A systematic comparison of integrated genomic platforms and bioinformatics pipelines for next generation sequencing in patients with rare neuromuscular disease.

Hadil Alrohaif

Thesis submitted for the degree of Doctor of Medicine (MD)

Newcastle University

Faculty of Medical Sciences

Institute of Genetic Medicine

August 2018

Abstract

Neuromuscular disorders are a group of genetically and phenotypically heterogeneous disorders and pose a challenge for molecular diagnosis. Next generation sequencing is increasingly used in research and clinical settings for diagnosis and disease gene discovery. Inconsistencies in bioinformatics pipelines, research and validation results suggest that bioinformatics tools for next generation sequencing are yet problematic and that further research is needed. In addition, sequencing data, bioinformatics tools, clinical data and databases of current knowledge of the human genome need to be integrated in an effective workflow that facilitates diagnosis and novel gene discoveries. Furthermore, optimising and standardising analysis workflow for next generation sequencing allows data from different projects and research sites to be shared and validated.

At the John Walton Muscular Dystrophy Research Centre, Newcastle University, three genomics platforms are used to analyse whole exome and whole genome sequencing data for patients with rare neuromuscular disease. These three platforms, namely: RD-Connect Genome-Phenome Analysis Platform (CNAG, Barcelona, academic), *seqr* (Broad Institute, Boston, academic) and the Clinical Sequence Analyser (CSA, WuXi NextCODE, commercial) use combinations of different bioinformatics tools and integrate different software applications and databases for variant annotation, filtering and prioritization.

Here, the aim was to compare the yield of genome sequencing over exome sequencing for patients with rare neuromuscular disorders and to assess the degree of agreement between the three genomic platforms and their respective bioinformatics pipelines. I also aimed to evaluate the value of using an integrated genomics platform in diagnosis and novel gene discovery in patients with rare neuromuscular disorders.

The analysis showed that whole genome sequencing offers more uniform coverage of coding regions in the genome and has the potential to detect additional coding variants in known neuromuscular disease genes that are missed by exome sequencing due to low coverage. Low coverage was associated with genomic features such as high GC-content and low sequence heterogeneity. The uniform coverage and sequencing methods used for whole genome sequencing may also lead to improved detection of InDels and copy number variants in this group of patients.

Analysis of the bioinformatics pipelines at the three sites using patient WES and WGS data revealed that the highest agreement was between the RD-Connect and the CSA platforms (75%). However, using high quality reference data revealed higher concordance rates (up to 91%).

As for variant output from the three genomics platforms, the mean variant concordance for all three platforms was 37%, and the highest pairwise concordance rate was 66% for *seqr* and RD-Connect. Looking at variant type, agreement in variant output was largely accounted for by single nucleotide variants and InDel agreement was significantly low. Comparing the variant output between the three platforms revealed very low agreement. This highlights variant annotation software and filtering algorithms as contributors to the discrepancy in variant output.

Whole exome sequencing data from molecularly undiagnosed families with limb girdle muscle weakness were used on the *seqr* platform to assess the utility of an integrated genomics platform in diagnosis and disease gene discovery in patients with rare neuromuscular disorders. This analysis showed that for 65.6% of families, a genetic diagnosis was proposed. This included a number of proposed novel genetic associations in neuromuscular disorders, including the recently published *MYMK* gene and the *FILIP1* gene, which is projected as a strong candidate for syndromic congenital myopathy.

Once a genetic diagnosis for a rare disease is established, phenotype-genotype correlations can be established. A group of patients with genetically confirmed GNE myopathy from Kuwait were studied. A description of clinical, genetic and epidemiological aspects of the disease in the Kuwaiti Bedouin population is given.

In conclusion, next generation sequencing undoubtedly continues to offer new insights in rare neuromuscular disorders. However, advances in bioinformatics need to match advances in sequencing technologies. Whole genome sequencing offers additional value over whole exome sequencing. Nevertheless, it remains costly and data interpretation is still problematic. A targeted approach to the analysis of whole genome sequencing data may be a more appropriate intermediate approach. Analysis pipelines require a standardised approach for development and validation. Moreover, bioinformatics algorithms remain an area for continued assessment and optimisation. This will maximise the benefit from research in next generation sequencing and enable data to be shared and compared. Lastly, integrated genomics platforms are an ideal interface between the researcher and all relevant genetic,

phenotypic, population and bioinformatics prediction data, for diagnosis and novel gene discoveries in patients with rare neuromuscular disorders.

"Although I cannot move and I have to speak through a computer, in my mind I am free."

Stephen Hawking (1942-2018)

Dedicated to people living with disabilities and to all those who support and care for them.

Acknowledgements

I would like to express my deepest appreciation to my supervisor and mentor professor Hanns Lochmüller. I would like to thank him for his expert supervision and guidance. I am grateful for his continued encouragement and advice in research and in professional development.

I am extremely grateful to my supervisor and friend Dr Ana Töpf for her continued support and day-to-day approachability in giving advice and answering questions. In particular, I'm grateful for her prioritising my work despite her difficult workload.

I thank Dr Phillip Lord for his input as an annual review panel member and for agreeing to supervise the final stages of my work.

I am grateful to Rachel Thompson for her approachability to answer questions related to the RD-Connect platform and her continued input and advice in relation to my project.

I am thankful to Professor Robert Lightowlers, Dr Mauro Santibanez-Koref, John Dawson, Dr Helen Griffin, Louise Pease, Dr Rita Baressi, Dr Teresinha Evangelista, Dr Chiara Marini-Bettello, Dr Michela Guglieri, Dr Tuomo Polvikoski, Professor Rita Horvath and all my colleagues in the muscle group for their support and useful discussions.

I would also like to acknowledge Isaac Walton and Rebecca Haigh's contribution in performing family segregation studies to support my work, which was done as part of their undergraduate training.

I am grateful to the teams at RD-Connect (CNAG), deCODE Genetics and the Broad Institute, especially to Steve Laurie, Nanna Vioarsdottir, Olafur Magnusson, Monkol Lek and Michael Wilson for answering queries regarding the platforms and bioinformatics pipelines and for processing patient and reference data included in this thesis.

I would also like to thank Dr Laila Bastaki, Dr Ali AlAjmi and the clinicians and laboratory staff at the Kuwait Medical Genetics Centre for their help in data collection that enabled me to put together a detailed description of GNE myopathy in Kuwait.

I am most grateful to Dr Andoni Urtizberea, at the Neuromuscular Reference Centre in Hendaye Hospital in France, for initiating and supporting the GNE myopathy work and for his continued encouragement. And a special thanks to Oksana Pogoryelova for all her input and help with the GNE myopathy work.

I am deeply thankful to my husband Ameer and to my children Mohammad and Sama for being there, for their love and support, and for putting up with our difficult living situation during the past few years away from home. I am also indebted to my family and friends in Kuwait for their love, support and their prayers.

My deepest gratitude goes to my rock and my "survival buddy" Sumaya. Her existence in my life over the past few years has been invaluable.

A special thanks goes to my dear friend Najwa for being there when I needed her and for her continued encouragement and support.

Finally, I would like to thank my friends, Selma, Salome, Katy, Olla, Michelle, Liz, Persefoni, Anne and Kath for helping me through stressful times, for encouraging and motivating me and for adding some fun to the past three years.

Declaration

I, Hadil Alrohaif, declare that the work presented in this thesis is a result of my own original research. I confirm that contributions from others are clearly acknowledged and that any published work is referenced. I certify that the thesis contains no material that has been submitted for any other academic award.

Table of contents

List of tables

List of figures

List of abbreviations

ACMG	American College for Medical Genetics and Genomics
API	Application Programming Interface
BAM	Binary alignment map
BP	Base pair
BQSR	Base quality score recalibration
BWA	Burrows-Wheeler Aligner
BWT	Burrows-Wheeler Transform
CGH-array	Comparative Genomics Hybridization
CK	Creatine kinase
CMS	Congenital myasthenic syndrome
CNAG	Centro Nacional de Análisis Genómico (National Centre for Genomics Analysis, Barcelona)
CNV	Copy number variant
CSA	Clinical Sequence Analyzer
DMRV	Distal myopathy with rimmed vacuoles
DNA	Deoxyribonucleic acid
EBI	European Bioinformatics Institute
ECG	Electrocardiopgarphy
EMG	Electromyopgraphy
ExAC	Exome Aggregation Consortium
GATK	Genome Analysis Tool Kit
GIAB	Genome in a Bottle
gnomAD	Genome Aggregation Database
GQ	Genotype quality
GTEx	Genotype Tissue Expression (database)
H&E	Haematoxylin Eosin
HDPS	Hadoop distributed file system
HGVS	Human Genome Variation Society
HIBM	Hereditary inclusion body myopathy
HPO	Human Phenotype Ontology
HSMN	hereditary sensory motor neuropathy
IGV	Integrated Genomics Viewer
InDels	Insertions/deletions
KB	Kilobyte
KMGC	Kuwait Medical Genetics Centre
LGMD	Limb girdle muscular dystrophy
MFM	Myofibrillar myopathy
MHC	Myosin heavy chain
MLPA	Multiple ligase probe amplification
MME	Matchmaker Exchange (project)
MRC	Medical Research Council
MRI	Magnetic resonance imaging
MSeqDR	Mitochondrial Disease Sequence Data Resource

NADH	Nicotinamide adenine dinucleotide
NCAM	Neural cell adhesion molecule
NGS	Next generation sequecning
NIST	National Intitute for Standards and Technology
NMD	Neuromuscular disorders
OLC	Overlap/Layout/Consensus (method)
OMIM	Online Mendelian Inheritance in Man (database)
PCR	Polymerase chain reaction
QC	Quality control
RNA	Ribonucleic acid
SMRT	Single molecule real time (system)
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SO	Sequence ontology
VEP	Variant Effect Predictor (tool)
VQSR	Variant quality score recalibration
VUS	Variants of unknown significance
WCD	Wheelchair dependant
WES	Whole exome sequencing
WGS	Whole genome sequencing
WMS	World Muscle Society

Chapter 1. Introduction

1.1 Rare genetic diseases

A rare or orphan disease is defined as one that affects less than 1 in 2000 individuals. In many cases, the aetiology of rare diseases is a genetic mutation or susceptibility. Between 6500 and 7000 rare diseases are currently characterized and this number is continuously rising in the current genomics era (Eurordis, 2005; Daunert *et al.*, 2017). These diseases are often chronic, progressive, degenerative and life threatening. They present a spectrum of disabilities that make patients and their families vulnerable and isolated. In addition, a considerable number of these diseases have no parent organization or investigator dedicated to research on their prevention, diagnosis or treatment. The National Institute of Health (NIH) Undiagnosed Disease Program, launched in 2008, reported that 33% of patients with a rare disease waited 1-5 years for a diagnosis and 15% waited more than five years (Gahl and Tifft, 2011). EURORDIS (http://www.eurordis.org) also reports similar statistics, where 25% of patients with a rare disease in Europe waited between 5 and 30 years for a diagnosis. 40% were initially misdiagnosed which resulted in 33% of them receiving incorrect treatment and in some cases an unnecessary invasive surgical procedure (Eurordis). A thorough genetic evaluation and molecular diagnosis in rare genetic disorders is thus warranted to direct care and counselling. In addition, certain molecular forms of a genetic disorder are specifically treated with a particular agent. For example, neonatal diabetes associated with mutations in the *KCNJ11* gene does not respond to insulin therapy and is best treated with sulfonylureas (Babiker *et al.*, 2016). In the neurogenetics context, children with epilepsy and mutations in the *SLC2A1* gene are managed through a ketogenic diet, a therapy that is less effective in other molecular forms of childhood epilepsy (Klepper, 2015). Furthermore, some rare genetic disorders are associated with medical conditions and complications and correct diagnosis will guide appropriate surveillance and intervention. It is also necessary for accurate genetic counselling and communication of recurrence risk. Lastly, molecular diagnosis for rare disorders provides insights into disease mechanisms directing medically relevant research and providing new candidate therapeutic targets (Gahl and Tifft, 2011; Efthymiou *et al.*, 2016).

1.2 Approach for molecular diagnosis for rare genetic diseases

A common traditional approach for diagnosis of genetic disorders is initiated with the clinicians' medical evaluation based on the patient's presenting symptoms, age of onset, disease progression, associated features, clinical examination findings and family history. This in most cases is followed by a tailored and tiered series of laboratory and imaging tests. At this stage, a clinician may put forward a suspected clinical diagnosis. This will then dictate the choice of genetic test, which may take the form of one of the following: targeted single gene sequencing (Sanger sequencing) or mutation analysis, methylation testing or chromosomal studies. In some instances, multiple single gene tests are considered and tiered based on availability and the patient's clinical picture. If these initial tests are negative or if the initial clinical evaluation does not suggest a particular genetic condition, chromosomal micro-array (CGH-array) or next generation sequencing (NGS) in the form of a gene panel, whole exome sequencing (WES) or whole genome sequencing (WGS) is indicated. These options are further discussed below and summarised in table 1. A further scenario is when the clinical evaluation does not point towards a genetic disorder, in which case no further genetic testing is indicated (Shashi *et al.*, 2014; Warman Chardon *et al.*, 2015).

Table 1: Molecular testing methods in rare inherited diseases.

	Single gene test	**CGH-array**	**Gene panel test**	**WES**	**WGS**
Indication	Phenotype commonly associated with one gene	Genetically heterogeneous disorders where CNV and structural rearrangements are implicated mechanisms	Genetically heterogeneous and phenotypically overlapping disorders	Rare diseases, non-specific features or those with high phenotypic or genetic heterogeneity	Rare diseases, non-specific features or those with high phenotypic or genetic heterogeneity
Genes	Single disease causing gene(s)	Whole genome. Can be tailored to target regions	10s-100 disease associated genes	~22,000 genes	Whole genome
Coverage	Excellent coverage	-	Deep coverage of protein-coding regions	Variable coverage of protein-coding regions	Uniform coverage of whole genome
Risk of VUS	Minimal	High	Variable but lower than WES and WGS	High, in coding regions	High, in coding and non-coding regions
Incidental findings	None	Possible	Variable number but less likely than WES and WGS	Possible	Possible
Time to return of result	2-4 weeks	2-6 weeks	6-8 weeks	9-12 weeks	>12 weeks
Cost per run	Relatively expensive per base	£100-500	£140	£280	£800

Sanger confirmation	Sanger sequencing used and is gold standard	Not required	Requires confirmation with Sanger sequencing	Requires confirmation with Sanger sequencing	Requires confirmation with Sanger sequencing

CGH; comparative genomics hybridisation, CNV; copy number variants, VUS; variants of unknown significance, WES; whole exome sequencing and WGS; whole genome sequencing.

1.2.1 Single gene testing

The choice of a single gene test is dictated by the patient's medical evaluation. Sanger sequencing is the current gold standard for mutation detection. The single gene testing approach is relevant in cases where there is a strong phenotype-genotype correlation, for genetic disorders with distinct phenotypes and where certain gene mutations have a relatively high frequency for a particular disorder in specific populations (Xue *et al.*, 2015). Otherwise, this approach can be time and resource consuming especially for genetic diseases that are known to have genetic and phenotypic heterogeneity, when it may not be the most appropriate initial test. For example, Stargardt disease, a juvenile onset hereditary macular degeneration, is commonly caused by mutations in the *ABCA4* gene and targeted sequencing of this gene has led to mutation detection in more than half of affected patients (Briggs *et al.*, 2001; Xin *et al.*, 2015). On the other hand, for the heterogeneous movement disorders only 5% are diagnosed by Sanger sequencing (Neveling *et al.*, 2013).

Furthermore, single gene testing can take the form of methylation analysis for disorders caused by epigenetic mechanisms (e.g. imprinting disorders), fragment analysis for trinucleotide repeat expansions and multiple ligase probe amplification (MLPA) for small copy number variants (CNV) such as exon-level deletions and duplications (Schouten *et al.*, 2002; Poole *et al.*, 2013; Xue *et al.*, 2015; Liu *et al.*, 2017b).

1.2.2 CGH-array

CGH-array is now considered a first tier test for developmental disorders such as intellectual disability, autistic spectrum disorder and multiple congenital anomalies. The technique is designed to detect cytogenetic abnormalities such as aneuploidies and structural rearrangements. Variants identified through CGH-array often require scrutiny before they are associated with disease. Nevertheless, using this method has proposed many candidate disease genes in neurodevelopmental disorders. It is worth noting that although CGH-array interrogates the whole genome, it does not identify all types of mutations. Thus, a negative

test does not mean absence of genetic defects and should be followed by further testing of relevant loci by gene screening or NGS, for example (South *et al.*, 2013; Cappuccio *et al.*, 2016).

1.2.3 Gene panel testing

Gene panel testing allows multiple genes, previously associated with a specific genetically heterogeneous disorder to be sequenced simultaneously. Although multiple genes (tens to hundreds) are sequenced, this translates to a single test for the patient and a comprehensive and effective diagnostic tool for the clinical service (Xue *et al.*, 2015). NGS gene panels, although dependant on the patient's initial diagnosis and thus the choice of disease panel, have shown high diagnostic yields for many genetic diseases. In a cohort of patients with limb girdle muscular dystrophy (LGMD), a genetically and phenotypically heterogeneous neuromuscular disorder, an NGS-based panel of 23 LGMD related genes and 15 genes associated with phenotypically overlapping disorders resulted in a diagnostic rate of 33%. This rate would have been difficult, if not impossible, to achieve using a single gene testing approach within the same time and resource frames (Kuhn *et al.*, 2016). Similarly, in a group of 400 patients with non-specific early onset epilepsy and severe developmental delay, 18% of cases had a molecular diagnosis using a gene panel approach. This rate was higher at 39% for patients with very early onset epilepsy (< 2 months) (Trump *et al.*, 2016).

Furthermore, gene panel testing may provide further insights into the phenotype-genotype correlations and the phenotypic spectrum of the genetic disease under investigation. This was demonstrated for inherited peripheral neuropathies where a targeted gene panel revealed a high diagnostic yield (31%) and unexpected phenotype-genotype variability, where some patients were found to have mutations in genes not indicated by their phenotype (Antoniadi *et al.*, 2015; Trump *et al.*, 2016).

1.2.4 Whole exome sequencing

WES is increasingly being used in clinical diagnostics for rare genetic diseases. WES captures the majority of the coding sequence of the human genome. For genetically heterogeneous disorders and for those with diagnostic uncertainty WES has revealed high diagnostic yield when compared to single gene testing and NGS-panel approaches. A post-hoc study comparing WES with Sanger sequencing for a number of heterogeneous genetic diseases including: blindness, deafness, movement disorders, mitochondrial diseases and colorectal cancer found that when combined with targeted analysis of WES data, diagnostic yield was

considerably higher for all diseases, with the exception of colorectal cancer (Neveling *et al.*, 2013). In a preceding study, the authors also compared yield for retinitis pigmentosa using a targeted NGS-panel and WES and found that WES was still superior (36% and 52% diagnostic yield for targeted panel sequencing and WES, respectively) (Neveling *et al.*, 2012). WES has also contributed to identification of *de novo* and novel mutations in rare genetic disorders. For autistic spectrum disorders (ASD), WES has played a major role in identifying *de novo* causative and risk variants in sporadic and familial forms of the disease (Iossifov *et al.*, 2014; Lee *et al.*, 2014b; Toma *et al.*, 2014; Fukai *et al.*, 2015; Hara *et al.*, 2015). Likewise, *de novo* mutations play a major role in intellectual disability and WES has been found to be an effective diagnostic tool for their detection. In a series of 100 patients with intellectual disability, *de novo* mutations were detected in 53 patients and these mutations were in novel candidate genes in 22 patients in this series (de Ligt *et al.*, 2012). Furthermore, WES identified likely causative mutations for patients with LGMD in genes previously associated with other myopathies rather than LGMD, thus expanding the clinical phenotype for those genes, increasing the number of genes associated with LGMD and providing insights into disease mechanism and aetiology (Ghaoui *et al.*, 2015).

1.2.5 Whole genome sequencing

WGS is increasingly being used in research into rare genetic disease and is now entering clinical settings for molecular diagnosis of these disorders. For example, inherited retinal dystrophies are most commonly investigated using a targeted gene panel approach. However, a study comparing the yield of WGS compared to the targeted panel showed that WGS had the added benefit of detecting large deletions and non-coding deleterious variants and suggested that WGS can result in a 29% increase in diagnostic yield (Ellingford *et al.*, 2016). WGS has also been successful not only in detecting novel variants in new candidate genes but has also contributed in defining previously unreported Mendelian disorders (Lupski *et al.*, 2010; Sobreira *et al.*, 2010; Wang *et al.*, 2013a; Gilissen *et al.*, 2014). Although the majority of mutations identified through WGS are within the coding regions of the DNA, WGS potentially has the added benefit of detecting non-coding pathogenic mutations and is expected to be the preferred test for a comprehensive assessment for patients with a rare disease (Warman Chardon *et al.*, 2015).

1.2.6 Diagnostic algorithm

Xue et al suggest an algorithm for molecular diagnosis of Mendelian disease as shown in figure 1 (Xue *et al.*, 2015). A single gene test where indicated is associated with high

sensitivity and the chance of finding variants of unknown significance (VUS) is small, making interpretation of the results less challenging and making the overall test time and cost-efficient. This approach, however, requires the clinicians to have clinical knowledge and expertise enabling them to select the correct gene or genes to sequence. NGS-gene panel testing simplifies the process of test selection for the clinician when faced with a heterogeneous disorder and again will have a higher diagnostic sensitivity when appropriately selected. However, as the number of genes sequenced increases, the number of VUS will also increases posing a challenge in data interpretation and determining the correct causal variant. In addition, the choice of genes to include in a panel test for a particular disorder varies amongst laboratories, where gene selection depends on many criteria including strength of gene-disease association, cost and availability of expertise for interpretation. In addition, some laboratories choose to include genes associated with overlapping phenotypes and diseases included in the differential diagnosis. For example, a gene panel test for LGMD may strictly include only genes known to cause a form of LGMD or may be expanded in another laboratory to include genes associated with myopathies and other neuromuscular disorders. It is also worth noting that with the ongoing identification of new genes that are associated with NMD, gene panels are quickly becoming outdated (Warman Chardon *et al.*, 2015; Xue *et al.*, 2015).

For clinical presentations that are non-specific, WES is the preferred and the most commonly used test. Studies have shown that WES can also be cost-effective when compared to sequential sequencing of multiple single genes. In addition, WES has the advantage of being "hypothesis-free". Nevertheless, WES still requires a skilful clinical evaluation by experienced clinicians. In addition, this clinical information should be made available to the laboratory geneticist to enable interpretation of the vast number of variants produced by WES. A further point to consider is the fact that WES does not cover the entire exome, including some exons of known disease-associated genes. It has been reported that up to 10% of exons do not have adequate coverage at a 20-fold-depth. In addition, due to technical limitations for WES, large deletions, expansion and structural rearrangement maybe be missed. These along with the fact that a disease may not be genetic in origin, contribute to the WES average diagnostic yield of approximately 30% (Need *et al.*, 2012; Ankala *et al.*, 2015; Warman Chardon *et al.*, 2015; Xue *et al.*, 2015). WGS offers a more comprehensive method for diagnosis and disease gene discovery. It offers a more uniform coverage of coding and non-coding regions with their regulatory elements such as promoters, enhancers and extended splice sites. Cost and bioinformatics analysis as well as VUS interpretation all remain a challenge. However, decreasing costs and increasing development in informatics are

sufficient to brand WGS a comprehensive test for diagnostics in rare Mendelian diseases (Xuan *et al.*, 2013; Warman Chardon *et al.*, 2015; Xue *et al.*, 2015; Shiao, 2016).

Despite this thorough and comprehensive approach to testing, up to 50% of patients with a suspected genetic disorder remain without a confirmatory molecular diagnosis (Shashi *et al.*, 2014). For some specific presentations, for example, intellectual disability, the percentage of molecularly undiagnosed children is up to 80% (van Karnebeek *et al.*, 2005; Rauch *et al.*, 2006; Shashi *et al.*, 2014). The low diagnostic yield of this current approach can, in part, be accounted for by limitations in NGS technologies and analysis pipelines discussed in further detail in the following sections.

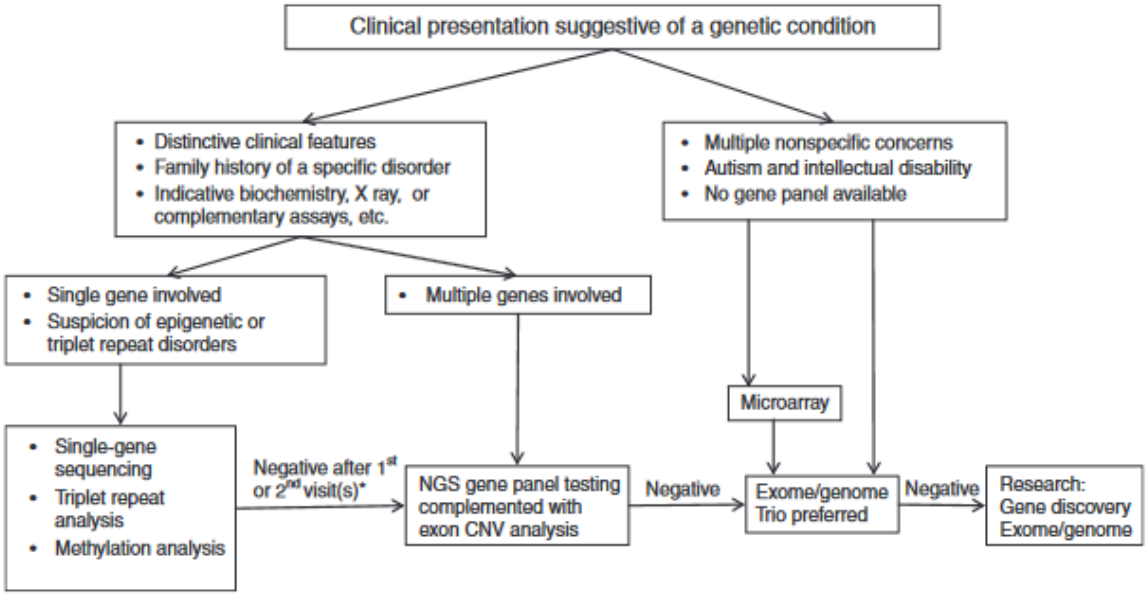


Figure1: Molecular genetics diagnostic algorithm for patients with a suspected genetic disorder (Xue *et al.*, 2015).

1.3 Next generation sequencing technologies

The Sanger method is considered the first generation in DNA sequencing and is still gold standard in diagnostics. NGS, initially introduced in 2005, refers to the range of newer technologies that have massively improved the throughput of DNA sequencing (Metzker, 2010). Technical aspects and system performance for the various NGS technologies since 2005 are discussed below (Xuan *et al.*, 2013; Pant *et al.*, 2014).

454 (Roche)

454 was the first NGS technology to be introduced. It is based on emulsion polymerase chain reactions (PCR), where beads carrying single stranded templates are confined to individual emulsion droplets and subjected to PCR amplification. These beads are then placed in wells where pyrosequencing takes place. In this latter process, sequencing by synthesis occurs where incorporation of a nucleotide leads to release of pyrophosphate causing a luminescence emission that is monitored in real time. This technology has the ability to produce the longest of read with the latest update claiming up to one kilobyte (kb). However, due to the lack of a terminating moiety, multiple incorporations of identical nucleotides can occur. This leads to problems when sequencing homopolymers; stretches of the same nucleotide. If three or more consecutive nucleotides are incorporated, the signal intensity does not necessarily correlate with the length of the homopolymer, leading to a high error rate in calling insertions and deletions. A further drawback of this system is its high cost when compared to other NGS technologies. The system is currently no longer supported by the developers.

Illumina (Solexa)

The Illumina sequencing system is the most widely used due to its high throughput and its cost effectiveness compared to other technologies. It uses an array-based DNA sequencing by synthesis method. Here a florescence is generated as nucleotides are incorporated, reversibly terminating the sequence reaction. This overcomes the issue of homopolymer sequencing. Its drawback however is that it generates short read lengths, 300 base pairs (bp), as the quality of the reads degrades after a relatively small number of sequencing. In addition, uneven coverage has been seen in AT-rich and GC-rich regions.

SoLiD (Applied Biosystems)

The SOLiD technology uses emulsion PCR as previously described. This is followed by ligation-based sequencing, where fluorescent octamer probes are ligated based on binding on complementary di-bases of the probe to those of the template. This di-base technology results in high accuracy in sequence calling. On the other hand, due to this unique technology, SOLiD is not as frequently used as many other technologies as the machine is not compatible with the most commonly used software for data analysis. In addition, SOLiD generates very short reads (up to 75bp).

Helicos (Biosciences)

Helicos was the first single-molecule sequencing system to be commercially introduced. Single-molecule sequencing technology meant that a single DNA or RNA molecule could be sequenced without prior amplification. This newer technology has the advantages of minimising sample handling and reducing amplification errors, making it potentially ideal for molecular diagnosis. The Helicos technology uses a base-by-base incorporation method and a florescence-labelled inhibitory moiety. It therefore overcomes the homopolymer problem, however, it is associated with a high error rate of ~3-5% mainly of deletions. This system did not receive much interest and was shut down in 2012.

Pacific Biosciences (PacBio)

The PacBio technology uses a "single molecule real time" (SMRT) system, where template-directed incorporation of fluorescence-labelled nucleotides is recorded in real time. PacBio also produces large read lengths (~8.5 kb). Despite its relatively high cost, it is becoming increasingly popular for applications that benefit from long reads.

Ion Torrent (Life Technologies)

Ion Torrent was invented by Jonathan Rothberg who also developed the earlier 454 system. Ion Torrent uses a similar system to 454 with the exception of the detection technology that relies on pH measurement, making it different from all other NGS technologies. It does not require optical sensor systems thus the costs are reduced. However, it can only reliably produce short reads making it more suitable for smaller targeted sequencing projects.

Oxford Nanopore Technologies

The nanopore sequencing technology is also a single molecule sequencing technology. The single-stranded molecule is passed through a nanometer-sized pore and sequence detection is by recording ionic current change or optical signal. This technology is currently being selectively tested using the MinION, a small USB attached sequencer that requires minimal sample preparation. However, it requires a disposable flow cell costing nearly $1000. The system still has a current high error rate. Nonetheless, the potential for long reads is promising.

1.4 NGS data flow and bioinformatics pipelines

Recent advances in NGS technology and the rapid drop in sequence time and cost has led to generation of vast amounts of data requiring interpretation through clinically targeted bioinformatics algorithms. In the context of clinical NGS, these algorithms can be divided into stages (figure 2) including: initial data processing, variant calling, variant annotation and variant prioritization (Bao *et al.*, 2014).

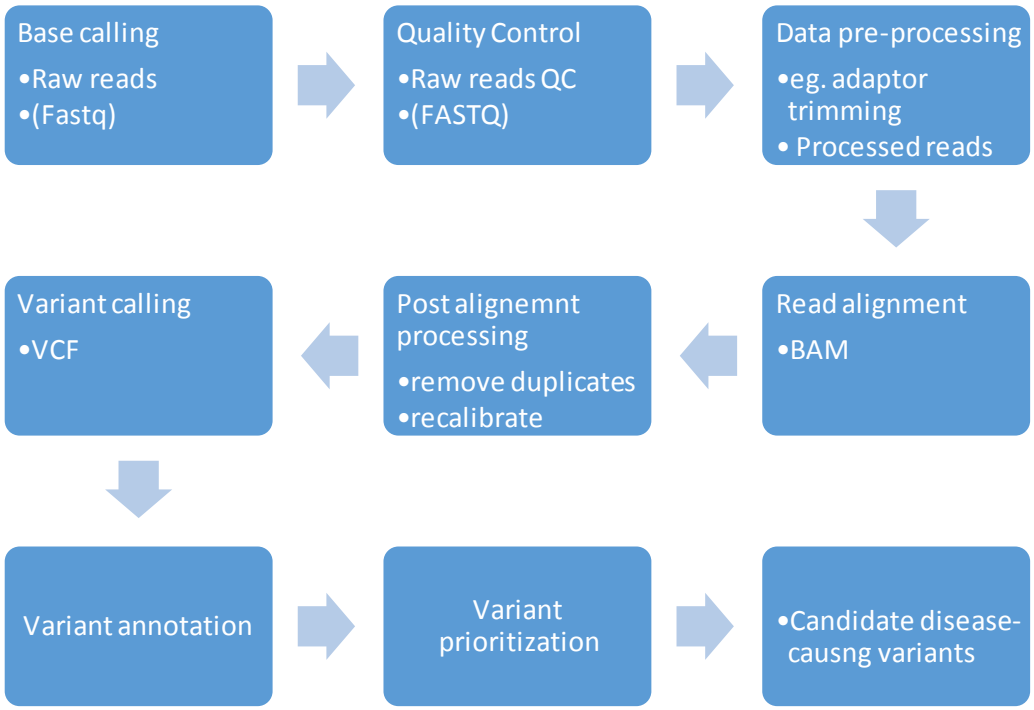


Figure 2: Clinical next generation sequencing analysis pipeline. BAM; binary alignment map, VCF; variant call format.

1.4.1 Base calling

The initial challenge in NGS is optimal base calling by the NGS platform. As previously discussed, NGS technologies vary in detection methods and operating algorithms and thus sensitivity and specificity. For example, InDels are more frequently erroneously called by the 454 platform, while substitution errors prevail at a rate of 1% in Illumina system and are dependent on read length. Platform-dependent base calling algorithms are constantly being developed to reduce error rates. Many use quality scores to estimate error probabilities for each base call. This can be expressed as a Phred score, where a Phred score of 20 indicates a 10^{-2} or 1% probability of an error in base calling. Base calling accuracy continues to be

challenging and requires further development in computational algorithms (Xuan *et al.*, 2013). Recent publications assessing accuracy of variant calling with multiple sequencing technologies have demonstrated inconsistencies and varying sources of error (Boland *et al.*, 2013; Zhang *et al.*, 2015; Laurie *et al.*, 2016).

Most NGS platforms will generate a Fastq file containing the raw data. This text file contains the sequence reads along with quality scores. Further quality control (QC) processing involves trimming low quality reads and contaminating bases, (e.g adaptor sequences) using software programs such as FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc).

1.4.2 Sequence read mapping

The next step is to map the reads in the Fastq file. This refers to the assembly of the reads to reconstruct the parent genome as accurately as possible. This can either be done via 1) alignment to a reference genome or 2) *de novo* assembly which uses no scaffold sequence to map reads. In *de novo* assembly, the reads group into contigs and contigs into scaffolds based on read overlap. The contigs give multiple sequence alignment of reads and the scaffold defines their order and orientation as well as gaps in the contigs. For *de novo* assembly software tools use either the "Overlap/Layout/Consensus" (OLC), "Greedy" or Bruijn graph algorithms all of which align reads based on pairwise overlapping sequences to reconstruct the parent genome (Miller *et al.*, 2010).

When aligning to a reference genome, the computational algorithm attempt to find the location of each short read in the NGS experiment in the reference genome. The most commonly used aligner, the Burrows-Wheeler Aligner (BWA), uses the Burrows Wheeler Transform (BWT). This rearranges and indexes repetitive reads together while still allowing the original reads to be searched. The final index can then be aligned to the reference genome (Huang *et al.*, 2013). Currently, there are three different BWA algorithms: BWA-ALN (BWA-backtrack), BWA-SW (Smith Waterman Alignment) and BWA-MEM (maximum exact matches). Each of these algorithms has its own strengths and weaknesses in terms of speed of performance, handling read length, ability to account for sequencing errors and ability to facilitate insertion and deletion (InDel) detection (Robinson *et al.*, 2017).

The aligned reads are then commonly stored as a binary alignment map (BAM) file format. Read mapping is a crucial step in the bioinformatics analysis pipeline as the subsequent

variant calling relies on the accuracy of mapping and mistakes in read mapping may lead to erroneously called variants.

The alignment based-approach is limited by incompleteness of genome assembly, interference from single nucleotide polymorphisms and base calling errors and structural variations in otherwise balanced genomes. In addition, alignment to a reference genome is difficult in regions of high diversity. All this makes variant calling on alignment-based reads more prone to false positives. In general, this may be overcome through use of longer paired-end reads (Nielsen *et al.*, 2011; Wu *et al.*, 2017).

De novo assembly methods allow for a comprehensive assembly of the sequenced genome irrespective of the reference genome. It may also offer a solution for mapping of known highly diverse regions. However, although *de novo* assembly has shown a small additional value in single nucleotide variant (SNV) discovery, the method is associated with a high rate of false positives for novel mutations and still requires further development (Nielsen *et al.*, 2011; Cao *et al.*, 2015; Wu *et al.*, 2017).

1.4.3 Post mapping data processing

Prior to variant calling, the sequenced and mapped reads require further quality control and filtering. This step varies between researchers and projects. However, for the majority of sequencing experiments, duplicate fragments are marked and removed. The most commonly used software application is Picard (https://broadinstitute.github.io/picard/) developed by the Broad Institute of Harvard and MIT. The tool also assesses mapping quality, allows for elimination of low quality reads and performs a local realignment. In addition to these steps, many projects perform a base quality score recalibration (BQSR) prior to variant calling. This step is aimed at detecting systematic error by the sequencing platform in assigning the initial base quality score (Pirooznia *et al.*, 2014).

1.4.4 Variant calling

Once reads are assembled and aligned to the reference, variant calling can then take place looking for SNV or short InDels that do not match the reference and thus may be disease causing. Commonly used variant calling software are SAMtools, the Genome Analysis Tool Kit (GATK) and the Short Oligonucleotide Alignment Program (SOAP) (Wang *et al.*, 2013b; Pirooznia *et al.*, 2014). These tools use different calling algorithms that can fall into one of three categories depending on how base calls are deducted from the high throughput sequencing data. The first algorithm, individual-based single marker caller (IBC), assigns

genotypes for a single individual at a single position. This is usually used for high-depth exome sequencing data. The second algorithm, population-based single marker callers (PBC), utilizes reads at a particular position from the whole sample to determine allele frequencies and polymorphisms. Genotypes are then called for each call for an individual based on the calculated allele frequencies. The third group of callers use linkage disequilibrium-aware calling algorithms (LDC). Here, the algorithm uses linkage disequilibrium data flanking each variant identified by IBC or PBC by several hundred kilobases, then phases variant calls into haplotypes. This information is then used in the algorithm to update genotypes. This latter method, although computationally demanding, has been used in combination with PBC in the 1000 Genomes Project (https://www.genomicsengland.co.uk/the-100000-genomes-project/) to interpret low coverage, genome-wide data. These algorithms then produce a variant call format (VCF) file (Abecasis *et al.*, 2010; Nielsen *et al.*, 2011; Lo *et al.*, 2015).

1.4.5 Variant annotation

The increasing use of NGS in research and in clinical setting means that a large amount of genetic variants are recognized and need to be assigned as disease causing or polymorphism. Initially, the variant type needs to be attributed and its location in the genome defined, depending on the version of the Human Genome Assembly (currently GRCh37 or GRCh38). RefGene (http://varianttools.sourceforge.net/Annotation/RefGene/) and (Ensembl http://www.ensembl.org) are the current gold standards for locating and categorizing variants. This is a crucial step in determining whether the variant has a potential to disrupt the protein sequence and thus alter its function.

NGS data analyses also require integration from other genomic projects. For example, The Encyclopedia of DNA Elements (ENCODE) project and the Roadmap Epigenomics Project (http://www.roadmapepigenomics.org/) provide additional genome-wide functional and regulatory element data. Also, RNA sequencing projects such as the one included in the RefSeq repository (https://www.ncbi.nlm.nih.gov/refseq/about/) are continuously adding to the list of new splice events and novel transcripts (McCarthy *et al.*, 2014; Li and Wang, 2015; Salgado *et al.*, 2016). This data can all be used to annotate variants.

It is important to note here that the choice of transcript set is an important step to consider for the variant annotation process. Currently, for human NGS annotation, the transcript sets offered by RefSeq and Ensembl are most commonly used (McCarthy *et al.*, 2014; Laurie *et al.*, 2016).

A further annotation element that is particularly relevant in the context of rare Mendelian diseases is the frequency of the variant in the general population. For this purpose, data from large-scale population-based projects such as the 1000 Genomes Project (http://www.internationalgenome.org/), the Exome Aggregation Consortium (Lek *et al.*, 2016) (ExAC) database (http://exac.broadinstitute.org/) and the Genome Aggregation Database (gnomAD) (http://gnomad.broadinstitute.org/) are utilised.

The next level of annotation is for variant predicted pathogenicity. This involves a prediction software tool such as SIFT, Polyphen2, Mutation Taster, CADD, UMD-Predictor and FATHMM. Many bioinformatics pipelines use a combination of these predictors. Variants are also annotated for association with disease. This information is usually extracted from databases such as ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), UNIPROT (http://www.uniprot.org/) and the Human Gene Mutation Database (HGMD, http://www.hgmd.cf.ac.uk/ac/index.php) (Li and Wang, 2015; Salgado *et al.*, 2016).

At the gene level, variants can be annotated for functional gene attributes, by using the Gene Ontology Consortium database (Ashburner *et al.*, 2000) for example. Genes can also be annotated for tissue expression levels through integrating projects and databases such as the Genotype Tissue Expression (GTEx) resource (https://gtexportal.org/home/) and the European Bioinformatics Institute's (EBI) Expression Atlas (https://www.ebi.ac.uk/gxa/home), which enables queries across numerous participating gene expression experiments. Databases for biological pathways, protein interactions and protein sequence features, such as KEGG Pathways (http://www.genome.jp/kegg/pathway.html), EBI IntAct (https://www.ebi.ac.uk/intact/), and EBI InterPro (https://www.ebi.ac.uk/interpro/), respectively, can also be used to annotate at the gene level. However, currently these are not integrated in annotation tools and are provided through external links (Li and Wang, 2015; Salgado *et al.*, 2016).

Variant annotation requires a complex algorithm and integration of a comprehensive combination of the above-mentioned resources. The most commonly used variant annotation tool is ANNOVAR (http://annovar.openbioinformatics.org/en/latest/) which integrates over 4000 public databases (Wang *et al.*, 2010; Bao *et al.*, 2014). Other commonly used annotation software are the Ensembl Variant Effect Predictor (VEP) (McLaren *et al.*, 2016), and SnpEff (Cingolani *et al.*, 2012).

1.4.6 Variant prioritization

For a WES experiment, an individual sample will have more than 30,000 variants that are different to the reference sequence and more than 10,000 of those are predicted to be non-synonymous SNVs, to cause a splice site alteration or to be small insertions or deletions (Gilissen *et al.*, 2012). For a WGS experiment the number of variants called is approximately 3 million (Wang and Xing, 2013). Following annotation, these variants need to be filtered to enable an effective search for disease-causing mutations. This step is required to prioritize and shortlist a smaller number of the most relevant variants. For this purpose, a number of variant prioritization platforms and interfaces have been developed. These tools allow filtration of variants based on the annotations attached to them in the previous step of the analysis pipeline. In the context of rare disease, variants are filtered for their population frequency, effect on protein, pathogenicity predictions, and inheritance model. Furthermore, variants from a WES and WGS experiment can be filtered and restricted to those falling within particular regions in the genome identified through linkage studies or in a candidate gene list known to be associated with a particular disease spectrum. More recently developed software allow for filtering variants based on protein interaction networks, cross-species phenotype comparisons and clinical relevance, where the relevance of the gene harbouring the variant is added to the filtering criteria of the variant prioritization algorithm (Franke *et al.*, 2006; Zemojtel *et al.*, 2014; Smedley *et al.*, 2015; Muller *et al.*, 2017). To enable this, data needs to be added in a standardized language using ontologies such as the Human Phenotype Ontology (Kohler *et al.*, 2014) (HPO) and the Online Mendelian Inheritance in Man (OMIM) database (https://omim.org/). Taking this a step further, variant prioritization can also incorporate genetic and phenotypic data shared across projects in the process of matchmaking. This is of particular importance in rare disease and for novel discoveries, where data sharing allows researchers to confirm their suspicion of a novel gene association by finding a phenotypically similar case with variants in the same gene of interest at another research site (Kirkpatrick *et al.*, 2015; Philippakis *et al.*, 2015).

Variant prioritization requires a standalone, user-friendly, interactive and flexible platform, which facilitates application of custom filters. Commercially and academically available web applications are continuously being developed and released and all share common features but also have specific strengths (Jalali Sefid Dashti and Gamieldien, 2017). For example, as well as filtering for functional variants, allele frequency and pathogenicity predictions, Exomiser allows variant prioritization by integrating cross species phenotype comparisons (Smedley *et al.*, 2015). The RD-Connect Genome-Phenome Analysis Platform (https://platform.rd-

connect.eu/) integrates Exomiser as well as variant/phenotype matchmaking tools, to further facilitate short listing of variants and genes (Thompson *et al.*, 2014). The platform now also utilizes haplotype data for analysis of large pedigrees and homogenous cohorts. PhenIX (Phenotypic Interpretation of Exomes) uses terms from the HPO to prioritize variants in genes with a known association to phenotypes similar to the individual under investigation (Zemojtel *et al.*, 2014). *seqr* (https://seqr.broadinstitute.org/) allows users to visualize the reads and ascertain the variant calls through integrating the Integrative Genomic Viewer (IGV) software (http://software.broadinstitute.org/software/igv/).

Researchers will have different preferences of variant prioritization methods and this relies on the type of NGS project. In addition, while some larger scale projects would benefit from automated workflows, smaller ones may require a more interactive filtering process (Wang and Xing, 2013; Jalali Sefid Dashti and Gamieldien, 2017).

Once candidate variants are prioritized, they then need to be validated by phenotypic data, Sanger sequencing, segregation analysis and functional studies (Coonrod *et al.*, 2011).

1.5 Benchmarking NGS bioinformatics pipelines

As NGS is increasingly being adopted in clinics, it is crucial to validate bioinformatics pipelines to ensure high sensitivity; the ability to correctly identify sites where the patient's DNA differs from the reference (true positives), and high specificity; the ability to correctly identify sites that match the reference (true negatives) (Zook *et al.*, 2014).
There is a lack of published guidelines on how to benchmark NGS bioinformatics pipelines and as a result, validation methods vary between researchers. Some labs will validate their pipeline results by selecting a number of sites for confirmation via Sanger sequencing (Beck *et al.*, 2016; Gao *et al.*, 2016), while others use SNP-array data on the same samples and assess concordance (Cottrell *et al.*, 2014). Not only do both methods carry a degree of selection bias, the lack of uniformity in validation methods makes it difficult to compare results across experiments, and thus identify inaccuracies and judge discrepancies. A more recent trend follows the use of well characterized reference data. For example, the National Institute of Standards and Technology (NIST) and the Genome in a Bottle (GIAB) consortium have created reference material for genome sequencing as well as guidance on benchmarking. The material is characterized to outstanding levels and publicly available for researchers to use when validating their sequencing technologies, bioinformatics pipelines and variant

detection methods. It includes consensus data from a number of sequencing technologies, library preparation methods and bioinformatics pipelines (Zook *et al.*, 2014; Zook *et al.*, 2016). These initiatives not only allow WES and WGS experiment benchmarking, they also enable comparison of experiments against a well-defined reference. Recent literature demonstrate the usefulness of the GIAB reference material in designing next generation sequencing experiments and comparing reliability of the many bioinformatics tools in read mapping and variant calling (Zook *et al.*, 2014; Cornish and Guda, 2015; Laurie *et al.*, 2016). The benchmarking materials have also facilitated the identification of regions in the genome where SNV and InDel calls are less reliable. This meant that it was possible to publish reliable genomic regions for benchmarking, where variant calls are expected to be highly consistent (Zook *et al.*, 2014; Laurie *et al.*, 2016).

Very recently, the Association for Molecular Pathology and the College of American Pathologists published a set of guidelines for validating NGS bioinformatics pipelines. Amongst the 17 recommendations mentioned in table 2, there is great emphasis on the use of a representative reference set of high quality variants produced by an orthogonal method (Roy *et al.*, 2018).

Table 2: Recommendation from the Association for Molecular Pathology and the College of American Pathologists for validating NGS bioinformatics pipelines (Roy *et al.*, 2018).

Recommendation number	Statement
1	Clinical laboratories offering NGS-based testing should perform their own validation of the bioinformatics pipeline
2	A qualified medical professional with appropriate training in NGS interpretation and certification must oversee and be involved in the validation process
3	Validation must be performed only after completion of design, development, optimization, and familiarization of the bioinformatics pipeline and its components
4	Bioinformatics pipeline validation should closely emulate the real-world environment of the laboratory in which the test is performed
5	Validation should include all individual components of the bioinformatics pipeline used in the analysis, and each component must be reviewed and approved by an appropriately qualified medical molecular professional and the laboratory director
6	The design and implementation of the bioinformatics pipeline must ensure the security of identifiable patient information and be compliant with all applicable laws at the local, state, and national levels
7	Validation of the NGS bioinformatics pipeline must be appropriate and applicable for the intended clinical use, specimen, and variant types detected of the NGS test
8	Laboratories must ensure that the design, implementation, and validation of the bioinformatics pipeline are compliant with applicable laboratory accreditation standards and regulations
9	The bioinformatics pipeline is part of the test procedure, and its components and processes must be documented according to laboratory accreditation standards and regulations
10	The identity of the sample must be preserved throughout each step of the NGS bioinformatics pipeline with a minimum of four unique identifiers, including a

Recommendation number	Statement
	unique location identifier within the content of each data file read and/or generated by the pipeline
11	Specific quality control and quality assurance parameters must be evaluated during validation and used to determine satisfactory performance of the bioinformatics pipeline
12	The methods used to alter or filter sequence reads at any point in the bioinformatics pipeline before interpretation must be validated to ensure that the data presented for interpretation accurately and reproducibly represent the sequence in the specimen, and full documentation of these methods must be kept as part of the test documentation according to laboratory accreditation standards and regulations
13	Laboratories must include specific measures to ensure that each data file generated in the bioinformatics pipeline maintains its integrity and provides alerts for or prevents the use of data files that have been altered in an unauthorized or unintended manner
14	*In silico* validation can be used to supplement the validation of the bioinformatics pipeline but shall not be used in lieu of end-to-end validation of the bioinformatics pipelines using human samples
15	Validation of the bioinformatics pipeline must include confirmation of a representative set of variants with high-quality independent data; appropriate validation metrics by variant type should be reported
16	Clinical laboratories must ensure the accuracy of software-generated HGVS variant nomenclature and annotations and have an alert in place to indicate when the software-generated nomenclature and annotations need to be manually reviewed and/or corrected, and documentation of any corrections must be maintained
17	Supplemental validation is required whenever a significant change is made to any component of the bioinformatics pipeline

HGVS; Human Genome Variation Society, NGS; next generation sequencing.

1.6 Advantages and drawbacks of NGS

NGS provides high throughput data in a single reaction. It has the advantage of relatively uniform coverage of larger proportions of the genome and in the case of WGS, uniform coverage across the whole genome. NGS allows hypothesis-free experiments and has facilitated novel disease gene discoveries. For example, loci for three autosomal dominant limb girdle muscular dystrophies had been identified through linkage studies for decades prior to WES identifying mutations in *DNAJB6*, *TNPO3* and *HNRPDL* as disease causing for LGMD1D, LGMD1F and LGMD1G, respectively (Sarparanta *et al.*, 2012; Torella *et al.*, 2013; Vieira *et al.*, 2014). In addition, it is the technique of choice in heterogeneous disorders (Lek and MacArthur, 2014) and is particularly useful for identifying variants in larger genes such as *TTN* (Toro *et al.*, 2013; Liu *et al.*, 2017a).
Nonetheless, NGS is a continuously developing entity and many points are to be considered when designing NGS experiments. First, despite the continuing decline in cost of sequencing, the costs remain high and vary depending on the setting, the technology and the analysis strategy (Sims *et al.*, 2014). The aim is thus to design a sequencing experiment that provides reliable results at the lowest costs possible.

The large volume of data is another drawback for analysis and storage. It has been suggested that with the decreasing cost of sequencing, it may be more appropriate to re-sequence a patient's DNA than to store the original sequencing data (Efthymiou *et al.*, 2016). However, the development of compressed forms of NGS data files, such as gVCF and CRAM, may potentially reduce storage costs (Lek and MacArthur, 2014).

A further key consideration in NGS experiment design is coverage. Coverage has been used to define the "depth of coverage" which is the number and length of high quality reads from an NGS experiment that represent each base in the reference genome. Coverage also refers to the "breadth of coverage" of a target genome, which represents the percentage of bases in the target that are sequenced a given number of times, and have a given minimum depth. Depth and coverage are terms that are often used interchangeably, as they are in this account. The higher the coverage, the higher the experiment cost. Sanger sequencing remains the gold standard and the most commonly used genetic test. It has high coverage for the target segment of the genome being sequenced as Sanger sequencing has longer reads that are derived from larger insert libraries and can be assembled using reliable computational algorithms.
However, from the genome point of view, Sanger sequencing is not a practical test and may not be cost effective in rare diseases with phenotypic and genetic heterogeneity. WES

provides a more comprehensive coverage of 1-2% of the genome where an estimated 85% of disease causing mutations are located. However, it has been reported that up to 10% of the exome in WES experiments is not adequately covered (has a depth of <20) (Rabbani *et al.*, 2014). The power to detect InDels in particular is directly related to uniformity of coverage. WGS has higher breadth of coverage compared to WES, in that it provides a more uniform coverage of coding and non-coding parts of the genome. However, WGS is still unreliable in detecting structural rearrangement and repeat expansions that are longer than the read length.

Technical limitations in NGS may result in inaccuracies. For example, capture kits are liable to reference bias. This occurs for variants in the heterozygous state, where the capture probes that match the reference sequence tend to enrich the reference allele, resulting in a false negative call (Abnizova *et al.*, 2017). In addition, at GC-rich sites and sites with repetitive elements, PCR amplification may result in areas with poor coverage, duplicate reads, off-target capture, and thus a decrease in uniformity of coverage. These technical limitations of NGS lead to lower sensitivity and thus false negative calls, and need to be accounted for (Treangen and Salzberg, 2011; Abnizova *et al.*, 2017; Roeh *et al.*, 2017).

False positives also need to be considered. Template amplification and sequencing can lead to false variant calls. To overcome this, Sanger sequencing can be used to confirm variants that are of interest. In addition, methods for variant quality assurance (for example, joint variant calling and variant quality score recalibration, VQSR) and filtering should be applied to the pipeline (Carson *et al.*, 2014; Efthymiou *et al.*, 2016).

A PCR-free WGS experiment will provide more uniform coverage of the genome (breadth) and thus require a smaller average depth to reliably cover the target sequence. Nonetheless, WES remains cheaper than WGS and despite the limitations discussed here, can achieve a near similar breadth of coverage for the coding genome. This makes WES more popular than WGS as reduced costs of the experiment allows researchers and clinical diagnostics services to include more samples (Sims *et al.*, 2014; Li *et al.*, 2015; Patwardhan *et al.*, 2015; Warman Chardon *et al.*, 2015; Evila *et al.*, 2016).

A further crucial point to consider in NGS data analysis is choice of bioinformatics tools. It has been suggested that the two most important differentiating steps in NGS analysis are alignment of reads to the reference genome and variant calling (Zook *et al.*, 2014; Cornish and Guda, 2015; Hofmann *et al.*, 2017). Many studies have been carried out in an attempt to identify the best combination of computational tools. A 2015 paper by Cornish *et al* used the reference GIAB data to compare performance of alignment tools and variant callers. The

authors used six aligners and five variant callers (figure 3) resulting in 30 combinations of pipelines. They found that Novoalign as the alignment tool and the GATK UnifiedGenotyper as the caller provided the best sensitivity. From the six alignment tools the Borrows-wheeler Aligner (BWA) had the best performance with all variant callers but results still varied depending on the variant caller in the pipeline. However, the GATK UnifiedGenotyper had sensitive results regardless of the aligner. The authors also noted that sensitivity for InDels was considerably lower and posed a difficulty for all pipelines (Cornish and Guda, 2015).

A further study compared bioinformatics pipelines on WES and WGS samples focusing on InDels only. They compared GATK's Unifiedgenotyper, HaplotypeCaller and Pindel, and found that for short indels with high read depths, validation data supported HaplotypeCaller. They suggested that Pindel was best suited for larger InDels at smaller read depths (Ghoneim *et al.*, 2014). A more recent study by Laurie et al in 2016, also used the GIAB reference genome to assess robustness of bioinformatics pipelines and found that for 70% of the human genome results of the various pipelines were consistent. However, this consistency was more so for SNV over InDels (Laurie *et al.*, 2016). The results of these studies highlight the need for continued development and validation of bioinformatics tools and pipelines in NGS data analysis.

Copy number variants (CNV) are an additional challenge for NGS experiments and the gold standard for their detection is still array-CGH (Wenric *et al.*, 2017). However, a number of bioinformatics tools have recently been developed and integrated into NGS analysis pipelines. Although such integration is becoming more and more accurate at mapping and detecting CNV, validation of many of these software tools has shown considerable variation and the algorithms require further development (D'Aurizio *et al.*, 2016; Nam *et al.*, 2016; Onsongo *et al.*, 2016; Tan *et al.*, 2017; Wenric *et al.*, 2017).

These technical issues are currently being addressed by combining NGS technologies with other methods, by increasing sequencing coverage and read lengths and improving bioinformatics software (Efthymiou *et al.*, 2016).

VUS are also a major issue for NGS analysis and distinguishing true pathogenic mutations from polymorphic or benign variants with no clinical significance is problematic. In addition, the "narrative potential" with the researcher building a plausible but false story about functional variants in the genome is also a problem (Lek and MacArthur, 2014). Both these issues are continuously being addressed through improved variant filtering and annotation methods, as well as integrating up-to-date human variation and phenotypic knowledge, and

following the most recent published guidelines on associating genetic variants with disease (MacArthur *et al.*, 2014).

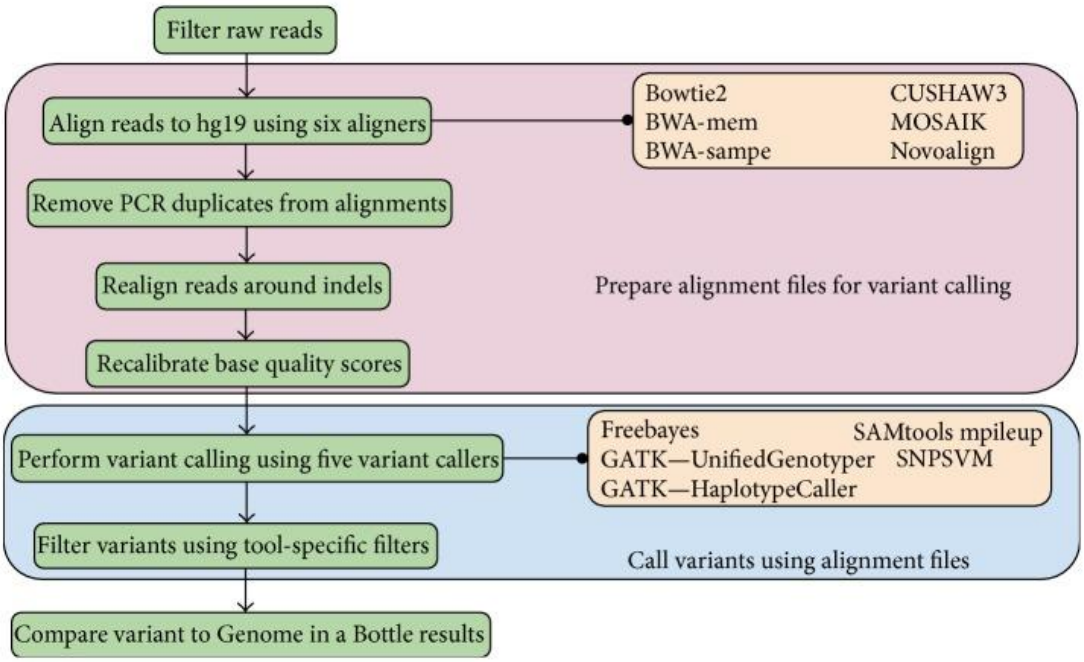


Figure 3: Bioinformatics analysis pipeline used in the Cornish et al study comparing six alignment tools and five variant callers using Genome in a Bottle as the reference (Cornish and Guda, 2015).

1.7 Data sharing

Sharing data is critical to scientific advancement. However, it has been hindered by traditional scientific practices. The Personal Genome Project (http://www.personalgenomes.org/) is an international collaboration led by Harvard Medical School since 2005. It aims to sequence the genomes of 100, 000 individuals from the USA, UK, Canada and Austria and to make genetic and phenotypic data publically available for use in scientific research. In doing so, the project aims at understanding how human genetic variation influences health and disease. In addition, biological samples from participants are stored in biobanks for potential uses in research. This would allow clinicians and scientists internationally access to large amounts of data and samples (Zarate *et al.*, 2016).

In rare disease research, the concept of data sharing becomes even more critical, when sharing of high quality genotype and phenotype data may lead to an improved analysis of genomic

data from cohorts with rare and heterogeneous disorders. Large amounts of patient data will result in a more accurate variant annotation, analysis and interpretation. Sharing data relies on active participation of clinicians and scientists and on the involvement of patients and public in the process.

A few examples of collaborative data sharing projects are well established while many others are in their early stages. Through genomic data sharing, the National Institute of Health's (NIH) National Cancer Institute (http://www.cancer.gov/) have used genomic data contributed through various collaborative projects in translational research and therapy development. The Mitochondrial Disease Sequence Data Resource (MSeqDR, https://mseqdr.org/), established by the United Mitochondrial Disease Foundation, provides a robust data resource. MSeqDR is a web-based, user-friendly portal allowing researchers to share and access sequence data from individuals and families with a suspected mitochondrial disease. Further integration with online tools and databases facilitates further interrogation of the shared phenotypic and sequence data (Falk *et al.*, 2015).

More recent data sharing projects in rare disease include The Beacon Project and The Matchmaker Exchange (MME) project. The Beacon Project (http://ga4gh.org/#/beacon) is an initiative to test the willingness of international sites to share genetic data. It is a simple web service that allows users to query whether or not other users have any genomes with a particular nucleotide at a particular position on the chromosome of interest. The MME (http://www.matchmakerexchange.org) recognises that a proportion of the undiagnosed cases of rare genetic diseases have a VUS in a novel gene. Finding an additional patient with a similar phenotype and a variant in the same gene would increase evidence for causation. The MME enables this through data sharing and linking a large number of projects, institutes and databases for rare disease as shown in figure 4 (Thompson *et al.*, 2014; Gonzalez *et al.*, 2015; Philippakis *et al.*, 2015).

The vast amounts of data produced by global research necessitates development of means to pool and integrate phenotypic, genomic and biological sample data together. This will enable interpretation of the integrated data, not only for disease gene discovery, but also for identifying Mendelian disease modifiers and for selection of patient cohorts for therapeutic trials (Thompson *et al.*, 2014).

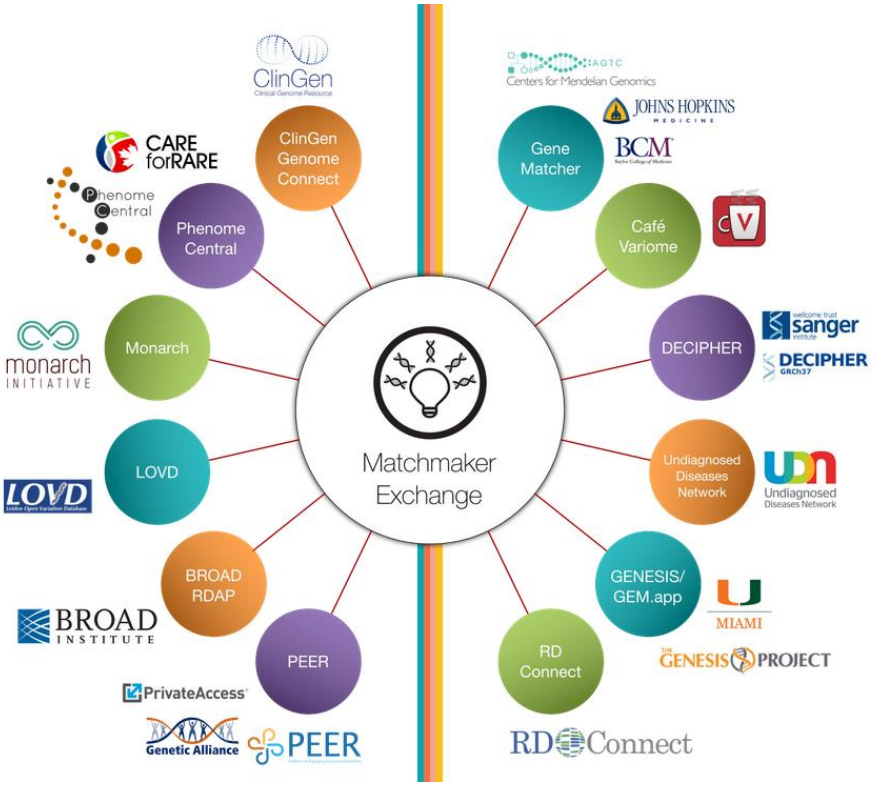


Figure 4: The Matchmaker Exchange participants and collaborators.
(http://www.matchmakerexchange.org**)**

1.8 Ethical considerations

NGS technology is continuously ascertaining its benefit in rare disease research and diagnosis. However, application of this advancing technology needs to be regulated through standard policies and ethical constraints in order to maximise the benefit of the data produced to patients and the public while protecting rights of privacy and data ownership of patients. In rare diseases, the clinical benefit of NGS may not be immediately obvious to the patient as its primary focus is on diagnosis rather than treatment of a rare disorder. Likewise, from the public's perspective it may not prove its utility and cost effectiveness in diagnosing a disease affecting a small number of individuals with limited treatment options. In addition, these treatment options where available are expensive and issues may be raised within governments on whether to fund therapies benefiting larger populations or those for a small number of patients with a rare disease (Kang, 2013; Curnutte *et al.*, 2016).

The most apparent ethical concern with regards to NGS in rare diseases is the uncertainty created by the large amounts of data that needs to be interpreted and associated with disease. VUS are the most challenging in terms of deciding what to report to patients and families.

Unexpected incidental findings not related to the disease under investigation are a particular challenge. A recent survey on the views of healthcare professionals in the UK on disclosure of prenatal genetic testing data found that the majority thought that VUS should not be reported to patients. However, the majority of participating health care professionals also agreed that reporting incidental findings associated with adult-onset disease risk is justified. Participants also supported involvement of parents in deciding what information to disclose (Shkedi-Rafid *et al.*, 2016). These views may also be relevant in the context of NGS testing for rare diseases. The American College for Medical Genetics and Genomics (ACMG) recommendations on reporting of incidental findings advise reporting of variants in the published consensus gene list. The ACMG also recommend that the matter should be an ongoing process for discussion and updating (Green *et al.*, 2013).

A further issue posed by NGS data for some rare disease patients is the issue of discrimination based on genetic data. Insurance companies may request and use genetic data on individuals before issuing their policies. In addition, patients may miss employment and educational opportunities. Sharing of genomic data should therefore be regulated using guidelines put together by policy makers in conjunction with clinicians, geneticists and patients (Desai and Jere, 2012; Fiore and Goodman, 2016).

A recent survey of patients with rare diseases found that data sharing was a main issue and that patients were particularly concerned about security of their data and its misuse (McCormack *et al.*, 2016). It is critical that genomic data is secure with restricted access and data sharing is regulated through governance frameworks that involve patients and their views in policymaking (Thompson *et al.*, 2014; Falk *et al.*, 2015).

All the above illustrate the importance of a comprehensive informed consent procedure prior to performing an NGS-based test. In addition, if patients' data and samples are to be stored and further used or shared then they should be informed and the purpose fully explained (Pelissier *et al.*, 2016).

1.9 NGS in rare neuromuscular disorders

Neuromuscular diseases (NMD) are a group of genetically and phenotypically heterogeneous disorders that arise from defects in the motor neuron, the neuromuscular junction or the muscle itself. The World Muscle Society (WMS, https://www.worldmusclesociety.org/)

categorizes NMD into 16 disease groups as shown in box 1. They acknowledge the difficulty of maintaining and updating the list of diseases and their associated genes in a printed format because of genetic and phenotypic heterogeneity. These are rather published online as the "muscle gene table" (http://www.musclegenetable.fr/). The heterogeneity, individual rarity, overlapping clinical presentations, varying age of onset and atypical presentations means that molecular diagnosis for NMD is, in many cases, challenging (Warman Chardon *et al.*, 2015). As in other rare disease groups, NGS has played a major role in diagnosis, gene discovery, disease pathophysiology and developing therapies (Bauche *et al.*, 2016; O'Connor *et al.*, 2016; Di Gioia *et al.*, 2017; Harris *et al.*, 2017; Johnson *et al.*, 2017). NMD NGS panels have demonstrated higher diagnostic yields when compared to sequential single gene testing, 46% and 19%, respectively (Ankala *et al.*, 2015). WES on a cohort of patients with limb girdle muscular dystrophy (LGMD) resulted in molecular diagnosis in 45% of patients. This rate was higher at 60% for trios. This study also identified mutations in genes associated with other muscle disorders such as myofibrillar myopathies (MFM) and congenital myasthenic syndromes (CMS), highlighting the phenotypic overlap and the genetic heterogeneity of NMD (Ghaoui *et al.*, 2015). A further WES study on a UK patient cohort with LGMD was able to achieve a molecular diagnosis in 37%, including identification of a novel gene. With regards to mutations in known NMD genes, the authors identify reasons as to why patients were not diagnosed via single gene testing as: atypical phenotypes, reassignment of pathogenicity of variants and somatic mosaicism (Harris *et al.*, 2017).

WGS has also played a role in novel gene discovery and in identifying functional InDels in NMD (Wang *et al.*, 2013a; Brewer *et al.*, 2016). And with large scale genomic projects such as the 100,000 Genomes (https://www.genomicsengland.co.uk/the-100000-genomes-project/), many patients with rare NMD are being sequenced along with patients with other rare diseases. As clinical and genomic data from these genomes is being interpreted, many new genetic associations are likely to be revealed (Efthymiou *et al.*, 2016).

Nonetheless, many patients with rare NMD remain undiagnosed. This, in part, is due to technical limitations in NGS such as coverage and limitations in detecting certain mutation types (such as trinucleotide repeat expansions and CNV), and in part due to limitations in analysis and interpretation of the data. These issues can be addressed by combining NGS with other mutation detection methods such as complementary Sanger sequencing and CGH-array, mutation validation methods such as family segregation and animal model studies, and an integrated diagnostic pathway, which includes clinical, population, genetic and matchmaking data (Biancalana and Laporte, 2015).

Box 1: NMD groups as categorised in the muscle gene table by the World Muscle Society (http://www.musclegenetable.fr/).

1. Muscular dystrophies
2. Congenital muscular dystrophies
3. Congenital myopathies
4. Distal myopathies
5. Other myopathies
6. Myotonic Syndromes
7. Ion channel muscle disease
8. Malignant hyperthermia
9. Metabolic myopathies
10. Hereditary cardiomyopathies
11. Congenital myasthenic syndromes
12. Motor neuron disease
13. Hereditary ataxias
14. Hereditary motor and sensory neuropathies
15. Hereditary paraplegias
16. Other neuromuscular disorders

1.10 Thesis aims and objectives

- Compare WES and WGS data for patients with NMD to assess the limitations of WES and the added yield of WGS when examining NMD genes.
- Compare three genomic platforms namely: the RD-Connect Genome-Phenome Analysis Platform, *seqr*, the Clinical Sequence Analyser (CSA), and their respective bioinformatics pipelines using WES and WGS data from patients with NMD.
- Demonstrate the utility of using an integrated genomics platform in diagnosing a cohort of patients with rare NMD in a research setting.
- Describe the genetic, demographic and clinical aspects for patients with GNE myopathy from Kuwait.

Chapter 2. WES and WGS comparison

2.1 Introduction

Rare disease research focuses on cost effective projects that will have the most impact on patient care and disease outcomes. This includes determining the molecular basis of rare genetic diseases. In theory, current genomics technology has the capability to identify all causes of genetic disease. However, the majority remain undiagnosed. In part, this is due to the rarity and heterogeneity of these disorders, the lack of a systematic coordinated international system to coordinate novel discoveries, and in part, due to technical limitations in sequencing technologies and analysis pipelines (Lupski *et al.*, 2010; Majewski and Rosenblatt, 2012; Toscano *et al.*, 2017; Volk and Kubisch, 2017).

With the decreasing costs of NGS, WGS has become feasible in many setting in research and increasingly in diagnostics. An obvious additional benefit for WGS over WES is that the former is powered to identifying potentially relevant variants in non-coding regions of the genome. Although WGS flourished identification of non-coding mutations in cancer genetics (Araya *et al.*, 2016; Khurana *et al.*, 2016; Gan *et al.*, 2018), this has not been as impressive for Mendelian disorders and the number of non-coding mutations identified using WGS has been limited (Protas *et al.*, 2017; Liskova *et al.*, 2018). In part, this may be due to the difficulty in interpreting the significance of non-coding variants and the need for their validation through laboratory-based methods (Biancalana and Laporte, 2015). Nonetheless, non-coding regulatory and splice-site mutations are known to be implicated in human disease and WGS will eventually identify many more (Warman Chardon *et al.*, 2015).

WGS is also expected to have higher sensitivity for InDel and CNV detection aided by the longer reads, the relatively uniform coverage and a PCR-free experiment (Fang *et al.*, 2014; Meienberg *et al.*, 2016; Trost *et al.*, 2018). However, with improved bioinformatics performance and read alignment methods, algorithm intersections and incorporation of CNV detection software into WES analysis pipelines, InDel and CNV detection rates are claimed to be comparable with WGS (D'Aurizio *et al.*, 2016; Rennert *et al.*, 2016; Kim *et al.*, 2017).

A recent study focused on the exome capture defined as reliably callable (Zook *et al.*, 2014) to compare WES and WGS variant agreement. The study used a number of bioinformatics pipelines consisting of different combinations of variant callers and alignment tools for WES and WGS data from the reference NA12878 GIAB sample. This revealed that there was high variant agreement between the two methods especially for SNV (98.03-99.46%). However, this was significantly lower for InDels at 65.76-84.85% (Laurie *et al.*, 2016).

In the context of rare disease, inherited retinal disease has been an extensively studied group of disorders. Although a common cause of visual impairment, individual retinal disorders are rare and known to be genetically and phenotypically heterogeneous. An undiagnosed cohort of patients with inherited retinal disorders underwent WES; this yielded a diagnosis in 50% of patients. For an additional 6% of the patients, pathogenic mutations were identified using WGS. These additional mutations were missed by WES either because the region where the variant was localised was not included in the capture kit, the variant was a large InDel not called by WES or the variant was called but filtered out due to low quality. WGS also identified a novel intronic variant in a known inherited retinal disease gene (*CHM*) (Carss *et al.*, 2017).

These mechanisms may also play a limiting role in WES for other rare inherited disorders. For example, alternative splicing creates a long muscle-specific isoform of GFPT1 (Selcen *et al.*, 2013). In the NeurOmics project (https://rd-neuromics.eu/), which included patients with CMS, it was discovered that the muscle specific exon of *GFPT1* was not included in the WES 62 Mb capture kit (Illumina Nextera Rapid expanded exome) used in the project (personal communication from Hanns Lochmüller and Ana Töpf).

In addition, a deep intronic mutation in the DMD gene has been found to create a cryptic splice site and introduce a pseudo-exon causing a frameshift in the gene and resulting in muscular dystrophy (Zaum *et al.*, 2017). Furthermore, for laminopathies, intronic splice site mutations have been reported in the *LMNA* gene and it has been suggested that these are responsible for some genotype-negative cases with a clinical phenotype consistent with laminopathies (Rogozhina *et al.*, 2016).

These types of mutations are expected to play an important role in monogenetic disorders, including NMD, through creating pseudo-exons, altering splice sites and affecting transcription regulation. These variants are likely to become more recognised as WGS moves into clinical practice (Vaz-Drago *et al.*, 2017).

Similarly, structural variations such as large InDels, trinucleotide repeat expansions and CNVs are known causative mechanisms in NMD and WGS is expected to increase the number of known pathogenic structural variations in this group of disorders (Laing, 2012). Nonetheless, although WGS provides superior variant detection, interpretation and cost remain an issue and deciding which technology to use requires careful consideration. Here, both WES and WGS were performed for 10 patients with NMDs. Focusing on coding regions, the additional yield of WGS over WES was assessed and limitations of WES identified.

2.2 Aims

- Assess the comprehensiveness of WES when compared to WGS for coverage of coding regions in the context of NMD by using the variant output from both.

- Identify genomic features leading to low coverage in WES.

- Examine the added benefit of WGS in detecting InDels and CNV in coding regions in NMD genes.

2.3 Methods

2.3.1 Ethical approval

Ethical approval for this project was granted by Newcastle University Research Office (ref. 2306/2015). Informed consent for research was obtained from all patients undergoing WES and WGS under the protocol for Newcastle MRC Centre Biobank for Neuromuscular Diseases (REC reference: 08/H0906/28 + 5).

2.3.2 Genomic platforms

a. Clinical Sequence Analyser

CSA is a commercial web interface developed by deCODE Genetics to integrate NGS into clinical practice. Patient samples are sequenced at deCODE Genetics using an Illumina platform, reads aligned using BWA and variants called by GATK UnifiedGenotyper. The pipeline produces an annotated VCF file in addition to a Genomic ordered Relational (GOR) architecture file (Guethbjartsson *et al.*, 2016) containing coverage and variant frequency statistics. Variants are annotated for their predicted effect (Ensembl VEP package) and population allele frequency (1000 Genome project, NHLBI GO Exome Sequencing Project, Icelandic population statistics). CSA allows rapid retrieval and visualisation of sequence raw data by integrating the Sequence Miner software. This allows confirmation of candidate variant calls in the raw reads (NEXTCODE-HEALTH, 2016).

b. The RD-Connect Genome-Phenome Analysis Platform

RD-Connect Genome-Phenome Analysis Platform receives Fastq files or BAM files (WES or WGS) from data deposited at European Genome Archive (EGA). BAM files are reverted back to Fastq files. These are then processed through a standardized bioinformatics pipeline as shown in figure 5. GATK HaplotypeCaller is used for variant calling. Variants are annotated (table 3) and the resulting genomic VCF (gVCF) file is uploaded on the RD-Connect genomic platform in a Hadoop Distributed File System (HDPS). This is a system that allows distributed storage and processing of large data sets. A search engine is then included and allows real-time queries through a web-based client interface. On the interface, this is taken even further by cross-linking data from various databases and registries and through integrating databases and providing application-programming interface (API) access to relevant third-party resources, all in an authorized secure manner. The interface allows authorized users to customize queries and filter variants using for example, genotype and genotype quality, pathogenicity predictions or allele frequencies (Laurie, 2016).

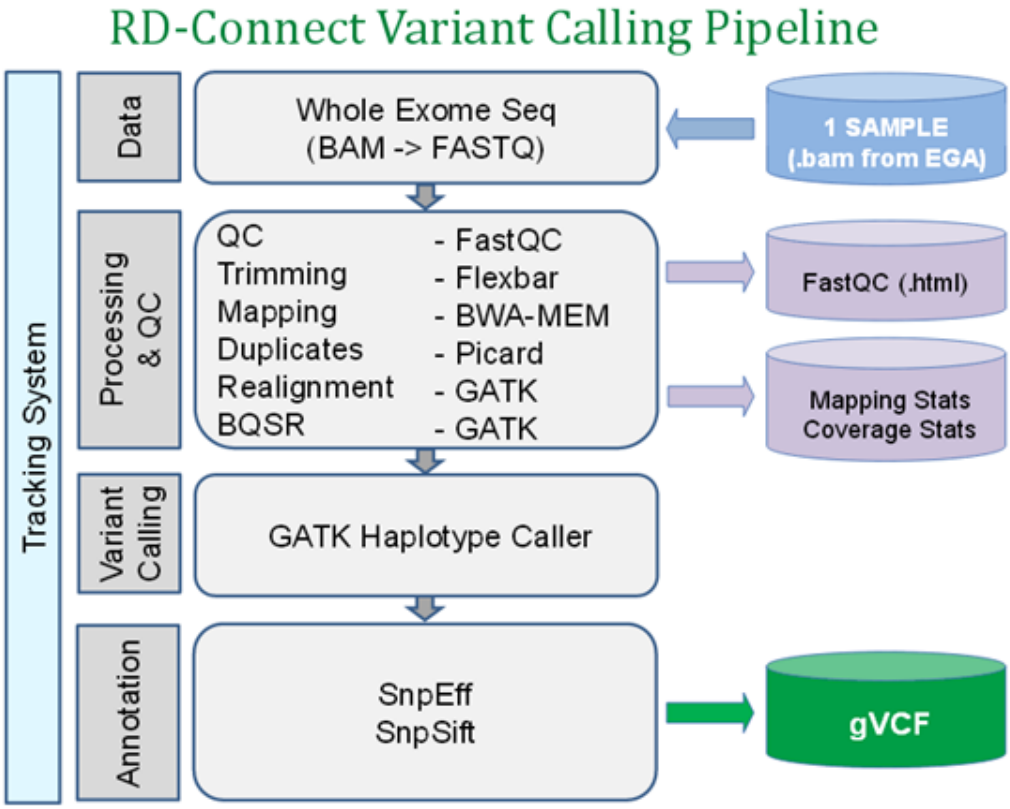


Figure 5: RD-Connect bioinformatics pipeline (Laurie, 2016).

Table 3: RD-Connect variant annotation tools and databases

Annotation	Tool
Call specific (quality score, read depth, etc.)	Bioinformatics algorithm
Functional annotation (gene name, transcript ID, amino acid change)	GRCh37-Ensembl 75
Variant Effect	SnpEff, Polyphen2, Mutation Taster, UMD, CADD
Population allele frequencies	1000 Genomes Project, ExAC, ESP6500

2.3.3 Patient samples

WES and WGS were performed and data processed for 10 patients with NMD (N1-10) according to the protocols in table 4 at deCODE Genetics (Iceland).

WES data for a cohort of 50 patients with neurodegenerative disorders was used as a control sample to assess coverage. Sequencing for these patients was also performed at deCODE Genetics and processed through the same WES pipeline.

Table 4: Sequencing and analysis pipeline at deCODE Genetics used for processing WES and WGS N1-10 samples (Personal communication: Nanna Vidarsdottir, deCODE Genetics).

Tool	WES	WGS
Sequencing method	Illumina Nextera Rapid expanded exome (62Mb target)	Illumina PCR-free whole genome sequencing
Mark duplicates	Picard 1.55	Picard 1.117
Aligner	BWA 0.6.2	BWA 0.7.10
Assembly	NCBI Build 37 of the human reference sequence (GRCh37/hg19)	NCBI Build 37 of the human reference sequence (GRCh37/hg19)
Caller	GATK Lite version 2.3-9 UnifiedGenotyper	GATK Lite version 2.3-9 UnifiedGenotyper
Reference data version	20130917.DCREF_1-1-0	DCREF_1-2-0
Sequence Miner	2.13.0 M5 (2014-09-24-1013)	5.21.1
CSA version	2.13.0M12	4.11.2
Damp in-house pipeline	clinseq_build37	clinseq_build37_v2

WES, whole exome sequencing; WGS, whole genome sequencing; BWA, Burrows-Wheeler Aligner; NCBI, National Centre for Biotechnology Information; GATK, Genome Analysis Tool Kit; CSA, Clinical Sequence Analyzer.

2.3.4 Analysis and filtering parameters

Variant output analysis was performed on the CSA platform using the filtering parameters shown in table 5 and focusing on coding variants.

Where variants were not consistent in WES and WGS for the same patient, Sequence Miner (version 5.7) was used to visualise the raw data for both and confirm the call. Exons in which variants were missed by WES were further studied for coverage. Read depth is provided for each variant position in the CSA output report and on the Sequence Miner.

The comparison of WES and WGS data for the same samples (N1-10) was repeated on the RD-Connect platform. Both data sets were processed through the same pipeline as described in figure 5 and table 3. Variant output analysis focused on coding regions in NMD genes and were filtered for variants with a population frequency of 0.01 or less, a read depth of 8 or more and variants with a predicted moderate to high impact on the protein.

For the major part of the analysis, output was restricted on both platforms to variants falling in NMD genes. This gene list was constructed based the 416 genes reported to be implicated in NMD by the World Muscle Society (WMS, http://musclegenetable.fr/) in July 2016 (appendix A).

Table 5: Filtering parameters used for WES and WGS comparison on the CSA

Section	Parameter	Default value
Variant effect and frequency filter	Recessive max allele freq	0.03
Variant effect and frequency filter	Dominant max allele freq	0.01
Variant effect and frequency filter	Recessive max gt freq	0.0001
Variant effect and frequency filter	VEP maximum impact	MODERATE
Variant quality	Variant filter	Include LowQual
Variant quality	Min Gt Likelihood	5
Variant quality	Min Het Call Percent	20
Variant quality	Min Hom Call Percent	66
Variant quality	Min read depth	8
ASMG Category settings	Cat1 clinical impact	pathogenic_only
ASMG Category settings	Cat1B distance	2
Genomeic range filter	Max dinstance for exome overlap	10
Genomeic range filter	exclude_repeat_regions	false
Genomeic range filter	Max dinstance for repeat overlap	2
Penetrance	case_delta	0
Penetrance	control_delta	0
Custom reference sources	Allele frequency file	None
Custom reference sources	max_cust_af	0.01
Custom reference sources	max_cust_gf	0.0001
Custom reference sources	Annotation variation file	None
Custom reference sources	Annotation region file	None
Custom reference sources	Exclusion variant file	None
Custom reference sources	Region file	None
Custom reference sources	region_file_usage	exclude
Custom reference sources	Max overlap distance	0
Custom reference sources	Gene panel info file	None
Gene reference	Gene coverage	Coding only
Gene reference	VEP genes	Ensembl

2.3.5 Coverage assessment

Variant metric data from the CSA and RD-Connect were used to assess WES read depth/coverage at variant sites identified through WGS but missed by WES. Coverage was determined at the chromosomal position for each variant and inferred to the respective exon. The mean coverage was estimated for the N1-10 samples and for the control sample (N=50). These means were then compared to the mean coverage at respective sites in the ExAC population data (version 0.3.1, http://exac.broadinstitute.org/).

2.3.6 Assessment of trinucleotide repeats calling

WES and WGS data for the N1-10 samples were compared for their ability to call trinucleotide repeat variants. These were assessed at loci known to be implicated in neurogenetic disorders via trinucleotide repeat expansions and for coding regions only. These loci are listed in table 6.

Table 6: Positions affected by trinucleotide repeat expansions and implicated in neurogenetic disorders.

Gene	Repeat	Disease (Clin/var)	Position on (GRch37)	Position in gene
ATN1	CAG	Dentatorubral pallidoluysian atrophy	chr12:7045892-7045894	Exon 5/10
ATN1	CAG	Dentatorubral pallidoluysian atrophy	chr12:7045880-7045882	Exon 5/10
HTT	CAG	Huntington's Chorea	chr4:3076604-3076606	Exon 1/67
AR	CAG	Bulbar-spinal atrophy, X-linked	chrX:66765160-66765162	Exon 1/5
ATXN1	CAG	Spinocerebellar ataxia 1	chr6:16327918-16327920	Exon 8/9
ATXN3	CAG	Spinocerebellar ataxia 3 (MJD)	chr14:92537382-92537384	Exon 10/11
CACNA 1A	CAG	Spinocerebellar ataxia 6, Episodic ataxia 2	chr19:13318673-13318675	Exon 47/47
ATXN7	CAG	Spinocerebellar ataxia 7	chr3:63898362-63898364	Exon 3/13
TBP	CAG	Spinocerebellar ataxia 17	chr6:170870996-170870998	Exon 3/8
FMR1	CGG	Fragile X syndrome	chrX:146993570-146993572	Exon 1/15
DMPK	GAA	Steinert myotonic dystrophy syndrome	chr19:46273520-46273522	Exon 15/15

2.3.7 Relationship between sequence specificity and read depth

The Ensembl BLAST tool for human GRCh37/hg19 (http://grch37.ensembl.org/Homo_sapiens/Tools/Blast) was used to assess whether variants were in regions with low sequence heterogeneity and thus may map to multiple sites in the genome. The analysis was performed by including 100, 75 and 50 bases upstream and downstream of the variant position and looking for matches in the genome for sequence lengths of 201, 150 and 101, respectively.

2.3.8 Relationship between GC content and read depth.

To assess whether the GC content had an impact on read depth for the N-10 samples, variants that were detected by WGS but missed by WES due to low coverage (read depth =/<10) were examined. UCSC Genome Browser (http://genome.ucsc.edu/) and version GRCh37/hg19 of the human genome were used to determine the GC content at particular variant positions. A GC content percentage is given on the bowser based on a 5-base window.

An additional 20 variant positions in NMD genes (table 7) with adequate coverage in WES for the N1-10 samples were randomly selected and assessed for GC content.

Table 7: WES variant position randomly selected from the N1-10 samples for assessment of the relation between GC content and coverage.

Gene	Exon	Position
ITPR1	44	chr3:4808290
MYH2	27	chr17:10432042
KCNE1	3	chr21:35821680
POMGNT1	20	chr1:46655645
PFKM	19	chr12:48537578
SBF1	37	chr22:50886728
TTN	42	chr2:179629358
HSPG2	3	chr1:22214036
VAPB	5	chr20:57016039
DDHD1	1	chr14:53619480
TBP	2	chr6:170871051
TTN	301	chr2:179404402
FLNA	38	chrX:153580986
TTN	325	chr2:179441295
KLHL41	1	chr2:170366686
ETFDH	6	chr4:159618760
GNE	11	chr9:36217445
AKAP9	9	chr7:91630298
MURC	2	chr9:103348329
TUBB3	1	chr16:89986130

2.3.9 Statistics

Statistical significance was assessed using single tailed student t-Test in Microsoft Excel, where the p-value was assessed against an alpha value of 0.01. Standard deviation was calculated for the mean values from the patient samples for WES and WGS separately.

The CORREL function in Microsoft Excel was used to calculate the Pearson Product-Moment correlation co-efficient. The syntax for the function is CORREL (array1, array2) for two sets of values (GC content percentage and coverage) where, array 1 is the independent variable (GC content percentage) and array2 the dependant one (coverage).

2.4 Results

2.4.1 WES and WGS comparison using CSA

WES and WGS were performed for ten cases (N1-10) according to the pipeline in table 4. Using the filtering parameters in table 5, the CSA was initially used to analyse the data with a focus on rare variants in known NMD-causing genes and with a moderate to high impact on the protein structure.

The mean number of coding variants and coding variants in NMD genes are shown in figure 6. For coding variants in NMD genes the difference in the number of variants between WES and WGS experiments was not significant (p-value = 0.06). Nonetheless, WES was limited by low coverage (read depth <10) at particular exons in NMD genes. This meant that a number of variants were present in the output report as possible candidates by WGS analysis but were missed by WES for the same patients.

Exons in which variants had a read depth of 10 or less are shown in figure 7. Read depth at these positions was then examined in 50 WES control samples. With the exception of exon 1 of the *AR* gene, coverage analysis of the remaining exons revealed consistent low coverage. These positions were then further examined in the ExAC database for read depth. For three of the genes, the exomes from the ExAC population also showed low read depth: for exon 1 of the *KCNC3* gene, for the majority of exon 1 of the *SLC10A4* gene and for the last exon (exon 36) of the *SLIT3* gene (table 8).

These exons were then examined for sequence features leading to low coverage. Variants that were called through WGS in these exons were not at sites affected by homopolymers or repeat

sequences based on the NCBI reference genome (GRCh37/hg19) and thus, these mechanisms cannot explain low coverage.

The second observation from this comparison was the presence of a CAG trinucleotide repeat expansion in exon 9 of the *ATXN3* gene. The expansion ranged from 7-10 repeats and was called by WGS in six individuals. Despite this region having adequate read depth and the repeat sequence being present in the WES raw reads (visualized through Sequence Miner on the CSA), the expansion was not called by the WES analysis pipeline of CSA. This indicated a limitation of the latter in calling the trinucleotide repeat expansion.

To examine this further, 13 coding loci in the human genome implicated in disease through trinucleotide repeat expansions (table 6) were investigated. Results showed that WGS called more trinucleotide repeat expansions at these sites than WES when the samples were analysed using CSA and the filters in table 6. This observation was still correct when loci with a mean read depth of less than 10 were excluded from the analysis as shown in figure 8.

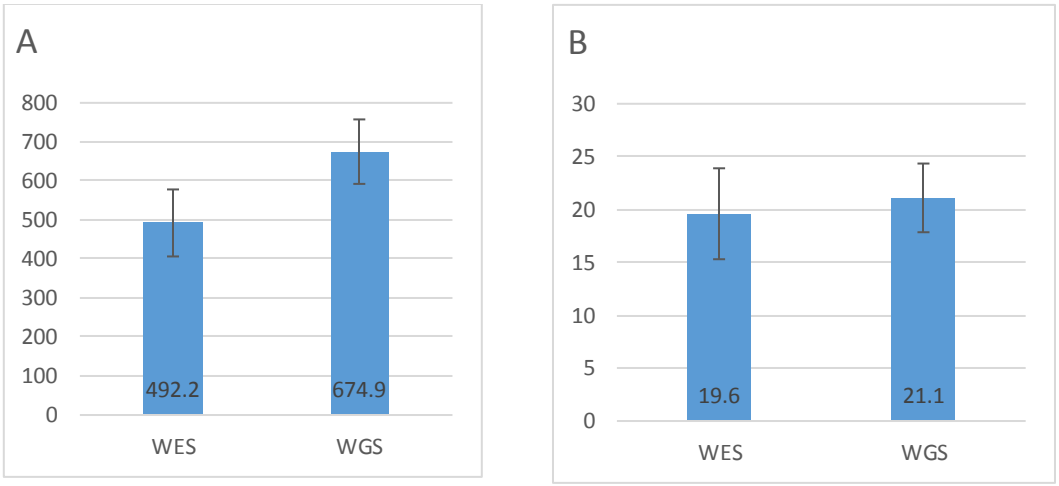


Figure 6: Mean number of coding variants for WES and WGS for 10 patient samples as outputted by the CSA. (n=10). A; all coding variants, B; coding variants in NMD genes.

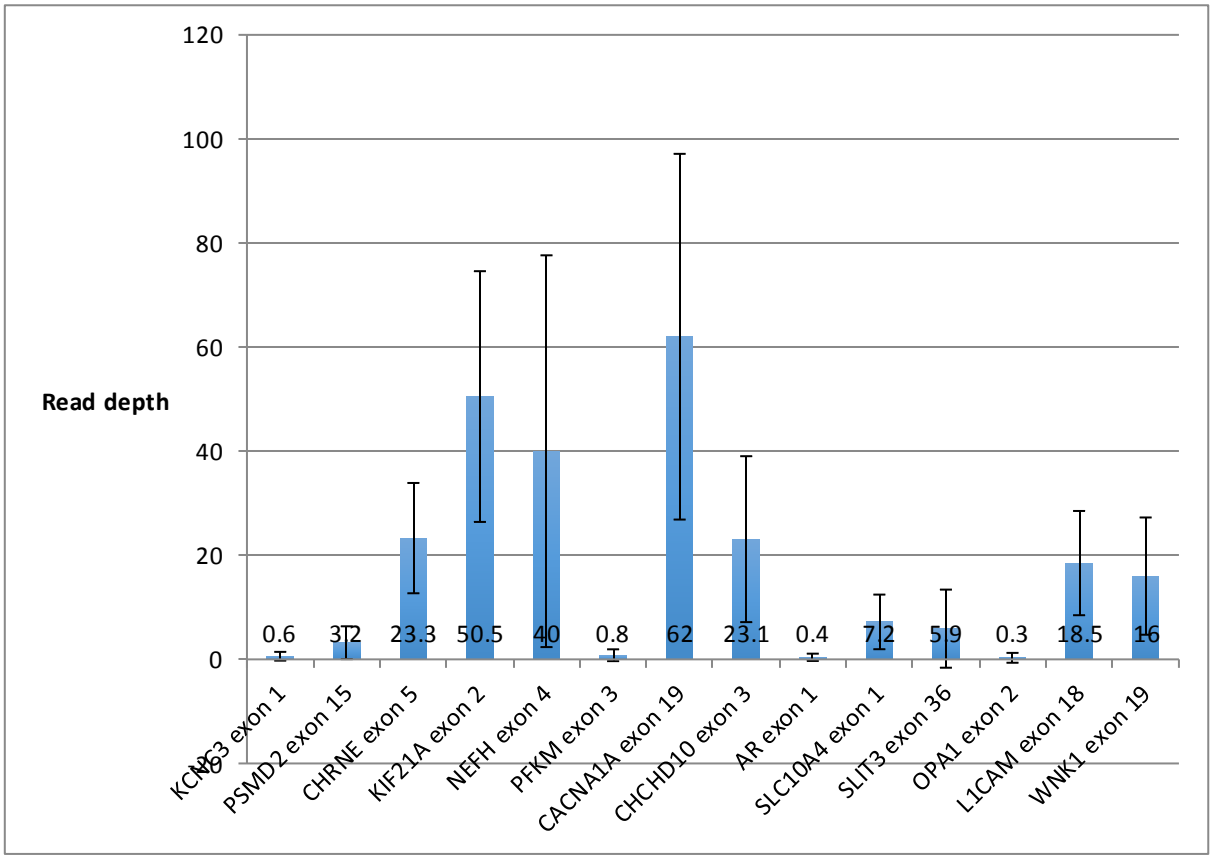


Figure 7: Mean read depth in WES for exons with variants proposed by WGS but missed by WES for the same patient. Error bars represent the spread of read depth across the sample (n=10).

Table 8: Coverage data for exons identified as having a read depth of 10 or less in WES of the N1-10 samples. Mean coverage is shown for randomly selected samples for patients with neurogenetic disorders and from the ExAC population.

Gene	Exon identified with low read depth	Mean read depth in Neuromics WES data set*	Mean read depth in 50 Neuromics WES samples**	Maximum coverage per base in exon in ExAC population (WES data)	Average read depth for all exons in ExAC population
KCNC3	Exon 1	0.6	<10	<10	30.07
PSMD2	Exon 15	3.2	<10	90	70.49
PFKM	Exon 3	0.8	<10	75	68.57
AR	Exon 1	0.4	>100	65	30.20
SLC10A4	Exon 1	7.2	>10 (13.8)	10 (for initial 2/3 of exon)	51.70
SLIT3	Exon 36	5.9	>10 (12.7)	<10	55.37
OPA1	Exon 2	0.3	<10	85	53.89

*, ** Read depth refers to that at the site of variants identified by WGS in each exon. Genes in bold font: consistently showing low coverage for the exons stated.

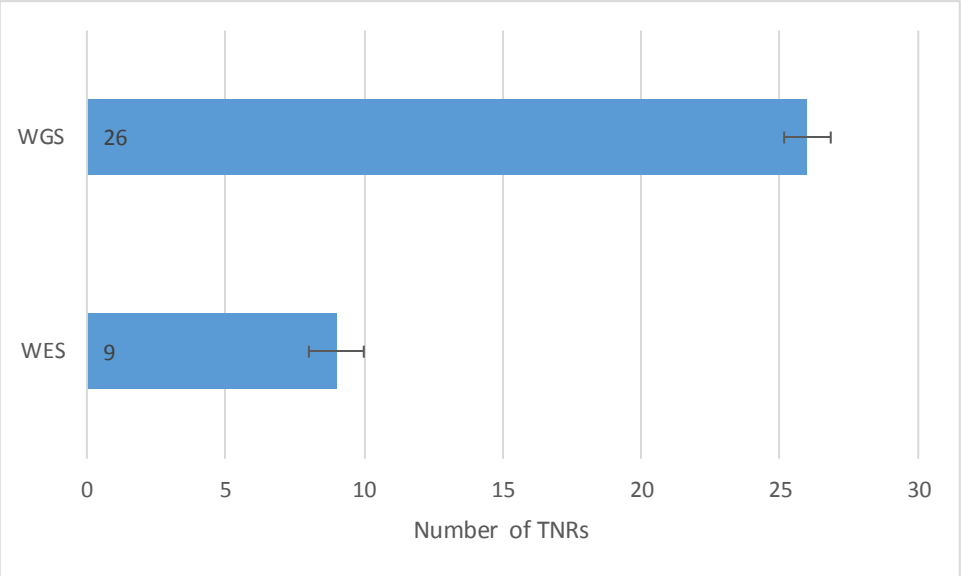


Figure 8: Number of coding trinucleotide repeats (TNRs) at known neurogenetic disease loci* from analysis of WES and WGS samples for patients (N1-10) on CSA. * Only loci with a mean read depth of 10 or more in WES were included.

2.4.2 WES and WGS comparison on the RD-Connect Genome-Phenome Analysis Platform.

The mean number of coding variants for the N1-10 samples WES and WGS in the RD-Connect platform output is given in figure 9. Variants were filtered for frequency, impact on protein structure and read depth as mentioned in the methods and the concordance of variants between WES and WGS was then assessed (figure 10). Overall, 65.6% of coding NMD gene variants were called by both WES and WGS on the RD-Connect platform. This was mainly accounted for by concordance in calling SNV (78.9%).

For coding variants in NMD genes on the RD-Connect platform, the difference in the number of variants called by WES and WGS for the N1-10 samples was not significant (P-value= 0.29 and 0.10 for SNV and InDel variant number, respectively).

In addition to the three regions identified by the CSA as having low coverage in WES, 11 additional variants were missed by WES due to low coverage (read depth =/<10) in WES. The sites of these variants were further examined in the ExAC population and their coverage remained low. These sites are listed in table 9. GC content and read specificity to align to the correct site in the genome are also shown for these loci.

To assess the relationship between GC content and coverage, an additional 20 positions with adequate coverage to call a variant in WES for the N1-10 samples were randomly selected. The line chart in figure 11 shows the relationship at all positions (n=31). The association of GC content and coverage at these positions shows a moderate negative trend (correlation coefficient =-0.51), suggesting that GC content and coverage are inversely related.

As for the specificity of these sites to match a single region in the genome, data in table 9 shows that sequences at these positions are not unique and align to more than one region in the genome. This data shows that a high GC content and similarity of sequence reads with more than one site of the reference genome are features of areas with low coverage in WES.

With regards to the limitation of WES in identifying TNRs shown in section 2.4.1, this difference between WES and WGS was not significant on the RD-Connect platform where only two additional TNR were called in the WGS experiment. This was a nine CAG repeat in exon 9 of the *ATXN3* gene and appeared in the output for two patients.

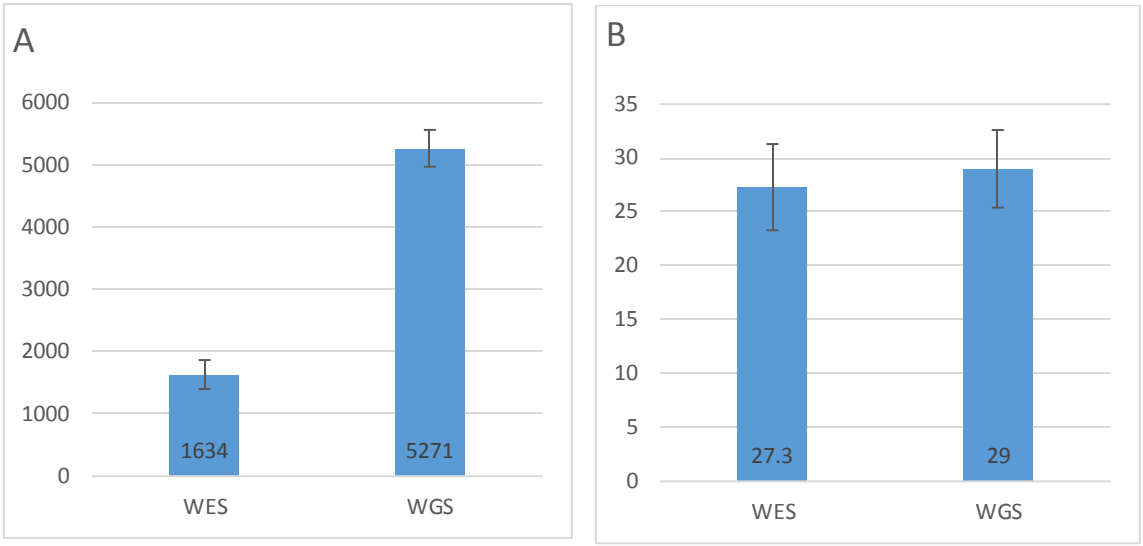


Figure 9: Mean number of coding variants for WES and WGS for 10 patient samples as outputted through the RD-Connect (n=10). A; all coding variants, B; coding variants in NMD genes.

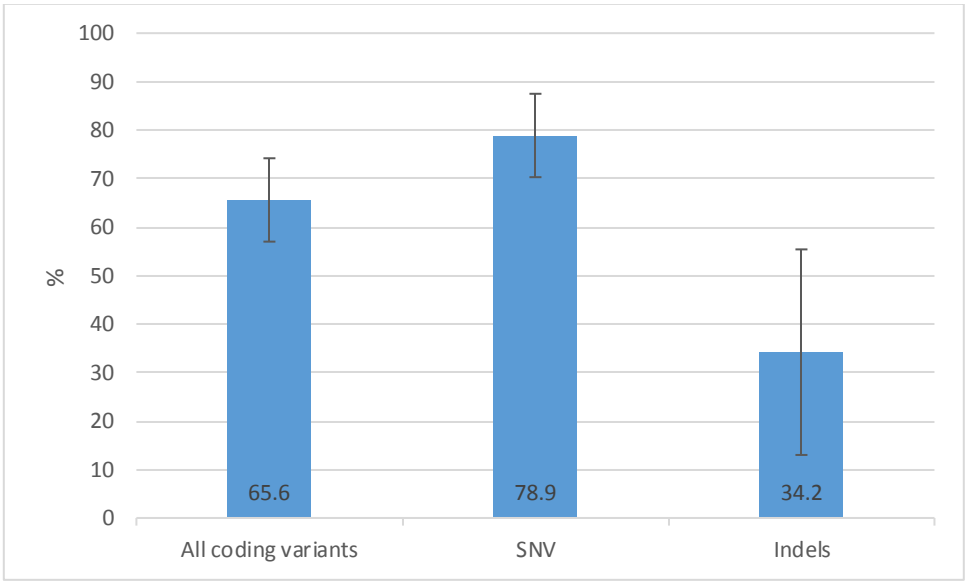


Figure 10: Concordance for coding variants in NMD genes between WES and WGS on the RD-Connect platform (n=10).

Table 9: Variant positions identified as having low coverage in WES for the N1-10 samples showing, coverage in the ExAC population, GC content and number of matches in the genome using the BLAST tool.

Variant position	Gene	Exon	ExAC coverage	GC content % *	Number of matches in the genome using BLAST		
					N=201	N= 151	N=101
chr3:63898360	ATXN7	exon 3	0	100	1	1	5
chr2:241696840	KIF1A	exon 27	10	60	2**	2	2
chr19:13318672	CACNA1A	exon 23	2	80	7	7	7
chr14:105173862	INF2	exon 8	10	80	7	6	8
chr12:112036753	ATXN2	exon 1	10	100	9	3	4
chr12:112036770	ATXN2	exon 1	10	80	3	3	4
chr15:23086364	NIPA1	exon 1	10	100	6**	6**	15**
chr12:32687343	FGD4	exon 1	2	40	3	3	9
chr12:112036796	ATXN2	exon 1	5	80	3	3	1
chr16:66583871	TK2	exon 1	5	100	1	1	2
chr17:4852305	PFN1	exon 1	0	100	9	6	9

* based on 5 base window, **E-value significant (=/<0.01), E-value: probability that the alignment between the query sequence and the subject sequence is due to chance (BLAST).

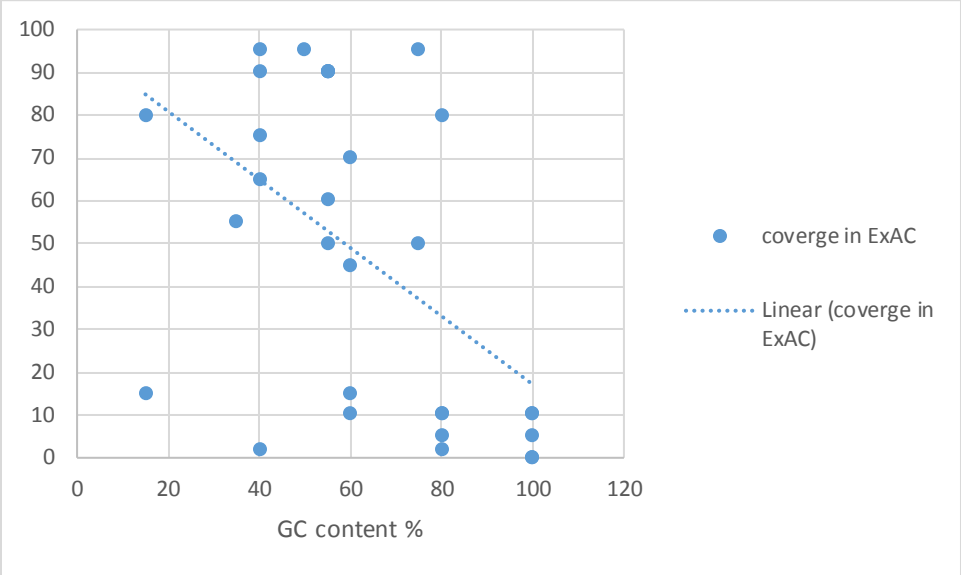


Figure 11: Relationship between GC content and coverage in ExAC WES data at variant positions from the RD-Connect platform output report for sample N1-10.

2.5 Discussion

Patients N1-10 had an undiagnosed NMD and were recruited for WES and later underwent WGS as part of the same project as WES did not propose any causative mutations. Although in this case WGS failed to identify relevant pathogenic mutations, comparison of the WES and WGS data was informative.

The sequencing was performed at deCODE Genetics (Iceland) and the data was uploaded on their analysis platforms the CSA. Later, NGS data for the N1-10 samples was shared and uploaded onto the RD-Connect Genome-Phenome Analysis Platform.

Comparison between WES and WGS data for these patients focused on coding regions in NMD genes. Analysis on the CSA showed that the number of rare, damaging variants in NMD genes was not significantly different between WES and WGS datasets. However, the analysis highlighted two limitations of WES. First, coding exons in known NMD genes were found to consistently have poor coverage and second, that WGS analysis pipeline was superior at calling trinucleotide repeats. The latter was due to a limitation of the WES bioinformatics pipeline as the repeats were present in the raw sequences for these patients but did not appear in the output report. The time gap of approximately 18 months between when WES and WGS experiments were carried out and processed at deCODE Genetics meant that pipeline tools were updated and more recent versions were used to process WGS data as shown in table 4. The difference in the pipelines is likely to have contributed to the

discrepancies in the output results between WES and WGS. Therefore, the comparison was repeated on the RD-Connect platform, where all WES and WGS experiments are processed through the same pipeline and re-processed following any software update.

On the RD-Connect platform, again, coding variant number in NMD genes were not significantly different between WES and WGS. However, WES and WGS agreed on only 65.6% of the variants. InDel agreement was much lower (34.4%). In addition, comparison of the variants proposed an additional 11 NMD gene exons that are poorly covered by WES, having a read depth of 10 or less.

The most important reason for low sensitivity in an NGS experiment is low coverage. It is generally agreed that sufficient coverage for variant calling is 20x or more. However, for WES experiments achieving uniform coverage for all coding exons is not currently possible as 5-10% of coding regions are not sufficiently covered (Dewey *et al.*, 2014; Lelieveld *et al.*, 2015; Warman Chardon *et al.*, 2015). In part, WES low coverage maybe attributed to coding regions that are missed in the exome capture kit. Earlier generations of exome capture platforms were able to capture 80.5% of coding regions (Parla *et al.*, 2011). Whereas, newer version have shown higher capture (up to 95%) and thus higher sensitivity in variant calling (Lelieveld *et al.*, 2015). Nonetheless, WGS is superior to WES in providing uniform coverage for more than 98% of coding regions. This even coverage not only is expected to result in higher sensitivity of SNV detection but also InDel and CNV detection (Medvedev *et al.*, 2010).

GC-content bias is another important cause of low coverage. Studying the relationship between GC content percentage and coverage at variant positions proposed on the RD-Connect showed moderate negative correlation (correlation coefficient -5.1), indicating that as GC content increases coverage decreases. In addition, seven of the low coverage regions identified (63.6%) correspond to exon 1 of their respective genes. The first exons are known to be GC-rich and their capture and coverage is known to be problematic due to PCR bias.

GC-bias occurs at a threshold of a GC content of 62%. Above that, loss of reads and thus low coverage may occur (Roeh *et al.*, 2017). This bias is directly related to the number of PCR cycles in the sequencing protocol (Lelieveld *et al.*, 2015). A PCR-free WGS is expected to completely overcome this issue and thus coverage of these GC-rich regions will be adequate and uniform leading to a higher sensitivity for variant detection (Meienberg *et al.*, 2016). It is important to note that solutions to overcome GC-bias may vary for different sequencing technologies. For example, on the Illumina systems, optimising conditions for library

preparations and PCR thermocycler has been proven to reduce PCR related bias (Aird *et al.*, 2011). However, these solutions are not effective on the SOLiD sequencing systems (Applied Biosystems) and Illumina is more suited for experiments requiring sequencing for GC rich parts of the genome including for WGS experiments (Roeh *et al.*, 2017).

For variant positions with low coverage, specificity to match to the correct site in the reference human genome was also assessed using the BLAST tool. This revealed that even at sequence reads of 200 bases, 9/11 sites had at least two hits in the genome. This may cause an incorrect read alignment and poor coverage at the correct site. Increasing the positional accuracy of read alignment maybe overcome by longer and paired-end reads and by intersecting results from more than one alignment tool. In addition, de novo assembly rather than alignment to the reference genome may also be relevant here. These solutions to improve read specificity in WES come with additional costs and computational complexity. The longer reads in WGS maybe a more efficient solution (Ratnakumar *et al.*, 2010; Techa-Angkoon *et al.*, 2017).

With regards to trinucleotide repeats, analysis on the CSA revealed that WGS was superior in detecting these variants. Examining the raw reads showed that the majority of these trinucleotide repeats were present but were not detected as variants by the WES analysis pipeline. As mentioned above, deCODE Genetics used an updated pipeline for the analysis of WGS data. Therefore, for these relatively short sequence variations, the limitation was in the analysis pipeline and the calling algorithm rather than the sequencing technology itself. This is further supported by the RD-Connect analysis, where there was no significant difference between the number of coding trinucleotide repeats called by WES and WGS.

In conclusion, WGS provides higher sensitivity in variant detection in coding regions by overcoming issues in coverage related to capture kits, GC-bias and read specificity. The uniform coverage of WGS also increases sensitivity to detect large InDels, CNVs and trinucleotide repeat expansions. These are all described pathogenic molecular mechanisms in inherited neuromuscular disorders (Laing, 2012). It is expected that WGS will replace array-based methods for detecting these structural variations. In addition, WGS is able to detect intronic variants that may account for a proportion of undiagnosed cases. Improvement in WES coverage through deep sequencing, improved capture kits and improved analysis algorithms are likely to improve variant detection in the exome capture region. However, WGS currently remains superior for the detection of intronic and structural variants (Lelieveld *et al.*, 2015; Meienberg *et al.*, 2015; Meienberg *et al.*, 2016; Zatz *et al.*, 2016).

Nonetheless, WGS costs remain high, analysis is complex and the technique carries a higher risk of incidental findings unrelated to the patients' presenting disease. These however maybe overcome through exon targeted analysis of WGS data or the use of a "virtual panel" consisting of genes related to the disease of interest (Meienberg *et al.*, 2016).

Chapter 3. Genomic Platform and bioinformatics pipeline comparison

3.1 Introduction

Development and continued innovation of NGS sequencing technologies have led to a decrease in costs and an increase in application in medical research and in diagnostics (Alioto *et al.*, 2015). Nonetheless, clinical use of NGS requires it to show high sensitivity and specificity when compared to the current gold standard, Sanger sequencing. Although NGS technologies offer high sensitivity across larger proportions of the genome, this may be low for particular loci due to genomic features and limitations of the sequencing technology leading to low coverage (Lelieveld *et al.*, 2015). In addition, challenges in analysis and interpretation of NGS data may contribute to this low sensitivity (Alioto *et al.*, 2015). Comparisons of the performance of sequencing platforms, and enrichment and capture methods have shown significant variability in target capture, coverage and thus variant calls. These issues are unceasingly addressed through innovative sequencing method development with a focus on increasing the depth and uniformity of coverage (Clark *et al.*, 2011; Lelieveld *et al.*, 2015).

NGS data analysis methods have also shown discrepancies. These may arise from the many steps of the analysis pipeline and may be due to variations and limitations in the algorithms for quality assessment, alignment, variant calling, variant annotation or variant filtration (Alioto *et al.*, 2015; Cornish and Guda, 2015; Laurie *et al.*, 2016; Allali *et al.*, 2017).

The two most important steps of the analysis pipeline are read alignment (giving an indexed BAM file) and variant calling (producing a VCF file). Previous work comparing various combinations of tools at these two steps has shown that performance is variable, and it has been suggested that the performance of variant callers is highly dependable on the read alignment tool in the algorithms tested (Cornish and Guda, 2015).

In clinical research and diagnostics, variant annotation and prioritization are equally vital to identifying variants that are implicated in the disease of interest and understanding of its biology and in developing novel therapeutics (Butkiewicz and Bush, 2016). Computational algorithms assign attributes to variants using a comprehensive transcript set. Filters can then be applied to the annotated variants to produce a shorter list of disease-relevant candidates for further investigation. Filtering is also performed using computational algorithms and

interfaces that generally differ depending on their developer, purpose and target users (Wang and Xing, 2013; Taylor *et al.*, 2015; Yang and Wang, 2015; Stark *et al.*, 2017).

Functional variant annotation may follow an algorithmic or non-algorithmic method. For the former, annotation takes a qualitative or quantitative approach. The qualitative variant annotation algorithm examines the sequence in which the variant is located and assigns its consequence accordingly. The algorithm uses a reference transcript catalogue for example, the Ensembl or Refseq transcript sets, and databases for genomic functional elements, such as those developed by the GENCODE and ENCODE projects. Initially, the algorithm identifies and annotates where the variant falls in the genome (exon, intron or untranslated region). It then uses known sequence motifs to identify and annotate for sequence features such as splice sites. Finally, it defines the sequence change (SNV or InDel) and assesses its consequence (amino acid substitution, premature stop, frameshift, or splice site modification). The result is that each variant in the VCF file gets one or more of the attributes listed in table 10. These attributes follow the standard Sequence Ontology (SO, http://www.sequenceontology.org/). The most commonly used annotation tools, such as ANNOVAR, SNPEff and VEP, use the qualitative algorithm (Pabinger *et al.*, 2014; Frankish *et al.*, 2015; Butkiewicz and Bush, 2016).

The quantitative algorithm is based on assigning an impact score to variants. This is based on statistical or machine learning algorithms that predict the impact of a variant. The algorithm requires a training set of disease associated variants to be used as a reference. Mutations provided in OMIM and HGMD are usually used for this purpose. Quantitative scores may also use sequence features such as motifs, domains and species conservation data. Software applications that attribute an impact or deleteriousness score to sequence variants include tools such as SIFT, Polyphen2, Mutation Taster and CADD (Butkiewicz and Bush, 2016).

Variants are then filtered and prioritised based on the associated annotations. Many software applications (genomics platforms) with complex computational algorithms are developed for this purpose. The applications have a user interface that allows users to manipulate and filter variants from the annotated VCF files. These software applications vary in the databases they integrate, filtering options, filtering algorithms and degrees of flexibility and interactivity (Alexander *et al.*, 2017; Jalali Sefid Dashti and Gamieldien, 2017; Muller *et al.*, 2017; Stark *et al.*, 2017).

NGS patient data used in this project utilise three genomics platforms developed at different academic or commercial institutions for the analysis of panel, WES, WGS and RNA

sequencing data. RD-Connect (CNAG, Barcelona, http://www.cnag.crg.eu/), the Broad Institute of Harvard and MIT and deCODE Genetics process the raw data in the form of Fastq files to the VCF/gVCF stage. These are then uploaded on the respective genomics platforms, the RD-Connect Genome-Phenome Analysis platform, *seqr* and the Clinical Sequence Analyzer (CSA). The data is then available for collaborating researchers to interrogate through a computer interface. The bioinformatics tools used at each site are specified in table 11. Overall, all three sites use a version of the BWA for read alignment and a GATK variant caller. The Broad Institute and deCODE Genetics use the VEP for variant annotation, while RD-Connect uses the SNPEff annotation tool.

Here, WES and WGS data from patients with rare NMD are used to assess the agreement of variant output from RD-Connect, *seqr* and CSA platforms. A reference genome, GIAB, is also used on the platforms as a mean of comparing the platforms against a set of high quality reference variants. This data is also used to assess concordance of the site-specific bioinformatics pipelines.

Table 10: Sequence ontology terms and identifiers assigned by tools using a qualitative annotation algorithm based on the Standard Sequence Ontology (Eilbeck *et al.*, 2005).

SO term	SO identifier
transcript_ablation	SO:0001893
splice_acceptor_variant	SO:0001574
splice_donor_variant	SO:0001575
stop_gained	SO:0001587
frameshift_variant	SO:0001589
stop_lost	SO:0001578
start_lost	SO:0002012
inframe_insertion	SO:0001821
inframe_deletion	SO:0001822
missense_variant	SO:0001583
splice_region_variant	SO:0001630
incomplete_terminal_codon_variant	SO:0001626
stop_retained_variant	SO:0001567
synonymous_variant	SO:0001819
coding_sequence_variant	SO:0001580
mature_miRNA_variant	SO:0001620
5_prime_UTR_variant	SO:0001623
3_prime_UTR_variant	SO:0001624
non_coding_transcript_exon_variant	SO:0001792
intron_variant	SO:0001627
NMD_transcript_variant	SO:0001621

non_coding_transcript_variant	SO:0001619
upstream_gene_variant	SO:0001631
downstream_gene_variant	SO:0001632
TFBS_ablation	SO:0001895
TFBS_amplification	SO:0001892
TF_binding_site_variant	SO:0001782

SO: Sequence Ontology, UTR: untranslated region, NMD: nonsense-mediated decay, TFBS: transcription factor binding site, TF: transcription factor.

Table 11: Genomic Platforms used for NGS data analysis.

Genomic Platform	Affiliation	Bioinformatics data processing			
		Pre-processing	Alignment	Variant calling	Variant annotation
RD-Connect	CNAG, Barcelona (Academic)	BAM> FASTQ FASTQC Picard	BWA-MEM	HaplotypeCaller	SNPEff
Clinical sequence Analyser (CSA)	WuXi NextCODE, Iceland (Commercial)	Picard	BWA	UnifiedGenotyper	VEP
Seqr	Broad Institute, Boston (Academic)	Picard	BWA	HaplotypeCaller	VEP

BWA; Burrows Wheeler Aligner, CNAG; National Centre for Genomic Analysis Barcelona, VEP; Variant Effect Predictor,

3.2 Aims

- Compare variant output from RD-Connect, *seqr* and CSA platforms for WES and WGS data from patients with NMD and assess concordance rates between the platforms for variants falling in NMD genes.
- Compare the bioinformatics pipelines used to process WES and WGS data prior to upload on the genomic platforms.
- Use a set of high confidence variant calls (GIAB) to assess agreement of the platforms and bioinformatics pipelines with each other and with the reference.

3.3 Methods

3.3.1 Ethical approval

Ethical approval for this project was granted by Newcastle University Research Office (ref. 2306/2015). Informed consent for research was obtained from all patients undergoing WES and WGS under the protocol for Newcastle MRC Centre Biobank for Neuromuscular Diseases (REC reference: 08/H0906/28 + 5).

3.3.2 WES and WGS samples

To maximise sample size, all WES and WGS data available for analysis on at least 2/3 platforms were used. All samples belonged to affected and non-affected individuals from families with NMD. Datasets were labelled: cohorts A, B, C, D and sample NA12878. A description of each cohort, the sequencing technology and bioinformatics pipeline is given below and summarised in table 12.

a. Cohorts A and B

Cohorts A and B consisted of data from 88 WES samples and 30 WGS respectively. These samples were sequenced at deCODE Genetics in Iceland and processed through the WES and WGS bioinformatics pipeline for deCODE Genetics shown in table 4. Data from these samples was also processed at RD-Connect according to the pipeline and tools shown in table 3 and figure 5. VCF/gVCF files for Cohorts A and B were uploaded on the CSA and the RD-Connect Genome-Phenome Analysis Platform and used to compare variant output from both platforms.

b. Cohort C

Data from 120 samples were sequenced at the Broad Institute of Harvard and MIT on an Illumina HighSeqXs platform. The capture kit used was the Agilent Sure-Select Human All Exon v.2.0 (44Mb). Read alignment was performed using BWA, variant calling via GATK and annotation using VEP. Data was then uploaded on the *seqr* platform for analysis. The data was also processed through the RD-Connect pipeline (table 3 and figure5) and uploaded onto the RD-Connect Genome-Phenome Analysis pipeline. Cohort B datasets were then used to compare variant output between *seqr* and the RD-Connect platform.

c. Cohort D

WES data from 27 individuals and a candidate mutation proposed through analysis of WES data on the *seqr* platform were used. These samples were sequenced at the Broad Institute of Harvard and MIT using an Illumina exome capture (38Mb) on the Illumina HighSeqXs platform. BWA, GATK and VEP were used for read alignment, variant calling and variant annotation respectively. These samples were also processed and uploaded on the RD-Connect platform. Data analysis on the latter was used to test whether the platform proposes the same candidate mutations as *seqr*.

d. Cohort E

WES data for 9 samples was used to compare all three bioinformatics pipelines and genomic platforms. Samples were sequenced at the Broad Institute and processed through all three pipelines. VCF/gVCF files for these samples were uploaded for analysis on *seqr*, CSA and the RD-Connect platform where variant output was compared. VCF files for Cohort E were also used to assess agreement of variant calling between the bioinformatics pipelines.

e. NA12878 reference genome

The reference genome for the NA12878 sample from the GIAB was used as ground truth to assess sensitivity of the bioinformatics pipelines by comparing pipeline outputs to the reference material published online. Platform outputs for the GIAB sample was also compared.

Fastq files for the NA12878 reference genome sample were downloaded from the European Nucleotide Archive (www.ebi.ac.uk.ena/data/view/ERP001229). The Fastq files were generated on the Illumina HiSeq2000 platform and corresponds to approximately 50x

coverage for the genome. Files were then sent to deCODE Genetics, the Broad Institute and RD-Connect for processing through site-specific bioinformatics pipelines and uploaded on the respective platforms (CSA, *seqr* or RD-Connect Genome Phenome Analysis Platform) for analysis of variant output. VCF files were requested from each site and used to assess agreement of the bioinformatics pipelines at the three sites in variant calling.

The reference VCF file for the NA12878 sample was downloaded from the Genome in a Bottle Resources (http://jimb.stanford.edu/giab-resources/). Version 3.3.2 of the high confidence variant calls was downloaded from

ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/ on 13.12.2017.

Table 12: Whole exome and genome sequencing data used for the comparison of genomics platforms and bioinformatics pipelines.

Dataset	Type of NGS data	Number of patients	Sequencing site	Processing pipelines	Data analysis platforms
Cohort A	WES	88	deCODE Genetics	deCODE Genetics and RD-Connect	CSA and RD-Connect Genome Phenome Analysis Platform
Cohort B	WGS	33	deCODE Genetics	deCODE Genetics and RD-Connect	CSA and RD-Connect Genome Phenome Analysis Platform
Cohort C	WES	120	The Broad Institute of Harvard and MIT	The Broad Institute of Harvard and MIT and RD-Connect	*seqr* and RD-Connect Genome Phenome Analysis Platform
Cohort D	WES	27	The Broad Institute of Harvard and MIT	The Broad Institute of Harvard and MIT and RD-Connect	*seqr* and RD-Connect Genome Phenome Analysis Platform
Cohort E	WES	9	The Broad Institute of Harvard and MIT	The Broad Institute of Harvard and MIT, deCODE and RD-Connect	*seqr,* CSA and RD-Connect Genome Phenome Analysis Platform
NA12878	WGS	1	Reference set of high confidence calls generated from 14 sequencing experiments.	The Broad Institute of Harvard and MIT, deCODE and RD-Connect	*seqr,* CSA and RD-Connect Genome Phenome Analysis Platform

3.3.3 Standardised filters for platform output assessment

Assessment of agreement in variant output between the RD-Connect platform, CSA and *seqr* was carried out by standardising the user-modifiable filters on the platforms. The filters in table 13 were applied on all three platforms. All passing variants were included irrespective of inheritance model.

Table 13: Standardized filters applied to compare the RD-Connect, *seqr* and CSA platforms.

Parameter	**Filter**
Variant quality (GQ)	50
Read depth	>8*
Variant effect	Moderate-high
Variant frequency in control population(s)	1%
Gene coverage	Coding and non-coding regions
Deleteriousness predictions	Include all variants

*With the exception of *seqr*, as the Broad Institute represent read depth in an overall sensitivity score for each variant.

3.3.4 Assessment of platform agreement

For each individual WES or WGS dataset, the filters above were applied on the relevant platforms. The total number of passing variants on each platform for a particular patient were noted and the mean calculated for each cohort. For NMD genes (appendix A), variant outputs for each individual from each platform under comparison were assessed for concordance. A mean concordance rate was then calculated for each cohort.

Cohort D was used to assess whether the RD-Connect platform agreed with *seqr* in proposing the same candidate causative mutations. On the RD-Connect platform, variants were filtered for those with a read depth of 8 or more, having a moderate to high impact, a maximum population frequency of 1%, and those located in known NMD genes.

3.3.5 Bioinformatics pipeline assessment

To assess the agreement of variants in VCF files produced by the site-specific bioinformatics pipelines for *seqr*, CSA and RD-Connect, the VCF comparison commands of VCFtools (v0.1.12a, Adam Auton and Anthony Marcketta 2009, http://vcftools.sourceforge.net/man_latest.html) were used. The tool was used on a Newcastle University remote server (Monolith). Files were analysed in the "*.vcf.gz"*/zipped format. The following commands were used to assess concordance for all variants, InDels and SNVs, respectively: *vcftools --gzvcf file1.vcf.gz --gzdiff file2.vcf.gz --out filename*, *vcftools --gzvcf file1.vcf.gz --gzdiff file2.vcf.gz --keep-only-indels --out filename and vcftools --gzvcf file1.vcf.gz --gzdiff file2.vcf.gz --remove-indels --out filename.*

For analysis of the reference genome sample (NA12878), VCFtools (v.0.1.15) was used on Newcastle University's Faculty of Medical Sciences remote cluster (fmsclustergw). The following commands were used for the updated VCFtools version:

vcftools --gzvcf file1.vcf.gz --gzdiff file2.vcf.gz --diff-site --out filename, vcftools --gzvcf file1.vcf.gz --gzdiff file2.vcf.gz --diff-site --keep-only-indels --out filename and *vcftools --gzvcf file1.vcf.gz --gzdiff file2.vcf.gz --diff-site --remove-indels --out filename*.

To retrieve the total number of variants and the number of SNV and InDels in each file the following commands were used:
vcftools –gzvcf file.vcf.gz --keep-only-indels --out filename and *vcftools –gzvcf file.vcf.gz --remove-indels --out filename*

3.4 Results

3.4.1 Two platform comparisons for Cohorts A, B, C, and D

WES data in Cohorts A (n=88) and C (n=128) were analysed and compared for RD-Connect with CSA and *seqr*, respectively. Variants passing the standardised filters in table 13 were included in the analysis. The mean number of variants from patient datasets in each cohort is shown in figures 12 and 13. Mean variant numbers are also given for variants in NMD genes. These figures show that the number of passing variants on the RD-Connect platform is significantly higher than for *seqr* and CSA. For passing variants in NMD genes, this remained significant only when comparing RD-Connect to *seqr*.

For each pair of platforms, concordance rates were calculated across the samples for variants falling in NMD genes. Concordance was also assessed for SNV and InDels separately. These are given in figure 14.

These figures show that mean concordance of variants for RD-Connect and CSA is 49%, and for RD-Connect and *seqr* is 34 % for the Cohort A and C, respectively. Concordance percentages are mostly accounted for by concordance in SNVs. InDel agreement averaged as low as 3%.

For Cohort C, a significantly higher number of InDels in NMD genes was present in the output report for RD-Connect compared to *seqr*. In this WES cohort, the total number of discrepant InDels was 796, for which RD-Connect called 727 (95%) and *seqr* 42 (5%). The passing InDels on the RD-Connect platform were further examined due to their significantly higher number (Figure 15). This revealed that these 727 InDels represented 104 unique positions only. Although intronic, these InDels were annotated by the RD-Connect pipeline as having a protein-altering feature.

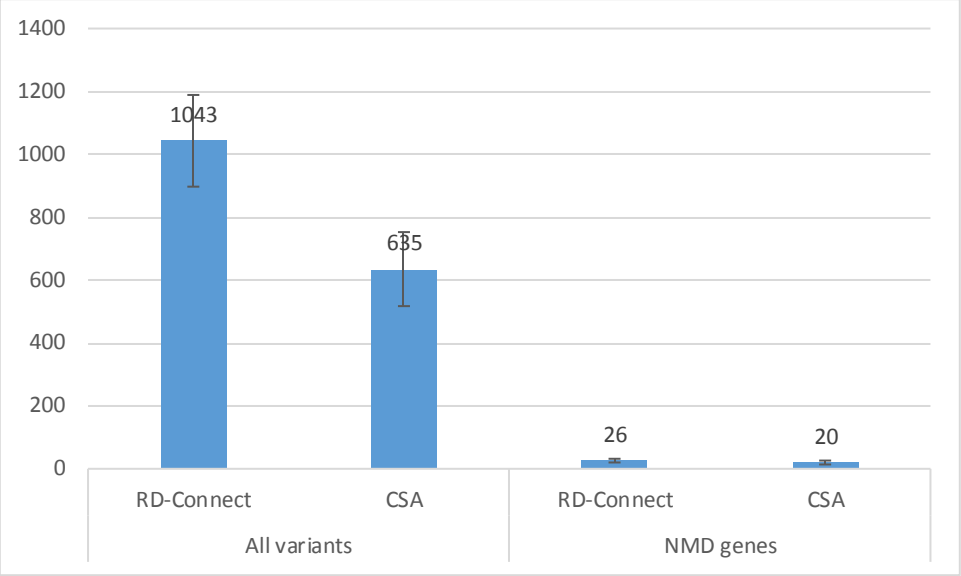


Figure 12: Mean number of variants in the output reports from RD-Connect and CSA for Cohort A (n=88)

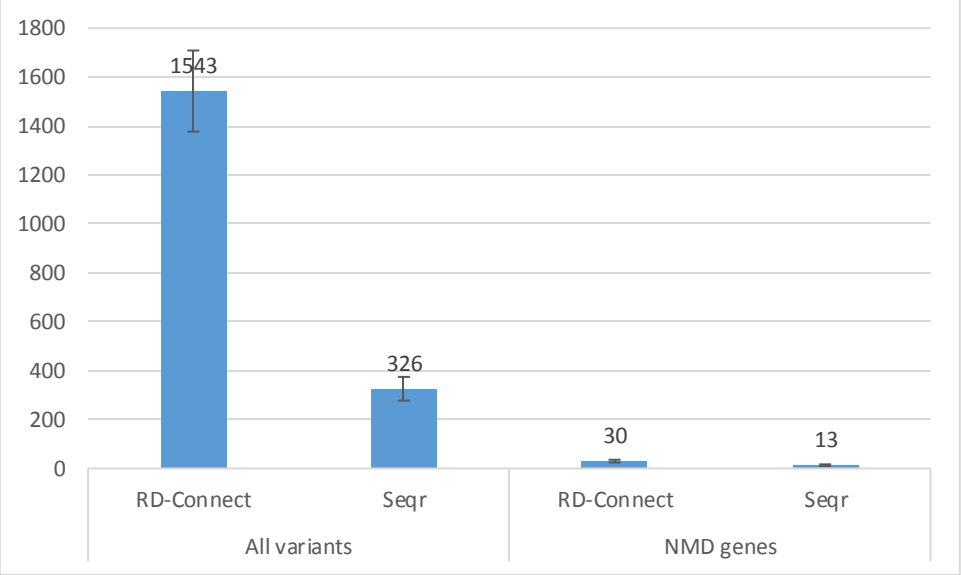


Figure 13: Mean number of variants in the output reports from RD-Connect and *seqr* for Cohort C (n=120)

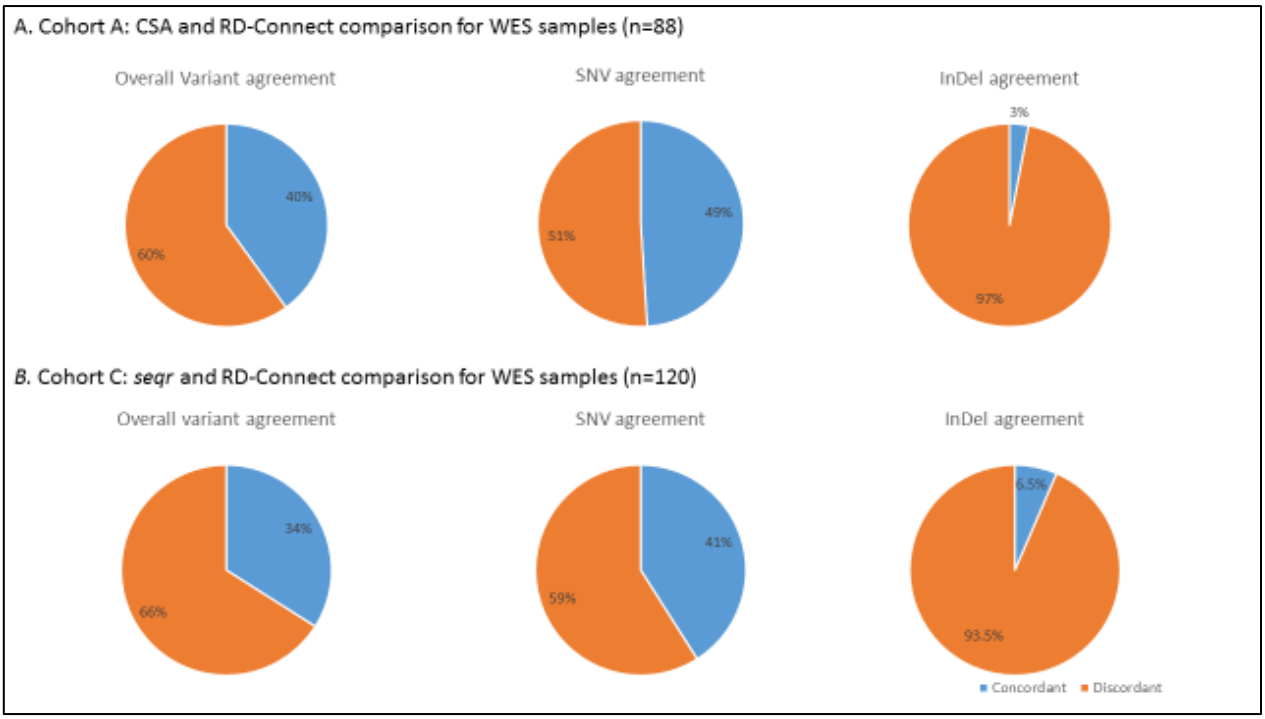


Figure 14: Two-platform variant output agreement for variants in NMD genes. A; Cohort A, B; Cohort C.

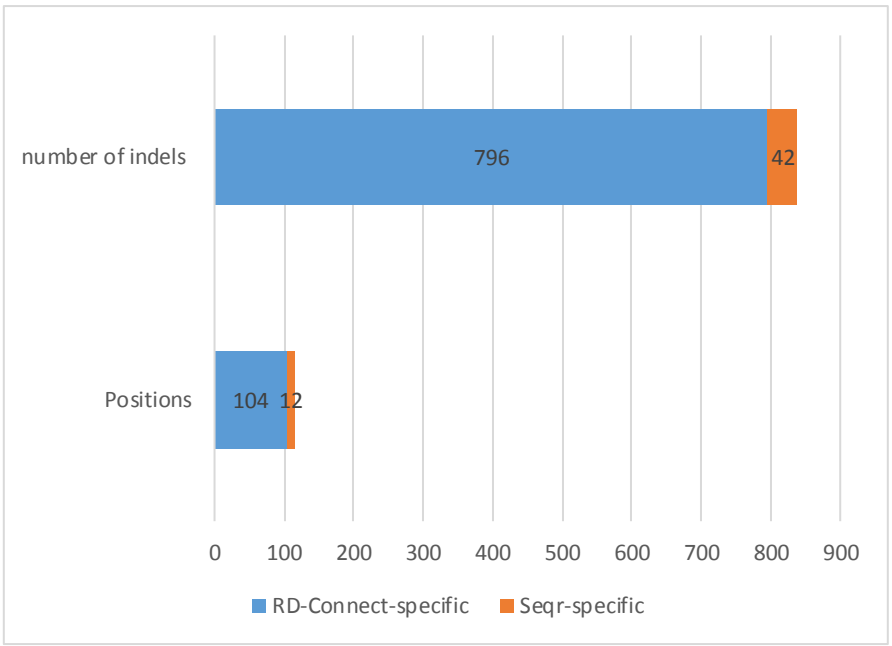


Figure 15: Discordant InDels for Cohort C samples in the RD-Connect and seqr output reports.

WGS data (Cohort B, n=30) was also compared for agreement on two platforms (CSA and RD-Connect) using standardised platform variant filters. Agreement between the platforms was extremely low and the two platforms only agreed on an average of 13% of variants. InDel agreement was 8%.

The mean number of variants produced by each sample in the cohort is shown in figure 16. Concordance rates for all variants, SNVs and InDels are given in figure 17. Variant output on the RD-Connect platform had 5 times more the number of variants in the output for the same patients from the CSA platform. The higher number of variants was also significant for those falling in NMD genes only.

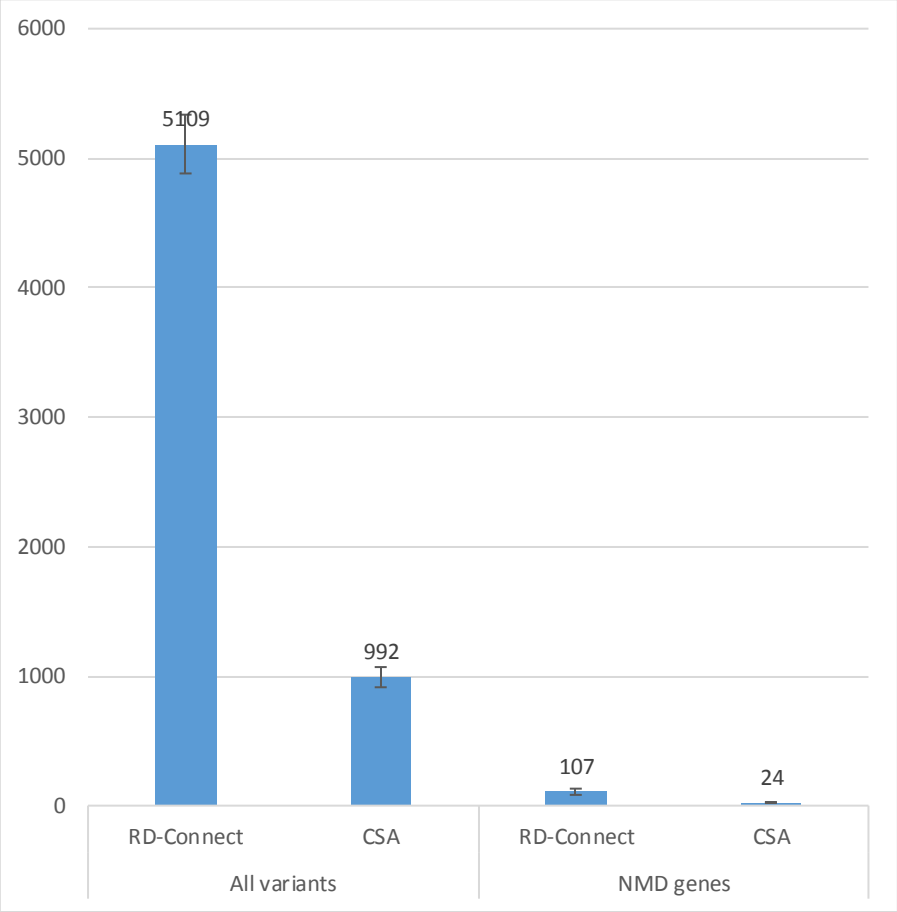


Figure 16: Mean number of variants for Cohort B WGS samples in the output report for CSA and RD-Connect platforms (n=30).

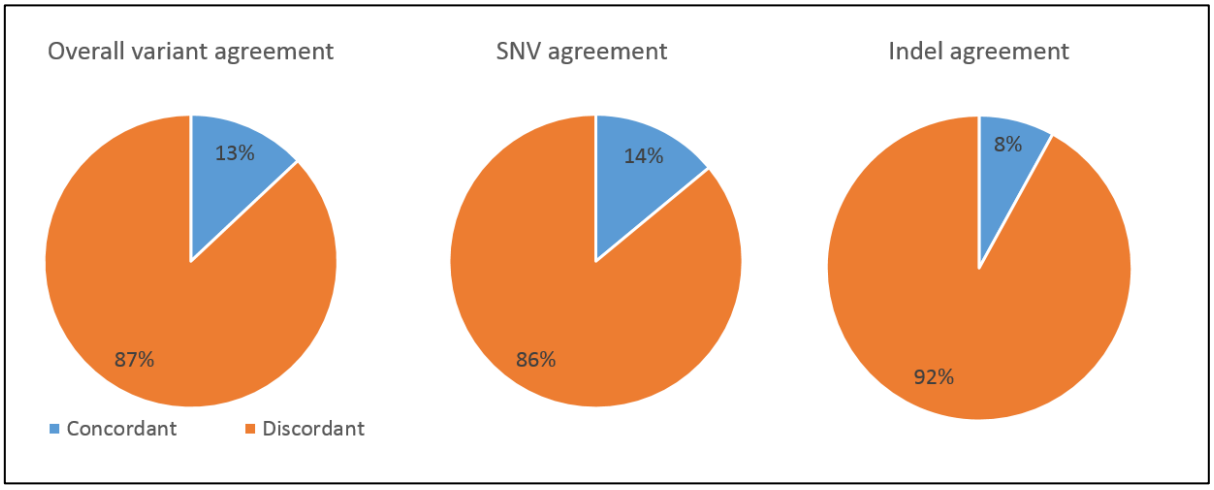


Figure 17: Variant agreement rates (concordance) between CSA and RD-Connect for Cohort B WGS samples (n=30).

For Cohort D, a small number of solved cases (n=27) were used to assess agreement of the *seqr* and RD-Connect platforms in proposing correct disease causing mutations.
For all but three of the 27 cases, RD-Connect proposed the same variants as causative mutations as *seqr*. The variants reported as disease causing following analysis on the *seqr* platform included: a one base pair (bp) frameshift insertion in exon 5 of the ANO5 gene (chr11:22242646), an intronic SNV in exon 9 of the CAPN3 gene (chr15:42695919), and a 9 base inframe deletion in intron 4 of the *SGCA* gene (chr17:48246421). The RD-Connect pipeline annotated these variants as having low impact on the protein. Annotations of these variants was similar for the Broad Institute's pipeline. On the seqr platform, these three mutations did not appear in the output report when the variant search was limited to those with moderate to high impact on the protein.

3.4.2 Three-platform WES comparison

WES Data from Cohort E were processed through the bioinformatics pipelines at the three sites and the variant calling files (VCF) were used to assess agreement of the pipelines in variant calling. The VCF files were then uploaded onto their respective platforms. Variant filtering was standardised on the platforms and the variants that appeared in the output report for each patient were compared for agreement between all three platforms.

a. Bioinformatics pipeline comparison

Figure 18 shows the average number of variants in each set of VCF files from RD-Connect, deCODE Genetics and the Broad Institute for patients in Cohort E.

The variant content in the VCF files was used to calculate pairwise concordance rates. This was calculated for the total number of variants from each platform and for SNV and InDels separately. Figure 19 illustrates the average concordance percentage between the site – specific VCF files in a pairwise manner. Figures 20 shows the mean number of variants that appear in two VCF files for each patient in the cohort.

This analysis revealed that there is a significant difference in the number of variants called by each of the three bioinformatics pipelines and that, while deCODE Genetics calls the highest number of variants, RD-Connect calls significantly more InDels than the other two platforms. The *seqr* pipeline called the lowest number of variants for this cohort.

Mean concordance for all variants called was highest for CSA and RD-Connect VCFs (75%). For all pairwise analyses, concordance of variants was mainly accounted for by SNV, as concordance for InDels was significantly low across all three platforms (figures 19 and 20).

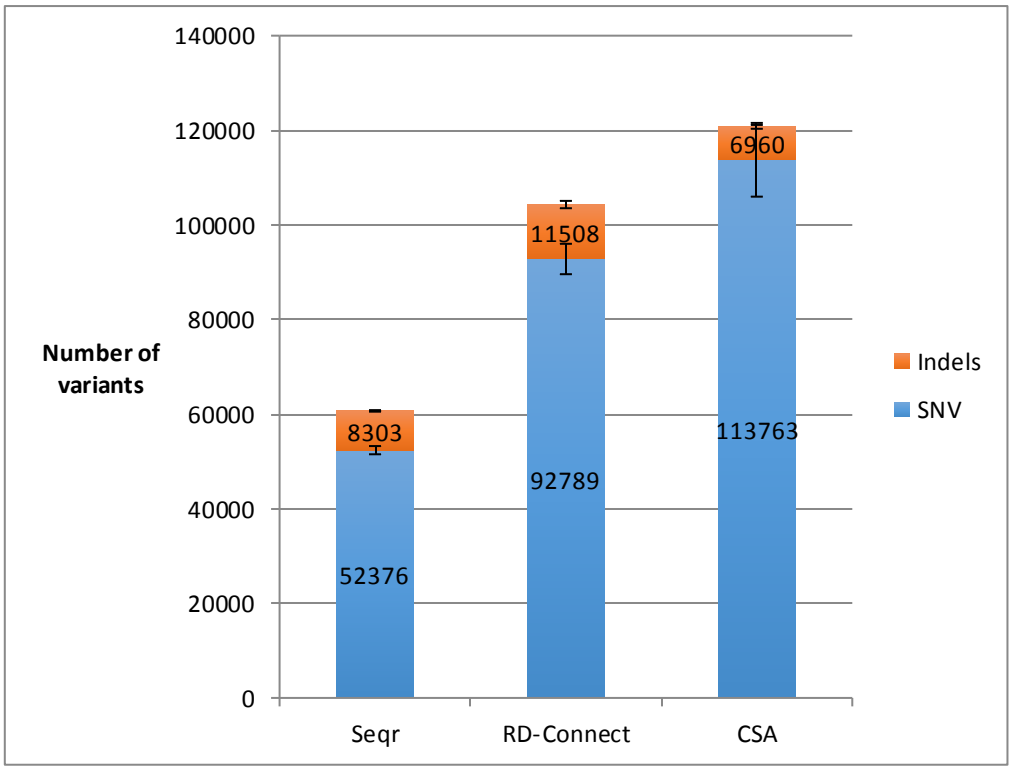


Figure 18: Mean number of variants in VCF files for Cohort E from the Broad Institute, RD-Connect and deCODE Genetics bioinformatics analysis pipelines.

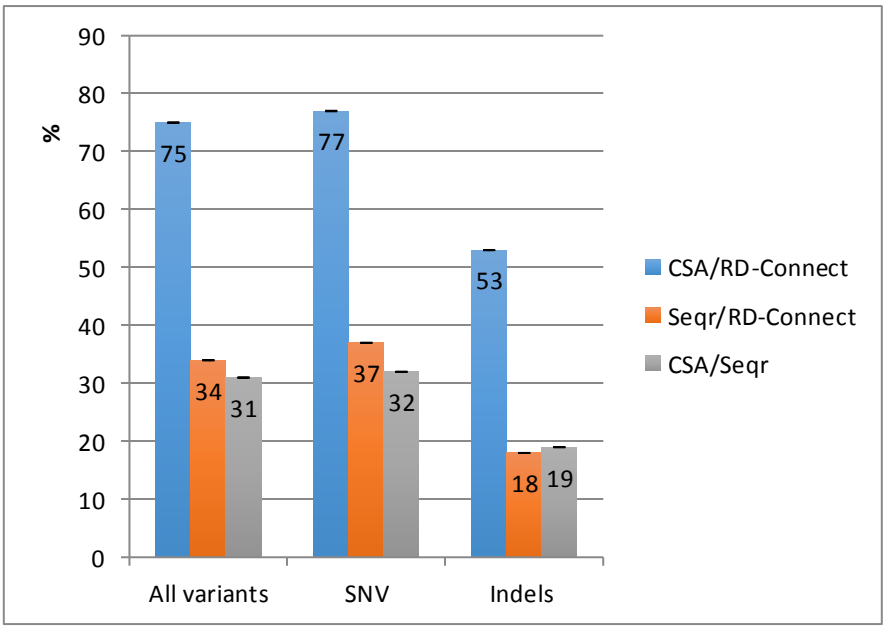
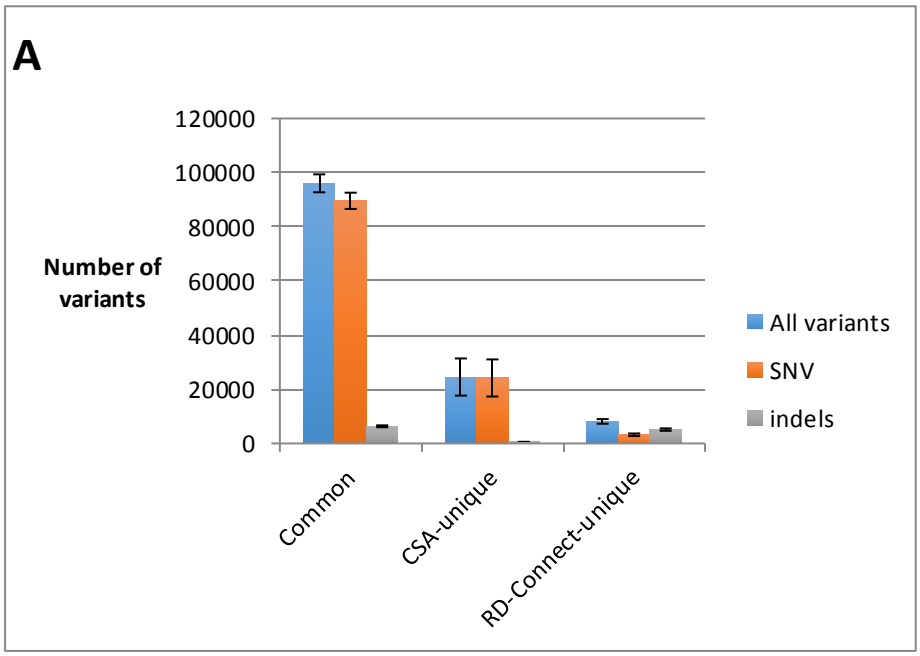
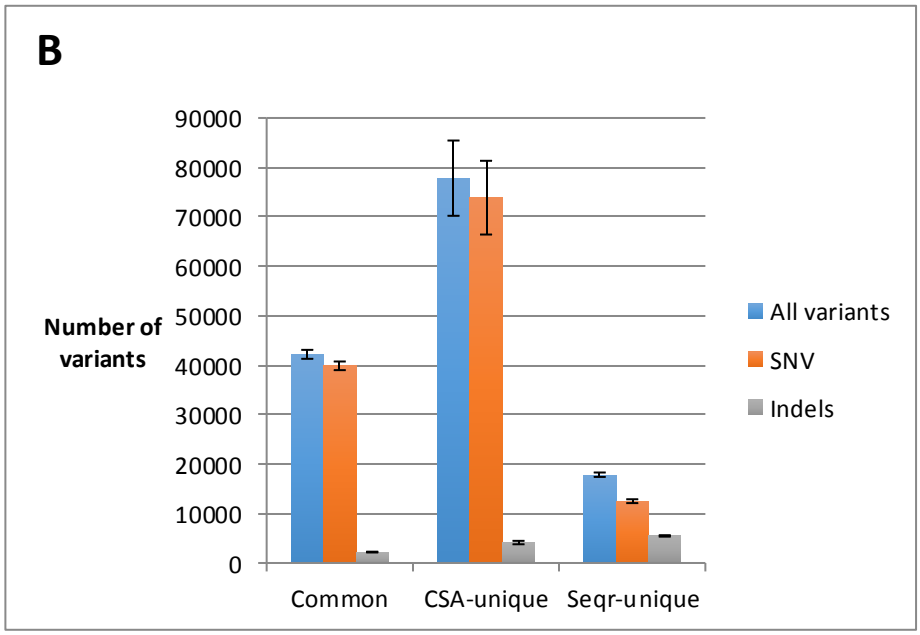
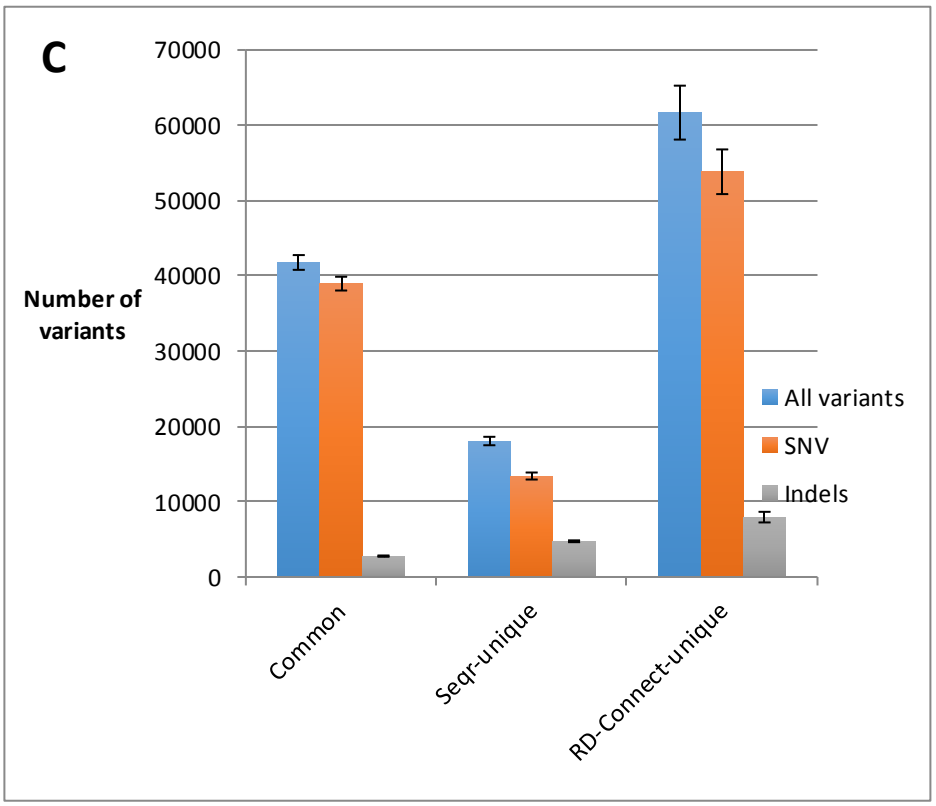


Figure 19: Mean pairwise variant concordance percentage in VCF files for Cohort E from the Broad Institute, RD-Connect and deCODE Genetics bioinformatics analysis pipelines.

Figure 20: Pairwise agreement of Cohort E (n=9) VCF samples shown in variant number for VCF files from deCODE Genetics and RD-Connect (A), deCODE Genetics and the Broad Institute (B), and RD-Connect and the Broad Institute (C).







b. Platform variant output comparison

Using the filtering options specified in table 13, variant output for Cohort E samples was analysed on each of the three platforms. The mean number of all variants and variants in NMD genes in the output reports are as shown in figure 21.

NMD gene variants were further examined. Overall, locus concordance for the three platforms averaged at 36.6% for all variants. With regards to pairwise concordance, CSA and *seqr* appeared to have the highest mean percentage concordance. However, this showed significant variation across this cohort as illustrated per sample in figure 22.

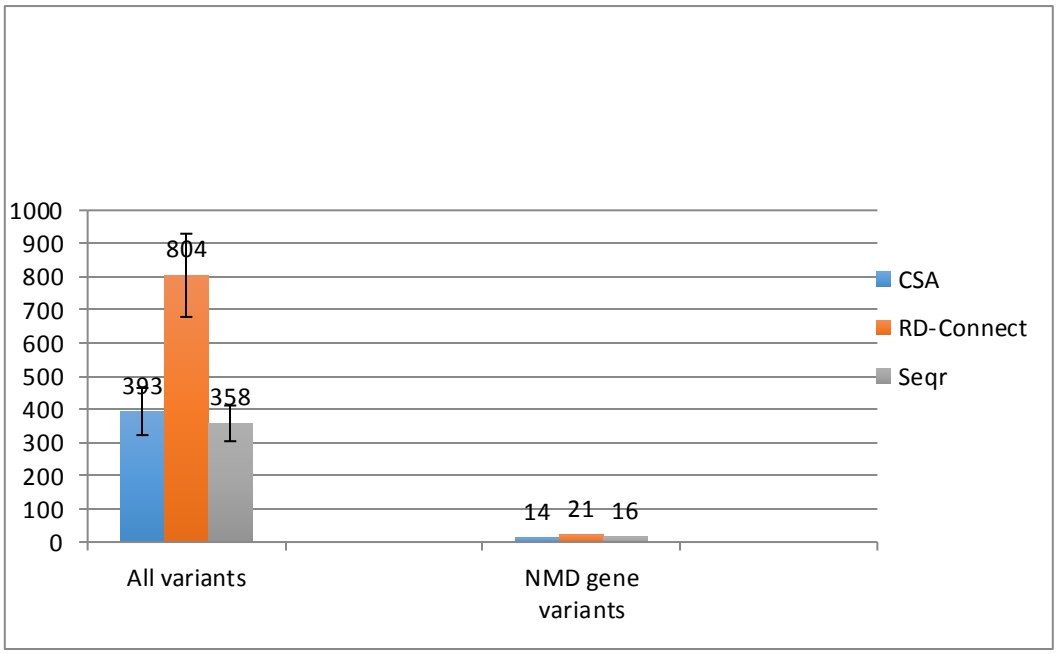


Figure 21: Mean number of variants in the output reports for CSA, RD-Connect and *seqr* for Cohort E WES samples. Standardised filters were applied on all three platforms.

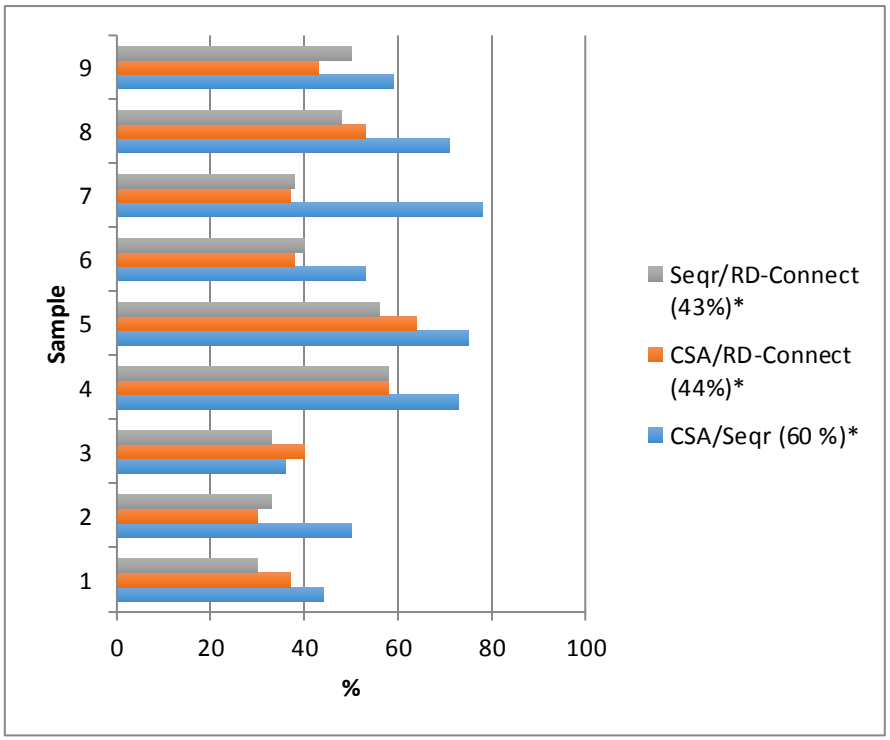


Figure 22: Pairwise agreement for variants in NMD genes in the output reports for Cohort E WES samples on the RD-Connect, CSA and *seqr* platforms. * Mean percentage agreement from all nine samples. NMD, neuromuscular disease; CSA, Clinical Sequence Analyzer.

3.4.3 Reference genome (GIAB) comparison on all three platforms

a. Bioinformatics pipeline comparison

VCF files produced for the NA12878 reference sample from RD-Connect, the Broad Institute and deCODE Genetics were compared using VCFtools. Total numbers of variants in each VCF file are shown in figure 23. VCF files were compared with the reference file and with one another (figures 24 and 25).

For the NA12878 sample, overall pairwise variant agreement between the three pipelines was as high as 91%, and 81% for agreement with the reference file. Variant agreement for InDels was comparatively low.

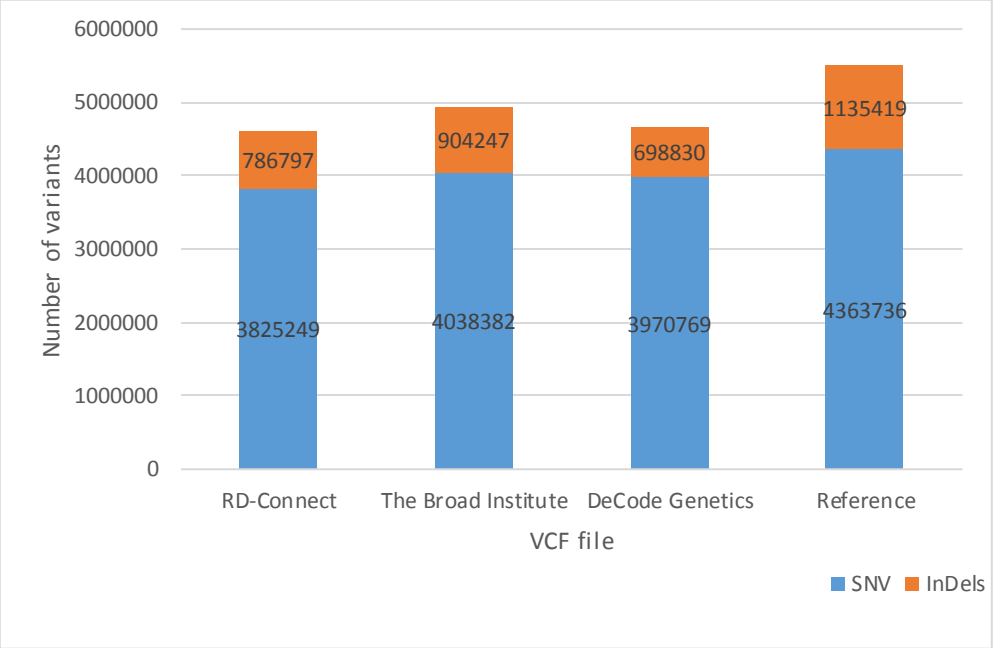


Figure 23: Number of variants in VCF files for the NA12878 WGS sample processed by RD-Connect, the Broad institute, deCODE Genetics and in the reference file.

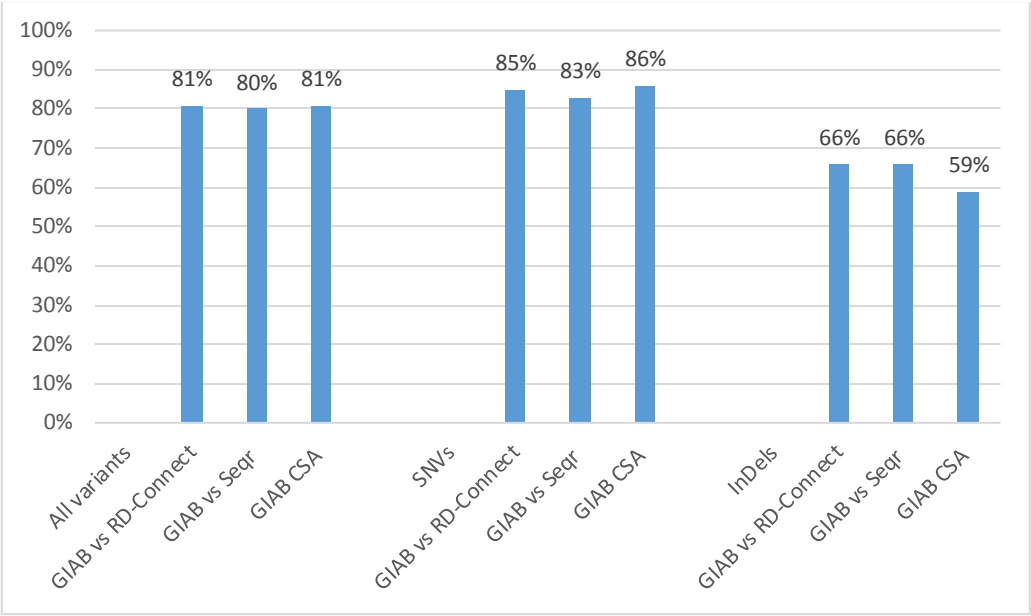


Figure 24: Variant agreement between the NA12878 VCF reference file and VCF files from RD-Connect, the Broad institute and deCODE Genetics.

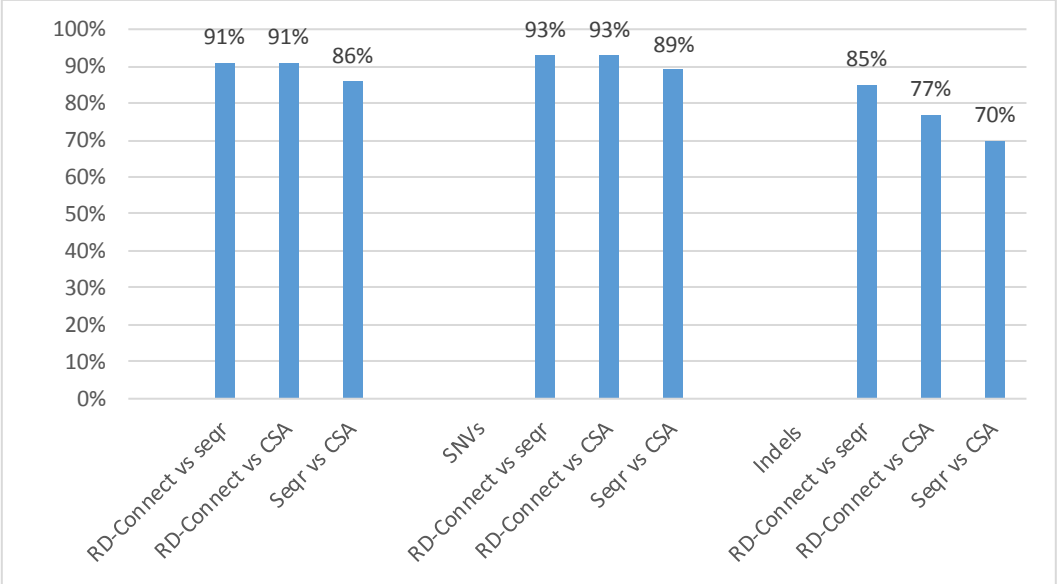


Figure 25: NA12878 variant agreement between VCF files from RD-Connect, the Broad institute and deCODE Genetics.

b. Platform output comparison

Using data for the NA12878 sample from the GIAB reference material, variant output on RD-Connect, *seqr* and CSA was examined using standardised filters (table 13).

The number of variants from each platform is shown in figure 26. NMD gene variant agreement is shown in figure 27.

Variant output on the RD-Connect platform contained a significantly higher number of variants. In addition, pair-wise variant output agreement was lower for comparisons involving the RD-Connect platform and was highest at 29% for *seqr* and CSA. Overall, all three platforms only agreed on 5% of variants.

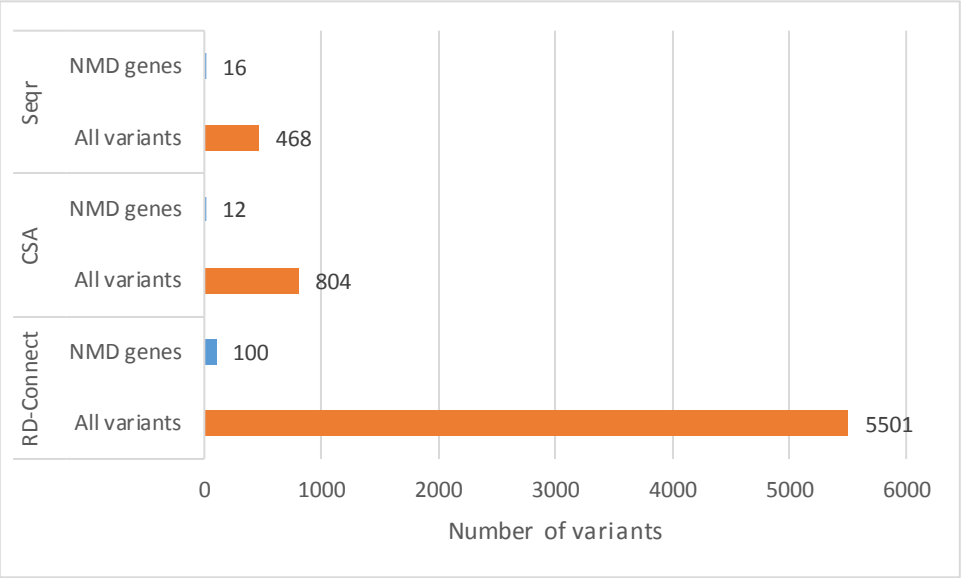


Figure 26: Number of variants for the NA12878 reference sample on the RD-Connect, CSA and *seqr* platforms using standardised variant filters.

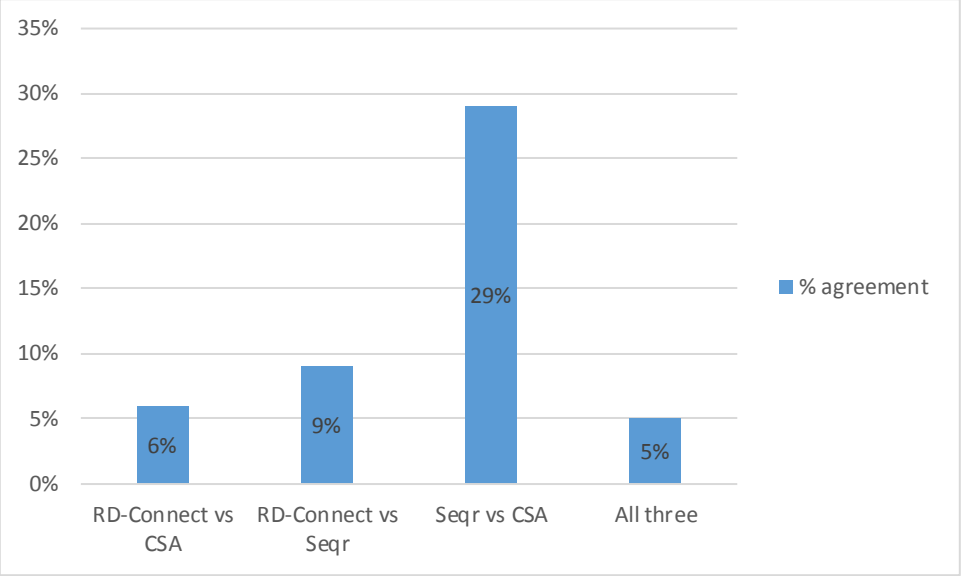


Figure 27: NMD genes variant agreement for the NA12878 sample on the RD-Connect, CSA and seqr platforms using standardised variant filters.

3.5 Discussion

Real WES and WGS data from patients with previously undiagnosed rare NMD was used to compare three bioinformatics pipelines at RD-Connect, the Broad Institute and deCODE Genetics and their respective genomic analysis platforms, the RD-Connect Phenome-Genome Analysis Platform, *seqr* and CSA. Filters were standardised for variant quality, population frequency and impact on protein on all three platforms.

WES data from two cohorts of patients (Cohorts A and C) were used assess pairwise agreement between the platforms. The RD-Connect platform had the highest number of variants in the output report and *seqr* had a significantly lower number. Variant agreement was low at 34% and 40% for RD-Connect and *seqr* and RD-Connect and CSA, respectively. Agreement was mainly accounted for by SNVs as InDel agreement was less than 10%.

When using WGS data (Cohort B) for patients with NMD, the discrepancies were magnified. The differences in variant numbers were greater, where the mean number of passing variants on RD-Connect was five times more that from CSA and the difference remained significant for passing variants in NMD genes. In addition, variant agreement between the two platforms was low (13%).

Comparison of WES samples (Cohort E) on all three platforms revealed significant discrepancies in variant numbers and concordance. The RD-Connect platform had more than twice the number of passing variants as the other two platforms. For NMD genes, RD-Connect still had the highest number of variants although the difference was not statistically significant. For these genes, variant agreement was highest for *seqr* and CSA (60%) although this was very variable across the sample. For passing variants falling in NMD genes, all three platforms only agreed on 36.6%.

Examination of the bioinformatics pipelines was carried out by comparing VCF files for the same patients that were processed at the three sites. This analysis also revealed discrepancies in overall VCF variant numbers and agreement. The mean number of variants in the VCF files from deCODE Genetics was higher than the other two sets of VCF files. The Broad Institute's pipeline called a significantly lower number of variants compared to the pipelines at the other two sites.

All three pipelines use a version of BWA to map reads to the human reference genome (build hg19). For variant calling RD-Connect and *seqr* use the GATK HaplotypeCaller while CSA uses a previous version in the GATK pipeline, UnifiedGenotyper. GATK developers recommend use of the HaplotypeCaller in their recent versions of the workflow and claim it to be as effective as the UnifiedGenotyper in calling SNV, but far more superior at calling

InDels (TheBroadInstitute). This may be represented in Cohort E exomes by the significantly fewer InDels called by deCODE Genetics' pipeline when compared to those called by the Broad Institute and RD-Connect.

It is also important to note here that for the *seqr* platform, the Broad Institute incorporates joint variant calling in the GATK HaplotypeCaller pipeline (Lek *et al.*, 2016). Joint calling is when variants (SNV and InDels) are called simultaneously from all BAM files in a cohort, and generate a single call file for the entire sample. Joint calling is proposed to have greater sensitivity for low frequency variants, where the whole cohort of samples is used to ascertain and call a variant with poor coverage. This variant would otherwise be filtered out when applying read depth and quality filters, however, is rescued by other samples in the cohort having a variant with adequate coverage at that position. In addition, joint calling is also set to reduce false positive calls by applying variant quality filtering uniformly across the whole sample. Joint calling also comes with the drawback of cost, computational complexity and the need for large hardware space. In addition, joint calling requires all cohort samples to be available at the same time which is not feasible in many diagnostic and research labs. This also means that addition of any sample(s) at a later stage would require the whole variant calling process to be repeated and that results may vary for each patient depending on the cohort their sample is processed with (TheBroadInstitute; Lek *et al.*, 2016). For Cohort E samples, the higher agreement of the RD-Connect and deCODE Genetics VCF files may provide some confidence in the variants called by these two pipelines. This may also suggest that, while joint calling used in the Broad Institute's pipeline is proposed to reduce false positives, it may also filter out real but low quality or rare variants that are not supported by the rest of the cohort.

Also worth noting, while genome alignment and variant calling are considered crucial steps in any sequencing experiment, further processing steps have also been found to impact the final set of variants called. GATK UnifiedGenotyper InDel calls have been found to be more sensitive to post-alignment processing steps, while these had little or no effect on pipelines using the GATK HaplotypeCaller (Tian *et al.*, 2016). This may also account for the discrepancy in the number of InDels called by all three pipelines and by CSA/deCODE Genetics specifically.

Overall, using the platforms to analyse Cohort E samples showed varying concordance and this was inconsistent with site-specific VCF file concordance rates for these samples. Platform NMD variant agreement rates were lower than overall variant numbers in VCF files

and although RD-Connect and deCODE Genetics's VCF file had higher agreement, the CSA platform output agreed more with *seqr*. These findings point towards an important role for annotation and filtering algorithms and gene prioritization methods in determining the final variant output from the genomic platforms.

A further observation from the VCF files for these nine samples is that the number of InDels called by the RD-Connect pipeline is significantly higher than the other two sites. In addition, there is a significant high number of passing InDels on the RD-Connect platform that are not supported by CSA or *seqr*. The RD-Connect-specific InDels in the original VCF file require validation. However, further examination of the discordant passing InDels on the RD-Connect platform for Cohort C exomes revealed that the InDels were incorrectly annotated as having a protein-altering feature and a moderate impact and thus, were not filtered out and appeared in the final variant output. As mentioned above, RD-Connect used the SNPEff (Cingolani *et al.*, 2012) as the annotation tool in the pipeline while the other two sites use Ensembl VEP (McLaren *et al.*, 2016). Findings of this analysis as well as those by other researchers at Newcastle University were reported back to the RD-Connect bioinformatics pipeline developers and plans are in place to replace SNPEff with the VEP in the pipeline (personal communication, Steve Laurie and Ana Topf).

As mentioned above, annotation tools perform differently depending on the transcript set used as a reference (McCarthy *et al.*, 2014). However, this was not an issue here as all three pipelines used the Ensembl transcript set. Therefore, annotation discordance is mostly related to the annotation software itself. A recent study (Yen *et al.*, 2017) compared annotation from SNPEff and Ensembl against those in ClinVar. The authors found that concordance between the tools was very high (99.5%) for SNV annotations in coding regions. However, concordance for InDels was significantly lower (<90%). The authors also compared both annotation tools using a large somatic mutation dataset (COSMIC). They found that in the case of somatic mutations, agreement was lower for SNV (<90%) and substantially lower for InDels (<15% for insertions mutations). It is also important to note here that half of the SNPEff errors were found to be due to random right-shifting of the variant. This error has been corrected in the newer version of SNPEff (4.2) (Yen *et al.*, 2017). These discordances are likely to contribute to discordances in output reports when filtering to prioritise variants in the context of rare disease. For the small sample of solved cases (n=27) in Cohort D, RD-Connect and *seqr* proposed the same candidate mutations. For this cohort, three variants were annotated as having a low impact by both platforms. However, due to the clinical relevance of the genes, they were further investigated and later assigned as pathogenic mutations.

Using WGS data for the reference GIAB NA12878 sample, the bioinformatics pipelines had a relatively high concordance with each other (approximately 90%) and with the reference VCF (approximately 80%). These results are comparable with ones from a recent study using the NA12878 reference sample to compare six combinations of read aligners and variant callers. They found that all pipelines performed well in variant calling particularly for calling SNVs (sensitivity >99%). InDels were more problematic for all pipelines (mean sensitivity 87.4%) (Laurie *et al.*, 2016).

Overall variant numbers in all three VCF files were not significantly different. The previously evident role of joint calling in the Broad's pipeline in eliminating noise and reducing variant numbers was no longer evident using the high quality calls in the NA12878 sample. This can be explained by the fact that the variants no longer need other samples in the joint calling process due to their high quality.

The high variant agreement of pipelines when using a reference set of high confidence variants contradicts findings of the analysis performed using real patient WES data. This highlights that variant calling is sample and project-specific and that errors may occur prior to the bioinformatics analysis stage. Thus, sample quality, sequencing platform, sequence coverage, in-house pre-alignment processing of the raw data and variations in quality control parameters may play a role in discrepancies between the bioinformatics pipelines and the final variant output (Clark *et al.*, 2011; Alioto *et al.*, 2015; Lelieveld *et al.*, 2015).

In addition, despite the high agreement of the VCF files reflecting the high performance of bioinformatics pipelines when using the reference NA12878 data, there was very low agreement when the VCF files were analysed on the corresponding three platforms. For variants in NMD genes, all three platforms only agreed on 5% of variants. This suggests a high contribution from variant annotation and filtering algorithms to the discrepancies between the three genomics platforms.

It is important to note here that NGS projects are usually designed for a specific purpose. For example, an NGS workflow designed for diagnostics may not be suitable for rare disease research and for the latter the pipelines are usually customised for the nature of disease and population under investigation and to overcome known sequencing challenges and artefacts. In addition, default parameters for a particular pipeline may change depending on the project. For example, imputation, computational linkage and pedigree incorporation and association tools maybe integrated into a pipeline for a project investigating disease in families. This customisation is proposed to increase the sensitivity and yield of the pipelines (Wittig *et al.*, 2015; Chung *et al.*, 2016; Peng *et al.*, 2016; Zucca *et al.*, 2016; Blauwendraat *et al.*, 2017).

In conclusion, it is clear that variations in bioinformatics tools, annotation software and filtering algorithms lead to discordances in variant outputs. This must be considered in the design and development of NGS projects particularly in the field of rare diseases when highly accurate and sensitive data is necessary for novel discoveries. Research into somatic mutation detection suggests that a workflow that intersects and combines data from more than one combination of aligner and variant caller in addition to intersecting calls from the same variant caller but different aligners, improves performance and increases call sensitivity and specificity (Chung *et al.*, 2016). This approach may also be appropriate for rare genetic disorders such as NMD.

Moreover, research involving comparisons of NGS data processing, alignment tools and variant callers is extensive and has led to impressive improvements. However, performance comparisons between annotation tools and platform filtering algorithms remain limited. In addition, although in-house sample preparation and DNA sequencing protocols are standardised, these vary between research and diagnostic labs and may bias comparisons of results from different sites. Furthermore, a set of high confidence variant calls is important for benchmarking bioinformatics pipelines in the development stage. However, use of real patient data may also provide valuable insights for assessment and development of the bioinformatics pipelines and variant prioritisation methods.

Chapter 4. Genomics platform utility in WES analysis in patients with limb girdle weakness.

4.1 Introduction

With the immense data produced through WES, differentiating sequence variants that are pathogenic from those that are polymorphisms is an ongoing challenge. This is further complicated by the computational complexity of bioinformatics algorithms used in data analysis and variant prioritization. Patient sequence data needs to be linked to phenotypic data, inheritance patterns, population frequencies and bioinformatics prediction tools. In addition, integration of databases for genomic features, gene expression, RNA sequencing data and protein interaction networks is essential. All this information needs to be incorporated in a defined yet flexible workflow accessible to researchers via an on-screen interface. The aim of this is to enable researchers to visualise, manipulate and filter sequence data to prioritize the most relevant variants for further investigation (Coonrod *et al.*, 2011; Jalali Sefid Dashti and Gamieldien, 2017).

LGMDs are a genetically and phenotypically heterogeneous group of disorders with limb girdle weakness being a common feature. Many of the LGMDs have additional features that are non-distinguishing and overlap with other LGMDs as well as other myopathies and neuropathies. MRI and histopathology have been traditionally used to guide molecular testing for LGMD. Sequential gene testing by Sanger sequencing in this group has led to lengthy diagnoses and low diagnostic yields (Lo *et al.*, 2008). Although yield is higher for gene panels, it depends on the genes included in the panel and on the nature of the patients tested. In addition, a gene panel may not account for other phenotypically overlapping disorders and for novel genes (Ghosh and Zhou, 2012; Efthymiou *et al.*, 2016).

WES in an effective data workflow and a user-friendly interface offers the opportunity to provide diagnosis and novel gene discoveries.

WES was performed on patients presenting with limb girdle weakness and suspected to have LGMD. These patients have been extensively investigated prior to WES and remained without a genetic diagnosis. The WES data was analysed using the *seqr* platform and a study of this cohort is presented here as an example of application of an integrated genomics platform in a cohort of patients with rare NMD.

4.2 Aims

- Study the genetic aetiology of undiagnosed patients with limb girdle weakness using WES.
- Identify the genetic cause in undiagnosed cases and propose novel candidates for NMD.
- Assess the added benefit of using and integrated genomics platform in analysis of WES data for patients with a rare NMD presenting with limb girdle weakness.

4.3 Patients and methods

4.3.1 Consent

Informed consent for research was obtained from all patients undergoing WES under the protocol for Newcastle MRC Centre Biobank for Neuromuscular Diseases (REC reference: 08/H0906/28 + 5).

4.3.2 Patients

Patient were selected for WES if they presented with limb girdle weakness and had been extensively screened for mutations in known genes as dictated by their phenotype. Recruitment occurred during 2014-2015. Informative family members were included where relevant and available.

4.3.3 Sequencing and bioinformatics pipeline

WES samples were sequenced at the Broad Institute of Harvard and MIT on an Illumina HighSeqXs platform. The capture kit used was the Agilent Sure-Select Human All Exon v.2.0 (44Mb). Read alignment was performed using BWA, variant calling via GATK, and annotation using VEP.

4.3.4 Platform analysis

Analysis of WES data was carried out on the *seqr* platform. Initially, variants were filtered for those that fit an inheritance model suggested by the family pedigree, allele frequency <0.01 in the ExAC and 1000 Genomes population, variants that were predicted to have a damaging effect on the protein and those that were in the NMD gene list (appendix A). Predicted damaging variants include nonsense, frameshift, essential splice site, in-frame and missense

variants. Any potential candidate variants were assessed further for their population frequencies and for missense variants, pathogenicity prediction (as given by Sift, Polyphen2, Mutation Taster and FATHMM tools on *seqr*). These variants were then also evaluated according to phenotype, reports of patients with the same variant or other variants in the same gene, and with tissue-specific expression data provided via GTEx (https://gtexportal.org) and integrated onto the *seqr* platform. Furthermore, *seqr* provides direct web links to PubMed, OMIM, NCBI Gene and the Protein Atlas and these were used to further examine the relevance of the variant as a disease candidate.

If the initial analysis did not propose any candidates, the analysis was repeated with less stringent filters for inheritance model and effect on protein and expanded to include the whole exome.

4.3.5 Evidence for pathogenicity

Where relevant and available, candidate diagnoses where evaluated with other test results such as MRI and muscle histology results and by Sanger sequencing and segregation studies in affected and non-affected family members.

For novel candidate genes, species conservation data, intolerance to loss of function mutations, association with disease, and tissue-specific expression were examined amongst other features using the online tools and databases in table 14.

Finally, candidates were discussed with a multidisciplinary team to accept or disprove the variant as a causative mutation.

Table 14: Online tools and databases used to gather evidence for variant pathogenicity and association with NMD.

Tool/Database	URL	Supporting evidence provided
UCSC Genome browser (hg19)	https://genome.ucsc.edu/	Conservation data. Variant localization. Genomic features.
ExAC Browser	http://exac.broadinstitute.org/	Allele and genotype frequencies. Loss of function mutation metrics.
Pubmed	https://www.ncbi.nlm.nih.gov/pubmed/	Gene function. Disease association.
OMIM	https://omim.org/	Disease association.
EMBL-EBI Expression Atlas GTEx portal	https://www.ebi.ac.uk/gxa/home https://gtexportal.org/home/	Tissue-specific gene expression.
STRING database	https://string-db.org/	Protein interaction network.

4.3.6 Segregation studies

Potential pathogenic variants were segregated in family members via Sanger sequencing. Primer design, primers, PCR and sequencing protocols are detailed in the appendices.

4.4 WES results

4.4.1 Patients

Patients were selected for WES when prior genetic screening of candidate genes did not reveal a molecular diagnosis. WES was performed on 178 individuals from 93 families with 125 affected with a NMD and 53 unaffected relatives. Seventy were single cases with no relatives undergoing WES and one family consisted of the unaffected parents only.

4.4.2 Molecular diagnosis in patients presenting with limb girdle weakness through analysis of WES data on the seqr platform.

Prior to the start of this project, WES data for this cohort was reviewed by team members at Newcastle University. Out of the 93 families, 33 (35.5%) had candidate variants proposed as causative, of which three were in novel unpublished genes. These are described in more detail in table 15.

WES data from the remaining 60 families were analysed and a molecular diagnosis was proposed for an additional 28 families, including variants in 11 novel candidate genes. Overall, 61 families (65.6%) had a proposed genetic diagnosis. This included 47 families with a proposed diagnosis in known NMD genes and 14 families with new candidate genes (77% and 22.6% of all families with a proposed diagnosis, respectively). A pie chart of the distribution of molecular diagnoses is given in figure 28. The highest number of proposed causative variants were in the *TTN* gene (19.7%) followed by the COL6 genes (*COL6A1, COL6A2* and *COL6A3*).

Fifty out of the 93 families (53.8%) consisted of a single affected individual. Families where at least one additional affected or unaffected individual also underwent WES were slightly more likely to have a candidate molecular diagnosis than singletons (67.4% and 62%, respectively).

The variants presented here are the best candidates proposed through WES. Moreover, for many of these variants, pathogenicity is yet to be confirmed through family segregation analyses, deep phenotyping, and functional work for the gene and the proposed mutations.

Table 15: Proposed genetic diagnosis from WES for 33 families presenting with limb girdle weakness using the *seqr* genomic platform prior to the work presented in this thesis.

Family	Gene	DNA change	Protein change	Variant reported/ novel	Evidence supporting pathogenicity
6	*DMD*	c.8146C>T	p.Gln2716Ter	novel	De novo novel variant. Fits with phenotype
11	*SRPK3**	c.1146G>A	p.Trp382Ter	novel	Novel nonsense variant. Segregates. Variants in the same gene discovered in phenotypically similar cases in other NGS projects.
12	*MYH7*	c.5560-2A>G	ESS	novel	Damaging variant. Fits with phenotype.
14	*DNM2*	c.1115T>C	p.Phe372Ser	novel	Damaging variant. Fits with phenotype.
18	*COL6A1*	c.472G>A	p.Asp158Asn	novel	Damaging variants. Segregate. Fit with phenotype.
		c.958_961delGGAG	p.Gly320ArgfsTer13	novel	
26	*FKTN*	c.203delA	p.Asn69MetfsTer10	novel	Damaging variants. Fit with phenotype.
		c.1098T>A	p.Asp366Glu	novel	
27	*RYR1*	c.8137G>A	p.Asp2713Asn	reported	Damaging variant. Fit with phenotype.
		c.14614T>G	p.Cys4872Gly	novel	
28	*KLF15**	c.103C>G	p.Leu35Val	novel	Damaging variant. Gene highly expressed in skeletal and cardiac muscle.

					Variant segregates in family.
29	*MUSK*	c.1119G>T	p.Leu373Phe	reported	Damaging variants. Segregate
		c.1141G>T	p.Glu381Ter	novel	
35	*VCP*	c.329G>A	p.Arg110His	reported pathogenic	Fits with phenotype Segregates
36	*COL6A2*	c.1769C>T	p.Thr590Met	reported	Damaging variants. Fit with phenotype.
		c.2192C>T	p.Thr731Met	novel	
38	*TTN*	c.107377+1G>A	ESS	reported	Damaging variants. Fit with phenotype. Segregate.
		c.71783delA	p.Phe23928SerfsTer11		
				novel	
40	*TTN*	c.76440_76444delGCAGA	p.Leu25481HisfsTer7	reported	Damaging variants. Fit with phenotype.
		c.35828dupA	p.Glu11945ArgfsTer6	novel	
42	*COL6A1*	c.362A>G***	p.Lys121Arg	novel	Damaging variant. Fit with phenotype. Segregate.
44	*STIM1*	c.20G>A	p.Gly7Asp	novel	Damaging variant. Fits with phenotype
55	*LAMA2*	c.2049_2050delAG	p.Arg683SerfsTer21	reported	Damaging variants. Fit with phenotype.
		c.6992+5G>A	ESS	reported	
58	*LMNA*	c.746G>A	p.Arg249Gln	novel	Damaging variant. Fits with phenotype

59	*DNM2*	c.1679_1681delAGA	p.Lys562del	novel	Damaging variant. Fits with phenotype
64	*STIM1*	c.40A>G	p.Ser14Gly	novel	Damaging variant. Fits with phenotype.
65	*TTN*	c.107377+1G>A	ESS	reported	Damaging variants. Fit with phenotype.
		c.71043G>A	p.Trp23681Ter	novel	
66	*MAMDC2**	c.2008_2009delAC	p.Thr670AsnfsTer9	novel	Damaging variant. Gene highly expressed in skeletal muscle. Segregate. Second family with similar phenotype and mutations in the same gene.
69	*GMPPB*	c.860G>A	p.Arg287Gln	reported	Damaging variant. Fit with phenotype.
		c.338G>A	p.Cys113Tyr	reported	
70	*MTM1*	c.1054-2A>T	ESS	novel	Damaging variant. Fit with phenotype. Proposed manifesting female carrier.
71	*TTN*	c.107377+1G>A	ESS	reported	Damaging variants. Fit with phenotype.
		c.60709A>T	p.Lys20237Ter	novel	
72	*COL6A3*	c.7447A>G	p.Lys2483Glu	reported	Damaging variant. Fits with phenotype
76	*TTN*	c.77593C>T	p.Arg25865Ter	reported likely pathogenic	Damaging variants. Fit with phenotype.
		c.80103T>A	p.Asn26701Lys	reported	

78	*ANO5*	c.155A>G	p.Asn52Ser	reported with uncertain significance	Segregate. Fit with phenotype.
		c.191dupA	p.Asn64LysfsTer15	reported pathogenic	
80	*LAMA2*	c.611C>T	p.Ser204Phe	novel	Damaging variant. Fit with phenotype.
		c.4533delT	p.Gly1512AlafsTer83	novel	
82	*PIEZO2*	c.2136G>A	p.Met712Ile	novel	Fit with arthrogryposis phenotype. Damaging variant. Segregates.
88	*COL6A3*	c.5665G>C	p.Gly1889Arg	novel	Damaging variant. Fit with phenotype. Segregates.
90	*COL6A2*	c.541G>A	p.Glu181Lys	reported pathogenic	Pathogenic variant. Fits with phenotype.
92	*MYH7*	c.5533C>T	p.Arg1845Trp	reported pathogenic	Pathogenic variant. Fits with phenotype Segregates.
93	*TTN*	c.48312+2_48312+15delTGAGTTTGAGCAG	ESS	novel	Damaging variants. Fit with phenotype.
		c.1933G>T	p.Glu645Ter	reported	

*Novel gene in neuromuscular disorders. **Reported pathogenicity from ClinVar or reported in control populations in ExAC/gnomAD at a frequency <0.001 in a heterozygous state. All variants listed here are in a heterozygous state unless stated as homozygous (marked with***). ESS, essential splice site.

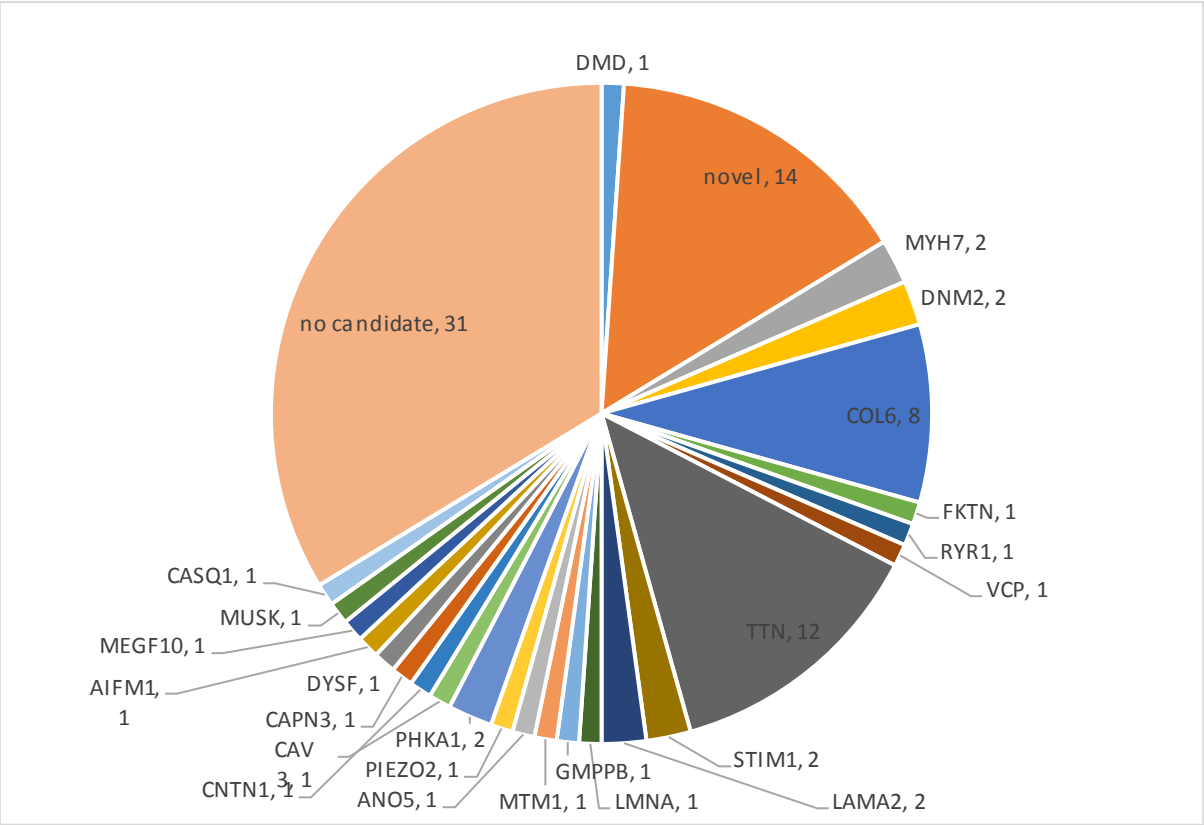


Figure 28: Number of families per candidate genes for a cohort with limb girdle weakness and WES data analysed on *seqr* (n=93).

4.4.3 Proposed genetic diagnosis as a result of this project

A. Variants in known NMD genes

As a result of the work presented in this thesis, a genetic diagnosis was proposed for 28 families following this analysis. For 17 of these families, variants were proposed in genes with a described association with NMD. These are further detailed in table 16.

Table 16: Families with proposed candidate mutations in known NMD genes identified through analysis of WES data on the *seqr* platform as a result of the work presented in this thesis.

Family	NMD gene	DNA change	Protein change	Variant novel/reported*	Evidence supporting pathogenicity
2	*CASQ1*	c.401G>A	p.Gly134Glu	novel	Damaging variant. Fits with phenotype.
24	*TTN*	c.5289T>A	p.Asp1763Glu	novel	Damaging variant. May fit with phenotype.
30	*TTN*	c.25561C>A c.36509A>T	p.Pro8521Thr p.Glu12170Val	novel reported	Damaging variants. May fit with phenotype.
31	*COL6A2*	c.1138C>T	p.Arg380Cys	novel	Damaging variant. Fits with phenotype.
32	*PHKA1*	c.890T>C	p.Leu297Pro	reported	Damaging variant. May fit with phenotype.
34	*TTN*	c.53789C>G	p.Pro17930Arg	novel	Damaging variant.
37	*CAV3*	c.136G>A	p.Ala46Thr	reported pathogenic	Fits with phenotype. Segregates.
46	*COL6A1*	c.957G>T c.1043C>T	p.Lys319Asn p.Ser348Leu	reported pathogenic reported	Fit with phenotype.

47	*CNTN1*	c.623A>T	p.Asn208Ile	novel	Damaging variant. CNTN1 mutations reported in AR lethal congenital myopathy in one family. AD inheritance proposed
48	*TTN*	c.35794G>T	p.Glu11932Ter	reported	Damaging variants. Fit with phenotype.
		c.28226_28227insA	p.Val9410GlyfsTer6	novel	
62	*TTN*	c.70947_70948insTTCC	p.Lys23650PhefsTer23	novel	Damaging variants. Fit with phenotype.
		c.61815A>G	p.Ile20605Met	reported	
67	*CAPN3*	c.759_761delGAA	p.Lys254del	reported pathogenic	Pathogenic mutation. Fits with phenotype.
74	*DYSF*	c.988G>C	p.Gly330Arg	reported	Fits with phenotype.
		c.2929C>T	p.Arg977Trp	reported pathogenic	
77	*AIFM1*	c.340G>A	p.Ala114Thr	reported	May fit with phenotype.
79	*TTN*	c.60223G>A	p.Val20075Met	novel	Damaging variants. Fit with phenotype.
		c.71705T>C	p.Ile23902Thr	reported	
81	*MEGF10*	c.352T>C	p.Cys118Arg	novel	Damaging variants. Fit with phenotype.
		c.1426+1G>T	ESS	novel	
91	*PHKA1*	c.3359T>A	p.Phe1120Tyr	novel	Damaging variant Fits with phenotype.

* Reported pathogenicity from ClinVar or reported heterozygous frequency in control populations (ExAC/GNOMAD) less than 0.001.

B. Variants in novel candidate NMD genes

For 11 families in this cohort variants in novel candidate genes were proposed to be causative. WES did not reveal any potential mutations in known NMD genes and the search was expanded across the whole exome. The families and the candidate genes are further described below.

a. MYMK

The patient (family 49) is a 70 year-old white male with juvenile onset limb girdle, proximal and distal myopathy. Clinical features are illustrated in Figure 29. Muscle weakness was mild, symmetric, slow progressing and more evident proximally. Facial muscle weakness, ptosis and lagophthalmos were also features of disease in this patient. In addition, he suffered with scoliosis, pectoralis muscle hypoplasia and cryptorchidism. The patient suffered severe gastrointestinal symptoms that led to food avoidance and weight loss. Progression of the myopathy has been unremarkable and at the age of 70 the patient remains ambulant and living independently.

Serum creatine kinase (CK) levels were mildly elevated (500-1000IU/L). EMG studies were suggestive of a chronic mildly active necrotising myopathy. Muscle MRI showed selective involvement in the lower limbs as shown in figure 30, and muscle histology showed non-specific myopathic changes (figure 31) (Alrohaif *et al.*, 2018).

Analysis of WES data for this patient revealed no pathogenic variants in known NMD genes. However, two compund heterozygous variants in the *MYMK* gene: c.271C>A (p.Pro91Thr) and c.553T>C (p.Cys185Arg) were identified. The proposed mode of inheritance was autosomal recessive.

The *MYMK* gene encodes the transmembrane protein 8 (TMEM8C), also known as myomaker. It is a highly conserved protein with muscle-specific expression during muscle development and regeneration (Millay *et al.*, 2013; Millay *et al.*, 2014). At the time of this analysis, the *MYMK* gene was not associated with disease. However, a recent publication described eight patients with phenotypic similarity to the patient described here and mutations in the *MYMK* gene (Di Gioia *et al.*, 2017). The gene is now associated with Carey-Fineman-Ziter syndrome, a congenital myopathy with facial weakness, cryptorchidism and gastrointestinal disturbance, amongst other consistent features. The patient described here is now also described in the literature, as mild case of Carey-Fineman-Ziter syndrome and mutations in the *MYMK* gene (Alrohaif *et al.*, 2018).

Figure 29: Clinical features in *MYMK*-related CFZS. (A) Top: Front and profile facial photos demonstrating lagophthalmos (left, patient attempting lid closure), muscle hypoplasia, retrognathia and broad nasal tip. Middle: Wasting of intrinsic hand muscles, contracture deformities of the right little finger and of the toes. Bottom: Scoliosis (left) and generalised muscle atrophy with pectoralis muscle hypoplasia (right). Reproduced with permission under CC-BY license, Wolter Kluwer.

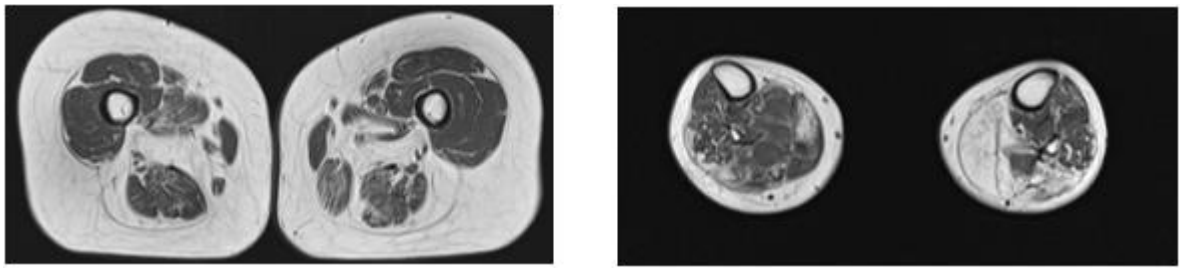


Figure 30: Muscle MRI images for a patient with *MYMK* gene mutations (Alrohaif *et al.*, 2018). A: MRI T2 weighted images of the thighs (left) showing severe fatty replacement of hamstrings, thigh adductors and sartorius muscles, with relative sparing of the gracilis and quadriceps muscles bilaterally, and of the calves (right) showing asymmetric involvement with more marked fatty replacement in muscle of the right leg. Gastrocnemius, and soleus muscles are severely affected and the tibialis anterior on the right is relatively spared. Reproduced with permission under CC-BY license, Wolter Kluwer.

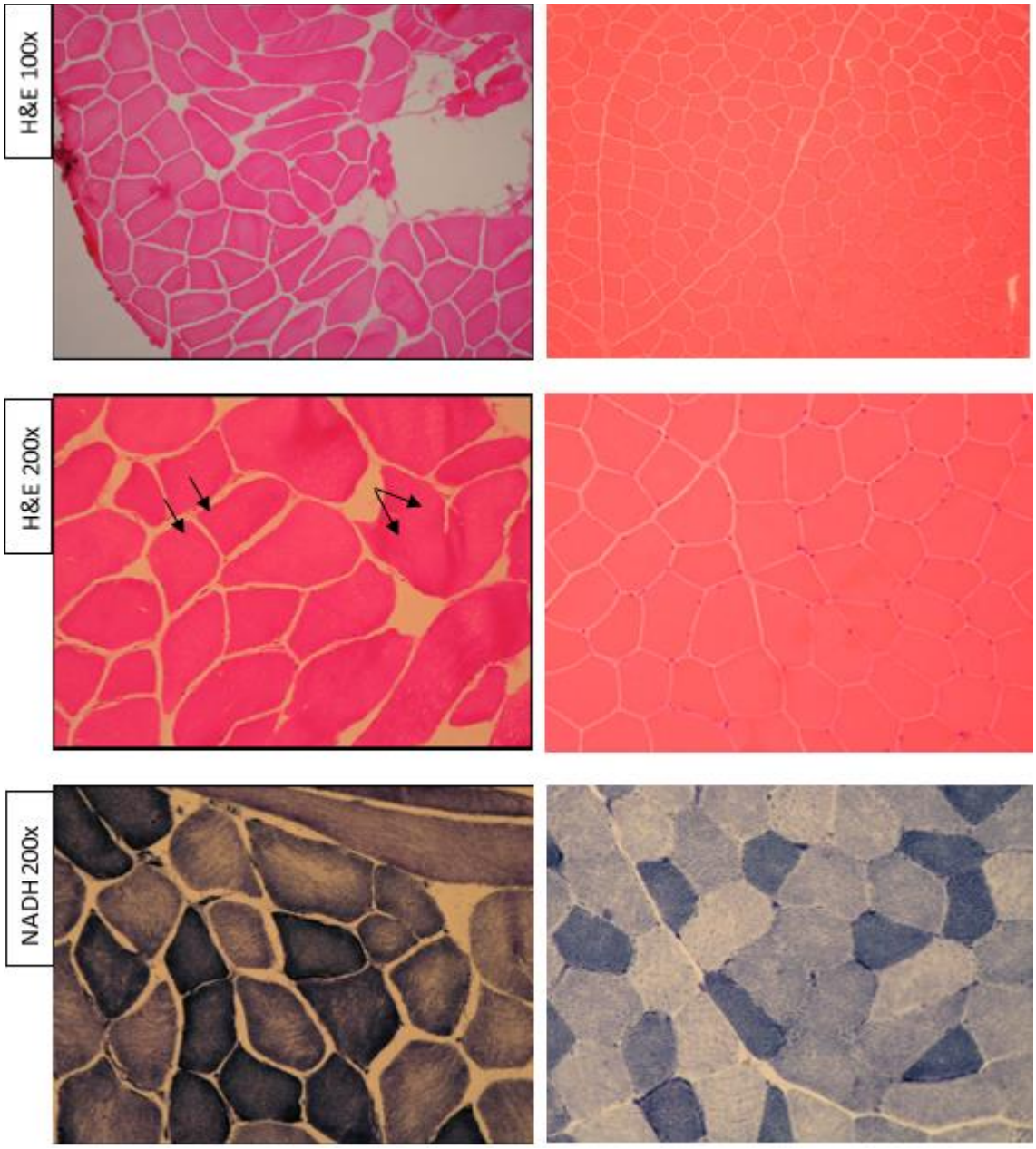


Figure 31: Needle biopsy of the left vastus lateralis from a patient with *MYMK* mutations. Patient images (left) and control images (right). H&E stain demonstrates fibre size variation (H&E 100x) and occasional internal nuclei (H&E 200x, arrows). Mild moth-eaten changes seen on NADH stain indicating uneven mitochondrial enzyme activity within the sarcoplasm (NADH 200x). Reproduced with permission under CC-BY license, Wolter Kluwer.

b. FILIP1

For family 53, the index case is the first child to consanguineous Pakistani parents living in the UK. She presented at birth with hypotonia, contractures and feeding difficulties. Her motor development was delayed and she sat independently at 18 months of age and took her first steps at 24 months. She was noted to have early kyphoscoliosis, a rigid spine, symmetrical scapular winging and Gowers' manoeuvre. She had contractures at the elbows, knees and long finger flexors. She also has dysmorphic facial features (figure 32), facial weakness, ptosis, a high arched palate and webbing of the neck. Her skin was rough and prone to hypertrophic scaring. She had mild learning difficulties with speech delay and required learning support at school.

Although in early childhood she suffered with frequent episodes of pneumonia, she had no respiratory muscle weakness. She underwent echocardiography and this did not reveal any abnormalities. There was no evidence of delayed growth and menarche was normal at the age of 12 years.

At the age of 15 years, she was walking unaided and able to climb stairs. Manual muscle testing using the MRC-muscle strength grading showed 5/5 in neck flexion, shoulder abduction, elbow flexion, elbow extension, hip flexion, hip extension, knee flexion, knee extension, dorsiflexion and plantar flexion bilaterally. At this stage of the disease, she was also complaining of exercise-induced myalgia. She also demonstrated a one hand Gower's manoeuvre to rise from the floor.

The patient had two affected younger sisters and one healthy brother.

Her CK levels have been persistently mildly elevated at 400-900 IU/L. One of the affected sisters however, had a significantly higher CK at 3400 IU/L.

Cultivated skin fibroblasts were used to investigate the possibility of a collagen VI disorder but no abnormality was detected in collagen VI expression.

A muscle biopsy was taken from the left quadriceps and histology showed type1 fibre predominance while immunohistochemistry showed patchy dystrophin deficiency (figure 33). Analysis of the dystrophin gene showed a heterozygous pathogenic non-sense mutation, c.8713C>T (p.Arg2905Ter). However, no X chromosome inactivation skewing was identified and this mutation did not explain the phenotype.

WES highlighted a homozygous non-sense novel variant, c.169C>T (p.Arg57Ter), in exon 2 of the *FILIP1* gene as a potential candidate for this patient.

FILIP1 codes for filamin A interacting protein 1 and is highly expressed in skeletal muscle. The protein has recently been shown to interact with myogenic transcription factors such as *Myog* (Chen *et al.*, 2018). Mutations in *FLNA*, which encodes the interacting protein, filamin A, is associated with a spectrum of disorders including Melnick-Needles, fronto-metaphyseal dysplasia and oto-palato-digital syndromes (Moutton *et al.*, 2016). The extra-muscular features in the patient overlap with features from these syndromes. For example, facial features such as frontal bossing, high arched palate and micrognathia, spinal deformities, scapular winging, joint contractures and learning disabilities are features of *FLNA*-related syndromes and are also features in our patient with *FILIP1* mutations. Furthermore, reported chromosome 6q microdeletion syndromes include deletion of the FILIP1 gene and the phenotypic features include, intellectual disability, facial dysmorphism, skeletal deformities and hypotonia (Becker *et al.*, 2012).

Segregation studies were performed to further strengthen the evidence for pathogenicity of the *FILIP1* variant in this this family. These revealed that both parents are heterozygous carriers, an affected sibling is homozygous and an unaffected sibling is heterozygous for the variant. Therefore, the variant segregates with the disease in this family.

Additional *FILIP1* variants were identified through WES in patient from a parallel sequencing project at Newcastle University. The index is a boy of German origin, born to consanguineous parents who are first cousins, and had prenatal onset disease with polyhydramnios. At birth, he presented with hypotonia, joint contractures and poor feeding. He had delayed motor development, delayed speech and generalised muscle weakness. He is also reported to have dysmorphic facial features. A homozygous missense variant, c.3398C>T (p.Pro1133Leu), in exon 5 of the *FILIP1* gene was identified using WES. The variant is novel, corresponds to a highly conserved amino acid and is predicted to be damaging by Polyphen2, SIFT and Mutation Taster. Deep phenotyping, muscle biopsy and segregation of the variant in this second family is currently ongoing.

FILIP1 is proposed to be a novel disease gene for NMD and particularly for syndromic congenital myopathy.

Figure 32: Disease features in a female patient with homozygous p.Arg2905Ter *FILIP1* variants. A and B, Dysmorphic features: Large forehead with prominent ridging of the metopic suture, large prominent ears, prominent nose and micrognathia, ptosis, myopathic face, neck webbing, kyphoscoliosis, elbow and knee flexion contractures and bilateral foot pronation. C, rigid spine and D, flexion contractures of the fingers.


C


D

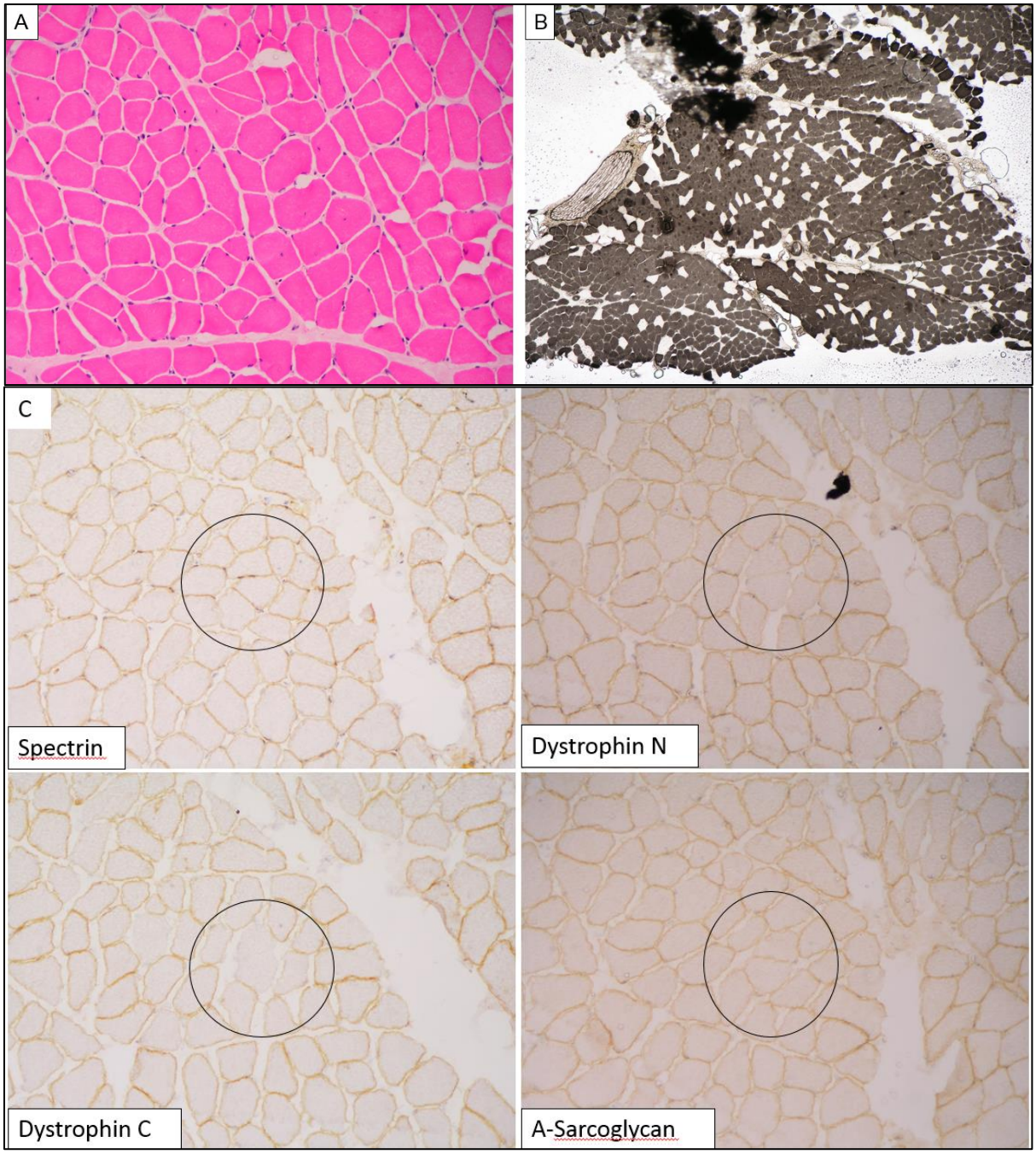


Figure 33: Muscle histology and immunohistochemistry for a female patient with *FILIP1* variants. A: H&E stain; fibre-size variation. B: ATPase pH4.3; type 1 fibre predominance. C: immunohistochemistry; patchy dystrophin deficiency.

c. TENM2

Family 21 is a consanguineous family from Yemen. The affected daughter presented in early childhood with frequent falls, and difficulties with walking and climbing stairs. She was found to have proximal weakness, joint hyper-laxity, scapular winging and lordosis. She also had facial muscle weakness. At 11 years old, she was ambulant but had difficulties climbing stairs and rising from a lying position. She had severe neck and truncal weakness with some spinal rigidity. Her proximal limb weakness remained mild and she had foot extensor weakness. She was found to have hypertrophic scarring.

CK levels and muscle histology were normal. However, the muscle biopsy was taken from a relatively unaffected muscle (medial gastrocnemius). EMG studies pointed towards a myopathic process and nerve conduction studies were normal.

WES performed for the affected child and her parents proposed a homozygous missense variant in the *TENM2* as a candidate gene in the affected child. Both parents were heterozygous carriers. The variant, c.5879G>A (p.Arg1960His) is rare (allele frequency in ExAC: 0.0001326) and predicted to be damaging by Polyphen, Mutation Taster and FATHMM.

TENM2 encodes teneurin transmembrane protein 2, which is proposed to have role in neural development and establishment of connectivity in the nervous system. By sequence similarity, the protein is also expected to act as a transcription inhibition factor (Silva *et al.*, 2011). *TENM2* is expressed in moderate to high levels in muscle, skin, heart, brain and peripheral nerves.

Segregation of the variant showed that an unaffected brother was heterozygous for the variant.

TENM2 is proposed as a novel candidate for congenital myopathy.

d. ***NRAP***

For family 9, WES was performed for one affected individual. She presented with a congenital myopathy, generalized muscle weakness that was more evident proximally, and involvement of the extra-occular muscles manifesting with ptosis and opthalmoplegia.

WES highlighted two heterozygous variants (c.3300+1G>A; essential splice site variant and c.741C>T; c.741C>T(p.=)) in the *NRAP* gene. The variants are rare with no homozygous individuals reported in the ExAC control population for either of them. The second variant is

a synonymous one and is not predicted to cause an amino acid change. Nonetheless, synonymous mutations are implicated in disease through mechanisms related to splicing or alteration of RNA structure and rate of translation (Sauna and Kimchi-Sarfaty, 2011; Supek *et al.*, 2014).

NRAP encodes the nebulin-related anchoring protein, which is an actin-binding cytoskeletal protein with high expression in skeletal and cardiac muscle (Truszkowska *et al.*, 2017). Homozygous NRAP mutations have been reported in a single individual with dilated cardiomyopathy (Truszkowska *et al.*, 2017), and a heterozygous mutation has been discussed in the context of BAG3-related myofibrillar myopathy (D'Avila *et al.*, 2016).

e. XIRP2

The patient (family 23) has a congenital myopathy with facial weakness and scoliosis and proximal muscle weakness and wasting. He is the only affected child to non-consanguineous healthy parents.

WES identified compound heterozygous variants in the *XIRP2* gene. The missense variants, c.3014A>G (p.Gln1005Arg) and c.2682T>A (p.Asn894Lys), are rare variants with no homozygous individuals reported in control populations. Bioinformatics pathogenicity predictions were conflicting with 1/5 predictions tools classifying the variants as damaging.

XIRP2 is highly expressed in skeletal and cardiac muscle. It belongs to the Xin gene family that are known to regulate intercalated disk development in cardiac muscle. Mutations in *XIRP2* have been associated with cardiac muscle and conduction defects (Long *et al.*, 2015). The protein has also been shown to interact with a number of proteins with known expression and importance in skeletal muscle development, function and disease (figure 34). Its unique and direct interaction with FLNC, a known NMD gene (Leber *et al.*, 2016), makes *XIRP2* a plausible candidate for muscle disease.

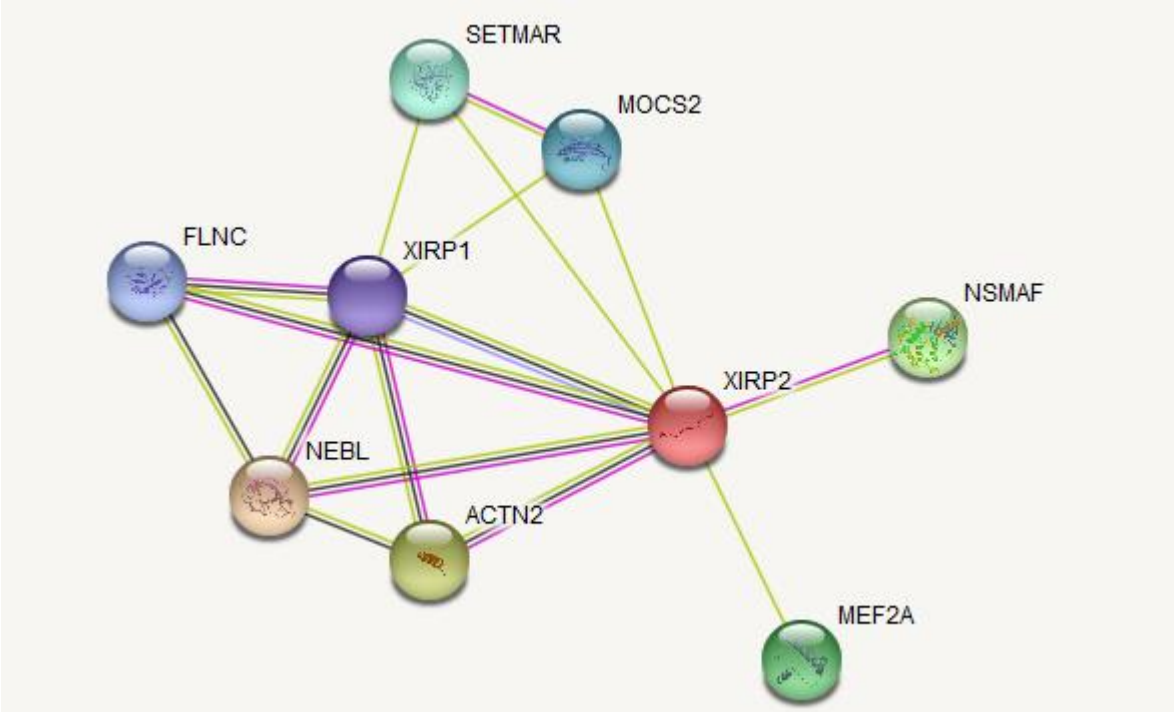


Figure 34: Protein interaction network for XIRP2. The image was produced on the STRING database (https://string-db.org/) using XIRP2 in homosapiens as the query protein. FLNC, filamin C; NEBL, nebulette; XIRP1, Xin-related protein 1; SETMAR, SET domain and mariner transposase fusion gene; MOCS2, Molybdenum cofactor synthesis 2; XIRP2, Xin-related protein 2; MEF2A, myocyte enhancer factor 2A and NSMAF, Neutral sphingomyelinase activation associated factor.

f. CAMK2B

The affected father and son (family 66) underwent WES. The clinical presentation was of progressive adult onset weakness and an autosomal dominant inheritance pattern. WES proposed a novel, damaging missense variant, c.559C>T (p.Arg187Cys), in both father and son in exon 8 of the *CAMK2B* gene.

CAMK2B encodes the calcium/calmodulin-dependent protein kinase II, which is a serine/threonine kinase that is highly expressed in skeletal muscle, heart, brain and peripheral nerves. CAMK2B has a protein kinase domain (https://www.ebi.ac.uk/interpro/protein/Q13554) where the variant identified here is located.

De novo mutations in *CAMK2B* have been associated with mental disability and neurodevelopmental disorders. These were *de novo* missense mutations that were found to affect the auto-phosphorylation of the protein in the brain (Kury *et al.*, 2017; Akita *et al.*, 2018). In the context of NMD, although, *CAMK2B* has not directly been associated with

muscle disease, the protein was proposed as one of the top candidates in a study searching for therapy-responsive biomarkers in a mouse model of Duchenne muscular dystrophy (Coenen-Stass *et al.*, 2015).

g. CCDC158

Family 75 has two affected siblings that underwent WES along with their parents. The siblings presented with proximal weakness, contractures, ptosis, dysphagia and respiratory impairment. WES revealed that both the affected individuals carry compound heterozygous variants in the *CCDC158* gene, c.2485C>T (p.Arg829Cys) and c.1094G>A (p.Arg365His). Each parent carries one of the variants only. Both variants are rare and predicted to be damaging.

CCDC158 codes for coiled coil domain containing protein 158. Function of the protein remains unknown and it has no known association with disease. However, a study looking into single nucleotide polymorphisms (SNP) in cattle and their association with carcass quality traits found that SNPs in *CCDC158* were associated with longissimus dorsi muscle area and muscle fat content and distribution (Lee *et al.*, 2013).

h. MYH13

Family 86 consisted of one female who presented with limb girdle weakness and was thought to have a vacuolar myopathy following muscle biopsy. WES found two *MHY13* missense variants, c.5108G>T (p.Arg1703Leu) and c.4021C>T (p.Arg1341Cys). Both variants are rare and predicted to be damaging.

MYH13 encodes a fast twitch myosin belonging to the extra-ocular myosin heavy chains (MYH-EO) (Mascarello *et al.*, 2016). However, moderate to high expression has been detected in other skeletal and cardiac muscles, as evident by experiments reported on Gtex Portal and the EBI Expression Atlas in humans and other mammals, making it a potential candidate gene for a generalized muscle disorder.

i. DSG2

One individual from family 89 underwent WES. He suffered with proximal limb weakness and myalgia. CK was raised to levels up to 1000 units per litre (U/L). He also had mild left ventricular impairment. His brother also suffered with cardiac disease, however, no further details are available regarding his diagnosis.

WES identified two heterozygous variants in the *DSG2* gene; a frame shift deletion (c.602_603delTA, p.Leu203GlyfsTer12) and a missense variant (c.604T>C, p.Ser202Pro). Both variants were absent in control populations and predicted to be damaging. Due to the proximity of these variants to each other, it was possible to examine the sequence reads via the Integrated Genomics Viewer (IGV) integrated on *seqr*. This revealed that both variants were in cis and thus the proposed mode of inheritance here would be that of an autosomal dominant (AD) one.

DSG2 codes for desmoglein-2. In cardiac muscle, desmoglein-2 is a transmembrane protein in the desmosomes of the intercalated discs, which are responsible for linking cells to intermediate filaments (Kessler *et al.*, 2017). Although *DSG2* is highly expressed in cardiac and skeletal muscle, localization within skeletal muscle is not known.

Homozygous and heterozygous mutations in *DSG2* have been associated with dilated cardiomyopathy 1BB and with arrythmogenic right ventricular dysplasia 10. However, to date, mutations in the gene have not been reported to cause a generalized muscle disorder.

j. WDR49

For family 17, WES was performed for three affected children, an affected father and an unaffected mother. The phenotype was of autosomal dominant LGMD.

WES identified a novel essential splice site variant at the border of exon 6 of the *WDR49* gene. The variant (c.1126-1G>A) is predicted to be damaging. All four affected individuals were heterozygous for the variant and the unaffected mother had the wildtype allele.

WDR49 encodes a WD-repeat domain protein, belonging to a family of proteins with diverse functions (Smith, 2008). WDR49 protein function and expression in muscle is not well characterized.

k. LARS

Family 10 consists of one individual who underwent WES. He complained of proximal weakness and wasting with calf muscle hypertrophy.

WES identified a heterozygous novel nonsense variant in the *LARS* gene (c.1228C>T; p.Gln410Ter). *LARS* encodes the leucyl-tRNA synthetase, an aminoacyl tRNA synthetase. This group of proteins are responsible for catalysing the specific attachment of amino acids to their respective tRNA (Han *et al.*, 2012). The gene is highly expressed in muscle; however, its specific role remains to be investigated.

4.5 Discussion

WES was performed for 93 families with at least one individual presenting with limb girdle weakness. These patients have been extensively examined and investigated through sequential single gene testing but remained without a molecular diagnosis and therefore, they were recruited in a WES project. Analysis of the WES data was carried out on the integrated genomics platform *seqr*. Analysis on the platform allowed filtering the VCF files to a smaller, manageable and more relevant number of variants. Variants were filtered for inheritance model, predicted damaging effect on the protein, population frequency and, in the initial analysis, for those falling in NMD genes. The output report was further examined for variant type, predicted or known pathogenicity, tissue-specific expression, gene function, allele frequency, gene constraint metrics and phenotypic associations. This was possible on one interface through the integrated databases and application programme interfaces (API) that link, and allow queries across, various software and web tools.

Using an integrated platform, a genetic diagnosis was proposed for 65.6% of families in this cohort. For 51% of all families, this was in a known NMD gene. WES also proposed 14 novel candidate genes for NMD by querying the entire exome on *seqr* for rare, damaging and potentially pathogenic variants, that segregate in the family with disease, and that are in genes known to be expressed in skeletal muscle tissue. For example, in consanguineous families (families 21 and 53), the report output on the platform was filtered to include variants that fit and autosomal recessive inheritance, in addition to filters for population frequency and predicted effect on the protein. This lead to the identification of the novel candidates *TENM2* and *FILIP1*. Using the data provided on the platform and the additional links to genomic

databases, the importance of these variants was evaluated. For example, expression of the *FILIP1* variant in muscle was found to be high in skeletal muscle (figure 35). In addition, the *TENM2* variant was predicted to be damaging by four out of the five pathogenicity prediction tools incorporated on the platform. This result was reported directly in the variant output report (figure 36). Further segregation of both variants in additional family members via Sanger sequencing confirmed that both segregate with disease. Moreover, the platform enables gene queries across all projects, and interrogating other NMD WES projects found an additional family with compound heterozygous variants in *FILIP1*.

As the cohort described here was recruited on a collaborative research basis, clinical data and bio-samples were limited for a considerable number of patients. In addition, many of the variants described here require validation prior to being assigned as disease causing. This should be done through reverse and deep phenotyping, re-evaluation of muscle imaging and biopsy tests, segregation within the family, and for novel candidates, identification of other families with variants in the same gene and a similar phenotype, and functional in vivo and in vitro studies in the laboratory (Coonrod *et al.*, 2011).

Overall, there was a small benefit in including affected and unaffected relatives when performing WES for patients with rare NMD. The benefit has been previously demonstrated in a cohort of 60 patients with LGMD, where trios WES analysis proposed a diagnosis in 60% compared to 40% when only the proband was sequenced (Ghaoui *et al.*, 2015). The analysis in this latter study was also performed on the *seqr* platform (known as xBrowse at the time). Inclusion of parents and family members, with the added knowledge of their disease status allows the researcher to apply additional filters to fit a particular inheritance model and to validate candidates based on their segregation in the family with disease (Ghaoui *et al.*, 2015).

The highest number of variants proposed in a single gene were in the *TTN* gene, where it was a candidate for 12 families (19.6% of families with a proposed diagnosis). Using WES and myopathy-gene targeted NGS methods, novel and rare *TTN* variants rank highest amongst all variants detected in genes of interest (Norton *et al.*, 2013; Evila *et al.*, 2016; Harris *et al.*, 2017). *TTN* is an extremely large gene (363 exons) and thus is prone to higher variability and assigning pathogenicity to these variants is difficult (Evila *et al.*, 2016; Harris *et al.*, 2017). Using an integrated platform allows for *TTN* variants to be evaluated in the context of up-to-date published knowledge of the human variome through the interface. However, for novel *TTN* variants, it is recommended that functional confirmation is carried out prior to assigning pathogenicity (Hackman *et al.*, 2017).

The proposed causative variants discussed here were the best candidates from the WES experiment for these patients. Patients with a variant in a novel candidate gene did not have

any pathogenic or predicting damaging mutations in known NMD genes. To establish pathogenicity, novel candidates require validation via familial analysis, RNA and protein studies, functional experiments in model systems and identifying additional patients with a similar phenotype and variants in the same gene (Coonrod *et al.*, 2011; Biancalana and Laporte, 2015).

For the remaining families with no proposed diagnosis, it is possible that the genetic defect is in a region not detectable by WES (for example introns, promoter or intergenic regions), is in a novel gene, or is a structural variation not detected by WES. In addition, it is also possible that the pathology is not genetic in origin, particularly for sporadic cases (Efthymiou *et al.*, 2016).

WES has substantially changed the way rare diseases are diagnosed and has led to many novel genetic associations. Nonetheless, validating variants in a WES experiment and assigning pathogenicity can be difficult. Integrated genomics platforms, such as the one used in this analysis, are becoming increasingly popular for variant prioritization in WES and WGS. The interfaces require minimal training and are increasingly being tailored to researchers' needs, making them practical and user friendly (Ghaoui *et al.*, 2015). The platforms allow users to manipulate NGS and customize the variant output report, adding flexibility to a rigid computational algorithm. The filtering options applied in this analysis as well as the links and tools used on the platform are also part of the CSA, RD-Connect, and many other integrated analysis platforms. The RD-Connect Genome-Phenome Analysis Platform integrates the Exomizer software application that allows users to further filter and prioritize variants using the Exomizer algorithm. This algorithm uses data from protein interaction networks and cross-species phenotype comparisons (Smedley *et al.*, 2015).

A highlight of *seqr* is the summary page provided for each variant. This page includes basic information about the gene, statistics on gene variation in control populations, disease associations, links to external databases and a visual summary of tissue-specific expression levels provided by GTEx. This information is also accessible on CSA and RD-Connect either directly on the platforms or via external links. In addition, *seqr* and CSA enable visualization of the sequence reads at a particular variant position via IGV and Sequence Miner integrated on *seqr* and CSA respectively.

Furthermore, *seqr*, RD-Connect and CSA allow searches in particular genes of interest across all projects on the platform. This is particularly important for validating novel candidates, when a second family with a similar phenotype and variants in the same gene would support the proposed pathogenicity.

Overall, it is evident that analysis of WES data on an integrated platform is a practical solution for rare disease diagnosis and disease gene discovery. The benefits of these platforms are also evident in clinical settings, where software platforms for WES data analysis have shown success and cost-effectiveness despite technical and bioinformatics challenges (Coonrod *et al.*, 2011).

A successful integrated platform should have a simple and user-friendly interface, separating the clinician or researcher from the complex computational algorithms carried out by the platform. In addition, the many options and filters for data manipulation and variant prioritization should be grouped on the interface in themes and organized in dropdown menus. Users should be able to trust that the platform is executing the expected analysis and errors should be reported with possible solutions. Finally, the bioinformatics pipeline and the integrated platform should be flexible enough to cope with different standard forms of input files and with updates in integrated tools and databases. This will ensure interoperability and allow data from different NGS projects to be analysed on the same platform (Coonrod *et al.*, 2011; Shyr *et al.*, 2014).

The example shown here has demonstrated the usefulness and practicality of an integrated platform for the analysis of WES in NMD. The analysis allowed candidate genes to be proposed for more than 60% of families in this difficult to diagnose cohort. Despite the proposed candidate being in a known NMD for most patients in the cohort, traditional single gene testing did not reveal a diagnosis. This highlights the genetic and phenotypic heterogeneity of NMD and the effectiveness of WES and integrated analysis. A number of novel genes, a novel mode of inheritance and phenotypic expansions of existing genetic associations were also proposed, and were all facilitated by data accessible from a single interface.

With continued improvements of sequencing technologies, bioinformatics tools, knowledge of the human genome and variome and phenotypic data, the integration of all this data in an effective and easy-to-use workflow, the use of integrated platforms will increasingly move into clinical settings (Coonrod *et al.*, 2011). This will undoubtedly lead to a molecular diagnosis of more patients with rare NMD and to more novel gene discoveries.

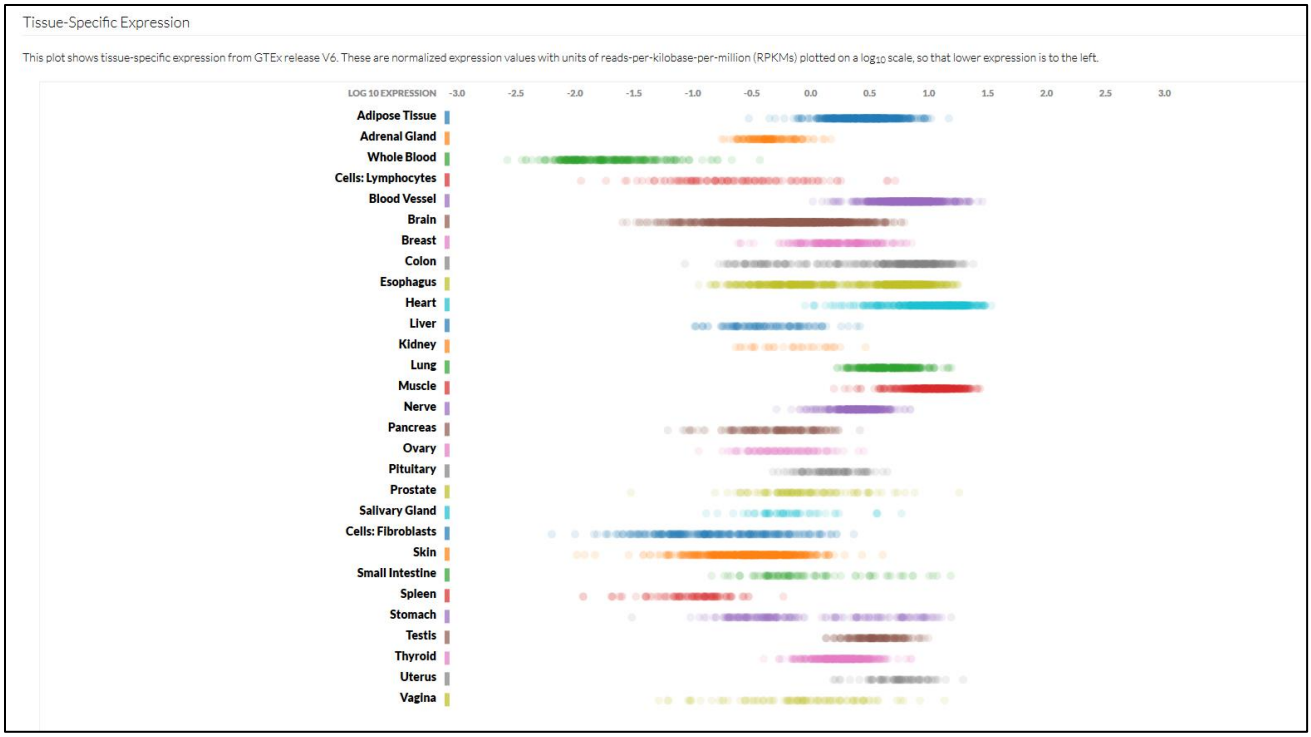


Figure 35: Screen shot of gene expression data provided on *seqr* for *FILIP1*, showing high expression in muscle (amongst other tissues).

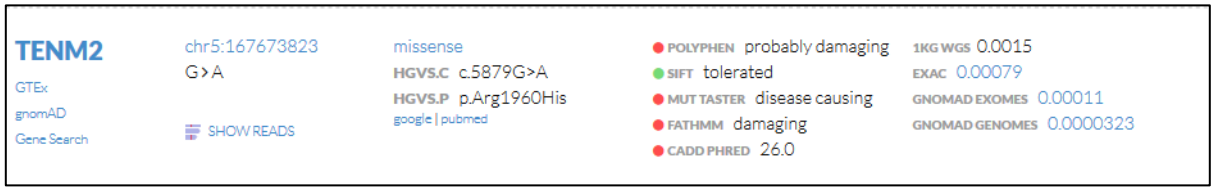


Figure 36: Screen shot from *seqr* variant output report for *TENM2*, the proposed novel candidate gene in family 21. Pathogenicity predictions are given in the initial variant output report.

Chapter 5. GNE myopathy in the Bedouin population of Kuwait

5.1 Introduction

GNE myopathy is also known as hereditary inclusion body myopathy (HIBM), distal myopathy with rimmed vacuoles (DMRV) and Nonaka myopathy. Clinically, it is an autosomal recessive juvenile to adult onset distal myopathy with relative quadriceps muscle sparing. Proximal muscles are also affected and progression is usually slow. The disease initially affects the tibialis anterior muscles, causing foot drop. Serum creatine Kinase (CK) levels are normal to mildly elevated. Rimmed vacuoles are usually seen in fibres of affected muscles in patients with GNE myopathy (Eisenberg *et al.*, 2003; Argov and Mitrani Rosenbaum, 2015; Chamova *et al.*, 2015).

GNE myopathy is caused by mutations in the *GNE* gene that encodes UDP-N-acetylglucosamine 2-epimerase/N-acetylmannosamine kinase. The enzyme catalyses the initial two rate-limiting steps in N-acetylneuraminic acid biosynthesis. The latter is a member of the sialic acid family. The gene is located on chromosome 9p12-13 and has multiple isoforms. Although ubiquitously expressed, highest expressions of the *GNE* gene are seen in the liver, lungs and kidneys. However, these organs are not affected in GNE myopathy. Enzyme levels in skeletal muscle is relatively low and the exact mechanism by which *GNE* mutations selectively result in a myopathy remains unknown (Hinderlich *et al.*, 1997; Stasche *et al.*, 1997).

GNE myopathy has shown population clustering and several founder mutations have been described: the Middle Eastern p.M743T mutation (Argov *et al.*, 2003; Argov and Mitrani Rosenbaum, 2015), the p.I618T mutation in the Roma/Gypsy population in Bulgaria (Argov and Mitrani Rosenbaum, 2015; Chamova *et al.*, 2015), and the two common mutations, p.N409T and p.A662V, with potential founder effects in Northern Britain (Chaouch *et al.*, 2014; Argov and Mitrani Rosenbaum, 2015). In Asian populations, GNE myopathy is more heterogeneous. However, a few mutations show higher prevalence and founder effects in Japan (p.V572L) (Cho *et al.*, 2014), Southern India (p.V727M) (Argov and Mitrani Rosenbaum, 2015), and China (p.D207V) (Zhao *et al.*, 2015).

Screening for the Middle Eastern mutations in Kuwait started in 2013. To date, the mutation has been described in five families of Arab Bedouin origin. Following establishing a genetic

diagnosis in these families, a retrospective study using data from the molecular laboratory and patient records was carried out to study GNE myopathy in the Kuwaiti population. A clinical, demographic and epidemiological description of the disease in the Kuwaiti population is given.

5.2 Aim

- Report on the epidemiological, clinical and genetics findings in patients with GNE myopathy from Kuwait.

5.3 Patients and methods

5.3.1 Ethical approval

Informed consent for participation in research was obtained from all patients according to protocols at the Kuwait Medical Genetics Centre (KMGC). Ethical approval for the project was granted by Newcastle University Ethics Committee (Ref: 3601/2018).

5.3.2 Patients, clinical evaluation and mutation analysis

GNE mutation screening commenced at KMGC in 2013. Patients presenting with distal myopathy and relative quadriceps sparing were tested. In addition, patients with undiagnosed distal myopathy were re-evaluated for the possibility of GNE-myopathy and screened where indicated.

Using the molecular laboratory register at KMGC, a retrospective analysis was performed for all patients screened for the p.M743T mutation, between January 2013 and August 2017. This revealed that 37 patients (figure 37) were screened for the mutation, fourteen of whom were found to have GNE myopathy. Clinical assessment of patients took place at KMGC and included thorough history and examination, clinical photography, family history and pedigree structure, cardiac evaluation, serum CK levels and muscle electrophysiology studies. Muscle MRI was performed for three patients and muscle biopsy for six patients in this cohort. Cardiac evaluations were carried out by specialist cardiologists and electrocardiography (ECG) +/- echocardiography performed as indicated.

Mutation analysis was performed for suspected cases using Sanger sequencing. Protocols for DNA extraction, PCR, primers and sequencing for the Middle Eastern p.M743T mutation at KMGC are given in the appendices.

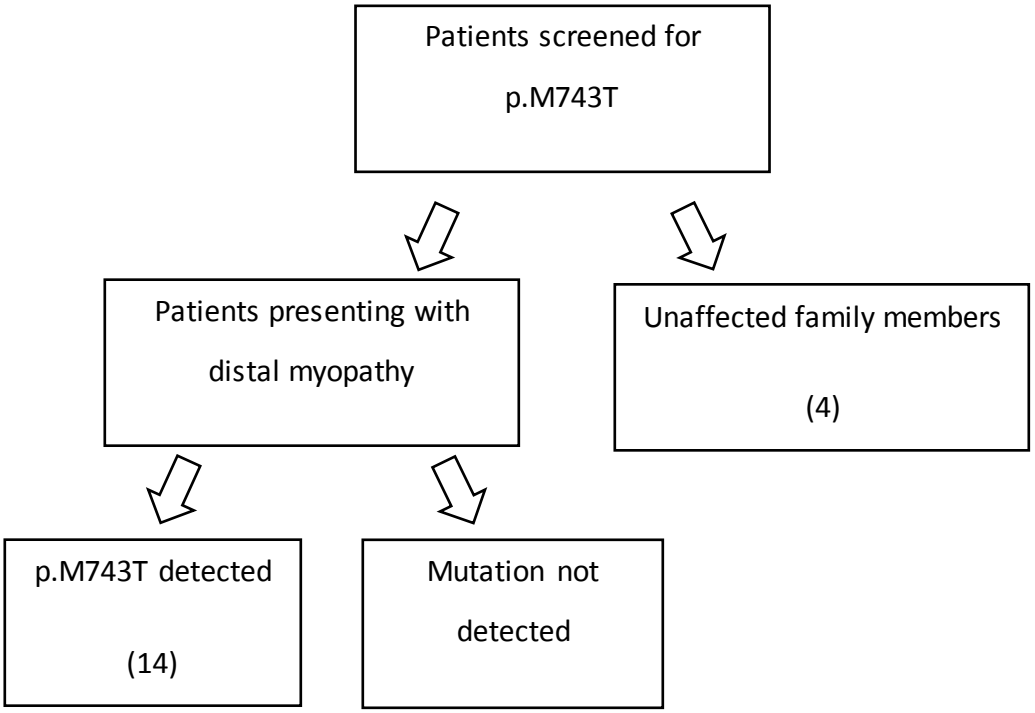


Figure 37: Patients screened for the p.M743T mutation at the Kuwait Medical Genetics Centre between January 2013 and August 2017.

5.3.3 Prevalence estimate and carrier frequency

KMGC is the only governmental clinical genetics centre in Kuwait and all cases suspected to have a genetic disorder are referred to KMGC clinics for evaluation and testing.

Therefore, the number of cases with a *GNE* mutation identified at the KMGC reflects the number of all known GNE myopathy cases in Kuwait. To date (01-August-2017), KMGC has identified fourteen cases with a mutation in the GNE gene since 2013.

To estimate carrier frequency in the Bedouin population of Kuwait, 100 DNA samples were screened for the p.M743T mutation. These samples were obtained from the DNA bank at KMGC and were for patients attending the centre for anything other than a neuromuscular

disorder. All DNA samples were from individuals of Bedouin origin as GNE myopathy has not been described in any other population group in Kuwait.

5.4 Results

5.4.1 Demographic and genetic findings

All fourteen patients identified as having GNE myopathy and the p.M473T mutation were of Bedouin Arab origin. The patients belonged to five apparently unrelated families (KW-1-5 in figure 38) from Kuwait, with one family originating from Saudi Arabia. Three additional cases (KW-5/5 and 6, and KW-4/3) were reported by family members to have muscle weakness, however, were not seen in clinic or tested for the mutation.

With the exception of KW-4/1, all patients were born to consanguineous parents and all are from large inbred Bedouin tribes. KW1-4 belong to one of the major Arab Bedouin tribes in the gulf and Middle East and KW-5 descend from an equally large but distinct tribe. Families KW-1 and 5 show pseudo-dominant inheritance, where offspring appear to directly inherit the disease from an affected parent. Pseudo-dominance highlights the complex consanguinity in these families.

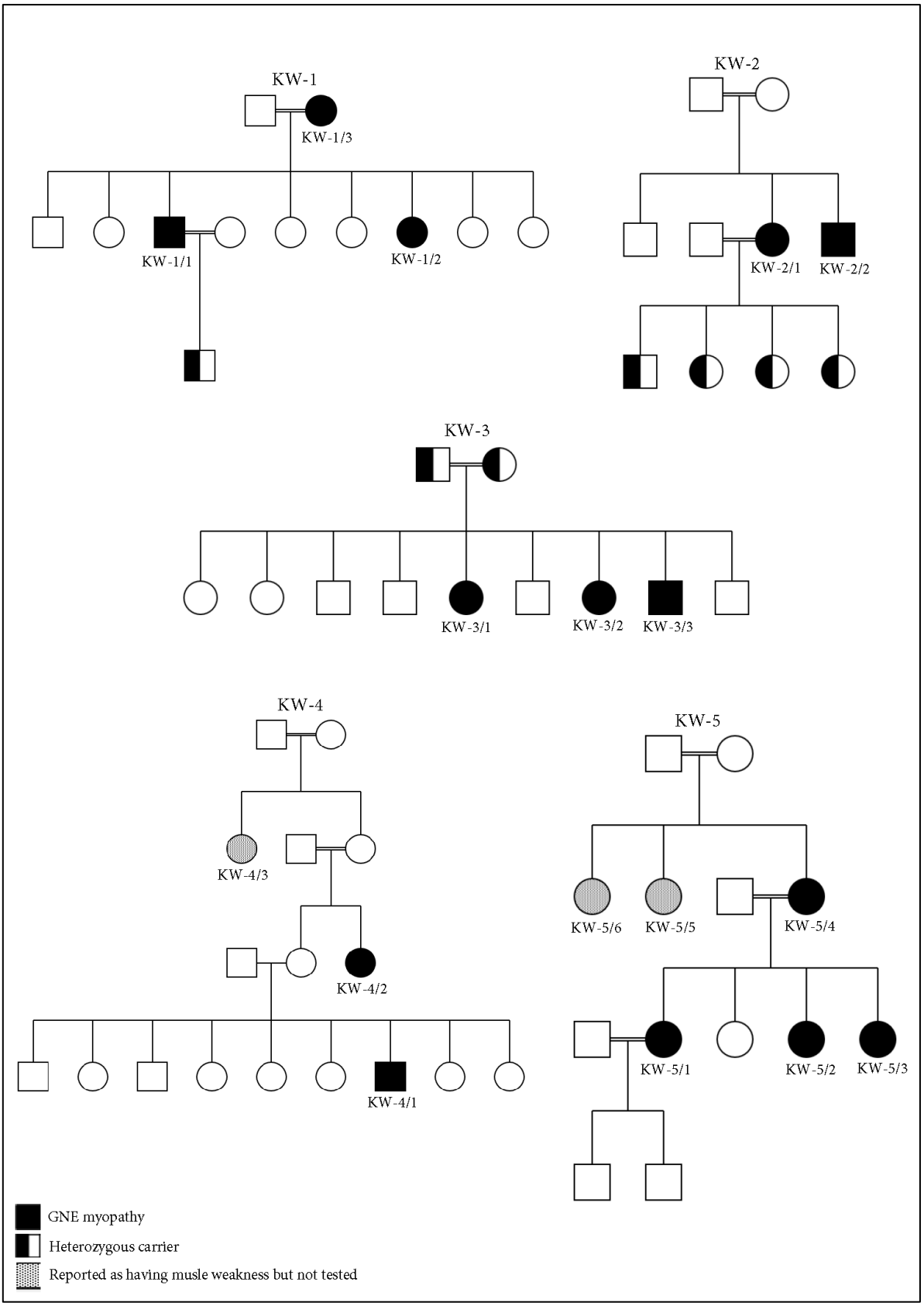


Figure 38: Pedigrees for families affected with GNE myopathy from Kuwait.

5.4.2 Clinical findings

Table 17 describes disease aspects in the five families. Mean age of symptom onset was 27.4 years (range 18-37 years). The tibialis anterior muscle and intrinsic hand muscles were significantly affected. All patients presented with distal weakness in upper and lower limbs, progressing to foot drop and hand grip weakness.

Proximal weakness was also prominent in these families and manifested as progressive walking difficulties and loss of ability to climb stairs. Seven out of fourteen patients have been followed up for 10 years or more since onset of symptoms, six of which have lost ambulation and are wheel chair dependant. From the remaining seven patients, two have lost ambulation remarkably early at 5-6 years from disease onset. At 10 to 15 years from disease onset, approximately 50% of patients remain ambulant.

All patients had relative sparing of the quadriceps muscle, with the exception of one (KW-5/4) who had markedly weak quadriceps bilaterally and lost ambulation early at the age of 30 years (5 years from disease onset).

All patients showed distal muscle wasting of the upper limbs. However, this varied in the lower limbs where individuals from two of the families had prominent pseudo-hypertrophied calves (KW-2 and 4). Atrophy and weakness were also evident in hip and shoulder girdle muscles and scapular winging and gait abnormalities were subsequently early features of the disease.

Atypical findings such as bilateral ophthalmoplegia, ptosis and tongue wasting were signs noted in case KW-2/2.

With the exception of KW-4/2, none of the patients had cardiac-related complaints or symptoms suggestive of respiratory insufficiency. All patients were nonetheless referred to cardiology specialists for assessment, and no abnormalities were reported back except for case KW-4/2. The latter patient complained of difficulty breathing and snoring. Her ECG and Holter monitoring showed sinus rhythm and sinus arrhythmia. Echocardiogram revealed normal ejection left ventricular size, motion and ejection fraction and heart valves, but grade 2/4 ventricular diastolic dysfunction and right atrial enlargement. Pulmonary function tests showed decreased vital capacity and decreased maximal respiratory pressures consistent with poor performance or respiratory muscle weakness. Overnight pulse oximetry was normal and swallowing assessment did not show any abnormality.

Serum CK levels were mildly elevated in this cohort, ranging from 195 units/litre (u/L) to 712 u/L. EMG studies were performed for one or more affected individuals from each family. Results were either myopathic (3 out of 8) or showed a mixed neuropathic and myopathic pattern (5 out of 8).

MRI studies (Figure 39) were performed for four individuals (KW-1/1, KW-2/2, KW-5/1 and 2) from three families and all showed selective and symmetrical muscle involvement. In the lower leg, muscles of the anterior and lateral compartments were atrophied, while posterior compartment muscles were relatively preserved. Proximally, the hamstrings were affected and the quadriceps were spared.

Muscle biopsy was performed for six out of the fourteen patients evaluated (KW-1/1, KW-3/1, KW-4/1 and 2, and KW-5/1 and 3). Histological and immunohistochemical analysis showed non-specific myopathic changes such as fibre-size variation, atrophy and regeneration. Muscle biopsy for patient KW-1/1 also showed some possible neurogenic findings (Figure 40). He underwent cervical, dorsolumbar and sacral MRI to exclude pathology contributing to neurogenic changes seen on his muscle biopsy but no abnormalities were detected.

Table 17: Clinical features for patients with GNE myopathy homozygous for the p.M743T mutation. All patients are of Bedouin Arab origin.

Case	Age at clinical onset	Relative quadriceps sparing	Clinical highlights	Serum CK levels (u/L)*	Ambulation (total years from disease onset at last examination)	EMG
KW-1/1	34	Yes	Beevor's sign	471	Ambulant (3)	mixed
KW-1/2	34	Yes	Hyperlordosis	264	Ambulant (7)	mixed
KW-1/3	30	Yes		N/A	WCD (30)	N/A
KW-2/1	34	Yes		190	WCD (6)	myopathic
KW-2/2	37	Yes	Calf hypertrophy, ophthalmoplegia, ptosis, asthenia, oligospermia, tongue wasting	609	Ambulant (8)	N/A
KW-3/1	25	Yes	-	411	WCD (15)	mixed
KW-3/2	24	Yes	-	195	Ambulant with aids (11)	mixed
KW-3/3	23	Early disease stages	-	N/A	Ambulant (6)	N/A
KW-4/1	21	Yes	Calf hypertrophy	500	Ambulant (8)	Myopathic

KW-4/2	29	Yes	Cardio-respiratory involvement, urinary incontinence	185	WCD (18)	Myopathic
KW-5/1	18	Yes	-	712	WCD (19)	mixed
KW-5/2	25	Yes	-	N/A	WCD (8)	N/A
KW-5/3	25	Yes	-	N/A	WCD (14)	N/A
KW-5/4	25	No, severely affected quadriceps	Markedly weak quadriceps	N/A	WCD (30), lost ambulation 5 years from disease onset	N/A

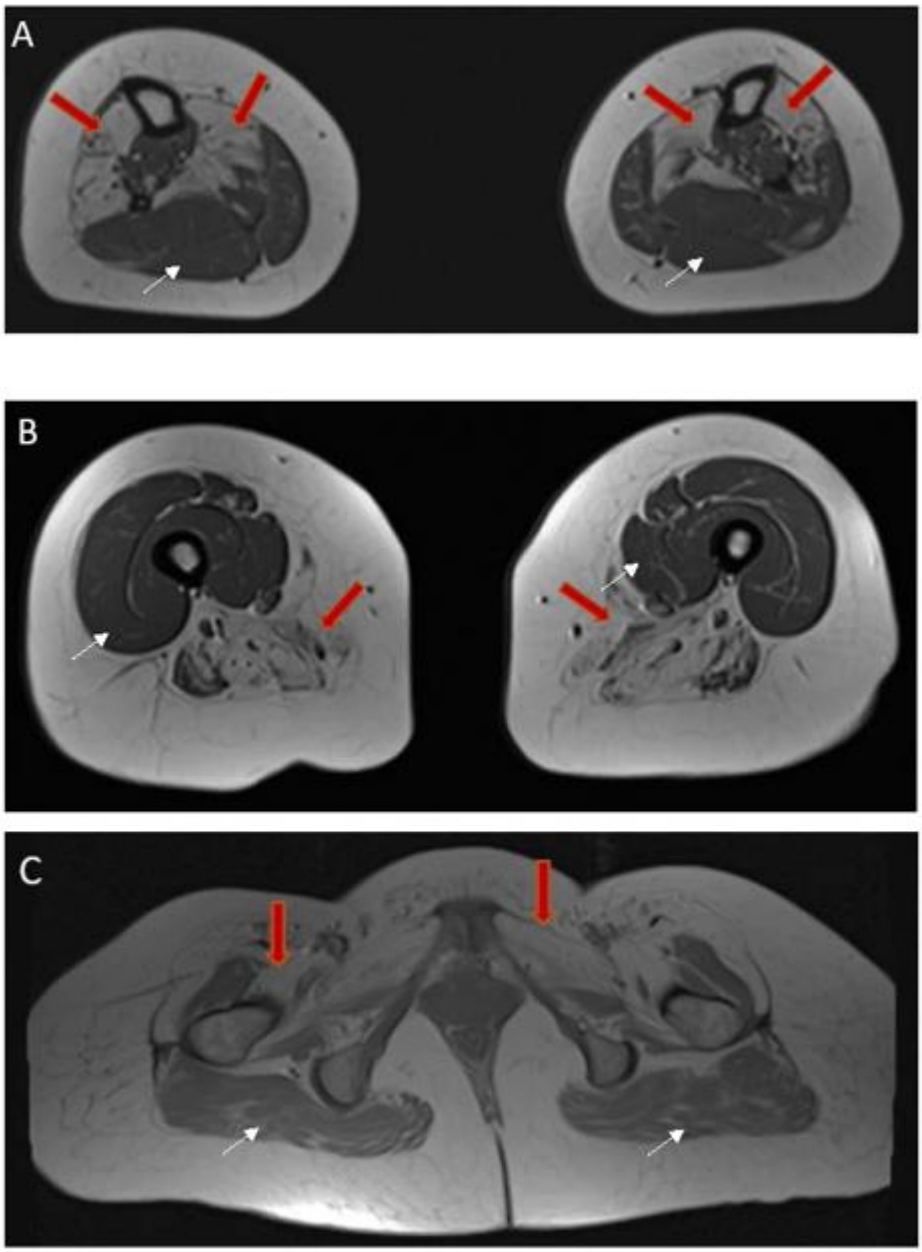
*reference CK values for males and females: 39-308 U/L and 26-192 U/L, respectively. CK; creatine kinase, WCD; wheelchair dependent.

Figure 39: MRI T1 weighted images, axial (A-C, for KW-2/1) and coronal (D, for KW-1/1) views. Images show the selective muscle involvement associated with GNE myopathy. Red and white arrows point to affected muscles with fatty replacement, and preserved muscles, respectively. A, Mid-calf, bilateral fatty tissue infiltration and atrophy of the anterior compartment (tibialis anterior, extensor digitorum muscles), lateral compartments (peroneus longus and brevis) and tibialis posterior from the deep posterior compartment. Posterior and deep posterior compartments (soleus, gastrocnemius and flexor digitorum longus) are preserved.

B, Mid-thigh, severe bilateral fatty tissue infiltration and atrophy of the posterior knee flexors, and medial hip adductor muscles. Quadriceps femoris and sartorius are spared.

C, pelvis, bilateral fatty tissue infiltration and atrophy of iliopsoas, pectineus, adductors magnus and brevis and paraspinal muscles. Both Gluteus maximus are relatively preserved.

D, bilateral severe fatty tissue infiltration and atrophy of thigh adductors. Psoas muscles are preserved.


A
B
C


D

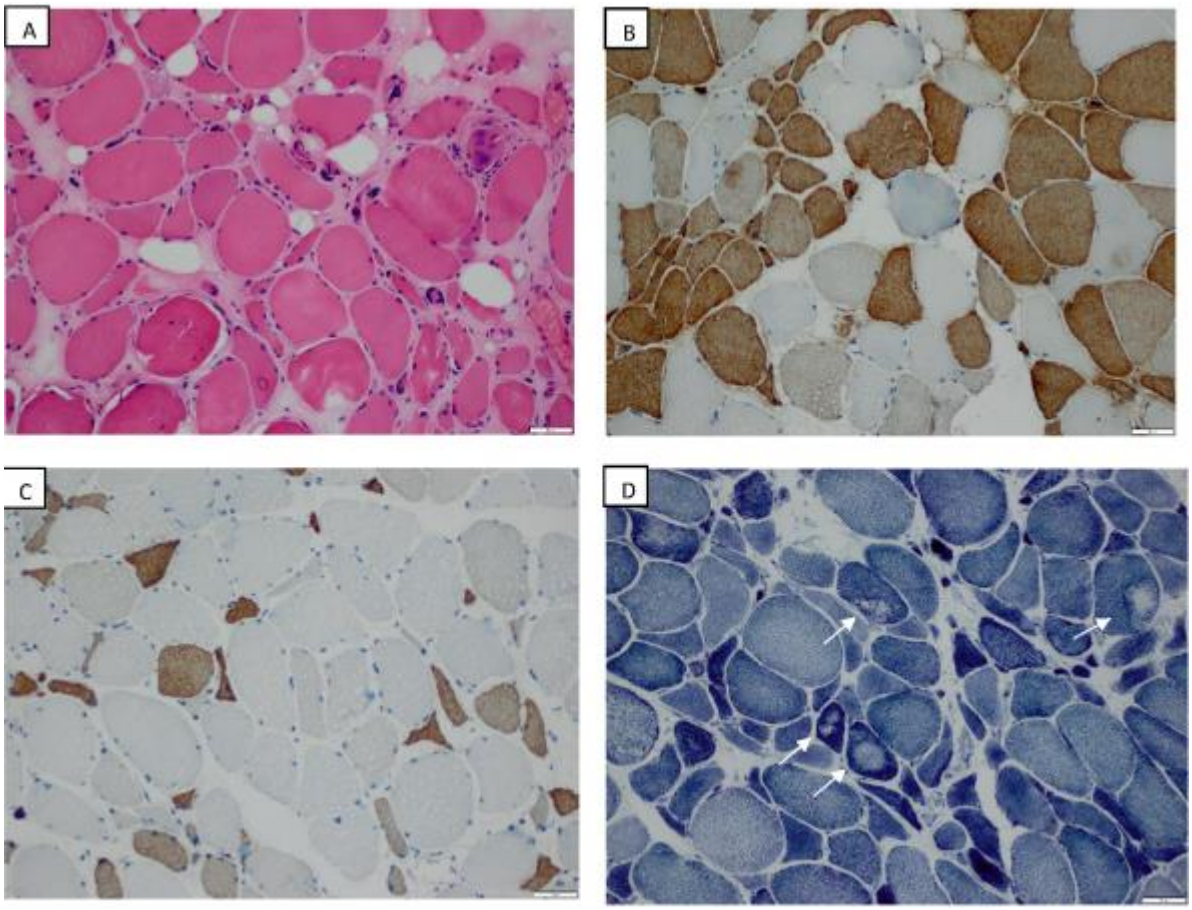


Figure 40: Histological and immunohistochemical findings from the hamstrings muscle for patient KW-1/1. A, H&E x200; fibre size variation and angulated atrophic fibres. B, myosin heavy chain (MHC) fast x200; Selective type 2 atrophy and type 1 fibre predominance. C, MHC neonatal; selective dark staining indicating regeneration in pathologic muscle fibres. D, NADH x200; intracytoplasmic cores (arrows) surrounded by occasional coarse NADH-positive granules.

5.4.3 Disease prevalence and p.M743T carrier frequency in the Kuwaiti population

According to national statistics (http://stat.paci.gov.kw/englishreports/#DataTabPlace:ColumnChartEduAge) the population of Kuwait is almost 4.4 million, from which around 100 000 constitute Bedouin Arab tribes. Therefore, we can estimate disease prevalence to be 14 per 100 000 (0.00014) in the Bedouin population in Kuwait and thus carrier frequency is expected to be 1 in 43 (0.0233). When screening 100 DNA samples belonging to Kuwaiti Bedouins for the Middle Eastern p.M743T mutation, no carriers were detected.

5.4.4 Patients with no GNE gene mutations in exon 12

Screening for the p.M743T mutation was performed in 37 cases (figure 37). 23/37 patients did not have the mutation. Four were tested because they were in a consanguineous marriage with, or they were related to an affected person (2 and 2, respectively). Eleven patients were not from Bedouin families. The remainder were descendant of Bedouin tribes distinct from the patients with GNE myopathy discussed above. These patients were from families either showing recessive (9), dominant (2) or sporadic (9) inheritance. Eight patients also showed sensory impairment in addition to distal weakness and the clinical picture appeared to fit with hereditary sensory motor neuropathy (HSMN). Features that were not consistent with GNE myopathy were: onset at birth (1), respiratory insufficiency (1), fluctuating weakness (3), vocal cord paralysis (1), and facial muscle weakness (1). Three patients had predominantly proximal weakness and two have a predominantly distal myopathy. For the latter two patients, GNE myopathy remains in the differential diagnoses.

5.5 Discussion

GNE Myopathy caused by the p.M743T mutation was first described in the Persian Jewish population from Iran. Soon after, the mutation was detected in Jewish communities from other countries in the Middle East (Argov and Mitrani Rosenbaum, 2015).

In Kuwait, GNE myopathy associated with the p.M743T mutation was first identified through sequencing of exon 12 of the *GNE* gene. The diagnosis was then considered for patients presenting with distal myopathy and relative quadriceps muscle sparing. Here, a description of the genetic, epidemiological aspects of five families of Bedouin Arab origin with GNE myopathy and the p.M743T mutation from Kuwait is given. This report further expands the

founder effect of the p.M743T mutation to the Arabian Gulf. There are also reports of p.M743T GNE myopathy in three unrelated Arab patients from Saudi Arabia (Personal communication, A. Urtizberea, Centre Neuromusculaire, Hendaye, France).

Due to high consanguinity and low genetic diversity in the studied cohort, sequencing of the whole *GNE* gene was not indicated (Alsmadi *et al.*, 2013). However, it is a possibility that other mutations are associated with GNE myopathy in the Kuwaiti population.

Pseudo-dominance was evident in this cohort as a mode of inheritance for GNE-myopathy in two families. Pseudo-dominance occurs when inheritance of an autosomal recessive disorder appears to mimic that of an autosomal dominant one, due to high frequency of the mutant allele within families. This phenomenon has previously been described in one pedigree of the Roma/Gypsies in Bulgaria, where cryptic consanguinity appeared as pseudo-dominant inheritance in GNE myopathy (Chamova *et al.*, 2015). It is also worth noting here that upon further investigation of extended family pedigrees, three out of five pedigrees (KW-1, 2 and 4) were found to share a fifth grandparent.

The carrier frequency for the p.M743T mutation in the Bedouin community was calculated to be 1 in 43. However, testing 100 DNA samples belonging to Bedouin families from Kuwait revealed no carriers. Assuming a 99% probability of observing at least one carrier, a larger sample of 200 individuals or more from the Bedouin population is required. Nevertheless, we now classify GNE-myopathy as the most common cause for inherited distal myopathy in Kuwait and recommend it be considered early in a diagnostic process in patients presenting with distal myopathy. We also recommend pre-marital counselling and carrier testing for the p.M743T mutation in the Bedouin community, due to the expected high carrier frequency in this group.

Clinically, there was variability in age of onset and disease progression rate amongst and within families despite harbouring the same mutation. This has been previously reported in other cohorts in the Middle East (Khademian *et al.*, 2013). In addition, severely affected quadriceps muscles progressing to early loss of ambulation, in an established quadriceps-sparing myopathy, was an early presenting feature in one patient in this cohort. Relative quadriceps sparing is a distinctive feature in GNE myopathy. If the quadriceps muscles are affected early in the disease course, diagnosis of GNE myopathy maybe missed or delayed and thus the diagnosis should not be dismissed in patients with weak quadriceps.

In addition, ophthalmoplegia, ptosis and tongue atrophy were found here in a patient with p.M743T- associated GNE myopathy. The possibility of two co-existing disorders in this patient remains high. However, atypical features associated with the GNE myopathy Middle Eastern p.M743T mutation have been reported and include marked quadriceps weakness, facial weakness and low frequency of proximal weakness (Argov *et al.*, 2003).

These findings may indicate a role for disease modifying genes or environmental factors. Dietary influences have previously been proposed as a disease-modifying mechanism in the Sangesar population of North Iran, where high dietary sialic acid content is thought to delay disease onset and produce a milder phenotype (Khademian *et al.*, 2013).

Furthermore, one of the patients suffered with cardiac and respiratory muscle dysfunction. Typically, GNE myopathy is not associated with cardio-respiratory disease. However, cohort based studies report varying degrees of structural and functional abnormalities in GNE myopathy patients. For example, patients homozygous for mutations in the kinase domain and at advanced disease stages were at higher risk of developing respiratory dysfunction (Mori-Yoshimura *et al.*, 2013). Also, an association of cardiac disease with GNE myopathy in the Roma/Gypsy population has been described (Chamova *et al.*, 2015).

On muscle biopsy, patients with GNE myopathy typically have rimmed vacuoles. Muscle biopsies from six patients here failed to show this. Rather, non-specific myopathic changes were seen. Although rimmed vacuoles are the most prominent feature on muscle biopsy in HIBM, non-specific myopathic features have been reported without rimmed vacuoles (Eisenberg *et al.*, 2003; Haghighi *et al.*, 2016). This can be attributed to the selective nature of the disease where relatively unaffected muscles are selected for biopsy (Haghighi *et al.*, 2016) or when the biopsy is done very early in the disease course. The latter possibility is because the accumulated "aggrephagy" producing the rimmed vacuoles results from the autophagy build up (Nishino *et al.*, 2015). Protein aggregates should nonetheless be seen earlier, however are difficult to visualise without immune-histochemical staining, for example P62 staining (Nishino *et al.*, 2015).

Further muscle biopsy findings included possible neuropathic changes such as fibre type grouping, selective fibre-type atrophy and the presence of target fibres. These were described in one patient. A potential explanation could be the decrease in sialylation, and its effects on proteins such as neural cell adhesion molecule (NCAM) or on proteases like neprilysin, proposed in some credible experiments (Malicdan *et al.*, 2008; Broccolini *et al.*, 2009). In

addition, NCAM is essential for muscle innervation and has been shown to be hypo-sialylated in GNE myopathy (Ricci *et al.*, 2006).

The patients described here underwent molecular testing for disorders such as calpainopathy, dysferlinopathy, mitochondrial myopathy and fascio-scapulo-humoural muscular dystrophy. In hindsight, these tests were unnecessary and added to costs per patient as well as causing diagnostic delays. However, with GNE myopathy being the most common cause of distal myopathy in the Bedouin community in Kuwait, initial screening for the p.M743T mutation is expected to decrease costs and reduce presentation-to-diagnosis time.

Finally, the exact mechanism of how *GNE* mutations cause myopathy remains under study. Normal sialylation in skeletal muscle is key to stabilisation and function of muscle glycoproteins and changes in sialylation of skeletal muscle glycoproteins may impact cell adhesion and signal transduction causing myofibrillar degeneration (Hinderlich *et al.*, 2004; Chamova *et al.*, 2015). Evidence for altered sialylation of skeletal muscle glycoproteins in GNE myopathy however, has been contradictory. This mechanism remains controversial and further research is needed (Hinderlich *et al.*, 2004; Huizing *et al.*, 2004; Broccolini *et al.*, 2008).

Nonetheless, sialic acid replacement therapy has been tested in a phase III trial for the treatment of GNE-myopathy (Mori-Yoshimura and Nishino, 2015; Suzuki *et al.*, 2017). The families described here represent a homogenous cohort for an ultra-rare disease. Continued identification and characterization of such patients is warranted to develop on the understanding of the natural history of the disease, define outcome measures and eventually recruitment into international disease registries, disease monitoring programmes and clinical trials (Pogoryelova *et al.*, 2018).

Chapter 6. Conclusions and future directions

Advances in NGS technology over the past few years have led to a decrease in per-base sequencing cost, increased accuracy and increased speed of generating sequence data. These advances have also led to identification of the mutated genes in many of the rare Mendelian diseases including rare NMD (Laing, 2012; Lee *et al.*, 2014a).

NMD are a complex group of disorders to diagnose due to their genetic heterogeneity. For example, mutations in at least 29 genes are responsible for causing the autosomal recessive limb girdle muscular dystrophies alone (WMS, Muscle Gene Table, May 2018 http://www.musclegenetable.fr/). In addition, multiple types of mutations in these genes result in a NMD, requiring laboratories to perform multiple diagnostic techniques to detect them (Laing, 2012).

Adding to the diagnostic complexity in this group is that multiple phenotypes are caused by mutations in the same gene. For instance, mutations in *RYR1* are associated with a range of muscle pathologies that are inherited as autosomal recessive, dominant or as an inherited susceptibility for malignant hyperthermia or rhabdomyolysis (Voermans *et al.*, 2016; Witherspoon and Meilleur, 2016). In addition, mutations in *LMNA* have the most clinically varied phenotypes, where the gene is associated with progeria, congenital malformations, lipodystrophy, dermopathy, cardiomyopathy, Charcot-Marie Tooth disease and a spectrum of muscular dystrophies (OMIM 150330). Furthermore, some of the genes associated with NMD, such as *DMD*, *TTN*, *RYR1* and *NEB*, are very large genes; sequencing of these genes via traditional methods (Sanger sequencing) is costly as well as time and labour exhaustive (Laing, 2012).

The genetic, phenotypic and mutation heterogeneity of NMD makes NGS an ideal method for diagnosis and novel gene discovery. For the genetically heterogeneous LGMDs, gene panel tests may reduce the time and cost associated with sequential gene testing and may replace the need for invasive tests such as muscle biopsies as well as increase diagnostic yield in this group of patients. In addition, WES and WGS are designed to cover all the genes associated with NMD, particularly the larger genes, as well as the remainder of the genome for novel gene discoveries. Also, with advances in the sequencing technologies and the computation analysis of sequencing data, these tests are powered to detect multiple mutation types including SNV, InDels and CNV (Laing, 2012; Lek and MacArthur, 2014; Efthymiou *et al.*, 2016). Nonetheless, the majority of patients with rare NMD remain without a molecular diagnosis. In part, this may be due to our limited knowledge of the human genome and that

the defect is an undiscovered novel genetic association. It is also possible that in some cases the disease is not genetic in origin or is a multigenic disorder. However, many of the undiagnosed cases can be attributed to limitations of the NGS sequencing technology and the computational analysis of the data. Comparisons of the most widely used sequencing platforms and bioinformatics analysis tools have shown that challenges arise at the many stages of the NGS process. From the sequencing technology itself, the initial processing of the data, genome alignment and variant calling, to the process of variant annotation and prioritisation (Clark *et al.*, 2011; Cornish and Guda, 2015; Lelieveld *et al.*, 2015; Laurie *et al.*, 2016; Yen *et al.*, 2017).

The work presented in this thesis utilised three NGS bioinformatics pipelines and genomic platforms used for data analysis and variant prioritization for patients with rare NMD. NGS data processing occurred at the Broad Institute (USA), deCODE Genetics (Iceland) and RD-Connect (CNAG, Barcelona) and the variant call files (VCF/gVCF) were then uploaded on *seqr*, CSA, and the RD-Connect Genome-Phenome Analysis Platform, respectively, for variant filtering and prioritisation.

The initial aim here was to compare data from WES and WGS experiments for patients with NMD and to assess the limitations of WES when examining NMD genes. The analysis used real patient data and revealed that WES was limited by low coverage in a number of coding regions across several NMD genes. Investigating these regions identified high GC content and low sequence heterogeneity as reasons for the low coverage. These were overcome by the longer reads and the PCR-free WGS experiment and therefore, additional potentially disease-causing variants were called in these regions.

WGS provides uniform coverage of the genome and thus higher sensitivity in variant detection in coding regions by overcoming issues related to capture kits, GC-bias and read specificity. This higher sensitivity in variant detection applies to SNVs, InDels, CNV and trinucleotide repeat expansions, which are all relevant mutations in NMD (Laing, 2012). In addition, some of the undiagnosed cases in NMD may be accounted for by mutations in non-coding regions. Nonetheless, data analysis from WGS is more challenging and the risk of VUS is higher. Therefore, as WGS costs continue to decline, and the technology is further used in research and diagnostics, virtual gene panels or exome targeted analysis of WGS data may be an appropriate initial approach (Meienberg *et al.*, 2016).

The next part of the analysis focused on a comparison between the bioinformatics pipelines at deCODE Genetics, the Broad Institute and RD-Connect and their respective variant

prioritisation platforms. The analysis used real patient data as well as the reference GIAB sample NA12878 and a number of important conclusions were made. The analysis highlighted that, although high quality reference data is useful for benchmarking and development of the NGS bioinformatics pipelines (Roy *et al.*, 2018), the use of real patient data is also informative and may provide insights on the NGS experiment as a whole and the effects of the quality of the sequencing on the downstream analysis. This was highlighted by the fact that the bioinformatics pipelines had high variant agreement rates when the reference sample was used (up to 91%) however, this was significantly lower when using patient samples (mean 75%). Using patient WGS data, the discrepancies between the platforms were magnified and the platform variant agreement was as low as 13%.

A further important conclusion is that variant discrepancies were more evident in the platform output reports than in the VCF files. This suggested that variant annotation and platform filtering algorithms vary considerably between the three sites. This was also highlighted when the reference genome was used.

Use of patient data highlighted the discrepancies and suggested the sequencing quality, annotation and filtering algorithms as contributors to these discrepancies. However, they cannot be used to assess sensitivity of the bioinformatics pipelines in detecting variants. In this context, use of high quality reference data may be more appropriate for performance comparison of multiple bioinformatics pipelines as sensitivity in variant detection can be accurately defined against a well-characterised reference (Cornish and Guda, 2015; Laurie *et al.*, 2016).

Communication is crucial between clinical researchers, clinical and molecular geneticists and software tool and analysis platform developers. During the course of the work presented in this thesis, errors and drawbacks of the platforms were reported back to their developers and subsequently actions were taken. For example, the high number of incorrectly annotated InDels in the RD-Connect Genome-Phenome Analysis Platform output reports was discussed with the team in Barcelona and plans were put in place to replace the annotation tool in their pipeline. In addition, the platform also now operates an internal frequency filter in the algorithm where variants with high allele frequency in the RD-Connect database are removed from the output report so that passing variants reflect only true rare ones (personal communication from Hanns Lochmüller).

Overall, the NGS data analysis pipeline comparison presented here constitutes an observational study that highlighted discrepancies and suggested levels in the pipeline where

issues may arise. A more analytical approach looking into the algorithms should be undertaken by bioinformaticians and software developers.

A further aim of the work presented here, was to assess the value of using an integrated genomics platform for WES variant prioritisation and diagnosis in patients with rare NMD. Analysis of WES data from 93 families with a phenotype of limb girdle weakness led to a candidate diagnosis in 65.6% of families including 14 novel candidate genes. This included the novel *MYMK* (Alrohaif *et al.*, 2018) gene and the strong candidate *FILIP1* (manuscript in preparation). These candidates were highlighted on the *seqr* platform due to the integrated tools and databases that enabled filtering variants for population frequency, variant quality, predicted damaging effect and inheritance model. In addition, evaluation of the variants in the output reports was feasible on the platform where information on disease associations and tissue expression levels as well as external links to public databases is incorporated. All candidates need to be further confirmed by Sanger sequencing, functional studies and by identifying additional confirmatory families, which requires cross-project data sharing and matchmaking such as, through RD-Connect (Lochmuller *et al.*, 2018).

Finally, following a confirmed genetic diagnosis in patients with GNE myopathy in the Bedouin population of Kuwait, a description of the clinical and demographic aspects in this cohort was given. Unusual phenotypic findings such as facial muscle and cardiorespiratory involvement, and severely affected quadriceps, were described in this cohort. Accurate molecular diagnosis is essential to phenotype-genotype correlation studies, which will aid in the understanding of the disease course and in the provision of homogenous cohorts for natural history studies and clinical trials. This study also highlighted the expected high carrier frequency for the Middle Eastern p.M743T mutation that will have implications on pre-marital counselling and testing the Bedouin population in Kuwait.

As with all rare diseases, NMD patients benefit from a molecular diagnosis. A molecular diagnosis leads to an accurate clinical diagnosis and directs disease management, surveillance and genetic counselling. In research, an accurate diagnosis allows for genotype-phenotype correlation studies and deeper insights into the course of a disease (Amburgey *et al.*, 2013; Pogoryelova *et al.*, 2018). In addition, most clinical trials into inherited disorders require a confirmed genetic diagnosis prior to recruitment in to the study (Bushby *et al.*, 2014; Goemans *et al.*, 2016). Furthermore, the genetic aetiology increases our knowledge of disease patho-mechanisms and provides novel therapeutic targets (Natrajan and Wilkerson, 2013; Borad *et al.*, 2016; Ruiz-Martinez *et al.*, 2017).

WES and WGS have had a remarkable impact on diagnosis and novel gene discoveries and are expected to continue to identify new genes associated with NMD. Understanding technical limitations in the sequencing methods and the downstream data analysis allows innovative methods to be developed to increase the accuracy of variant calling and annotation and the efficacy of variant prioritisation methods. These improvements may solve a proportion of the undiagnosed cases. In addition, use of orthogonal experiments such as metabolomics, transcriptomics or proteomics will increase the number of novel discoveries as well as confirming pathogenicity by investigating the functional effects of variants (Boycott *et al.*, 2017). A further proportion is likely to be solved by data sharing and matchmaking through systematic collaborations between researchers and clinicians worldwide (Philippakis *et al.*, 2015; Lochmuller *et al.*, 2018).

Chapter 7. References

Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E. and McVean, G.A. (2010) 'A map of human genome variation from population-scale sequencing', *Nature*, 467(7319), pp. 1061-73.

Abnizova, I., Boekhorst, R.t. and Orlov, Y.L. (2017) 'Computational Errors and Biases in Short Read Next Generation

Sequencing', *Journal of Proteomics & Bioinformatics*, 10(1), pp. 1-17.

Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C. and Gnirke, A. (2011) 'Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries', *Genome Biol*, 12(2), p. R18.

Akita, T., Aoto, K., Kato, M., Shiina, M., Mutoh, H., Nakashima, M., Kuki, I., Okazaki, S., Magara, S., Shiihara, T., Yokochi, K., Aiba, K., Tohyama, J., Ohba, C., Miyatake, S., Miyake, N., Ogata, K., Fukuda, A., Matsumoto, N. and Saitsu, H. (2018) 'De novo variants in CAMK2A and CAMK2B cause neurodevelopmental disorders', *Ann Clin Transl Neurol*, 5(3), pp. 280-296.

Alexander, J., Mantzaris, D., Georgitsi, M., Drineas, P. and Paschou, P. (2017) 'Variant Ranker: a web-tool to rank genomic data according to functional significance', *BMC Bioinformatics*, 18(1), p. 341.

Alioto, T.S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M.D., Hovig, E., Heisler, L.E., Beck, T.A., Simpson, J.T., Tonon, L., Sertier, A.S., Patch, A.M., Jager, N., Ginsbach, P., Drews, R., Paramasivam, N., Kabbe, R., Chotewutmontri, S., Diessl, N., Previti, C., Schmidt, S., Brors, B., Feuerbach, L., Heinold, M., Grobner, S., Korshunov, A., Tarpey, P.S., Butler, A.P., Hinton, J., Jones, D., Menzies, A., Raine, K., Shepherd, R., Stebbings, L., Teague, J.W., Ribeca, P., Giner, F.C., Beltran, S., Raineri, E., Dabad, M., Heath, S.C., Gut, M., Denroche, R.E., Harding, N.J., Yamaguchi, T.N., Fujimoto, A., Nakagawa, H., Quesada, V., Valdes-Mas, R., Nakken, S., Vodak, D., Bower, L., Lynch, A.G., Anderson, C.L., Waddell, N., Pearson, J.V., Grimmond, S.M., Peto, M., Spellman, P., He, M., Kandoth, C., Lee, S., Zhang, J., Letourneau, L., Ma, S., Seth, S., Torrents, D., Xi, L., Wheeler, D.A., Lopez-Otin, C., Campo, E., Campbell, P.J., Boutros, P.C., Puente, X.S., Gerhard, D.S., Pfister, S.M., McPherson, J.D., Hudson, T.J., Schlesner, M., Lichter, P., Eils, R., Jones, D.T. and Gut, I.G. (2015) 'A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing', *Nat Commun*, 6, p. 10001.

Allali, I., Arnold, J.W., Roach, J., Cadenas, M.B., Butz, N., Hassan, H.M., Koci, M., Ballou, A., Mendoza, M., Ali, R. and Azcarate-Peril, M.A. (2017) 'A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome', *BMC Microbiol*, 17(1), p. 194.

Alrohaif, H., Topf, A., Evangelista, T., Lek, M., McArthur, D. and Lochmuller, H. (2018) 'Whole-exome sequencing identifies mutations in MYMK in a mild form of Carey-Fineman-Ziter syndrome', *Neurol Genet*, 4(2), p. e226.

Alsmadi, O., Thareja, G., Alkayal, F., Rajagopalan, R., John, S.E., Hebbar, P., Behbehani, K. and Thanaraj, T.A. (2013) 'Genetic substructure of Kuwaiti population reveals migration history', *PLoS One*, 8(9), p. e74913.

Amburgey, K., Bailey, A., Hwang, J.H., Tarnopolsky, M.A., Bonnemann, C.G., Medne, L., Mathews, K.D., Collins, J., Daube, J.R., Wellman, G.P., Callaghan, B., Clarke, N.F. and Dowling, J.J. (2013) 'Genotype-phenotype correlations in recessive RYR1-related myopathies', *Orphanet J Rare Dis*, 8, p. 117.

Ankala, A., da Silva, C., Gualandi, F., Ferlini, A., Bean, L.J., Collins, C., Tanner, A.K. and Hegde, M.R. (2015) 'A comprehensive genomic approach for neuromuscular diseases gives a high diagnostic yield', *Ann Neurol*, 77(2), pp. 206-14.
Antoniadi, T., Buxton, C., Dennis, G., Forrester, N., Smith, D., Lunt, P. and Burton-Jones, S. (2015) 'Application of targeted multi-gene panel testing for the diagnosis of inherited peripheral neuropathy provides a high diagnostic yield with unexpected phenotype-genotype variability', *BMC Med Genet*, 16, p. 84.
Araya, C.L., Cenik, C., Reuter, J.A., Kiss, G., Pande, V.S., Snyder, M.P. and Greenleaf, W.J. (2016) 'Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations', *Nat Genet*, 48(2), pp. 117-25.
Argov, Z., Eisenberg, I., Grabov-Nardini, G., Sadeh, M., Wirguin, I., Soffer, D. and Mitrani-Rosenbaum, S. (2003) 'Hereditary inclusion body myopathy: the Middle Eastern genetic cluster', *Neurology*, 60(9), pp. 1519-23.
Argov, Z. and Mitrani Rosenbaum, S. (2015) 'GNE Myopathy: Two Clusters with History and Several Founder Mutations', *J Neuromuscul Dis*, 2(s2), pp. S73-S76.
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.
Babiker, T., Vedovato, N., Patel, K., Thomas, N., Finn, R., Mannikko, R., Chakera, A.J., Flanagan, S.E., Shepherd, M.H., Ellard, S., Ashcroft, F.M. and Hattersley, A.T. (2016) 'Successful transfer to sulfonylureas in KCNJ11 neonatal diabetes is determined by the mutation and duration of diabetes', *Diabetologia*.
Bao, R., Huang, L., Andrade, J., Tan, W., Kibbe, W.A., Jiang, H. and Feng, G. (2014) 'Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing', *Cancer Inform*, 13(Suppl 2), pp. 67-82.
Bauche, S., O'Regan, S., Azuma, Y., Laffargue, F., McMacken, G., Sternberg, D., Brochier, G., Buon, C., Bouzidi, N., Topf, A., Lacene, E., Remerand, G., Beaufrere, A.M., Pebrel-Richard, C., Thevenon, J., El Chehadeh-Djebbar, S., Faivre, L., Duffourd, Y., Ricci, F., Mongini, T., Fiorillo, C., Astrea, G., Burloiu, C.M., Butoianu, N., Sandu, C., Servais, L., Bonne, G., Nelson, I., Desguerre, I., Nougues, M.C., Boeuf, B., Romero, N., Laporte, J., Boland, A., Lechner, D., Deleuze, J.F., Fontaine, B., Strochlic, L., Lochmuller, H., Eymard, B., Mayer, M. and Nicole, S. (2016) 'Impaired Presynaptic High-Affinity Choline Transporter Causes a Congenital Myasthenic Syndrome with Episodic Apnea', *Am J Hum Genet*, 99(3), pp. 753-761.
Beck, T.F., Mullikin, J.C. and Biesecker, L.G. (2016) 'Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants', *Clin Chem*, 62(4), pp. 647-54.
Becker, K., Di Donato, N., Holder-Espinasse, M., Andrieux, J., Cuisset, J.M., Vallee, L., Plessis, G., Jean, N., Delobel, B., Thuresson, A.C., Anneren, G., Ravn, K., Tumer, Z., Tinschert, S., Schrock, E., Jonch, A.E. and Hackmann, K. (2012) 'De novo microdeletions of chromosome 6q14.1-q14.3 and 6q12.1-q14.1 in two patients with intellectual disability - further delineation of the 6q14 microdeletion syndrome and review of the literature', *Eur J Med Genet*, 55(8-9), pp. 490-7.
Biancalana, V. and Laporte, J. (2015) 'Diagnostic use of Massively Parallel Sequencing in Neuromuscular Diseases: Towards an Integrated Diagnosis', *J Neuromuscul Dis*, 2(3), pp. 193-203.
Blauwendraat, C., Faghri, F., Pihlstrom, L., Geiger, J.T., Elbaz, A., Lesage, S., Corvol, J.C., May, P., Nicolas, A., Abramzon, Y., Murphy, N.A., Gibbs, J.R., Ryten, M., Ferrari, R., Bras, J., Guerreiro, R., Williams, J., Sims, R., Lubbe, S., Hernandez, D.G., Mok, K.Y., Robak, L.,

Campbell, R.H., Rogaeva, E., Traynor, B.J., Chia, R., Chung, S.J., Hardy, J.A., Brice, A., Wood, N.W., Houlden, H., Shulman, J.M., Morris, H.R., Gasser, T., Kruger, R., Heutink, P., Sharma, M., Simon-Sanchez, J., Nalls, M.A., Singleton, A.B. and Scholz, S.W. (2017) 'NeuroChip, an updated version of the NeuroX genotyping platform to rapidly screen for variants associated with neurological diseases', *Neurobiol Aging*, 57, pp. 247 e9-247 e13.
Boland, J.F., Chung, C.C., Roberson, D., Mitchell, J., Zhang, X., Im, K.M., He, J., Chanock, S.J., Yeager, M. and Dean, M. (2013) 'The new sequencer on the block: comparison of Life Technology's Proton sequencer to an Illumina HiSeq for whole-exome sequencing', *Hum Genet*, 132(10), pp. 1153-63.
Borad, M.J., Egan, J.B., Condjella, R.M., Liang, W.S., Fonseca, R., Ritacca, N.R., McCullough, A.E., Barrett, M.T., Hunt, K.S., Champion, M.D., Patel, M.D., Young, S.W., Silva, A.C., Ho, T.H., Halfdanarson, T.R., McWilliams, R.R., Lazaridis, K.N., Ramanathan, R.K., Baker, A., Aldrich, J., Kurdoglu, A., Izatt, T., Christoforides, A., Cherni, I., Nasser, S., Reiman, R., Cuyugan, L., McDonald, J., Adkins, J., Mastrian, S.D., Valdez, R., Jaroszewski, D.E., Von Hoff, D.D., Craig, D.W., Stewart, A.K., Carpten, J.D. and Bryce, A.H. (2016) 'Clinical Implementation of Integrated Genomic Profiling in Patients with Advanced Cancers', *Sci Rep*, 6(1), p. 25.
Boycott, K.M., Rath, A., Chong, J.X., Hartley, T., Alkuraya, F.S., Baynam, G., Brookes, A.J., Brudno, M., Carracedo, A., den Dunnen, J.T., Dyke, S.O.M., Estivill, X., Goldblatt, J., Gonthier, C., Groft, S.C., Gut, I., Hamosh, A., Hieter, P., Hohn, S., Hurles, M.E., Kaufmann, P., Knoppers, B.M., Krischer, J.P., Macek, M., Jr., Matthijs, G., Olry, A., Parker, S., Paschall, J., Philippakis, A.A., Rehm, H.L., Robinson, P.N., Sham, P.C., Stefanov, R., Taruscio, D., Unni, D., Vanstone, M.R., Zhang, F., Brunner, H., Bamshad, M.J. and Lochmuller, H. (2017) 'International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases', *Am J Hum Genet*, 100(5), pp. 695-705.
Brewer, M.H., Chaudhry, R., Qi, J., Kidambi, A., Drew, A.P., Menezes, M.P., Ryan, M.M., Farrar, M.A., Mowat, D., Subramanian, G.M., Young, H.K., Zuchner, S., Reddel, S.W., Nicholson, G.A. and Kennerson, M.L. (2016) 'Whole Genome Sequencing Identifies a 78 kb Insertion from Chromosome 8 as the Cause of Charcot-Marie-Tooth Neuropathy CMTX3', *PLoS Genet*, 12(7), p. e1006177.
Briggs, C.E., Rucinski, D., Rosenfeld, P.J., Hirose, T., Berson, E.L. and Dryja, T.P. (2001) 'Mutations in ABCR (ABCA4) in patients with Stargardt macular degeneration or cone-rod degeneration', *Invest Ophthalmol Vis Sci*, 42(10), pp. 2229-36.
Broccolini, A., Gidaro, T., De Cristofaro, R., Morosetti, R., Gliubizzi, C., Ricci, E., Tonali, P.A. and Mirabella, M. (2008) 'Hyposialylation of neprilysin possibly affects its expression and enzymatic activity in hereditary inclusion-body myopathy muscle', *J Neurochem*, 105(3), pp. 971-81.
Broccolini, A., Gidaro, T., Morosetti, R. and Mirabella, M. (2009) 'Hereditary inclusion-body myopathy: clues on pathogenesis and possible therapy', *Muscle Nerve*, 40(3), pp. 340-349.
Bushby, K., Finkel, R., Wong, B., Barohn, R., Campbell, C., Comi, G.P., Connolly, A.M., Day, J.W., Flanigan, K.M., Goemans, N., Jones, K.J., Mercuri, E., Quinlivan, R., Renfroe, J.B., Russman, B., Ryan, M.M., Tulinius, M., Voit, T., Moore, S.A., Lee Sweeney, H., Abresch, R.T., Coleman, K.L., Eagle, M., Florence, J., Gappmaier, E., Glanzman, A.M., Henricson, E., Barth, J., Elfring, G.L., Reha, A., Spiegel, R.J., O'Donnell M, W., Peltz, S.W. and McDonald, C.M. (2014) 'Ataluren treatment of patients with nonsense mutation dystrophinopathy', *Muscle Nerve*, 50(4), pp. 477-87.
Butkiewicz, M. and Bush, W.S. (2016) 'In Silico Functional Annotation of Genomic Variation', *Curr Protoc Hum Genet*, 88, p. Unit 6 15.
Cao, H., Wu, H., Luo, R., Huang, S., Sun, Y., Tong, X., Xie, Y., Liu, B., Yang, H., Zheng, H., Li, J., Li, B., Wang, Y., Yang, F., Sun, P., Liu, S., Gao, P., Huang, H., Sun, J., Chen, D., He, G., Huang,

W., Huang, Z., Li, Y., Tellier, L.C., Liu, X., Feng, Q., Xu, X., Zhang, X., Bolund, L., Krogh, A., Kristiansen, K., Drmanac, R., Drmanac, S., Nielsen, R., Li, S., Wang, J., Yang, H., Li, Y., Wong, G.K. and Wang, J. (2015) 'De novo assembly of a haplotype-resolved human genome', *Nat Biotechnol*, 33(6), pp. 617-22.
Cappuccio, G., Vitiello, F., Casertano, A., Fontana, P., Genesio, R., Bruzzese, D., Ginocchio, V.M., Mormile, A., Nitsch, L., Andria, G. and Melis, D. (2016) 'New insights in the interpretation of array-CGH: autism spectrum disorder and positive family history for intellectual disability predict the detection of pathogenic variants', *Ital J Pediatr*, 42, p. 39.
Carson, A.R., Smith, E.N., Matsui, H., Braekkan, S.K., Jepsen, K., Hansen, J.B. and Frazer, K.A. (2014) 'Effective filtering strategies to improve data quality from population-based whole exome sequencing studies', *BMC Bioinformatics*, 15, p. 125.
Carss, K.J., Arno, G., Erwood, M., Stephens, J., Sanchis-Juan, A., Hull, S., Megy, K., Grozeva, D., Dewhurst, E., Malka, S., Plagnol, V., Penkett, C., Stirrups, K., Rizzo, R., Wright, G., Josifova, D., Bitner-Glindzicz, M., Scott, R.H., Clement, E., Allen, L., Armstrong, R., Brady, A.F., Carmichael, J., Chitre, M., Henderson, R.H.H., Hurst, J., MacLaren, R.E., Murphy, E., Paterson, J., Rosser, E., Thompson, D.A., Wakeling, E., Ouwehand, W.H., Michaelides, M., Moore, A.T., Webster, A.R. and Raymond, F.L. (2017) 'Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease', *Am J Hum Genet*, 100(1), pp. 75-90.
Chamova, T., Guergueltcheva, V., Gospodinova, M., Krause, S., Cirak, S., Kaprelyan, A., Angelova, L., Mihaylova, V., Bichev, S., Chandler, D., Naydenov, E., Grudkova, M., Djukmedzhiev, P., Voit, T., Pogoryelova, O., Lochmuller, H., Goebel, H.H., Bahlo, M., Kalaydjieva, L. and Tournev, I. (2015) 'GNE myopathy in Roma patients homozygous for the p.I618T founder mutation', *Neuromuscul Disord*, 25(9), pp. 713-8.
Chaouch, A., Brennan, K.M., Hudson, J., Longman, C., McConville, J., Morrison, P.J., Farrugia, M.E., Petty, R., Stewart, W., Norwood, F., Horvath, R., Chinnery, P.F., Costigan, D., Winer, J., Polvikoski, T., Healy, E., Sarkozy, A., Evangelista, T., Pogoryelova, O., Eagle, M., Bushby, K., Straub, V. and Lochmuller, H. (2014) 'Two recurrent mutations are associated with GNE myopathy in the North of Britain', *J Neurol Neurosurg Psychiatry*, 85(12), pp. 1359-65.
Chen, R., Jiang, T., She, Y., Xie, S., Zhou, S., Li, C., Ou, J. and Liu, Y. (2018) 'Comprehensive analysis of lncRNAs and mRNAs with associated co-expression and ceRNA networks in C2C12 myoblasts and myotubes', *Gene*, 647, pp. 164-173.
Cho, A., Hayashi, Y.K., Monma, K., Oya, Y., Noguchi, S., Nonaka, I. and Nishino, I. (2014) 'Mutation profile of the GNE gene in Japanese patients with distal myopathy with rimmed vacuoles (GNE myopathy)', *J Neurol Neurosurg Psychiatry*, 85(8), pp. 914-7.
Chung, R.H., Tsai, W.Y., Kang, C.Y., Yao, P.J., Tsai, H.J. and Chen, C.H. (2016) 'FamPipe: An Automatic Analysis Pipeline for Analyzing Sequencing Data in Families for Disease Studies', *PLoS Comput Biol*, 12(6), p. e1004980.
Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012) 'A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3', *Fly (Austin)*, 6(2), pp. 80-92.
Clark, M.J., Chen, R., Lam, H.Y., Karczewski, K.J., Chen, R., Euskirchen, G., Butte, A.J. and Snyder, M. (2011) 'Performance comparison of exome DNA sequencing technologies', *Nat Biotechnol*, 29(10), pp. 908-14.
Coenen-Stass, A.M., McClorey, G., Manzano, R., Betts, C.A., Blain, A., Saleh, A.F., Gait, M.J., Lochmuller, H., Wood, M.J. and Roberts, T.C. (2015) 'Identification of novel, therapy-responsive protein biomarkers in a mouse model of Duchenne muscular dystrophy by aptamer-based serum proteomics', *Sci Rep*, 5, p. 17014.

Coonrod, E.M., Margraf, R.L. and Voelkerding, K.V. (2011) 'Translating exome sequencing from research to clinical diagnostics', *Clin Chem Lab Med*, 50(7), pp. 1161-8.
Cornish, A. and Guda, C. (2015) '59103A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference', *Biomed Res Int*, 2015, p. 456479.
Cottrell, C.E., Al-Kateb, H., Bredemeyer, A.J., Duncavage, E.J., Spencer, D.H., Abel, H.J., Lockwood, C.M., Hagemann, I.S., O'Guin, S.M., Burcea, L.C., Sawyer, C.S., Oschwald, D.M., Stratman, J.L., Sher, D.A., Johnson, M.R., Brown, J.T., Cliften, P.F., George, B., McIntosh, L.D., Shrivastava, S., Nguyen, T.T., Payton, J.E., Watson, M.A., Crosby, S.D., Head, R.D., Mitra, R.D., Nagarajan, R., Kulkarni, S., Seibert, K., Virgin, H.W.t., Milbrandt, J. and Pfeifer, J.D. (2014) 'Validation of a next-generation sequencing assay for clinical molecular oncology', *J Mol Diagn*, 16(1), pp. 89-105.
Curnutte, M.A., Frumovitz, K.L., Bollinger, J.M., Cook-Deegan, R.M., McGuire, A.L. and Majumder, M.A. (2016) 'Developing context-specific next-generation sequencing policy', *Nat Biotechnol*, 34(5), pp. 466-70.
D'Aurizio, R., Pippucci, T., Tattini, L., Giusti, B., Pellegrini, M. and Magi, A. (2016) 'Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2', *Nucleic Acids Res*, 44(20), p. e154.
D'Avila, F., Meregalli, M., Lupoli, S., Barcella, M., Orro, A., De Santis, F., Sitzia, C., Farini, A., D'Ursi, P., Erratico, S., Cristofani, R., Milanesi, L., Braga, D., Cusi, D., Poletti, A., Barlassina, C. and Torrente, Y. (2016) 'Exome sequencing identifies variants in two genes encoding the LIM-proteins NRAP and FHL1 in an Italian patient with BAG3 myofibrillar myopathy', *J Muscle Res Cell Motil*, 37(3), pp. 101-15.
Daunert, S., Sittampalam, G.S. and Goldschmidt-Clermont, P.J. (2017) 'Twenty-First Century Diseases: Commonly Rare and Rarely Common?', *Antioxid Redox Signal*, 27(9), pp. 511-516.
de Ligt, J., Willemsen, M.H., van Bon, B.W., Kleefstra, T., Yntema, H.G., Kroes, T., Vulto-van Silfhout, A.T., Koolen, D.A., de Vries, P., Gilissen, C., del Rosario, M., Hoischen, A., Scheffer, H., de Vries, B.B., Brunner, H.G., Veltman, J.A. and Vissers, L.E. (2012) 'Diagnostic exome sequencing in persons with severe intellectual disability', *N Engl J Med*, 367(20), pp. 1921-9.
Desai, A.N. and Jere, A. (2012) 'Next-generation sequencing: ready for the clinics?', *Clin Genet*, 81(6), pp. 503-10.
Dewey, F.E., Grove, M.E., Pan, C., Goldstein, B.A., Bernstein, J.A., Chaib, H., Merker, J.D., Goldfeder, R.L., Enns, G.M., David, S.P., Pakdaman, N., Ormond, K.E., Caleshu, C., Kingham, K., Klein, T.E., Whirl-Carrillo, M., Sakamoto, K., Wheeler, M.T., Butte, A.J., Ford, J.M., Boxer, L., Ioannidis, J.P., Yeung, A.C., Altman, R.B., Assimes, T.L., Snyder, M., Ashley, E.A. and Quertermous, T. (2014) 'Clinical interpretation and implications of whole-genome sequencing', *JAMA*, 311(10), pp. 1035-45.
Di Gioia, S.A., Connors, S., Matsunami, N., Cannavino, J., Rose, M.F., Gilette, N.M., Artoni, P., de Macena Sobreira, N.L., Chan, W.M., Webb, B.D., Robson, C.D., Cheng, L., Van Ryzin, C., Ramirez-Martinez, A., Mohassel, P., Leppert, M., Scholand, M.B., Grunseich, C., Ferreira, C.R., Hartman, T., Hayes, I.M., Morgan, T., Markie, D.M., Fagiolini, M., Swift, A., Chines, P.S., Speck-Martins, C.E., Collins, F.S., Jabs, E.W., Bonnemann, C.G., Olson, E.N., Carey, J.C., Robertson, S.P., Manoli, I. and Engle, E.C. (2017) 'A defect in myoblast fusion underlies Carey-Fineman-Ziter syndrome', *Nat Commun*, 8, p. 16077.
Efthymiou, S., Manole, A. and Houlden, H. (2016) 'Next-generation sequencing in neuromuscular diseases', *Curr Opin Neurol*, 29(5), pp. 527-36.
Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R. and Ashburner, M. (2005) 'The Sequence Ontology: a tool for the unification of genome annotations', *Genome Biol*, 6(5), p. R44.

Eisenberg, I., Grabov-Nardini, G., Hochner, H., Korner, M., Sadeh, M., Bertorini, T., Bushby, K., Castellan, C., Felice, K., Mendell, J., Merlini, L., Shilling, C., Wirguin, I., Argov, Z. and Mitrani-Rosenbaum, S. (2003) 'Mutations spectrum of GNE in hereditary inclusion body myopathy sparing the quadriceps', *Hum Mutat*, 21(1), p. 99.
Ellingford, J.M., Barton, S., Bhaskar, S., Williams, S.G., Sergouniotis, P.I., O'Sullivan, J., Lamb, J.A., Perveen, R., Hall, G., Newman, W.G., Bishop, P.N., Roberts, S.A., Leach, R., Tearle, R., Bayliss, S., Ramsden, S.C., Nemeth, A.H. and Black, G.C. (2016) 'Whole Genome Sequencing Increases Molecular Diagnostic Yield Compared with Current Diagnostic Testing for Inherited Retinal Disease', *Ophthalmology*.
Eurordis *Survey of the delay in diagnosis for 8 rare diseases in Europe (EurordisCare2) Factsheet*. Available at: http://www.eurordis.org/sites/default/files/publications/Fact_Sheet_Eurordiscare2.pdf.
Eurordis (2005) *Rare diseases: understanding this public health priority*. Available at: http://www.eurordis.org/IMG/pdf/princeps_document-EN.pdf (Accessed: 19/04/2016).
Evila, A., Arumilli, M., Udd, B. and Hackman, P. (2016) 'Targeted next-generation sequencing assay for detection of mutations in primary myopathies', *Neuromuscul Disord*, 26(1), pp. 7-15.
Falk, M.J., Shen, L., Gonzalez, M., Leipzig, J., Lott, M.T., Stassen, A.P., Diroma, M.A., Navarro-Gomez, D., Yeske, P., Bai, R., Boles, R.G., Brilhante, V., Ralph, D., DaRe, J.T., Shelton, R., Terry, S.F., Zhang, Z., Copeland, W.C., van Oven, M., Prokisch, H., Wallace, D.C., Attimonelli, M., Krotoski, D., Zuchner, S. and Gai, X. (2015) 'Mitochondrial Disease Sequence Data Resource (MSeqDR): a global grass-roots consortium to facilitate deposition, curation, annotation, and integrated analysis of genomic data for the mitochondrial disease clinical and research communities', *Mol Genet Metab*, 114(3), pp. 388-96.
Fang, H., Wu, Y., Narzisi, G., O'Rawe, J.A., Barron, L.T., Rosenbaum, J., Ronemus, M., Iossifov, I., Schatz, M.C. and Lyon, G.J. (2014) 'Reducing INDEL calling errors in whole genome and exome sequencing data', *Genome Med*, 6(10), p. 89.
Fiore, R.N. and Goodman, K.W. (2016) 'Precision medicine ethics: selected issues and developments in next-generation sequencing, clinical oncology, and ethics', *Curr Opin Oncol*, 28(1), pp. 83-7.
Franke, A., Wollstein, A., Teuber, M., Wittig, M., Lu, T., Hoffmann, K., Nurnberg, P., Krawczak, M., Schreiber, S. and Hampe, J. (2006) 'GENOMIZER: an integrated analysis system for genome-wide association data', *Hum Mutat*, 27(6), pp. 583-8.
Frankish, A., Uszczynska, B., Ritchie, G.R., Gonzalez, J.M., Pervouchine, D., Petryszak, R., Mudge, J.M., Fonseca, N., Brazma, A., Guigo, R. and Harrow, J. (2015) 'Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction', *BMC Genomics*, 16 Suppl 8, p. S2.
Fukai, R., Hiraki, Y., Yofune, H., Tsurusaki, Y., Nakashima, M., Saitsu, H., Tanaka, F., Miyake, N. and Matsumoto, N. (2015) 'A case of autism spectrum disorder arising from a de novo missense mutation in POGZ', *J Hum Genet*, 60(5), pp. 277-9.
Gahl, W.A. and Tifft, C.J. (2011) 'The NIH Undiagnosed Diseases Program: lessons learned', *JAMA*, 305(18), pp. 1904-5.
Gan, K.A., Carrasco Pro, S., Sewell, J.A. and Bass, J.I.F. (2018) 'Identification of Single Nucleotide Non-coding Driver Mutations in Cancer', *Front Genet*, 9, p. 16.
Gao, J., Wu, H., Wang, L., Zhang, H., Duan, H., Lu, J. and Liang, Z. (2016) 'Validation of targeted next-generation sequencing for RAS mutation detection in FFPE colorectal cancer tissues: comparison with Sanger sequencing and ARMS-Scorpion real-time PCR', *BMJ Open*, 6(1), p. e009532.

Ghaoui, R., Cooper, S.T., Lek, M., Jones, K., Corbett, A., Reddel, S.W., Needham, M., Liang, C., Waddell, L.B., Nicholson, G., O'Grady, G., Kaur, S., Ong, R., Davis, M., Sue, C.M., Laing, N.G., North, K.N., MacArthur, D.G. and Clarke, N.F. (2015) 'Use of Whole-Exome Sequencing for Diagnosis of Limb-Girdle Muscular Dystrophy: Outcomes and Lessons Learned', *JAMA Neurol*, 72(12), pp. 1424-32.
Ghoneim, D.H., Myers, J.R., Tuttle, E. and Paciorkowski, A.R. (2014) 'Comparison of insertion/deletion calling algorithms on human next-generation sequencing data', *BMC Res Notes*, 7, p. 864.
Ghosh, P.S. and Zhou, L. (2012) 'The diagnostic utility of a commercial limb-girdle muscular dystrophy gene test panel', *J Clin Neuromuscul Dis*, 14(2), pp. 86-7.
Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A., Leach, R., Klein, R., Tearle, R., Bo, T., Pfundt, R., Yntema, H.G., de Vries, B.B., Kleefstra, T., Brunner, H.G., Vissers, L.E. and Veltman, J.A. (2014) 'Genome sequencing identifies major causes of severe intellectual disability', *Nature*, 511(7509), pp. 344-7.
Gilissen, C., Hoischen, A., Brunner, H.G. and Veltman, J.A. (2012) 'Disease gene identification strategies for exome sequencing', *Eur J Hum Genet*, 20(5), pp. 490-7.
Goemans, N.M., Tulinius, M., van den Hauwe, M., Kroksmark, A.K., Buyse, G., Wilson, R.J., van Deutekom, J.C., de Kimpe, S.J., Lourbakos, A. and Campion, G. (2016) 'Long-Term Efficacy, Safety, and Pharmacokinetics of Drisapersen in Duchenne Muscular Dystrophy: Results from an Open-Label Extension Study', *PLoS One*, 11(9), p. e0161955.
Gonzalez, M., Falk, M.J., Gai, X., Postrel, R., Schule, R. and Zuchner, S. (2015) 'Innovative genomic collaboration using the GENESIS (GEM.app) platform', *Hum Mutat*, 36(10), pp. 950-6.
Green, R.C., Berg, J.S., Grody, W.W., Kalia, S.S., Korf, B.R., Martin, C.L., McGuire, A.L., Nussbaum, R.L., O'Daniel, J.M., Ormond, K.E., Rehm, H.L., Watson, M.S., Williams, M.S. and Biesecker, L.G. (2013) 'ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing', *Genet Med*, 15(7), pp. 565-74.
Guethbjartsson, H., Georgsson, G.F., Guethjonsson, S.A., Valdimarsson, R., Sigurethsson, J.H., Stefansson, S.K., Masson, G., Magnusson, G., Palmason, V. and Stefansson, K. (2016) 'GORpipe: a query tool for working with sequence data based on a Genomic Ordered Relational (GOR) architecture', *Bioinformatics*, 32(20), pp. 3081-3088.
Hackman, P., Udd, B., Bonnemann, C.G. and Ferreiro, A. (2017) '219th ENMC International Workshop Titinopathies International database of titin mutations and phenotypes, Heemskerk, The Netherlands, 29 April-1 May 2016', *Neuromuscul Disord*, 27(4), pp. 396-407.
Haghighi, A., Nafissi, S., Qurashi, A., Tan, Z., Shamshiri, H., Nilipour, Y., Haghighi, A., Desnick, R.J. and Kornreich, R. (2016) 'Genetics of GNE myopathy in the non-Jewish Persian population', *Eur J Hum Genet*, 24(2), pp. 243-51.
Han, J.M., Jeong, S.J., Park, M.C., Kim, G., Kwon, N.H., Kim, H.K., Ha, S.H., Ryu, S.H. and Kim, S. (2012) 'Leucyl-tRNA synthetase is an intracellular leucine sensor for the mTORC1-signaling pathway', *Cell*, 149(2), pp. 410-24.
Hara, M., Ohba, C., Yamashita, Y., Saitsu, H., Matsumoto, N. and Matsuishi, T. (2015) 'De novo SHANK3 mutation causes Rett syndrome-like phenotype in a female patient', *Am J Med Genet A*, 167(7), pp. 1593-6.
Harris, E., Topf, A., Barresi, R., Hudson, J., Powell, H., Tellez, J., Hicks, D., Porter, A., Bertoli, M., Evangelista, T., Marini-Betollo, C., Magnusson, O., Lek, M., MacArthur, D., Bushby, K., Lochmuller, H. and Straub, V. (2017) 'Exome sequences versus sequential gene testing in the UK highly specialised Service for Limb Girdle Muscular Dystrophy', *Orphanet J Rare Dis*, 12(1), p. 151.

Hinderlich, S., Salama, I., Eisenberg, I., Potikha, T., Mantey, L.R., Yarema, K.J., Horstkorte, R., Argov, Z., Sadeh, M., Reutter, W. and Mitrani-Rosenbaum, S. (2004) 'The homozygous M712T mutation of UDP-N-acetylglucosamine 2-epimerase/N-acetylmannosamine kinase results in reduced enzyme activities but not in altered overall cellular sialylation in hereditary inclusion body myopathy', *FEBS Lett*, 566(1-3), pp. 105-9.
Hinderlich, S., Stasche, R., Zeitler, R. and Reutter, W. (1997) 'A bifunctional enzyme catalyzes the first two steps in N-acetylneuraminic acid biosynthesis of rat liver. Purification and characterization of UDP-N-acetylglucosamine 2-epimerase/N-acetylmannosamine kinase', *J Biol Chem*, 272(39), pp. 24313-8.
Hofmann, A.L., Behr, J., Singer, J., Kuipers, J., Beisel, C., Schraml, P., Moch, H. and Beerenwinkel, N. (2017) 'Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers', *BMC Bioinformatics*, 18(1), p. 8.
Huang, L., Popic, V. and Batzoglou, S. (2013) 'Short read alignment with populations of genomes', *Bioinformatics*, 29(13), pp. i361-70.
Huizing, M., Rakocevic, G., Sparks, S.E., Mamali, I., Shatunov, A., Goldfarb, L., Krasnewich, D., Gahl, W.A. and Dalakas, M.C. (2004) 'Hypoglycosylation of alpha-dystroglycan in patients with hereditary IBM due to GNE mutations', *Mol Genet Metab*, 81(3), pp. 196-202.
Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., Smith, J.D., Paeper, B., Nickerson, D.A., Dea, J., Dong, S., Gonzalez, L.E., Mandell, J.D., Mane, S.M., Murtha, M.T., Sullivan, C.A., Walker, M.F., Waqar, Z., Wei, L., Willsey, A.J., Yamrom, B., Lee, Y.H., Grabowska, E., Dalkic, E., Wang, Z., Marks, S., Andrews, P., Leotta, A., Kendall, J., Hakker, I., Rosenbaum, J., Ma, B., Rodgers, L., Troge, J., Narzisi, G., Yoon, S., Schatz, M.C., Ye, K., McCombie, W.R., Shendure, J., Eichler, E.E., State, M.W. and Wigler, M. (2014) 'The contribution of de novo coding mutations to autism spectrum disorder', *Nature*, 515(7526), pp. 216-21.
Jalali Sefid Dashti, M. and Gamieldien, J. (2017) 'A practical guide to filtering and prioritizing genetic variants', *Biotechniques*, 62(1), pp. 18-30.
Johnson, K., Topf, A., Bertoli, M., Phillips, L., Claeys, K.G., Stojanovic, V.R., Peric, S., Hahn, A., Maddison, P., Akay, E., Bastian, A.E., Lusakowska, A., Kostera-Pruszczyk, A., Lek, M., Xu, L., MacArthur, D.G. and Straub, V. (2017) 'Identification of GAA variants through whole exome sequencing targeted to a cohort of 606 patients with unexplained limb-girdle muscle weakness', *Orphanet J Rare Dis*, 12(1), p. 173.
Kang, P.B. (2013) 'Ethical issues in neurogenetic disorders', *Handb Clin Neurol*, 118, pp. 265-76.
Kessler, E.L., Nikkels, P.G. and van Veen, T.A. (2017) 'Disturbed Desmoglein-2 in the intercalated disc of pediatric patients with dilated cardiomyopathy', *Hum Pathol*, 67, pp. 101-108.
Khademian, H., Mehravar, E., Urtizberea, J., Sagoo, S., Sandoval, L., Carbajo, R., Darvish, B., Valles-Ayoub, Y. and Darvish, D. (2013) 'Prevalence of GNE p.M712T and hereditary inclusion body myopathy (HIBM) in Sangesar population of Northern Iran', *Clin Genet*, 84(6), pp. 589-92.
Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M.A. and Gerstein, M. (2016) 'Role of non-coding sequence variants in cancer', *Nat Rev Genet*, 17(2), pp. 93-108.
Kim, B.Y., Park, J.H., Jo, H.Y., Koo, S.K. and Park, M.H. (2017) 'Optimized detection of insertions/deletions (INDELs) in whole-exome sequencing data', *PLoS One*, 12(8), p. e0182272.
Kirkpatrick, B.E., Riggs, E.R., Azzariti, D.R., Miller, V.R., Ledbetter, D.H., Miller, D.T., Rehm, H., Martin, C.L. and Faucett, W.A. (2015) 'GenomeConnect: matchmaking between patients,

clinical laboratories, and researchers to improve genomic knowledge', *Hum Mutat*, 36(10), pp. 974-8.
Klepper, J. (2015) 'GLUT1 deficiency syndrome and ketogenic diet therapies: missing rare but treatable diseases?', *Dev Med Child Neurol*, 57(10), pp. 896-7.
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.
Kuhn, M., Glaser, D., Joshi, P.R., Zierz, S., Wenninger, S., Schoser, B. and Deschauer, M. (2016) 'Utility of a next-generation sequencing-based gene panel investigation in German patients with genetically unclassified limb-girdle muscular dystrophy', *J Neurol*, 263(4), pp. 743-50.
Kury, S., van Woerden, G.M., Besnard, T., Proietti Onori, M., Latypova, X., Towne, M.C., Cho, M.T., Prescott, T.E., Ploeg, M.A., Sanders, S., Stessman, H.A.F., Pujol, A., Distel, B., Robak, L.A., Bernstein, J.A., Denomme-Pichon, A.S., Lesca, G., Sellars, E.A., Berg, J., Carre, W., Busk, O.L., van Bon, B.W.M., Waugh, J.L., Deardorff, M., Hoganson, G.E., Bosanko, K.B., Johnson, D.S., Dabir, T., Holla, O.L., Sarkar, A., Tveten, K., de Bellescize, J., Braathen, G.J., Terhal, P.A., Grange, D.K., van Haeringen, A., Lam, C., Mirzaa, G., Burton, J., Bhoj, E.J., Douglas, J., Santani, A.B., Nesbitt, A.I., Helbig, K.L., Andrews, M.V., Begtrup, A., Tang, S., van Gassen, K.L.I., Juusola, J., Foss, K., Enns, G.M., Moog, U., Hinderhofer, K., Paramasivam, N., Lincoln, S., Kusako, B.H., Lindenbaum, P., Charpentier, E., Nowak, C.B., Cherot, E., Simonet, T., Ruivenkamp, C.A.L., Hahn, S., Brownstein, C.A., Xia, F., Schmitt, S., Deb, W., Bonneau, D., Nizon, M., Quinquis, D., Chelly, J., Rudolf, G., Sanlaville, D., Parent, P., Gilbert-Dussardier, B., Toutain, A., Sutton, V.R., Thies, J., Peart-Vissers, L., Boisseau, P., Vincent, M., Grabrucker, A.M., Dubourg, C., Tan, W.H., Verbeek, N.E., Granzow, M., Santen, G.W.E., Shendure, J., Isidor, B., Pasquier, L., Redon, R., Yang, Y., State, M.W., Kleefstra, T., Cogne, B., Petrovski, S., Retterer, K., Eichler, E.E., Rosenfeld, J.A., Agrawal, P.B., et al. (2017) 'De Novo Mutations in Protein Kinase Genes CAMK2A and CAMK2B Cause Intellectual Disability', *Am J Hum Genet*, 101(5), pp. 768-788.
Laing, N.G. (2012) 'Genetics of neuromuscular disorders', *Crit Rev Clin Lab Sci*, 49(2), pp. 33-48.
Laurie, S. (2016) *RD-Connect Update: WP5*. Presentation
Laurie, S., Fernandez-Callejo, M., Marco-Sola, S., Trotta, J.R., Camps, J., Chacon, A., Espinosa, A., Gut, M., Gut, I., Heath, S. and Beltran, S. (2016) 'From Wet-Lab to Variations: Concordance and Speed of Bioinformatics Pipelines for Whole Genome and Whole Exome Sequencing', *Hum Mutat*, 37(12), pp. 1263-1271.
Leber, Y., Ruparelia, A.A., Kirfel, G., van der Ven, P.F., Hoffmann, B., Merkel, R., Bryson-Richardson, R.J. and Furst, D.O. (2016) 'Filamin C is a highly dynamic protein associated with fast repair of myofibrillar microdamage', *Hum Mol Genet*, 25(13), pp. 2776-2788.
Lee, H., Deignan, J.L., Dorrani, N., Strom, S.P., Kantarci, S., Quintero-Rivera, F., Das, K., Toy, T., Harry, B., Yourshaw, M., Fox, M., Fogel, B.L., Martinez-Agosto, J.A., Wong, D.A., Chang, V.Y., Shieh, P.B., Palmer, C.G., Dipple, K.M., Grody, W.W., Vilain, E. and Nelson, S.F. (2014a) 'Clinical exome sequencing for genetic identification of rare Mendelian disorders', *JAMA*, 312(18), pp. 1880-7.

Lee, H., Lin, M.C., Kornblum, H.I., Papazian, D.M. and Nelson, S.F. (2014b) 'Exome sequencing identifies de novo gain of function missense mutation in KCND2 in identical twins with autism and seizures that slows potassium channel inactivation', *Hum Mol Genet*, 23(13), pp. 3481-9.
Lee, J.Y., Lee, J.H., Yeo, J.S. and Kim, J.J. (2013) 'A SNP Harvester Analysis to Better Detect SNPs of CCDC158 Gene That Are Associated with Carcass Quality Traits in Hanwoo', *Asian-Australas J Anim Sci*, 26(6), pp. 766-71.
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. and MacArthur, D.G. (2016) 'Analysis of protein-coding genetic variation in 60,706 humans', *Nature*, 536(7616), pp. 285-91.
Lek, M. and MacArthur, D. (2014) 'The Challenge of Next Generation Sequencing in the Context of Neuromuscular Diseases', *J Neuromuscul Dis*, 1(2), pp. 135-149.
Lelieveld, S.H., Spielmann, M., Mundlos, S., Veltman, J.A. and Gilissen, C. (2015) 'Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions', *Hum Mutat*, 36(8), pp. 815-22.
Li, M.H., Abrudan, J.L., Dulik, M.C., Sasson, A., Brunton, J., Jayaraman, V., Dugan, N., Haley, D., Rajagopalan, R., Biswas, S., Sarmady, M., DeChene, E.T., Deardorff, M.A., Wilkens, A., Noon, S.E., Scarano, M.I., Santani, A.B., White, P.S., Pennington, J., Conlin, L.K., Spinner, N.B., Krantz, I.D. and Vetter, V.L. (2015) 'Utility and limitations of exome sequencing as a genetic diagnostic tool for conditions associated with pediatric sudden cardiac arrest/sudden cardiac death', *Hum Genomics*, 9, p. 15.
Li, M.J. and Wang, J. (2015) 'Current trend of annotating single nucleotide variation in humans--A case study on SNVrap', *Methods*, 79-80, pp. 32-40.
Liskova, P., Dudakova, L., Evans, C.J., Rojas Lopez, K.E., Pontikos, N., Athanasiou, D., Jama, H., Sach, J., Skalicka, P., Stranecky, V., Kmoch, S., Thaung, C., Filipec, M., Cheetham, M.E., Davidson, A.E., Tuft, S.J. and Hardcastle, A.J. (2018) 'Ectopic GRHL2 Expression Due to Non-coding Mutations Promotes Cell State Transition and Causes Posterior Polymorphous Corneal Dystrophy 4', *Am J Hum Genet*, 102(3), pp. 447-459.
Liu, J.S., Fan, L.L., Zhang, H., Liu, X., Huang, H., Tao, L.J., Xia, K. and Xiang, R. (2017a) 'Whole-Exome Sequencing Identifies Two Novel TTN Mutations in Chinese Families with Dilated Cardiomyopathy', *Cardiology*, 136(1), pp. 10-14.
Liu, Q., Zhang, P., Wang, D., Gu, W. and Wang, K. (2017b) 'Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing', *Genome Med*, 9(1), p. 65.
Lo, H.P., Cooper, S.T., Evesson, F.J., Seto, J.T., Chiotis, M., Tay, V., Compton, A.G., Cairns, A.G., Corbett, A., MacArthur, D.G., Yang, N., Reardon, K. and North, K.N. (2008) 'Limb-girdle muscular dystrophy: diagnostic evaluation, frequency and clues to pathogenesis', *Neuromuscul Disord*, 18(1), pp. 34-44.

Lo, Y., Kang, H.M., Nelson, M.R., Othman, M.I., Chissoe, S.L., Ehm, M.G., Abecasis, G.R. and Zollner, S. (2015) 'Comparing variant calling algorithms for target-exon sequencing in a large sample', *BMC Bioinformatics*, 16, p. 75.
Lochmuller, H., Badowska, D.M., Thompson, R., Knoers, N.V., Aartsma-Rus, A., Gut, I., Wood, L., Harmuth, T., Durudas, A., Graessner, H., Schaefer, F. and Riess, O. (2018) 'RD-Connect, NeurOmics and EURenOmics: collaborative European initiative for rare diseases', *Eur J Hum Genet*.
Long, P.A., Larsen, B.T., Evans, J.M. and Olson, T.M. (2015) 'Exome Sequencing Identifies Pathogenic and Modifier Mutations in a Child With Sporadic Dilated Cardiomyopathy', *J Am Heart Assoc*, 4(12).
Lupski, J.R., Reid, J.G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D.C., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D.A., McGuire, A.L., Zhang, F., Stankiewicz, P., Halperin, J.J., Yang, C., Gehman, C., Guo, D., Irikat, R.K., Tom, W., Fantin, N.J., Muzny, D.M. and Gibbs, R.A. (2010) 'Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy', *N Engl J Med*, 362(13), pp. 1181-91.
MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A., Barrett, J.C., Biesecker, L.G., Conrad, D.F., Cooper, G.M., Cox, N.J., Daly, M.J., Gerstein, M.B., Goldstein, D.B., Hirschhorn, J.N., Leal, S.M., Pennacchio, L.A., Stamatoyannopoulos, J.A., Sunyaev, S.R., Valle, D., Voight, B.F., Winckler, W. and Gunter, C. (2014) 'Guidelines for investigating causality of sequence variants in human disease', *Nature*, 508(7497), pp. 469-76.
Majewski, J. and Rosenblatt, D.S. (2012) 'Exome and whole-genome sequencing for gene discovery: the future is now!', *Hum Mutat*, 33(4), pp. 591-2.
Malicdan, M.C., Noguchi, S. and Nishino, I. (2008) 'Recent advances in distal myopathy with rimmed vacuoles (DMRV) or hIBM: treatment perspectives', *Curr.Opin.Neurol*, 21(5), pp. 596-600.
Mascarello, F., Toniolo, L., Cancellara, P., Reggiani, C. and Maccatrozzo, L. (2016) 'Expression and identification of 10 sarcomeric MyHC isoforms in human skeletal muscles of different embryological origin. Diversity and similarity in mammalian species', *Ann Anat*, 207, pp. 9-20.
McCarthy, D.J., Humburg, P., Kanapin, A., Rivas, M.A., Gaulton, K., Cazier, J.B. and Donnelly, P. (2014) 'Choice of transcripts and software has a large effect on variant annotation', *Genome Med*, 6(3), p. 26.
McCormack, P., Kole, A., Gainotti, S., Mascalzoni, D., Molster, C., Lochmuller, H. and Woods, S. (2016) ''You should at least ask'. The expectations, hopes and fears of rare disease patients on large-scale data and biomaterial sharing for genomics research', *Eur J Hum Genet*.
McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P. and Cunningham, F. (2016) 'The Ensembl Variant Effect Predictor', *Genome Biol*, 17(1), p. 122.
Medvedev, P., Fiume, M., Dzamba, M., Smith, T. and Brudno, M. (2010) 'Detecting copy number variation with mated short reads', *Genome Res*, 20(11), pp. 1613-22.
Meienberg, J., Bruggmann, R., Oexle, K. and Matyas, G. (2016) 'Clinical sequencing: is WGS the better WES?', *Hum Genet*, 135(3), pp. 359-62.
Meienberg, J., Zerjavic, K., Keller, I., Okoniewski, M., Patrignani, A., Ludin, K., Xu, Z., Steinmann, B., Carrel, T., Rothlisberger, B., Schlapbach, R., Bruggmann, R. and Matyas, G. (2015) 'New insights into the performance of human whole-exome capture platforms', *Nucleic Acids Res*, 43(11), p. e76.
Metzker, M.L. (2010) 'Sequencing technologies - the next generation', *Nat Rev Genet*, 11(1), pp. 31-46.

Millay, D.P., O'Rourke, J.R., Sutherland, L.B., Bezprozvannaya, S., Shelton, J.M., Bassel-Duby, R. and Olson, E.N. (2013) 'Myomaker is a membrane activator of myoblast fusion and muscle formation', *Nature*, 499(7458), pp. 301-5.
Millay, D.P., Sutherland, L.B., Bassel-Duby, R. and Olson, E.N. (2014) 'Myomaker is essential for muscle regeneration', *Genes Dev*, 28(15), pp. 1641-6.
Miller, J.R., Koren, S. and Sutton, G. (2010) 'Assembly algorithms for next-generation sequencing data', *Genomics*, 95(6), pp. 315-27.
Mori-Yoshimura, M. and Nishino, I. (2015) '[Sialic Acid Replacement Therapy for Distal Myopathy with Rimmed Vacuoles]', *Brain Nerve*, 67(9), pp. 1115-23.
Mori-Yoshimura, M., Oya, Y., Hayashi, Y.K., Noguchi, S., Nishino, I. and Murata, M. (2013) 'Respiratory dysfunction in patients severely affected by GNE myopathy (distal myopathy with rimmed vacuoles)', *Neuromuscul Disord*, 23(1), pp. 84-8.
Moutton, S., Fergelot, P., Naudion, S., Cordier, M.P., Sole, G., Guerineau, E., Hubert, C., Rooryck, C., Vuillaume, M.L., Houcinat, N., Deforges, J., Bouron, J., Deves, S., Le Merrer, M., David, A., Genevieve, D., Giuliano, F., Journel, H., Megarbane, A., Faivre, L., Chassaing, N., Francannet, C., Sarrazin, E., Stattin, E.L., Vigneron, J., Leclair, D., Abadie, C., Sarda, P., Baumann, C., Delrue, M.A., Arveiler, B., Lacombe, D., Goizet, C. and Coupry, I. (2016) 'Otopalatodigital spectrum disorders: refinement of the phenotypic and mutational spectrum', *J Hum Genet*, 61(8), pp. 693-9.
Muller, H., Jimenez-Heredia, R., Krolo, A., Hirschmugl, T., Dmytrus, J., Boztug, K. and Bock, C. (2017) 'VCF.Filter: interactive prioritization of disease-linked genetic variants from sequencing data', *Nucleic Acids Res*, 45(W1), pp. W567-W572.
Nam, J.Y., Kim, N.K., Kim, S.C., Joung, J.G., Xi, R., Lee, S., Park, P.J. and Park, W.Y. (2016) 'Evaluation of somatic copy number estimation tools for whole-exome sequencing data', *Brief Bioinform*, 17(2), pp. 185-92.
Natrajan, R. and Wilkerson, P. (2013) 'From integrative genomics to therapeutic targets', *Cancer Res*, 73(12), pp. 3483-8.
Need, A.C., Shashi, V., Hitomi, Y., Schoch, K., Shianna, K.V., McDonald, M.T., Meisler, M.H. and Goldstein, D.B. (2012) 'Clinical application of exome sequencing in undiagnosed genetic conditions', *J Med Genet*, 49(6), pp. 353-61.
Neveling, K., Collin, R.W., Gilissen, C., van Huet, R.A., Visser, L., Kwint, M.P., Gijsen, S.J., Zonneveld, M.N., Wieskamp, N., de Ligt, J., Siemiatkowska, A.M., Hoefsloot, L.H., Buckley, M.F., Kellner, U., Branham, K.E., den Hollander, A.I., Hoischen, A., Hoyng, C., Klevering, B.J., van den Born, L.I., Veltman, J.A., Cremers, F.P. and Scheffer, H. (2012) 'Next-generation genetic testing for retinitis pigmentosa', *Hum Mutat*, 33(6), pp. 963-72.
Neveling, K., Feenstra, I., Gilissen, C., Hoefsloot, L.H., Kamsteeg, E.J., Mensenkamp, A.R., Rodenburg, R.J., Yntema, H.G., Spruijt, L., Vermeer, S., Rinne, T., van Gassen, K.L., Bodmer, D., Lugtenberg, D., de Reuver, R., Buijsman, W., Derks, R.C., Wieskamp, N., van den Heuvel, B., Ligtenberg, M.J., Kremer, H., Koolen, D.A., van de Warrenburg, B.P., Cremers, F.P., Marcelis, C.L., Smeitink, J.A., Wortmann, S.B., van Zelst-Stams, W.A., Veltman, J.A., Brunner, H.G., Scheffer, H. and Nelen, M.R. (2013) 'A post-hoc comparison of the utility of sanger sequencing and exome sequencing for the diagnosis of heterogeneous diseases', *Hum Mutat*, 34(12), pp. 1721-6.
NEXTCODE-HEALTH (2016) *Clinical Sequence Analyzer. A technical whilepaper*.
Nielsen, R., Paul, J.S., Albrechtsen, A. and Song, Y.S. (2011) 'Genotype and SNP calling from next-generation sequencing data', *Nat Rev Genet*, 12(6), pp. 443-51.
Nishino, I., Carrillo-Carrasco, N. and Argov, Z. (2015) 'GNE myopathy: current update and future therapy', *J Neurol Neurosurg Psychiatry*, 86(4), pp. 385-92.

Norton, N., Li, D., Rampersaud, E., Morales, A., Martin, E.R., Zuchner, S., Guo, S., Gonzalez, M., Hedges, D.J., Robertson, P.D., Krumm, N., Nickerson, D.A. and Hershberger, R.E. (2013) 'Exome sequencing and genome-wide linkage analysis in 17 families illustrate the complex contribution of TTN truncating variants to dilated cardiomyopathy', *Circ Cardiovasc Genet*, 6(2), pp. 144-53.
O'Connor, E., Topf, A., Muller, J.S., Cox, D., Evangelista, T., Colomer, J., Abicht, A., Senderek, J., Hasselmann, O., Yaramis, A., Laval, S.H. and Lochmuller, H. (2016) 'Identification of mutations in the MYO9A gene in patients with congenital myasthenic syndrome', *Brain*, 139(Pt 8), pp. 2143-53.
Onsongo, G., Baughn, L.B., Bower, M., Henzler, C., Schomaker, M., Silverstein, K.A. and Thyagarajan, B. (2016) 'CNV-RF Is a Random Forest-Based Copy Number Variation Detection Method Using Next-Generation Sequencing', *J Mol Diagn*, 18(6), pp. 872-881.
Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J. and Trajanoski, Z. (2014) 'A survey of tools for variant analysis of next-generation genome sequencing data', *Brief Bioinform*, 15(2), pp. 256-78.
Pant, S., Weiner, R. and Marton, M.J. (2014) 'Navigating the rapids: the development of regulated next-generation sequencing-based clinical trial assays and companion diagnostics', *Front Oncol*, 4, p. 78.
Parla, J.S., Iossifov, I., Grabill, I., Spector, M.S., Kramer, M. and McCombie, W.R. (2011) 'A comparative analysis of exome capture', *Genome Biol*, 12(9), p. R97.
Patwardhan, A., Harris, J., Leng, N., Bartha, G., Church, D.M., Luo, S., Haudenschild, C., Pratt, M., Zook, J., Salit, M., Tirch, J., Morra, M., Chervitz, S., Li, M., Clark, M., Garcia, S., Chandratillake, G., Kirk, S., Ashley, E., Snyder, M., Altman, R., Bustamante, C., Butte, A.J., West, J. and Chen, R. (2015) 'Achieving high-sensitivity for clinical applications using augmented exome sequencing', *Genome Med*, 7, p. 71.
Pelissier, A., Peyron, C. and Bejean, S. (2016) 'Next-generation sequencing in clinical practice: from the patients' preferences to the informed consent process', *Public Health*.
Peng, H.H., Chang, N.C., Chen, K.T., Lu, J.J., Chang, P.Y., Chang, S.C., Wu-Chou, Y.H., Chou, Y.T., Phang, W. and Cheng, P.J. (2016) 'Nonsynonymous variants in MYH9 and ABCA4 are the most frequent risk loci associated with nonsyndromic orofacial cleft in Taiwanese population', *BMC Med Genet*, 17(1), p. 59.
Philippakis, A.A., Azzariti, D.R., Beltran, S., Brookes, A.J., Brownstein, C.A., Brudno, M., Brunner, H.G., Buske, O.J., Carey, K., Doll, C., Dumitriu, S., Dyke, S.O., den Dunnen, J.T., Firth, H.V., Gibbs, R.A., Girdea, M., Gonzalez, M., Haendel, M.A., Hamosh, A., Holm, I.A., Huang, L., Hurles, M.E., Hutton, B., Krier, J.B., Misyura, A., Mungall, C.J., Paschall, J., Paten, B., Robinson, P.N., Schiettecatte, F., Sobreira, N.L., Swaminathan, G.J., Taschner, P.E., Terry, S.F., Washington, N.L., Zuchner, S., Boycott, K.M. and Rehm, H.L. (2015) 'The Matchmaker Exchange: a platform for rare disease gene discovery', *Hum Mutat*, 36(10), pp. 915-21.
Pirooznia, M., Kramer, M., Parla, J., Goes, F.S., Potash, J.B., McCombie, W.R. and Zandi, P.P. (2014) 'Validation and assessment of variant calling pipelines for next-generation sequencing', *Hum Genomics*, 8, p. 14.
Pogoryelova, O., Cammish, P., Mansbach, H., Argov, Z., Nishino, I., Skrinar, A., Chan, Y., Nafissi, S., Shamshiri, H., Kakkis, E. and Lochmuller, H. (2018) 'Phenotypic stratification and genotype-phenotype correlation in a heterogeneous, international cohort of GNE myopathy patients: First report from the GNE myopathy Disease Monitoring Program, registry portion', *Neuromuscul Disord*, 28(2), pp. 158-168.
Poole, R.L., Docherty, L.E., Al Sayegh, A., Caliebe, A., Turner, C., Baple, E., Wakeling, E., Harrison, L., Lehmann, A., Temple, I.K. and Mackay, D.J. (2013) 'Targeted methylation testing

of a patient cohort broadens the epigenetic and clinical description of imprinting disorders', *Am J Med Genet A*, 161a(9), pp. 2174-82.
Protas, M.E., Weh, E., Footz, T., Kasberger, J., Baraban, S.C., Levin, A.V., Katz, L.J., Ritch, R., Walter, M.A., Semina, E.V. and Gould, D.B. (2017) 'Mutations of conserved non-coding elements of PITX2 in patients with ocular dysgenesis and developmental glaucoma', *Hum Mol Genet*, 26(18), pp. 3630-3638.
Rabbani, B., Tekin, M. and Mahdieh, N. (2014) 'The promise of whole-exome sequencing in medical genetics', *J Hum Genet*, 59(1), pp. 5-15.
Ratnakumar, A., McWilliam, S., Barris, W. and Dalrymple, B.P. (2010) 'Using paired-end sequences to optimise parameters for alignment of sequence reads against related genomes', *BMC Genomics*, 11, p. 458.
Rauch, A., Hoyer, J., Guth, S., Zweier, C., Kraus, C., Becker, C., Zenker, M., Huffmeier, U., Thiel, C., Ruschendorf, F., Nurnberg, P., Reis, A. and Trautmann, U. (2006) 'Diagnostic yield of various genetic approaches in patients with unexplained developmental delay or mental retardation', *Am J Med Genet A*, 140(19), pp. 2063-74.
Rennert, H., Eng, K., Zhang, T., Tan, A., Xiang, J., Romanel, A., Kim, R., Tam, W., Liu, Y.C., Bhinder, B., Cyrta, J., Beltran, H., Robinson, B., Mosquera, J.M., Fernandes, H., Demichelis, F., Sboner, A., Kluk, M., Rubin, M.A. and Elemento, O. (2016) 'Development and validation of a whole-exome sequencing test for simultaneous detection of point mutations, indels and copy-number alterations for precision cancer care', *NPJ Genom Med*, 1.
Ricci, E., Broccolini, A., Gidaro, T., Morosetti, R., Gliubizzi, C., Frusciante, R., Di Lella, G.M., Tonali, P.A. and Mirabella, M. (2006) 'NCAM is hyposialylated in hereditary inclusion body myopathy due to GNE mutations', *Neurology*, 66(5), pp. 755-758.
Robinson, K.M., Hawkins, A.S., Santana-Cruz, I., Adkins, R.S., Shetty, A.C., Nagaraj, S., Sadzewicz, L., Tallon, L.J., Rasko, D.A., Fraser, C.M., Mahurkar, A., Silva, J.C. and Dunning Hotopp, J.C. (2017) 'Aligner optimization increases accuracy and decreases compute times in multi-species sequence data', *Microb Genom*, 3(9), p. e000122.
Roeh, S., Weber, P., Rex-Haffner, M., Deussing, J.M., Binder, E.B. and Jakovcevski, M. (2017) 'Sequencing on the SOLiD 5500xl System - in-depth characterization of the GC bias', *Nucleus*, 8(4), pp. 370-380.
Rogozhina, Y., Mironovich, S., Shestak, A., Adyan, T., Polyakov, A., Podolyak, D., Bakulina, A., Dzemeshkevich, S. and Zaklyazminskaya, E. (2016) 'New intronic splicing mutation in the LMNA gene causing progressive cardiac conduction defects and variable myopathy', *Gene*, 595(2), pp. 202-206.
Roy, S., Coldren, C., Karunamurthy, A., Kip, N.S., Klee, E.W., Lincoln, S.E., Leon, A., Pullambhatla, M., Temple-Smolkin, R.L., Voelkerding, K.V., Wang, C. and Carter, A.B. (2018) 'Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists', *J Mol Diagn*, 20(1), pp. 4-27.
Ruiz-Martinez, J., Azcona, L.J., Bergareche, A., Marti-Masso, J.F. and Paisan-Ruiz, C. (2017) 'Whole-exome sequencing associates novel CSMD1 gene mutations with familial Parkinson disease', *Neurol Genet*, 3(5), p. e177.
Salgado, D., Bellgard, M.I., Desvignes, J.P. and Beroud, C. (2016) 'How to Identify Pathogenic Mutations among All Those Variations: Variant Annotation and Filtration in the Genome Sequencing Era', *Hum Mutat*, 37(12), pp. 1272-1282.
Sarparanta, J., Jonson, P.H., Golzio, C., Sandell, S., Luque, H., Screen, M., McDonald, K., Stajich, J.M., Mahjneh, I., Vihola, A., Raheem, O., Penttila, S., Lehtinen, S., Huovinen, S., Palmio, J., Tasca, G., Ricci, E., Hackman, P., Hauser, M., Katsanis, N. and Udd, B. (2012)

'Mutations affecting the cytoplasmic functions of the co-chaperone DNAJB6 cause limb-girdle muscular dystrophy', *Nat Genet*, 44(4), pp. 450-5, S1-2.
Sauna, Z.E. and Kimchi-Sarfaty, C. (2011) 'Understanding the contribution of synonymous mutations to human disease', *Nat Rev Genet*, 12(10), pp. 683-91.
Schouten, J.P., McElgunn, C.J., Waaijer, R., Zwijnenburg, D., Diepvens, F. and Pals, G. (2002) 'Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification', *Nucleic Acids Res*, 30(12), p. e57.
Selcen, D., Shen, X.M., Milone, M., Brengman, J., Ohno, K., Deymeer, F., Finkel, R., Rowin, J. and Engel, A.G. (2013) 'GFPT1-myasthenia: clinical, structural, and electrophysiologic heterogeneity', *Neurology*, 81(4), pp. 370-8.
Shashi, V., McConkie-Rosell, A., Rosell, B., Schoch, K., Vellore, K., McDonald, M., Jiang, Y.H., Xie, P., Need, A. and Goldstein, D.B. (2014) 'The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders', *Genet Med*, 16(2), pp. 176-82.
Shiao, Y.H. (2016) 'Editorial: Marching Toward 100% Whole Genome Sequencing', *Front Genet*, 7, p. 41.
Shkedi-Rafid, S., Fenwick, A., Dheensa, S., Wellesley, D. and Lucassen, A.M. (2016) 'What results to disclose, when, and who decides? Healthcare professionals' views on prenatal chromosomal microarray analysis', *Prenat Diagn*, 36(3), pp. 252-9.
Shyr, C., Kushniruk, A. and Wasserman, W.W. (2014) 'Usability study of clinical exome analysis software: top lessons learned and recommendations', *J Biomed Inform*, 51, pp. 129-36.
Silva, J.P., Lelianova, V.G., Ermolyuk, Y.S., Vysokov, N., Hitchen, P.G., Berninghausen, O., Rahman, M.A., Zangrandi, A., Fidalgo, S., Tonevitsky, A.G., Dell, A., Volynski, K.E. and Ushkaryov, Y.A. (2011) 'Latrophilin 1 and its endogenous ligand Lasso/teneurin-2 form a high-affinity transsynaptic receptor pair with signaling capabilities', *Proc Natl Acad Sci U S A*, 108(29), pp. 12113-8.
Sims, D., Sudbery, I., Ilott, N.E., Heger, A. and Ponting, C.P. (2014) 'Sequencing depth and coverage: key considerations in genomic analyses', *Nat Rev Genet*, 15(2), pp. 121-32.
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.
Smith, T.F. (2008) 'Diversity of WD-repeat proteins', *Subcell Biochem*, 48, pp. 20-30.
Sobreira, N.L., Cirulli, E.T., Avramopoulos, D., Wohler, E., Oswald, G.L., Stevens, E.L., Ge, D., Shianna, K.V., Smith, J.P., Maia, J.M., Gumbs, C.E., Pevsner, J., Thomas, G., Valle, D., Hoover-Fong, J.E. and Goldstein, D.B. (2010) 'Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene', *PLoS Genet*, 6(6), p. e1000991.
South, S.T., Lee, C., Lamb, A.N., Higgins, A.W. and Kearney, H.M. (2013) 'ACMG Standards and Guidelines for constitutional cytogenomic microarray analysis, including postnatal and prenatal applications: revision 2013', *Genet Med*, 15(11), pp. 901-9.
Stark, Z., Dashnow, H., Lunke, S., Tan, T.Y., Yeung, A., Sadedin, S., Thorne, N., Macciocca, I., Gaff, C., Oshlack, A., White, S.M. and James, P.A. (2017) 'A clinically driven variant prioritization framework outperforms purely computational approaches for the diagnostic analysis of singleton WES data', *Eur J Hum Genet*, 25(11), pp. 1268-1272.
Stasche, R., Hinderlich, S., Weise, C., Effertz, K., Lucka, L., Moormann, P. and Reutter, W. (1997) 'A bifunctional enzyme catalyzes the first two steps in N-acetylneuraminic acid

biosynthesis of rat liver. Molecular cloning and functional expression of UDP-N-acetyl-glucosamine 2-epimerase/N-acetylmannosamine kinase', *J Biol Chem*, 272(39), pp. 24319-24.
Supek, F., Minana, B., Valcarcel, J., Gabaldon, T. and Lehner, B. (2014) 'Synonymous mutations frequently act as driver mutations in human cancers', *Cell*, 156(6), pp. 1324-1335.
Suzuki, N., Izumi, R., Kato, M., Warita, H. and Aoki, M. (2017) '[Therapeutic development for GNE myopathy.]', *Clin Calcium*, 27(3), pp. 429-434.
Tan, R., Wang, J., Wu, X., Juan, L., Zheng, L., Ma, R., Zhan, Q., Wang, T., Jin, S., Jiang, Q. and Wang, Y. (2017) 'ERDS-exome: a Hybrid Approach for Copy Number Variant Detection from Whole-exome Sequencing Data', *IEEE/ACM Trans Comput Biol Bioinform*.
Taylor, J.C., Martin, H.C., Lise, S., Broxholme, J., Cazier, J.B., Rimmer, A., Kanapin, A., Lunter, G., Fiddy, S., Allan, C., Aricescu, A.R., Attar, M., Babbs, C., Becq, J., Beeson, D., Bento, C., Bignell, P., Blair, E., Buckle, V.J., Bull, K., Cais, O., Cario, H., Chapel, H., Copley, R.R., Cornall, R., Craft, J., Dahan, K., Davenport, E.E., Dendrou, C., Devuyst, O., Fenwick, A.L., Flint, J., Fugger, L., Gilbert, R.D., Goriely, A., Green, A., Greger, I.H., Grocock, R., Gruszczyk, A.V., Hastings, R., Hatton, E., Higgs, D., Hill, A., Holmes, C., Howard, M., Hughes, L., Humburg, P., Johnson, D., Karpe, F., Kingsbury, Z., Kini, U., Knight, J.C., Krohn, J., Lamble, S., Langman, C., Lonie, L., Luck, J., McCarthy, D., McGowan, S.J., McMullin, M.F., Miller, K.A., Murray, L., Nemeth, A.H., Nesbit, M.A., Nutt, D., Ormondroyd, E., Oturai, A.B., Pagnamenta, A., Patel, S.Y., Percy, M., Petousi, N., Piazza, P., Piret, S.E., Polanco-Echeverry, G., Popitsch, N., Powrie, F., Pugh, C., Quek, L., Robbins, P.A., Robson, K., Russo, A., Sahgal, N., van Schouwenburg, P.A., Schuh, A., Silverman, E., Simmons, A., Sorensen, P.S., Sweeney, E., Taylor, J., Thakker, R.V., Tomlinson, I., Trebes, A., Twigg, S.R., Uhlig, H.H., Vyas, P., Vyse, T., Wall, S.A., Watkins, H., Whyte, M.P., Witty, L., et al. (2015) 'Factors influencing success of clinical genome sequencing across a broad spectrum of disorders', *Nat Genet*, 47(7), pp. 717-26.
Techa-Angkoon, P., Sun, Y. and Lei, J. (2017) 'A sensitive short read homology search tool for paired-end read sequencing data', *BMC Bioinformatics*, 18(Suppl 12), p. 414.
TheBroadInstitute. Available at: https://software.broadinstitute.org/gatk/guide/article?id=4150 (Accessed: 03/02/2017).
Thompson, R., Johnston, L., Taruscio, D., Monaco, L., Beroud, C., Gut, I.G., Hansson, M.G., t Hoen, P.B., Patrinos, G.P., Dawkins, H., Ensini, M., Zatloukal, K., Koubi, D., Heslop, E., Paschall, J.E., Posada, M., Robinson, P.N., Bushby, K. and Lochmuller, H. (2014) 'RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research', *J Gen Intern Med*, 29 Suppl 3, pp. S780-7.
Tian, S., Yan, H., Kalmbach, M. and Slager, S.L. (2016) 'Impact of post-alignment processing in variant discovery from whole exome data', *BMC Bioinformatics*, 17(1), p. 403.
Toma, C., Torrico, B., Hervas, A., Valdes-Mas, R., Tristan-Noguero, A., Padillo, V., Maristany, M., Salgado, M., Arenas, C., Puente, X.S., Bayes, M. and Cormand, B. (2014) 'Exome sequencing in multiplex autism families suggests a major role for heterozygous truncating mutations', *Mol Psychiatry*, 19(7), pp. 784-90.
Torella, A., Fanin, M., Mutarelli, M., Peterle, E., Del Vecchio Blanco, F., Rispoli, R., Savarese, M., Garofalo, A., Piluso, G., Morandi, L., Ricci, G., Siciliano, G., Angelini, C. and Nigro, V. (2013) 'Next-generation sequencing identifies transportin 3 as the causative gene for LGMD1F', *PLoS One*, 8(5), p. e63536.
Toro, C., Olive, M., Dalakas, M.C., Sivakumar, K., Bilbao, J.M., Tyndel, F., Vidal, N., Farrero, E., Sambuughin, N. and Goldfarb, L.G. (2013) 'Exome sequencing identifies titin mutations causing hereditary myopathy with early respiratory failure (HMERF) in families of diverse ethnic origins', *BMC Neurol*, 13, p. 29.
Toscano, A., Barca, E. and Musumeci, O. (2017) 'Update on diagnostics of metabolic myopathies', *Curr Opin Neurol*, 30(5), pp. 553-562.

Treangen, T.J. and Salzberg, S.L. (2011) 'Repetitive DNA and next-generation sequencing: computational challenges and solutions', *Nat Rev Genet*, 13(1), pp. 36-46.
Trost, B., Walker, S., Wang, Z., Thiruvahindrapuram, B., MacDonald, J.R., Sung, W.W.L., Pereira, S.L., Whitney, J., Chan, A.J.S., Pellecchia, G., Reuter, M.S., Lok, S., Yuen, R.K.C., Marshall, C.R., Merico, D. and Scherer, S.W. (2018) 'A Comprehensive Workflow for Read Depth-Based Identification of Copy-Number Variation from Whole-Genome Sequence Data', *Am J Hum Genet*, 102(1), pp. 142-155.
Trump, N., McTague, A., Brittain, H., Papandreou, A., Meyer, E., Ngoh, A., Palmer, R., Morrogh, D., Boustred, C., Hurst, J.A., Jenkins, L., Kurian, M.A. and Scott, R.H. (2016) 'Improving diagnosis and broadening the phenotypes in early-onset seizure and severe developmental delay disorders through gene panel analysis', *J Med Genet*.
Truszkowska, G.T., Bilinska, Z.T., Muchowicz, A., Pollak, A., Biernacka, A., Kozar-Kaminska, K., Stawinski, P., Gasperowicz, P., Kosinska, J., Zielinski, T. and Ploski, R. (2017) 'Homozygous truncating mutation in NRAP gene identified by whole exome sequencing in a patient with dilated cardiomyopathy', *Sci Rep*, 7(1), p. 3362.
van Karnebeek, C.D., Scheper, F.Y., Abeling, N.G., Alders, M., Barth, P.G., Hoovers, J.M., Koevoets, C., Wanders, R.J. and Hennekam, R.C. (2005) 'Etiology of mental retardation in children referred to a tertiary care center: a prospective study', *Am J Ment Retard*, 110(4), pp. 253-67.
Vaz-Drago, R., Custodio, N. and Carmo-Fonseca, M. (2017) 'Deep intronic mutations and human disease', *Hum Genet*, 136(9), pp. 1093-1111.
Vieira, N.M., Naslavsky, M.S., Licinio, L., Kok, F., Schlesinger, D., Vainzof, M., Sanchez, N., Kitajima, J.P., Gal, L., Cavacana, N., Serafini, P.R., Chuartzman, S., Vasquez, C., Mimbacas, A., Nigro, V., Pavanello, R.C., Schuldiner, M., Kunkel, L.M. and Zatz, M. (2014) 'A defect in the RNA-processing protein HNRPDL causes limb-girdle muscular dystrophy 1G (LGMD1G)', *Hum Mol Genet*, 23(15), pp. 4103-10.
Voermans, N.C., Snoeck, M. and Jungbluth, H. (2016) 'RYR1-related rhabdomyolysis: A common but probably underdiagnosed manifestation of skeletal muscle ryanodine receptor dysfunction', *Rev Neurol (Paris)*, 172(10), pp. 546-558.
Volk, A.E. and Kubisch, C. (2017) 'The rapid evolution of molecular genetic diagnostics in neuromuscular diseases', *Curr Opin Neurol*, 30(5), pp. 523-528.
Wang, K., Kim, C., Bradfield, J., Guo, Y., Toskala, E., Otieno, F.G., Hou, C., Thomas, K., Cardinale, C., Lyon, G.J., Golhar, R. and Hakonarson, H. (2013a) 'Whole-genome DNA/RNA sequencing identifies truncating mutations in RBCK1 in a novel Mendelian disease with neuromuscular and cardiac involvement', *Genome Med*, 5(7), p. 67.
Wang, K., Li, M. and Hakonarson, H. (2010) 'ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data', *Nucleic Acids Res*, 38(16), p. e164.
Wang, S. and Xing, J. (2013) 'A primer for disease gene prioritization using next-generation sequencing data', *Genomics Inform*, 11(4), pp. 191-9.
Wang, Z., Liu, X., Yang, B.Z. and Gelernter, J. (2013b) 'The role and challenges of exome sequencing in studies of human diseases', *Front Genet*, 4, p. 160.
Warman Chardon, J., Beaulieu, C., Hartley, T., Boycott, K.M. and Dyment, D.A. (2015) 'Axons to Exons: the Molecular Diagnosis of Rare Neurological Diseases by Next-Generation Sequencing', *Curr Neurol Neurosci Rep*, 15(9), p. 64.
Wenric, S., Sticca, T., Caberg, J.H., Josse, C., Fasquelle, C., Herens, C., Jamar, M., Max, S., Gothot, A., Caers, J. and Bours, V. (2017) 'Exome copy number variation detection: Use of a pool of unrelated healthy tissue as reference sample', *Genet Epidemiol*, 41(1), pp. 35-40.
Witherspoon, J.W. and Meilleur, K.G. (2016) 'Review of RyR1 pathway and associated pathomechanisms', *Acta Neuropathol Commun*, 4(1), p. 121.

Wittig, M., Anmarkrud, J.A., Kassens, J.C., Koch, S., Forster, M., Ellinghaus, E., Hov, J.R., Sauer, S., Schimmler, M., Ziemann, M., Gorg, S., Jacob, F., Karlsen, T.H. and Franke, A. (2015) 'Development of a high-resolution NGS-based HLA-typing and analysis pipeline', *Nucleic Acids Res*, 43(11), p. e70.
Wu, L., Yavas, G., Hong, H., Tong, W. and Xiao, W. (2017) 'Direct comparison of performance of single nucleotide variant calling in human genome with alignment-based and assembly-based approaches', *Sci Rep*, 7(1), p. 10963.
Xin, W., Xiao, X., Li, S., Jia, X., Guo, X. and Zhang, Q. (2015) 'Identification of Genetic Defects in 33 Probands with Stargardt Disease by WES-Based Bioinformatics Gene Panel Analysis', *PLoS One*, 10(7), p. e0132635.
Xuan, J., Yu, Y., Qing, T., Guo, L. and Shi, L. (2013) 'Next-generation sequencing in the clinic: promises and challenges', *Cancer Lett*, 340(2), pp. 284-95.
Xue, Y., Ankala, A., Wilcox, W.R. and Hegde, M.R. (2015) 'Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: single-gene, gene panel, or exome/genome sequencing', *Genet Med*, 17(6), pp. 444-51.
Yang, H. and Wang, K. (2015) 'Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR', *Nat Protoc*, 10(10), pp. 1556-66.
Yen, J.L., Garcia, S., Montana, A., Harris, J., Chervitz, S., Morra, M., West, J., Chen, R. and Church, D.M. (2017) 'A variant by any name: quantifying annotation discordance across tools and clinical databases', *Genome Med*, 9(1), p. 7.
Zarate, O.A., Brody, J.G., Brown, P., Ramirez-Andreotta, M.D., Perovich, L. and Matz, J. (2016) 'Balancing Benefits and Risks of Immortal Data: Participants' Views of Open Consent in the Personal Genome Project', *Hastings Cent Rep*, 46(1), pp. 36-45.
Zatz, M., Passos-Bueno, M.R. and Vainzof, M. (2016) 'Neuromuscular disorders: genes, genetic counseling and therapeutic trials', *Genet Mol Biol*, 39(3), pp. 339-48.
Zaum, A.K., Stuve, B., Gehrig, A., Kolbel, H., Schara, U., Kress, W. and Rost, S. (2017) 'Deep intronic variants introduce DMD pseudoexon in patient with muscular dystrophy', *Neuromuscul Disord*, 27(7), pp. 631-634.
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, G., Wang, J., Yang, J., Li, W., Deng, Y., Li, J., Huang, J., Hu, S. and Zhang, B. (2015) 'Comparison and evaluation of two exome capture kits and sequencing platforms for variant calling', *BMC Genomics*, 16, p. 581.
Zhao, J., Wang, Z., Hong, D., Lv, H., Zhang, W., Chen, J. and Yuan, Y. (2015) 'Mutational spectrum and clinical features in 35 unrelated mainland Chinese patients with GNE myopathy', *J Neurol Sci*, 354(1-2), pp. 21-6.
Zook, J.M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C.E., Alexander, N., Henaff, E., McIntyre, A.B., Chandramohan, D., Chen, F., Jaeger, E., Moshrefi, A., Pham, K., Stedman, W., Liang, T., Saghbini, M., Dzakula, Z., Hastie, A., Cao, H., Deikus, G., Schadt, E., Sebra, R., Bashir, A., Truty, R.M., Chang, C.C., Gulbahce, N., Zhao, K., Ghosh, S., Hyland, F., Fu, Y., Chaisson, M., Xiao, C., Trow, J., Sherry, S.T., Zaranek, A.W., Ball, M., Bobe, J., Estep, P., Church, G.M., Marks, P., Kyriazopoulou-Panagiotopoulou, S., Zheng, G.X., Schnall-Levin, M., Ordonez, H.S., Mudivarti, P.A., Giorda, K., Sheng, Y., Rypdal, K.B. and Salit, M. (2016) 'Extensive sequencing of seven human genomes to characterize benchmark reference materials', *Sci Data*, 3, p. 160025.

Zook, J.M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W. and Salit, M. (2014) 'Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls', *Nat Biotechnol*, 32(3), pp. 246-51.
Zucca, S., Villaraggia, M., Gagliardi, S., Grieco, G.S., Valente, M., Cereda, C. and Magni, P. (2016) 'Analysis of amplicon-based NGS data from neurological disease gene panels: a new method for allele drop-out management', *BMC Bioinformatics*, 17(Suppl 12), p. 339.

Chapter 8. Appendices

A. Muscle Gene Table

The NMD gene list used in the analyses presented in this thesis was compiled from the Muscle Gene Table provided by the World Muscle Society at http://musclegenetable.fr/ in July 2016, and consists of the 416 genes listed below.

AARS	BICD2	DMPK	GBA2	KIAA0196	NDRG1	PRX	SPAST	TTN
AARS2	BIN1	DNAJB2	GBE1	KIF1A	NDUFAF1	PSEN2	SPEG	TTPA
ABCC9	BSCL2	DNAJB6	GDAP1	KIF1B	NEB	PTPLA	SPG11	TTR
ABHD5	C10orf2	DNM2	GFPT1	KIF21A	NEFH	PTRF	SPG20	TUBB3
ACADVL	C12orf65	DNMT1	GJA5	KIF5A	NEFL	PUS1	SPG21	UBA1
ACTA1	C19orf12	DOK7	GJB1	KLHL40	NEXN	PYGM	SPG7	UBQLN2
ACTC1	C9orf72	DOLK	GJB3	KLHL41	NGF	RAB7A	SPTBN2	VAPB
ACTN2	CACNA1A	DPAGT1	GJC2	KLHL9	NIPA1	RAPSN	SPTLC1	VCL
ACVR1	CACNA1C	DPM1	GLE1	L1CAM	NOP56	RBCK1	SPTLC2	VCP
ADCK3	CACNA1S	DPM2	GMPPB	LAMA2	NPPA	RBM20	SQSTM1	VMA21
AFG3L2	CACNB2	DPM3	GNB4	LAMA4	NT5C2	REEP1	STIM1	VRK1
AGL	CACNB4	DSC2	GNE	LAMB2	NUP155	RNF216	SUCLA2	WNK1
AGRN	CAPN3	DSG2	GPD1L	LAMP2	OPA1	RRM2B	SURF1	YARS

AIFM1	CASQ1	DSP	GYG1	LARGE	OPTN	RTN2	SYNE1	YARS2
AKAP9	CASQ2	DTNA	GYS1	LDB3	ORAI1	RYR1	SYNE2	ZFYVE26
ALDH3A2	CAV3	DUX4	HCN4	LDHA	PABPN1	RYR2	SYT2	ZFYVE27
ALG13	CCDC88C	DYNC1H1	HEXB	LIMS2	PDK3	SACS	TARDBP	
ALG14	CFL2	DYSF	HINT1	LITAF	PDYN	SBF1	TAZ	
ALG2	CHAT	EEF2	HK1	LMNA	PEX7	SBF2	TBP	
ALS2	CHCHD10	EGR2	HNRNPDL	LMOD3	PFKM	SCN1B	TCAP	
AMPD2	CHKB	ELOVL4	HOXD10	LPIN1	PFN1	SCN2B	TDP1	
ANG	CHMP2B	ELOVL5	HSPB1	LRSAM1	PGAM2	SCN3B	TECPR2	
ANK2	CHRNA1	EMD	HSPB3	MARS	PGK1	SCN4A	TFG	
ANKRD1	CHRNB1	ENO3	HSPB8	MATR3	PGM1	SCN4B	TGFB3	
ANO5	CHRND	ENTPD1	HSPD1	MED25	PHKA1	SCN5A	TGM6	
AP4B1	CHRNE	ERBB3	HSPG2	MEGF10	PHOX2A	SDHA	TIA1	
AP4E1	CHRNG	ERLIN2	IFRD1	MFN2	PHYH	SEPN1	TK2	
AP4M1	CLCN1	ETFA	IGHMBP2	MPZ	PIP5K1C	SEPT9	TMEM43	
AP4S1	CLN3	ETFB	IKBKAP	MRE11A	PKP2	SETX	TMEM5	
AP5Z1	CNBP	ETFDH	ILK	MRPL3	PLEC	SGCA	TMPO	
APTX	CNTN1	EXOSC3	INF2	MRPL44	PLEKHG5	SGCB	TNNC1	
AR	COL6A1	EXOSC8	ISCU	MSTN	PLN	SGCD	TNNI2	

ARHGEF10	COL6A2	EYA4	ISPD	MTM1	PLP1	SGCE	TNNI3
ARL6IP1	COL6A3	FA2H	ITGA7	MTMR2	PMP22	SGCG	TNNT1
ASAH1	COLQ	FBLN5	ITPR1	MTO1	PNPLA2	SH3TC2	TNNT2
ATL1	COX15	FBXO38	JPH2	MTPAP	PNPLA6	SIGMAR1	TNNT3
ATM	COX6A1	FGD4	JUP	MURC	PNPLA8	SIL1	TNPO3
ATP2A1	CPT2	FGF14	KARS	MUSK	POLG	SLC12A6	TOR1A
ATP7A	CRYAB	FHL1	KBTBD13	MYBPC3	POLG2	SLC1A3	TOR1AIP1
ATXN1	CSRP3	FIG4	KCNA1	MYH2	POMGNT1	SLC22A5	TPM1
ATXN10	CTDP1	FKRP	KCNA5	MYH3	POMGNT2	SLC25A20	TPM2
ATXN2	CYP2U1	FKTN	KCNC3	MYH6	POMK	SLC25A4	TPM3
ATXN3	CYP7B1	FLNA	KCND3	MYH7	POMT1	SLC33A1	TRAPPC11
ATXN7	DAG1	FLNC	KCNE1	MYH8	POMT2	SLC52A2	TRIM2
ATXN8OS	DCTN1	FUS	KCNE2	MYL2	PPP2R2B	SLC52A3	TRIM32
B3GALNT2	DDHD1	FXN	KCNE3	MYL3	PREPL	SLC5A7	TRIM54
B3GNT1	DDHD2	GAA	KCNH2	MYLK2	PRKAG2	SMCHD1	TRIM63
B4GALNT1	DES	GAN	KCNJ2	MYOT	PRKCG	SMN1	TRPV4
BAG3	DHTKD1	GARS	KCNJ5	MYOZ2	PRPH	SNTA1	TSFM
BEAN1	DMD	GATAD1	KCNQ1	MYPN	PRPS1	SOD1	TTBK2

B. PCR and Sequencing protocols

i. Newcastle University

Variant segregation for *FILIP1* and *TENM2* variants.

Variant segregation was performed by two Newcastle University undergraduate students, Isaac Walton and Rebecca Haigh.

Primer pairs were designed using Primer3 (http://primer3.ut.ee/). For *FILIP1*, the primers amplified a 242bp region including the variant at chr6:76124519 where the nonsense variant (A>G) was detected by WES.

The regions were amplified in DNA samples from the index, father, mothers, unaffected paternal uncle, affected sister and unaffected brother.

For the *TENM2* variant, chr5:167673823 C>G, the primers amplified a 377bp region in the index, parents and unaffected sibling.

A HotStarTaq PCR protocol was used according to manufacturer's instructions. Amplification was confirmed using a 2% agarose gel run at 100 volts for 40 minutes in TAE buffer. The amplified DNA was then purified using an ExoSap protocol according to manufacturer's instructions. Sanger sequencing was then performed by Eurofins Genomics and the chromatograms were returned. Variant location was manually determined on the chromatogram and segregation of the variant was evaluated in the family. Primer pairs and the chromatogram image are shown in supplementary figures 1 and 2 for the *FILIP1* and *TENM2* variants respectively.

Supplementary figure 1: A, Forward and reverse primers used for the segregation of the *FILIP1* variant at position chr6:76124520 (output from Primer3, http://primer3.ut.ee/). B, chromatogram image for variant segregation in the family. The variant position is underlined in yellow.

A

Template masking not selected

No mispriming library specified

Using 1-based sequence positions OLIGO start len tm gc% any_th 3'_th hairpin seq

LEFT PRIMER 385 20 58.73 55.00 0.00 0.00 0.00 TGACTCTACCTGCAACTCCC

POSITION PENALTY, QUALITY: 92.000000, 0.267121

RIGHT PRIMER 626 20 58.56 55.00 0.00 0.00 0.00 ATATCTCCTGTCCCAAGCCC

POSITION PENALTY, QUALITY: 112.000000, 0.439543

SEQUENCE SIZE: 1001

INCLUDED REGION SIZE: 1001

PRODUCT SIZE: 242, PAIR ANY_TH COMPL: 0.00, PAIR 3'_TH COMPL: 0.00 TARGETS (start len EXCLUDED REGIONS (start len

1 AAGCAAAATTAGCAATTGACCTGAACTCTCCATTTTAATAATCACTGCATTTCTACATTT

61 TTCAGTCACAGTCAATGTCTAAATTAGATAACCTAATTGATTACCTACTGACCGAAAGGA

X

121 TCTGTATTAAATCACCACCAAAATCATATAATGAAGAAATTTTCTGCTTCTAGAGAAGTT

181 TGAAGTTGGAATCAGCACTTTTTCATTTAGCATTCATTCAACAAATATTTATTTAGTGCA

241 AACTTGTGCCAGGCACTAGGGATTTAGAAATAAGGGAAACATTCTCCTTTCCCCATAAAG

X

301 TTCATTACTCAGAGAGGGGAACAGATTTACATTCTCTTATTTTATAGCTTAGAAACCAAA

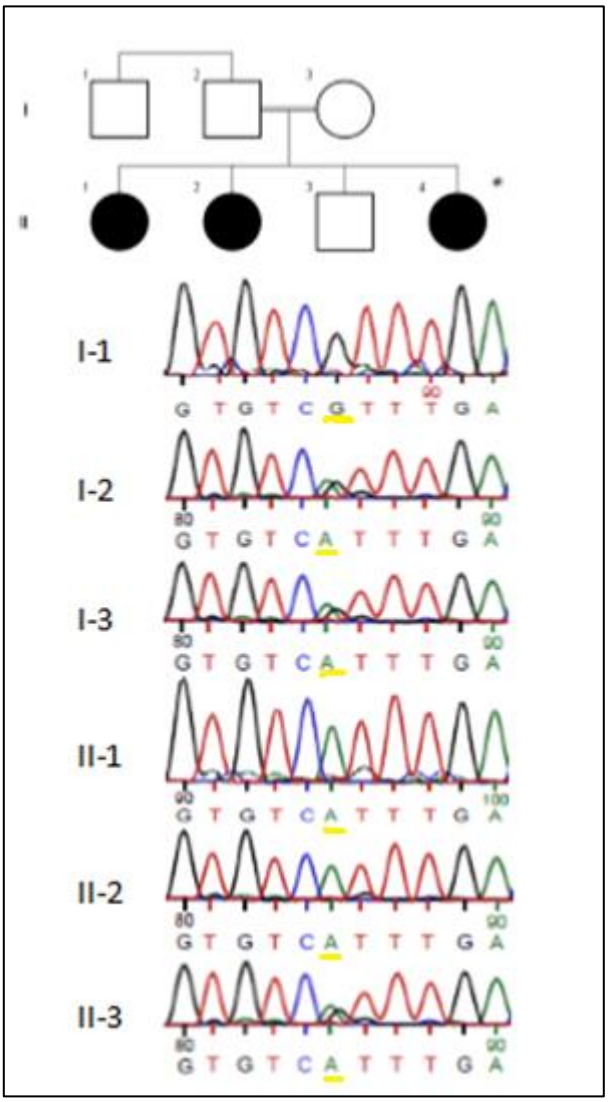
361 GACACAATAACAGTTGGGAAAGTTTGACTCTACCTGCAACTCCCCTTCCATTATACTGAG

>>>>>>>>>>>>>>>>>>>>

421 TAGTTGGATGAGGTCTTCTTTGGATAACTCCAGGGATTTCTTAGTTTTTCGTTCACATTC

```
                          ****************************

481 TCCAGATGTTTTTAGGTGTCGTTTGACAGTTCCTGAGGCCATGACATCATCCTCCTTCCT

    ************************************************************

541 ATTTGATTTCTTCTTCTTTTTTGCATCTTCTGAGAGACTTTTTTTCACCAGCATTGCCGAT

    **********

601 GATGGAGGGCTTGGGACAGGAGATATGCCCATCAGATGCACTTTCACCACCTTGGTTTCG

        <<<<<<<<<<<<<<<<<<<<

661 AGATCTCATTCCCACCTACAACACATACATAAAAAAAGATAAGATGTTCTTTACCACCAT

721 CAAAATTATTTTTGCCTAAATACTTAGAAACTTATAGCAGCTTTACCACATCGCCAATAA

                                    X

781 GGAAAAATCAAATTAACATGTCATTAAAATTTTCAGGAAACCCCTTTGATGGTCTTCAGA

841 AAAAATGAAAGTAAGATTTTAACCTGACAAACCCCTCAAGCCAAATGATCAAGGTCAGTA

901 TCAATAGTGCTAAGTCATATTGCTGTTACATAGCCCCCATAATAATATATCAATATTGGT

961 TCATTGATTGTAACAAACGTACTATACTATCCTAATGGAGA

KEYS (in order of precedence):

XXXXXX excluded region

****** target

>>>>>> left primer

<<<<<< right primer
```

B



Supplementary figure 2: A, forward and reverse primers used for the segregation of the *TENM2* variant at position chr5:167673823 (output from Primer3, http://primer3.ut.ee/). B, chromatogram image variant segregation in the father (top), affected individual (centre) and unaffected sibling (bottom).

A

```
Template masking not selected

No mispriming library specified

Using 1-based sequence positions

OLIGO            start  len      tm     gc%   any th  3' th hairpin seq

LEFT PRIMER        110   21   57.88   47.62     0.00   0.00    0.00
TCTGGTCTAAGCTTCCCATCT

RIGHT PRIMER       486   20   59.24   50.00     0.00   0.00    0.00
ACTTGTAGAACACCTGGCGT

SEQUENCE SIZE: 601

INCLUDED REGION SIZE: 601



PRODUCT SIZE: 377, PAIR ANY_TH COMPL: 0.00, PAIR 3'_TH COMPL: 0.00

TARGETS (start len


    1 TTACAAGATCCTGAGTATTTTGCATGCACATTAAATTTGGGAAGCACTGGTCTAGGCAAC




   61 CACCATGATCCATCAAAATTGGCATGTCCAGTGAAATAAATCTGTTTAATCTGGTCTAAG
                                                          >>>>>>>>>>>


  121 CTTCCCATCTTTCATGACTTTTGAGTTGACATTTGGGGAAATGGACTAAATTCTGTCTGT
      >>>>>>>>>>                  ******************************
```

```
181 TTGAATGCTTGGTCCTCAAAAGCCAACTGATATGTTCTGTTCCCCTTTCCTTTCTGCAGT
    ************************************************************


241 CCATGGTCCTCCTGCTTCAGAGCCAACGTCAGTATATATTTGAGTATGACTCCTCTGACC
    ************************************************************


301 GCCTCCTTGCCGTCACCATGCCCAGCGTGGCCCGGCACAGCATGTCCACACACACCTCCA
    ************************************************************


361 TCGGCTACATCCGTAATATTTACAACCCGCCTGAAAGCAATGCTTCGGTCATCTTTGACT
    ************************************************************


421 ACAGTGATGACGGCCGCATCCTGAAGACCTCCTTTTTGGGCACCGGACGCCAGGTGTTCT
    ******************************                <<<<<<<<<<<<<<


481 ACAAGTATGGGAAACTCTCCAAGTTATCAGAGATTGTCTACGACAGTACCGCCGTCACCT
    <<<<<<


541 TCGGGTATGACGAGACCACTGGTGTCTTGAAGATGGTCAACCTCCAAAGTGGGGGCTTCT




601 C
```

KEYS (in order of precedence):

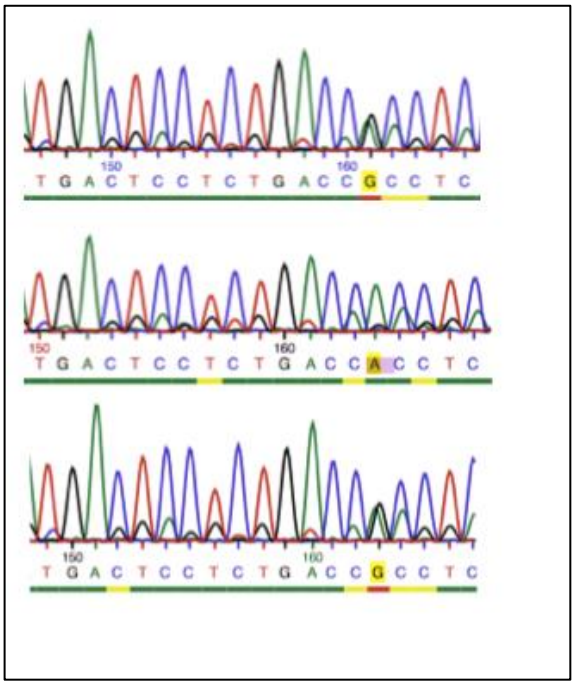
XXXXXX excluded region

****** target

>>>>>> left primer

<<<<<< right primer

B



ii. Kuwait Medical Genetics Centre

Target mutation analysis via Sanger sequencing for the Middle Eastern mutation (pM743T) in exon 12 of the GNE gene at KMGC is performed for all suspected cases of GNE myopathy. Total genomic DNA is extracted from approximately 4 millilitres of peripheral blood to be processed by using the Maxwell® 16 System (Promega). Amplification of target sequence is performed by standard protocols of polymerase chain reaction (PCR), using apparatus GeneAmp (Applied BioSystem). Amplification is performed using a pair of primers: "sense", 5'-GAACAGCTTTGGGTCTTGGG-3' and "antisense", 5'-CTGGACTACACAACACGCAG-3'. The GoTaq® Green MasterMix (Promega) is used in all the amplifications. For direct sequencing, the purified PCR products are prepared according to manufacturer's instructions using a 3130 Genetic Analyzer (Applied BioSystem) automatic sequencer. The sequencing results are then analysed in SeqScape V2.6 software (Applied BioSystem) and compared with the reference database (NM_005476).