert, S.C. (1997) 'Mutations in the Gene Encoding the Inwardly-Rectifying Renal Potassium Channel, ROMK, Cause the Antenatal Variant of Bartter Syndrome: Evidence for Genetic Heterogeneity', *Human Molecular Genetics*, 6(1), pp. 17-26.

Kasperaviciute, D., Catarino, C.B., Heinzen, E.L., Depondt, C., Cavalleri, G.L., Caboclo, L.O., Tate, S.K., Jamnadas-Khoda, J., Chinthapalli, K., Clayton, L.M., Shianna, K.V., Radtke, R.A., Mikati, M.A., Gallentine, W.B., Husain, A.M., Alhusaini, S., Leppert, D., Middleton, L.T., Gibson, R.A., Johnson, M.R., Matthews, P.M., Hosford, D., Heuser, K., Amos, L., Ortega, M.,

Zumsteg, D., Wieser, H.G., Steinhoff, B.J., Kramer, G., Hansen, J., Dorn, T., Kantanen, A.M., Gjerstad, L., Peuralinna, T., Hernandez, D.G., Eriksson, K.J., Kalviainen, R.K., Doherty, C.P., Wood, N.W., Pandolfo, M., Duncan, J.S., Sander, J.W., Delanty, N., Goldstein, D.B. and Sisodiya, S.M. (2010) 'Common genetic variation and susceptibility to partial epilepsies: a genome-wide association study', *Brain*, 133(Pt 7), pp. 2136–47.

Kass, H.R., Winesett, S.P., Bessone, S.K., Turner, Z. and Kossoff, E.H. (2016) 'Use of dietary therapies amongst patients with GLUT1 deficiency syndrome', *Seizure*, 35, pp. 83-7.

Kassahun, Y., Perrone, R., De Momi, E., Berghofer, E., Tassi, L., Canevini, M.P., Spreafico, R., Ferrigno, G. and Kirchner, F. (2014) 'Automatic classification of epilepsy types using ontology-based and genetics-based machine learning', *Artif Intell Med*, 61(2), pp. 79-88.

Kasteleijn-Nolst Trenite, D.G., Schmitz, B., Janz, D., Delgado-Escueta, A.V., Thomas, P., Hirsch, E., Lerche, H., Camfield, C., Baykan, B., Feucht, M., Martinez-Juarez, I.E., Duron, R.M., Medina, M.T., Rubboli, G., Jerney, J., Hermann, B., Yacubian, E., Koutroumanidis, M., Stephani, U., Salas-Puig, J., Reed, R.C., Woermann, F., Wandschneider, B., Bureau, M., Gambardella, A., Koepp, M.J., Gelisse, P., Gurses, C., Crespel, A., Nguyen-Michel, V.H., Ferlazzo, E., Grisar, T., Helbig, I., Koeleman, B.P., Striano, P., Trimble, M., Buono, R., Cossette, P., Represa, A., Dravet, C., Serafini, A., Berglund, I.S., Sisodiya, S.M., Yamakawa, K. and Genton, P. (2013) 'Consensus on diagnosis and management of JME: From founder's observations to current trends', *Epilepsy Behav*, 28 Suppl 1, pp. S87–90.

Kaufman, L. and Rousseeuw, P.J. (2005) *Finding Groups in Data : An Introduction to Cluster Analysis*. Newy York, UNITED STATES: John Wiley & Sons, Incorporated.

Keet, C.M. (2013) 'Open World Assumption', in Dubitzky, W., Wolkenhauer, O., Cho, K.-H. and Yokota, H. (eds.) *Encyclopedia of Systems Biology*. New York, NY: Springer New York, pp. 1567-1567.

Keylock, A., Hong, Y., Saunders, D., Omoyinmi, E., Mulhern, C., Roebuck, D., Brogan, P., Ganesan, V. and Eleftheriou, D. (2018) 'Moyamoya-like cerebrovascular disease in a child with a novel mutation in myosin heavy chain 11', *Neurology*, 90(3), pp. 136-138.

Kilstrup-Nielsen, C., Rusconi, L., La Montanara, P., Ciceri, D., Bergo, A., Bedogni, F. and Landsberger, N. (2012) 'What we know and would like to know about CDKL5 and its involvement in epileptic encephalopathy', *Neural Plast*, 2012, p. 728267.

Kim, E., Lee, Y., Lee, H.J., Kim, J.S., Song, B.S., Huh, J.W., Lee, S.R., Kim, S.U., Kim, S.H., Hong, Y., Shim, I. and Chang, K.T. (2010) 'Implication of mouse Vps26b-Vps29-Vps35 retromer complex in sortilin trafficking', *Biochem Biophys Res Commun*, 403(2), pp. 167-71.

Kirby, J.C., Speltz, P., Rasmussen, L.V., Basford, M., Gottesman, O., Peissig, P.L., Pacheco, J.A., Tromp, G., Pathak, J., Carrell, D.S., Ellis, S.B., Lingren, T., Thompson, W.K., Savova, G., Haines, J., Roden, D.M., Harris, P.A. and Denny, J.C. (2016) 'PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability', *J Am Med Inform Assoc*, 23(6), pp. 1046-1052.

Kirsten, F., Chirag, M., Robin, P.G.-K., Michael, C.F., Matthew, F.H., Katrina, W. and Michael, A. (2023) 'Developmental regression in children: Current and future directions', *Cortex*, 169, pp. 5-17.

Kjeldsen, M.J., Corey, L.A., Christensen, K. and Friis, M.L. (2003) 'Epileptic seizures and syndromes in twins: the importance of genetic factors', *Epilepsy Res*, 55(1-2), pp. 137–46.

Klein, O., Pierre-Kahn, A., Boddaert, N., Parisot, D. and Brunelle, F. (2003) 'Dandy-Walker malformation: prenatal diagnosis and prognosis', *Childs Nerv Syst*, 19(7-8), pp. 484-9.

Kline, A.D., Moss, J.F., Selicorni, A., Bisgaard, A.M., Deardorff, M.A., Gillett, P.M., Ishman, S.L., Kerr, L.M., Levin, A.V., Mulder, P.A., Ramos, F.J., Wierzba, J., Ajmone, P.F., Axtell, D., Blagowidow, N., Cereda, A., Costantino, A., Cormier-Daire, V., FitzPatrick, D., Grados, M., Groves, L., Guthrie, W., Huisman, S., Kaiser, F.J., Koekkoek, G., Levis, M., Mariani, M., McCleery, J.P., Menke, L.A., Metrena, A., O'Connor, J., Oliver, C., Pie, J., Piening, S., Potter, C.J., Quaglio, A.L., Redeker, E., Richman, D., Rigamonti, C., Shi, A., Tümer, Z., Van Balkom, I.D.C. and Hennekam, R.C. (2018) 'Diagnosis and management of Cornelia de Lange syndrome: first international consensus statement', *Nat Rev Genet*, 19(10), pp. 649-666.

Knight, E.M.P., Amin, S., Bahi-Buisson, N., Benke, T.A., Cross, J.H., Demarest, S.T., Olson, H.E., Specchio, N., Fleming, T.R., Aimetti, A.A., Gasior, M. and Devinsky, O. (2022) 'Safety and efficacy of ganaxolone in patients with CDKL5 deficiency disorder: results from the double-blind phase of a randomised, placebo-controlled, phase 3 trial', *Lancet Neurol*, 21(5), pp. 417-427.

Knowles, J.K., Helbig, I., Metcalf, C.S., Lubbers, L.S., Isom, L.L., Demarest, S., Goldberg, E.M., George, A.L., Jr., Lerche, H., Weckhuysen, S., Whittemore, V., Berkovic, S.F. and Lowenstein, D.H. (2022) 'Precision medicine for genetic epilepsy on the horizon: Recent advances, present challenges, and suggestions for continued progress', *Epilepsia*, 63(10), pp. 2461-2475.

Kohler, S. (2009) *The Phenomizer*. Available at: https://compbio.charite.de/phenomizer/ (Accessed: 2024-02-18).

Kohler, S., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B., Danis, D., Gourdine, J.P., Gargano, M., Harris, N.L., Matentzoglu, N., McMurry, J.A., Osumi-Sutherland, D., Cipriani, V., Balhoff, J.P., Conlin, T., Blau, H., Baynam, G., Palmer, R., Gratian, D., Dawkins, H., Segal, M., Jansen, A.C., Muaz, A., Chang, W.H., Bergerson, J., Lauledeerkind, S.J.F., Yuksel, Z., Beltran, S., Freeman, A.F., Sergouniotis, P.I., Durkin, D., Storm, A.L., Hanauer, M., Brudno, M., Bello, S.M., Sincan, M., Rageth, K., Wheeler, M.T., Oegema, R., Lourghi, H., Della Rocca, M.G., Thompson, R., Castellanos, F., Priest, J., Cunningham-Rundles, C., Hegde, A., Lovering, R.C., Hajek, C., Olry, A., Notarangelo, L., Similuk, M., Zhang, X.A., Gomez-Andres, D., Lochmuller, H., Dollfus, H., Rosenzweig, S., Marwaha, S., Rath, A., Sullivan, K., Smith, C., Milner, J.D., Leroux, D., Boerkoel, C.F., Klion, A., Carter, M.C., Groza, T., Smedley, D., Haendel, M.A., Mungall, C. and Robinson, P.N. (2019) 'Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources', *Nucleic Acids Res*, 47(D1), pp. D1018-D1027.

Kohler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C., Brown, D.L., Brudno, M., Campbell, J., FitzPatrick, D.R., Eppig, J.T., Jackson, A.P., Freson, K., Girdea, M., Helbig, I., Hurst, J.A., Jahn, J., Jackson, L.G., Kelly, A.M., Ledbetter, D.H., Mansour, S., Martin, C.L., Moss, C., Mumford, A., Ouwehand, W.H., Park, S.M., Riggs, E.R., Scott, R.H., Sisodiya, S., Van Vooren, S., Wapner, R.J., Wilkie, A.O., Wright, C.F., Vulto-van Silfhout, A.T., de Leeuw, N., de Vries, B.B., Washingthon, N.L., Smith, C.L., Westerfield, M., Schofield, P., Ruef, B.J., Gkoutos, G.V., Haendel, M., Smedley, D., Lewis, S.E. and Robinson, P.N. (2014) 'The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data', *Nucleic Acids Res*, 42(Database issue), pp. D966-74.

Kohler, S., Gargano, M., Matentzoglu, N., Carmody, L.C., Lewis-Smith, D., Vasilevsky, N.A., Danis, D., Balagura, G., Baynam, G., Brower, A.M., Callahan, T.J., Chute, C.G., Est, J.L., Galer, P.D., Ganesan, S., Griese, M., Haimel, M., Pazmandi, J., Hanauer, M., Harris, N.L., Hartnett, M.J., Hastreiter, M., Hauck, F., He, Y., Jeske, T., Kearney, H., Kindle, G., Klein, C., Knoflach, K., Krause, R., Lagorce, D., McMurry, J.A., Miller, J.A., Munoz-Torres, M.C., Peters, R.L., Rapp, C.K., Rath, A.M., Rind, S.A., Rosenberg, A.Z., Segal, M.M., Seidel, M.G., Smedley, D., Talmy, T., Thomas, Y., Wiafe, S.A., Xian, J., Yuksel, Z., Helbig, I., Mungall, C.J., Haendel, M.A. and Robinson, P.N. (2021) 'The Human Phenotype Ontology in 2021', *Nucleic Acids Res*, 49(D1), pp. D1207-D1217.

Kohler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dolken, S., Ott, C.E., Mundlos, C., Horn, D., Mundlos, S. and Robinson, P.N. (2009) 'Clinical diagnostics in human genetics with semantic similarity searches in ontologies', *Am J Hum Genet*, 85(4), pp. 457-64.

Kohler, S., Vasilevsky, N.A., Engelstad, M., Foster, E., McMurry, J., Ayme, S., Baynam, G., Bello, S.M., Boerkoel, C.F., Boycott, K.M., Brudno, M., Buske, O.J., Chinnery, P.F., Cipriani, V., Connell, L.E., Dawkins, H.J., DeMare, L.E., Devereau, A.D., de Vries, B.B., Firth, H.V., Freson, K., Greene, D., Hamosh, A., Helbig, I., Hum, C., Jahn, J.A., James, R., Krause, R., SJ, F.L., Lochmuller, H., Lyon, G.J., Ogishima, S., Olry, A., Ouwehand, W.H., Pontikos, N., Rath, A., Schaefer, F., Scott, R.H., Segal, M., Sergouniotis, P.I., Sever, R., Smith, C.L., Straub, V., Thompson, R., Turner, C., Turro, E., Veltman, M.W., Vulliamy, T., Yu, J., von Ziegenweidt, J., Zankl, A., Zuchner, S., Zemojtel, T., Jacobsen, J.O., Groza, T., Smedley, D., Mungall, C.J., Haendel, M. and Robinson, P.N. (2017) 'The Human Phenotype Ontology in 2017', *Nucleic Acids Res*, 45(D1), pp. D865-D876.

Koko, M., Krause, R., Sander, T., Bobbili, D.R., Nothnagel, M., May, P., Lerche, H. and Epi, C. (2021) 'Distinct gene-set burden patterns underlie common generalized and focal epilepsies', *EBioMedicine*, 72, p. 103588.

Komodo Health (2024). Available at: https://www.komodohealth.com (Accessed: 2023-03-24).

Koolen, D.A., Pfundt, R., Linda, K., Beunders, G., Veenstra-Knol, H.E., Conta, J.H., Fortuna, A.M., Gillessen-Kaesbach, G., Dugan, S., Halbach, S., Abdul-Rahman, O.A., Winesett, H.M., Chung, W.K., Dalton, M., Dimova, P.S., Mattina, T., Prescott, K., Zhang, H.Z., Saal, H.M., Hehir-Kwa, J.Y., Willemsen, M.H., Ockeloen, C.W., Jongmans, M.C., Van der Aa, N., Failla, P., Barone, C., Avola, E., Brooks, A.S., Kant, S.G., Gerkes, E.H., Firth, H.V., Õunap, K., Bird, L.M., Masser-Frye, D., Friedman, J.R., Sokunbi, M.A., Dixit, A., Splitt, M., Kukolich, M.K., McGaughran, J., Coe, B.P., Flórez, J., Nadif Kasri, N., Brunner, H.G., Thompson, E.M., Gecz, J., Romano, C., Eichler, E.E., de Vries, B.B.A. and Study, D.D.D. (2016) 'The Koolen-de Vries syndrome: a phenotypic comparison of patients with a 17q21.31 microdeletion versus a KANSL1 sequence variant', *European Journal of Human Genetics*, 24(5), pp. 652-659.

Krey, I., Platzer, K., Esterhuizen, A., Berkovic, S.F., Helbig, I., Hildebrand, M.S., Lerche, H., Lowenstein, D., Møller, R.S., Poduri, A., Sadleir, L., Sisodiya, S.M., Weckhuysen, S., Wilmshurst, J.M., Weber, Y., Lemke, J.R., Berkovic, S.F., Cross, J.H., Helbig, I., Lerche, H., Lowenstein, D., Mefford, H.C., Perucca, P., Tan, N.C., Caglayan, H., Helbig, K., Singh, G., Weber, Y. and Weckhuysen, S. (2022) 'Current practice in diagnostic genetic testing of the epilepsies', *Epileptic Disord*, 24(5), pp. 765-786.

Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., McDermott, M.G., Monteiro, C.D.,

Gundersen, G.W. and Ma'ayan, A. (2016) 'Enrichr: a comprehensive gene set enrichment analysis web server 2016 update', *Nucleic Acids Res*, 44(W1), pp. W90-7.

Kwan, P., Arzimanoglou, A., Berg, A.T., Brodie, M.J., Hauser, W.A., Mathern, G., Moshe, S.L., Perucca, E., Wiebe, S. and French, J. (2010) 'Definition of drug resistant epilepsy: consensus proposal by the ad hoc Task Force of the ILAE Commission on Therapeutic Strategies', *Epilepsia*, 51(6), pp. 1069-77.

Kwan, P. and Brodie, M.J. (2000) 'Early identification of refractory epilepsy', *N Engl J Med*, 342(5), pp. 314-9.

Lagae, L., Sullivan, J., Knupp, K., Laux, L., Polster, T., Nikanorova, M., Devinsky, O., Cross, J.H., Guerrini, R., Talwar, D., Miller, I., Farfel, G., Galer, B.S., Gammaitoni, A., Mistry, A., Morrison, G., Lock, M., Agarwal, A., Lai, W.W., Ceulemans, B. and Group, F.A.D.S. (2020) 'Fenfluramine hydrochloride for the treatment of seizures in Dravet syndrome: a randomised, double-blind, placebo-controlled trial', *Lancet*, 394(10216), pp. 2243–2254.

Lai, M.C. and Baron-Cohen, S. (2015) 'Identifying the lost generation of adults with autism spectrum conditions', *Lancet Psychiatry*, 2(11), pp. 1013-27.

Lal, D., May, P., Perez-Palma, E., Samocha, K.E., Kosmicki, J.A., Robinson, E.B., Møller, R.S., Krause, R., Nürnberg, P., Weckhuysen, S., De Jonghe, P., Guerrini, R., Niestroj, L.M., Du, J., Marini, C., Balling, R., Barisic, N., Baulac, S., Caglayan, H., Craiu, D.C., De Jonghe, P., Depienne, C., Guerrini, R., Helbig, I., Hjalgrim, H., Hoffman-Zacharska, D., Jähn, J., Klein, K.M., Koeleman, B.P.C., Komarek, V., Krause, R., Leguern, E., Lehesjoki, A.-E., Lemke, J.R., Lerche, H., Linnankivi, T., Marini, C., May, P., Muhle, H., Pal, D.K., Palotie, A., Rosenow, F., Schubert-Bast, S., Selmer, K., Serratosa, J.M., Stephani, U., Štěrbová, K., Striano, P., Suls, A., Talvik, T., von Spiczak, S., Weber, Y.G., Weckhuysen, S., Zara, F., Ware, J.S., Kurki, M., Gormley, P., Tang, S., Wu, S., Biskup, S., Poduri, A., Neubauer, B.A., Koeleman, B.P.C., Helbig, K.L., Weber, Y.G., Helbig, I., Majithia, A.R., Palotie, A., Daly, M.J. and Euro, E.-R.E.S.C. (2020) 'Gene family information facilitates variant interpretation and identification of disease-associated genes in neurodevelopmental disorders', *Genome Medicine*, 12(1), p. 28.

Lal, D., Ruppert, A.K., Trucks, H., Schulz, H., de Kovel, C.G., Kasteleijn-Nolst Trenite, D., Sonsma, A.C., Koeleman, B.P., Lindhout, D., Weber, Y.G., Lerche, H., Kapser, C., Schankin, C.J., Kunz, W.S., Surges, R., Elger, C.E., Gaus, V., Schmitz, B., Helbig, I., Muhle, H., Stephani, U., Klein, K.M., Rosenow, F., Neubauer, B.A., Reinthaler, E.M., Zimprich, F., Feucht, M., Moller, R.S., Hjalgrim, H., De Jonghe, P., Suls, A., Lieb, W., Franke, A., Strauch, K., Gieger, C., Schurmann, C., Schminke, U., Nurnberg, P., Consortium, E. and Sander, T. (2015) 'Burden analysis of rare microdeletions suggests a strong impact of neurodevelopmental genes in genetic generalised epilepsies', *PLoS Genet*, 11(5), p. e1005226.

Lamberink, H.J., Otte, W.M., Geerts, A.T., Pavlovic, M., Ramos-Lizana, J., Marson, A.G., Overweg, J., Sauma, L., Specchio, L.M., Tennison, M., Cardoso, T.M.O., Shinnar, S., Schmidt, D., Geleijns, K. and Braun, K.P.J. (2017) 'Individualised prediction model of seizure recurrence and long-term outcomes after withdrawal of antiepileptic drugs in seizure-free patients: a systematic review and individual participant data meta-analysis', *Lancet Neurol*, 16(7), pp. 523-531.

Landais, R., Sultan, M. and Thomas, R.H. (2024) 'The promise of AI Large Language Models for Epilepsy care', *Epilepsy Behav*, 154, p. 109747.

Landolfi, A., Barone, P. and Erro, R. (2021) 'The Spectrum of PRRT2-Associated Disorders: Update on Clinical Features and Pathophysiology', *Front Neurol*, 12, p. 629747.

Leach, J.P., Mohanraj, R. and Borland, W. (2012) 'Alcohol and drugs in epilepsy: pathophysiology, presentation, possibilities, and prevention', *Epilepsia*, 53 Suppl 4, pp. 48-57.

Leal, S.M. and Ott, J. (2000) 'Effects of stratification in the analysis of affected-sib-pair data: benefits and costs', *Am J Hum Genet*, 66(2), pp. 567–75.

Lee, C.C., Huang, C.C. and Hsu, K.S. (2015) 'The phospholipid-binding protein SESTD1 negatively regulates dendritic spine density by interfering with Rac1-Trio8 signaling pathway', *Sci Rep*, 5, p. 13250.

Lee, C.M., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Gonzalez, J.N., Hinrichs, A.S., Lee, B.T., Nassar, L.R., Powell, C.C., Raney, B.J., Rosenbloom, K.R., Schmelter, D., Speir, M.L., Zweig, A.S., Haussler, D., Haeussler, M., Kuhn, R.M. and Kent, W.J. (2020) 'UCSC Genome Browser enters 20th year', *Nucleic Acids Res*, 48(D1), pp. D756-d761.

Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J. and Pavlidis, P. (2004) 'Coexpression analysis of human genes across many microarray data sets', *Genome Res*, 14(6), pp. 1085-94.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., Tukiainen, T., Birnbaum, D.P., Kosmicki, J.A., Duncan, L.E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D.N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M.I., Moonshine, A.L., Natarajan, P., Orozco, L., Peloso, G.M., Poplin, R., Rivas, M.A., Ruano-Rubio, V., Rose, S.A., Ruderfer, D.M., Shakir, K., Stenson, P.D., Stevens, C., Thomas, B.P., Tiao, G., Tusie-Luna, M.T., Weisburd, B., Won, H.H., Yu, D., Altshuler, D.M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J.C., Gabriel, S.B., Getz, G., Glatt, S.J., Hultman, C.M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M.I., McGovern, D., McPherson, R., Neale, B.M., Palotie, A., Purcell, S.M., Saleheen, D., Scharf, J.M., Sklar, P., Sullivan, P.F., Tuomilehto, J., Tsuang, M.T., Watkins, H.C., Wilson, J.G., Daly, M.J., MacArthur, D.G. and Exome Aggregation, C. (2016) 'Analysis of protein-coding genetic variation in 60,706 humans', *Nature*, 536(7616), pp. 285-91.

Lemke, J.R., Lal, D., Reinthaler, E.M., Steiner, I., Nothnagel, M., Alber, M., Geider, K., Laube, B., Schwake, M., Finsterwalder, K., Franke, A., Schilhabel, M., Jahn, J.A., Muhle, H., Boor, R., Van Paesschen, W., Caraballo, R., Fejerman, N., Weckhuysen, S., De Jonghe, P., Larsen, J., Moller, R.S., Hjalgrim, H., Addis, L., Tang, S., Hughes, E., Pal, D.K., Veri, K., Vaher, U., Talvik, T., Dimova, P., Guerrero Lopez, R., Serratosa, J.M., Linnankivi, T., Lehesjoki, A.E., Ruf, S., Wolff, M., Buerki, S., Wohlrab, G., Kroell, J., Datta, A.N., Fiedler, B., Kurlemann, G., Kluger, G., Hahn, A., Haberlandt, D.E., Kutzer, C., Sperner, J., Becker, F., Weber, Y.G., Feucht, M., Steinbock, H., Neophythou, B., Ronen, G.M., Gruber-Sedlmayr, U., Geldner, J., Harvey, R.J., Hoffmann, P., Herms, S., Altmuller, J., Toliat, M.R., Thiele, H., Nurnberg, P., Wilhelm, C., Stephani, U., Helbig, I., Lerche, H., Zimprich, F., Neubauer, B.A., Biskup, S. and von Spiczak, S. (2013) 'Mutations in GRIN2A cause idiopathic focal epilepsy with rolandic spikes', *Nat Genet*, 45(9), pp. 1067–72.

Leonard, H., Downs, J., Benke, T.A., Swanson, L., Olson, H. and Demarest, S. (2022a) 'CDKL5 deficiency disorder: clinical features, diagnosis, and management', *Lancet Neurol*, 21(6), pp. 563-576.

Leonard, H., Junaid, M., Wong, K., Aimetti, A.A., Pestana Knight, E. and Downs, J. (2022b) 'Influences on the trajectory and subsequent outcomes in CDKL5 deficiency disorder', *Epilepsia*, 63(2), pp. 352-363.

Leonard, H., Junaid, M., Wong, K., Demarest, S. and Downs, J. (2021) 'Exploring quality of life in individuals with a severe developmental and epileptic encephalopathy, CDKL5 Deficiency Disorder', *Epilepsy Res*, 169, p. 106521.

Leppa, V.M., Kravitz, S.N., Martin, C.L., Andrieux, J., Le Caignec, C., Martin-Coignard, D., DyBuncio, C., Sanders, S.J., Lowe, J.K., Cantor, R.M. and Geschwind, D.H. (2016) 'Rare Inherited and De Novo CNVs Reveal Complex Contributions to ASD Risk in Multiplex Families', *Am J Hum Genet*, 99(3), pp. 540-554.

Lerche, H., Berkovic, S.F., Lowenstein, D.H., Euro, E.-C.C., Epi, P.G.X.C. and Epi, K.C.E.P.G.P. (2019) 'Intestinal-Cell Kinase and Juvenile Myoclonic Epilepsy', *N Engl J Med*, 380(16), p. e24.

Lerer, I., Sagi, M., Meiner, V., Cohen, T., Zlotogora, J. and Abeliovich, D. (2005) 'Deletion of the ANKRD15 gene at 9p24.3 causes parent-of-origin-dependent inheritance of familial cerebral palsy', *Hum Mol Genet*, 14(24), pp. 3911-20.

Leu, C., Stevelink, R., Smith, A.W., Goleva, S.B., Kanai, M., Ferguson, L., Campbell, C., Kamatani, Y., Okada, Y., Sisodiya, S.M., Cavalleri, G.L., Koeleman, B.P.C., Lerche, H., Jehi, L., Davis, L.K., Najm, I.M., Palotie, A., Daly, M.J., Busch, R.M., Epi, C. and Lal, D. (2019) 'Polygenic burden in focal and generalized epilepsies', *Brain*, 142(11), pp. 3473-3481.

Lewis-Smith D, Ganesan S, Galer PD, Krause R, Thomas RH, Helbig I and Epi25 Collaborative (2021) 'An Expanded Genotype-Phenotype Analysis of 11,500 People with Epilepsy', *American Epilepsy Society Annual Meeting*. Chicago, 2021-12-05. Available at: https://cms.aesnet.org/abstractslisting/an-expanded-genotype-phenotype-analysis-of-11-500-people-with-epilepsy (Accessed: 2022-01-06).

Lewis-Smith, D., Craig, D.P., Thomas, N.J.P., Hamandi, K. and Thomas, R.H. (2020a) 'Chapter 16 - Molecular genetic management of epilepsy', in Kumar, D. (ed.) *Clinical Molecular Medicine*. Academic Press, pp. 289–308.

Lewis-Smith, D., Ellis, C.A., Helbig, I. and Thomas, R.H. (2020b) 'Early-onset genetic epilepsies reaching adult clinics', *Brain*, 143(3), p. e19.

Lewis-Smith, D., Galer, P.D., Balagura, G., Kearney, H., Ganesan, S., Cosico, M., O'Brien, M., Vaidiswaran, P., Krause, R., Ellis, C.A., Thomas, R.H., Robinson, P.N. and Helbig, I. (2021a) 'Modeling seizures in the Human Phenotype Ontology according to contemporary ILAE concepts makes big phenotypic data tractable', *Epilepsia*, 62(6), pp. 1293-1305.

Lewis-Smith, D., Ganesan, S., Galer, P.D., Helbig, K.L., McKeown, S.E., O'Brien, M., Khankhanian, P., Kaufman, M.C., Gonzalez, A.K., Felmeister, A.S., Krause, R., Ellis, C.A. and Helbig, I. (2021b) 'Phenotypic homogeneity in childhood epilepsies evolves in gene-specific patterns across 3251 patient-years of clinical data', *Eur J Hum Genet*, 29(11), pp. 1690-1700.

Lewis-Smith, D., Jaiser, S.R. and Thomas, R.H. (2022) 'Autoimmune musicogenic bilateral temporal lobe epilepsy', *Epileptic Disord*, 24(5), pp. 961-964.

Lewis-Smith, D., Parthasarathy, S., Xian, J., Kaufman, M.C., Ganesan, S., Galer, P.D., Thomas, R.H. and Helbig, I. (2022) 'Computational analysis of neurodevelopmental phenotypes: Harmonization empowers clinical discovery', *Hum Mutat*, 43(11), pp. 1642-1658.

Lewis-Smith, D.K., M; Xian, J; Ganesan, S; Fitzgerald, M; Abend, N; Thomas, RH; Helbig, I (2022) 'Panning for gold: electronic health records in the epilepsy clinic', *Epilepsy Professional*, (Spring), pp. 12-22.

Lhatoo, S.D., Bernasconi, N., Blumcke, I., Braun, K., Buchhalter, J., Denaxas, S., Galanopoulou, A., Josephson, C., Kobow, K., Lowenstein, D., Ryvlin, P., Schulze-Bonhage, A., Sahoo, S.S., Thom, M., Thurman, D., Worrell, G., Zhang, G.Q. and Wiebe, S. (2020) 'Big data in epilepsy: Clinical and research considerations. Report from the Epilepsy Big Data Task Force of the International League Against Epilepsy', *Epilepsia*, 61(9), pp. 1869-1883.

Li, B., Wang, J.Z., Feltus, F.A., Zhou, J. and Luo, F. (2010) 'Effectively integrating information content and structural relationship to improve the GO-based similarity measure between proteins', *arXiv*, p. arXiv:1001.0958.

Li, H., Zhao, P., Xu, Q., Shan, S., Hu, C., Qiu, Z. and Xu, X. (2016) 'The autism-related gene SNRPN regulates cortical and spine development via controlling nuclear receptor Nr4a1', *Sci Rep*, 6, p. 29878.

Li, K.K., Mao, C.Y., Zhang, J.G., Ma, Q., Wang, Y.J., Liu, X.H., Bao, T. and Guo, W. (2019) 'Overexpression of U three protein 14a (UTP14a) is associated with poor prognosis of esophageal squamous cell carcinoma', *Thorac Cancer*, 10(11), pp. 2071-2080.

Li, Q. and Wang, K. (2017) 'InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines', *Am J Hum Genet*, 100(2), pp. 267-280.

Li, W., Schneider, A.L. and Scheffer, I.E. (2021) 'Defining Dravet syndrome: An essential pre-requisite for precision medicine trials', *Epilepsia*, 62(9), pp. 2205-2217.

Liao, W.W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., Buonaiuto, S., Chang, X.H., Cheng, H., Chu, J., Colonna, V., Eizenga, J.M., Feng, X., Fischer, C., Fulton, R.S., Garg, S., Groza, C., Guarracino, A., Harvey, W.T., Heumos, S., Howe, K., Jain, M., Lu, T.Y., Markello, C., Martin, F.J., Mitchell, M.W., Munson, K.M., Mwaniki, M.N., Novak, A.M., Olsen, H.E., Pesout, T., Porubsky, D., Prins, P., Sibbesen, J.A., Sirén, J., Tomlinson, C., Villani, F., Vollger, M.R., Antonacci-Fulton, L.L., Baid, G., Baker, C.A., Belyaeva, A., Billis, K., Carroll, A., Chang, P.C., Cody, S., Cook, D.E., Cook-Deegan, R.M., Cornejo, O.E., Diekhans, M., Ebert, P., Fairley, S., Fedrigo, O., Felsenfeld, A.L., Formenti, G., Frankish, A., Gao, Y., Garrison, N.A., Giron, C.G., Green, R.E., Haggerty, L., Hoekzema, K., Hourlier, T., Ji, H.P., Kenny, E.E., Koenig, B.A., Kolesnikov, A., Korbel, J.O., Kordosky, J., Koren, S., Lee, H., Lewis, A.P., Magalhães, H., Marco-Sola, S., Marijon, P., McCartney, A., McDaniel, J., Mountcastle, J., Nattestad, M., Nurk, S., Olson, N.D., Popejoy, A.B., Puiu, D., Rautiainen, M., Regier, A.A., Rhie, A., Sacco, S., Sanders, A.D., Schneider, V.A., Schultz, B.I., Shafin, K., Smith, M.W., Sofia, H.J., Abou Tayoun, A.N., Thibaud-Nissen, F., Tricomi, F.F., et al. (2023) 'A draft human pangenome reference', *Nature*, 617(7960), pp. 312-324.

Lim, E.T., Würtz, P., Havulinna, A.S., Palta, P., Tukiainen, T., Rehnström, K., Esko, T., Mägi, R., Inouye, M., Lappalainen, T., Chan, Y., Salem, R.M., Lek, M., Flannick, J., Sim, X., Manning, A., Ladenvall, C., Bumpstead, S., Hämäläinen, E., Aalto, K., Maksimow, M., Salmi, M., Blankenberg, S., Ardissino, D., Shah, S., Horne, B., McPherson, R., Hovingh, G.K., Reilly,

M.P., Watkins, H., Goel, A., Farrall, M., Girelli, D., Reiner, A.P., Stitziel, N.O., Kathiresan, S., Gabriel, S., Barrett, J.C., Lehtimäki, T., Laakso, M., Groop, L., Kaprio, J., Perola, M., McCarthy, M.I., Boehnke, M., Altshuler, D.M., Lindgren, C.M., Hirschhorn, J.N., Metspalu, A., Freimer, N.B., Zeller, T., Jalkanen, S., Koskinen, S., Raitakari, O., Durbin, R., MacArthur, D.G., Salomaa, V., Ripatti, S., Daly, M.J., Palotie, A. and for the Sequencing Initiative Suomi, P. (2014) 'Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population', *PLoS Genet*, 10(7), p. e1004494.

Lim, Z., Wong, K., Downs, J., Bebbington, K., Demarest, S. and Leonard, H. (2018) 'Vagus nerve stimulation for the treatment of refractory epilepsy in the CDKL5 Deficiency Disorder', *Epilepsy Res*, 146, pp. 36-40.

Lim, Z., Wong, K., Olson, H.E., Bergin, A.M., Downs, J. and Leonard, H. (2017) 'Use of the ketogenic diet to manage refractory epilepsy in CDKL5 disorder: Experience of >100 patients', *Epilepsia*, 58(8), pp. 1415-1422.

Lin, D. (1998) *15th International Conference on Machine Learning*. Madison, WI, USA. Morgan Kaufmann.

Lin, N., Wu, B., Jansen, R., Gerstein, M. and Zhao, H. (2004) 'Information assessment on predicting protein-protein interactions', *BMC Bioinformatics*, 5(1), p. 154.

Lindy, A.S., Stosser, M.B., Butler, E., Downtain-Pickersgill, C., Shanmugham, A., Retterer, K., Brandt, T., Richard, G. and McKnight, D.A. (2018) 'Diagnostic outcomes for genetic testing of 70 genes in 8565 patients with epilepsy and neurodevelopmental disorders', *Epilepsia*, 59(5), pp. 1062-1071.

Loddenkemper, T., Kellinghaus, C., Wyllie, E., Najm, I.M., Gupta, A., Rosenow, F. and Luders, H.O. (2005) 'A proposal for a five-dimensional patient-oriented epilepsy classification', *Epileptic Disord*, 7(4), pp. 308-16.

Loiseau, P., Duche, B., Cordova, S., Dartigues, J.F. and Cohadon, S. (1988) 'Prognosis of benign childhood epilepsy with centrotemporal spikes: a follow-up study of 168 patients', *Epilepsia*, 29(3), pp. 229-35.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., Fleming, J., Siminoff, L., Traino, H., Mosavel, M., Barker, L., Jewell, S., Rohrer, D., Maxim, D., Filkins, D., Harbach, P., Cortadillo, E., Berghuis, B., Turner, L., Hudson, E., Feenstra, K., Sobin, L., Robb, J., Branton, P., Korzeniewski, G., Shive, C., Tabor, D., Qi, L., Groch, K., Nampally, S., Buia, S., Zimmerman, A., Smith, A., Burges, R., Robinson, K., Valentino, K., Bradbury, D., Cosentino, M., Diaz-Mayoral, N., Kennedy, M., Engel, T., Williams, P., Erickson, K., Ardlie, K., Winckler, W., Getz, G., DeLuca, D., MacArthur, D., Kellis, M., Thomson, A., Young, T., Gelfand, E., Donovan, M., Meng, Y., Grant, G., Mash, D., Marcus, Y., Basile, M., Liu, J., Zhu, J., Tu, Z., Cox, N.J., Nicolae, D.L., Gamazon, E.R., Im, H.K., Konkashbaev, A., Pritchard, J., Stevens, M., Flutre, T., Wen, X., Dermitzakis, E.T., Lappalainen, T., Guigo, R., Monlong, J., Sammeth, M., Koller, D., Battle, A., Mostafavi, S., McCarthy, M., Rivas, M., Maller, J., Rusyn, I., Nobel, A., Wright, F., Shabalin, A., Feolo, M., Sharopova, N., et al. (2013) 'The Genotype-Tissue Expression (GTEx) project', *Nat Genet*, 45(6), pp. 580-585.

Lopez-Nunez, O., Alaggio, R., Ranganathan, S., Schmitt, L., John, I., Church, A.J. and Picarsic, J. (2020) 'New molecular insights into the pathogenesis of lipoblastomas: clinicopathologic, immunohistochemical, and molecular analysis in pediatric cases', *Hum Pathol*, 104, pp. 30-41.

López-Rivera, J.A., Leu, C., Macnee, M., Khoury, J., Hoffmann, L., Coras, R., Kobow, K., Bhattarai, N., Pérez-Palma, E., Hamer, H., Brandner, S., Rössler, K., Bien, C.G., Kalbhenn, T., Pieper, T., Hartlieb, T., Butler, E., Genovese, G., Becker, K., Altmüller, J., Niestroj, L.-M., Ferguson, L., Busch, R.M., Nürnberg, P., Najm, I., Blümcke, I. and Lal, D. (2022) 'The genomic landscape across 474 surgically accessible epileptogenic human brain lesions', *Brain*, 146(4), pp. 1342-1356.

Lord, P.W., Stevens, R.D., Brass, A. and Goble, C.A. (2003a) 'Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation', *Bioinformatics*, 19(10), pp. 1275-83.

Lord, P.W., Stevens, R.D., Brass, A. and Goble, C.A. (2003b) 'Semantic similarity measures as tools for exploring the gene ontology', *Pac Symp Biocomput*, pp. 601-12.

Louis, D.N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W.K., Ohgaki, H., Wiestler, O.D., Kleihues, P. and Ellison, D.W. (2016) 'The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary', *Acta Neuropathol*, 131(6), pp. 803-20.

Lu, C., Liao, W., Huang, Y., Huang, Y. and Luo, Y. (2022) 'Increased expression of NOP14 is associated with improved prognosis due to immune regulation in colorectal cancer', *BMC Gastroenterol*, 22(1), p. 207.

Luders, H., Acharya, J., Baumgartner, C., Benbadis, S., Bleasel, A., Burgess, R., Dinner, D.S., Ebner, A., Foldvary, N., Geller, E., Hamer, H., Holthausen, H., Kotagal, P., Morris, H., Meencke, H.J., Noachtar, S., Rosenow, F., Sakamoto, A., Steinhoff, B.J., Tuxhorn, I. and Wyllie, E. (1998) 'Semiological seizure classification', *Epilepsia*, 39(9), pp. 1006-13.

Lüders, H., Najm, I. and Wyllie, E. (2003) 'Reply to “Of Cabbages and Kings: Some Considerations on Classifications, Diagnostic Schemes, Semiology, and Concepts”', *Epilepsia*, 44(1), pp. 6–7.

Luders, H., Vaca, G.F., Akamatsu, N., Amina, S., Arzimanoglou, A., Baumgartner, C., Benbadis, S.R., Bleasel, A., Bermeo-Ovalle, A., Bozorgi, A., Carreno, M., Devereaux, M., Francione, S., Losarcos, N.G., Hamer, H., Holthausen, H., Jamal-Omidi, S., Kalamangalam, G., Kanner, A.M., Knake, S., Lacuey, N., Lhatoo, S., Lim, S.H., Londono, L.V., Mani, J., Matsumoto, R., Miller, J.P., Noachtar, S., Palmini, A., Park, J., Rosenow, F., Shahid, A., Schuele, S., Steinhoff, B.J., Akos Szabo, C., Tandon, N., Terada, K., Boas, W.V.E., Widdess-Walsh, P. and Kahane, P. (2019) 'Classification of paroxysmal events and the four-dimensional epilepsy classification system', *Epileptic Disord*, 21(1), pp. 1-29.

Luders, H.O., Amina, S., Baumgartner, C., Benbadis, S., Bermeo-Ovalle, A., Devereaux, M., Diehl, B., Edwards, J., Baca-Vaca, G.F., Hamer, H., Ikeda, A., Kaiboriboon, K., Kellinghaus, C., Koubeissi, M., Lardizabal, D., Lhatoo, S., Luders, J., Mani, J., Mayor, L.C., Miller, J., Noachtar, S., Pestana, E., Rosenow, F., Sakamoto, A., Shahid, A., Steinhoff, B.J., Syed, T., Tanner, A. and Tsuji, S. (2012) 'Modern technology calls for a modern approach to classification of epileptic seizures and the epilepsies', *Epilepsia*, 53(3), pp. 405–11.

Ma, Q.L., Yang, F., Frautschy, S.A. and Cole, G.M. (2012) 'PAK in Alzheimer disease, Huntington disease and X-linked mental retardation', *Cell Logist*, 2(2), pp. 117-125.

MacKay, C.I., Wong, K., Demarest, S.T., Benke, T.A., Downs, J. and Leonard, H. (2021) 'Exploring genotype-phenotype relationships in the CDKL5 deficiency disorder using an international dataset', *Clin Genet*, 99(1), pp. 157-165.

Malwade, S., Gasthaus, J., Bellardita, C., Andelic, M., Moric, B., Korshunova, I., Kiehn, O., Vasistha, N.A. and Khodosevich, K. (2022) 'Identification of Vulnerable Interneuron Subtypes in 15q13.3 Microdeletion Syndrome Using Single-Cell Transcriptomics', *Biol Psychiatry*, 91(8), pp. 727-739.

Manford, M., Hart, Y.M., Sander, J.W. and Shorvon, S.D. (1992) 'National General Practice Study of Epilepsy (NGPSE): partial seizure patterns in a general population', *Neurology*, 42(10), pp. 1911-7.

Mangatt, M., Wong, K., Anderson, B., Epstein, A., Hodgetts, S., Leonard, H. and Downs, J. (2016) 'Prevalence and onset of comorbidities in the CDKL5 disorder differ from Rett syndrome', *Orphanet J Rare Dis*, 11, p. 39.

Manning, M. and Hudgins, L. (2010) 'Array-based technology and recommendations for utilization in medical genetics practice for detection of chromosomal abnormalities', *Genet Med*, 12(11), pp. 742-5.

Manning, M. and Hudgins, L. (2020) 'Addendum: Array-based technology and recommendations for utilization in medical genetics practice for detection of chromosomal abnormalities', *Genet Med*, 22(12), p. 2126.

Marini, C., Scheffer, I.E., Crossland, K.M., Grinton, B.E., Phillips, F.L., McMahon, J.M., Turner, S.J., Dean, J.T., Kivity, S., Mazarib, A., Neufeld, M.Y., Korczyn, A.D., Harkin, L.A., Dibbens, L.M., Wallace, R.H., Mulley, J.C. and Berkovic, S.F. (2004) 'Genetic architecture of idiopathic generalized epilepsy: clinical genetic analysis of 55 multiplex families', *Epilepsia*, 45(5), pp. 467-78.

Marson, A., Burnside, G., Appleton, R., Smith, D., Leach, J.P., Sills, G., Tudur-Smith, C., Plumpton, C., Hughes, D.A., Williamson, P., Baker, G.A., Balabanova, S., Taylor, C., Brown, R., Hindley, D., Howell, S., Maguire, M., Mohanraj, R. and Smith, P.E. (2021a) 'The SANAD II study of the effectiveness and cost-effectiveness of levetiracetam, zonisamide, or lamotrigine for newly diagnosed focal epilepsy: an open-label, non-inferiority, multicentre, phase 4, randomised controlled trial', *Lancet*, 397(10282), pp. 1363-1374.

Marson, A., Burnside, G., Appleton, R., Smith, D., Leach, J.P., Sills, G., Tudur-Smith, C., Plumpton, C., Hughes, D.A., Williamson, P., Baker, G.A., Balabanova, S., Taylor, C., Brown, R., Hindley, D., Howell, S., Maguire, M., Mohanraj, R. and Smith, P.E. (2021b) 'The SANAD II study of the effectiveness and cost-effectiveness of valproate versus levetiracetam for newly diagnosed generalised and unclassifiable epilepsy: an open-label, non-inferiority, multicentre, phase 4, randomised controlled trial', *Lancet*, 397(10282), pp. 1375-1386.

Marson, A.G., Al-Kharusi, A.M., Alwaidh, M., Appleton, R., Baker, G.A., Chadwick, D.W., Cramp, C., Cockerell, O.C., Cooper, P.N., Doughty, J., Eaton, B., Gamble, C., Goulding, P.J., Howell, S.J., Hughes, A., Jackson, M., Jacoby, A., Kellett, M., Lawson, G.R., Leach, J.P., Nicolaides, P., Roberts, R., Shackley, P., Shen, J., Smith, D.F., Smith, P.E., Smith, C.T., Vanoli, A., Williamson, P.R. and group, S.S. (2007a) 'The SANAD study of effectiveness of

carbamazepine, gabapentin, lamotrigine, oxcarbazepine, or topiramate for treatment of partial epilepsy: an unblinded randomised controlled trial', *Lancet*, 369(9566), pp. 1000-15.

Marson, A.G., Al-Kharusi, A.M., Alwaidh, M., Appleton, R., Baker, G.A., Chadwick, D.W., Cramp, C., Cockerell, O.C., Cooper, P.N., Doughty, J., Eaton, B., Gamble, C., Goulding, P.J., Howell, S.J., Hughes, A., Jackson, M., Jacoby, A., Kellett, M., Lawson, G.R., Leach, J.P., Nicolaides, P., Roberts, R., Shackley, P., Shen, J., Smith, D.F., Smith, P.E., Smith, C.T., Vanoli, A., Williamson, P.R. and group, S.S. (2007b) 'The SANAD study of effectiveness of valproate, lamotrigine, or topiramate for generalised and unclassifiable epilepsy: an unblinded randomised controlled trial', *Lancet*, 369(9566), pp. 1016-26.

Martinez-Juarez, I.E., Alonso, M.E., Medina, M.T., Duron, R.M., Bailey, J.N., Lopez-Ruiz, M., Ramos-Ramirez, R., Leon, L., Pineda, G., Castroviejo, I.P., Silva, R., Mija, L., Perez-Gosiengfiao, K., Machado-Salas, J. and Delgado-Escueta, A.V. (2006) 'Juvenile myoclonic epilepsy subsyndromes: family studies and long-term follow-up', *Brain*, 129(Pt 5), pp. 1269–80.

Martins Custodio, H., Clayton, L.M., Bellampalli, R., Pagni, S., Silvennoinen, K., Caswell, R., Brunklaus, A., Guerrini, R., Koeleman, B.P.C., Lemke, J.R., Møller, R.S., Scheffer, I.E., Weckhuysen, S., Zara, F., Zuberi, S., Kuchenbaecker, K., Balestrini, S., Mills, J.D. and Sisodiya, S.M. (2023) 'Widespread genomic influences on phenotype in Dravet syndrome, a 'monogenic' condition', *Brain*, 146(9), pp. 3885-3897.

Masino, A.J., Dechene, E.T., Dulik, M.C., Wilkens, A., Spinner, N.B., Krantz, I.D., Pennington, J.W., Robinson, P.N. and White, P.S. (2014) 'Clinical phenotype-based gene prioritization: an initial study using semantic similarity and the human phenotype ontology', *BMC Bioinformatics*, 15(1), p. 248.

May, P., Girard, S., Harrer, M., Bobbili, D.R., Schubert, J., Wolking, S., Becker, F., Lachance-Touchette, P., Meloche, C., Gravel, M., Niturad, C.E., Knaus, J., De Kovel, C., Toliat, M., Polvi, A., Iacomino, M., Guerrero-Lopez, R., Baulac, S., Marini, C., Thiele, H., Altmuller, J., Jabbari, K., Ruppert, A.K., Jurkowski, W., Lal, D., Rusconi, R., Cestele, S., Terragni, B., Coombs, I.D., Reid, C.A., Striano, P., Caglayan, H., Siren, A., Everett, K., Moller, R.S., Hjalgrim, H., Muhle, H., Helbig, I., Kunz, W.S., Weber, Y.G., Weckhuysen, S., Jonghe, P., Sisodiya, S.M., Nabbout, R., Franceschetti, S., Coppola, A., Vari, M.S., Kasteleijn-Nolst Trenite, D., Baykan, B., Ozbek, U., Bebek, N., Klein, K.M., Rosenow, F., Nguyen, D.K., Dubeau, F., Carmant, L., Lortie, A., Desbiens, R., Clement, J.F., Cieuta-Walti, C., Sills, G.J., Auce, P., Francis, B., Johnson, M.R., Marson, A.G., Berghuis, B., Sander, J.W., Avbersek, A., McCormack, M., Cavalleri, G.L., Delanty, N., Depondt, C., Krenn, M., Zimprich, F., Peter, S., Nikanorova, M., Kraaij, R., van Rooij, J., Balling, R., Ikram, M.A., Uitterlinden, A.G., Avanzini, G., Schorge, S., Petrou, S., Mantegazza, M., Sander, T., LeGuern, E., Serratosa, J.M., Koeleman, B.P.C., Palotie, A., Lehesjoki, A.E., Nothnagel, M., Nurnberg, P., Maljevic, S., Zara, F., Cossette, P., Krause, R., Lerche, H., Epicure, C., Euro, E.C.C., et al. (2018) 'Rare coding variants in genes encoding GABA(A) receptors in genetic generalised epilepsies: an exome-based case-control study', *Lancet Neurol*, 17(8), pp. 699-708.

Mazandu, G.K. and Mulder, N.J. (2012) 'A topology-based metric for measuring term similarity in the gene ontology', *Adv Bioinformatics*, 2012, p. 975783.

McCormack, M., Gui, H., Ingason, A., Speed, D., Wright, G.E.B., Zhang, E.J., Secolin, R., Yasuda, C., Kwok, M., Wolking, S., Becker, F., Rau, S., Avbersek, A., Heggeli, K., Leu, C., Depondt, C., Sills, G.J., Marson, A.G., Auce, P., Brodie, M.J., Francis, B., Johnson, M.R., Koeleman, B.P.C., Striano, P., Coppola, A., Zara, F., Kunz, W.S., Sander, J.W., Lerche, H.,

Klein, K.M., Weckhuysen, S., Krenn, M., Gudmundsson, L.J., Stefansson, K., Krause, R., Shear, N., Ross, C.J.D., Delanty, N., Consortium, E., Pirmohamed, M., Carleton, B.C., Canadian Pharmacogenomics Network for Drug, S., Cendes, F., Lopes-Cendes, I., Liao, W.P., O'Brien, T.J., Sisodiya, S.M., Epi, P.G.X.C., Cherny, S., Kwan, P., Baum, L., International League Against Epilepsy Consortium on Complex, E. and Cavalleri, G.L. (2018) 'Genetic variation in CFH predicts phenytoin-induced maculopapular exanthema in European-descent patients', *Neurology*, 90(4), pp. e332-e341.

McDonald, D.A., Shenkar, R., Shi, C., Stockton, R.A., Akers, A.L., Kucherlapati, M.H., Kucherlapati, R., Brainer, J., Ginsberg, M.H., Awad, I.A. and Marchuk, D.A. (2011) 'A novel mouse model of cerebral cavernous malformations based on the two-hit mutation hypothesis recapitulates the human disease', *Hum Mol Genet*, 20(2), pp. 211-22.

McKusick-Nathans Institute of Genetic Medicine, J.H.U.B., MD), 2020 (2020) 'Online Mendelian Inheritance in Man, OMIM®'. 2023-05-15. Dec 2020. Available at: https://omim.org/.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P. and Cunningham, F. (2016) 'The Ensembl Variant Effect Predictor', *Genome Biol*, 17(1), p. 122.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P. and Cunningham, F. (2010) 'Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor', *Bioinformatics*, 26(16), pp. 2069-70.

McTague, A., Nair, U., Malhotra, S., Meyer, E., Trump, N., Gazina, E.V., Papandreou, A., Ngoh, A., Ackermann, S., Ambegaonkar, G., Appleton, R., Desurkar, A., Eltze, C., Kneen, R., Kumar, A.V., Lascelles, K., Montgomery, T., Ramesh, V., Samanta, R., Scott, R.H., Tan, J., Whitehouse, W., Poduri, A., Scheffer, I.E., Chong, W.K.K., Cross, J.H., Topf, M., Petrou, S. and Kurian, M.A. (2018) 'Clinical and molecular characterization of KCNT1-related severe early-onset epilepsy', *Neurology*, 90(1), pp. e55-e66.

Medical Research Council Antiepileptic Drug Withdrawal Study, G., Bessant, P., Chadwick, D., Eaton, B., Taylor, J., Holland, A., Joannou, J., Johnson, A.L., Oldfield, L., Reader, N.P., Gumpert, E.J.W., Jacoby, A., Cuckle, H. and Warlow, C. (1991) 'Randomised study of antiepileptic drug withdrawal in patients in remission', *Lancet*, 337(8751), pp. 1175-1180.

Mefford, H.C. (2014) 'CNVs in Epilepsy', *Curr Genet Med Rep*, 2(3), pp. 162-167.

Mefford, H.C., Muhle, H., Ostertag, P., von Spiczak, S., Buysse, K., Baker, C., Franke, A., Malafosse, A., Genton, P., Thomas, P., Gurnett, C.A., Schreiber, S., Bassuk, A.G., Guipponi, M., Stephani, U., Helbig, I. and Eichler, E.E. (2010) 'Genome-wide copy number variation in epilepsy: novel susceptibility loci in idiopathic generalized and focal epilepsies', *PLoS Genet*, 6(5), p. e1000962.

Mefford, H.C., Yendle, S.C., Hsu, C., Cook, J., Geraghty, E., McMahon, J.M., Eeg-Olofsson, O., Sadleir, L.G., Gill, D., Ben-Zeev, B., Lerman-Sagie, T., Mackay, M., Freeman, J.L., Andermann, E., Pelakanos, J.T., Andrews, I., Wallace, G., Eichler, E.E., Berkovic, S.F. and Scheffer, I.E. (2011) 'Rare copy number variants are an important cause of epileptic encephalopathies', *Ann Neurol*, 70(6), pp. 974-85.

Miceli, F., Millevert, C., Soldovieri, M.V., Mosca, I., Ambrosino, P., Carotenuto, L., Schrader, D., Lee, H.K., Riviello, J., Hong, W., Risen, S., Emrick, L., Amin, H., Ville, D., Edery, P., de Bellescize, J., Michaud, V., Van-Gils, J., Goizet, C., Willemsen, M.H., Kleefstra, T., Møller,

R.S., Bayat, A., Devinsky, O., Sands, T., Korenke, G.C., Kluger, G., Mefford, H.C., Brilstra, E., Lesca, G., Milh, M., Cooper, E.C., Taglialatela, M. and Weckhuysen, S. (2022) 'KCNQ2 R144 variants cause neurodevelopmental disability with language impairment and autistic features without neonatal seizures through a gain-of-function mechanism', *EBioMedicine*, 81, p. 104130.

Miehe, S., Bieberstein, A., Arnould, I., Ihdene, O., Rütten, H. and Strübing, C. (2010) 'The phospholipid-binding protein SESTD1 is a novel regulator of the transient receptor potential channels TRPC4 and TRPC5', *J Biol Chem*, 285(16), pp. 12426-34.

Miller, D.T., Adam, M.P., Aradhya, S., Biesecker, L.G., Brothman, A.R., Carter, N.P., Church, D.M., Crolla, J.A., Eichler, E.E., Epstein, C.J., Faucett, W.A., Feuk, L., Friedman, J.M., Hamosh, A., Jackson, L., Kaminsky, E.B., Kok, K., Krantz, I.D., Kuhn, R.M., Lee, C., Ostell, J.M., Rosenberg, C., Scherer, S.W., Spinner, N.B., Stavropoulos, D.J., Tepperberg, J.H., Thorland, E.C., Vermeesch, J.R., Waggoner, D.J., Watson, M.S., Martin, C.L. and Ledbetter, D.H. (2010) 'Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies', *Am J Hum Genet*, 86(5), pp. 749-64.

Miller, L.L., Pellock, J.M., DeLorenzo, R.J., Meyer, J.M. and Corey, L.A. (1998) 'Univariate genetic analyses of epilepsy and seizures in a population-based twin study: the Virginia Twin Registry', *Genet Epidemiol*, 15(1), pp. 33–49.

Minardi, R., Licchetta, L., Baroni, M.C., Pippucci, T., Stipa, C., Mostacci, B., Severi, G., Toni, F., Bergonzini, L., Carelli, V., Seri, M., Tinuper, P. and Bisulli, F. (2020) 'Whole-exome sequencing in adult patients with developmental and epileptic encephalopathy: It is never too late', *Clin Genet*, 98(5), pp. 477-485.

Mishra, S., Sabhlok, S., Panda, P.K. and Khatri, I. (2015) 'Management of Midline Facial Clefts', *Journal of Maxillofacial and Oral Surgery*, 14(4), pp. 883-890.

Mistry, M. and Pavlidis, P. (2008) 'Gene Ontology term overlap as a measure of gene functional similarity', *BMC Bioinformatics*, 9, p. 327.

Mitchell, J.W., Noble, A., Baker, G., Batchelor, R., Brigo, F., Christensen, J., French, J., Gil-Nagel, A., Guekht, A., Jette, N., Kälviäinen, R., Leach, J.P., Maguire, M., O'Brien, T., Rosenow, F., Ryvlin, P., Tittensor, P., Tripathi, M., Trinka, E., Wiebe, S., Williamson, P.R. and Marson, T. (2022) 'Protocol for the development of an international Core Outcome Set for treatment trials in adults with epilepsy: the EPilepsy outcome Set for Effectiveness Trials Project (EPSET)', *Trials*, 23(1), p. 943.

Mizutani, K., Miyata, M., Shiotani, H., Kameyama, T. and Takai, Y. (2022) 'Nectin-2 in general and in the brain', *Mol Cell Biochem*, 477(1), pp. 167-180.

Monlong, J., Girard, S.L., Meloche, C., Cadieux-Dion, M., Andrade, D.M., Lafreniere, R.G., Gravel, M., Spiegelman, D., Dionne-Laporte, A., Boelman, C., Hamdan, F.F., Michaud, J.L., Rouleau, G., Minassian, B.A., Bourque, G. and Cossette, P. (2018) 'Global characterization of copy number variants in epilepsy patients from whole genome sequencing', *PLoS Genet*, 14(4), p. e1007285.

Montanucci, L., Lewis-Smith, D., Collins, R.L., Niestroj, L.M., Parthasarathy, S., Xian, J., Ganesan, S., Macnee, M., Brunger, T., Thomas, R.H., Talkowski, M., Epi, C., Helbig, I., Leu,

C. and Lal, D. (2023) 'Genome-wide identification and phenotypic characterization of seizure-associated copy number variations in 741,075 individuals', *Nat Commun*, 14(1), p. 4392.

Morales, J., Pujar, S., Loveland, J.E., Astashyn, A., Bennett, R., Berry, A., Cox, E., Davidson, C., Ermolaeva, O., Farrell, C.M., Fatima, R., Gil, L., Goldfarb, T., Gonzalez, J.M., Haddad, D., Hardy, M., Hunt, T., Jackson, J., Joardar, V.S., Kay, M., Kodali, V.K., McGarvey, K.M., McMahon, A., Mudge, J.M., Murphy, D.N., Murphy, M.R., Rajput, B., Rangwala, S.H., Riddick, L.D., Thibaud-Nissen, F., Threadgold, G., Vatsan, A.R., Wallin, C., Webb, D., Flicek, P., Birney, E., Pruitt, K.D., Frankish, A., Cunningham, F. and Murphy, T.D. (2022) 'A joint NCBI and EMBL-EBI transcript set for clinical genomics and research', *Nature*, 604(7905), pp. 310-315.

Moreau, C., Rebillard, R.M., Wolking, S., Michaud, J., Tremblay, F., Girard, A., Bouchard, J., Minassian, B., Laprise, C., Cossette, P. and Girard, S.L. (2020) 'Polygenic risk scores of several subtypes of epilepsies in a founder population', *Neurol Genet*, 6(3), p. e416.

Mori, Y., Downs, J., Wong, K., Anderson, B., Epstein, A. and Leonard, H. (2017) 'Impacts of caring for a child with the CDKL5 disorder on parental wellbeing and family quality of life', *Orphanet J Rare Dis*, 12(1), p. 16.

Mulkey, S.B., Ben-Zeev, B., Nicolai, J., Carroll, J.L., Grønborg, S., Jiang, Y.H., Joshi, N., Kelly, M., Koolen, D.A., Mikati, M.A., Park, K., Pearl, P.L., Scheffer, I.E., Spillmann, R.C., Taglialatela, M., Vieker, S., Weckhuysen, S., Cooper, E.C. and Cilio, M.R. (2017) 'Neonatal nonepileptic myoclonus is a prominent clinical feature of KCNQ2 gain-of-function variants R201C and R201H', *Epilepsia*, 58(3), pp. 436-445.

Mullen, S.A., Carvill, G.L., Bellows, S., Bayly, M.A., Trucks, H., Lal, D., Sander, T., Berkovic, S.F., Dibbens, L.M., Scheffer, I.E. and Mefford, H.C. (2013) 'Copy number variants are frequent in genetic generalized epilepsy with intellectual disability', *Neurology*, 81(17), pp. 1507-14.

Muller, A., Helbig, I., Jansen, C., Bast, T., Guerrini, R., Jahn, J., Muhle, H., Auvin, S., Korenke, G.C., Philip, S., Keimer, R., Striano, P., Wolf, N.I., Pust, B., Thiels, C., Fogarasi, A., Waltz, S., Kurlemann, G., Kovacevic-Preradovic, T., Ceulemans, B., Schmitt, B., Philippi, H., Tarquinio, D., Buerki, S., von Stulpnagel, C. and Kluger, G. (2016) 'Retrospective evaluation of low long-term efficacy of antiepileptic drugs and ketogenic diet in 39 patients with CDKL5-related epilepsy', *Eur J Paediatr Neurol*, 20(1), pp. 147-51.

Musante, L., Costa, P., Zanus, C., Faletra, F., Murru, F.M., Bianco, A.M., La Bianca, M., Ragusa, G., Athanasakis, E., d'Adamo, A.P., Carrozzi, M. and Gasparini, P. (2022) 'The Genetic Diagnosis of Ultrarare DEEs: An Ongoing Challenge', *Genes (Basel)*, 13(3).

Myers, K.A., McGlade, A., Neubauer, B.A., Lal, D., Berkovic, S.F., Scheffer, I.E. and Hildebrand, M.S. (2018) 'KANSL1 variation is not a major contributing factor in self-limited focal epilepsy syndromes of childhood', *PLoS One*, 13(1), p. e0191546.

Nascimento, F.A., Hood, V., Yap, S.V., Sheikh, I.S., Anne Meskis, M. and Thiele, E.A. (2023) 'Evaluating adult care in Dravet syndrome upon transferring from pediatrics in the U.S.: A caregiver-based survey study', *Epilepsy Behav*, 147, p. 109368.

National Research Council (2011) *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. Washington (DC).

Ndimubanzi, P.C., Carabin, H., Budke, C.M., Nguyen, H., Qian, Y.J., Rainwater, E., Dickey, M., Reynolds, S. and Stoner, J.A. (2010) 'A systematic review of the frequency of neurocyticercosis with a focus on people with epilepsy', *PLoS Negl Trop Dis*, 4(11), p. e870.

Nelson, K.B. and Ellenberg, J.H. (1976) 'Predictors of epilepsy in children who have experienced febrile seizures', *N Engl J Med*, 295(19), pp. 1029-33.

Nesbitt, G., McKenna, K., Mays, V., Carpenter, A., Miller, K., Williams, M. and Investigators, E. (2013) 'The Epilepsy Phenome/Genome Project (EPGP) informatics platform', *Int J Med Inform*, 82(4), pp. 248-59.

NHS England (2024) *NHS Genomic Medicine Service*. Available at: https://www.england.nhs.uk/genomics/nhs-genomic-med-service/ (Accessed: 2024-03-24).

NHS Health Education England (2020) *Requesting whole genome sequencing: information for clinicians*. Available at: https://www.genomicseducation.hee.nhs.uk/supporting-the-nhs-genomic-medicine-service/requesting-whole-genome-sequencing-information-for-clinicians/ (Accessed: Dec 2020).

Nicholl, J.S. (2003) 'Of Cabbages and Kings: Some considerations on classifications, diagnostic schemes, semiology, and concepts', *Epilepsia*, 44(7), p. 988.

Niestroj, L.M., Perez-Palma, E., Howrigan, D.P., Zhou, Y., Cheng, F., Saarentaus, E., Nürnberg, P., Stevelink, R., Daly, M.J., Palotie, A. and Lal, D. (2020) 'Epilepsy subtype-specific copy number burden observed in a genome-wide study of 17 458 subjects', *Brain*, 143(7), pp. 2106-2118.

Ning, L., Jansen, R. and Gerstein, M. (2002) 'Toward a systematic definition of protein function that scales to the genome level: defining function in terms of interactions', *Proc IEEE*, 90(12), pp. 1848-1858.

Northrup, H., Aronow, M.E., Bebin, E.M., Bissler, J., Darling, T.N., de Vries, P.J., Frost, M.D., Fuchs, Z., Gosnell, E.S., Gupta, N., Jansen, A.C., Jozwiak, S., Kingswood, J.C., Knilans, T.K., McCormack, F.X., Pounders, A., Roberds, S.L., Rodriguez-Buritica, D.F., Roth, J., Sampson, J.R., Sparagana, S., Thiele, E.A., Weiner, H.L., Wheless, J.W., Towbin, A.J., Krueger, D.A. and International Tuberous Sclerosis Complex Consensus, G. (2021) 'Updated International Tuberous Sclerosis Complex Diagnostic Criteria and Surveillance and Management Recommendations', *Pediatr Neurol*, 123, pp. 50-66.

O'Roak, B.J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J.J., Girirajan, S., Karakoc, E., Mackenzie, A.P., Ng, S.B., Baker, C., Rieder, M.J., Nickerson, D.A., Bernier, R., Fisher, S.E., Shendure, J. and Eichler, E.E. (2011) 'Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations', *Nat Genet*, 43(6), pp. 585-9.

O'Roak, B.J., Vives, L., Fu, W., Egertson, J.D., Stanaway, I.B., Phelps, I.G., Carvill, G., Kumar, A., Lee, C., Ankenman, K., Munson, J., Hiatt, J.B., Turner, E.H., Levy, R., O'Day, D.R., Krumm, N., Coe, B.P., Martin, B.K., Borenstein, E., Nickerson, D.A., Mefford, H.C., Doherty, D., Akey, J.M., Bernier, R., Eichler, E.E. and Shendure, J. (2012) 'Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders', *Science*, 338(6114), pp. 1619-22.

Ohta, T., Gray, T.A., Rogan, P.K., Buiting, K., Gabriel, J.M., Saitoh, S., Muralidhar, B., Bilienska, B., Krajewska-Walasek, M., Driscoll, D.J., Horsthemke, B., Butler, M.G. and

Nicholls, R.D. (1999) 'Imprinting-mutation mechanisms in Prader-Willi syndrome', *Am J Hum Genet*, 64(2), pp. 397-413.

Okur, V., Hamm, L., Kavus, H., Mebane, C., Robinson, S., Levy, B. and Chung, W.K. (2021) 'Clinical and genomic characterization of 8p cytogenomic disorders', *Genet Med*, 23(12), pp. 2342-2351.

Oliver, K.L., Ellis, C.A., Scheffer, I.E., Ganesan, S., Leu, C., Sadleir, L.G., Heinzen, E.L., Mefford, H.C., Bass, A.J., Curtis, S.W., Harris, R.V., Whiteman, D.C., Helbig, I., Ottman, R., Epstein, M.P., Bahlo, M. and Berkovic, S.F. (2022) 'Common risk variants for epilepsy are enriched in families previously targeted for rare monogenic variant discovery', *EBioMedicine*, 81, p. 104079.

Oliver, K.L., Scheffer, I.E., Bennett, M.F., Grinton, B.E., Bahlo, M. and Berkovic, S.F. (2023) 'Genes4Epilepsy: An epilepsy gene resource', *Epilepsia*, 64(5), pp. 1368-1375.

Olson, H., Shen, Y., Avallone, J., Sheidley, B.R., Pinsky, R., Bergin, A.M., Berry, G.T., Duffy, F.H., Eksioglu, Y., Harris, D.J., Hisama, F.M., Ho, E., Irons, M., Jacobsen, C.M., James, P., Kothare, S., Khwaja, O., Lipton, J., Loddenkemper, T., Markowitz, J., Maski, K., Megerian, J.T., Neilan, E., Raffalli, P.C., Robbins, M., Roberts, A., Roe, E., Rollins, C., Sahin, M., Sarco, D., Schonwald, A., Smith, S.E., Soul, J., Stoler, J.M., Takeoka, M., Tan, W.H., Torres, A.R., Tsai, P., Urion, D.K., Weissman, L., Wolff, R., Wu, B.L., Miller, D.T. and Poduri, A. (2014) 'Copy number variation plays an important role in clinical epilepsy', *Ann Neurol*, 75(6), pp. 943-58.

Ortega-Moreno, L., Giráldez, B.G., Soto-Insuga, V., Losada-Del Pozo, R., Rodrigo-Moreno, M., Alarcón-Morcillo, C., Sánchez-Martín, G., Díaz-Gómez, E., Guerrero-López, R. and Serratosa, J.M. (2017) 'Molecular diagnosis of patients with epilepsy and developmental delay using a customized panel of epilepsy genes', *PLoS One*, 12(11), p. e0188978.

Otte, W.M., Singla, M., Sander, J.W. and Singh, G. (2013) 'Drug therapy for solitary cysticercus granuloma: a systematic review and meta-analysis', *Neurology*, 80(2), pp. 152-62.

Pal, D. and Helbig, I. (2015) 'Commentary: Pathogenic EFHC1 mutations are tolerated in healthy individuals dependent on reported ancestry', *Epilepsia*, 56(2), pp. 195–6.

Pal, D.K., Evgrafov, O.V., Tabares, P., Zhang, F., Durner, M. and Greenberg, D.A. (2003) 'BRD2 (RING3) is a probable major susceptibility gene for common juvenile myoclonic epilepsy', *Am J Hum Genet*, 73(2), pp. 261-70.

Palmini, A., Akamatsu, N., Bast, T., Bauer, S., Baumgartner, C., Benbadis, S., Bermeo-Ovalle, A., Beyenburg, S., Bleasel, A., Bozorgi, A., Brazdil, M., Carreno, M., Delanty, N., Devereaux, M., Duncan, J.S., Fernandez-Baca Vaca, G., Garcia Losarcos, N., Ghanma, L., Gil-Nagel, A., Hamer, H., Holthausen, H., Omidi, S.J., Kahane, P., Kalamangalam, G., Kanner, A., Knake, S., Kovac, S., Kraemer, G., Kurlemann, G., Lacuey, N., Landazuri, P., Hui Lim, S., LoRusso, G., Luders, H., Mani, J., Matsumoto, R., Miller, J., Noachtar, S., O'Dwyer, R., Park, J., Reif, P.S., Remi, J., Rosenow, F., Sakamoto, A., Schubert-Bast, S., Schuele, S., Shahid, A., Steinhoff, J.B., Strzelczyk, A., Szabo, C.A., Tandon, N., Terada, K., Toledo, M., van Emde Boas, W., Walker, M. and Widdess-Walsh, P. (2020) 'From theory to practice: Critical points in the 2017 ILAE classification of epileptic seizures and epilepsies', *Epilepsia*, 61(2), pp. 350–353.

Panayiotopoulos, C.P. (2011) 'The new ILAE report on terminology and concepts for organization of epileptic seizures: a clinician's critical view and contribution', *Epilepsia*, 52(12), pp. 2155–60.

Panoutsopoulos, A.A., De Crescenzo, A.H., Lee, A., Lu, A.M., Ross, A.P., Borodinsky, L.N., Marcucio, R., Trainor, P.A. and Zarbalis, K.S. (2020) 'Pak1ip1 Loss-of-Function Leads to Cell Cycle Arrest, Loss of Neural Crest Cells, and Craniofacial Abnormalities', *Front Cell Dev Biol*, 8, p. 510063.

Parrini, E., Ramazzotti, A., Dobyns, W.B., Mei, D., Moro, F., Veggiotti, P., Marini, C., Brilstra, E.H., Dalla Bernardina, B., Goodwin, L., Bodell, A., Jones, M.C., Nangeroni, M., Palmeri, S., Said, E., Sander, J.W., Striano, P., Takahashi, Y., Van Maldergem, L., Leonardi, G., Wright, M., Walsh, C.A. and Guerrini, R. (2006) 'Periventricular heterotopia: phenotypic heterogeneity and correlation with Filamin A mutations', *Brain*, 129(Pt 7), pp. 1892-906.

Pathak, S., Miller, J., Morris, E.C., Stewart, W.C.L. and Greenberg, D.A. (2018) 'DNA methylation of the BRD2 promoter is associated with juvenile myoclonic epilepsy in Caucasians', *Epilepsia*, 59(5), pp. 1011-1019.

Pekar, V. and Staab, S. (2002). Available at: https://aclanthology.org/C02-1090.

Peljto, A.L., Barker-Cummings, C., Vasoli, V.M., Leibson, C.L., Hauser, W.A., Buchhalter, J.R. and Ottman, R. (2014) 'Familial risk of epilepsy: a population-based study', *Brain*, 137(Pt 3), pp. 795-805.

Peng, J., Zhou, Y. and Wang, K. (2021) 'Multiplex gene and phenotype network to characterize shared genetic pathways of epilepsy and autism', *Sci Rep*, 11(1), p. 952.

Pérez-Palma, E., Helbig, I., Klein, K.M., Anttila, V., Horn, H., Reinthaler, E.M., Gormley, P., Ganna, A., Byrnes, A., Pernhorst, K., Toliat, M.R., Saarentaus, E., Howrigan, D.P., Hoffman, P., Miquel, J.F., De Ferrari, G.V., Nürnberg, P., Lerche, H., Zimprich, F., Neubauer, B.A., Becker, A.J., Rosenow, F., Perucca, E., Zara, F., Weber, Y.G. and Lal, D. (2017) 'Heterogeneous contribution of microdeletions in the development of common generalised and focal epilepsies', *J Med Genet*, 54(9), pp. 598-606.

Pesquita, C. (2017) 'Semantic Similarity in the Gene Ontology', in Dessimoz, C. and Škunca, N. (eds.) *The Gene Ontology Handbook*. New York, NY: Springer New York, pp. 161-173.

Pesquita, C., Faria, D., Bastos, H., Ferreira, A.E., Falcao, A.O. and Couto, F.M. (2008) 'Metrics for GO based protein semantic similarity: a systematic evaluation', *BMC Bioinformatics*, 9 Suppl 5(Suppl 5), p. S4.

Pesquita, C., Faria, D., Bastos, H.P., Falcão, A.O. and Couto, F.M. (2007). Available at: https://webpages.ciencias.ulisboa.pt/~fjcouto/files/workshop%20cpesquita-ismb2007.pdf.

Pesquita, C., Faria, D., Falcao, A.O., Lord, P. and Couto, F.M. (2009) 'Semantic similarity in biomedical ontologies', *PLoS Comput Biol*, 5(7), p. e1000443.

Phillips, H.A., Scheffer, I.E., Berkovic, S.F., Hollway, G.E., Sutherland, G.R. and Mulley, J.C. (1995) 'Localization of a gene for autosomal dominant nocturnal frontal lobe epilepsy to chromosome 20q 13.2', *Nat Genet*, 10(1), pp. 117–8.

Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., Thiruvahindrapuram, B., Xu, X., Ziman, R., Wang, Z., Vorstman, J.A., Thompson, A., Regan, R., Pilorge, M.,

Pellecchia, G., Pagnamenta, A.T., Oliveira, B., Marshall, C.R., Magalhaes, T.R., Lowe, J.K., Howe, J.L., Griswold, A.J., Gilbert, J., Duketis, E., Dombroski, B.A., De Jonge, M.V., Cuccaro, M., Crawford, E.L., Correia, C.T., Conroy, J., Conceição, I.C., Chiocchetti, A.G., Casey, J.P., Cai, G., Cabrol, C., Bolshakova, N., Bacchelli, E., Anney, R., Gallinger, S., Cotterchio, M., Casey, G., Zwaigenbaum, L., Wittemeyer, K., Wing, K., Wallace, S., van Engeland, H., Tryfon, A., Thomson, S., Soorya, L., Rogé, B., Roberts, W., Poustka, F., Mouga, S., Minshew, N., McInnes, L.A., McGrew, S.G., Lord, C., Leboyer, M., Le Couteur, A.S., Kolevzon, A., Jiménez González, P., Jacob, S., Holt, R., Guter, S., Green, J., Green, A., Gillberg, C., Fernandez, B.A., Duque, F., Delorme, R., Dawson, G., Chaste, P., Café, C., Brennan, S., Bourgeron, T., Bolton, P.F., Bölte, S., Bernier, R., Baird, G., Bailey, A.J., Anagnostou, E., Almeida, J., Wijsman, E.M., Vieland, V.J., Vicente, A.M., Schellenberg, G.D., Pericak-Vance, M., Paterson, A.D., Parr, J.R., Oliveira, G., Nurnberger, J.I., Monaco, A.P., Maestrini, E., Klauck, S.M., Hakonarson, H., Haines, J.L., Geschwind, D.H., Freitag, C.M., Folstein, S.E., Ennis, S., et al. (2014) 'Convergence of genes and cellular pathways dysregulated in autism spectrum disorders', *Am J Hum Genet*, 94(5), pp. 677-94.

Porubsky, D., Harvey, W.T., Rozanski, A.N., Ebler, J., Höps, W., Ashraf, H., Hasenfeld, P., Paten, B., Sanders, A.D., Marschall, T., Korbel, J.O. and Eichler, E.E. (2023) 'Inversion polymorphism in a complete human genome assembly', *Genome Biol*, 24(1), p. 100.

Postma, A., Minderhoud, C.A., Otte, W.M., Jansen, F.E., Gunning, W.B., Verhoeven, J.S., Jongmans, M.J., Zinkstok, J.R. and Brilstra, E.H. (2024) 'Understanding neurodevelopmental trajectories and behavioral profiles in SCN1A-related epilepsy syndromes', *Epilepsy Behav*, 154.

Pressler, R.M., Cilio, M.R., Mizrahi, E.M., Moshe, S.L., Nunes, M.L., Plouin, P., Vanhatalo, S., Yozawitz, E., de Vries, L.S., Puthenveettil Vinayan, K., Triki, C.C., Wilmshurst, J.M., Yamamoto, H. and Zuberi, S.M. (2021) 'The ILAE classification of seizures and the epilepsies: Modification for seizures in the neonate. Position paper by the ILAE Task Force on Neonatal Seizures', *Epilepsia*, 62(3), pp. 615-628.

Pruett, D.G., Shaw, D.M., Chen, H.H., Petty, L.E., Polikowsky, H.G., Kraft, S.J., Jones, R.M. and Below, J.E. (2021) 'Identifying developmental stuttering and associated comorbidities in electronic health records and creating a phenome risk classifier', *J Fluency Disord*, 68, p. 105847.

R Core Team (2017) 'R: a language and environment for statistical computing'. R Foundation for Statistical Computing.

Rada, R., Mili, H., Bicknell, E. and Blettner, M. (1989) 'Development and application of a metric on semantic nets', *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1), pp. 17-30.

Radu, M., Semenova, G., Kosoff, R. and Chernoff, J. (2014) 'PAK signalling during the development and progression of cancer', *Nat Rev Cancer*, 14(1), pp. 13-25.

Rajpurkar, P., Chen, E., Banerjee, O. and Topol, E.J. (2022) 'AI in health and medicine', *Nat Med*, 28(1), pp. 31-38.

Rao, A., Vg, S., Joseph, T., Kotte, S., Sivadasan, N. and Srinivasan, R. (2018) 'Phenotype-driven gene prioritization for rare diseases using graph convolution on heterogeneous networks', *BMC Med Genomics*, 11(1), p. 57.

Rare-X (2024). Available at: https://rare-x.org/about/ (Accessed: 2024-03-24).

Rathore, C., Prakash, S. and Makwana, P. (2020) 'Prevalence of photoparoxysmal response in patients with epilepsy: Effect of the underlying syndrome and treatment status', *Seizure*, 82, pp. 39-43.

Ravindra, V.M., Karsy, M., Schmidt, R.H., Taussky, P., Park, M.S. and Bollo, R.J. (2016) 'Rapid de novo aneurysm formation after clipping of a ruptured middle cerebral artery aneurysm in an infant with an MYH11 mutation', *J Neurosurg Pediatr*, 18(4), pp. 463-470.

Reinthaler, E.M., Lal, D., Lebon, S., Hildebrand, M.S., Dahl, H.H., Regan, B.M., Feucht, M., Steinböck, H., Neophytou, B., Ronen, G.M., Roche, L., Gruber-Sedlmayr, U., Geldner, J., Haberlandt, E., Hoffmann, P., Herms, S., Gieger, C., Waldenberger, M., Franke, A., Wittig, M., Schoch, S., Becker, A.J., Hahn, A., Männik, K., Toliat, M.R., Winterer, G., Lerche, H., Nürnberg, P., Mefford, H., Scheffer, I.E., Berkovic, S.F., Beckmann, J.S., Sander, T., Jacquemont, S., Reymond, A., Zimprich, F. and Neubauer, B.A. (2014) '16p11.2 600 kb Duplications confer risk for typical and atypical Rolandic epilepsy', *Hum Mol Genet*, 23(22), pp. 6069-80.

Reis, M.C., Patrun, J., Ackl, N., Winter, P., Scheifele, M., Danek, A. and Nolte, D. (2022) 'A Severe Dementia Syndrome Caused by Intron Retention and Cryptic Splice Site Activation in STUB1 and Exacerbated by TBP Repeat Expansions', *Front Mol Neurosci*, 15, p. 878236.

Ren, P., Sun, X., Zhang, C., Wang, L., Xing, B. and Du, X. (2019) 'Human UTP14a promotes angiogenesis through upregulating PDGFA expression in colorectal cancer', *Biochem Biophys Res Commun*, 512(4), pp. 871-876.

Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. and Kircher, M. (2019) 'CADD: predicting the deleteriousness of variants throughout the human genome', *Nucleic Acids Res*, 47(D1), pp. D886-D894.

Resnik, P. (1995) *14th International Joint Conference on Artificial Intelligence*. San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Resnik, P. (1999) 'Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language', *J Artif Int Res*, 11(1), pp. 95-130.

Rett, A. (1966) 'On a unusual brain atrophy syndrome in hyperammonemia in childhood', *Wien Med Wochenschr*, 116(37), pp. 723-6.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H.L. and Committee, A.L.Q.A. (2015) 'Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology', *Genet Med*, 17(5), pp. 405-24.

Richter, M., Murtaza, N., Scharrenberg, R., White, S.H., Johanns, O., Walker, S., Yuen, R.K.C., Schwanke, B., Bedürftig, B., Henis, M., Scharf, S., Kraus, V., Dörk, R., Hellmann, J., Lindenmaier, Z., Ellegood, J., Hartung, H., Kwan, V., Sedlacik, J., Fiehler, J., Schweizer, M., Lerch, J.P., Hanganu-Opatz, I.L., Morellini, F., Scherer, S.W., Singh, K.K. and Calderon de Anda, F. (2019) 'Altered TAOK2 activity causes autism-related neurodevelopmental and cognitive abnormalities through RhoA signaling', *Mol Psychiatry*, 24(9), pp. 1329-1350.

Riggs, E.R., Andersen, E.F., Cherry, A.M., Kantarci, S., Kearney, H., Patel, A., Raca, G., Ritter, D.I., South, S.T., Thorland, E.C., Pineda-Alvarez, D., Aradhya, S. and Martin, C.L. (2020) 'Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen)', *Genet Med*, 22(2), pp. 245-257.

Riggs, E.R., Andersen, E.F., Cherry, A.M., Kantarci, S., Kearney, H., Patel, A., Raca, G., Ritter, D.I., South, S.T., Thorland, E.C., Pineda-Alvarez, D., Aradhya, S. and Martin, C.L. (2021) 'Correction: Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen)', *Genet Med*, 23(11), p. 2230.

Riney, K., Bogacz, A., Somerville, E., Hirsch, E., Nabbout, R., Scheffer, I.E., Zuberi, S.M., Alsaadi, T., Jain, S., French, J., Specchio, N., Trinka, E., Wiebe, S., Auvin, S., Cabral-Lim, L., Naidoo, A., Perucca, E., Moshe, S.L., Wirrell, E.C. and Tinuper, P. (2022) 'International League Against Epilepsy classification and definition of epilepsy syndromes with onset at a variable age: position statement by the ILAE Task Force on Nosology and Definitions', *Epilepsia*, 63(6), pp. 1443-1474.

Robinson, P.N., Kohler, S., Bauer, S., Seelow, D., Horn, D. and Mundlos, S. (2008) 'The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease', *Am J Hum Genet*, 83(5), pp. 610-5.

Robinson, P.N., Ravanmehr, V., Jacobsen, J.O.B., Danis, D., Zhang, X.A., Carmody, L.C., Gargano, M.A., Thaxton, C.L., Core, U.N.C.B., Karlebach, G., Reese, J., Holtgrewe, M., Kohler, S., McMurry, J.A., Haendel, M.A. and Smedley, D. (2020) 'Interpretable Clinical Genomics with a Likelihood Ratio Paradigm', *Am J Hum Genet*, 107(3), pp. 403-417.

Rodriguez-Garcia, M.A., Gkoutos, G.V., Schofield, P.N. and Hoehndorf, R. (2017) 'Integrating phenotype ontologies with PhenomeNET', *J Biomed Semantics*, 8(1), p. 58.

Rong, M., Benke, T., Zulfiqar Ali, Q., Aledo-Serrano, Á., Bayat, A., Rossi, A., Devinsky, O., Qaiser, F., Ali, A.S., Fasano, A., Bassett, A.S. and Andrade, D.M. (2023) 'Adult Phenotype of SYNGAP1-DEE', *Neurol Genet*, 9(6), p. e200105.

Rosenow, F., Akamatsu, N., Bast, T., Bauer, S., Baumgartner, C., Benbadis, S., Bermeo-Ovalle, A., Beyenburg, S., Bleasel, A., Bozorgi, A., Brazdil, M., Carreno, M., Delanty, N., Devereaux, M., Duncan, J., Fernandez-Baca Vaca, G., Francione, S., Garcia Losarcos, N., Ghanma, L., Gil-Nagel, A., Hamer, H., Holthausen, H., Omidi, S.J., Kahane, P., Kalamangalam, G., Kanner, A., Knake, S., Kovac, S., Krakow, K., Kramer, G., Kurlemann, G., Lacuey, N., Landazuri, P., Lim, S.H., Londono, L.V., LoRusso, G., Luders, H., Mani, J., Matsumoto, R., Miller, J., Noachtar, S., O'Dwyer, R., Palmini, A., Park, J., Reif, P.S., Remi, J., Sakamoto, A.C., Schmitz, B., Schubert-Bast, S., Schuele, S., Shahid, A., Steinhoff, B., Strzelczyk, A., Szabo, C.A., Tandon, N., Terada, K., Toledo, M., van Emde Boas, W., Walker, M. and Widdess-Walsh, P. (2020) 'Could the 2017 ILAE and the four-dimensional epilepsy classifications be merged to a new "Integrated Epilepsy Classification"?', *Seizure*, 78, pp. 31-37.

Roshandel, D., Sanders, E.J., Shakeshaft, A., Panjwani, N., Lin, F., Collingwood, A., Hall, A., Keenan, K., Deneubourg, C., Mirabella, F., Topp, S., Zarubova, J., Thomas, R.H., Talvik, I., Syvertsen, M., Striano, P., Smith, A.B., Selmer, K.K., Rubboli, G., Orsini, A., Ng, C.C., Møller, R.S., Lim, K.S., Hamandi, K., Greenberg, D.A., Gesche, J., Gardella, E., Fong, C.Y., Beier, C.P., Andrade, D.M., Jungbluth, H., Richardson, M.P., Pastore, A., Fanto, M., Pal, D.K. and

Strug, L.J. (2023) 'SLCO5A1 and synaptic assembly genes contribute to impulsivity in juvenile myoclonic epilepsy', *NPJ Genom Med*, 8(1), p. 28.

Ross, A.P., Mansilla, M.A., Choe, Y., Helminski, S., Sturm, R., Maute, R.L., May, S.R., Hozyasz, K.K., Wójcicki, P., Mostowska, A., Davidson, B., Adamopoulos, I.E., Pleasure, S.J., Murray, J.C. and Zarbalis, K.S. (2013) 'A mutation in mouse Pak1ip1 causes orofacial clefting while human PAK1IP1 maps to 6p24 translocation breaking points associated with orofacial clefting', *PLoS One*, 8(7), p. e69333.

Royal College of Paediatrics and Child Health, Department of Health and Scottish Government (2009) *Boys UK-WHO Neonatal and infant close monitoring growth chart: 23 weeks gestation to 2 years corrected age*. Available at: https://www.rcpch.ac.uk/sites/default/files/Boys_neonatal_and_infant_close_monitoring_growth_chart.pdf (Accessed: 2023-11-07).

Sahoo, S.S., Lhatoo, S.D., Gupta, D.K., Cui, L., Zhao, M., Jayapandian, C., Bozorgi, A. and Zhang, G.Q. (2014) 'Epilepsy and seizure ontology: towards an epilepsy informatics infrastructure for clinical research and patient care', *J Am Med Inform Assoc*, 21(1), pp. 82-9.

Samocha, K.E., Kosmicki, J.A., Karczewski, K.J., O'Donnell-Luria, A.H., Pierce-Hoffman, E., MacArthur, D.G., Neale, B.M. and Daly, M.J. (2017) 'Regional missense constraint improves variant deleteriousness prediction', *bioRxiv*, p. 148353.

Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A., Wall, D.P., MacArthur, D.G., Gabriel, S.B., DePristo, M., Purcell, S.M., Palotie, A., Boerwinkle, E., Buxbaum, J.D., Cook, E.H., Gibbs, R.A., Schellenberg, G.D., Sutcliffe, J.S., Devlin, B., Roeder, K., Neale, B.M. and Daly, M.J. (2014) 'A framework for the interpretation of de novo mutation in human disease', *Nat Genet*, 46(9), pp. 944-950.

Sanchez Fernandez, I., Loddenkemper, T., Gainza-Lein, M., Sheidley, B.R. and Poduri, A. (2019) 'Diagnostic yield of genetic tests in epilepsy: A meta-analysis and cost-effectiveness study', *Neurology*, 92(5), pp. e418-28.

Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S., Goldberg, A.P., Jinlu, C., Keaney, J.F., 3rd, Klei, L., Mandell, J.D., Moreno-De-Luca, D., Poultney, C.S., Robinson, E.B., Smith, L., Solli-Nowlan, T., Su, M.Y., Teran, N.A., Walker, M.F., Werling, D.M., Beaudet, A.L., Cantor, R.M., Fombonne, E., Geschwind, D.H., Grice, D.E., Lord, C., Lowe, J.K., Mane, S.M., Martin, D.M., Morrow, E.M., Talkowski, M.E., Sutcliffe, J.S., Walsh, C.A., Yu, T.W., Ledbetter, D.H., Martin, C.L., Cook, E.H., Buxbaum, J.D., Daly, M.J., Devlin, B., Roeder, K. and State, M.W. (2015) 'Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci', *Neuron*, 87(6), pp. 1215-1233.

Scheffer, I.E., Berkovic, S., Capovilla, G., Connolly, M.B., French, J., Guilhoto, L., Hirsch, E., Jain, S., Mathern, G.W., Moshe, S.L., Nordli, D.R., Perucca, E., Tomson, T., Wiebe, S., Zhang, Y.H. and Zuberi, S.M. (2017) 'ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology', *Epilepsia*, 58(4), pp. 512-521.

Scheffer, I.E. and Berkovic, S.F. (1997) 'Generalized epilepsy with febrile seizures plus. A genetic disorder with heterogeneous clinical phenotypes', *Brain*, 120 (Pt 3), pp. 479-90.

Scheffer, I.E., French, J., Hirsch, E., Jain, S., Mathern, G.W., Moshe, S.L., Perucca, E., Tomson, T., Wiebe, S., Zhang, Y.H. and Zuberi, S.M. (2016) 'Classification of the epilepsies: New concepts for discussion and debate-Special report of the ILAE Classification Task Force of the Commission for Classification and Terminology', *Epilepsia Open*, 1(1-2), pp. 37–44.

Scheffer, I.E., Phillips, H.A., O'Brien, C.E., Saling, M.M., Wrennall, J.A., Wallace, R.H., Mulley, J.C. and Berkovic, S.F. (1998) 'Familial partial epilepsy with variable foci: a new partial epilepsy syndrome with suggestion of linkage to chromosome 2', *Ann Neurol*, 44(6), pp. 890–9.

Schlicker, A., Domingues, F.S., Rahnenfuhrer, J. and Lengauer, T. (2006) 'A new measure for functional similarity of gene products based on Gene Ontology', *BMC Bioinformatics*, 7, p. 302.

Schoech, A.P., Jordan, D.M., Loh, P.R., Gazal, S., O'Connor, L.J., Balick, D.J., Palamara, P.F., Finucane, H.K., Sunyaev, S.R. and Price, A.L. (2019) 'Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection', *Nat Commun*, 10(1), p. 790.

Schol-Gelok, S., Janssens, A.C., Tiemeier, H., Liu, F., Lopez-Leon, S., Zorkoltseva, I.V., Axenovich, T.I., van Swieten, J.C., Uitterlinden, A.G., Hofman, A., Aulchenko, Y.S., Oostra, B.A. and van Duijn, C.M. (2010) 'A genome-wide screen for depression in two independent Dutch populations', *Biol Psychiatry*, 68(2), pp. 187-96.

Schulz, H., Ruppert, A.K., Zara, F., Madia, F., Iacomino, M., M, S.V., Balagura, G., Minetti, C., Striano, P., Bianchi, A., Marini, C., Guerrini, R., Weber, Y.G., Becker, F., Lerche, H., Kapser, C., Schankin, C.J., Kunz, W.S., Moller, R.S., Oliver, K.L., Bellows, S.T., Mullen, S.A., Berkovic, S.F., Scheffer, I.E., Caglayan, H., Ozbek, U., Hoffmann, P., Schramm, S., Tsortouktzidis, D., Becker, A.J. and Sander, T. (2019) 'No evidence for a BRD2 promoter hypermethylation in blood leukocytes of Europeans with juvenile myoclonic epilepsy', *Epilepsia*, 60(5), pp. e31–e36.

Schulz, M.H., Kohler, S., Bauer, S. and Robinson, P.N. (2011) 'Exact score distribution computation for ontological similarity searches', *BMC Bioinformatics*, 12, p. 441.

Seefelder, M. and Kochanek, S. (2021) 'A meta-analysis of transcriptomic profiles of Huntington's disease patients', *PLoS One*, 16(6), p. e0253037.

Selvarajah, A., Gorodetsky, C., Marques, P., Zulfiqar Ali, Q., Berg, A.T., Fasano, A. and Andrade, D.M. (2022) 'Progressive Worsening of Gait and Motor Abnormalities in Older Adults With Dravet Syndrome', *Neurology*, 98(22), pp. e2204-e2210.

Selvarajah, A., Zulfiqar-Ali, Q., Marques, P., Rong, M. and Andrade, D.M. (2021) 'A systematic review of adults with Dravet syndrome', *Seizure*, 87, pp. 39-45.

Semenova, G. and Chernoff, J. (2017) 'Targeting PAK1', *Biochem Soc Trans*, 45(1), pp. 79-88.

Sevilla, J.L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J.M., Martinez-Cruz, L.A., Corrales, F.J. and Rubio, A. (2005) 'Correlation between gene expression and GO semantic similarity', *IEEE/ACM Trans Comput Biol Bioinform*, 2(4), pp. 330-8.

Shaw, D.M., Polikowsky, H.P., Pruett, D.G., Chen, H.H., Petty, L.E., Viljoen, K.Z., Beilby, J.M., Jones, R.M., Kraft, S.J. and Below, J.E. (2021) 'Phenome risk classification enables

phenotypic imputation and gene discovery in developmental stuttering', *Am J Hum Genet*, 108(12), pp. 2271-2283.

Sheidley, B.R., Malinowski, J., Bergner, A.L., Bier, L., Gloss, D.S., Mu, W., Mulhern, M.M., Partack, E.J. and Poduri, A. (2022) 'Genetic testing for the epilepsies: A systematic review', *Epilepsia*, 63(2), pp. 375-387.

Shen, F., Wang, L. and Liu, H. (2017) 'Phenotypic Analysis of Clinical Narratives Using Human Phenotype Ontology', *Stud Health Technol Inform*, 245, pp. 581-585.

Shiotani, H., Miyata, M., Kameyama, T., Mandai, K., Yamasaki, M., Watanabe, M., Mizutani, K. and Takai, Y. (2021) 'Nectin-2α is localized at cholinergic neuron dendrites and regulates synapse formation in the medial habenula', *J Comp Neurol*, 529(2), pp. 450-477.

Shyr, C., Tarailo-Graovac, M., Gottlieb, M., Lee, J.J.Y., van Karnebeek, C. and Wasserman, W.W. (2014) 'FLAGS, frequently mutated genes in public exomes', *BMC Medical Genomics*, 7(1), p. 64.

Silvennoinen, K., de Lange, N., Zagaglia, S., Balestrini, S., Androsova, G., Wassenaar, M., Auce, P., Avbersek, A., Becker, F., Berghuis, B., Campbell, E., Coppola, A., Francis, B., Wolking, S., Cavalleri, G.L., Craig, J., Delanty, N., Johnson, M.R., Koeleman, B.P.C., Kunz, W.S., Lerche, H., Marson, A.G., O'Brien, T.J., Sander, J.W., Sills, G.J., Striano, P., Zara, F., van der Palen, J., Krause, R., Depondt, C., Sisodiya, S.M. and Epi, P.G.X.C. (2019) 'Comparative effectiveness of antiepileptic drugs in juvenile myoclonic epilepsy', *Epilepsia Open*, 4(3), pp. 420-430.

Simoes, S., Guo, J., Buitrago, L., Qureshi, Y.H., Feng, X., Kothiya, M., Cortes, E., Patel, V., Kannan, S., Kim, Y.H., Chang, K.T., Hussaini, S.A., Moreno, H., Di Paolo, G., Andersen, O.M. and Small, S.A. (2021) 'Alzheimer's vulnerable brain region relies on a distinct retromer core dedicated to endosomal recycling', *Cell Rep*, 37(13), p. 110182.

Simons Foundation (2024) *SFARI Gene*. Available at: https://gene.sfari.org/ (Accessed: 06 March 2024).

Simons Searchlight (2024). Available at: https://www.simonssearchlight.org (Accessed: 2024-03-24).

Singh, T., Poterba, T., Curtis, D., Akil, H., Al Eissa, M., Barchas, J.D., Bass, N., Bigdeli, T.B., Breen, G., Bromet, E.J., Buckley, P.F., Bunney, W.E., Bybjerg-Grauholm, J., Byerley, W.F., Chapman, S.B., Chen, W.J., Churchhouse, C., Craddock, N., Cusick, C.M., DeLisi, L., Dodge, S., Escamilla, M.A., Eskelinen, S., Fanous, A.H., Faraone, S.V., Fiorentino, A., Francioli, L., Gabriel, S.B., Gage, D., Gagliano Taliun, S.A., Ganna, A., Genovese, G., Glahn, D.C., Grove, J., Hall, M.H., Hämäläinen, E., Heyne, H.O., Holi, M., Hougaard, D.M., Howrigan, D.P., Huang, H., Hwu, H.G., Kahn, R.S., Kang, H.M., Karczewski, K.J., Kirov, G., Knowles, J.A., Lee, F.S., Lehrer, D.S., Lescai, F., Malaspina, D., Marder, S.R., McCarroll, S.A., McIntosh, A.M., Medeiros, H., Milani, L., Morley, C.P., Morris, D.W., Mortensen, P.B., Myers, R.M., Nordentoft, M., O'Brien, N.L., Olivares, A.M., Ongur, D., Ouwehand, W.H., Palmer, D.S., Paunio, T., Quested, D., Rapaport, M.H., Rees, E., Rollins, B., Satterstrom, F.K., Schatzberg, A., Scolnick, E., Scott, L.J., Sharp, S.I., Sklar, P., Smoller, J.W., Sobell, J.L., Solomonson, M., Stahl, E.A., Stevens, C.R., Suvisaari, J., Tiao, G., Watson, S.J., Watts, N.A., Blackwood, D.H., Børglum, A.D., Cohen, B.M., Corvin, A.P., Esko, T., Freimer, N.B., Glatt, S.J., Hultman, C.M., McQuillin, A., Palotie, A., Pato, C.N., Pato, M.T., Pulver, A.E., St Clair, D., et al. (2022) 'Rare

coding variants in ten genes confer substantial risk for schizophrenia', *Nature*, 604(7906), pp. 509-516.

Singhal, A. (2001) 'Modern Information Retrieval: A Brief Overview', *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4), pp. 35--43.

Singleton, M.V., Guthery, S.L., Voelkerding, K.V., Chen, K., Kennedy, B., Margraf, R.L., Durtschi, J., Eilbeck, K., Reese, M.G., Jorde, L.B., Huff, C.D. and Yandell, M. (2014) 'Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families', *Am J Hum Genet*, 94(4), pp. 599-610.

Sisodiya, S.M. and Mefford, H.C. (2011) 'Genetic contribution to common epilepsies', *Curr Opin Neurol*, 24(2), pp. 140-5.

Skotte, L., Fadista, J., Bybjerg-Grauholm, J., Appadurai, V., Hildebrand, M.S., Hansen, T.F., Banasik, K., Grove, J., Albiñana, C., Geller, F., Bjurström, C.F., Vilhjálmsson, B.J., Coleman, M., Damiano, J.A., Burgess, R., Scheffer, I.E., Pedersen, O.B.V., Erikstrup, C., Westergaard, D., Nielsen, K.R., Sørensen, E., Bruun, M.T., Liu, X., Hjalgrim, H., Pers, T.H., Mortensen, P.B., Mors, O., Nordentoft, M., Dreier, J.W., Børglum, A.D., Christensen, J., Hougaard, D.M., Buil, A., Hviid, A., Melbye, M., Ullum, H., Berkovic, S.F., Werge, T. and Feenstra, B. (2022) 'Genome-wide association study of febrile seizures implicates fever response and neuronal excitability genes', *Brain*, 145(2), pp. 555-568.

Slater, L.T., Russell, S., Makepeace, S., Carberry, A., Karwath, A., Williams, J.A., Fanning, H., Ball, S., Hoehndorf, R. and Gkoutos, G.V. (2022) 'Evaluating semantic similarity methods for comparison of text-derived phenotype profiles', *BMC Med Inform Decis Mak*, 22(1), p. 33.

Smedley, D., Jacobsen, J.O., Jager, M., Kohler, S., Holtgrewe, M., Schubach, M., Siragusa, E., Zemojtel, T., Buske, O.J., Washington, N.L., Bone, W.P., Haendel, M.A. and Robinson, P.N. (2015) 'Next-generation diagnostics and disease-gene discovery with the Exomiser', *Nat Protoc*, 10(12), pp. 2004-15.

Smedley, D., Oellrich, A., Kohler, S., Ruef, B., Sanger Mouse Genetics, P., Westerfield, M., Robinson, P., Lewis, S. and Mungall, C. (2013) 'PhenoDigm: analyzing curated annotations to associate animal models with human diseases', *Database (Oxford)*, 2013, p. bat025.

SNOMED International (2020) *SNOMED Clinical Terms*. Available at: https://www.snomed.org/snomed-ct/sct-worldwide (Accessed: 2020-12-10).

Son, J.H., Xie, G., Yuan, C., Ena, L., Li, Z., Goldstein, A., Huang, L., Wang, L., Shen, F., Liu, H., Mehl, K., Groopman, E.E., Marasa, M., Kiryluk, K., Gharavi, A.G., Chung, W.K., Hripcsak, G., Friedman, C., Weng, C. and Wang, K. (2018) 'Deep Phenotyping on Electronic Health Records Facilitates Genetic Diagnosis by Clinical Exomes', *Am J Hum Genet*, 103(1), pp. 58-73.

Song, J., Bergen, S.E., Di Florio, A., Karlsson, R., Charney, A., Ruderfer, D.M., Stahl, E.A., Chambert, K.D., Moran, J.L., Gordon-Smith, K., Forty, L., Green, E.K., Jones, I., Jones, L., Scolnick, E.M., Sklar, P., Smoller, J.W., Lichtenstein, P., Hultman, C., Craddock, N., Landén, M., Smoller, J.W., Perlis, R.H., Lee, P.H., Castro, V.M., Hoffnagle, A.G., Sklar, P., Stahl, E.A., Purcell, S.M., Ruderfer, D.M., Charney, A.W., Roussos, P., Michele Pato, C.P., Medeiros, H., Sobel, J., Craddock, N., Jones, I., Forty, L., Florio, A.D., Green, E., Jones, L., Gordon-Smith, K., Landen, M., Hultman, C., Jureus, A., Bergen, S., McCarroll, S., Moran, J., Smoller, J.W., Chambert, K. and Belliveau, R.A. (2016) 'Genome-wide association study identifies SESTD1

as a novel risk gene for lithium-responsive bipolar disorder', *Mol Psychiatry*, 21(9), pp. 1290-7.

Sørensen, T. (1948) 'A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons', *Kongelige Danske Videnskabernes Selskab*, 5(4), pp. 1-34.

Specchio, N., Wirrell, E.C., Scheffer, I.E., Nabbout, R., Riney, K., Samia, P., Guerreiro, M., Gwer, S., Zuberi, S.M., Wilmshurst, J.M., Yozawitz, E., Pressler, R., Hirsch, E., Wiebe, S., Cross, H.J., Perucca, E., Moshe, S.L., Tinuper, P. and Auvin, S. (2022) 'International League Against Epilepsy classification and definition of epilepsy syndromes with onset in childhood: Position paper by the ILAE Task Force on Nosology and Definitions', *Epilepsia*, 63(6), pp. 1398-1442.

Speed, D., O'Brien, T.J., Palotie, A., Shkura, K., Marson, A.G., Balding, D.J. and Johnson, M.R. (2014) 'Describing the genetic architecture of epilepsy through heritability analysis', *Brain*, 137(Pt 10), pp. 2680-9.

Stamberger, H., Crosiers, D., Balagura, G., Bonardi, C.M., Basu, A., Cantalupo, G., Chiesa, V., Christensen, J., Dalla Bernardina, B., Ellis, C.A., Furia, F., Gardiner, F., Giron, C., Guerrini, R., Klein, K.M., Korff, C., Krijtova, H., Leffler, M., Lerche, H., Lesca, G., Lewis-Smith, D., Marini, C., Marjanovic, D., Mazzola, L., McKeown Ruggiero, S., Mochel, F., Ramond, F., Reif, P.S., Richard-Mornas, A., Rosenow, F., Schropp, C., Thomas, R.H., Vignoli, A., Weber, Y., Palmer, E., Helbig, I., Scheffer, I.E., Striano, P., Moller, R.S., Gardella, E. and Weckhuysen, S. (2022) 'Natural History Study of STXBP1-Developmental and Epileptic Encephalopathy Into Adulthood', *Neurology*, 99(3), pp. e221-e233.

Stansauk, J., Fidell, A., Benke, T., Schaffer, M. and Demarest, S.T. (2023) 'Analysis of electrocardiograms in individuals with CDKL5 deficiency disorder', *Am J Med Genet A*, 191(1), pp. 108-111.

Stefanski, A., Calle-Lopez, Y., Leu, C., Perez-Palma, E., Pestana-Knight, E. and Lal, D. (2021) 'Clinical sequencing yield in epilepsy, autism spectrum disorder, and intellectual disability: A systematic review and meta-analysis', *Epilepsia*, 62(1), pp. 143-151.

Stefanski, A., Pérez-Palma, E., Mrdjen, M., McHugh, M., Leu, C. and Lal, D. (2022) 'Identification and quantification of oligogenic loss-of-function disorders', *Genet Med*, 24(3), pp. 729-735.

Steinlein, O.K., Mulley, J.C., Propping, P., Wallace, R.H., Phillips, H.A., Sutherland, G.R., Scheffer, I.E. and Berkovic, S.F. (1995) 'A missense mutation in the neuronal nicotinic acetylcholine receptor alpha 4 subunit is associated with autosomal dominant nocturnal frontal lobe epilepsy', *Nat Genet*, 11(2), pp. 201-3.

Steinlin, M. (2013) 'Cerebrovascular disorders in childhood', *Handb Clin Neurol*, 112, pp. 1053-64.

Steriade, C., Sperling, M.R., DiVentura, B., Lozano, M., Shellhaas, R.A., Kessler, S.K., Dlugos, D. and French, J. (2022) 'Proposal for an updated seizure classification framework in clinical trials', *Epilepsia*, 63(3), pp. 565-572.

Stosic, M., Levy, B. and Wapner, R. (2018) 'The Use of Chromosomal Microarray Analysis in Prenatal Diagnosis', *Obstet Gynecol Clin North Am*, 45(1), pp. 55-68.

Strehlow, V., Swinkels, M.E., Thomas, R.H., Rapps, N., Syrbe, S., Dorn, T. and Lemke, J.R. (2016) 'Generalized Epilepsy and Myoclonic Seizures in 22q11.2 Deletion Syndrome', *Mol Syndromol*, 7(4), pp. 239-246.

Stritt, S., Nurden, P., Turro, E., Greene, D., Jansen, S.B., Westbury, S.K., Petersen, R., Astle, W.J., Marlin, S., Bariana, T.K., Kostadima, M., Lentaigne, C., Maiwald, S., Papadia, S., Kelly, A.M., Stephens, J.C., Penkett, C.J., Ashford, S., Tuna, S., Austin, S., Bakchoul, T., Collins, P., Favier, R., Lambert, M.P., Mathias, M., Millar, C.M., Mapeta, R., Perry, D.J., Schulman, S., Simeoni, I., Thys, C., Consortium, B.-B., Gomez, K., Erber, W.N., Stirrups, K., Rendon, A., Bradley, J.R., van Geet, C., Raymond, F.L., Laffan, M.A., Nurden, A.T., Nieswandt, B., Richardson, S., Freson, K., Ouwehand, W.H. and Mumford, A.D. (2016) 'A gain-of-function variant in DIAPH1 causes dominant macrothrombocytopenia and hearing loss', *Blood*, 127(23), pp. 2903-14.

Subaran, R.L., Conte, J.M., Stewart, W.C. and Greenberg, D.A. (2015) 'Pathogenic EFHC1 mutations are tolerated in healthy individuals dependent on reported ancestry', *Epilepsia*, 56(2), pp. 188-94.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H., Konkel, M.K., Malhotra, A., Stütz, A.M., Shi, X., Casale, F.P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M.J.P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H.Y.K., Mu, X.J., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J.M., Kong, Y., Lameijer, E.W., McCarthy, S., Flicek, P., Gibbs, R.A., Marth, G., Mason, C.E., Menelaou, A., Muzny, D.M., Nelson, B.J., Noor, A., Parrish, N.F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E.E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A.A., Untergasser, A., Walker, J.A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M.A., McCarroll, S.A., Mills, R.E., Gerstein, M.B., Bashir, A., Stegle, O., Devine, S.E., Lee, C., Eichler, E.E. and Korbel, J.O. (2015) 'An integrated map of structural variation in 2,504 human genomes', *Nature*, 526(7571), pp. 75-81.

Sun, L., Jiang, Z., Acosta-Rodriguez, V.A., Berger, M., Du, X., Choi, J.H., Wang, J., Wang, K.W., Kilaru, G.K., Mohawk, J.A., Quan, J., Scott, L., Hildebrand, S., Li, X., Tang, M., Zhan, X., Murray, A.R., La Vine, D., Moresco, E.M.Y., Takahashi, J.S. and Beutler, B. (2017) 'HCFC2 is needed for IRF1- and IRF2-dependent Tlr3 transcription and for survival during viral infections', *J Exp Med*, 214(11), pp. 3263-3277.

Suzuki, T., Behnam, M., Ronasian, F., Salehi, M., Shiina, M., Koshimizu, E., Fujita, A., Sekiguchi, F., Miyatake, S., Mizuguchi, T., Nakashima, M., Ogata, K., Takeda, S., Matsumoto, N. and Miyake, N. (2018) 'A homozygous NOP14 variant is likely to cause recurrent pregnancy loss', *J Hum Genet*, 63(4), pp. 425-430.

Suzuki, T., Delgado-Escueta, A.V., Aguan, K., Alonso, M.E., Shi, J., Hara, Y., Nishida, M., Numata, T., Medina, M.T., Takeuchi, T., Morita, R., Bai, D., Ganesh, S., Sugimoto, Y., Inazawa, J., Bailey, J.N., Ochoa, A., Jara-Prado, A., Rasmussen, A., Ramos-Peek, J., Cordova, S., Rubio-Donnadieu, F., Inoue, Y., Osawa, M., Kaneko, S., Oguni, H., Mori, Y. and Yamakawa, K. (2004) 'Mutations in EFHC1 cause juvenile myoclonic epilepsy', *Nat Genet*, 36(8), pp. 842–9.

Suzuki, T., Koike, Y., Ashikawa, K., Otomo, N., Takahashi, A., Aoi, T., Kamatani, N., Nakamura, Y., Kubo, M., Kamatani, Y., Momozawa, Y., Terao, C. and Yamakawa, K. (2021)

'Genome-wide association study of epilepsy in a Japanese population identified an associated region at chromosome 12q24', *Epilepsia*, 62(6), pp. 1391-1400.

Sweeting, M.J., Sutton, A.J. and Lambert, P.C. (2004) 'What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data', *Stat Med*, 23(9), pp. 1351-75.

Symonds, J.D., Zuberi, S.M., Stewart, K., McLellan, A., O'Regan, M., MacLeod, S., Jollands, A., Joss, S., Kirkpatrick, M., Brunklaus, A., Pilz, D.T., Shetty, J., Dorris, L., Abu-Arafeh, I., Andrew, J., Brink, P., Callaghan, M., Cruden, J., Diver, L.A., Findlay, C., Gardiner, S., Grattan, R., Lang, B., MacDonnell, J., McKnight, J., Morrison, C.A., Nairn, L., Slean, M.M., Stephen, E., Webb, A., Vincent, A. and Wilson, M. (2019) 'Incidence and phenotypes of childhood-onset genetic epilepsies: a prospective population-based national cohort', *Brain*, 142(8), pp. 2303-2318.

Takata, A., Nakashima, M., Saitsu, H., Mizuguchi, T., Mitsuhashi, S., Takahashi, Y., Okamoto, N., Osaka, H., Nakamura, K., Tohyama, J., Haginoya, K., Takeshita, S., Kuki, I., Okanishi, T., Goto, T., Sasaki, M., Sakai, Y., Miyake, N., Miyatake, S., Tsuchida, N., Iwama, K., Minase, G., Sekiguchi, F., Fujita, A., Imagawa, E., Koshimizu, E., Uchiyama, Y., Hamanaka, K., Ohba, C., Itai, T., Aoi, H., Saida, K., Sakaguchi, T., Den, K., Takahashi, R., Ikeda, H., Yamaguchi, T., Tsukamoto, K., Yoshitomi, S., Oboshi, T., Imai, K., Kimizu, T., Kobayashi, Y., Kubota, M., Kashii, H., Baba, S., Iai, M., Kira, R., Hara, M., Ohta, M., Miyata, Y., Miyata, R., Takanashi, J.I., Matsui, J., Yokochi, K., Shimono, M., Amamoto, M., Takayama, R., Hirabayashi, S., Aiba, K., Matsumoto, H., Nabatame, S., Shiihara, T., Kato, M. and Matsumoto, N. (2019) 'Comprehensive analysis of coding variants highlights genetic complexity in developmental and epileptic encephalopathy', *Nat Commun*, 10(1), p. 2506.

Takumi, T. and Tamada, K. (2018) 'CNV biology in neurodevelopmental disorders', *Curr Opin Neurobiol*, 48, pp. 183-192.

Tan, A., Abecasis, G.R. and Kang, H.M. (2015) 'Unified representation of genetic variants', *Bioinformatics*, 31(13), pp. 2202-2204.

Tangarorang, J., Leonard, H., Epstein, A. and Downs, J. (2019) 'A framework for understanding quality of life domains in individuals with the CDKL5 deficiency disorder', *Am J Med Genet A*, 179(2), pp. 249-256.

The Gene Ontology Consortium, Aleksander, S.A., Balhoff, J., Carbon, S., Cherry, J.M., Drabkin, H.J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N.L., Hill, D.P., Lee, R., Mi, H., Moxon, S., Mungall, C.J., Muruganugan, A., Mushayahama, T., Sternberg, P.W., Thomas, P.D., Van Auken, K., Ramsey, J., Siegele, D.A., Chisholm, R.L., Fey, P., Aspromonte, M.C., Nugnes, M.V., Quaglia, F., Tosatto, S., Giglio, M., Nadendla, S., Antonazzo, G., Attrill, H., dos Santos, G., Marygold, S., Strelets, V., Tabone, C.J., Thurmond, J., Zhou, P., Ahmed, S.H., Asanitthong, P., Luna Buitrago, D., Erdol, M.N., Gage, M.C., Ali Kadhum, M., Li, K.Y.C., Long, M., Michalak, A., Pesala, A., Pritazahra, A., Saverimuttu, S.C.C., Su, R., Thurlow, K.E., Lovering, R.C., Logie, C., Oliferenko, S., Blake, J., Christie, K., Corbani, L., Dolan, M.E., Drabkin, H.J., Hill, D.P., Ni, L., Sitnikov, D., Smith, C., Cuzick, A., Seager, J., Cooper, L., Elser, J., Jaiswal, P., Gupta, P., Jaiswal, P., Naithani, S., Lera-Ramirez, M., Rutherford, K., Wood, V., De Pons, J.L., Dwinell, M.R., Hayman, G.T., Kaldunski, M.L., Kwitek, A.E., Laulederkind, S.J.F., Tutaj, M.A., Vedi, M., Wang, S.-J., D'Eustachio, P., Aimo, L., Axelsen, K., Bridge, A., Hyka-Nouspikel, N., Morgat, A., Aleksander, S.A., Cherry, J.M., Engel, S.R., Karra, K., Miyasato, S.R., Nash, R.S., Skrzypek, M.S., Weng, S., Wong, E.D., Bakker, E., et al. (2023) 'The Gene Ontology knowledgebase in 2023', *Genetics*, 224(1).

Thomas, R.H. and Berkovic, S.F. (2014) 'The hidden genetics of epilepsy-a clinically important new paradigm', *Nat Rev Neurol*, 10(5), pp. 283-92.

Tin, A., Fohner, A.E., Yang, Q., Brody, J.A., Davies, G., Yao, J., Liu, D., Caro, I., Lindbohm, J.V., Duggan, M.R., Meirelles, O., Harris, S.E., Gudmundsdottir, V., Taylor, A.M., Henry, A., Beiser, A.S., Shojaie, A., Coors, A., Fitzpatrick, A.L., Langenberg, C., Satizabal, C.L., Sitlani, C.M., Wheeler, E., Tucker-Drob, E.M., Bressler, J., Coresh, J., Bis, J.C., Candia, J., Jennings, L.L., Pietzner, M., Lathrop, M., Lopez, O.L., Redmond, P., Gerszten, R.E., Rich, S.S., Heckbert, S.R., Austin, T.R., Hughes, T.M., Tanaka, T., Emilsson, V., Vasan, R.S., Guo, X., Zhu, Y., Tzourio, C., Rotter, J.I., Walker, K.A., Ferrucci, L., Kivimäki, M., Breteler, M.M.B., Cox, S.R., Debette, S., Mosley, T.H., Gudnason, V.G., Launer, L.J., Psaty, B.M., Seshadri, S. and Fornage, M. (2023) 'Identification of circulating proteins associated with general cognitive function among middle-aged and older adults', *Commun Biol*, 6(1), p. 1117.

Tobochnik, S., Fahlstrom, R., Shain, C., Winawer, M.R. and Investigators, E. (2017) 'Familial aggregation of focal seizure semiology in the Epilepsy Phenome/Genome Project', *Neurology*, 89(1), pp. 22-28.

Trakadis, Y.J., Buote, C., Therriault, J.F., Jacques, P.E., Larochelle, H. and Levesque, S. (2014) 'PhenoVar: a phenotype-driven approach in clinical genomics for the diagnosis of polymalformative syndromes', *BMC Med Genomics*, 7(1), p. 22.

Trinka, E., Cock, H., Hesdorffer, D., Rossetti, A.O., Scheffer, I.E., Shinnar, S., Shorvon, S. and Lowenstein, D.H. (2015) 'A definition and classification of status epilepticus--Report of the ILAE Task Force on Classification of Status Epilepticus', *Epilepsia*, 56(10), pp. 1515-23.

Truty, R., Patil, N., Sankar, R., Sullivan, J., Millichap, J., Carvill, G., Entezam, A., Esplin, E.D., Fuller, A., Hogue, M., Johnson, B., Khouzam, A., Kobayashi, Y., Lewis, R., Nykamp, K., Riethmaier, D., Westbrook, J., Zeman, M., Nussbaum, R.L. and Aradhya, S. (2019) 'Possible precision medicine implications from genetic testing using combined detection of sequence and intragenic copy number variants in a large cohort with childhood epilepsy', *Epilepsia Open*, 4(3), pp. 397-408.

Tudorache, T., Nyulas, C., Noy, N.F. and Musen, M.A. (2013) 'WebProtege: A Collaborative Ontology Editor and Knowledge Acquisition Tool for the Web', *Semant Web*, 4(1), pp. 89-99.

Turro, E., Astle, W.J., Megy, K., Graf, S., Greene, D., Shamardina, O., Allen, H.L., Sanchis-Juan, A., Frontini, M., Thys, C., Stephens, J., Mapeta, R., Burren, O.S., Downes, K., Haimel, M., Tuna, S., Deevi, S.V.V., Aitman, T.J., Bennett, D.L., Calleja, P., Carss, K., Caulfield, M.J., Chinnery, P.F., Dixon, P.H., Gale, D.P., James, R., Koziell, A., Laffan, M.A., Levine, A.P., Maher, E.R., Markus, H.S., Morales, J., Morrell, N.W., Mumford, A.D., Ormondroyd, E., Rankin, S., Rendon, A., Richardson, S., Roberts, I., Roy, N.B.A., Saleem, M.A., Smith, K.G.C., Stark, H., Tan, R.Y.Y., Themistocleous, A.C., Thrasher, A.J., Watkins, H., Webster, A.R., Wilkins, M.R., Williamson, C., Whitworth, J., Humphray, S., Bentley, D.R., Nihr BioResource for the, G.P., Kingston, N., Walker, N., Bradley, J.R., Ashford, S., Penkett, C.J., Freson, K., Stirrups, K.E., Raymond, F.L. and Ouwehand, W.H. (2020) 'Whole-genome sequencing of patients with rare diseases in a national health system', *Nature*, 583(7814), pp. 96-102.

Uddin, M., Unda, B.K., Kwan, V., Holzapfel, N.T., White, S.H., Chalil, L., Woodbury-Smith, M., Ho, K.S., Harward, E., Murtaza, N., Dave, B., Pellecchia, G., D'Abate, L., Nalpathamkalam, T., Lamoureux, S., Wei, J., Speevak, M., Stavropoulos, J., Hope, K.J., Doble, B.W., Nielsen, J., Wassman, E.R., Scherer, S.W. and Singh, K.K. (2018) 'OTUD7A Regulates

Neurodevelopmental Phenotypes in the 15q13.3 Microdeletion Syndrome', *Am J Hum Genet*, 102(2), pp. 278-295.

Uemura, N., Okumura, A., Negoro, T. and Watanabe, K. (2002) 'Clinical features of benign convulsions with mild gastroenteritis', *Brain Dev*, 24(8), pp. 745-9.

UK Biobank (2007). Available at: www.ukbiobank.ac.uk (Accessed: 2022-01-06).

Umeki, I., Niihori, T., Abe, T., Kanno, S.I., Okamoto, N., Mizuno, S., Kurosawa, K., Nagasaki, K., Yoshida, M., Ohashi, H., Inoue, S.I., Matsubara, Y., Fujiwara, I., Kure, S. and Aoki, Y. (2019) 'Delineation of LZTR1 mutation-positive patients with Noonan syndrome and identification of LZTR1 binding to RAF1-PPP1CB complexes', *Hum Genet*, 138(1), pp. 21-35.

UniProt (2023) 'UniProtKB O76039 · CDKL5_HUMAN'. 2023-11-22. Available at: https://www.uniprot.org/uniprotkb/O76039/.

Vadlamudi, L., Andermann, E., Lombroso, C.T., Schachter, S.C., Milne, R.L., Hopper, J.L., Andermann, F. and Berkovic, S.F. (2004) 'Epilepsy in twins: insights from unique historical data of William Lennox', *Neurology*, 62(7), pp. 1127–33.

Vadlamudi, L., Milne, R.L., Lawrence, K., Heron, S.E., Eckhaus, J., Keay, D., Connellan, M., Torn-Broers, Y., Howell, R.A., Mulley, J.C., Scheffer, I.E., Dibbens, L.M., Hopper, J.L. and Berkovic, S.F. (2014) 'Genetics of epilepsy: The testimony of twins in the molecular era', *Neurology*, 83(12), pp. 1042-8.

Van Bergen, N.J., Massey, S., Quigley, A., Rollo, B., Harris, A.R., Kapsa, R.M.I. and Christodoulou, J. (2022) 'CDKL5 deficiency disorder: molecular insights and mechanisms of pathogenicity to fast-track therapeutic development', *Biochem Soc Trans*, 50(4), pp. 1207-1224.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler, D., Gabriel, S. and DePristo, M.A. (2013) 'From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline', *Curr Protoc Bioinformatics*, 43(1110), pp. 11.10.1-11.10.33.

van Diessen, E., van Amerongen, R.A., Zijlmans, M. and Otte, W.M. (2024) 'Potential merits and flaws of large language models in epilepsy care: A critical review', *Epilepsia*.

van Karnebeek, C.D., Tiebout, S.A., Niermeijer, J., Poll-The, B.T., Ghani, A., Coughlin, C.R., 2nd, Van Hove, J.L., Richter, J.W., Christen, H.J., Gallagher, R., Hartmann, H. and Stockler-Ipsiroglu, S. (2016) 'Pyridoxine-Dependent Epilepsy: An Expanding Clinical Spectrum', *Pediatr Neurol*, 59, pp. 6-12.

Varley, J., Taylor, J. and Irani, S.R. (2018) 'Autoantibody-mediated diseases of the CNS: Structure, dysfunction and therapy', *Neuropharmacology*, 132, pp. 71-82.

Verma, A., Damrauer, S.M., Naseer, N., Weaver, J., Kripke, C.M., Guare, L., Sirugo, G., Kember, R.L., Drivas, T.G., Dudek, S.M., Bradford, Y., Lucas, A., Judy, R., Verma, S.S., Meagher, E., Nathanson, K.L., Feldman, M., Ritchie, M.D., Rader, D.J. and For The Penn Medicine, B. (2022) 'The Penn Medicine BioBank: Towards a Genomics-Enabled Learning Healthcare System to Accelerate Precision Medicine in a Diverse Population', *J Pers Med*, 12(12).

Vervoort, L. and Vermeesch, J.R. (2022) 'The 22q11.2 Low Copy Repeats', *Genes (Basel)*, 13(11).

Villar, J., Papageorghiou, A.T., Pang, R., Ohuma, E.O., Cheikh Ismail, L., Barros, F.C., Lambert, A., Carvalho, M., Jaffer, Y.A., Bertino, E., Gravett, M.G., Altman, D.G., Purwar, M., Frederick, I.O., Noble, J.A., Victora, C.G., Bhutta, Z.A., Kennedy, S.H., International, F. and Newborn Growth Consortium for the 21st, C. (2014) 'The likeness of fetal growth and newborn size across non-isolated populations in the INTERGROWTH-21st Project: the Fetal Growth Longitudinal Study and Newborn Cross-Sectional Study', *Lancet Diabetes Endocrinol*, 2(10), pp. 781-92.

Vlaskamp, D.R.M., Bassett, A.S., Sullivan, J.E., Robblee, J., Sadleir, L.G., Scheffer, I.E. and Andrade, D.M. (2019) 'Schizophrenia is a later-onset feature of PCDH19 Girls Clustering Epilepsy', *Epilepsia*, 60(3), pp. 429-440.

Vreeman, D.J., McDonald, C.J. and Huff, S.M. (2010) 'LOINC(R) - A Universal Catalog of Individual Clinical Observations and Uniform Representation of Enumerated Collections', *Int J Funct Inform Personal Med*, 3(4), pp. 273-291.

Waisbren, S.E., Noel, K., Fahrbach, K., Cella, C., Frame, D., Dorenbaum, A. and Levy, H. (2007) 'Phenylalanine blood levels and clinical outcomes in phenylketonuria: a systematic literature review and meta-analysis', *Mol Genet Metab*, 92(1-2), pp. 63-70.

Wakefield, J. (2009) 'Bayes factors for genome-wide association studies: comparison with P-values', *Genet Epidemiol*, 33(1), pp. 79-86.

Wallace, R.H., Wang, D.W., Singh, R., Scheffer, I.E., George, A.L., Jr., Phillips, H.A., Saar, K., Reis, A., Johnson, E.W., Sutherland, G.R., Berkovic, S.F. and Mulley, J.C. (1998) 'Febrile seizures and generalized epilepsy associated with a mutation in the Na+-channel beta1 subunit gene SCN1B', *Nat Genet*, 19(4), pp. 366–70.

Wang, B., Cao, C., Liu, X., He, X., Zhuang, H., Wang, D. and Chen, B. (2020) 'BRCA1-associated protein inhibits glioma cell proliferation and migration and glioma stem cell self-renewal via the TGF-β/PI3K/AKT/mTOR signalling pathway', *Cell Oncol (Dordr)*, 43(2), pp. 223-235.

Wang, H., Qi, W., Zou, C., Xie, Z., Zhang, M., Naito, M.G., Mifflin, L., Liu, Z., Najafov, A., Pan, H., Shan, B., Li, Y., Zhu, Z.J. and Yuan, J. (2021) 'NEK1-mediated retromer trafficking promotes blood-brain barrier integrity by regulating glucose metabolism and RIPK1 activation', *Nat Commun*, 12(1), p. 4826.

Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S. and Chen, C.F. (2007a) 'A new method to measure the semantic similarity of GO terms', *Bioinformatics*, 23(10), pp. 1274-81.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F., Hakonarson, H. and Bucan, M. (2007b) 'PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data', *Genome Res*, 17(11), pp. 1665-74.

Wang, Y. and Tang, B.L. (2006) 'SNAREs in neurons--beyond synaptic vesicle exocytosis (Review)', *Mol Membr Biol*, 23(5), pp. 377-84.

Warda, A.S., Freytag, B., Haag, S., Sloan, K.E., Görlich, D. and Bohnsack, M.T. (2016) 'Effects of the Bowen-Conradi syndrome mutation in EMG1 on its nuclear import, stability and nucleolar recruitment', *Hum Mol Genet*, 25(24), pp. 5353-5364.

Washington, N.L., Haendel, M.A., Mungall, C.J., Ashburner, M., Westerfield, M. and Lewis, S.E. (2009) 'Linking human diseases to animal models using ontology-based phenotype annotation', *PLoS Biol*, 7(11), p. e1000247.

Weaving, L.S., Christodoulou, J., Williamson, S.L., Friend, K.L., McKenzie, O.L., Archer, H., Evans, J., Clarke, A., Pelka, G.J., Tam, P.P., Watson, C., Lahooti, H., Ellaway, C.J., Bennetts, B., Leonard, H. and Gecz, J. (2004) 'Mutations of CDKL5 cause a severe neurodevelopmental disorder with infantile spasms and mental retardation', *Am J Hum Genet*, 75(6), pp. 1079-93.

Westbury, S.K., Turro, E., Greene, D., Lentaigne, C., Kelly, A.M., Bariana, T.K., Simeoni, I., Pillois, X., Attwood, A., Austin, S., Jansen, S.B., Bakchoul, T., Crisp-Hihn, A., Erber, W.N., Favier, R., Foad, N., Gattens, M., Jolley, J.D., Liesner, R., Meacham, S., Millar, C.M., Nurden, A.T., Peerlinck, K., Perry, D.J., Poudel, P., Schulman, S., Schulze, H., Stephens, J.C., Furie, B., Robinson, P.N., van Geet, C., Rendon, A., Gomez, K., Laffan, M.A., Lambert, M.P., Nurden, P., Ouwehand, W.H., Richardson, S., Mumford, A.D., Freson, K. and Consortium, B.-B. (2015) 'Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders', *Genome Med*, 7(1), p. 36.

Westfall, P.H. and Wolfinger, R.D. (1997) 'Multiple Tests with Discrete Distributions', *Am Stat*, 51(1), pp. 3-8.

Whiffin, N., Ware, J.S. and O'Donnell-Luria, A. (2019) 'Improving the Understanding of Genetic Variants in Rare Disease With Large-scale Reference Populations', *JAMA*, 322(13), pp. 1305-1306.

White, J., Mazzeu, J.F., Hoischen, A., Jhangiani, S.N., Gambin, T., Alcino, M.C., Penney, S., Saraiva, J.M., Hove, H., Skovby, F., Kayserili, H., Estrella, E., Vulto-van Silfhout, A.T., Steehouwer, M., Muzny, D.M., Sutton, V.R., Gibbs, R.A., Lupski, J.R., Brunner, H.G., van Bon, B.W. and Carvalho, C.M. (2015) 'DVL1 frameshift mutations clustering in the penultimate exon cause autosomal-dominant Robinow syndrome', *Am J Hum Genet*, 96(4), pp. 612-22.

Whitney, R., Nair, A., McCready, E., Keller, A.E., Adil, I.S., Aziz, A.S., Borys, O., Siu, K., Shah, C., Meaney, B.F., Jones, K. and RamachandranNair, R. (2021) 'The spectrum of epilepsy in children with 15q13.3 microdeletion syndrome', *Seizure*, 92, pp. 221-229.

Wienke, J., Visser, L.L., Kholosy, W.M., Keller, K.M., Barisa, M., Poon, E., Munnings-Tomes, S., Himsworth, C., Calton, E., Rodriguez, A., Bernardi, R., van den Ham, F., van Hooff, S.R., Matser, Y.A.H., Tas, M.L., Langenberg, K.P.S., Lijnzaad, P., Borst, A.L., Zappa, E., Bergsma, F.J., Strijker, J.G.M., Verhoeven, B.M., Mei, S., Kramdi, A., Restuadi, R., Sanchez-Bernabeu, A., Cornel, A.M., Holstege, F.C.P., Gray, J.C., Tytgat, G.A.M., Scheijde-Vermeulen, M.A., Wijnen, M., Dierselhuis, M.P., Straathof, K., Behjati, S., Wu, W., Heck, A.J.R., Koster, J., Nierkens, S., Janoueix-Lerosey, I., de Krijger, R.R., Baryawno, N., Chesler, L., Anderson, J., Caron, H.N., Margaritis, T., van Noesel, M.M. and Molenaar, J.J. (2024) 'Integrative analysis of neuroblastoma by single-cell RNA sequencing identifies the NECTIN2-TIGIT axis as a target for immunotherapy', *Cancer Cell*, 42(2), pp. 283-300.e8.

Wigglesworth, S., Neligan, A., Dickson, J.M., Pullen, A., Yelland, E., Anjuman, T. and Reuber, M. (2023) 'The incidence and prevalence of epilepsy in the United Kingdom 2013–2018: A retrospective cohort study of UK primary care data', *Seizure*, 105, pp. 37-42.

Wiley, L.K., Shortt, J.A., Roberts, E.R., Lowery, J., Kudron, E., Lin, M., Mayer, D., Wilson, M., Brunetti, T.M., Chavan, S., Phang, T.L., Pozdeyev, N., Lesny, J., Wicks, S.J., Moore, E.T., Morgenstern, J.L., Roff, A.N., Shalowitz, E.L., Stewart, A., Williams, C., Edelmann, M.N., Hull, M., Patton, J.T., Axell, L., Ku, L., Lee, Y.M., Jirikowic, J., Tanaka, A., Todd, E., White, S., Peterson, B., Hearst, E., Zane, R., Greene, C.S., Mathias, R., Coors, M., Taylor, M., Ghosh, D., Kahn, M.G., Brooks, I.M., Aquilante, C.L., Kao, D., Rafaels, N., Crooks, K.R., Hess, S., Barnes, K.C. and Gignoux, C.R. (2024) 'Building a vertically integrated genomic learning health system: The biobank at the Colorado Center for Personalized Medicine', *Am J Hum Genet*, 111(1), pp. 11-23.

Wilson, K., Newbury, D.F. and Kini, U. (2023) 'Analysis of exome data in a UK cohort of 603 patients with syndromic orofacial clefting identifies causal molecular pathways', *Human Molecular Genetics*, 32(11), pp. 1932-1942.

Winawer, M.R. (2006) 'Phenotype definition in epilepsy', *Epilepsy Behav*, 8(3), pp. 462-76.

Winawer, M.R., Marini, C., Grinton, B.E., Rabinowitz, D., Berkovic, S.F., Scheffer, I.E. and Ottman, R. (2005) 'Familial clustering of seizure types within the idiopathic generalized epilepsies', *Neurology*, 65(4), pp. 523-8.

Winawer, M.R., Martinelli Boneschi, F., Barker-Cummings, C., Lee, J.H., Liu, J., Mekios, C., Gilliam, T.C., Pedley, T.A., Hauser, W.A. and Ottman, R. (2002) 'Four new families with autosomal dominant partial epilepsy with auditory features: clinical description and linkage to chromosome 10q24', *Epilepsia*, 43(1), pp. 60-7.

Winawer, M.R., Ottman, R., Hauser, W.A. and Pedley, T.A. (2000) 'Autosomal dominant partial epilepsy with auditory features: defining the phenotype', *Neurology*, 54(11), pp. 2173-6.

Winterhalder, R., McCabe, J., Young, C., Lamb, K., Sawhney, I., Jory, C., O'Dwyer, M. and Shankar, R. (2022) 'Bone health, intellectual disability and epilepsy: An observational community-based study', *Acta Neurol Scand*, 145(6), pp. 753-761.

Wirrell, E., Tinuper, P., Perucca, E. and Moshe, S.L. (2022a) 'Introduction to the epilepsy syndrome papers', *Epilepsia*, 63(6), pp. 1330-1332.

Wirrell, E.C., Nabbout, R., Scheffer, I.E., Alsaadi, T., Bogacz, A., French, J.A., Hirsch, E., Jain, S., Kaneko, S., Riney, K., Samia, P., Snead, O.C., Somerville, E., Specchio, N., Trinka, E., Zuberi, S.M., Balestrini, S., Wiebe, S., Cross, J.H., Perucca, E., Moshe, S.L. and Tinuper, P. (2022b) 'Methodology for classification and definition of epilepsy syndromes with list of syndromes: Report of the ILAE Task Force on Nosology and Definitions', *Epilepsia*, 63(6), pp. 1333-1348.

Wojcik, M.H., Okada, K., Prabhu, S.P., Nowakowski, D.W., Ramsey, K., Balak, C., Rangasamy, S., Brownstein, C.A., Schmitz-Abe, K., Cohen, J.S., Fatemi, A., Shi, J., Grant, E.P., Narayanan, V., Ho, H.H. and Agrawal, P.B. (2018) 'De novo variant in KIF26B is associated with pontocerebellar hypoplasia with infantile spinal muscular atrophy', *Am J Med Genet A*, 176(12), pp. 2623-2629.

Wolf, P. (2003) 'Of cabbages and kings: some considerations on classifications, diagnostic schemes, semiology, and concepts', *Epilepsia*, 44(1), pp. 1-4; discussion 4-13.

World Health Organization (1992) *The ICD-10 classification of mental and behavioural disorders : clinical descriptions and diagnostic guidelines*. Geneva: World Health Organization,.

Wu, H., Su, Z., Mao, F., Olman, V. and Xu, Y. (2005) 'Prediction of functional modules based on comparative genome analysis and Gene Ontology application', *Nucleic Acids Res*, 33(9), pp. 2822-37.

Wu, X., Zhu, L., Guo, J., Zhang, D.Y. and Lin, K. (2006) 'Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations', *Nucleic Acids Res*, 34(7), pp. 2137-50.

Wu, Z. and Palmer, M. (1994) *32nd Annual Meeting of the Association for Computational Linguistics*. Las Cruces, New Mexico, USA.

Xia, C., Ma, W., Stafford, L.J., Marcus, S., Xiong, W.C. and Liu, M. (2001) 'Regulation of the p21-activated kinase (PAK) by a human Gbeta -like WD-repeat protein, hPIP1', *Proc Natl Acad Sci U S A*, 98(11), pp. 6174-9.

Xian, J., Parthasarathy, S., Ruggiero, S.M., Balagura, G., Fitch, E., Helbig, K., Gan, J., Ganesan, S., Kaufman, M.C., Ellis, C.A., Lewis-Smith, D., Galer, P., Cunningham, K., O'Brien, M., Cosico, M., Baker, K., Darling, A., Veiga de Goes, F., El Achkar, C.M., Doering, J.H., Furia, F., Garcia-Cazorla, A., Gardella, E., Geertjens, L., Klein, C., Kolesnik-Taylor, A., Lammertse, H., Lee, J., Mackie, A., Misra-Isrie, M., Olson, H., Sexton, E., Sheidley, B., Smith, L., Sotero, L., Stamberger, H., Syrbe, S., Thalwitzer, K.M., van Berkel, A., van Haelst, M., Yuskaitis, C., Weckhuysen, S., Prosser, B., Son Rigby, C., Demarest, S., Pierce, S., Zhang, Y., Moller, R.S., Bruining, H., Poduri, A., Zara, F., Verhage, M., Striano, P. and Helbig, I. (2022) 'Assessing the landscape of STXBP1-related disorders in 534 individuals', *Brain*, 145(5), pp. 1668-1683.

Xian, J., Thalwitzer, K.M., McKee, J., Sullivan, K.R., Brimble, E., Fitch, E., Toib, J., Kaufman, M.C., deCampo, D., Cunningham, K., Pierce, S.R., Goss, J., Rigby, C.S., Syrbe, S., Boland, M., Prosser, B., Fitter, N., Ruggiero, S.M. and Helbig, I. (2023) 'Delineating clinical and developmental outcomes in STXBP1-related disorders', *Brain*, 146(12), pp. 5182-5197.

Xie, K., Gallagher, R.S., Shinohara, R.T., Xie, S.X., Hill, C.E., Conrad, E.C., Davis, K.A., Roth, D., Litt, B. and Ellis, C.A. (2023) 'Long-term epilepsy outcome dynamics revealed by natural language processing of clinic notes', *Epilepsia*, 64(7), pp. 1900-1909.

Xiong, L., Labuda, M., Li, D.S., Hudson, T.J., Desbiens, R., Patry, G., Verret, S., Langevin, P., Mercho, S., Seni, M.H., Scheffer, I., Dubeau, F., Berkovic, S.F., Andermann, F., Andermann, E. and Pandolfo, M. (1999) 'Mapping of a gene determining familial partial epilepsy with variable foci to chromosome 22q11-q12', *Am J Hum Genet*, 65(6), pp. 1698–710.

Yagi, H., Furutani, Y., Hamada, H., Sasaki, T., Asakawa, S., Minoshima, S., Ichida, F., Joo, K., Kimura, M., Imamura, S., Kamatani, N., Momma, K., Takao, A., Nakazawa, M., Shimizu, N. and Matsuoka, R. (2003) 'Role of TBX1 in human del22q11.2 syndrome', *Lancet*, 362(9393), pp. 1366-73.

Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. (2011) 'GCTA: a tool for genome-wide complex trait analysis', *Am J Hum Genet*, 88(1), pp. 76-82.

Yang, R., Wang, M., Zhang, G., Li, Y., Wang, L. and Cui, H. (2021) 'POU2F2 regulates glycolytic reprogramming and glioblastoma progression via PDPK1-dependent activation of PI3K/AKT/mTOR pathway', *Cell Death Dis*, 12(5), p. 433.

Yang, T.C., Lu, M.H., Wang, W.J. and Chen, J.Y. (2023) 'CEBPB/POU2F2 modulates endothelin 1 expression in prehypertensive SHR vascular smooth muscle cells', *J Mol Endocrinol*, 71(1).

Yang, X. and Cheyette, B.N. (2013) 'SEC14 and spectrin domains 1 (Sestd1) and Dapper antagonist of catenin 1 (Dact1) scaffold proteins cooperatively regulate the Van Gogh-like 2 (Vangl2) four-pass transmembrane protein and planar cell polarity (PCP) pathway during embryonic development in mice', *J Biol Chem*, 288(28), pp. 20111-20.

Yasuda, Y., Hashimoto, R., Fukai, R., Okamoto, N., Hiraki, Y., Yamamori, H., Fujimoto, M., Ohi, K., Taniike, M., Mohri, I., Nakashima, M., Tsurusaki, Y., Saitsu, H., Matsumoto, N., Miyake, N. and Takeda, M. (2014) 'Duplication of the NPHP1 gene in patients with autism spectrum disorder and normal intellectual ability: a case series', *Ann Gen Psychiatry*, 13, p. 22.

Yu, H., Jansen, R., Stolovitzky, G. and Gerstein, M. (2007) 'Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications', *Bioinformatics*, 23(16), pp. 2163-73.

Yu, W., Qiu, Z., Gao, N., Wang, L., Cui, H., Qian, Y., Jiang, L., Luo, J., Yi, Z., Lu, H., Li, D. and Liu, M. (2011) 'PAK1IP1, a ribosomal stress-induced nucleolar protein, regulates cell proliferation via the p53-MDM2 loop', *Nucleic Acids Res*, 39(6), pp. 2234-48.

Zacher, P., Mayer, T., Brandhoff, F., Bartolomaeus, T., Le Duc, D., Finzel, M., Heinze, A., Horn, S., Klockner, C., Korber, G., Hentschel, J., Kalita, M., Krey, I., Nastainczyk-Wulf, M., Platzer, K., Rebstock, J., Popp, B., Stiller, M., Teichmann, A.C., Jamra, R.A. and Lemke, J.R. (2021) 'The genetic landscape of intellectual disability and epilepsy in adults and the elderly: a systematic genetic work-up of 150 individuals', *Genet Med*, 23(8), pp. 1492-1497.

Zemojtel, T., Kohler, S., Mackenroth, L., Jager, M., Hecht, J., Krawitz, P., Graul-Neumann, L., Doelken, S., Ehmke, N., Spielmann, M., Oien, N.C., Schweiger, M.R., Kruger, U., Frommer, G., Fischer, B., Kornak, U., Flottmann, R., Ardeshirdavani, A., Moreau, Y., Lewis, S.E., Haendel, M., Smedley, D., Horn, D., Mundlos, S. and Robinson, P.N. (2014) 'Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome', *Sci Transl Med*, 6(252), p. 252ra123.

Zhang, P., Zhang, J., Sheng, H., Russo, J.J., Osborne, B. and Buetow, K. (2006) 'Gene functional similarity search tool (GFSST)', *BMC Bioinformatics*, 7(1), p. 135.

Zhang, Z., Ye, M., Li, Q., You, Y., Yu, H., Ma, Y., Mei, L., Sun, X., Wang, L., Yue, W., Li, R., Li, J. and Zhang, D. (2019) 'The Schizophrenia Susceptibility Gene OPCML Regulates Spine Maturation and Cognitive Behaviors through Eph-Cofilin Signaling', *Cell Reports*, 29(1), pp. 49-61.e7.

Zhao, C. and Wang, Z. (2018) 'GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms', *Sci Rep*, 8(1), p. 15107.

Zhuang, J., Xie, M., Yao, J., Fu, W., Zeng, S., Jiang, Y., Wang, Y., Xie, Y., Wang, G. and Chen, C. (2023) 'A de novo PAK1 likely pathogenic variant and a de novo terminal 1q microdeletion in a Chinese girl with global developmental delay, severe intellectual disability, and seizures', *BMC Med Genomics*, 16(1), p. 3.

Zuberi, S.M., Wirrell, E., Yozawitz, E., Wilmshurst, J.M., Specchio, N., Riney, K., Pressler, R., Auvin, S., Samia, P., Hirsch, E., Galicchio, S., Triki, C., Snead, O.C., Wiebe, S., Cross, J.H., Tinuper, P., Scheffer, I.E., Perucca, E., Moshe, S.L. and Nabbout, R. (2022) 'ILAE classification and definition of epilepsy syndromes with onset in neonates and infants: Position statement by the ILAE Task Force on Nosology and Definitions', *Epilepsia*, 63(6), pp. 1349-1397.

Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R. and Lander, E.S. (2014) 'Searching for missing heritability: designing rare variant association studies', *Proc Natl Acad Sci U S A*, 111(4), pp. E455-64.

Appendix A. Professional accomplishments within the Wellcome Trust Clinical PhD Fellowship and preceding Guarantors of Brain Entry Fellowship

8.1 Relevant publications

8.1.1 Primary or senior author peer-reviewed publications

2020

Lewis-Smith, D., Ellis, C.A., Helbig, I. and Thomas, R.H. (2020a) 'Early-onset genetic epilepsies reaching adult clinics', *Brain*, 143(3), p. e19.

Lewis-Smith, D. and Thomas, R.H. (2020) 'The prevalence of genetically diagnosable epilepsies in young adulthood: How many should we be looking for?', *Epilepsia*, 61(9), pp. 2053-2054.

2021

Lewis-Smith, D., Galer, P.D., Balagura, G., Kearney, H., Ganesan, S., Cosico, M., O'Brien, M., Vaidiswaran, P., Krause, R., Ellis, C.A., Thomas, R.H., Robinson, P.N. and Helbig, I. (2021a) 'Modeling seizures in the Human Phenotype Ontology according to contemporary ILAE concepts makes big phenotypic data tractable', *Epilepsia*, 62(6), pp. 1293-1305.

Lewis-Smith, D.,† Ganesan, S.,† Galer, P.D., Helbig, K.L., McKeown, S.E., O'Brien, M., Khankhanian, P., Kaufman, M.C., Gonzalez, A.K., Felmeister, A.S., Krause, R., Ellis, C.A. and Helbig, I. (2021b) 'Phenotypic homogeneity in childhood epilepsies evolves in gene-specific patterns across 3251 patient-years of clinical data', *Eur J Hum Genet*, 29(11), pp. 1690-1700.

(†Equal contributions)

2022

Lewis-Smith, D., Parthasarathy, S., Xian, J., Kaufman, M.C., Ganesan, S., Galer, P.D., Thomas, R.H. and Helbig, I. (2022) 'Computational analysis of neurodevelopmental phenotypes: Harmonization empowers clinical discovery', *Hum Mutat*, 43(11), pp. 1642-1658.

Lewis-Smith, D., Jaiser, S.R. and Thomas, R.H. (2022) 'Autoimmune musicogenic bilateral temporal lobe epilepsy', *Epileptic Disord*, 24(5), pp. 961-964.

Fowkes, R., Elwan, M., Akay, E., Mitchell, C.J., Thomas, R.H. and Lewis-Smith, D. (2022) 'A review of the clinical spectrum of BRAT1 disorders and case of developmental and epileptic encephalopathy surviving into adulthood', *Epilepsy Behav Rep*, 19, p. 100549.

2023

Montanucci,† L., Lewis-Smith, D.,† Collins, R.L.,† Niestroj, L.M.,† Parthasarathy, S., Xian, J., Ganesan, S., Macnee, M., Brunger, T., Thomas, R.H., Talkowski, M., Epi, C., Helbig, I., Leu, C. and Lal, D. (2023) 'Genome-wide identification and phenotypic characterization of seizure-associated copy number variations in 741,075 individuals', *Nat Commun*, 14(1), p. 4392.

8.1.2 Co-authorship of peer-reviewed publications

2020

Galer, P.D., Ganesan, S., Lewis-Smith, D., McKeown, S.E., Pendziwiat, M., Helbig, K.L., Ellis, C.A., Rademacher, A., Smith, L., Poduri, A., Seiffert, S., von Spiczak, S., Muhle, H., van Baalen, A., Group, N.S., Investigators, E., Consortium, E.-R.E.S., Genomics, R., Innovation, N., Thomas, R.H., Krause, R., Weber, Y. and Helbig, I. (2020) 'Semantic Similarity Analysis Reveals Robust Gene-Disease Relationships in Developmental and Epileptic Encephalopathies', *Am J Hum Genet*, 107(4), pp. 683-697.

2021

Kohler, S., Gargano, M., Matentzoglu, N., Carmody, L.C., Lewis-Smith, D., Vasilevsky, N.A., Danis, D., Balagura, G., Baynam, G., Brower, A.M., Callahan, T.J., Chute, C.G., Est, J.L., Galer, P.D., Ganesan, S., Griese, M., Haimel, M., Pazmandi, J., Hanauer, M., Harris, N.L., Hartnett, M.J., Hastreiter, M., Hauck, F., He, Y., Jeske, T., Kearney, H., Kindle, G., Klein, C., Knoflach, K., Krause, R., Lagorce, D., McMurry, J.A., Miller, J.A., Munoz-Torres, M.C., Peters, R.L., Rapp, C.K., Rath, A.M., Rind, S.A., Rosenberg, A.Z., Segal, M.M., Seidel, M.G., Smedley, D., Talmy, T., Thomas, Y., Wiafe, S.A., Xian, J., Yuksel, Z., Helbig, I., Mungall, C.J., Haendel, M.A. and Robinson, P.N. (2021) 'The Human Phenotype Ontology in 2021', *Nucleic Acids Res*, 49(D1), pp. D1207-D1217.

Accogli, A., Wiegand, G., Scala, M., Cerminara, C., Iacomino, M., Riva, A., Carlini, B., Camerota, L., Belcastro, V., Prontera, P., Fernandez-Jaen, A., Bebek, N., Scudieri, P., Baldassari, S., Salpietro, V., Novelli, G., De Luca, C., von Stulpnagel, C., Kluger, F., Kluger, G.J., Wohlrab, G.C., Ramantani, G., Lewis-Smith, D., Thomas, R.H., Lai, M., Verrotti, A., Striano, S., Depienne, C., Minetti, C., Benfenati, F., Brancati, F., Zara, F. and Striano, P. (2021) 'Clinical and Genetic Features in Patients With Reflex Bathing Epilepsy', *Neurology*, 97(6), pp. e577-e586.

Crawford, K., Xian, J., Helbig, K.L., Galer, P.D., Parthasarathy, S., Lewis-Smith, D., Kaufman, M.C., Fitch, E., Ganesan, S., O'Brien, M., Codoni, V., Ellis, C.A., Conway, L.J., Taylor, D., Krause, R. and Helbig, I. (2021) 'Computational analysis of 10,860 phenotypic annotations in individuals with SCN2A-related disorders', *Genet Med*, 23(7), pp. 1263-1272.

Gulcebi, M.I., Bartolini, E., Lee, O., Lisgaras, C.P., Onat, F., Mifsud, J., Striano, P., Vezzani, A., Hildebrand, M.S., Jimenez-Jimenez, D., Junck, L., Lewis-Smith, D., Scheffer, I.E., Thijs, R.D., Zuberi, S.M., Blenkinsop, S., Fowler, H.J., Foley, A., Epilepsy Climate Change, C. and Sisodiya, S.M. (2021) 'Climate change and epilepsy: Insights from clinical and basic science studies', *Epilepsy Behav*, 116, p. 107791.

2022

Parenti, I., Leitao, E., Kuechler, A., Villard, L., Goizet, C., Courdier, C., Bayat, A., Rossi, A., Julia, S., Bruel, A.L., Tran Mau-Them, F., Nambot, S., Lehalle, D., Willems, M., Lespinasse, J., Ghoumid, J., Caumes, R., Smol, T., El Chehadeh, S., Schaefer, E., Abi-Warde, M.T., Keren, B., Afenjar, A., Tabet, A.C., Levy, J., Maruani, A., Aledo-Serrano, A., Garming, W., Milleret-Pignot, C., Chassevent, A., Koopmans, M., Verbeek, N.E., Person, R., Belles, R., Bellus, G., Salbert, B.A., Kaiser, F.J., Mazzola, L., Convers, P., Perrin, L., Piton, A., Wiegand, G., Accogli, A., Brancati, F., Benfenati, F., Chatron, N., Lewis-Smith, D., Thomas, R.H., Zara, F., Striano, P., Lesca, G. and Depienne, C. (2022) 'The different clinical facets of SYN1-related neurodevelopmental disorders', *Front Cell Dev Biol*, 10, p. 1019715.

Stanley, K., Hostyk, J., Tran, L., Amengual-Gual, M., Dugan, P., Clark, J., Choi, H., Tchapyjnikov, D., Perucca, P., Fernandes, C., Andrade, D., Devinsky, O., Cavalleri, G.L., Depondt, C., Sen, A., O'Brien, T., Heinzen, E., Loddenkemper, T., Goldstein, D.B., Mikati, M.A. and Delanty, N. (2022) 'Genomic analysis of "microphenotypes" in epilepsy', *Am J Med Genet A*, 188(1), pp. 138-146.

Stamberger, H., Crosiers, D., Balagura, G., Bonardi, C.M., Basu, A., Cantalupo, G., Chiesa, V., Christensen, J., Dalla Bernardina, B., Ellis, C.A., Furia, F., Gardiner, F., Giron, C., Guerrini, R., Klein, K.M., Korff, C., Krijtova, H., Leffler, M., Lerche, H., Lesca, G., Lewis-Smith, D., Marini, C., Marjanovic, D., Mazzola, L., McKeown Ruggiero, S., Mochel, F., Ramond, F., Reif, P.S., Richard-Mornas, A., Rosenow, F., Schropp, C., Thomas, R.H., Vignoli, A., Weber, Y., Palmer, E., Helbig, I., Scheffer, I.E., Striano, P., Moller, R.S., Gardella, E. and Weckhuysen, S. (2022) 'Natural History Study of STXBP1-Developmental and Epileptic Encephalopathy Into Adulthood', *Neurology*, 99(3), pp. e221-e233.

Xian, J., Parthasarathy, S., Ruggiero, S.M., Balagura, G., Fitch, E., Helbig, K., Gan, J., Ganesan, S., Kaufman, M.C., Ellis, C.A., Lewis-Smith, D., Galer, P., Cunningham, K., O'Brien, M., Cosico, M., Baker, K., Darling, A., Veiga de Goes, F., El Achkar, C.M., Doering, J.H., Furia, F., Garcia-Cazorla, A., Gardella, E., Geertjens, L., Klein, C., Kolesnik-Taylor, A., Lammertse, H., Lee, J., Mackie, A., Misra-Isrie, M., Olson, H., Sexton, E., Sheidley, B., Smith, L., Sotero, L., Stamberger, H., Syrbe, S., Thalwitzer, K.M., van Berkel, A., van Haelst, M., Yuskaitis, C., Weckhuysen, S., Prosser, B., Son Rigby, C., Demarest, S., Pierce, S., Zhang, Y., Moller, R.S., Bruining, H., Poduri, A., Zara, F., Verhage, M., Striano, P. and Helbig, I. (2022) 'Assessing the landscape of STXBP1-related disorders in 534 individuals', *Brain*, 145(5), pp. 1668-1683.

Elwan, M., Fowkes, R., Lewis-Smith, D., Winder, A., Baker, M.R. and Thomas, R.H. (2022) 'Late-onset cluster seizures and intellectual disability associated with a novel truncation variant in SMC1A', *Epilepsy Behav Rep*, 19, p. 100556.

2023

McLeod, F., Dimtsi, A., Marshall, A.C., Lewis-Smith, D., Thomas, R., Clowry, G.J. and Trevelyan, A.J. (2023) 'Altered synaptic connectivity in an in vitro human model of STXBP1 encephalopathy', *Brain*, 146(3), pp. 850-857.

Daniali, M., Galer, P.D., Lewis-Smith, D., Parthasarathy, S., Kim, E., Salvucci, D.D., Miller, J.M., Haag, S. and Helbig, I. (2023) 'Enriching representation learning using 53 million patient notes through human phenotype ontology embedding', *Artif Intell Med*, 139, p. 102523.

International League Against Epilepsy Consortium on Complex Epilepsies (2023) 'GWAS meta-analysis of over 29,000 people with epilepsy identifies 26 risk loci and subtype-specific genetic architecture', *Nat Genet*, 55(9), pp. 1471-1482.

Carmody, L.C., Gargano, M.A., Toro, S., Vasilevsky, N.A., Adam, M.P., Blau, H., Chan, L.E., Gomez-Andres, D., Horvath, R., Kraus, M.L., Ladewig, M.S., Lewis-Smith, D., Lochmuller, H., Matentzoglu, N.A., Munoz-Torres, M.C., Schuetz, C., Seitz, B., Similuk, M.N., Sparks, T.N., Strauss, T., Swietlik, E.M., Thompson, R., Zhang, X.A., Mungall, C.J., Haendel, M.A. and Robinson, P.N. (2023) 'The Medical Action Ontology: A tool for annotating and analyzing treatments and clinical management of human disease', *Med*, 4(12), pp. 913-927 e3.

Gargano, M.A., Matentzoglu, N., Coleman, B., Addo-Lartey, E.B., Anagnostopoulos, A.V., Anderton, J., Avillach, P., Bagley, A.M., Bakštein, E., Balhoff, J.P., Baynam, G., Bello, S.M., Berk, M., Bertram, H., Bishop, S., Blau, H., Bodenstein, D.F., Botas, P., Boztug, K., Čady, J., Callahan, T.J., Cameron, R., Carbon, S.J., Castellanos, F., Caufield, J.H., Chan, L.E., Chute, C.G., Cruz-Rojo, J., Dahan-Oliel, N., Davids, J.R., de Dieuleveult, M., de Souza, V., de Vries, B.B.A., de Vries, E., DePaulo, J.R., Derfalvi, B., Dhombres, F., Diaz-Byrd, C., Dingemans, A.J.M., Donadille, B., Duyzend, M., Elfeky, R., Essaid, S., Fabrizzi, C., Fico, G., Firth, H.V., Freudenberg-Hua, Y., Fullerton, J.M., Gabriel, D.L., Gilmour, K., Giordano, J., Goes, F.S., Moses, R.G., Green, I., Griese, M., Groza, T., Gu, W., Guthrie, J., Gyori, B., Hamosh, A., Hanauer, M., Hanušová, K., He, Y.O., Hegde, H., Helbig, I., Holasová, K., Hoyt, C.T., Huang, S., Hurwitz, E., Jacobsen, J.O.B., Jiang, X., Joseph, L., Keramatian, K., King, B., Knoflach, K., Koolen, D.A., Kraus, M.L., Kroll, C., Kusters, M., Ladewig, M.S., Lagorce, D., Lai, M.C., Lapunzina, P., Laraway, B., Lewis-Smith, D., Li, X., Lucano, C., Majd, M., Marazita, M.L., Martinez-Glez, V., McHenry, T.H., McInnis, M.G., McMurry, J.A., Mihulová, M., Millett, C.E., Mitchell, P.B., Moslerová, V., Narutomi, K., Nematollahi, S., Nevado, J., et al. (2023) 'The Human Phenotype Ontology in 2024: phenotypes around the world', *Nucleic Acids Res*, 52(D1), pp. D1333-D1346.

8.1.3 Commissioned publications

Lewis-Smith, D., Craig, D.P., Thomas, N.J.P., Hamandi, K. and Thomas, R.H. (2020b) 'Molecular genetic management of epilepsy', in Kumar, D. (ed.) *Clinical Molecular Medicine*. Academic Press, pp. 289-308.

Lewis-Smith, D.K., M; Xian, J; Ganesan, S; Fitzgerald, M; Abend, N; Thomas, RH; Helbig, I (2022) 'Panning for gold: electronic health records in the epilepsy clinic', *Epilepsy Professional*, (Spring), pp. 12-22.

8.2 Conference abstracts

8.2.1 Invited platform presentations

American Academy of Neurology Annual Meeting 2022

April 2022, Seattle, WA, USA: An expanded genotype-phenotype analysis of 11,500 people with epilepsy.

Dianalund International Congress on Epileptology 2022
April 2022, Dianalund, Denmark: The phenotypic landscape of *SCN8A* epilepsies.

Second European *SCN8A* and *SCN2A* Researchers and Families Meeting 2023
May 2023, Køge, Denmark: The phenotypic landscape of *SCN8A* epilepsies.

Coalition to Cure *CHD2* Researchers and Families Meeting 2023
June 2023, Lake Buena Vista, FL, USA: The phenotypic landscape of *CHD2*.

Neurogenetics section of the annual meeting of la Sociedad Española de Neurología 2023
November 2023, Valencia, Spain: Deep phenotyping in neurological diseases.

8.2.2 Other platform presentations

American Epilepsy Society Annual Meeting 2020
December 2020, virtual: Ultra-rare Genetic Variants Influence Clinical Features of the Epilepsies.

American Epilepsy Society Annual Meeting 2021
December 2021, Chicago, IL, USA: An expanded genotype-phenotype analysis of 11,500 people with epilepsy. *Won the Young Investigator Award and selected for invited presentation at the American Academy of Neurology 2022.*

ILAE British Branch Annual Scientific Meeting 2021
October 2021, virtual: The clinical features of 11,544 people with epilepsy according to diagnostic variants in established epilepsy genes. *Won the Clinical Presentation prize*

8.2.3 Poster presentations

ILAE 35th International Congress on Epileptology 2023
September 2023, Dublin, Ireland: Discovery of 25 seizure-associated copy number variants and their clinical characterization in 562 carriers.

ILAE British Branch Annual Scientific Meeting 2023
October 2023, Gateshead, UK: Discovery of 25 seizure-associated copy number variants and their clinical characterization in 562 carriers.

Dianalund International Congress on Epileptology 2022

April 2022, Dianalund, Denmark: The phenotypic landscape of *GABA*-opathies.

ILAE British Branch Annual Scientific Meeting 2021

October 2021, virtual: A case of musicogenic temporal lobe epilepsy.

ILAE 34th International Congress on Epileptology 2021

July 2021, virtual: Analysis of electronic medical records reveals the clinical phenotypic evolution of genetic epilepsies through time.

Newcastle University Translational and Clinical Research Institute Live! 2021

June 2021, virtual: Computational analysis of clinical data to illustrate the clinical spectrum and outcomes of genetic epilepsies.

ClinGen Retreat 2021

June 2021, virtual: Narrowing the nomenclature gap – How to make HPO, OMIM, and other frameworks applicable to the epilepsies and neurodevelopmental disorders.

Association of British Neurologists Annual Meeting 2020

October 2020, virtual: How can we maximise discovery potential from sparse clinical information?

ILAE British Branch Annual Scientific Meeting 2019

October 2019, Birmingham, UK: When 'seizure' won't cut the mustard - Explaining the relationship between different seizures to non-specialists and computers for algorithmic interpretation.

Association of British Neurologists Annual Meeting 2019

May 2019, Edinburgh, UK: Adult phenotypes of paediatric-onset genetic epilepsies.

EMBO Phenotyping Neurological Syndromes for Systems Genetics Course

October 2018, Luxembourg: Familial epilepsies in Antioquia: The GAGE (Genetic Architecture of Generalised Epilepsies) Study.

8.3 Prizes and awards

ILAE Leadership Development Programme Scholarship 2022

American Epilepsy Society Annual Meeting Young Investigator Award 2021

ILAE (British Branch) Annual Scientific Meeting Clinical Presentation Prize 2021

ILAE 34th International Epilepsy Congress Bursary 2020

Wellcome Trust Translational Partnership 2020: co-applicant with Dr Faye McLeod awarded for Investigating the impact of anti-epileptic drugs in developing human brain slice cultures.

Great North Clinical Records Network Grant 2018

North of England Neurological Association Tidswell Travel Grant 2018

8.4 Professional roles within epileptology

8.4.1 Membership of professional bodies and qualifications

Member of the ILAE Epilepsiome Task Force, Genetics Commission, 2019 – present.

Member of the ILAE SNOMED Task Force, Big Data Commission, 2020 – present.

Member of the ClinGen Epilepsy Gene Curation Expert Panel, 2021 – present.

Member of Epi25 Collaborative Strategy and Phenotyping committees, 2021 – present.

Member of the ClinGen Epilepsy Sodium Channel Variant Curation Expert Panel, 2022 – present.

Member of the ILAE Consortium on Complex Epilepsies, 2018 – present.

Member of the EPIGEN, 2018 – present.

Member of the Association of British Neurologists, 2011 – present.

Member of the ILAE (British Branch), 2018 – present.

Member of the American Epilepsy Society, 2019 – present.

Vice-chairperson of the Young Epilepsy Section of the ILAE (British Chapter), 2019–2022.

Contributor to the UK Epilepsy Priority Setting Partnership (Epilepsy Research UK and James Lind Alliance), 2022.

Completion of Higher Specialty Training in Neurology, October 2023.

8.4.2 Peer review

I have reviewed original research on behalf of the editors at Epilepsia, Seizure, Genomics, European Journal of Human Genetics, Molecular Genetics & Genomic Medicine.

8.4.3 Grant review

I have reviewed post-doctoral fellowship proposals on behalf NeuroInsight 2023 (Republic of Ireland).

Appendix B. Collaborators providing data contributing to this work

8.5 Members of the Epi25 Collaborative

8.5.1 Columbia University Institute for Genomic Medicine analysis group

Joshua E. Motelow[1], Gundula Povysil[1], Ryan S. Dhindsa[1], Kate E. Stanley[1], Andrew S. Allen[2], David B. Goldstein[1]

8.5.2 Epi25 sequencing, analysis, project management, and browser development at the Broad Institute

Yen-Chen Anne Feng[3-6], Daniel P. Howrigan[3, 5, 6], Liam E. Abbott[3, 5, 6], Katherine Tashman[3, 5, 6], Felecia Cerrato[6], Caroline Cusick[6], Tarjinder Singh[3, 5, 6], Henrike Heyne[3, 5, 6], Andrea E. Byrnes[3, 5, 6], Claire Churchhouse[3, 5, 6], Nick Watts[3, 5, 6], Matthew Solomonson[3, 5, 6], Dennis Lal[6-8], Namrata Gupta[5], Benjamin M.Neale[3, 5, 6]

8.5.3 Epi25 executive committee

Samuel F. Berkovic[9], Holger Lerche[10], David B. Goldstein[1], Daniel H. Lowenstein[11]

8.5.4 Epi25 strategy, phenotyping, analysis, informatics, and project management committees

Samuel F. Berkovic[9], Holger Lerche[10], David B. Goldstein[1], Daniel H. Lowenstein[11], Gianpiero L. Cavalleri[12, 13], Patrick Cossette[14], Chris Cotsapas[15], Peter De Jonghe[16-18], Tracy Dixon-Salazar[19], Renzo Guerrini[20], Hakon Hakonarson[21], Erin L. Heinzen[1], Ryan S. Dhindsa[1], Kate E. Stanley[1], Ingo Helbig[21-23], Patrick Kwan[24, 25], Anthony G. Marson[26], Slavé Petrovski[25, 27], Sitharthan Kamalakaran[1], Sanjay M. Sisodiya[28],Randy Stewart[29], Sarah Weckhuysen[16-18], Chantal Depondt[30], Dennis J. Dlugos[21, 22], Ingrid E. Scheffer[9], Pasquale Striano[31, 32], Catharine Freyer[11], Roland Krause[33], Patrick May[33], Kevin McKenna[11], Brigid M. Regan[9], Caitlin A. Bennett[9], Stephanie L. Leech[9], Costin Leu[6, 7, 28], David Lewis-Smith[21,23,150,151]

8.5.5 Authors from individual Epi25 cohorts:

Australia: Melbourne (AUSAUS)

Samuel F. Berkovic[9], Ingrid E. Scheffer[9], Brigid M. Regan[9], Caitlin A. Bennett[9], Stephanie L. Leech[9]

Australia: Royal Melbourne (AUSRMB)

Terence J. O'Brien[24, 25], Slavé Petrovski[25, 27], Marian Todaro[24, 25], Patrick Kwan[24, 25]

Belgium: Antwerp (BELATW)

Sarah Weckhuysen[16-18], Peter De Jonghe[16-18], Hannah Stamberger[16-18]

Belgium: Brussels (BELULB)

Chantal Depondt[30]

Canada: Andrade (CANUTN)

Danielle M. Andrade34, 35, Quratulain Zulfiqar Ali34, Tara R. Sadoway35

Switzerland: Bern (CHEUBB)

Heinz Krestel[15], André Schaller[36]

Cyprus (CYPCYP)

Savvas S. Papacostas[37], Ioanna Kousiappa[37], George A. Tanteles[38], Christou Yiolanda[37]

Czech Republic: Prague (CZEMTH)

Katalin Štěrbová[39], Markéta Vlčková[39], Lucie Sedláčková[39], Petra Laššuthová[39]

Germany: Frankfurt/Marburg (DEUPUM)

Karl Martin Klein[40-43], Felix Rosenow[40-42], Philipp S. Reif[40-42], Susanne Knake[41, 42]

Germany: Giessen (DEUUGS)

Bernd A. Neubauer[44], Friedrich Zimprich[45], Martha Feucht[44], Eva Reinthaler[45]

Germany: Bonn (DEUUKB)

Wolfram S. Kunz[46, 47], Gábor Zsurka[46, 47], Rainer Surges[47], Tobias H. Baumgartner[47], Randi von Wrede[47]

Germany: Kiel (DEUUKL)

Ingo Helbig[21, 48, 49], Karl Martin Klein[40-43], Manuela Pendziwiat[48, 49], Hiltrud Muhle[49], Annika Rademacher[49], Andreas van Baalen[49], Sarah von Spiczak[49, 50], Ulrich Stephani[49], Zaid Afawi[51], Amos D. Korczyn[52], Moien Kanaan[53], Christina Canavati[53], Gerhard Kurlemann[54], Karen Müller-Schlüter[55], Gerhard Kluger[56, 57], MartinHäusler[58], Ilan Blatt[52, 59]

Germany: Leipzig (DEUULG)

Johannes R. Lemke[60], Ilona Krey[60]

Germany: Tuebingen (DEUUTB)

Holger Lerche[10], Yvonne G. Weber[10, 61], Stefan Wolking[10], Felicitas Becker[10, 62], Stephan Lauxmann[10], Christian Bosselmann[10], Josua Kegele[10], Christian Hengsbach[10],Sarah Rau[10], Bernhard J. Steinhoff[63], Andreas Schulze-Bonhage[64], Ingo Borggräfe[65, 66], Christoph J. Schankin[67, 68], Susanne Schubert-Bast[64], Herbert Schreiber[69], Thomas Mayer[70], Rudolf Korinthenberg[71], Knut Brockmann[72], Markus Wolff[73], GerhardKurlemann[74], Dieter Dennig[75], Rene Madeleyn[76], Christian Bosselmann[10], Josua Kegele[10]

Finland: Kuopio (FINKPH)

Reetta Kälviäinen[77, 78], Anni Saarela[77], Oskari Timonen[78]

Finland: Helsinki (FINUVH)

Tarja Linnankivi[79], Anna-Elina Lehesjoki[80]

France: Lyon (FRALYU)

Sylvain Rheims[81, 82], Gaetan Lesca[83], Philippe Ryvlin[84], Louis Maillard[85], Luc Valton[86], Philippe Derambure[87],Fabrice Bartolomei[88], Edouard Hirsch[89], Véronique Michel[90], Francine Chassoux[91]

Wales: Swansea (GBRSWU)

Mark I. Rees[92], Seo-Kyung Chung[92, 93], William O. Pickrell[94], Robert H.W. Powell[95], Mark D. Baker[94], BeataFonferko-Shadrach[94], Charlotte Lawthom[96], Joe Anderson[96]

UK: UCL (GBRUCL)

Sanjay M. Sisodiya[28], Natascha Schneider[28], Simona Balestrini[28], Sara Zagaglia[28], Vera Braatz[28]

UK: Imperial/Liverpool (GBRUNL)

Anthony G. Marson[26], Michael R. Johnson[97], Pauls Auce[98], Graeme J. Sills[99]

Hong Kong (HKGHKK)

Patrick Kwan[24, 25, 100], Larry W. Baum[101], Pak C. Sham[101], Stacey S. Cherny[102], Colin H.T. Lui[103]

Ireland: Dublin (IRLRCI)

Gianpiero L. Cavalleri[12, 13], Norman Delanty[12, 13, 104], Colin P. Doherty[13, 105], Arif Shukralla[13, 104], Hany El-Naggar[12, 13, 104], Peter Widdess-Walsh[12, 13, 104]

Croatia (HRVUZG)

Nina Barišić[106]

Italy: Milan (ITAICB)

Laura Canafoglia[107], Silvana Franceschetti[108], Barbara Castellotti[109], Tiziana Granata[110], Francesca Ragona[110]

Italy: Genova (ITAIGI)

Pasquale Striano[31, 32], Federico Zara[31, 32], Michele Iacomino[31], Antonella Riva[31, 32], Francesca Madia[31, 32], Maria Stella Vari[31], Vincenzo Salpietro[31], Marcello Scala[31], Maria Margherita Mancardi[31], Nobili Lino[31], Elisa Amadori[31], Thea Giacomini[31]

Italy: Bologna (ITAUBG)

Francesca Bisulli[111, 112], Tommaso Pippucci[113], Laura Licchetta[111, 112], Raffaella Minardi[111],Paolo Tinuper[111, 112],Lorenzo Muccioli[111, 112], Barbara Mostacci[111]

Italy: Catanzaro (ITAUMC)

Antonio Gambardella[114], Angelo Labate[114], Grazia Annesi[115], Lorella Manna[115], Monica Gagliardi[115]

Italy: Florence (ITAUMR)

Renzo Guerrini[20], Elena Parrini[20], Davide Mei[20], Annalisa Vetro[20], Claudia Bianchini[20], Martino Montomoli[20],Viola Doccini[20], Carmen Barba[20]

Japan: Fukuoka (JPNFKA)

Shinichi Hirose[116], Atsushi Ishii[116]

Japan: RIKEN Institute (JPNRKI)

Toshimitsu Suzuki[117, 118], Yushi Inoue[119], Kazuhiro Yamakawa[117, 118]

Lebanon: Beirut (LEBABM)

Ahmad Beydoun[120], Wassim Nasreddine[120], Nathalie Khoueiry-Zgheib[121]

Lithuania (LTUUHK)

Birute Tumiene[122, 123], Algirdas Utkus[122, 123]

New Zealand: Otago (NZLUTO)

Lynette G. Sadleir[124], Chontelle King[124]

Turkey: Bogazici (TURBZU)

S. Hande Caglayan[125], Mutluay Arslan[126], Zuhal Yapıcı[127], Pınar Topaloglu[127], Bulent Kara[128], Uluc Yis[128],Dilsad Turkdogan[128], Aslı Gundogdu-Eken[125]

Turkey: Istanbul (TURIBU)

Nerses Bebek[129, 130], Sibel Uğur-İşeri[130], Betül Baykan[129], Barış Salman[130], Garen Haryanyan[129], EmrahYücesan[131], Yeşim Kesim[129], Çiğdem Özkara[132]

Taiwan (TWNCGM)

Meng-Han Tsai[133], Chen-Jui Ho[133], Chih-Hsiang Lin[133], Kuang-Lin Lin[134], I-Jun Chou[134]

USA: BCH (USABCH)

Annapurna Poduri[135, 136], Beth R. Shiedley[135, 136], Catherine Shain[135, 136]

USA: Baylor College of Medicine (USABLC)

Jeffrey L. Noebels[137], Alicia Goldman[137]

USA: Cleveland Clinic (USACCF)

Robyn M. Busch[8, 138, 139], Lara Jehi[138, 139], Imad M. Najm[138, 139], Dennis Lal[8, 138], Lisa Ferguson[138], Jean Khoury[138, 139]

USA: Cincinnati Children's Hospital Medical Center (USACCH)

Tracy A. Glauser[140], Peggy O. Clark[140]

USA: Philadelphia/CHOP (USACHP) and Philadelphia/Rowan (USACRW)

Russell J. Buono[21, 141, 142], Thomas N. Ferraro[141], Michael R. Sperling[142], Dennis J. Dlugos[21, 143], Warren Lo[143] , Michael Privitera[144], Jacqueline A. French[145], Patrick Cossette[14], Steven Schachter[146], Hakon Hakonarson[21]

USA: EPGP (USAEGP)

Daniel H. Lowenstein[11], Ruben I. Kuzniecky[147], Dennis J. Dlugos[21, 22], Orrin Devinsky[145]

USA: NYU HEP (USAHEP)

Daniel H. Lowenstein[11], Ruben I. Kuzniecky[147], Jacqueline A. French[145], Manu Hegde[11]

USA: Nationwide Children's Hospital (USANCH)

David A. Greenberg[148]

USA: Penn/CHOP (USAUPN)

Ingo Helbig[21-23], Colin A. Ellis[22], Ethan Goldberg[21, 149], Katherine L. Helbig[21, 149], David Lewis-Smith[21,23,150,151], Mahgenn Cosico[21, 149], PriyaVaidiswaran[21, 149], Eryn Fitch[21, 149]

Kenya: Kilifi; South Africa: Aguincourt; Ghana: Kintampo (KENKIL, GHAKNT, ZAFAGN)

Charles RJC Newton[152, 153, 154], Symon M Kariuki[152, 153, 154], Ryan G Wagner[155, 156, 157], Seth Owusu-Agyei[158, 15

USA: Massachusetts General Hospital (USAMGH)

Andrew J. Cole[160], Christopher M. McGraw[160], S. Anthony Siena[161]

USA: Vanderbilt University Medical Centre (USAVAN)

Lea Davis[162-65], Donald Hucks[162, 165], Annika Faucon[165], David Wu[165], Bassel W. Abou-Khalil[165], Kevin Haas[165], Randip S. Taneja[165]

Affiliations

[1] Institute for Genomic Medicine, Columbia University, New York, NY 10032, USA

[2] Division of Integrative Genomics, Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27710,USA

[3] Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard MedicalSchool, Boston, MA 02114, USA

[4] Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital andHarvard Medical School, Boston, MA 02114, USA

[5] Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA02142, USA

[6] Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

[7] Genomic Medicine Institute, Cleveland Clinic, Cleveland, OH 44195, USA

[8] Center for Genomic Medicine, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA

[9] Epilepsy Research Centre, Department of Medicine, University of Melbourne, Victoria, Australia

[10] Department of Neurology and Epileptology, Hertie Institute for Clinical Brain Research, University of Tübingen, 72076Tübingen, Germany

[11] Department of Neurology, University of California, San Francisco, CA 94110, USA

[12] School of Pharmacy and Biomolecular Sciences, The Royal College of Surgeons in Ireland, Dublin, Ireland

[13] The FutureNeuro Research Centre, Ireland

[14] University of Montreal, Montreal, QC H3T 1J4, Canada

[15] School of Medicine, Yale University, New Haven, CT 06510, USA

[16] Neurogenetics Group, Center for Molecular Neurology, VIB, Antwerp, Belgium

[17] Laboratory of Neurogenetics, Institute Born-Bunge, University of Antwerp, Belgium

[18] Division of Neurology, Antwerp University Hospital, Antwerp, Belgium

[19] LGS Foundation, NY 11716, USA

[20] Pediatric Neurology, Neurogenetics and Neurobiology Unit and Laboratories, Children's Hospital A. Meyer, University ofFlorence, Italy

[21] Division of Neurology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[22] Perelman School of Medicine, University of Pennsylvania, PA 19104, USA

[23] Department of Biomedical and Health Informatics (DBHi), Children's Hospital of Philadelphia, Philadelphia, PA, USA

[24] Department of Neuroscience, Central Clinical School, Monash University, Alfred Hospital, Melbourne, Australia[25] Departments of Medicine and Neurology, University of Melbourne, Royal Melbourne Hospital, Parkville, Australia[26] Department of Pharmacology and Therapeutics, University of Liverpool, Liverpool, UK

[27] Centre for Genomics Research, Precision Medicine and Genomics, IMED Biotech Unit, AstraZeneca, Cambridge, UK

[28] Department of Clinical and Experimental Epilepsy, UCL Queen Square Institute of Neurology, London, UK and ChalfontCentre for Epilepsy, Chalfont St Peter, UK

[29] National Institute of Neurological Disorders and Stroke, MD 20852, USA

[30] Department of Neurology, Université Libre de Bruxelles, Brussels, Belgium

[31] IRCCS "G. Gaslini" Institute, Genova, Italy

[32] Department of Neurosciences, Rehabilitation, Ophthalmology, Genetics, Maternal and Child Health, University of Genoa, Italy

[33] Luxembourg Centre for Systems Biomedicine, University Luxembourg, Esch-sur-Alzette, Luxembourg

[34] Department of Neurology, Toronto Western Hospital, Toronto, ON M5T 2S8, Canada

[35] University Health Network, University of Toronto, Toronto, ON, Canada

[36] Institute of Human Genetics, Bern University Hospital, Bern, Switzerland

[37] Department of Neurobiology and Clinical Neurophysiology (Epilepsy Center), The Cyprus Institute of Neurology and Genetics, 2370 Nicosia, Cyprus

[38] Department of Clinical Genetics, The Cyprus Institute of Neurology and Genetics, 2370 Nicosia, Cyprus

[39] Department of Paediatric Neurology, 2nd Faculty of Medicine, Charles University and Motol Hospital, Prague, CzechRepublic

[40] Epilepsy Center Frankfurt Rhine-Main, Center of Neurology and Neurosurgery, Goethe University Frankfurt, Frankfurt,Germany

[41] Epilepsy Center Hessen-Marburg, Department of Neurology, Philipps University Marburg, Marburg, Germany

[42] LOEWE Center for Personalized Translational Epilepsy Research (CePTER), Goethe University Frankfurt, Germany

[43] Departments of Clinical Neurosciences, Medical Genetics and Community Health Sciences, Hotchkiss Brain Institute & Alberta Children's Hospital Research Institute, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada

[44] Pediatric Neurology, Children's Hospital, Giessen, Germany

[45] Department of Neurology, Medical University of Vienna, Vienna, Austria

[46] Institute of Experimental Epileptology and Cognition Research, University Bonn, 53127 Bonn, Germany

[47] Department of Epileptology, University Bonn, 53127 Bonn, Germany

[48] Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany

[49] Department of Neuropediatrics, Christian-Albrechts-University of Kiel, 24105 Kiel, Germany

[50] DRK-Northern German Epilepsy Centre for Children and Adolescents, 24223 Schwentinental-Raisdorf, Germany

[51] Sackler School of Medicine, Tel-Aviv University, Ramat Aviv, Israel

[52] Tel-Aviv University Sackler Faculty of Medicine, Ramat Aviv 69978, Israel

[53] Hereditary Research Lab, Bethlehem University, Bethlehem, Palestine

[54] Bonifatius Hospital Lingen, Neuropediatrics Wilhelmstrasse 13, 49808 Lingen, Germany

[55] Epilepsy Center for Children, University Hospital Neuruppin, Brandenburg Medical School, Neuruppin, Germany[56] Neuropediatric Clinic and Clinic for Neurorehabilitation, Epilepsy Center for Children and Adolescents, Vogtareuth,Germany

[57] Research Institute Rehabilitation / Transition / Palliation, PMU Salzburg, Austria

[58] Division of Neuropediatrics and Social Pediatrics, Department of Pediatrics, University Hospital, RWTH Aachen,Aachen, Germany

[59] Department of Neurology, Sheba Medical Center, Ramat Gan, Israel

[60] Institute of Human Genetics, Leipzig, Germany

[61] Department of Epileptology and Neurology, University of Aachen, Aachen, Germany

[62] Department of Neurology, University of Ulm, Ulm, Germany

[63] Epilepsy Center Kork, Kehl, Germany

[64] Department of Epileptology, University Hospital Freiburg, Freiburg, Germany

[65] Department of Pediatric Neurology, Ludwig Maximilians University, Munchen, German

[66] Dr von Hauner Childrens Hospital, Munich, Germany

[67] Department of Neurology, Ludwig Maximilians University, Munchen, Germany

[68] Department of Neurology, University Hospital Bern, Bern, Switzerland

[69] Neurological Practice Center & Neuropoint Patient Academy, Ulm, Germany

[70] Epilepsy Center Kleinwachau, Radeberg, Germany

[71] Department of Neuropediatrics and Muscular Disorders, University Hospital Freiburg, Freiburg, Germany

[72] University Children's Hospital, Göttingen, Germany

[73] Department of Pediatric Neurology & Developmental Medicine, University Childrens Hospital, Tubingen, Germany

[74] Department of Neuropediatrics, Westfalische Wilhelms University, Munster, Germany

[75] Private Neurological Practice, Stuttgart, Germany

[76] Department of Pediatrics, Filderklinik, Filderstadt, Germany

[77] Epilepsy Center, Kuopio University Hospital, Kuopio, Finland

[78] Institute of Clinical Medicine, University of Eastern Finland, Finland

[79] Child Neurology, New Children's Hospital and Pediatric Research Center, University of Helsinki and Helsinki UniversityHospital, Helsinki, Finland

[80] Medicum, University of Helsinki, Helsinki, Finland and Folkhälsan Research Center, Helsinki, Finland

[81] Department of Functional Neurology and Epileptology, Hospices Civils de Lyon and University of Lyon, France

[82] Lyon's Neuroscience Research Center, INSERM U1028 / CNRS UMR 5292, Lyon, France

[83] Department of Clinical Genetics, Hospices Civils de Lyon and University of Lyon, Lyon, France

[84] Department of Clinical Neurosciences, Centre Hospitalo-Universitaire Vaudois, Lausanne, Switzerland

[85] Neurology Department, University Hospital of Nancy, UMR 7039, CNRS, Lorraine University, Nancy, France

[86] Clinical Neurophysiology and Epilepsy Unit, Department of Neurology, University Hospital of Toulouse. Toulouse,France

[87] Department of Clinical Neurophysiology, Lille University Medical Center, EA 1046, University of Lille, France

[88] Clinical Neurophysiology and Epileptology Department. Timone Hospital, Marseille, France

[89] Department of Neurology, University Hospital of Strasbourg, Strasbourg, France

[90] Department of Neurology, Hôpital Pellegrin, Bordeaux, France

[91] Epilepsy Unit, Department of Neurosurgery, Centre Hospitalier Sainte-Anne and University Paris Descartes, Paris,France

[92] Faculty of Medicine and Health, University of Sydney, Sydney, Australia

[93] Kids Neuroscience Centre, Kids Research, Children Hospital at Westmead, Sydney, New South Wales, Australia

[94] Swansea University Medical School, Swansea University SA2 8PP, UK

[95] Department of Neurology, Morriston Hospital, Swansea Bay University Health Board, Swansea, Wales UK

[96] Neurology Department, Aneurin Bevan University Health Board, Newport, Wales, UK

[97] Division of Brain Sciences, Imperial College London, London, UK

[98] Department of Neurology, Walton Centre NHS Foundation Trust, Liverpool, UK

[99] School of Life Sciences, University of Glasgow, Glasgow, UK

[100] Department of Medicine and Therapeutics, Chinese University of Hong Kong, Hong Kong, China

[101] Department of Psychiatry, University of Hong Kong, Hong Kong, China

[102] Department of Epidemiology and Preventive Medicine and Department of Anatomy and Anthropology, Sackler Facultyof Medicine, Tel Aviv University, Israel

[103] Department of Medicine, Tseung Kwan O Hospital, Hong Kong, China

[104] The Department of Neurology, Beaumont Hospital, Dublin, Ireland

[105] Neurology Department, St. James Hospital, Dublin, Ireland

[106] Department of Pediatric University Hospital centre Zagreb, Croatia

[107] Integrated Diagnostics for Epilepsy, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy

[108] Neurophysiology, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy

[109] Unit of Genetics of Neurodegenerative and Metabolic Diseases, Fondazione IRCCS Istituto Neurologico Carlo Besta,Milan, Italy

[110] Department of Pediatric Neuroscience, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy

[111] IRCCS, Institute of Neurological Sciences of Bologna, Bologna, Italy

[112] Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna, Italy

[113] Medical Genetics Unit, Polyclinic Sant'Orsola-Malpighi University Hospital, Bologna, Italy

[114] Institute of Neurology, Department of Medical and Surgical Sciences, University "Magna Graecia", Catanzaro, Italy

[115] Institute of Molecular Bioimaging and Physiology, CNR, Section of Germaneto, Catanzaro, Italy

[116] Department of Pediatrics, School of Medicine, Fukuoka University

[117] Laboratory for Neurogenetics, RIKEN Center for Brain Science, Saitama, Japan

[118] Department of Neurodevelopmental Disorder Genetics, Institute of Brain Science, Nagoya City University GraduateSchool of Medical Science, Nagoya, Aichi, Japan

[119] National Epilepsy Center, Shizuoka Institute of Epilepsy and Neurological Disorder, Shizuoka, Japan

[120] Department of Neurology, American University of Beirut Medical Center, Beirut, Lebanon

[121] Department of Pharmacology and Toxicology, American University of Beirut Faculty of Medicine, Beirut, Lebanon

[122] Institute of Biomedical Sciences, Faculty of Medicine, Vilnius University, Vilnius, Lithuania

[123] Centre for Medical Genetics, Vilnius University Hospital Santaros Klinikos, Vilnius, Lithuania

[124] Department of Paediatrics and Child Health, University of Otago, Wellington

[125] Department of Molecular Biology and Genetics, Bogaziçi University, Istanbul, Turkey

[126] Department of Child Neurology, Gulhane Education and Research Hospital, Health Sciences University, Ankara,Turkey

[127] Department of Child Neurology, Istanbul Faculty of Medicine, Istanbul University, Istanbul, Turkey

[128] Department of Child Neurology, Medical School, Marmara University, Istanbul, Turkey

[129] Department of Neurology, Istanbul Faculty of Medicine, Istanbul University, Istanbul, Turkey

[130] Department of Genetics, Aziz Sancar Institute of Experimental Medicine, Istanbul University, Istanbul, Turkey

[131] Bezmialem Vakif University, Institute of Life Sciences and Biotechnology, Istanbul, Turkey

[132] Department of Neurology, Faculty of Medicine, Cerrahpaşa University Istanbul, Istanbul, Turkey

[133] Department of Neurology, Kaohsiung Chang Gung Memorial Hospital, Kaohsiung, Taiwan

[134] Department of Pediatric Neurology, Chang Gung Memorial Hospital, Taoyuan, Taiwan

[135] Epilepsy Genetics Program, Department of Neurology, Boston Children's Hospital, Boston, MA 02115, USA

[136] Department of Neurology, Harvard Medical School, Boston, MA 02115, USA

[137] Neurology, Baylor College of Medicine

[138] Cleveland Clinic Epilepsy Center, Neurological Institute, Cleveland Clinic, Cleveland, OH 44195, USA

[139] Department of Neurology, Neurological Institute, Cleveland Clinic, Cleveland, OH 44195, USA

[140] Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

[141] Cooper Medical School of Rowan University, Camden, NJ 08103, USA

[142] Thomas Jefferson University, Philadelphia, PA 19107, USA

[143] Nationwide Children's Hospital, Columbus, OH, 3205, USA

[144] University of Cincinnati, Cincinnati, OH 45220, USA

[145] Department of Neurology, New York University, New York NY, USA

[146] Beth Israel Deaconess/Harvard, Boston, MA 02115, USA

[147] Department of Neurology, Hofstra-Northwell Medical School, New York, NY 11549, USA

[148] Department of Pediatrics, Ohio State Wexner Medical Center, Battelle Center for Mathematical Medicine. Columbus,Ohio, USA.

[149] The Epilepsy NeuroGenetics Initiative (ENGIN), Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[150] Translational and Clinical Research Institute, Newcastle University, Newcastle Upon Tyne, UK

[151] Department of Clinical Neurosciences, Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle Upon Tyne, UK

[152] Neuroscience Unit, KEMRI-Wellcome Trust Research Programme, Kilifi, Kenya

[153] Department of Public Health, Pwani University, Kilifi, Kenya

[154] Department of Psychiatry, University of Oxford, Oxford, UK

[155] MRC/Wits Rural Public Health & Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[156] Department of Epidemiology and Global Health, Umeå University, Umeå, Sweden

[157] Department of Clinical Sciences, Neurosciences, Umeå University, Umeå, Sweden

[158] Kintampo Health Research Centre, Ghana Health Service, Kintampo, Ghana

[159] University of Health and Allied Science in Ho, Ghana.

[160] Neurology, Massachusetts General Hospital, Boston, MA, USA

[161] Medical School, Nova Southeastern University, Fort Lauderdale, FL, USA

[162] Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

[163] Department of Psychiatry and Behavioral Sciences, Vanderbilt University Medical Center, Nashville, TN, USA

[164] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA

[165] Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA

8.6 Collaborators contributing data about adults with CDKL5 Deficiency Disorder

Angel Aledo-Serrano, Helen Leonard, Allan Bayat, Mohamed Junaid, Eveline Hagebeuk, Christina D Fenger, Juliana Laze, Alessandra Rossi, Dragan Marjanovic, Marina Trivisano, Nicola Specchio, Beatriz Gonzalez-Giraldez, Julio Lama, Ilona Krey, Konrad Platzer, Elise Brischoux-Boucher, Catherine Sarret, Lysa Boissé Lomax, Caterina Zanus, Luciana Musante,

Paola Costa, Patrick Moloney, Norman Delanty, Angelo Russo, Bitten Schönewolf-Greulich , Anne-Marie Bisgaard, Colette Berger, Elena Freri, Satoru Takahashi, Pia Zacher, Julien Jung, Scott Demarest, Eric Marsh, Alan Percy, Jeff Neul, Heather Olson, Lindsay Swanson, Stefano Meletti, Maria Cristina Cioclu, Quratulain Z Ali, Ana Suller, Rafael Toledano, Irene Garcia-Morales, Alvaro Beltran-Corbellini, Antonio Gil-Nagel, Xiaoming Zhang, Roberto Previtali, Anne F Højte, Jenny Downs, Ingo Helbig, Gaetan Lesca, Guido Rubboli, Danielle Andrade, Elena Gardella, Elia Pestana, Orrin Devinsky, Timothy Benke, Rhys Thomas, Rikke S. Møller