The role of genetic variation in determining m.3243A>G variant heteroplasmy

Alia Saeed

Supervised by:

Dr. Sarah J. Pickett

Professor Gavin Hudson

Professor Robert W. Taylor

This thesis is submitted for the degree of Doctor of Philosophy

Mitochondrial Research Group

Biosciences Institute, Faculty of Medical Sciences

June 2024

Author's declaration

This thesis is submitted for the degree of Doctor of Philosophy at Newcastle University. I, Alia Saeed, declare that the work described here is my own, unless where clearly acknowledged and stated otherwise. I certify that I have not submitted any of the material in this thesis for a degree qualification at this or any other university.

Alia Saeed

Abstract

Mitochondrial DNA (mtDNA) is a maternally inherited, multi-copy genome that has the potential to be heteroplasmic, meaning that there can be different populations of mtDNA in a single cell. This is caused by the presence of different alleles, some of which can be pathogenic. Different alleles can be acquired throughout our lifetime in the context of aging for example, due to somatic mutation, but also inherited through the maternal lineage. Heteroplasmy levels vary between individuals, tissues, and cells, and there are various hypotheses that try to explain the variability in the level of pathogenic variants a disease carrier or affected mother can transmit to her offspring. The exact mechanisms involved in this variability are yet unclear however, genetic bottlenecks, mtDNA segregation, random genetic drift, and selection are the major candidates. The m.3243A>G variant is the most common, heteroplasmic pathogenic mitochondrial variant that is associated with several mitochondrial disorders, most notably MELAS (Mitochondrial Encephalopathy, Lactic Acidosis, and Stroke like episodes). Knowledge of the genetic factors that influence levels of pathogenic mtDNA variants may contribute to drug discovery that manipulates heteroplasmy levels, as well as aid clinicians and genetic counsellors with providing accurate success rates in case mitochondrial donation therapies are considered.

Family-based heritability studies have estimated that ~72% of the variance in m.3243A>G, can be attributed to additive genetic factors; the aim of this project was to identify these factors to further our knowledge of the pathways that influence m.3243A>G variability between individuals. To elucidate this, nuclear genome wide association studies (GWAS) were performed with the variable, age corrected m.3243A>G variant allele levels as the phenotype, within a cohort of 408 individuals carrying the m.3243A>G mutation from a multicentre cohort, which includes samples collected from centres across the UK, Italy, and Germany. Additional m.3243A>G carriers were identified using whole genome sequencing data from two large publicly available datasets (UK Biobank (UKBB) and 100,000 genomes project (100kGP Genomics England)). None of the GWAS analyses yielded a significant association peak. META analysis combining GWAS data of the large public cohorts revealed a peak approaching significance on chromosome eight, led by SNP with rs1512802 (8:5882269G>C) (-log10(Pvalue) = 6.9).

A mtDNA-GWAS was undertaken to determine whether mtDNA sequence context influences m.3243A>G levels, documenting an association in the 100kGP cohort which

mapped to haplogroup U (m.16356T>C, -log10(Pvalue) = 3.5). This association was not observed in the META analysis, which combined the mtDNA GWAS analyses on 100kGP and the UKBB cohorts, suggesting that the association might have been a false positive. These results indicate that sample size is a significant limitation of this study, necessitating the identification of further m.3243A>G carrier samples to increase analysis detection power.

Acknowledgments

However I phrase this, it will not capture the depth of my feelings, but I will ink your names down as it is the least I can do.

I owe a profound debt of gratitude to Dr. Sarah Pickett, my primary supervisor, for your guidance, patience, help, and support throughout the past three challenging years – thank you. A huge thank you to my supervisors, Prof. Gavin Hudson and Prof. Rob Taylor, for your invaluable guidance and advice whenever needed, and for your banter that lightened this journey! My panel members, Prof. Heather Cordell and Prof. Bobby McFarland, for your insightful and constructive feedback during our annual meetings.

Thanks to Dr. Stuart Cannon and Dr. Kashyap Patel for kindly taking the time to perform the analysis on the UKBB cohort. Of course, a heartfelt thank you to the patients who gave consent for their samples to be collected, making this research possible, you are at the forefront of our thoughts as we conduct our research!

It was a pleasure to be part of the Pickett lab and the wider WCMR group, working alongside such inspiring and respected professionals. I'm grateful to Dr. Angela Pyle, the first person I met in person at the centre, who immediately put my heart at ease – thank you for being such a thoughtful, kind, and positive presence. Dr. Uwe Richter, for your career advice and the thought-provoking scientific and cultural chats!

To the beloved past and present members of "big office big lads", your tea breaks, and office chats were incredibly annoying during write-up, but you (including the breaks and chats) are a huge part of what made these past three years fly by. Our muddy walk to the now-deceased Sycamore tree is a beautiful memory that I will forever cherish. A shout-out goes to the M4.046 GC ladies.

Dr. Yasmin Tang, Dania Hammadi, Varvara Koraki Foli, Dr. Ana Andreijovic, a special thank you to each of you for being the beautiful human beings that you are. I feel blessed to have crossed paths with such intelligent and kind young women. I look forward to seeing all of us reach our goals and do whatever brings us joy in life!

To the friends who have withstood the test of time and distance: as I write this, we have not seen each other in seven years, yet you have remained the source of tummy-aching laughter, comfort, and advice that I needed every single time. Osama Al Amer, Mohammad Salame, I could not ask for better friends. To the lovely people of our little

Russian-speaking society, the Arab society of NU, and Vita pals and on-site team, thank you for making Newcastle city centre feel safe and homey, even on weekends!

Last but certainly not the least, a deep bow of gratitude to my treasured family. I would not have reached this milestone without you. Your encouragement and belief in me, even when I had none left, pushed me more than you can imagine. Your warmth, care, and love transcended oceans. I love you all.

Adel, my one and only big-baby brother, a separate thank you goes to you. Despite being the youngest, you (always) knew what to say, and how to comfort, encourage, and push me forward! You are my life-long friend, and I could not have wished for a better one. I'm so proud of you, and I have no doubt that you will achieve everything that you truly set your mind to (even I made it!)

Mum and Dad, from the deepest bottom of my heart, thank you. For dedicating yourselves to our family, for going above and beyond to secure a good life for your now not-so-little kiddos. Thank you for being the haven to which I always yearned to return, no matter where life took me. You are the roots that keep me growing, and you are the water to my soul. I'm the luckiest to be your daughter.

I dedicate this to both of you.

List of figures

List of tables

List of abbreviations

100kGP 100,000 genomes project

ATP Adenosine triphosphate

BVSR Bayesian variable selection regression

CAP-R Chloramphenicol resistance

chQTL Chromatin quantitiative trait loci

CrJ Cristae junction

CVD Cardiovascular diseases

DdCBEs Deaminase Cytosine Base Editors

Dloop Displacement loop

eQTL Expression quantitiative trait loci

ER Endoplasmic reticulum

ES Effect size

EtBr Ethedium bromide

ETC Electron transport chain

FLD Fisher's Linear Discriminant

GE Genomics England

GMM Generalised mixed modelling

GWAS Genome wide association studies

HEFA Human UK Fertilization and Embryo Authority

HPC High performance computing

IBD Identity by descent estimates

IMM Inner mitochondrial membrane

IMS Intermembrane space

IVF In-vitro fertilisation

LD Linkage disequilibrium

LHON Leber hereditary optic neuropathy

LMM Linear mixed modelling

LOCO Leave one chromosome out

LOD Logarithm of the odds

LRT Likelihood ratio test

MAC Minor allele count

MAF Minor allele frequency

MAM Mitochondria-associated membrane

Mb Mega bases

MCMC Markov Chain Monte Carlo method

MCS Mitochondrial contact sites

MELAS Mitochondrial encephalopathy with lactic acidosis and stroke-like episodes

meQTL Methylation quantitiative trait loci

molQTL Molecular quantitative trait loci

MT-TL1 Mitochondrially encoded tRNA leucine 1

mtDNA Mitochondrial DNA

mtEF Mitochondrial elongation factor

mTERF Mitochondrial transcription termination factor

mtGIF Mitochondrial genomic inflation factor

mtIF Mitochondrial initiation factor

mtLSU Mitochondrial large subunit

mtPCs Mitochondrial principal components

MTRF1L mitochondrial release factor 1

mtSSB Mitochondrial single-stranded-binding protein

mtSSU Mitochondrial small subunit

mtZFN Mitochondrial-targeted zinc-finger nuclease

OMM Outer mitochondrial membrane

OXPHOS Oxidative phosphorylation

PCA Principal component analysis

PCR Polymerase chain reaction

PEO Progressive external ophthalmoplegia

PGC Primordial germ-line cell

PGD Prenatal genetic diagnosis

PIP Posterior inclusion probability

PIPi Inorganic pyrophosphatase

POLG DNA polymerase gamma

pQTL Protein quantitiative trait loci

QC Quality control

REML Restricted maximum likelihood

RITOLS Ribonucleotide incorporation throughout the lagging strand model

ROS Reactive oxygen species

RRM Realized relationship matrix

rRNA Ribosomal RNA

SCI-LITE Single cell combinatorial indexing leveraged to interrogate targeted expression

SNP Single nucleotide polymorphism

SPA Saddle point approximation

SSS Shotgun stochastic search

TCA Tricarboxylic acid cycle

TK Thymidine kinase

tRNA Transfer RNA

UKBB UK Bio bank

WGS Whole genome sequencing

WTCCC Wellcome Trust Case Control Consortium

Table of content

Chapter 1. General introduction

1.1. Mitochondrial Biology

1.1.1 Origin and Evolution

Once existing as free-living prokaryotes relying on the abundant carbon dioxide and water as a source of energy, over 2 billion years ago, an endosymbiotic relationship between two cells (an archaeon and an alpha protobacterium) occurred (Sagan, 1967). This "revolution" in the emergence of complex life was the key to the further development of cellular diversity and complexity, presenting a more efficient energy-generation mechanism that the archaeon's anaerobic metabolism could not have achieved, by that facilitating survival in the new oxygen saturated atmosphere (Muller and Radic, 2016). This was the starting point of a path to evolving into aerobic, eukaryotic cells (Martin and Müller, 1998).

Evidence supporting this theory includes the fact that the mitochondria contain their own DNA, known as mitochondrial DNA (mtDNA), which is completely separate from the nuclear genome, circular as it is in bacteria, and shares similarities in the genes it encodes with bacterial genomes (Oborník, 2019). Over evolutionary timescales, much of the endosymbiont's original genome, known as the mitochondrial genome, appears to have been lost or transferred to the nuclear genome. This created a complex system of mito-nuclear coordination essential for the normal operation of the cell (Lane, 2017). Over time, through a process of gene transfer and gene reduction, this bacterium became an integral part of the host cell, unable to freely exist on its own yet protected from the outside world and able to thrive in its novel environment (Garg, Zimorski and Martin, 2016; Lazcano and Peretó, 2017).

1.1.2 Mitochondrial Structure, Dynamics, and Function

Each mitochondrion is surrounded by two distinct membranes: the phospholipidic outer mitochondrial membrane (OMM), and the cardiolipin rich, inner mitochondrial membrane (IMM) that folds into cristae, which increases the surface area available for ATP generation via the electron transport chain (ETC) (Vaillant-Beuchot et al., 2021). These cristae are the site of oxidative phosphorylation, where ATP is generated through nutrient oxidation coupled with electron transfer (Wilson, 2017) (**Figure 1.1**). Beyond energy production, mitochondria have essential roles in several other cellular functions such as, the regulation of cellular metabolism, steroid synthesis, calcium signalling, and the induction of programmed cell death by apoptosis (Manoli, Alesci and Chrousos, 2007; Spinelli and Haigis, 2018; Haas, 2019). Additionally, changes in ETC affect mitochondrial dynamics in terms of fission and fusion, and mitophagy, a mitochondrial quality control measure. Defects in these processes are associated with various diseases, for example *PARKIN* gene – related early onset Parkinson's disease; mutations in this gene impair the E3 ubiquitin ligase activity of parkin, leading to a failure in marking damaged mitochondria for autophagic degradation. As a result, dysfunctional mitochondria accumulate within cells (Lücking et al., 2000; Clausen et al., 2024).

Human mitochondria contain their own genome, mitochondrial DNA (mtDNA), which is distinct from the nuclear genome. It has a high mutation rate, and although mtDNA repair does happen, it is not as efficient for the mtDNA as the nuclear DNA (nDNA) (Allkanjari and Baldock, 2021). Some of these mutations are linked to different mitochondrial diseases, all which will be discussed in greater detail in sections to follow (**Section 1.2.1** on mtDNA and **Section 1.3** on mitochondrial disease) (Vercellino and Sazanov, 2022). Mitochondria also have their own ribosomes, which synthesize the 13 proteins encoded by the mtDNA, all essential for the oxidative phosphorylation (OXHPOS) system (De Silva et al., 2015).

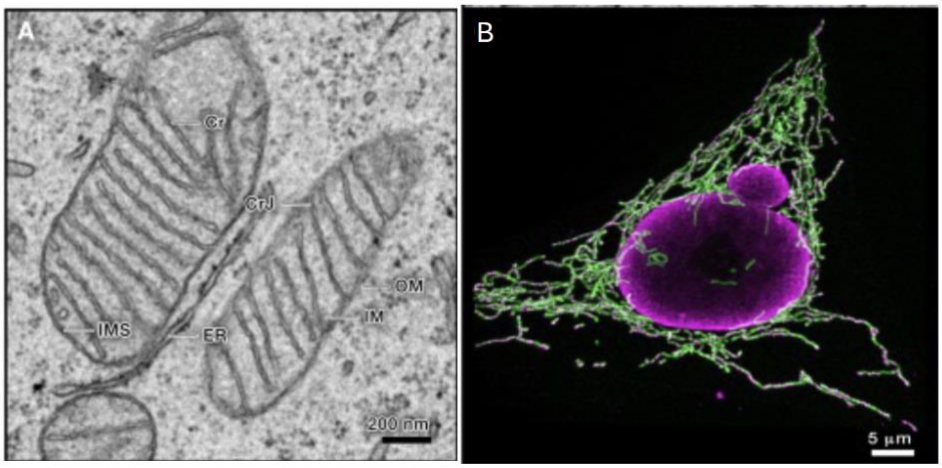

Figure 1.1: Mitochondrial Structure and interactions. A) A transmission electron micrograph of a human embryonic kidney cell line illustrates mitochondrial structure features including the inner membrane (IM), outer membrane (OM), intermembrane space (IMS), cristae (Cr), crista junction (CrJ), and endoplasmic reticulum (ER). (B) An immunofluorescence microscopy image showcases the mitochondrial network (green) and both mtDNA and nDNA (magenta), revealing mtDNA nucleoids within a fused mitochondrial network. [Figure taken from (Suomalainen and Nunnari, 2024)].

1.1.2.1 Fission

Fission is the separation of both the outer and inner membranes, followed by their rejoining to produce two daughter mitochondria. In order to ensure the functionality of each daughter mitochondrion, the amount of protein in the matrix, as well as the intermembrane space needs to be preserved. The scission of the OMM is finely orchestrated by a member of the dynamin GTPase protein, called DRP1. This is mostly found in the cytosol but is recruited to the OMM by an adaptor protein called FIS1; situated in the OMM with the majority of its protein-protein binding sites protruding to the cytosol. MFF is another OMM protein, whose exact function is yet unknown, but is crucial for fission as its loss blocks fission (Mozdy, McCaffery and Shaw, 2000; Fannjiang et al., 2004; Scott and Youle, 2010; Chapman, Ng and Nicholls, 2020). The fission of OMM is well characterised however, the mechanisms of IMM fission are yet to be fully elucidated (**Figure 1.2-A**).

1.1.2.2 Fusion

The proteins involved in mitochondrial fusion are as follows: the OMM proteins; mitofusin 1 (MFN1), and mitofusin 2 (MFN2), and the only IMM protein; optic atrophy factor 1 (OPA1) (Hoppins et al., 2011). These proteins utilize the energy from GTP hydrolysis to merge adjacent mitochondria, which facilitates the exchange of mtDNA, proteins, and metabolites. The fusion of the outer mitochondrial membrane (OMM) is mediated by MFN1 and MFN2, which can form either heterodimers (MFN1 binding to MFN2), or homodimers (MFN2 binding to MFN2) (Hall et al., 2014). IMM fusion is facilitated by OPA1; in its absence only the OMMs are fused, leading to metabolic, mitochondrial disturbances (Chen, Chomyn and Chan, 2005) (**Figure 1.2-B**).

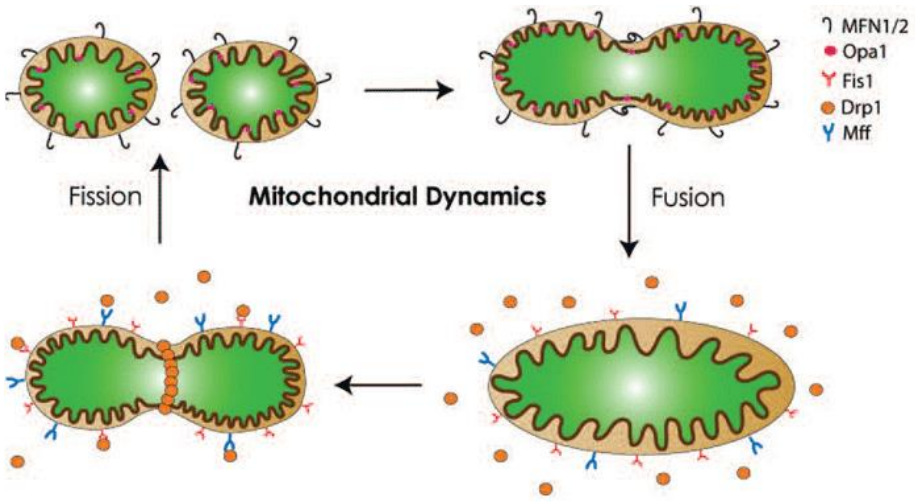

Figure 1.2: Mitochondrial dynamics. (**A**) To initiate fission, the mt OMM needs to recruit Drp1 that binds to the protruding adaptor protein Fis1, and Mff. Drp1 then forms a 'belt' like structure that separates the mitochondria. (**B**) Fusion of the OMM is mediated by MFN1/2, where the formation of dimers pulls the two membranes together; IMM fusion is mediated by Opa1 protein. [Figure taken from (Tieu and Imm, 2014)].

1.1.2.3 Oxidative Phosphorylation and ATP Synthesis

Oxidative phosphorylation (OXPHOS) is at the core of mitochondrial function as the process that converts metabolically derived carbon substrates into ATP, the main energy currency of the cell (Nolfi-Donegan, Braganza and Shiva, 2020). There are around 1136 proteins comprising the mitochondrial proteome, but only 13 of them are encoded by the mtDNA (Rath et al., 2021a). Nuclear genes encode the majority of proteins required for OXPHOS complex construction, organization, maintenance, and function regulation (Wilson, 2017). Variants in these nuclear-encoded genes alter OXPHOS performance and hence exacerbate or help counteract mitochondrial dysfunction caused by mtDNA mutations (Horan, Gemmell and Wolff, 2013). Therefore, the study of mito-nuclear

communication is vital, as there is significant variation in OXPHOS efficiency among individuals carrying pathogenic mtDNA variants.

In the case of glucose metabolism, OXPHOS is preceded by glycolysis, and tricarboxylic acid (TCA) cycle, also known as the Krebs cycle. Glycolysis takes place outside the mitochondria, in the cellular cytosol, and is the process that breaks down glycose into three to five NADH molecules, two pyruvate molecules, with two ATP molecules as a biproduct (Mitchell, 1961; Hatefi, 1985; Yellen, 2018).

NADH and pyruvate molecules are then transported into the mitochondria and are utilised by the TCA cycle that takes place in the matrix. Through a series of redox reactions, this produces two additional ATP molecules, along with NADH and $FADH_2$ molecules that are key for the subsequent electron transport chain (Martínez-Reyes and Chandel, 2020).

The electron transport chain (ETC) is the heart of oxidative phosphorylation. It consists of a sequence of protein complexes and mobile electron carriers embedded within the inner mitochondrial membrane (IMM) as illustrated in **Figure 1.3**. These complexes numbered I through IV, along with ATP synthase (complex V), execute the final and most energy-efficient step of cellular respiration. The process begins with the oxidation of NADH and $FADH_2$ by complexes I and II, respectively. This oxidation causes the release of electrons that pass through the chain, from one ETC complex to another with the help of two electron carriers: coenzyme Q (also referred to as ubiquinone, UQ), and cytochrome *c* (cyt *c*) (Protasoni and Zeviani, 2021).

The ETC is completed by the reduction of oxygen to water by complex IV. Energy released through electron transfer is used to transport protons from the mitochondrial matrix to the intermembrane space, creating a proton gradient between the IMS and matrix. Such a gradient is suitable for ATP synthesis, facilitated by the ATP synthase complex (complex V) that allows protons to flow back through the IMM into the matrix, catalysing the conversion of ADP to ATP through a process referred to as chemiosmosis (Ramchandani et al., 2021).

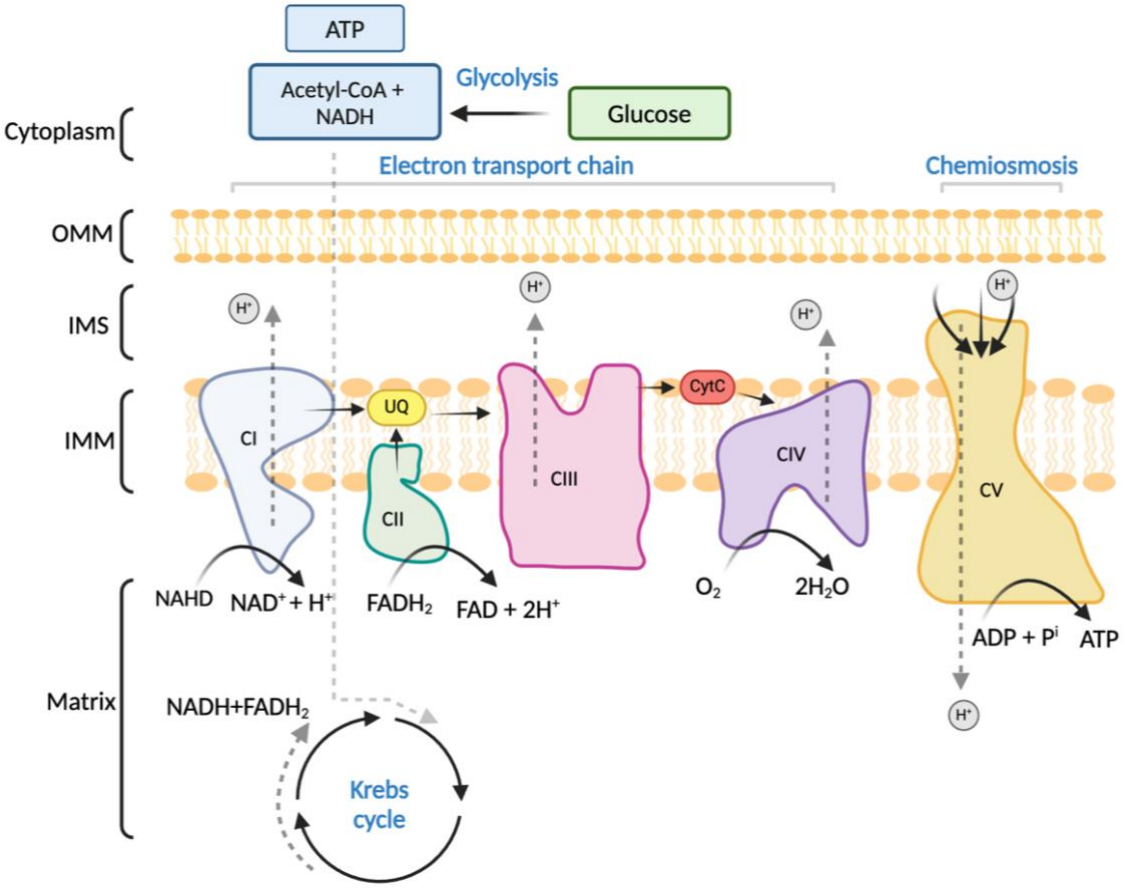

Figure 1.3: Three stages of the cellular respiration of glucose, with emphasis on OXPHOS. Electrons arising from cellular metabolism enter either complex I or complex II through hydrogen donations made by NADH and FADH2, respectively. They are later transferred to coenzyme Q, known as ubiquinone (UQ), hence allowing electron transport from complexes I or II to III. In this phase, electrons are then transferred to complex IV by cyt *c*. Cytochrome *c* oxidase (complex IV) also facilitates electron reduction of O2 given that oxygen is the terminal electron acceptor. The configuration from one complex into another induces the translocation of protons across the membrane to the intermembrane space. This electrochemical gradient gets exploited by complex V, thus synthesizing ATP. [Figure generated with Biorender.com].

1.1.2.4 Apoptosis

Apoptosis, or programmed cell death, is a vital biological event required for maintaining tissue equilibrium, organogenesis, and the removal of compromised or dangerous cells (Kakarla et al., 2020). Intrinsic apoptosis is mainly mediated by mitochondria, a process triggered by a diverse range of cellular stressors, including DNA damage, oxidative pressure, and UV radiation (Nagata, 2018). Key regulators of apoptosis, such as cytochrome *c* and apoptotic protease-activating factor 1 (Apaf-1), are released from mitochondria into the cytosol, where they activate caspase cascades leading to cell demise (Heitzer, Auinger and Speicher, 2020; Obeng, 2021). When cytochrome *c* assembles with Apaf-1 it forms the apoptosome, which activates procaspase 9 into

caspase 9, initiating the activation of downstream effector caspases (Fitzgerald et al., 2022). Altogether, this leads to chromosomal fragmentation. However, it is worth noting that mtDNA remains intact and does not fragment during apoptosis.

Moreover, mito-nuclear interactions play a pivotal role in modulating apoptotic sensitivity. Nuclear-encoded factors, such as B-cell lymphoma 2 (Bcl-2) family proteins, regulate mitochondrial outer membrane permeabilization and apoptotic susceptibility (Hwang, Lee and Paik, 2022).

1.1.2.5 Generation of Reactive Oxygen Species (ROS)

Naturally, superoxide anion, hydrogen peroxide, and hydroxyl radical are by-products of mitochondrial respiration, collectively known as reactive oxygen species (ROS) (Hernansanz-Agustín and Enríquez, 2021). While playing crucial roles in cell signalling and host defence, an overproduction of ROS can overpower the cell's antioxidant defences, leading to oxidative stress and cellular damage. ROS are predominantly produced within mitochondria at complexes I and III of the electron transport chain during OXPHOS, marking OXPHOS as a significant source of ROS in cells.

Mutations leading to OXPHOS dysfunction typically cause electron leakage from the ETC and increased ROS production (Hernansanz-Agustín & Enríquez, 2021). The extent of ROS generation is influenced by various factors; these include the nuclear genetic background of the cells, the specific topology of proton translocation and the inner membrane potential (Li et al., 2022). Besides, compromised antioxidant defence mechanisms such as diminished levels of glutathione and similar variations, amplify the damage caused by ROS in cells (Vianello et al., 2020). Nuclear-encoded factors cooperate with pathways by controlling the generation and detoxification of mitochondrial ROS, having an impact on cellular redox equilibrium (Vianello et al., 2020). In this regard, the upregulation of antioxidant genes is essential for neutralizing ROS creation. Furthermore, ROS-responsive elements in nuclear genes can be mutated and this mechanism may also alter the reactivity of the cell to oxidative stress, affecting the pathogenesis of various mitochondrial disorders (Heitzer, Auinger & Speicher, 2020). Therefore, the interpretation of the interaction between mitochondrial and nuclear genomes in ROS equilibrium is key in discovering intracellular pathogenic mechanisms of mitochondrial dysfunction.

1.1.2.6 Iron-Sulfur biogenesis and Calcium buffering

In addition to their known functions in energy production and metabolism, mitochondria maintain iron-sulfur (Fe-S) clusters and calcium among the key essential regulatory functions (Mühlenhoff et al., 2020). Intracellularly, iron is utilised in three different forms: Fe-S clusters, heme synthesis, and mono/di iron proteins. Both Fe-S clusters and the synthesis of heme take place in the inner mitochondrial membrane (Petronek, Spitz and Allen, 2021). Fe-S clusters are necessary cofactors for many enzymes required in crucial operations such as DNA replication and repair, transcription, translation, and metabolism. The malfunction of this biogenesis mechanism severely affects mitochondrial activity and is linked to various complications, including neurodegenerative conditions and severe metabolic challenges (Tifoun et al., 2021).

1.1.2.7 Calcium buffering

Calcium storage and stability is a key mitochondrial function. Mitochondria serve as calcium stores, managing cytosolic levels and facilitating well-regulated intracellular signalling. The mitochondria absorb calcium ions through a unique pore situated on the internal membrane known as the Mitochondrial Calcium Uniporter (Supinski, Schroder and Callahan, 2020). This procedure assists in stabilizing the flow of cytosolic calcium and influencing cell metabolism. Proper control of calcium is required to avoid the many cellular end routes dependent on calcium, such as apoptosis and necrosis.

1.1.2.8 Inter-organelle Communication: Signal Transduction, Vesicle Transport, and Membrane Contact Sites

Communication among organelles is essential for coordinating different cellular functionalities as well as stimulating responses to various stimuli (Jain & Zoncu, 2022) (**Figure 1.4**). Mitochondria dynamically interact with other organelles, including the endoplasmic reticulum (ER), the nucleus, and the plasma membrane. Such interrelations include signal transduction pathways, vesicle-mediated molecule transmission, and membrane contact sites. Mitochondria release significant signal molecules such as reactive oxygen species and mitochondria-derived peptides in coordinating cellular responses to stress, energy requirements, and cell destiny resolution (Krupinska et al., 2020). Signalling transmits the stimulation to various cellular events that transpire in

processes such as apoptosis, inflammatory responses, and metabolism. Mitochondrial signalling pathways are implicated in most diseases, including cancer and metabolic syndromes (Amorim et al., 2022; Popov, 2020). Vesicles permit the transmission of molecules between the mitochondrion and various cellular compartments. Mitochondria-associated membranes (MAMs) are linking systems that function in lipid distribution, calcium signalling, and protein trafficking between the mitochondria and another organelle. MAM deregulation has been associated with impaired mitochondrial metabolism, ER stress, and neurodegenerative processes (Liu & Yang, 2022). Membrane contact sites, specifically two organelles' membranous membrane touch, permit immediate communication and the translation and synthesis of reactants of different metabolites and signal molecules. Mitochondria establish contact sites (MCSs) with various organelles, this occurs via unique tethering proteins that connect the spaces of adjacent radicals. Dysfunctional MCSs have been associated with increased incidences of metabolic diseases, neurodegeneration diseases, and increased viral load (Barazzuol, Giamogante & Calì, 2021).

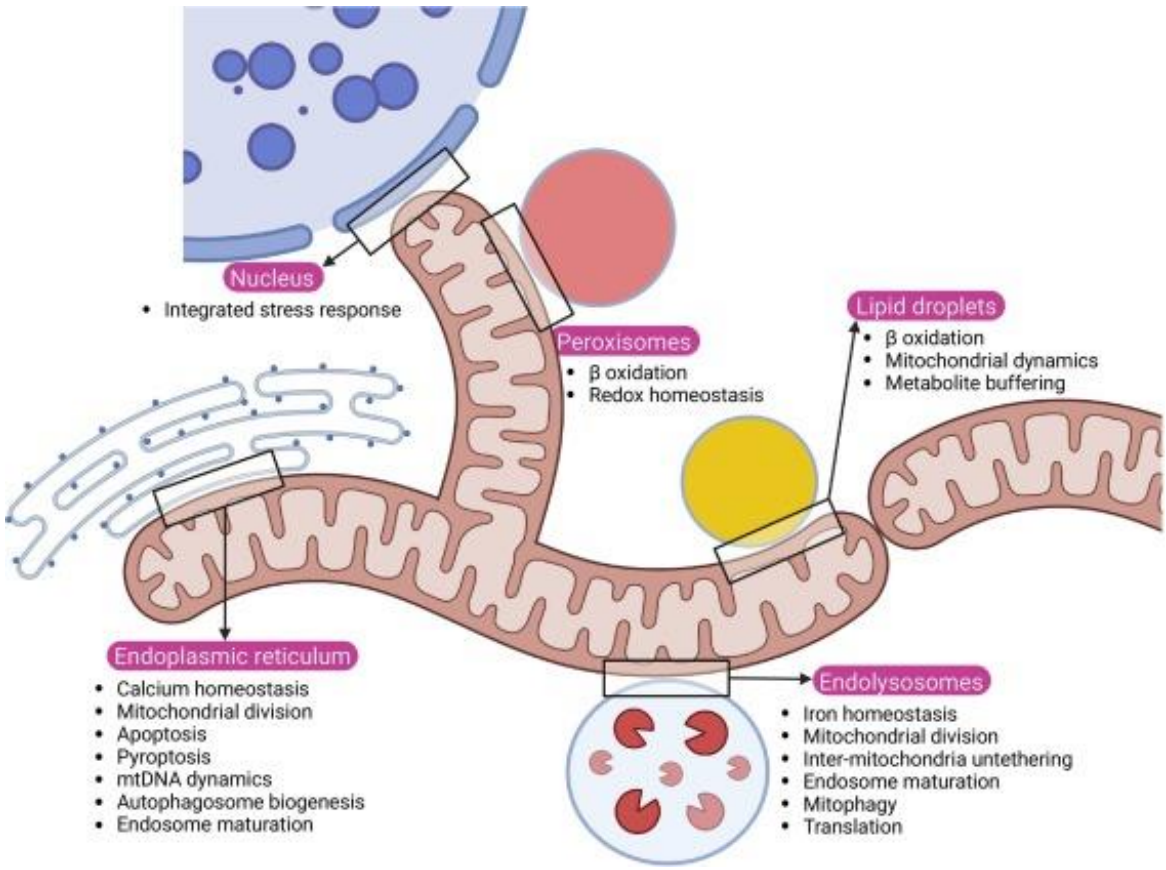

Figure 1.4: Mitochondrial contact sites. Schematic presents some functions that are mediated by contact sites between the mitochondria and various other cellular components, primarily via the flux of different metabolites. [Figure reproduced from(Collier et al., 2023)].

1.2. Mitochondrial genetics

1.2.1 mtDNA structure

Reflecting their bacterial origin, mitochondria contain their own double stranded, circular DNA (mtDNA). 16,569 bp in length, the human mtDNA contains 37 genes which encode 13 proteins that constitute core components of four of the five OXPHOS subunits (excluding complex II that is entirely nuclear encoded), 22 transfer RNAs (tRNA), and 2 ribosomal RNAs (rRNA) all required for the expression of these proteins (Anderson et al., 1981a). Most of the genes are encoded on the outer heavy, outer strand, named due to its higher G/C nucleotide content and thus, higher molecular weight (in comparison to the inner, light strand which is A/T rich). The mtDNA control region (1124bp), which has the highest substitution frequency, is the only non-coding region of the genome and serves as a promoter for both the H and L strands, and harbours the displacement (D) loop (**Figure 1.5**. The individual letters in the mtDNA are designated for tRNAs, and they have been referred to as punctuation marks in the 'tRNA punctuation model of processing' proposed by Ojala, Montoya and Attardi, (1981). The tRNA punctuation model of processing describes the mechanism by which mitochondrial precursor RNA transcripts are converted into mature tRNAs, rRNAs, and mRNAs. In mtDNA, genes for these RNA molecules are often arranged in a continuous sequence, with tRNA genes acting as punctuation marks. Enzymes recognize these tRNA sequences (by their nucleotide sequences) and cleave the long precursor transcript at these specific sites, thereby releasing individual tRNAs along with rRNAs and mRNAs. This precise cleavage facilitates the further processing and maturation of these RNA molecules, ensuring the proper function of mitochondrial gene expression. This model underscores the vital role of tRNA sequences in guiding the accurate processing of mitochondrial RNA ahead of translation (Lopez Sanchez et al., 2011).

The mitochondrial genetic code has some minor differences to the nuclear genome. Most notable is the use of two stop codons, which are "AGA" and "AGG" in mtDNA but are "UAA," "UAG," and "UGA" in the nuclear genome (Yamamoto et al., 2020). Lastly, "AUA," codes for methionine in mtDNA while coding for isoleucine in nDNA. Most mitochondrial protein synthesis is governed by the mitochondrial genome, but over 99% of proteins required for mitochondrial structure and functioning are encoded by nDNA (Calvo, Clauser and Mootha, 2016). Mitochondrial proteins that are encoded by the nuclear

genome are synthesized in the cytosol and successively imported into the mitochondria using well-defined import pathways. Proteins destined for the mitochondrial inner membrane and matrix are characterized by N-terminal presequences (Supinski et al., 2020). The sequences facilitate their import by specific translocases that are present in the outer and inner membranes of the mitochondria, TOM and TIM, respectively. Following translocation, these sequences are removed by the mitochondrial processing peptidases to ensure the correct maturation of the respective proteins in the matrix (La Morgia et al., 2020). This symbiotic relationship between the nuclear genome and the mitochondrial import machinery demonstrates the intricate mechanisms by which the nuclear and mitochondrial genomes interact to ensure cellular homeostasis (Walker & Moraes, 2022).

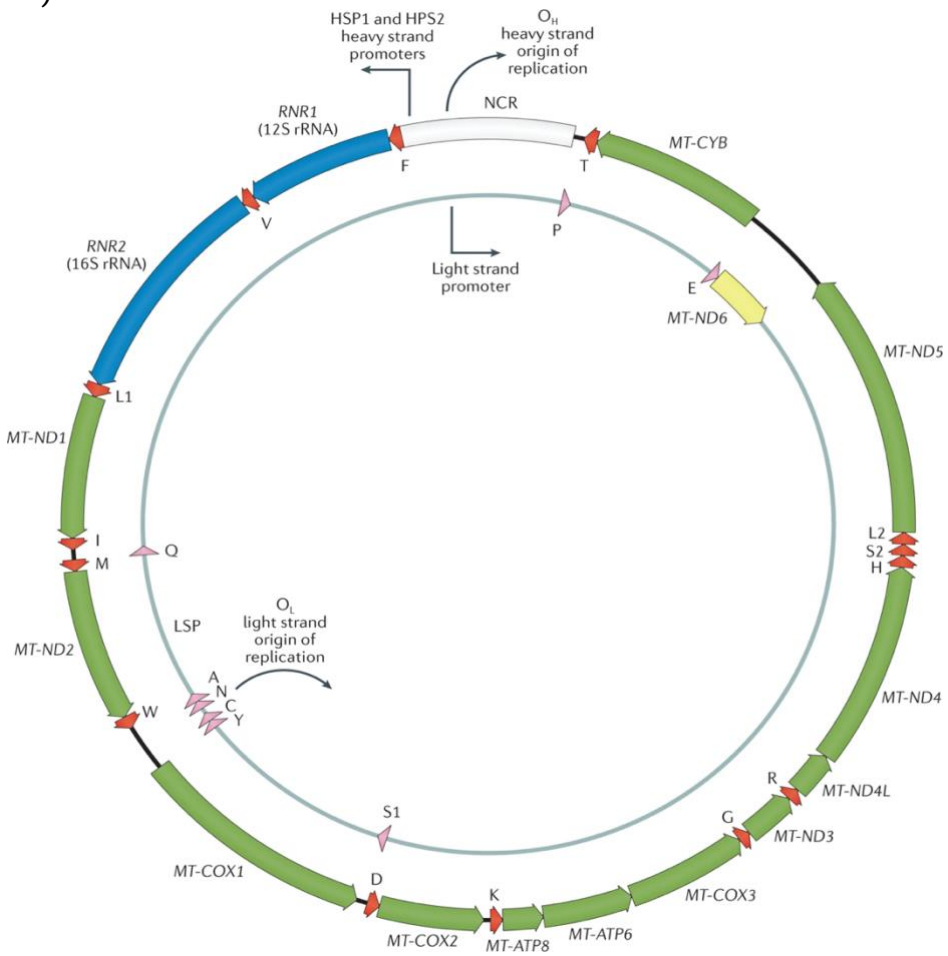

Figure 1.5: The mammalian mitochondrial genome. Counterclockwise, figure depicts the mitochondrial double stranded DNA with labels of mtDNA genes. HSP1 and HSP2: heavy strand promoter regions; LSP: light strand promoter region; *RNR1* and *RNR2* genes: that respectively encode for 12S and 16S rRNAs; *MT-ND1, MT-ND2, MT-ND3, MT-ND4, MT-ND5, and MT-ND6:* genes encoding subunits of Complex I: ND1, ND2, ND3, ND4L, ND4, ND5, and ND6; OL: light strand origin of replication; *MT-COI, MT-COII, MT-COIII*: encoding subunits of Complex IV: COI, COII, COIII; *MT-ATP6* and *MT-ATP8*: encoding the two subunits of Complex V ATPase 8 and 6; *MT-CYB*: encoding Cyt b protein, a polypeptide that forms one subunit of the respiratory chain Complex III; OH: heavy strand origin of replication; and the non-coding region (NCR), that includes the displacement D-loop; Single letters present the 22 tRNA genes [Figure obtained from Stewart and Chinnery, (2021)].

1.2.2 Mitochondrial central dogma

1.2.2.1 Replication

Mitochondrial DNA replication is governed by a different set of mechanisms compared to nuclear DNA. The replication of mtDNA is carried out by DNA polymerase γ and specific associated proteins (Falkenberg, Larsson and Gustafsson, 2007; Chapman, Ng and Nicholls, 2020; Falkenberg and Gustafsson, 2020). Errors in mtDNA replication are linked to numerous mutations that result in mitochondrial diseases and have also been suggested as a contributing factor to the aging process (Fontana & Gahlon, 2020). Replication commences at the core of mtDNA nucleoids, which are discrete spheres that are roughly ~100nm in diameter, each containing mtDNA and its associated proteins (Robinow and Kellenberger, 1994; Lee and Han, 2017), and necessitates a strictly regulated concourse between mitochondrial and nuclear-encoded factors, again emphasising the importance of mito-nuclear communication (Roy et al., 2022).
MtDNA replication is a critical cellular process that is central to cellular health and disease (Peeva et al., 2018). This requires the input of carefully assembled nucDNA-encoded proteins (Fontana & Gahlon, 2020). The precise molecular mechanism of replication is still under debate; three hypothesized models provide an understanding of the synchronous and asynchronous replication of mtDNA (**Figure 1.6**). The strand displacement model is one of the examples of the asynchronous models (Zinovkina, 2019). In this model, replication begins at a single origin and proceeds in a unidirectional manner. A new maternal H-strand is synthesized, and the stabilization of the displaced strand depends on mitochondrial single-stranded DNA-binding protein (mtSSB). Okazaki fragments are absent, and the mtDNA replication process occurs differently from nucDNA replication. The second model, the ribonucleotide incorporation throughout the lagging strand (RITOLS) model (**Figure 1.6-B**), has more similarities with the strand displacement model. However, in this case, RNA, as opposed to DNA, is bound to the maternal H-strand, leading to the process not requiring mtSSB (Holt & Reyes, 2012). The synchronous replication model (**Figure 1.6**) suggests that there is a two-direction replication that occurs from a specific origin and proceeds in both directions (Abraham et al., 2020). Simultaneous leading and lagging strand synthesis occurs, with Okazaki fragments incorporated into the lagging strand. The synchronous model is distinguished from the asynchronous models by the coordinated approach as well as the regulatory aspects that determine how the replication process occurs (Hämäläinen et al., 2019)

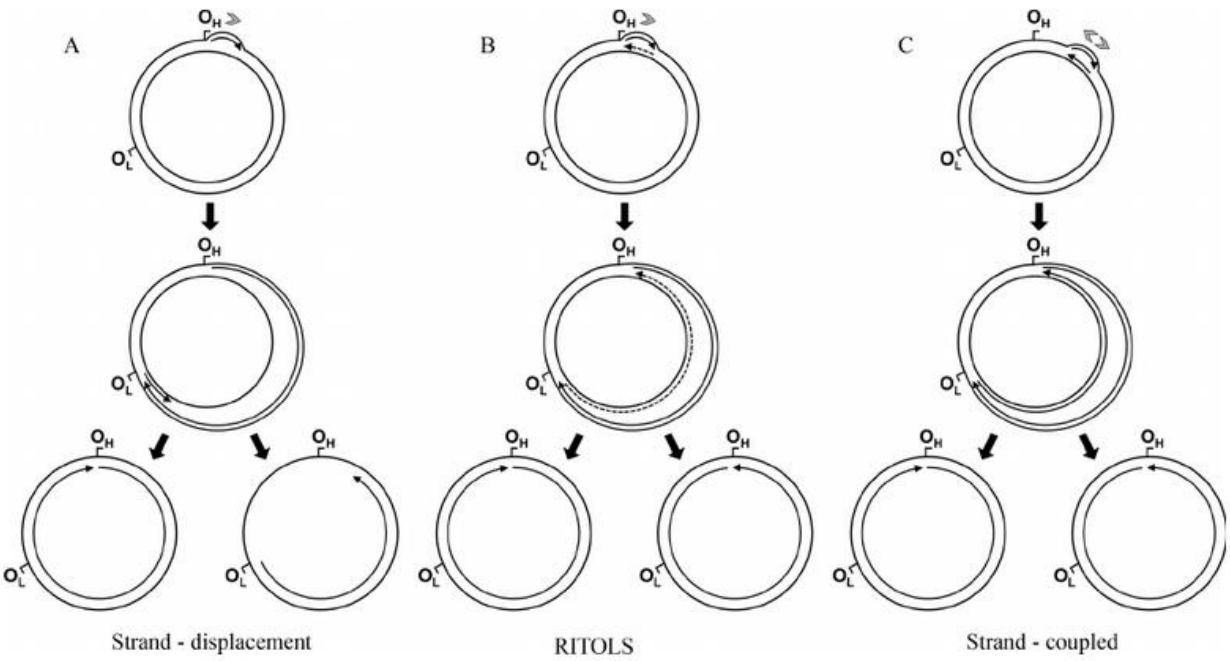

Figure 1.6: MtDNA replication. All models of mtDNA replication: the strand displacement (A); the RNA incorporated throughout the lagging strand (RITOLS) (B); and the leading and lagging strand-coupled (C) models. Arrows associated with replicating mtDNA are in a 5'–3' direction; continuous and dashed lines indicate DNA and RNA, respectively (only the long stretches of RNA as described for the RITOLS model are shown; no possible short RNA primers of the other models are present). Grey arrowheads indicate the quantity and the directionality of replication forks produced at the origin, according to each model. [Figure obtained from Zinovkina (2019)].

Mitochondrial basal replication machinery consists of five proteins and a DNA substrate, this is depicted in **Figure 1.7**. DNA polymerase gamma (POLγ or POLG) which is the only DNA polymerase active in the mitochondria, is a heterotrimer that consists of one catalytic POLγA (encoded by *POLG* gene), and two monomers of the processivity subunit POLγB (encoded by *POLG2*). In addition to DNA helicase twinkle, and mitochondrial single-stranded DNA-binding protein mtSSB (Wanrooij and Falkenberg, 2010). POLγA, mainly functions in proofreading during replication (Lim, Longley and Copeland, 1999). One POLγB monomer works by enhancing the replication rate, whereas the second subunit, closest to POLγA, stimulates enzyme-DNA interaction. The twinkle helicase moves ahead of the polymerase, unwinding the molecule and creating the mtDNA replication fork (Milenkovic et al., 2013). MtSSB has an essential function in protecting the ssDNA from nucleases and making sure that strands do not refold. Additionally, mtSSB stimulates primer recognition, which enhances mtDNA synthesis (Thömmes et al., 1995).

In addition to its role in transcription, POLRMT serves a role in both leading and lagging strand replication at the heavy strand origin of replication (OH) and the light strand (OL), where it functions as a primase by generating RNA primers needed for replication (Wanrooij et al., 2008).

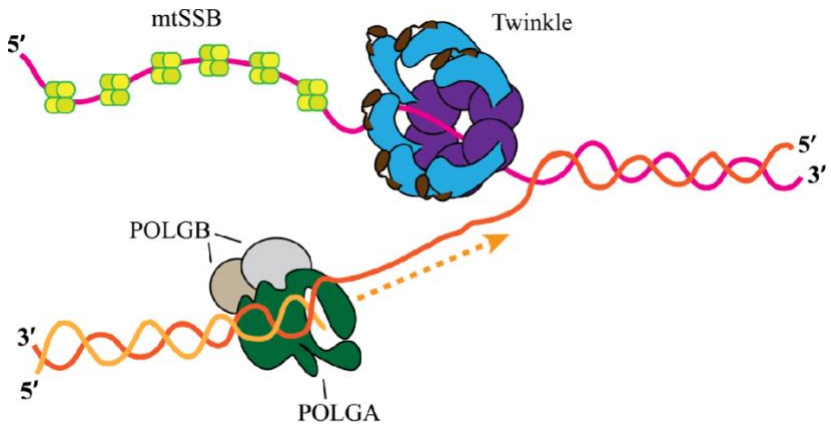

Figure 1.7: Mitochondrial replisome. An orange arrow with dashed lines depicts leading-strand synthesis by human DNA polγ and its components, the catalytic POLGA and two monomers of the accessory POLGB. TWINKLE moves ahead of the polymerase unfolding mtDNA strands, while mtSSB attaches to ssDNA preventing the refolding of DNA strands and protecting from nucleases. Meanwhile, POLRMT acts as a primase, creating primers at OL and OH necessary to initiate replication [Obtained from Farnum (2013)].

1.2.2.2 Transcription

Mitochondrial transcription involves a complex of several proteins, including POLRMT, TFAM, TFB2M, and MTERF-1, which have distinct functions during the initiation, elongation, and termination of mtDNA transcription (Fisher and Clayton, 1988; Tiranti, 1997; Falkenberg et al., 2002; Rebelo, Dillon and Moraes, 2011; Falkenberg, Larsson and Gustafsson, 2024). The isolation of POLRMT demonstrated that the protein is the enzyme responsible for the initiation of mtDNA transcription and as stated in the previous section, the synthesis of transcription primers that are degraded after initiation and replaced with fresh primers to allow for the elongation of newly polysomal RNA (Fontana & Gahlon, 2020). The second factor is TFAM, which has been identified as a critical determinant of the stability, packaging, and replication of mitochondrial DNA (Yakubovskaya et al., 2010). TFAM is particularly relevant to mammalian neurodegenerative diseases because of its role in the regulation of mtDNA copy number and transcription initiation (Song et al., 2024). Using knockdown experiments, upon the reintroduction of TFAM, the regulation of inflammatory responses due to the

accumulation of cytoplasmic, free mtDNA molecules, was restored (Liu et al., 2024). It is suggested that LIF2 motif of TFAM (Leucine rich sequence that helps TFAM bind to mRNA) binds to autophagy marker, LC3B; this activates the TFAM-mediated lysosomal activation pathway which degrades the leaked mtDNA molecules, by that exhibiting a protective mechanism by regulating mtDNA-driven inflammation (Liu et al., 2024). Additionally, TFAM interacts with two other proteins, TFB2M, and MTERF-1, to facilitate transcription, elongation, and termination. Transcriptionally, TFB2M is essential for the initiation of transcription, where it binds POLRMT in the presence of TFAM to initiate transcription (**Figure 1.8**).

Once initiation is completed, TFB2M is dissociated, and the elongation complex is attached to mtDNA (Barshad et al., 2018). Finally, MTERF-1 is responsible for the transcription termination process to ensure that each terminal segment is transcribed correctly. It is essential for the expression and maintenance of the mtDNA genome as well as cellular bioenergetics (D'Souza & Minczuk, 2018).

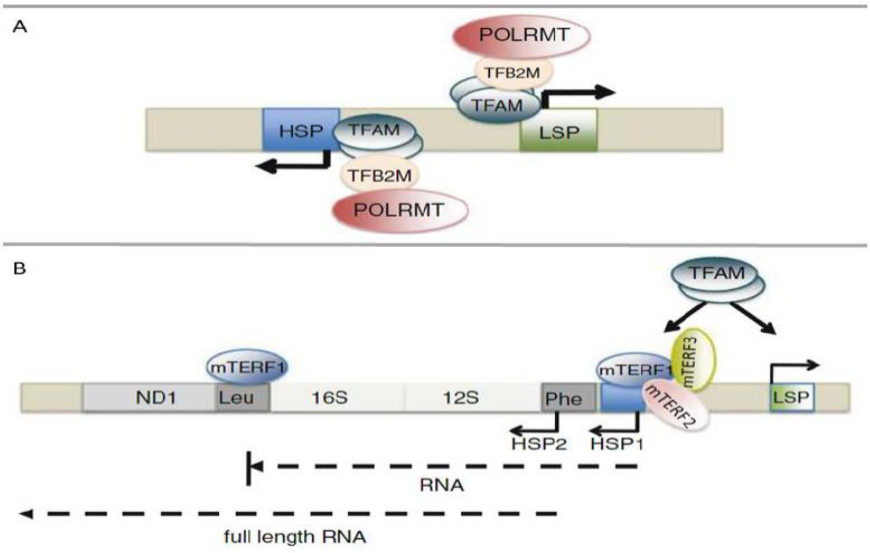

Figure 1.8: Mitochondrial DNA transcription. Mitochondrial transcription initiation involves bi-directional transcription of mtDNA mediated by TFAM, TFB2M, and POLRMT. The process starts with the binding of TFAM upstream from the HSP and LSP, setting off a torsional wave that unwinds the mtDNA helix. This unwinding then facilitates the binding of TBM2M and POLRMT to the promoter. The localization and dynamics of these factors have been explored in vivo, providing insights into the transcription initiation mechanism of mitochondrial genes within the nucleoid context. [Adapted from Rebelo, Dillon & Moraes (2011)].

1.2.2.3 Translation

Primarily, mitochondrial translation is facilitated by the mitochondrial ribosome, which like the cytosolic ribosome is composed of two subunits, the large (mtLSU) and small (mtSSU) (Farge & Falkenberg, 2019). It is also highly specialized since it ultimately translates crucial components of the OXPHOS system (Boczonadi, Ricci & Horvath, 2018). However, the mitoribosome's composition presents adaptations to the matrix environment of the mitochondria and is a product of mitochondrial and nuclear genome-encoded ribosomal components. **Figure 1.9** shows the steps involved in mitochondrial translation.

Human mitochondrial translation comprises four distinct phases: initiation, elongation, termination, and recycling. This process guarantees the generation of the 13 mtDNA encoded proteins, key components of the OXPHOS system (Iannello et al., 2019). The following discussion explores these subsides, emphasizing the fundamental mechanisms and their relevance.

Initiation: The initiation (phase 1 in **Figure 1.9**) of mitochondrial translation occurs when the mitochondrial ribosomal subunits, together with initiation factors (mtIF2 and mtIF3) and mitochondria specific protein mS37, form the pre-initiation complexes. mtPIC-1 is formed upon the binding of mtIF3 and mS37, then the mtPIC-2 upon the binding of mtIF2 (Mai et al., 2017). This facilitates the establishment of interactions within the initiation complex and the recruitment of the initiation complex to the mt-tRNA start codon (AUG or AUA) positioned in the ribosomal P site. Unless mtIF2 binds fMet-tRNA^{Met} to mt-mRNA, transcription is not initiated, and the mRNA is released (D'Souza and Minczuk, 2018; Khawaja et al., 2020). Upon the successful recruitment of all necessary elements, polypeptide chain synthesis is initiated.

Elongation: During the elongation phase, amino acids are sequentially added to the growing polypeptide chain. This step is enabled by mitochondrial elongation factors that physically guide aminoacyl-tRNAs to the A (acceptor) site of the ribosome (mtEF-G1). This is an energy demanding step that is facilitated by the hydrolysis of the active form EFTU·GTP. EFTU·GTP then is released from the ribosome in its inactive form known as EF-TS, that is then activated by the addition of a GTP molecule (Wang et al., 2021). The precise matching of the tRNA anticodon with the mRNA codon initiates the formation of a peptide bond, allowing the ribosome to translocate along the mRNA (D'Souza &

Minczuk, 2018). This movement ensures that the polypeptide chain is synthesized in alignment with the genetic information.

Termination and Recycling: Finally, termination is activated when an mRNA stop codon is presented at the mitoribosome A site, leading to the release of the newly synthesized polypeptide (De Silva et al., 2015). This phase involves mitochondrial release factors, such as mitochondrial release factor 1 (MTRF1L), which specifically binds to the mRNA stop codons and catalyses the release of the polypeptide and the tRNA from the ribosomal E site. Subsequently, ribosome recycling factors break down the post-termination ribosomal complex, disassembling the ribosomal subunits (with the help of EF-G2mt and MRRF) to make them available for initiating another translation round. This recycling stage is essential for maintaining translation efficiency, ensuring a supply of free ribosomes for new translation cycles (Hämäläinen et al., 2019). The coordination between these phases guarantees the swift and precise synthesis of mitochondrial proteins, crucial for both mitochondrial and cellular functionality.

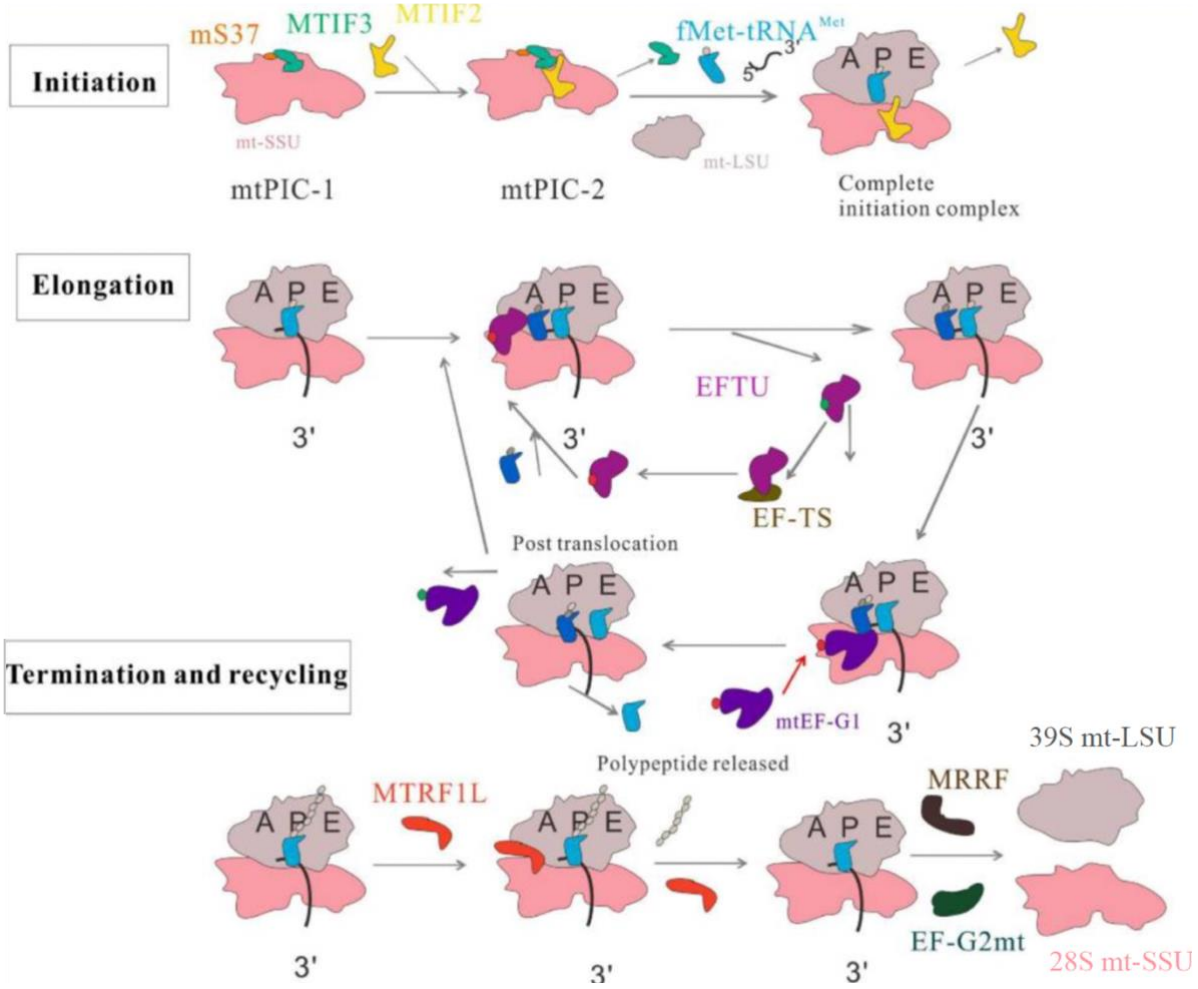

Figure 1.9: Human mitochondrial translation. The process is divided into several phases, including initiation, elongation, termination, and recycling. In the initiation phase, two identified pre-initiation assembly steps, mitochondrial preinitiation steps 1 and 2, have to occur for successful initiation. In the elongation phase, the aminoacyl-tRNA is transferred with the help of GTP to the A site of the mitochondrial ribosome. Meanwhile, the P site (peptidyl tRNA site) holds the tRNA carrying the growing polypeptide chain. Upon the completion of translation, tRNA is transferred to the ribosomal E site (exit site) ahead of being released from the ribosome. The termination of translation is initiated by MTRF1L, and the disassembly of the ribosomes is led by MRRF and EF-G2mt. [Diagram from (Wang et al., 2021)].

1.2.3 Heteroplasmy

The mitochondrial genome is polyploid; with copies relating to energy demand, one cell can host up to a few thousand mitochondria, each carrying multiple mtDNA molecules. Homoplasmy is when all the molecules in a cell or tissue are identical. On the other hand, heteroplasmy reflects the presence of different mtDNA populations within a single cell or tissue (Hauswirth and Laipis, 1982a). In an analysis of 56,434 mtDNA WGS data from gnomAD v3.1 database, it was estimated that 85% of unique mtDNA variants are homoplasmic (Laricchia et al., 2022). Although the majority of the identified pathogenic

mtDNA variants are heteroplasmic (Hong et al., 2023), it is important to note that not all heteroplasmies are pathogenic or linked to disease. In fact, it is estimated that almost every control or "healthy" individual harbours at least one heteroplasmic variant at an allele frequency between 0.5-1.5% (Wei et al., 2019a; Stewart and Chinnery, 2021a). Another recent study performed in the UK Biobank, found that 30.5% participants had at least one detectable heteroplasmy, which affected one of 10,161 sites (Hong et al., 2023). There are various hypotheses that try to explain the variability in heteroplasmy levels between individuals; random segregation, selection, and genetic bottlenecks being the major candidates.

1.2.3.1 The threshold effect

Heteroplasmy gives rise to a phenomenon known as the threshold level, which reflects the proportion of variant needed to manifest a phenotype (**Figure 1.10**). In other words, it is the level of pathogenic variants at which the wild type mtDNA can no longer compensate for the damaging effects (disrupted OXPHOS efficiency) caused by the pathogenic variant (Wallace, 1992; Rossignol et al., 1999).

This was first investigated as the 'mutation load' effect by Wallace, (1986). For his investigations, Wallace used cybrid cells with a 16S rRNA gene mutation, and investigated chloramphenicol (CAP) resistance, which is an antibiotic that inhibits mitochondrial protein synthesis by targeting the mitochondrial ribosome. CAP resistance was found to occur only when >10% of the mtDNA carried the mutant, CAP-R variant. *Transmitochondrial* cybrid cells, or simply cybrid cells, are cells that have had their mtDNA content entirely depleted, and then replenished from a donor cell (King and Attardi, 1989a) (more detail on cybrids in last paragraph of **Section 1.2.4.2.1**).

This threshold has been known to differ for each pathogenic variant (Shoffner et al., 1990a; Rossignol et al., 2003). For example, the most recent estimation of the threshold for m.3243A>G in muscle fibres is ~83% (Ahmed et al., 2022).

Rossignol et al., (1999) has also suggested that the threshold for the same variant may differ between tissues; given the variable level of tissue sensitivity to defective OXPHOS. Moreover, this threshold is unlikely to be static; it is likely to depend on a range of factors affecting– from mito-nuclear genetic interplays, environmental impacts and distinct metabolic requirements of tissues.

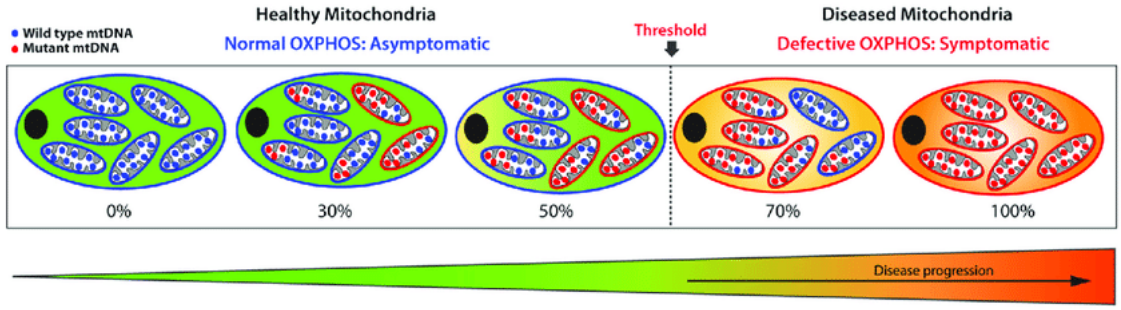

Figure 1.10: Mitochondrial heteroplasmy and the threshold effect. Cartoon depicts different states of mitochondrial mtDNA content, where they can be completely either wild or mutant type, and are thus called homoplasmic, or a mixture of both, which is called heteroplasmic. When the level of mtDNA mutant types becomes intolerable and mitochondria's OXPHOS efficiency is disrupted, this marks the crossing of that specific variant's functional threshold. [Figure obtained from (Li et al., 2021)].

1.2.3.2 Tissue Distribution and Specificity

Heteroplasmy is thought to be influenced by various factors, including mito-nuclear genetic variation, tissue-specific energy demand, and replication advantage of distinct mtDNA variants (Gupta et al., 2023). It is established that the level of pathogenic mtDNA variants differs between tissues. It is also well known that tRNA point mutations such as, m.3243A>G, and deletions have higher levels in post mitotic tissues such as skeletal muscle, compared to the constantly dividing tissues like blood and the epithelium (Chinnery et al., 1999; Stewart and Chinnery, 2021b). This is partially explained by the non-dividing nature of postmitotic tissues, which consequently makes muscle tissue heteroplasmy the most accurate measure due to its stability. On the other hand, heteroplasmy levels in mitotic tissues such as blood were found to be different for each variant, some are stable over time, such as the m.8344A>G variant, whereas others show a negative selection pattern, such as m.3243A>G, whose level decreases with age, which necessitates age-correction upon measurement (Grady et al., 2018; Bernardino Gomes et al., 2021).

Tissue segregation of pathogenic variants varies across individuals however, it was found that m.14487T>C, for example, has the same segregation pattern in monozygotic twins, suggesting the role of nuclear control in tissue segregation (Spyropoulos et al., 2013; Maeda et al., 2016).

1.2.3.3 mtDNA Clonal Expansion

Clonal expansion happens when a certain mtDNA variant in a cell or tissue increases over time in proportion to the rest (Kowald & Kirkwood, 2018). There are multiple theories that try to explain how clonal expansion happens, some of them hypothesise this to be a random process that is independent from cell division and is indifferent to the impact of variants on cellular function or fitness. Where some mitochondrial variants are 'by chance', replicated or passed on more frequently (Bernardino Gomes et al., 2021). Simulation studies have shown that relaxed replication alone is enough to cause changes in variant mtDNA proportions leading to a random genetic drift (Elson et al., 2001b). Additionally, single cell genotyping in two mouse models of human mtDNA disease revealed that variance of heteroplasmy increases equally over time in both mitotic (spleen) and post-mitotic (brain) tissues, implying the important role of relaxed replication in heteroplasmy variance in the absence of cell division (Glynos et al., 2023).

Other hypotheses suggest the presence of selective factors that determine the dynamics of clonal expansion, resulting a strict mtDNA replication. For example, in Diaz (2002), mtDNA deletions were found to be preferentially replicated, presumably due to their faster replication rates compared to full-length molecules. However, this phenomenon was not observed *in situ* and was rather attributed to the experimental conditions, which involved the use of ethidium bromide, a substance known to be mutagenic to nDNA (more details in **Section 1.2.4.2.1**). Another study on mtDNA deletions identified the perinuclear niche as the subcellular origin of clonally expanded mtDNA deletions (Vincent et al., 2018). The observed foci with increased mitochondrial molecules, as well as OXPHOS deficiency; were partially explained by the physical proximity to the nucleus (perinuclear), offering enhanced mito-nuclear signalling (Davis and Clayton, 1996).

The nuclear genetic background, in particular, is likely to influence the dynamics of clonal expansion in a complex manner, through both the replication rate of individual mtDNA and the selective pressures predisposing some variants over the others. Mito-nuclear interactions are crucial in informing the dynamics of clonal expansion because nuclear-encoded factors like Polγ must be in balance for normal replication, repair, and transcription of mtDNA. Therefore, variations in these factors are likely to lead to differences in the replication of variants, thereby influencing the rate of clonal expansion (Trifunov et al., 2018). Ma and O'Farrell, (2016) observed selective expansion of certain

mitochondrial single nucleotide variants (mtSNVs) in drosophila, which were advantageous on the cellular level rather than the whole organism, which is why it is referred to as 'selfish replication'. The level of variants differ between tissues and over time (Goto, Nonaka and Horai, 1990; Rahman et al., 2001). The m.3243A>G for example, is a 'variant with a moderate selection in replicating tissue', where tissues do not show a complete clearance of the variant nor a complete stability over time (Bernardino Gomes et al., 2021). On the other hand, mtDNA deletion disease such as Pearson's disease, shows a strong selection in blood yet paradoxically, is clonally expanded in muscle tissue (McShane et al., 1991; Grady et al., 2014). As mentioned in previous sections, the exact drivers of this differential, tissue-specific expansions are yet to be determined.

Recent work by Kotrys et al., (2024), provides evidence supporting the non-randomness surrounding heteroplasmy variability. Using a novel, SCI-LITE (single-cell combinatorial indexing leveraged to interrogate targeted expression) method, intracellular heteroplasmy was measured in base edited cell lineages within standard culture conditions. They suggest that non-synonymous mtDNA mutations are negatively selected, and that this happens at the level of cellular fitness rather than intracellularly and is fully driven by the conditions surrounding the cell, i.e., in cases where the accumulation of non-synonymous variants is advantageous, a maintenance of non-synonymous mutations is observed. This was explored in dividing tissues, and in artificial environments which do not always fully translate into what is happening in organisms. This, as well, does not explain the observed mutation-specific shifts in blood, where m.3243A>G decreases with age, and heteroplasmic LHON-associated mutations for example, remain stable over time. Providing another instance where additional, heritable factors may be involved.

1.2.3.4 The bottleneck effect

Extreme inter-generational mtDNA heteroplasmy shifts were first observed in Holstein cows; suggesting the existence of a 'mitochondrial bottleneck' (Hauswirth and Laipis, 1982b) (**Figure 1.11**). It was noted that if mtDNA populations were to be uniformly distributed to daughter cells, it would have been unlikely to observe progeny with the variant mtDNA molecules as the dominant population within a short time span as one generation, which is when the idea of a more random segregation pattern appeared

(Hauswirth and Laipis, 1982b; Olivo et al., 1983). The significant reduction in mitochondrial copy number (CN) during meiosis (Jenuth et al., 1996a; Cree et al., 2008a), may lead to a random assortment of mtDNA variants, drastically altering the proportion of mutated versus normal mtDNA, resulting a large difference in heteroplasmy levels across the generated oocytes (Howell et al., 1992). This was later observed in LHON pedigrees, where a rapid shift towards the variant allele at m.11788G>A was observed (Cree et al., 2008a).

The transmission of mtDNA pathogenic variants was found to follow different patterns, and this is explained by multiple factors: Blok et al., (1997) suggested that mtDNA variants affect the size of the bottleneck; variants such as m.8993T>G/C showed more rapid segregation than any other pathogenic variant (Wilson et al., 2016).

A smaller mtDNA CN would mean a tighter bottleneck, which would be more likely to yield oocytes with either extremely high and low variant levels; those with very high levels may be unviable. On the other hand, a greater mtDNA CN results a wider bottleneck and a less rapid segregation, which explains the presence of individuals with more similar heteroplasmy levels in the same pedigree, something that applies to m.3243A>G. It is plausible that these variations cause a difference in mtDNA CN, either by selection, or as a compensation due to the faulty respiration. By studying preimplantation mouse embryos, 70% of the observed heteroplasmic variability was explained by the random distribution of mtDNA molecules during bottlenecks, suggesting that there are additional factors responsible for the remaining 30% (Cree et al., 2008b).

In humans, it is established that genetic bottlenecks take place during the development of female germline cells (oogenesis) (Floros et al., 2018). Primordial germline cells (PGCs) undergo a period of severe reduction in the number of mitochondria, which is then followed by immense proliferation as they migrate throughout the embryo, on their way to developing into primary oocytes in the gonads (Floros et al., 2018). This is believed to mark the point when the bottleneck occurs.

Additionally, an *in vitro* study on the development of early mammalian germ cells showed that low oxygen levels were able to simulate a mtDNA bottleneck by reducing mtDNA content; the reduced cellular oxygen consumption was stimulated by the pathogenic variants which may contribute to the variation in transmission between pathogenic variants (Pezet et al., 2021).

The bottleneck effect and maternal inheritance are significant determinants of the dynamics of mitochondrial DNA heteroplasmy. Such information is crucial for understanding how mtDNA variations cause diseases, from the aspect of maternal inheritance (Zhang, Burr & Chinnery, 2018). Particularly that there is evidence suggesting that segregation rates (bottlenecks), of the same, pathogenic variants are affected by common mtDNA variants, which explains the observed geographic as well as inter-familial differences (Zhang, Burr & Chinnery, 2018).

The non-coding control region of the mitochondrial genome harbours LSP and HSP which play a vital role in mtDNA transcription, and it is believed to be the reason why deleterious variation in this region of the mtDNA are rarely inherited (Wei et al., 2020). It was thought that mature oocytes that underwent bottlenecks of the same mutation, would have similar heteroplasmy levels however, a study carried out by Pallotti et al., (2014) looked at two Italian families carriers for m.3243A>G mutation, in the first family, the mother transmitted intermediate, largely distributed pathogenic variant levels ranging from 10% to 75%; whereas in the second family, the pattern was much more skewed where one offspring had a pathogenic variant level of 81% and the four of his siblings had 0%. Indicating that variant segregation at bottlenecks in not random, and a range of 'mutant loads' can be obtained from the same pathogenic variant, suggesting that bottlenecks may indeed be under selective pressures potentially exerted by the nuclear DNA.

Moreover, an investigation of mother-child pairs reported a dichotomous selection pattern for the m.3243A>G variant; children with high variant level have a statistically significant descending pattern when looking back at the mothers level (evidence for positive selection), and vice versa for children with low levels (suggesting negative selection) (Franco et al., 2022), however, the mechanistic explanation for this observation is unknown. Considering the severe biochemical effect of this mutation, which is rarely seen at extremely high heteroplasmy, a selection in favour seems to be disadvantageous, a potential reasoning for this might be a compensatory reaction; where in response to the mutation led decreased OXPHOS efficiency, mitochondria try to increase their mtDNA content (Khrapko and Turnbull, 2014). Thus far, several studies have reported the presence of selection in the inheritance of pathogenic mtDNA variants however, inheritance still has a random genetic drift component. Something that was confirmed in

a study looking at different pathogenic variant carrier oocytes, zygotes, and blastomeres retrieved from patients going through PGD and IVF (Otten et al., 2018). Where pathogenic variants with significantly less severe biochemical consequences, such as m.8993T>G, and m.14478T>C, in contrast to m.3243A>G, show no selection against high mutation load, and their inheritance is thought to be predominately led by randomness.

Franco and colleagues suggest that on the population level, a positive selection for high frequency nascent mtDNA variants, and a negative selection for low frequency variants would have provided a protective mechanism, preventing the accumulation of these debilitating variants; by removing the low frequency variants, and uplifting the high frequency portion to a level that deems the embryo unfit for further development (Franco et al., 2022a).

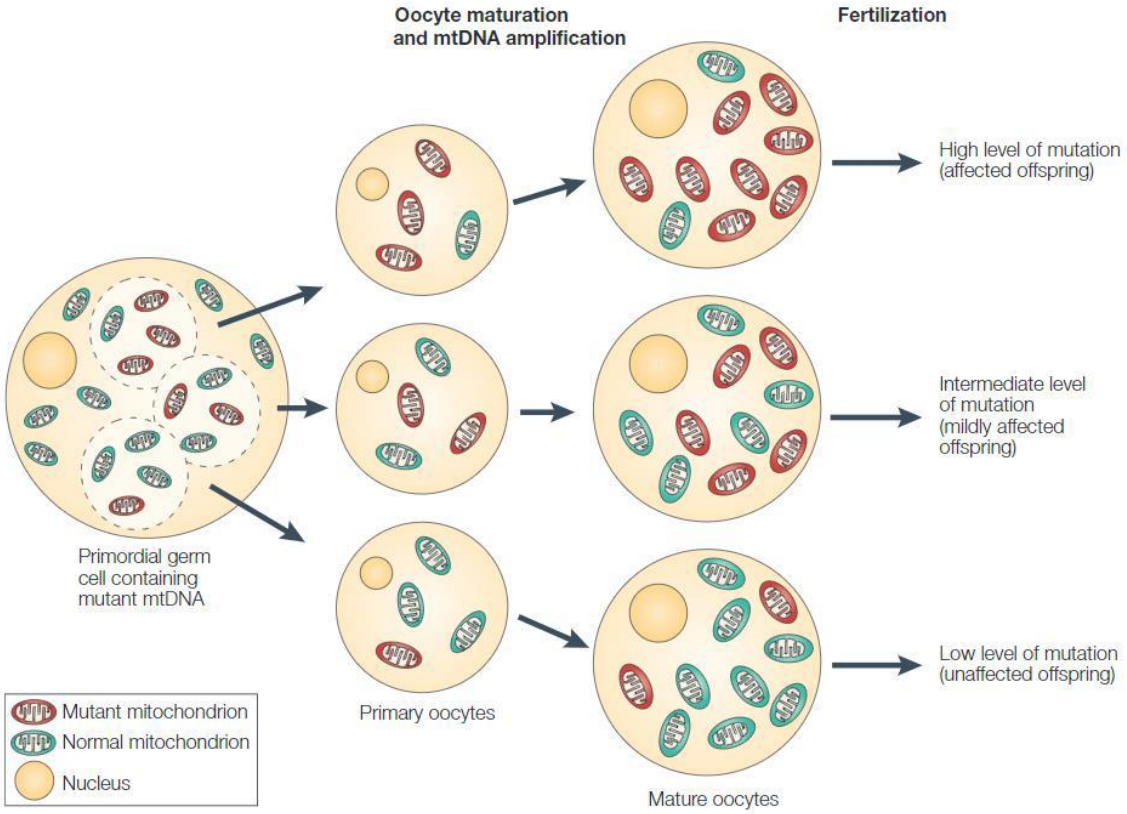

Figure 1.11: A genetic bottleneck affects mtDNA segregation. The genetic bottlenecks and is one of the explanations for highly variable levels of heteroplasmy seen in mature oocytes. As the PGC cells proliferate and develop into oogonia (primary oocytes) they first undergo a reduction in mitochondrial numbers, and it is estimated that this reduction accounts for 70% of the variability in heteroplasmy seen in offspring (Wei et al., 2020). The outcome of PGC proliferation and this bottleneck is mature oocytes with a wide range of pathogenic variant levels; when fertilised they can develop into embryos with an average heteroplasmy that are either higher, lower, or intermediate when compared to the mother. [Figure taken from Taylor and Turnbull, (2005)].

1.2.4 Inherited Mitochondrial DNA Variants

1.2.4.1 Inherited Non-Pathogenic Variants

Combinations of mtDNA SNPs that are conserved within populations and are passed on from a common maternal ancestor, are termed haplogroups. The relationship between haplogroups is represented by a phylogenetic tree, which has a theoretical origin at 'mitochondrial Eve', who is estimated to have lived more than ~200,000 years ago in Africa **(Figure 1.12)** (Cann, Stoneking and Wilson, 1987; van Oven and Kayser, 2009a). Haplogroups that represent the African ancestry have the greatest sequence variation, which supports the hypothesis that all modern humans had an African common ancestor (Chen et al., 1995).

Migration from Africa, and the geographical isolation of populations gave rise to two large branches from the African haplogroup L3; haplogroups M and N, which together encompass all of the modern non-L haplogroups (Wilson and Cann, 1992). Haplogroup R is the root of all European haplogroups, with haplogroup H being the most common (Richards et al., 2002). Understandably, haplogroups are often used to study human evolution, ancestry, migration patterns, and disease (Merriwether et al., 1991; Taylor and Turnbull, 2005; Guha et al., 2013).

Although many of the SNPs that define haplogroups are likely to have little functional consequence, there is evidence to suggest that subtle functional differences between haplogroups may exist. This is thought to be a way that enabled the mitochondria to adapt to the bioenergetic needs of the populations in their new environments. For example, the macrohaplogroup N emerged with two amino acid variants: ND3 gene variant m.10389G>A and ATP6 nucleotide m.8701G>A. These changes influence mitochondrial membrane potential and calcium regulation, potentially improving coupling efficiency in colder climates (Kazuno et al., 2006; Ruiz-Pesini and Wallace, 2006; Wallace, 2015). On the other hand, the European haplogroup J, derived from macrohaplogroup N, was formed by reversing the ND3 m.10389G>A variant and gaining a new ND5 m.13708G>A variant (Ruiz-Pesini and Wallace, 2006). Wei et al., (2017b) used sequence diversity estimates on 30,506 individuals and concluded that pathogenic mtDNA variants are more common on more recent mitochondrial subclades, compared to older, macro haplogroups; which confirms the evolutionary, protective selection against low frequency nascent mtDNA variants mentioned in **Section 1.2.3.4** (Franco et al., 2022a).

The correlation between haplogroups and various common diseases has been discussed in literature (Taylor and Turnbull, 2005; Hudson et al., 2007; 2013; 2014; Horan, Gemmell and Wolff, 2013), some haplogroups seem to play a protective role in disease whereas others increase disease susceptibility.

Such findings highlight the complex dynamics between inherited mtDNA variants, nuclear genetic background, as well as mtDNA sequence variation in shaping the risk and expression of diseases (Horan, Gemmell and Wolff, 2013). Something that will be outlined in more detail in **Chapter Six**.

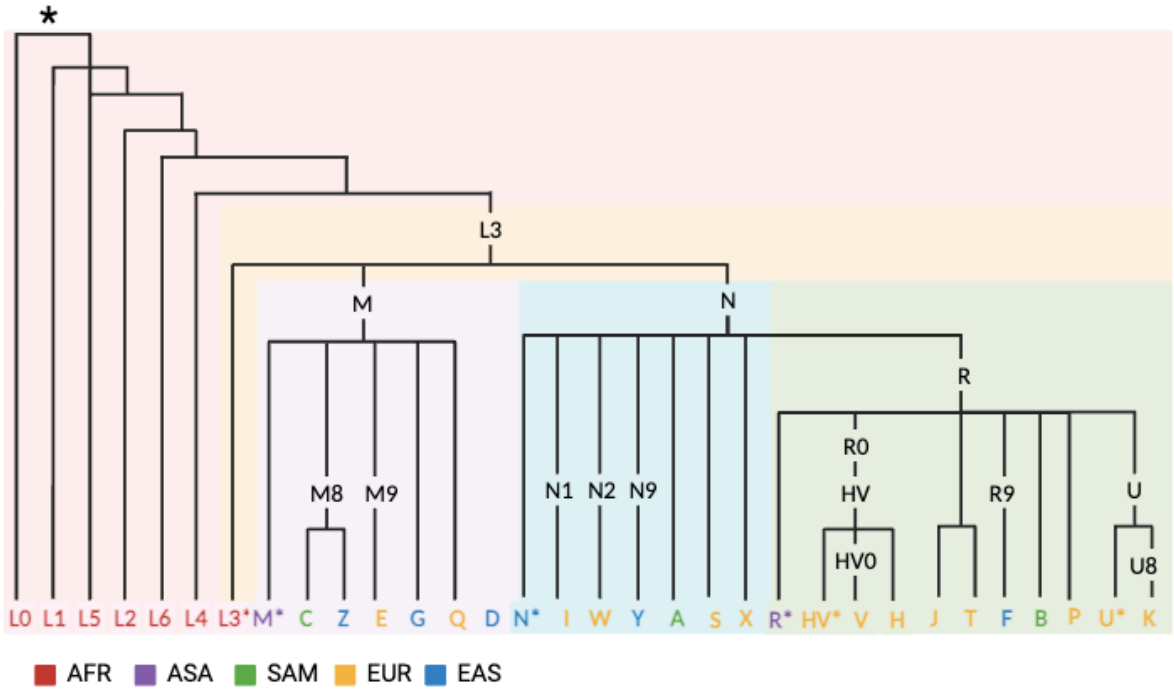

Figure 1.12: Simplified phylogenetic tree. All letters of the alphabet except O are used. The star reflects the root of the tree, the African origin of all modern humans descending from the ancestral Eve. Haplogroup L3 represents the oldest haplogroup that is closest to mitochondrial Eve. Haplogroups M, and N are the haplogroups that emerged out of Africa and that encompass all modern haplogroups. Colours represent the continental distribution of these haplogroups where: AFR = Africa (red), ASA = Southern Asia (purple), SAM = South America (green), EUR = Europe (yellow), and EAS = East Asia (blue). [Figure recreated using Biorender.com from van Oven and Kayser, (2009), with information from Palanichamy et al., (2004); and Takeda et al., (2023)].

1.2.4.2 Pathogenic mtDNA variants

Pathogenic alleles are found in more than 1 in 200 live births, and they arise *de novo* in at least 1 in every 1000 births (Elliott et al., 2008). More than 300 pathogenic mtDNA variants have been identified to date (Li et al., 2019). Pathogenic homoplasmic variants

are difficult to classify and have been generally poorly investigated due to their high population frequency, which makes it difficult to differentiate them from benign haplogroup markers (McFarland et al., 2002). The non-coding displacement D-loop is a mutation hotspot, compared to coding regions such as rRNA and tRNA genes which are 'mutational deserts'; possibly due to their critical roles in mtDNA translation (Elliott et al., 2008; Stewart and Chinnery, 2021c).

The single point mtDNA variant m.11778G>A, associated with LHON disease (Wallace et al., 1988), and m.8344A>G associated with myoclonic epilepsy and ragged red-fibres (MERRF) (Shoffner et al., 1990b), were the first few pathogenic mtDNA variants to be ever reported following the sequencing of the mitochondrial genome in 1981 (Anderson et al., 1981b).

1.2.4.2.1 The pathogenic m.3243A>G variant

The **m.3243A>G** variant (UUR) is the most common heteroplasmic, pathogenic mtDNA variant (Goto, Nonaka and Horai, 1990). It resides in the *MTTL1* gene, encoding one of the mitochondrial tRNA-Leucine genes, specifically mt-tRNA$^{Leu(UUR)}$ (**Figure 1.13**). The m.3243A>G variant is associated with a range of clinical phenotypes, from diabetes to deafness as well as syndromes such as mitochondrial encephalomyopathy lactic acidosis and stroke-like episodes (MELAS) (Whittaker et al., 2009, Mancuso et al., 2014, Nesbitt et al., 2013).

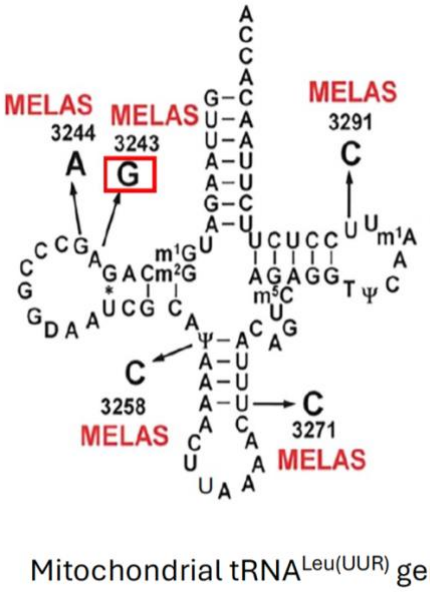

Figure 1.13: Secondary structure of mt-tRNA$^{Leu(UUR)}$. Figure depicts the location of MELAS associated variants on the secondary structure of mt-tRNA$^{leu(UUR)}$, with m.3243A>G marked with a red box. [Obtained from (Tetsuka et al., 2021)].

Due to its high frequency, and extremely wide variety of associated clinical features (see **Section 1.3.2** below), the m.3243A>G mutation is an extensively investigated pathogenic variant in mtDNA, particularly due to its yet elusive molecular pathology (Nesbitt and McFarland, 2011).

The pathogenic mechanism of m.3243A>G can be described based on three main hypotheses: (1) impaired transcription of the mitochondrially encoded 16S rRNA, (2) decreased translational efficiency, and (3) reduced aminoacylation levels of the $tRNA^{Leu\ (UUR)}$. Importantly, all three would significantly contribute to the development of mitochondrial dysfunction, lack of energy production, and, consequently, elevated oxidative stress (Blakely *et al*., 2013). Thus, the m.3243A>G mutation is thought to impact a variety of aspects of cellular metabolism and functioning. Below are the core models published to date.

First, researchers proposed that the m.3243A>G mutation disrupts the transcription termination of mitoribosomal subunit (16S) rRNA. Using cell culture, mitochondria with m.3243A>G mutation were found to have an impaired transcription termination due to a disrupted binding efficiency of transcription termination factors like mTERF1, resulting in reduced termination efficiency. This suggested the impaired ability of mitochondria to produce the necessary rRNA needed for translation (Hess et al., 1991).

Further studies using cybrid (ρ0) cells fused with m.3243A>G-carrying mitochondria identified a novel RNA transcript (RNA19), which supported the previous hypothesis suggesting that m.3243A>G mutation disrupts transcription termination. Cells with elevated levels of m.3243A>G presented with a perturbed OXPHOS function, a decreased level of mitochondrially translated proteins, and an increase in RNA19 (King et al., 1992). RNA19 transcripts were subsequently found to be a component of the mitoribosomes, affecting translation and contributing to disease (Schon et al., 1992).

In the same year, Chomyn and colleagues reported that they did not find evidence for the mentioned defect in transcription termination or the presence of RNA19. Nonetheless, they noted a reduced efficiency in mTERF's binding to MT-TL1 during the transcription termination process, supporting the finding of Hess et al., (1991). Flierl et al. (1997) found no evidence of RNA19 or transcription termination defects either, instead they reported that mitochondrially encoded proteins from cells harbouring the m.3243A>G mutation

lacked leucine. M.3243A>G variant was also found to reduce the efficacy of the aminoacylation of mt-tRNA$^{Leu\ (UUR),}$ which decreases its stability leading to a perturbed protein translation (Park, 2003).

It was hypothesized that the instability of the mitochondrially encoded proteins in m.3243A>G carrier cells is due to mis-incorporated amino acids (Janssen, Maassen and van den Ouweland, 1999; Yasukawa, 2001). Where the mutant mt-tRNA$^{Leu\ (UUR)}$ not only recognizes UUR (R = A or G) leucine codons but also UUY (Y = C or U) phenylalanine codons. This altered translation is believed to cause leucine to be incorporated into positions typically occupied by phenylalanine (Yasukawa, 2001), reducing the stability of mitochondrial proteins and increasing their susceptibility to degradation (Janssen, Maassen and van den Ouweland, 1999).

Some studies, such as Janssen et al. (2007) did not find evidence for m.3243A>G variant led mis-incorporation of amino acids. The disruption in UUR decoding rather suggests a loss-of-function of mt-tRNA$^{Leu\ (UUR)}$, whereas the mis-incorporation of amino acids suggests a gain-of-function. Sasarman et al. (2008) proposed that both loss- and gain-of-function mechanisms together contribute to the pathogenic effects of the m.3243A>G variant.

Most studies mentioned have employed the unstable *trans*mitochondrial cybrid cells (discussed in **Section** 1.**2.3.1**) (King and Attardi, 1989b). These cells are typically created using ethidium bromide (EtBr) treated thymidine kinase (TK) inactive osteosarcoma cells (43TK- cells). Treatment with EtBr exhausts their mtDNA, resulting in what are known as ρ0 cells. ρ0 cells are then reintroduced with external mitochondria to alter the mtDNA population within the cells. However, EtBr is also mutagenic to nuclear DNA, and so differences in downstream function may not be wholly attributed to the re-introduced mtDNA. *Trans*mitochondrial cybrid cells, like other immortalized cell lines, often depend on glycolysis for ATP production, which can result in a decreased response to OXPHOS inefficiency (Inak et al., 2021). These factors complicate the comparison of results from *trans*mitochondrial cybrid cell lines to mechanisms observed in vitro (Sasarman, Antonicka and Shoubridge, 2008).

As an alternative to cybrids, immortalized patient-derived myoblasts have been used, potentially offering a more accurate depiction of in vitro mechanisms. The fact they are patient-derived means that the recreation of the patient-specific nuclear and

mitochondrial genome is possible, by that offering maximum replication of *in-vitro*, patient-specific conditions (Sasarman, Antonicka and Shoubridge, 2008; Inak et al., 2021).

1.3. Mitochondrial disease

1.3.1 Clinical Manifestation of Mitochondrial Disease

Mitochondrial diseases present a wide clinical spectrum and might affect virtually any organ system at any age. One of the hallmarks of mitochondrial diseases is clinical heterogeneity, and their variable severity even in the same extended family and mutation carriers (Gorman *et al*., 2016) (**Table 1.1**). Mitochondrial diseases can be caused by mutations in either the mitochondrial or the nuclear genome.

This complexity somewhat explains why this group of diseases is hard to diagnose. The most commonly observed symptoms in adults are those associated with the central nervous system, muscle weakness and myopathies, whereas in paediatrics it is hypotonia, psychomotor delay, cardiorespiratory failure and lactic acidosis.

Mitochondrial variations have been associated with several late-onset common diseases like Parkinson's and Alzheimer's (Hutchin and Cortopassi, 1995; Hudson et al., 2013a), type 2 diabetes (Wang et al., 2001; Tang et al., 2006), and cancer (Canter et al., 2005; Wallace, 2012). Which make the identification of mtDNA variants important not only for molecular diagnosis of mitochondrial diseases but for a number of more common, complex diseases.

At last, it is worth mentioning that 50% and 80-90% of mitochondrial disease adult and paediatric patients, respectively, lack a molecular diagnosis as the genetic analyses carried out often fail to identify the causative mtDNA or even nuclear-disease-causing variants (Zeviani and Donato, 2004). The unusual nature of mitochondrial genetics, the limited methods available to manipulate mtDNA, and the lack of suitable disease models have stood in the way and hindered the ability to find a target for therapy or even the prevention of disease progression if symptoms are identified early on (Tuppen et al., 2010).This however, is a fast-moving field with novel methods constantly emerging and holding a great promise (Gammage et al., 2014; Bacman et al., 2013; Mok et al., 2020; Silva-Pinheiro and Minczuk, 2022).

Table 1.1: The clinical phenotypes most frequently observed in mitochondrial diseases. [Table reproduced from Zeviani and Donato, (2004)]

Neurological manifestations	Systemic manifestations
Neuromuscular	Heart
Ophthalmoplegia	Cardiomyopathy
Myopathy	Cardiac conduction defects
Exercise intolerance	
Peripheral sensory-motor	Endocrine system
Neuropathy	Diabetes
Central nervous system (CNS)	Exocrine pancreas dysfunction
Myelopathy	Hypoparathyroidism
Headache	Multiple endocrinopathy
Stroke	Short stature
Seizures	
Dementia	Blood
	Pancytopenia
Movement disorders	Sideroblastic anaemia
Ataxia	
Dystonia	Mesenchymal organs
Parkinsonism	Hepatopathy
Myoclonus	Nephropathy
	Intestinal pseudo-obstruction
Eye	
Blindness	Metabolism
Optic neuropathy	Metabolic acidosis
Pigmentary retinopathy	Nausea and vomiting
Cataract	
Ear	
Sensorineural deafness	

1.3.2 m.3243A>G-Related disease

The population frequency of m.3243A>G is 140-250 per 100,000 however, disease prevalence is ~ 1 in 5000, which reflects the asymptomatic portion of m.3243A>G carriers in the population and can be attributed to their low variant levels (Nesbitt and McFarland, 2011). The m.3243A>G variant is associated with a spectrum of clinical manifestations, among which are mitochondrial encephalopathy with lactic acidosis and stroke-like episodes (MELAS) syndrome, progressive external ophthalmoplegia (PEO), diabetes, and deafness. m.3243A>G affects multiple organs and systems, which results in a wide array of symptoms, such as muscle weakness, neurological impairments, or endocrine dysfunctions (Nesbitt et al., 2013). Traditionally, the spectrum of diseases associated with the m.3243A>G variant falls into two categories – syndromic and non-syndromic presentations – as outlined in Mancuso et al. (2014). Given the vast clinical heterogeneity seen in patients with m.3243A>G for example, syndromic diagnosis is not always possible and thus phenotypic descriptions are used instead. For example, in some

people, it may first manifest as diabetes and deafness simultaneously (van den Ouweland et al., 1992), with stroke-like episodes following shortly thereafter. PEO, which is an ophthalmological term denoting progressive weakness of eye muscles and ptosis, as well as other related symptoms can also arise (Moraes et al., 1993).

Notably, elevated mutation load in different tissues, such as blood, urine, and muscle do not always correlate with the diversity and severity of symptoms. Individuals with high variant levels may sometimes be relatively asymptomatic, which highlights the phenotypic heterogeneity of m.3243A>G, something that remains to be poorly understood (Grady et al., 2018). Grady et al., (2018) estimated that age and age-corrected blood 3243A>G levels account for only ~25% of the observed phenotypic variability.

Studies suggest that nuclear polymorphic background affects the phenotypic expression of mtDNA variants, which highlights the importance of studying mito-nuclear interactions. For example, in LHON, even though all children inherit the homoplasmic point mutation, only 10% of the females and 50% of males develop blindness, suggesting the influence of external factors such as nDNA variation (Taylor and Turnbull, 2005; Pickett et al., 2018). Supporting this, linkage was identified between X chromosome haplotype and mutations in the mitochondrial *MTND* gene, by that causing the visual impairment in LHON and explaining the observed sex bias however, the exact nDNA variants are yet to be identified (Hudson et al., 2005;Carelli et al., 2016).

1.3.3 Mitochondrial disease diagnosis and treatment

Diagnosing mitochondrial diseases is challenging because of the wide range of clinical symptoms and the complicated genetic composition of the mitochondria (Parikh *et al*., 2009). Typically, the combination of the clinical assessment of the state of the patient, biochemical tests, and molecular genetic testing is used to detect the signs of mitochondrial dysfunction (Thompson et al., 2023). Current treatments only alleviate disease symptoms. As a measure to prevent the transmission of mitochondrial disease, mitochondrial replacement therapies, or mitochondrial donation was developed (Craven et al., 2010; Tachibana et al., 2013). First approved by the UK parliament in 2015 (Kmietowicz, 2015a), then licenced by the Human UK Fertilization and Embryo Authority (HFEA) in 2017 (Kmietowicz, 2015b; Craven et al., 2018); this technique involves the transfer of the nuclear genome from an oocyte (or zygote) into the donors' enucleated

egg cell (or zygote) that harbours healthy mitochondrial populations (Tachibana et al., 2009). Nevertheless, the incomplete understanding of mitochondrial biology impedes the achievement of full recovery from mitochondrial diseases, particularly those caused by the m.3243A>G variant due to the multifactorial nature of disease progression and not fully uncovered role of nuclear DNA (Chinnery et al., 2014; Lightowlers, Taylor and Turnbull, 2015).

1.4 Nuclear DNA and mtDNA crosstalk in mitochondrial function and disease

The role of functional synergy between nucDNA and mtDNA is paramount in cellular energy production, particularly given that 99% of the genes essential to mitochondria are encoded by the nuclear genome, including subunits of a whole respiratory complex (CII) (Calvo, Clauser and Mootha, 2016; Rath et al., 2021b). Processes in the mitochondria, such as translation, are also governed by the two genomes; and thus, the organelle's functional integrity depends on efficient communication between the two via both anterograde (nucleus to mitochondria) and retrograde (mitochondria to nucleus) signalling. Disruption of this fine network leads to mitochondrial dysfunction and eventually disease expression (Horan and Cooper, 2014).

Embryos derived from mitochondrial donation experiments on embryonic stem cell lines reported mtDNA heteroplasmy reversion (Kang et al., 2016; Hudson, Takeda and Herbert, 2019). These recurrences were reported to happen at a rate of 15%, and an explanation for this might be that the nuclear genetic background is "favouring" certain mtDNA variants which would provide further evidence for nuclear-mitochondrial genetic interplay (Wei and Chinnery, 2020). nucDNA can affect mtDNA translation throughout embryo development by exerting selective forces, as well as throughout an individual's lifetime (Wei et al., 2019b). Importantly, a recent GWAS study by a team from 23andMe®, identified 20 nuclear loci associated with non-pathogenic mtDNA heteroplasmy, accounting for 20% of the observed heritability (Nandakumar et al., 2021). The identified loci were surrounding *TFAM* and *TWNK* genes vital for mtDNA replication, and others associated with mitochondrial funcition and quality control (*CLEC16A*, *PRKAB1)*. This raises the potential that nDNA variability can regulate mtDNA heteroplasmy by modifying the mitochondrial replication capacity, once again highlighting that mitochondrial function involves significant genetic interdependence between the two genomes.

Another illustrative example of this cooperative adaptation is seen in Saccharomyces cerevisiae systems, where mtDNA from different ecological niches can lead to various phenotypes depending on the nuclear genetic background (Nguyen *et al*., 2020). This highlights a selective pressure for optimised mito-nuclear interactions, with profound implications for evolutionary trajectories and adaptation processes.

1.4.1 Nuclear-Mitochondrial DNA crosstalk

Mitochondrial maintenance disorders exemplify the complex outcomes arising from disruptions in the interactions between nucDNA and mtDNA (DiMauro et al., 2013; Viscomi and Zeviani, 2017). These disorders highlight the crucial role of genetic crosstalk in maintaining mitochondrial function and integrity (Bonnen *et al*., 2013). Exchange of mtDNAs between different yeast strains was found to not only impact growth rates also demonstrated that strains with native mtDNA configurations are fitter than those with altered mito-nuclear combinations (Lehtonen et a, 2016; El-Hattab, Craigen & Scaglia¸2017). This fitness disparity underscores the influence of natural mito-nuclear interactions on evolutionary fitness, particularly in response to environmental changes. For example, RNA differential expression analyses on different cancer cell types revealed a completely broken association between nuclear proteins and the mitochondrial P9 site methylation – which is a post transcriptional processing that yields a mature, functional tRNA (Idaghdour and Hodgkinson, 2017). This was found to be nucDNA genotype specific, where nuclear mutations in KIRC cohort (kidney and renal clear cell carcinoma) were identified as significant patient survival predictions.

A study by Bellizzi et al. (2012) has shown a correlation between the methylation of nDNA and mtDNA haplogroups, where nuclear *trans*mitochondrial cybrid cells harbouring the J haplogroup had increased levels of nDNA methylation compared to other haplogroups. This is believed to be initiated by the increase of reactive oxygen species as a result of mitochondrial malfunction. This proves that mtDNA can also modulate the nucDNA in a retrograde manner through epigenetics (Horan, Gemmell and Wolff, 2013b). Such findings further emphasise the deep interdependency between the nuclear and mitochondrial genomes.

Furthermore, in cases of impaired mitochondrial function, such as in cancer cells, depletion of mtDNA can lead to significant alterations in nuclear gene expression, impacting pathways involved in energy metabolism, cell signalling, and apoptosis (Di Nottia *et al*., 2021).

1.5. Genetic Tools to Investigate Complex Diseases

1.5.1 Prior Work

Almost all diseases have a genetic component. Some, such as cystic fibrosis or sickle cell anaemia, are caused by mutations in a single, well-known gene (monogenic diseases). On the other hand, ~90-95% of diseases, including some mitochondrial, are polygenic and are also influenced by factors like lifestyle and environment; these are referred to as complex diseases. Genetic assessments such as linkage analysis, association studies, and heritability estimates have been instrumental in demystifying various Mendelian as well as complex diseases (Caspi *et al*., 2010). Typically, rare variants with a large effect size are associated with Mendelian disorders, whereas common variants with small effect sizes are involved in common, polygenic complex diseases, each method is powerful at detecting variants with a certain frequency/ effect size (**Figure 1.14**).

The impetus for studying the m.3243A>G variant heteroplasmy and its nuclear modifiers is largely built on two pivotal findings: firstly, using variance components to estimate heritability, Pickett et al., (2019) estimated that 72% of m.3243A>G variant allele variability can be attributed to additive, nuclear genetic factors, highlighting the importance of exploring the role of nuclear genetic variation in influencing m.3243A>G levels. Subsequently, Boggan et al., (2022) identified a locus on chromosome 7q22 that is linked to m.3243A>G-related mitochondrial encephalopathy, further highlighting the significance of nuclear genetic factors in mitochondrial disorders.

In this section I will outline the tools used in this thesis, mentioning their strengths and drawbacks with a focus on genome wide association studies (GWAS) as the main analysis approach.

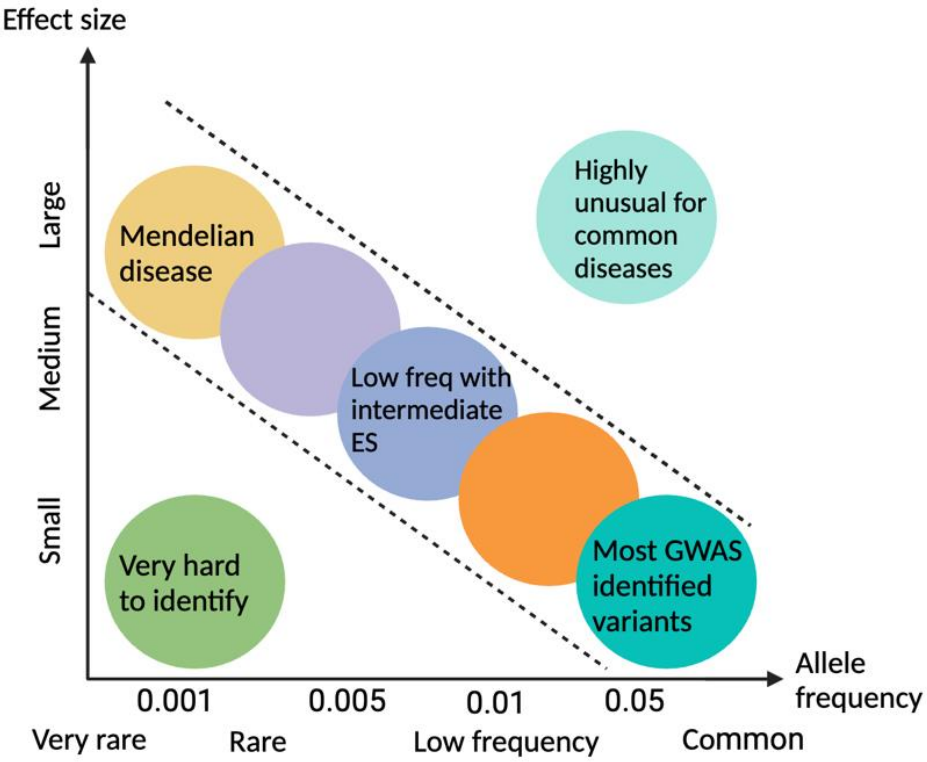

Figure 1.14: Relationship between variant frequency and effect size (ES). Area between the dashed lines is where the ongoing research is focused. [Figure recreated from (Manolio et al., 2009a; Hertel et al., 2013)].

1.5.2 Heritability Studies

Heritability studies provide crucial insights into how much of the variation in disease phenotypes and/or variant levels, can be attributed to genetic factors versus environmental influences (Farrar *et al.*, 2013). There are two types of heritability studies: broad sense heritability, and narrow sense heritability. The former estimates the percentage of genetic contribution to the phenotypic variability in pedigrees and family data, whilst including factors with dominant, additive, as well as epistatic effects. The latter is typically conducted on populations and focuses on common genetic factors with only an additive effect. Some methods sum the proportion of heritability contribution and effect sizes of only significant GWAS SNPs, called GWAS heritability however, analyses that consider the contribution of all measured, additive genomic regions, referred to as SNP heritability, provide a more biologically accurate reflection (Yang et al., 2010a; Matthews and Turkheimer, 2022). Despite utilising all SNPs for estimating SNP heritability, the disparity between that and estimates of broad sense heritability is still vast for many phenotypes. This disparity has often been referred to as the issue of 'missing heritability'

(Maher, 2008), where, the broad sense heritability on family data yields a greater percentage of variation attributed to genetics, compared to SNP-based, or narrow sense estimations. There are multiple explanations for this, the missed variants in GWAS analyses due to low frequency and miniscule effect size, which means they end up being ungenotyped, is the most prominent reasoning (Zuk et al., 2014; MN et al., 2021; Matthews and Turkheimer, 2022). There have been studies suggesting that broad sense heritability has overinflated values, and that is because it considers all possible non-linear or non-additive factors (mentioned previously), something that the current designs of molecular heritability studies cannot detangle (Manolio et al., 2009b; Matthews and Turkheimer, 2022). Some researchers are convinced that these are two fundamentally different sets of analysis and that we should not even aim to have the retrieved estimates similar to one another; particularly that attempts to obtain this would require the overall restructuring of analysis methods (Longino H.E., 2013). Heritability analyses are discussed further in **Section 1.5.5**.

1.5.3 Linkage Analysis

Linkage analysis looks into identifying chromosomal regions that are shared amongst individuals with the same phenotype. The underlying premise is that, in genomic regions that contain phenotype-influencing variation, such individuals would have higher identity by descent estimates (IBD) than would be expected by chance, as parts of their genome co-segregate with the phenotype of interest (Taylor E.W. et al., 1997) (**Figure 1.15**). The likelihood of the data is calculated assuming the loci are linked or not, if LOD (estimates of multipoint logarithm of the odds), equals to 3.3, then evidence of significant linkage is said to be identified (Nyholt, 2000). Considering that linkage analyses utilise family data, population stratification is not a consideration even when multiple families are analysed – this is because each family is investigated for segregation patterns as a single entity.

Linkage analysis can be parametric and non-parametric, and this reflects whether inheritance pattern is given as an input to the analysis. Non-parametric linkage analyses are useful in cases when the inheritance pattern is unknown however, parametric analyses are generally more powerful due to the increased sensitivity and greater analysis power (Penrose, 1952; Abecasis et al., 2002; Ott, Wang and Leal, 2015a; 2015b).

Although traditionally used in the context of monogenic diseases, linkage proved to be valuable for exploring complex, polygenic diseases (Sturtevant, 1913; Pulst, 1999; Slatkin, 2008a). In the case of mitochondrial, m.3243A>G related pathologies, linkage analysis, as previously shown by Boggan *et al.*, (2022), can reveal nuclear regions harbouring variants that potentially modulate mitochondrial disease phenotypes, highlighting a complex relationship.

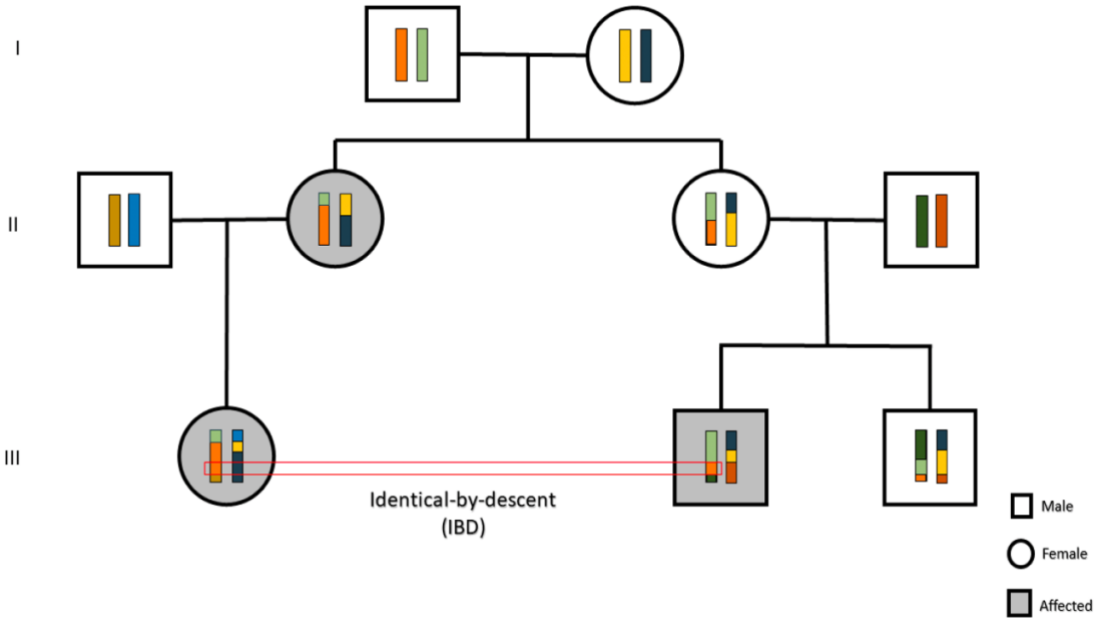

Figure 1.15: Pedigree depicting allele tracking via identitiy by descent (IBD). Coloured fragments represent chromosomal regions, it can be noticed that all affected individuals share the same fragment (orange), that they inherited from the ancestor in generation I. [Figure retrieved from (Boggan, 2022)].

1.5.4 Genome wide association studies (GWAS)

In large scale studies, typically investigating complex diseases, linkage analyses have been limited in their ability to find underlying causative variants, partly because the regions identified are typically very large and contain many potential causative genes (Hirschhorn and Daly, 2005). GWA studies test for association between a phenotype and millions of individual loci across the genome, it is a technique based on the concept of linkage disequilibrium, where it calculates the deviation from the expected random segregation of variants across a population and so has a higher resolution, narrowing genomic regions down much further (Joiret et al., 2019) (**Figure 1.16** and **Table 1.2**). GWAS became a critical tool of investigation in complex diseases through successful identification of genetic variants that increase disease susceptibility such as the 2007 study by the Wellcome Trust

Case Control Consortium, that identified variants associated with seven common diseases including type 1 and type 2 diabetes (more detail in **Chapter Four**) (Burton et al., 2007). GWAS is essential in my research for identifying specific m.3243A>G nuclear genetic modifiers of variant's allele level. By employing this technique, we can bridge the knowledge gap on how nuclear variants influence mitochondrial processes.

Unlike linkage analysis, GWA studies can be greatly confounded by population stratification (Price et al., 2010a; Hellwege et al., 2017). Instead of detecting true, causal variants, analysis aims to detect variants that are shared amongst the study population i.e. in linkage disequilibrium (LD). Population Stratification is typically accounted for using methods such as principal component analysis (PCA), which groups the population into variance explaining clusters, and by that, identifies population outliers (Patterson, Price and Reich, 2006; Price et al., 2006). Additionally, the presence of related individuals without providing records of that (cryptic relatedness), also impacts the accuracy of association analysis results (Sun and Dimitromanolakis, 2012). Mixed models have been proposed as suitable approaches in population-based studies to account for both confounding factors (Yu et al., 2006).

Given that regression analysis underlies GWASs, a rule of thumb reported by Green in 1991 is that the number of observations (individual samples) should be at least 50 plus eight multiplied by the number of predictors (SNPs) ($50 + 8(n_{predictor})$) for testing the overall model (Green, 1991). This builds up the reasoning to why GWASs often struggle with having sufficient sample sizes. For reproducible, significant GWAS results, large sample sizes are often required to ensure that the underlying regression model has enough power. Power calculations can be conducted ahead of analysis to determine the necessary sample size to reach the desired detection power (Moore, Jacobson and Fingerlin, 2019) (**Section 4.2.5**).

GWA study designs can either include cases and controls for binary traits or measure quantitative traits across the entire sample. Researchers also have the option to use either population-based or family-based designs, although there is a shift against family-based designs as they are typically highly underpowered (due to sample size and genetic diversity limitations), and were often resorted to as a way to avoid population structure, which given the emergence of methods to account for such confounding, is not a good justification (McCarthy et al., 2008; Holmes, Ala-Korpela and Smith, 2017). However, study design is also dictated by the studied disease, its inheritance patterns and penetrance,

where in some cases, family-based designs are most suited. There are numerous software packages used to conduct a GWAS; this is something that is thoroughly discussed in **Chapter Three**.

Cohort construction depends on multiple factors: 1) required sample size, 2) analysis question and 3) availability of pre-collected data. This is a step that should be carefully considered to avoid bias. For example, participants enrolled (selected) based on their clinical diagnosis from hospitals and healthcare centres, or from cohorts encompassing individuals with rare diseases such as, Genomics England; exhibit ascertainment compared to cohorts recruited from the population (such as the UK Biobank) (Uffelmann et al., 2021). If not careful, using non-random data can lead to a collider bias, which is when two variables influence a third variable and the third variable is used as a conditional, which causes spurious associations (Cole and Hernán, 2002; Uffelmann et al., 2021).

Either genotype data that is retrieved from a suitable microarray, usually followed by imputation to increase marker density, or sequencing data can be used in GWAS. However, with the decreasing cost of sequencing technologies, GWASs are increasingly relying on the robust sequencing data (Salomon et al., 2016).

Fine mapping is a post-GWAS analysis process that uses LD data along with the retrieved GWAS summary statistics and increases the resolution of variation found within any association peak. There are several sophisticated software packages that perform these analyses; in this thesis, FINEMAP was the software of choice (**Section 4.1.3**) (Benner et al., 2016a).

Upon the *in-silico* identification of an associated variant using GWA, various *in-vitro* functional analyses can step in to elucidate the biological implications of the pinpointed variant. It is worth mentioning that only 2-3% of GWAS fine mapped variants fall within coding genes (Visscher et al., 2017a). The remaining portion fall outside coding regions. One method for identifying target genes of genetic variants involves molecular quantitative trait loci (molQTL) analysis. This technique links genetic variants to specific molecular functions; for instance, eQTL analysis connects loci to RNA expression levels and there are several other techniques that can link variants to other molecular functions such as, pQTL (loci associated with protein abundance) (Chick et al., 2016), meQTL (loci associated with methylation levels) (Mulder et al., 2021), chQTL (loci associated with

chromatin accessibility, which reflects gene regulation) (Keele et al., 2020) (Li et al., 2016; Barbeira et al., 2021; Uffelmann et al., 2021).

Each of the above-mentioned method's unique approach makes it ideal for identifying different types of genetic variations that contribute to complex diseases. GWASs are particularly effective at detecting common variants with modest effect sizes. In contrast, linkage analysis excels in pinpointing rare variants with larger effect sizes (Ott et al., 2015). Below is a comparison table of the two approaches (**Table 1.2)**.

Approximately 300,000 associations with diseases, disorders, quantitative traits, and genomic traits have been identified by GWAS (Sollis et al., 2023). One prevalent example is from type 2 diabetes (T2D); its well-defined mode of inheritance and population prevalence made it easy to collect large, extended pedigrees (Vaxillaire and Froguel, 2006). This made T2D at the forefront of diseases studied by different genetic analyses. In 2003, Reynisdottir et al., identified regions of suggestive linkage to T2D on chromosomes 5 and 10; later *TCF7L2* was identified as the causative gene on chromosome 10. As association analyses developed and emerged from candidate gene approach to more unbiased approaches, *PPARG* as well as *KCNJ11* were identified and are currently targets for anti-diabetes medications (Gloyn et al., 2003). Further examples will be discussed in **Section 4.1.1**.

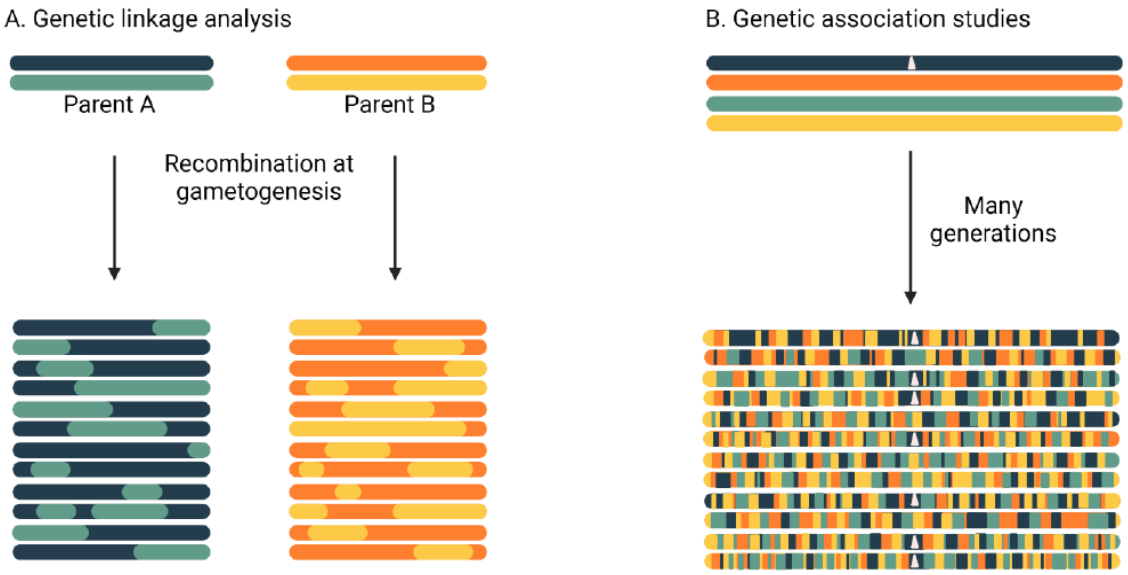

Figure 1.16: A comparison between Genetic linkage and association analysis from the aspect of resolution. During gametogenesis, recombination leads to an independent assortment of alleles, (**A**) linkage analysis relies on deviations from independent assortment observed in family pedigrees. Analysis points at large regions and this is attributed to the fact that the specific 'causal' allele has segregated over a very small number of generations. (**B**) Association analysis utilises large samples to compare phenotype sharing and co-segregation of alleles over generations that undergo multiple rounds of recombination. The white triangle in the figure shows a co-segregating allele that is carried on the blue, much smaller fragment on the population level, hence offering by that higher analysis resolution. [Obtained from Semagn, Bjørnstad & Xu (2010)].

Table 1.2: Comparison between linkage and association analysis techniques used in complex diseases. [Compiled from Joiret et al., (2019), March, (1999), Stranger, Stahl and Raj, (2011)]

	Linkage Analysis	**Association Analysis**
Methodology	Phenotype and genetic data from individuals within pedigrees are collected. Analysis identifies genomic regions that are shared among affected individuals.	Tests whether the frequency of SNPs differs between affected individuals and controls.
Data studied	Related individuals	Unrelated or related
Range of effect detected	Long (≤ 5 Mb)	Short (≤100 Kb)
Number of markers required	5 ~ 10,000 (Depending on the size of the region and number of alleles for each marker)	> 100,000
Suitable applications	Rare dominant traits, and common traits but lower resolution than GWAS	Common traits
Typical detection abilities	Rare variants with large effect size	Common variants with small effect size
Drawbacks	Disease heterogeneity decreases statistical power thus, disease penetrance should be relatively similar across the cases. Low power to detect genomic regions with small phenotypic effect.	Fails at identifying rare variants/ variants with a strong effect. Population stratification and cryptic relatedness affect reliability of results. If a marker lies at a distance from disease locus and/or locus information content is not identified, then in case sequencing data is not available, a higher density SNP array is required.
	Genotyping errors reduce power for both. Good data quality control is essential to ensure no false positives or false negatives.	

1.6. Project Rationale and Aims

The pathogenic m.3243A>G mitochondrial DNA variant is characterised by its clinical heterogeneity, (Xia *et al*., 2016; Parikh *et al*., 2015). Levels of the pathogenic m.3243A>G variant can be markedly different across individuals in the same pedigree. This heterogeneity poses a significant diagnostic and clinical challenge.

As previously outlined, genetic bottlenecks, segregation, selection, are all phenomena that, to a degree, may explain this variability. It has been established that nuclear, heritable factors explain up to 72% of observed heterogeneity of m.3243A>G (Pickett et al., 2019a). However, the precise mechanisms of this nuclear influence remain poorly understood. Therefore, an in-depth investigation into how nuclear genetic variations can modulate the levels of m.3243A>G heteroplasmy is essential. The findings from such research could lead to developing therapeutic strategies that significantly reduce the impact of the m.3243A>G mutation. Furthermore, understanding the influence of nuclear factors in the context of m.3243A>G could provide a foundational model for studying other mitochondrial pathogenic variants, extending its significance and application in the broader field of mitochondrial disorders. Ultimately, advancing our understanding in this area could lead to more accurate prognostics, enhanced therapeutic strategies, and hopefully, a better quality of life for affected individuals and families.

Hypothesis: I hypothesise that heritable genetic variants modulate the variability in m.3243A>G levels across individuals and that these may be identified using methods from complex disease genetics, such as GWAS.

Aims:

1. **To define m.3243A>G carrier and obligate carriers in population cohorts to enable GWAS analyses.** In addition to samples from the Newcastle-based, multi-centre cohort, which predominantly comprises patients with diabetes and neurological diseases selected for having m.3243A>G variant; I aim to further increase the sample size available for analyses by identifying additional 3243A>G carriers from the UK Biobank and the 100,000 genomes project (Genomics England). This is crucial for increasing analysis detection power. Additionally, it enables the comparison of m.3243A>G levels across all the three cohorts, which is essential for robust interpretation of analysis results.

2. **Evaluate the suitability of various genetic association analysis methods and refine an optimal analysis framework that suits all cohorts.** Considering the differences in pedigree and population structures across cohorts along with the substantial variation in the distribution of m.3243A>G levels among individuals, selecting an appropriate method is crucial. This objective involves conducting analyses using different statistical approaches to determine the most reliable method that effectively incorporates covariates, thereby minimising the risk of errors such as type I errors.
3. **To identify nuclear variants that are associated with detectable m.3243A>G levels.** Through GWA analysis, I aim to identify nuclear variants that modulate m.3243A>G levels in individuals. This objective is key for advancing our understanding about the genetic architecture underlying variant heteroplasmy, which is important for understanding differential disease susceptibility and severity in carriers of m.3243A>G.
4. **Investigate the role of mtDNA variation in modulating m.3243A>G heteroplasmy.** Using mitochondrial GWAS (miWAS), I will investigate whether sequence variation in the mitochondrial genome is associated with m.3243A>G levels and whether m.3243A>G is more commonly seen on different haplogroup backgrounds.

Chapter 2. Materials and methods

2.1. Cohort structures

2.1.1 Multicentre cohort

This patient cohort contains data from 488 individuals carrying m.3243A>G, and includes samples collected from The UK Mitochondrial Disease Patient Cohort – Newcastle, University College London (UCL), Exeter Centre of Excellence for Diabetes Research, The German Network for Mitochondrial Diseases, and The Nationwide Italian Collaborative Network of Mitochondrial Diseases. The number of samples, along with the median age at last clinical assessment, are summarised in **Table 2.1**. Samples were known for having m.3243A>G however, to determine variant allele levels, m.3243A>G pyrosequencing was performed by Dr Roisin Boggan (**Section 2.2.1**). Of these samples, 445 individuals were SNP genotyped with the remaining excluded for having poor quality, or little DNA. Quality control (QC) steps on genotyped data were then performed (see **Section 2.4.1** for more detail), 408 samples remained and were taken forwards for this project.

The Newcastle and North Tyneside Research Ethics Committee (13/NE/0326) provided ethical approval for 258 samples, and patients provided written informed consent prior to their inclusion. Tissue samples from these patients were obtained with ethical permission from the Newcastle Mitochondrial Research Biobank (REC reference 16/NE/0267- Application Ref: MRBOC ID 016). Additionally, ethical approval for the inclusion of 54 patient tissue samples received from UCL was provided by the Queen Square Research Ethics Committee, London, UK (09/H0716/76). A total of 110 individuals were referred for genetic testing from routine clinical care to the Exeter Genomics Laboratory at the Royal Devon and Exeter Hospital, and the study was approved by the North Wales ethics committee (17/WA/0327). Samples from 56 individuals were obtained from the German network for mitochondrial disorders "mitoNET", with funding from the German Ministry of Education and Research (01GM1906A, 01GM1906B). Ethical approval for the clinical Registry (mitoREGISTRY) was obtained from the Ethics Committee of the LMU Munich (182-09), and approval for the Biobank (mitoSAMPLE) was secured from the Ethics committee of the Technical University Munich (200/15 S-SR). Furthermore, ten individuals from the University of Pisa, enrolled in the "Nationwide Italian Collaborative Network of Mitochondrial Diseases", provided written consent for their inclusion in the study.

Table 2.1: Summary of data included in the analyses. Table includes summary of data in all three cohorts used in this project: the multicentre cohort (rows two to six), 100k genomes project (Genomics England), and the UK biobank.

Sub cohort	**N (M, F)**	**Median age at first assessment (IQR)**	**Median age-adjusted m.3243A>G variant levels (IQR, range)**
UK Mitochondrial Disease Patient Cohort	258 (105, 153)	42.8 (24.3)	0.648 (0.556, 0.002-1.000)
University College London	54 (22, 32)	44 (24)	0.938 (0.394, 0.074-1.000)
Exeter Centre of Excellence for Diabetes Research	110 (34, 76)	40 (17)	0.816 (0.324, 0.002-1.000)
German Network for Mitochondrial Disease	56 (25, 31)	40.5 (20.5)	0.637 (0.465, 0.058-1.000)
The Nationwide Italian Collaborative Network of Mitochondrial Diseases	10 (6, 4)	45 (9.75)	0.696 (0.169, 0.490-1.000)
Genomics England (100,000 Genomes Project)	176 (81,95)	35 (33.5)	0.078 (0.222,0.016-1.000)
UKBB	147 (72,75)	56.5 (14.5)	0.23 (0.27, 0.033-1.000)

2.1.2 Genomics England (100kGP)

Through application number (RR97), access to 61,140 tiered and quality controlled rare disease genome datasets (release V12) were available. Data belong to individuals recruited for having rare, possibly hereditary disease symptoms, with or without a molecular diagnosis, along with their family members. To identify m.3243A>G carrier samples, Dr Dasha Deen (Bioinformatician) designed a mtDNA heteroplasmy calling pipeline (**Section 2.2.2**). This identified 116 individuals with age corrected m.3243A>G levels ≥ 1%. Using family IDs, the data were extracted from their relatives (134 individuals) which were subsequently used in family tracing (**Section 2.2.3**). From a group of 134

relatives, 60 obligate carrier samples were identified, these are relatives who are not necessarily clinically affected, but based on their position within the family are carrying the m.3243A>G variant at low levels. The total number of m.3243A>G samples was 176, with a median age of 35 (IQR = 33.5) (**Table 2.1**).

2.1.3 UKBB

Data in the UKBB belong to individuals recruited from the UK, aged between 40 and 69, who have provided a set of health-related data via questionnaires, physical measurements and - most importantly - a variety of samples including blood.

Using application number 9072, our collaborators Dr Stuart Cannon and Dr Kashyap Patel (Exeter University), applied an in-house pipeline to identify m.3243A>G samples within the data available at that time (200,000 participants in the UKBB in May 2023). This led to the identification of 144 samples with variant age-corrected m.3243A>G levels ≥ 1% and using family tracing, 3 additional obligate carriers were identified. The total number of retrieved samples was 147, with a median age of 56.5 (IQR = 14.5) (**Table 2.1**).

2.2. Methods of estimating m.3243A>G levels

2.2.1 Pyrosequencing

Pyrosequencing is a sequencing by synthesis technique, with an upper limit of 400 bases, often used in molecular diagnostic settings. Compared to Sanger sequencing, it is much faster and more cost efficient, providing a quantitative analysis that permits the accurate determination of fraction analysis, or in this case, the percentage of mtDNA variant heteroplasmy (Fernandes and Zhang, 2014).

PCR amplified strands are attached to beads and immobilised into wells, once the sequencing primer is added, along with the appropriate annealing buffer, DNA synthesis is initiated. Upon the addition of dNTPs, inorganic pyrophosphatase is released (PIP_i); which by turn undergoes reactions with ATP sulfurylase as well as luciferase enzymes. This emits a bright light that is detected by sensors. The immobilisation of strands allows the purification and washing each time a new dNTP is added (Harrington et al., 2013). The intensity of light produced is proportional to the amount of emitted PIP_i, which reflects the number of identical bases that were added to the reaction. This is quantified by the height of peaks in the resultant pyro-grams (**Figure 2.1**). It is important to note that our reaction was designed to happen on the reverse strand, i.e. if it is an individual who

carries wild type variant at 3243 position, then dATPs will be aligning to SNP T at 3243, if it is an individual carrying the mutation, then light will be detected upon the addition of dGTPs.

Using PyroMark® Q24 system from Qiagen (mutant mtDNA test sensitivity > 3%), Dr Boggan carried out the pre-pyrosequencing PCR as well as the pyrosequencing reaction and determined m.3243A>G variant levels on samples from University College London, the German Network for Mitochondrial Diseases, the Nationwide Italian Collaborative Network of Mitochondrial Diseases, and the Exeter Centre of Excellence for Diabetes Research (**Table 2.2**). The same assay was performed on samples from UK Mitochondrial Disease Patient Cohort however, these were performed by the NHS Highly Specialised Mitochondrial Diagnostic Laboratory, Newcastle upon Tyne NHS Foundation Trust. In total, 445 samples were pyrosequenced.

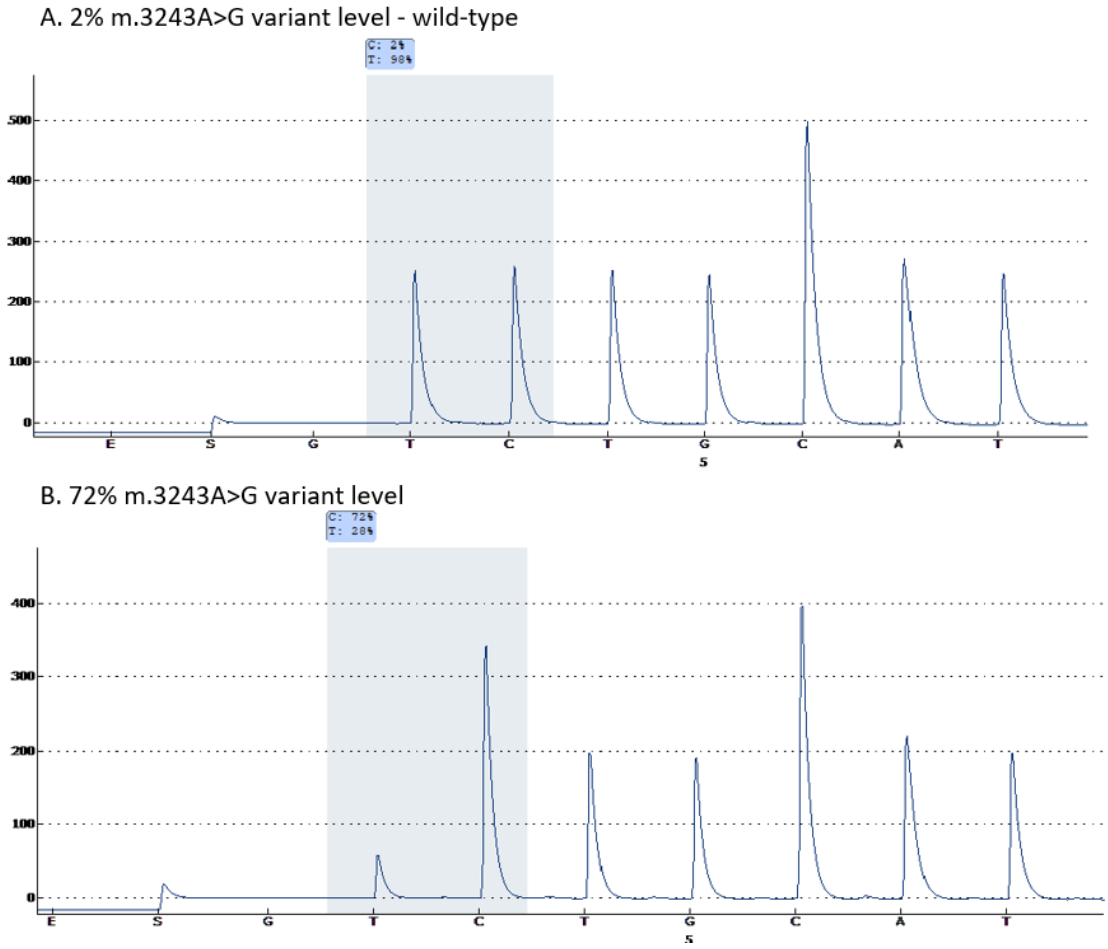

Figure 2.1: Pyro-grams depicting m.3243A>G variant levels. The sequence for m.3243A>G pyrosequencing is T/CCTGCCATCTT, in (A) The proportion of variants C:T is very alike, indicating that this is a sample with wild-type at 3243, compared to (B) where the ratio of C:T is skewed, reflecting an individual with 72% m.3243A>G variant level in their blood. [Figure retrieved from PyroMark software by Dr Roisin Boggan].

Table 2.2: PCR and pyrosequencing primer information. Pyrosequencing reaction was performed in line with instructions in PyroMark Q24 User Manual 01/2009 pages 33-40.

Primers and ref	Sequence	Length (bp)	Tm (C°)
Sequencing (m.3243A_G_Rev_Seq) *RefSeq NC_012920.1*	5' ATG CGA TTA CCG GGC 3'	15	52.3
Forward PCR (m.3243A_G_FBio)	5' /5Biosg/TAA GGC CTA CTT CAC AAA GCG 3'	21	55
Reverse PCR (m.3243A_G_R)	5' GCG ATT AGA ATG GGT ACA ATG AG 3'	23	53.5

2.2.2 mtDNA Variant calling using WGS data from blood samples

For 100kGP data, Dr Deen created a pipeline utilising Mutserve (v1.1), a software package designed specifically for mitochondrial variant detection (Weissensteiner et al., 2016a). Unlike standard variant calling software, which assume that the genome is diploid, mutserve deals with nuances specific to mtDNA including polyploidy and circularity. It also enables the detection of sites with a 1% allele frequency on each strand.

The pipeline used mitochondrial WGS (release v12) data GRh38 to call mitochondrial SNPs. Individuals with m.3243A>G levels ≥ 1% were selected, after passing a quality cut off of Phred 30, and coverage ≥ 100X.

Our collaborators in Exeter wrote their own pipeline for UKBB WGS data (GRh37) relying on MitoHPC software (Battle et al., 2022). Their cut-off for selecting individuals with m.3243A>G was also ≥ 1%. This pipeline excluded variants with a minor allele count (MAC) ≤ 5, minor allele frequency (MAF) < 0.01%, and coverage ≤ 200X (Cannon et al., 2023). Both software packages ensure specificity and sensitivity for mtDNA (Dierckxsens, Mardulyn and Smits, 2020), as a screen for contamination, they utilise haplogroup data, and provide coverage statistics (**Section 2.4.3.A-2** for more details).

2.2.3 Age correction of m.3243A>G heteroplasmy

Blood is one of the most non-invasive, easily obtained patient tissue samples, and blood corrected m.3243A>G levels have been the most reliable, and commonly used measure of heteroplasmy in the clinical assessment of mitochondrial patients. Grady et al., (2018)

determined that blood heteroplasmy declines with age at a rate of ~2.3% per year. This had to be accounted for given that data from 100kGP, UKBB, as well as most of the multicentre cohort come from sequenced/ genotyped blood samples. To do this, the blood age correction formula proposed by Grady et al., (2018), which adjusts for the 2.3% annual heteroplasmy decline (0.977), and accounts for the rapid decline in early age through the addition of 12, was applied.

$$Age\ adjusted\ blood\ \ level = \frac{Blood\ heteroplasmy}{0.977^{(age+12)}}$$

For the portion of samples that lacked blood levels of m.3243A>G in the multicentre cohort, estimates from urine samples were used (n=18). Due to gender differences in urine cellular content and mtDNA copy numbers, mutation load (variant heteroplasmy) in urine from males is 20% higher compared to females (Grady et al., 2018). This indicates the need to correct for sex when using urine samples and Dr Boggan used the adjusting formulae proposed by Grady et al., (2018) below:

$$Male\ adjusted\ urine\ level = \ logit^{-1}\left(\left(\frac{logit(urine\ heteroplasmy)}{0.791}\right) - 0.625\right)$$

$$Female\ adjusted\ urine\ level = \ logit^{-1}\left(\left(\frac{logit(urine\ heteroplasmy)}{0.791}\right) + \ 0.608\right)$$

Five samples within the multicentre cohort had neither blood nor urine estimates available, and so the level of m.3243A>G was estimated from muscle tissue. Given that it is a post-mitotic tissue, variant heteroplasmies remain largely the same over time and thus, no adjustments are needed (Grady et al., 2018).

2.2.4 Family tracing

R package kinship2 was used to identify pedigrees in 100kGP data (Sinnwell, Therneau and Schaid, 2014). This showed that the 250 individuals (116 m.3243A>G carriers and 134 relatives) belonged to 105 different pedigrees. Some filler individuals were added to enable pedigree drawing; these were given simple, sequential number IDs: 1, 2, 3 and so on. Age, age adjusted m.3243A>G levels, as well as the major and minor haplogroups

were included as additional information in pedigree drawings, the latter providing an additional quality control step for relatedness. Pedigrees were valuable for identifying the obligate carriers amongst the relatives; totalling 60 individuals, these are relatives who are clinically asymptomatic, but based on their position within the family are likely to carry the m.3243A>G variant at low levels (**Figure 2.2**). A minimum value of m.3243A>G allele frequency (1%) was assumed and assigned to these individuals, which then underwent age-correction using the formula above. Obligate carriers were consequently included in all analyses.

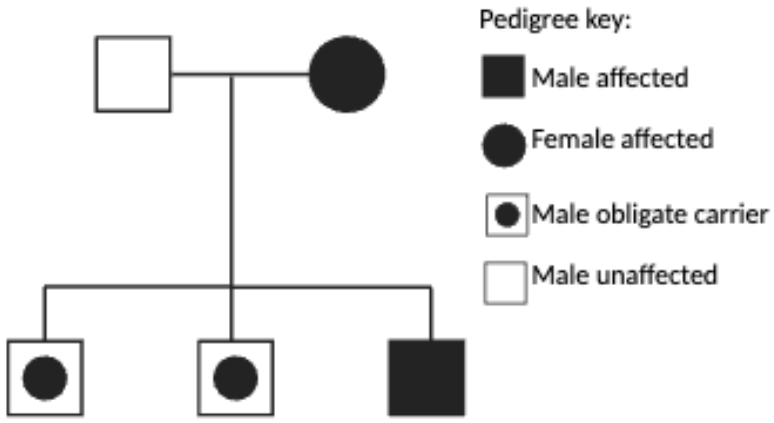

Figure 2.2: A pedigree depicting an example of a family case in Genomics England. Pedigree shows a family of three children. One of the siblings is shaded reflecting m.3243A>G blood level above the chosen ≥1% threshold. The other two siblings have been classified as 'obligate carriers'; this pedigree is a good representation of what this project is trying to unveil, the molecular genetic cause that has led to only one of the three siblings having detectable levels of the m.3243A>G variant.

2.3. Implemented software

A list of software packages utilised along with their version numbers and references is shown in **Table 2.3**. Briefly, the command line tools PLINK (versions 1.9 and 2), BCFTOOLS and VCFTOOLS were used to prepare genotyping/ sequencing data for analysis, as well as perform different quality control procedures. Three different GWAS software, FaSTLMM, REGENIE, and SAIGE were implemented and tested in **Chapter Three** of this thesis. For META analysis, GWAMA software was used. Ahead of the META, to ensure consistency in genomic assemblies across cohorts, LiftOver tool available from the UCSC website was used to lift over SNP coordinates.

Each software has its own way of referring to reference and alternative alleles, for example: REGENIE uses ALLELE 0 and ALLELE 1, SAIGE uses A1 and A2, for reference and alternative allele, respectively. On the other hand, PLINK v1.9 assigns A1 as the minor allele, and A2 as the reference allele. In order to avoid inaccurate alignment of effect direction in my analyses, it was necessary to standardize the reference and alternative alleles across all studies. For that, MungeSumStats software, which aligns data to a reference genome (build 38 or 37) and reformats the summary statistics, was used. Finemap was the software of choice for fine mapping analysis following the META, and LDAK (SumHer) was implemented to estimate SNP based heritability.

Table 2.3: List of the used software.

Purpose	Software	Reference
Data manipulation and quality control	BCFTOOLS (v11.2.0) VCFTOOLS (v0.1.16) PLINK (v1.9) PLINK (v2)	(Danecek et al., 2021) (Danecek et al., 2011) (Purcell et al., 2007) (Chang et al., 2015)
Visualisation of results and plotting	R (R studio V2023.3.1.446)	(Posit team, 2023)
GWAS	FaSTLMM (v2.07) SAIGE (v0.35.8.3) REGENIE (v3.0.1)	(Lippert et al., 2011a) (Zhou et al., 2018a) (Mbatchou et al., 2021a)
META	GWAMA (v2.2.2)	(Mägi & Morris, 2010)
Lifting over SNP coordinates between genome builds	LiftOver (v3.19) (https://genome.ucsc.edu/cgi-bin/hgLiftOver)	(Karchin et al., 2005)
GWAS results standardisation and (ref/alt) allele flipping	MungeSumStats (v1.10.1)	(Murphy, Schilder and Skene, 2021)
Fine mapping	FINEMAP (v1.4.2)	(Benner et al., 2016a)
SNP based heritability estimates	LDAK (SumHer)	(Speed, Holmes and Balding, 2020)

2.4. Methods of determining nuclear DNA variation and quality control (QC)

2.4.1 SNP genotyping and imputation

Quality-controlled, SNP genotyped, and imputed data from 408 individuals from the multicentre cohort were readily available by the time my project began. DNA samples from the multicentre cohort were SNP genotyped (genome build GRCh37 (hg19)) using the UKB_WCSG array, known as the UK Biobank Axiom® Array which is designed by the UK Biobank Array Design Group, and is widely used in research (Mizrahi-Man et al., 2022). The array carries 845,487 probes covering in total 825,928 markers, some with known disease associations and a MAF < 1%.

To ensure that the probe with the best call quality for each marker was taken forwards, initial QC was performed within the Axiom analysis suite (AxAS: version 4.0.3.3) using internal Axiom® quality metrics (Fisher's Linear Discriminant, FLD ≥4; Call rate ≥ 97; Heterozygous ratio offset ≥ 0; Homozygous ratio offset ≥ 0) all outlined in: (Boggan et al., 2022b). Data from 654,115 SNPs was exported in linkage format. Excluding PCA analysis, the per individual QC was performed as described in (Boggan et al., 2022b) (**Section 2.4.3.A**). The same data was taken forwards and a third round of QC was performed using PLINK (v1.9), all outlined in **Section 2.4.3.B** below.

GWA studies require dense SNP data. To increase resolution, and avoid missing any associated variants, statistical imputation was performed by Dr Pickett. This was performed using Michigan imputation server, and the Haplotype Reference Consortium (Version r1.1.2016) reference panel, which consists of 64,940 haplotypes of primarily European ancestry (Das et al., 2016). The output was 5,579,969 SNPs with imputation quality R^2>0.3.

The available WGS data in UKBB (200,000 individuals) were used by the Exeter team to identify m.3243A>G carrier samples (**Section 2.2.2** above). Genotyping data of the same 200,000 individuals were used for GWA analysis. UKBB data were also genotyped and imputed using the same methods outlined above, QC steps within the AxAS suite where samples with a poor quality due to, contamination for example, were identified and excluded. In addition to an array of per individual QCs that are thoroughly discussed in the following paper: (Bycroft et al., 2018).

2.4.2 *WGS*

Genomics England used Illumina's technology and Platypus variant caller (v0.1.5) on 61,676 rare disease individuals within Genomics England data release v12 (Rimmer et al., 2014). For my analysis, after running Dr Deen's pipeline and identifying individuals with m.3243A>G levels ≥ 1% (176 carrier and obligate carrier individuals), individual IDs were used to extract their nucDNA sequences. These data subsequently underwent a series of quality control steps, as outlined below.

2.4.3 *Quality control steps*

2.4.3.A Per individual QC (on 100kGP carrier and obligate carrier data)

2.4.3.A.1 Checking for discordant sex

As a check for discordant sex, X chromosome homozygosity was calculated. One reported female (**Figure 2.3**) deviated from expected values (1 for males and <0.2 for females (Wang et al., 2019)) and was consequently excluded from the downstream analysis.

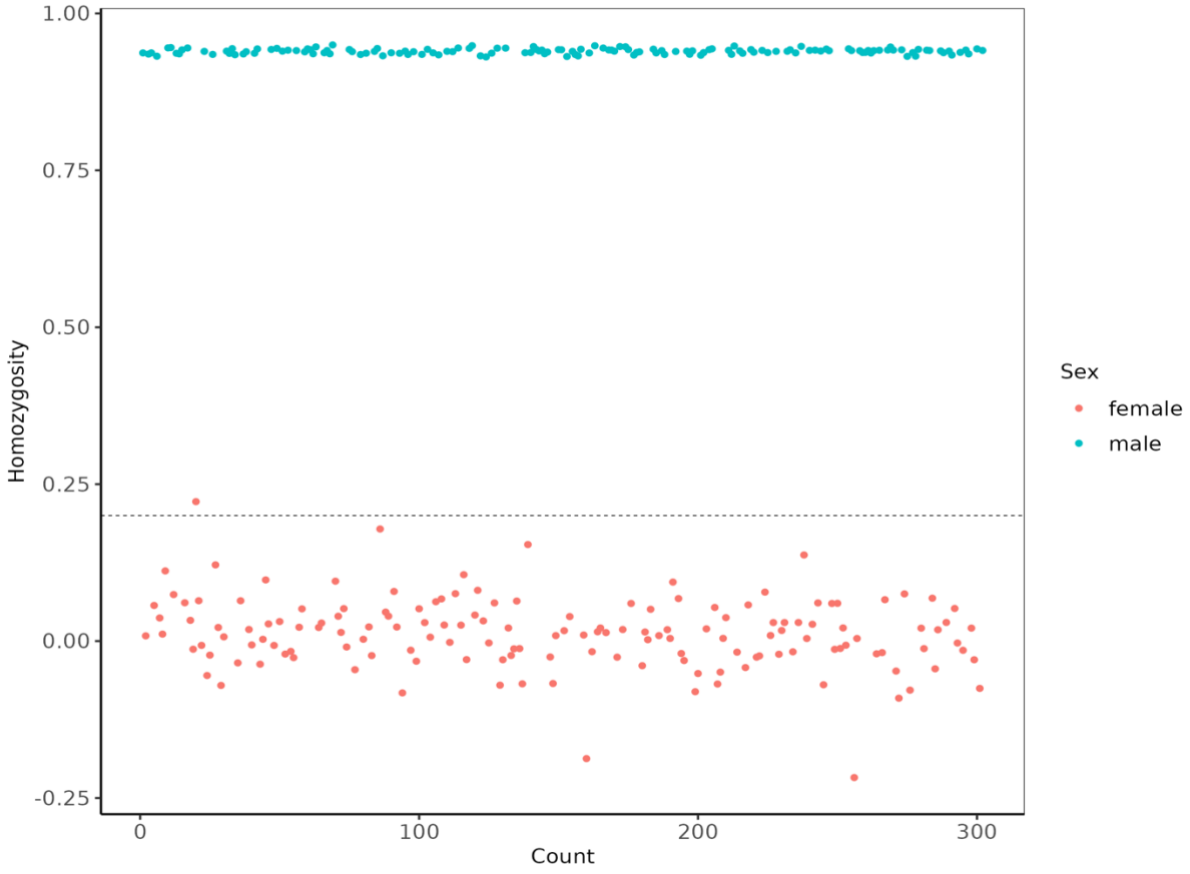

Figure 2.3: Discordant sex checks. Figure shows results from running PLINK sex checks; y-axis represents the F estimates of heterozygosity plotted against the frequency/count of samples. One clinically reported female had values above 0.2 and was excluded from downstream analysis.

2.4.3.A.2 Sample sequencing contamination

Haplocheck detects contamination patterns in sequencing studies by analysing the mitochondrial DNA (Weissensteiner et al., 2021); out of 176 (116 carriers and 60 obligate carriers), 13 samples were identified as contaminated by the software. Of those, four samples showed >1.4% contamination and were discarded. Contamination levels for the remaining nine samples were below the threshold and thus, were retained.

2.4.3.A.3 Identity by descent (IBD)

IBD is a pairwise analysis which measures the estimated proportion of two individuals' genomes that share either 0, 1, or 2 alleles inherited from a common ancestor. It takes the genome of each individual and compares it to everyone in the dataset to check the proportion of alleles shared. This was performed to check the relationships in the data and avoid any pedigree errors (Wang et al., 2019). Z0 is the fraction of the genome that shares 0 alleles from a common ancestor, Z1 the fraction sharing 1 copy, and Z2 the fraction sharing 2 copies.

A parent-offspring pair would have an expected Z1 score of 1, a pair of full siblings (including dizygotic twins) would have an expected Z1 score of ~0.50 and Z0 score of ~0.25 , and a pair of half siblings or second-degree relatives would have an expected Z1 and Z0 score of ~0.50. On the other hand, unrelated individuals have a Z0 score of 1 (**Figure 2.4**).

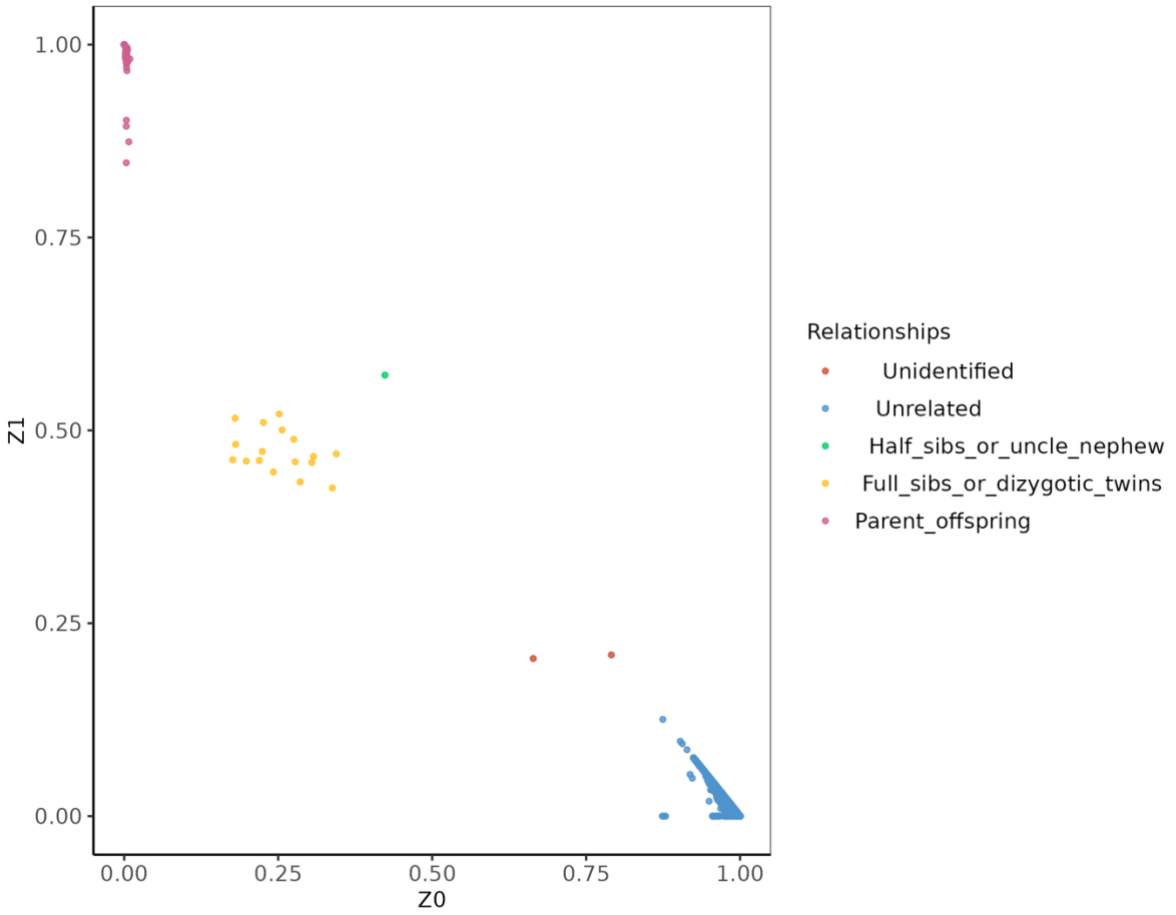

Figure 2.4: Identity by descent analysis. Plot depicts the IBD sharing of m.3243A>G carrier individuals in 100kGP cohort. This confirms all the reported familial relationships. Two individuals have an unidentified relationship however, considering the decreasing IBD scores and their closeness to the unrelated group of individuals, these might be third-degree relatives as IBD analysis is unreliable in identifying those.

2.4.3.A.4 Per individual missingness and heterozygosity rates

Individuals with heterozygosity rates that are considerably above, or below the mean values indicate: sample contamination, consanguinity (levels extremely below the thresholds), and population structure (levels extremely above the threshold) (Marees et al., 2018). Of 176 samples tested, nine individuals were not within the assigned thresholds for heterozygosity however, two individuals belonged to the same family and thus were retained whilst excluding the remaining seven. As it comes to per individual missingness of SNP reads, the generally accepted threshold is ≤ 0.02 (Marees et al., 2018), in this case, none of the samples exceeded a missingness of 0.006 (**Figure 2.5**).

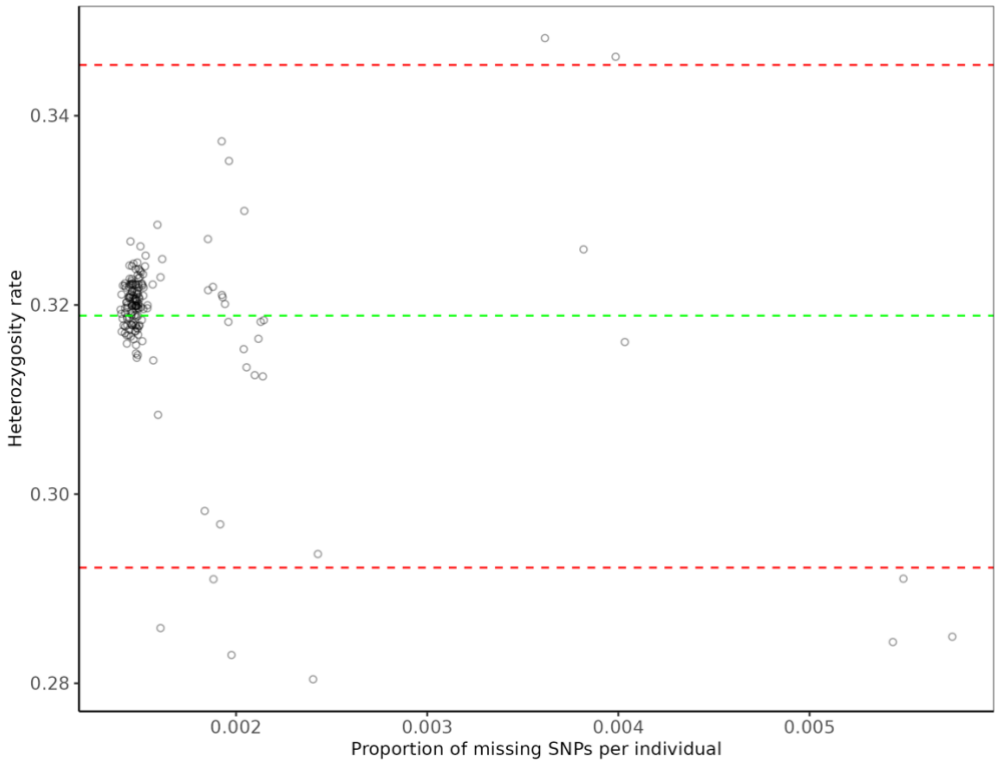

Figure 2.5: Per individual missingness and heterozygosity rates. Green dashed line presents the mean heterozygosity rate and the red lines are the upper and lower limits (mean ±3*standard deviation) used as thresholds. Nine individuals were found outside the limits; one pair was from the same family, indicating a level of consanguinity which explained their decreased heterozygosity values. The remaining seven were excluded as their heterozygosity could not be explained by consanguinity, and thus were an indication of contamination and potentially population structure in the two samples above the thresholds. These were consequently removed. The recommended threshold for missingness is 0.02; none of the samples was close to exceeding this threshold.

2.4.3.A.5 Principal component analysis (PCA) (on 100kGP data and multicentre cohort)

PCA is a statistical method used to reduce the dimensionality of multivariate data into components that explain the variation in data. In genetics, it is used to estimate genetic components' effect on the observed phenotypic variance however, it is also a method of genomic control that uses ancestry informative markers (Stranger, Stahl and Raj, 2011). The genomic inflation factor (λ), which is an indicative of how different/ similar GWAS results are compared to what is expected, is calculated by dividing the median of the first quantile in the chi-square distribution over 0.456, a value that reflects the median of the distribution under the null hypothesis (that is in the absence of inflation) (Devlin and Roeder, 1999). In case chi-square statistics are not available, then they can be obtained by running the retrieved per-SNP P values through the inverse chi-squared distribution

function. Values closer to 1 and up to 1.10 suggest no evidence of inflation. Inflation (values over 1.10) can be due to many factors including population stratification (PS), which is the difference in allele frequencies between populations. This can be accounted for by excluding PC eigenvalue outlier samples, or by including them as analysis covariates.

As GWA studies without corrections made for population structure can lead to false associations, this was a critical step (Price et al., 2010b).

Ahead of running the GWAS, data from m.3243A>G carrier individuals who passed the upstream QC analysis in both 100kGP (164 samples), and the multicentre cohort (408 samples), was combined with reference data from 1000 genomes project (build 37 for the multicentre cohort and build 38 for Genomics England), PCA was performed with PLINK producing eigenvectors and eigenvalues as output. Values were plotted and compared to those from the 1000 Genomes Project which contains individuals from five genetically distinct populations; African, European, East Asian, AD mixed American, and South-Asian (Auton et al., 2015a).

The majority of the individuals in all three cohorts appeared European (**Figure 2.6**), however, there are individuals from a variety of nuclear genetic backgrounds providing evidence that m.3243A>G does not occur exclusively within individuals of European ancestry as was once thought (this is discussed in more detail in **Section 3.3.1 Figures A and B**). As will be discussed in more detail in **Section 3.3.2**, association analyses performed after the exclusion of outlier samples yielded the best inflation factors, meaning it was the optimal way of correcting for PS, in comparison to including PCs as covariates. Based on this result, outliers of PC1 (<0) and PC2 (>-0.0175) values, such as individuals with family IDs E06, UCL_PED_009, and UCL_PED_014 as well as those seen when plotting PC2 and PC3 (PC3> 0.004) were excluded from downstream analysis, total in multicentre cohort is 24 individuals (**Figure 6**). The same was performed on the 100kGP data (shown in **Figure 3.3**), and the UKBB (by Dr Cannon) resulting in the exclusion of 28 and 4 individuals, respectively. UKBB PCA plots are not shown as due to privacy regulations, they were not approved for export out of the environment.

2.4.3.A.6 Haplogroup determination

Mitochondrial DNA sequence variations that are inherited in clusters across individuals of the same population are referred to as haplogroups (Torroni et al., 1996). They reflect ancestral background and dispite increased human migration, still somewhat reflect geographical distribution (Hägg et al., 2021). Haplogroup frequencies, and hence the variants that define them, vary between population groups (Biffi et al., 2010). Mutserve, which is the software used in Dr Deen's Genomics England heteroplasmy calling pipeline, includes Haplocheck (Weissensteiner et al., 2021). This tool is used as a way to detect contamination in samples; however, these data were also used to provide additional insight into the mitochondrial genetic ancestries of individuals, testing whether this is associated with the variability of m.3243A>G levels (**Chapter Five**). For the multicentre cohort, Haplogrep (v2) was used to retrieve individual haplogroup data (Weissensteiner, Pacher, et al., 2016).

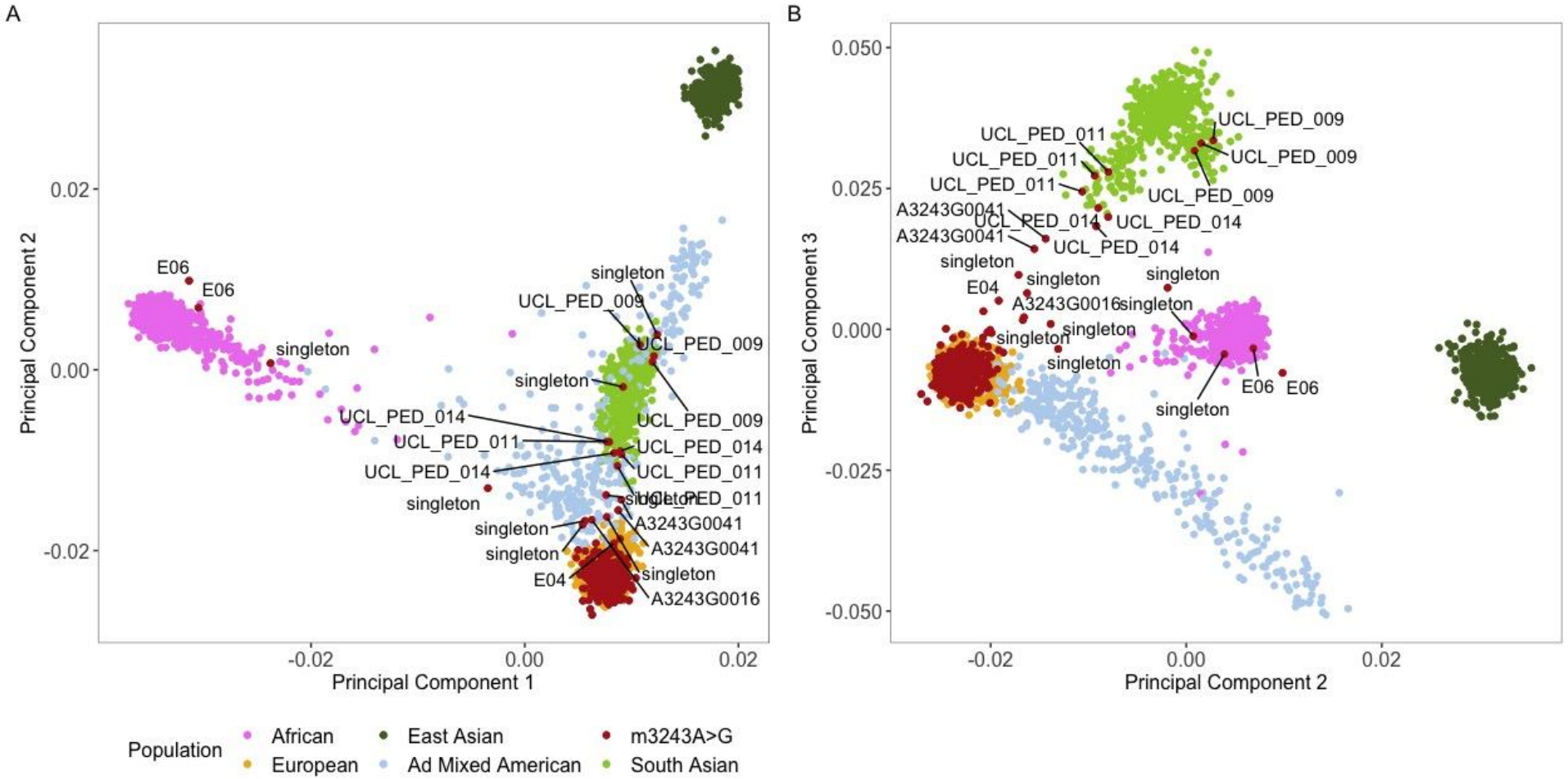

Figure 2.6: Principal component analysis in multicentre cohort. (**A**) presents PC1 against PC2, and (**B**) is PC2 against PC3, in both plots, each point represents an individual, however labels present their family IDs. Labels belong to European population outlier individuals (24 in total) which were excluded from downstream analysis. Labels of individual families are consistent in both plots.

2.4.3.B Per SNP QC

2.4.3.B.1 Minor allele frequency and missingness

GWAS analyses are typically performed on genotype data retrieved from commercial platforms usually covering common variants. Given that the technique has been best at detecting common variants with a small effect size, PLINK v1.9 was used to filter both Genomics England WGS data, and the imputed, multicentre genotyping data for a MAF ≥ 5% and SNP missingness <5%. This left a total of 6,957,718 and 5,575,537 SNPs in Genomics England and the multicentre cohort, respectively.

2.4.3.B.2 Creation of relationship matrix files using linkage disequilibrium pruning

GWAS software packages require a relationship matrix file as an input in order to incorporate information about relatedness into the model. Genetic relationship matrix is created using sparse SNP information (pruned for linkage disequilibrium).

Deviations from the expected association of SNPs at one or more loci is termed as linkage disequilibrium (LD) (Lewontin and Kojima, 1960). This is typically due to SNPs being in a close proximity, which decreases their chance of crossing over during recombination in meiosis. LD is often used by evolutionary biologists to understand evolutionary and demographic events as each genetic ancestry has its own LD pattern (Slatkin, 2008b). To enhance the precision of analysis results and eliminate any duplicated data, LD pruning, which is a method that involves the removal of one SNP from a SNP pair that is in high LD within the dataset, is a recommended procedure (Dudbridge and Newcombe, 2015). To do so, PLINK was used with the following parameters: --indep-pairwise 50 10 0.1 where 50 is window size in kb, 10 is step size, and 0.1 is r^2 (measure of LD) value. This retained 294,595 SNPs in the multicentre cohort and 382,300 in 100kGP. **Figure 2.7** below presents a summary flow diagram of the quality control steps performed on the 100kGP data and the multicentre cohort.

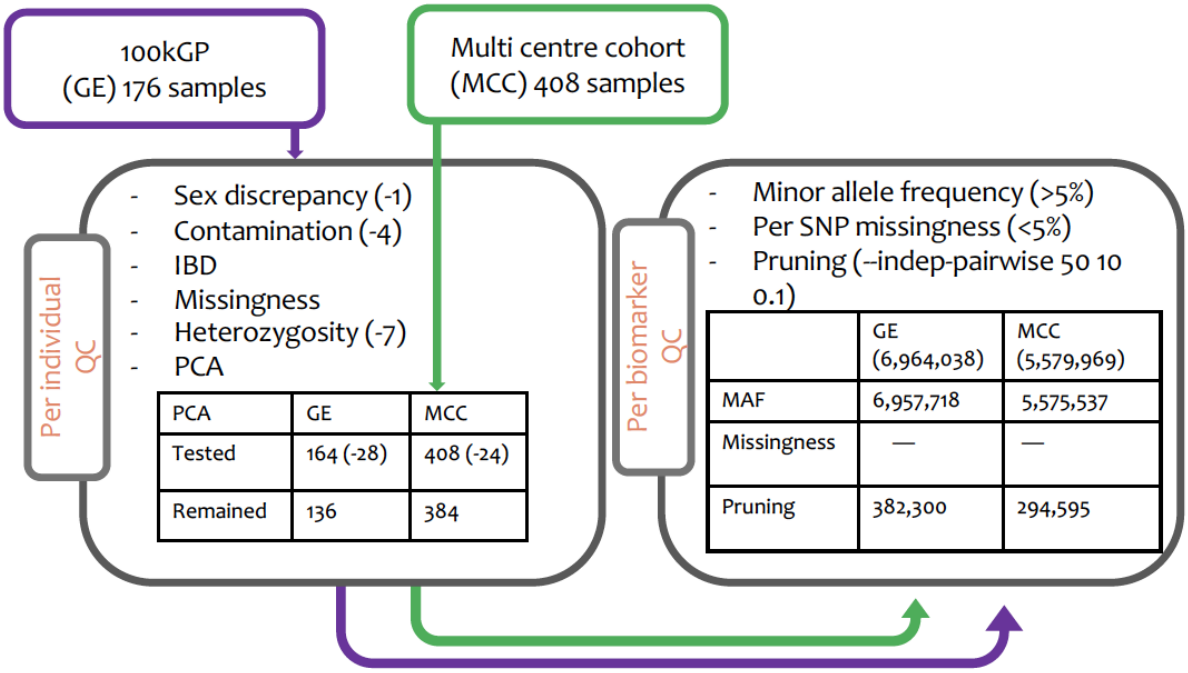

Figure 2.7: Flow chart depicting the QC performed on both 100kGP and the multi centre cohort. Multicentre cohort consists of 488 individuals however, as mentioned in **Section 2.1.1** genotyping data was available for only 445 samples. Dr Boggan performed two rounds of QC using Axiom analysis suite, as well as the necessary per individual QC using PLINK; which left 408 samples. I took the 408 QC'd data forward for my project and performed PC analysis and the second round of per SNP QCs. For the 100kGP data, I performed both per individual and per SNP QC; the number of individuals excluded at each QC step performed on 100kGP is indicated. QC performed on both cohorts are outlined in the tables. Total number of individuals that were carried forwards for analyses is 136 (excluding 28 outliers), and 384 (excluding 24 outliers) in 100kGP and the multicentre cohort, respectively.

2.4.3.C Lifting over SNP coordinates between assemblies

Genomics England data (100kGP) uses genome build 38, to ensure consistency, lifting over SNP coordinates from build GRCh37 to GRCh38 for both the multicentre cohort as well as UKBB data was performed using LiftOver, the web based UCSC lifting over tool **http://genome.ucsc.edu** (Kent et al., 2002; Kuhn, Haussler and Kent, 2013). Files containing chromosome, position and position -1 data were uploaded as inputs and an output with the updated positions was given. This was performed prior to META analysis.

2.5. Statistical tests

2.5.1 Linear mixed modelling

Mixed models, which underly one of the tested and investigated methods for my GWASs (in **Chapter Three** FASTLMM software), are called so because they have both fixed and random effects. As described in Winter, (2013) random effects are non-systematic parameters that have a predictable effect on data, such as principal components which reflect a population's structure; whereas a fixed effect is typically what we are trying to test; it is systematic and is something that tends to be "exhausted" within the population. In this scenario it is individuals' alleles at a given genomic locus, as a single locus can have two allele versions, and both varieties would certainly appear in the data. Both effects can be of a continuous or categorical nature, however, it is often recommended to resort to a logistic regression when the outcome is binary, avoiding biased estimates and incorrect inferences (more detail in **Section 2.5.2**). Linear regressions assume independence in the data points, if data belongs to groups (i.e. families, populations), a linear mixed model would be suited as it would account for the nested structure in the data (Gałecki and Burzykowski, 2013) (**Figure 2.8**). In case a simple linear regression is considered, the resultant model would fail to accurately predict the structure leading to false positives. A linear regression vs linear mixed models can be presented using the formulae below (Fox, 2002):

$Y = \beta_0 + \beta_1 X + \varepsilon$ where: Y: Dependant variable

X: Independent variable

β_0: The intercept

β_1 : Coefficient of the independent variable

ε: Error term

$y_i = X_i\beta + Z_i b_i + \varepsilon_i$ where: y_i: Dependant variable vector in the ith group

X_i: Fixed effects in observations of group i.

β: Coefficient of fixed effects

Z_i: Random effects in group i observations

b_i: Coefficients of random effects

ε_i: Vector of errors in observations of group i

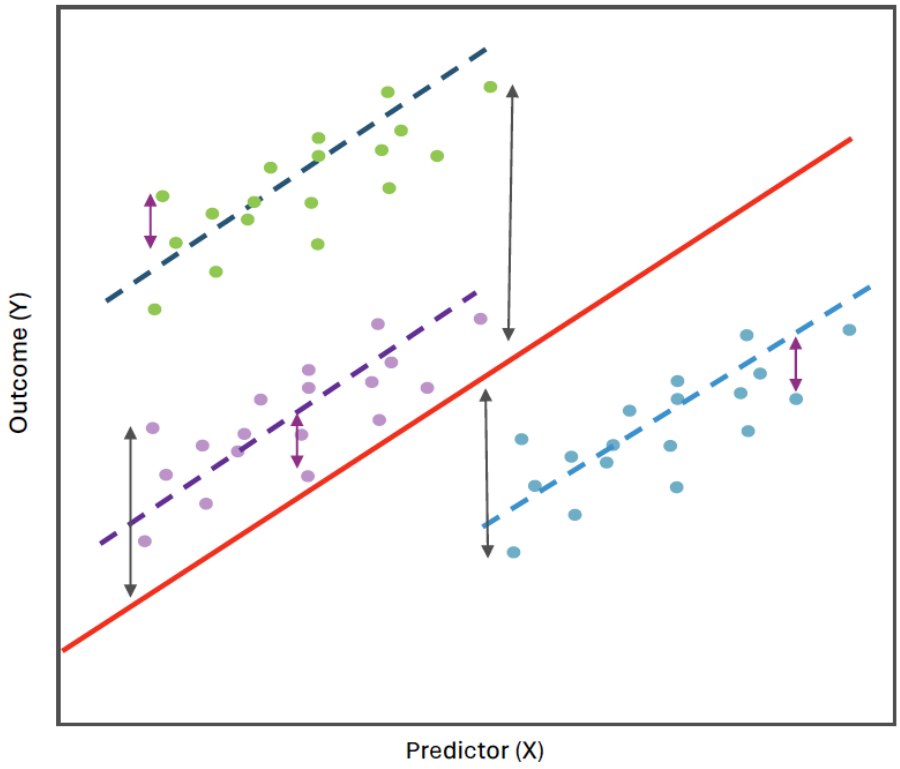

Figure 2.8: A linear regression on nested data points. Plot depicts a linear regression line (red) that runs through nested/grouped data points resulting in increased residuals indicating that the model fails to appropriately model the data. A LMM would result in multiple regression lines that would ensure a smaller residual, a smaller standard error value, and thus, a smaller p value. Leading to the best model prediction for each group (relatedness in the context of my GWAS).

2.5.2 Generalised mixed models

Both GMM and LMM allow the incorporation of random effects to account for correlated or nested data. However, GMM offer more flexibility in terms of the output variable where it allows for it to be both continuous (linear regression) or binary (logistic regression). Additionally, a LMM assumes a normal distribution for the output variable (also called as the error distribution) compared to different output distributions for GMM such as binomial or Poisson (Nettle, 2019). In summary, when there is structure in the data that needs to be accounted for, but the outcome cannot be normally distributed, maybe because of its categorical/ binary nature, then a generalised mixed model should be chosen. This is the model that underlies two GWAS software tested in **Chapter Three**, REGENIE and SAIGE.

2.5.3 META analysis

META analysis started appearing in the literature at much higher rates in the late 1970s, and it was Glass who first defined the term as: "*The statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings*" (GLASS, 1976).

A META is looked at as a method effective at assessing results from a group of studies, to later make conclusions about that particular body of research (Lean et al., 2009). Ideally, these studies should be picked at random in order to lead to precise effect estimates, however, there are concerns about this since, for example, large significant studies are published more often than studies with negative/ insignificant results and this can lead to selection bias. Despite these limitations, a META analysis of GWAS results is an effective tool in increasing analysis detection power. Heterogeneity (I^2) is a measure of variation across studies, and it often directs us to the type of statistical model that would best suite our data (random or mixed effects META). A thorough overview of heterogeneity estimates as well as META analysis is in **Sections 4.2.3 and 4.2.4**.

2.5.4 Fine mapping analysis

An associated region can harbour thousands of variants that are in complex LD, and correlation patterns (Benner et al., 2016). Fine mapping following a GWAS or a META analysis, aims to pinpoint at the "causal" variant(s), possibly associated with the molecular mechanisms leading to the studied disease or trait, to reduce the number of SNPs for follow-up studies. If done in combination with functional annotations, this can also lead to the identification of disease target genes (Spain and Barrett, 2015). There are multiple methods for fine mapping. Considering the complexity of my studied phenotype, I decided to investigate software that allow the possibility of multiple causalities in a locus, compared to one causal variant. I chose to use FINEMAP (v1.4.2) for my analysis (Benner et al., 2016), which is thoroughly discussed in **Section 4.2.6**.

2.5.5 Power analysis

Power analysis provide an estimate of the detection power given parameters such as, sample size, minor allele frequency, and effect size (ES); in addition to an estimate of the required sample size to detect variants with a certain level of significance, using a certain ES and MAF. GENPWR package in R (v1.0.4) (Moore, Jacobson and Fingerlin, 2019), was used to calculate the power of both GWAS and META analysis (see **Section 4.3.1** for details).

Chapter 3. GWA analysis optimisation

3.1 Introduction

The focus of this chapter is GWAS optimisation. The various software and study designs that were examined prior to deciding on the parameters that yielded the most optimal results for analysis will be evaluated.

3.1.1 m.3243A>G investigations leading up to GWAS

Heritability studies utilising variance components methods, adjusting for mother's variant levels by including them as covariates, identified that 72% of m.3243A>G level variability is explained by additive genetic factors (Pickett et al., 2019). To investigate this, our team carried out linkage analysis, identifying a linkage peak (LOD ≥ 3.3) on chromosome one (**Figure 3.1**); results point at large areas within a chromosome. On the other hand, association analyses which rely on linkage disequilibrium and deviations from the expected random segregation of variants, step in to offer finer mapping of variants either within previously identified linkage regions, or in an unbiased scan of the genome with no *a priori* knowledge (Visscher et al., 2012; Joiret et al., 2019). To identify these variants and infer the percentage of heritability that that they account for (by dividing their combined effect sizes over the total heritability), GWA analysis was then the ideal continuation of our investigation.

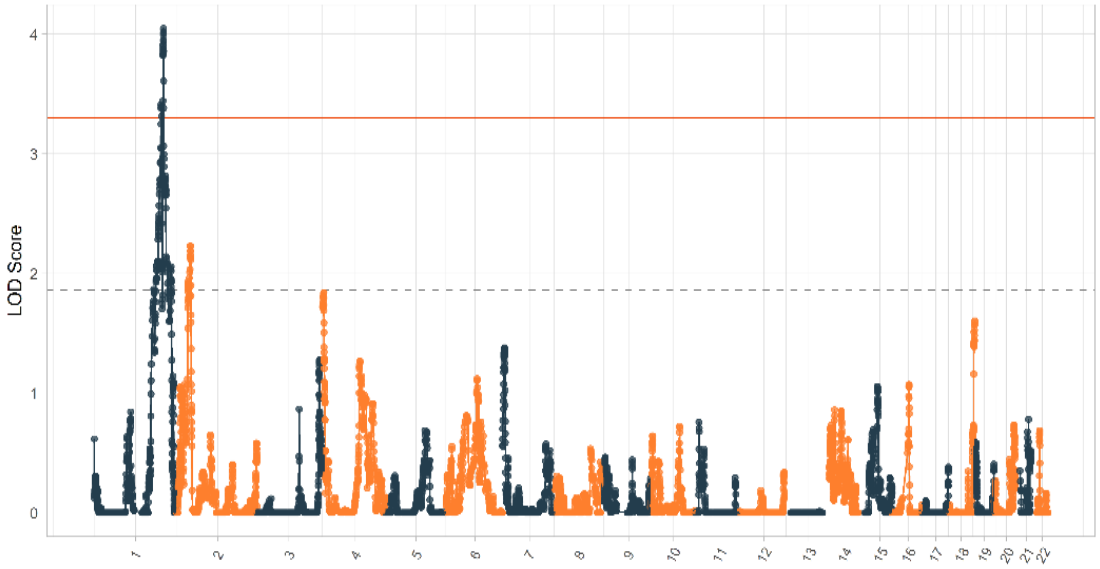

Figure 3.1: Linkage analysis logarithm of the odds (LOD) score results. Linkage analysis for m.3243A>G levels on 65 individuals from 45 pedigrees was carried out to estimate the position of the contributing nuclear genetic factors. As a result, one region on chromosome 1 (179,262,018 to 224,482,984) was identified with LOD score >3.3, and another region on chromosome 2 (26854157 to 56854157) was identified as suggestive of significance. [Figure obtained from analysis carried out by Dr Boggan].

3.1.2 Evolution of GWAS

In 1986, Bodmer suggested that using population data could provide linkage signals that are closer to the causative variant i.e., achieving finer mapping (Bodmer, 1986). Foreseeing the implications of future developments in the field of sequencing, Risch and Merikangas (1996) stated that "*the future of the genetics of complex diseases is likely to require large-scale testing by association analysis*". They also stated that overcoming technological constraints and having access to more "polymorphism data" would enable the performance of association analysis without the prerequisite of initial linkage analysis, unveiling associations that identify variants with subtle effects that linkage analysis overlooks, thereby enhancing our understanding of disease aetiology (Visscher et al., 2012).

The completion of the human genome project (Powledge, 2003) improved our understanding of common variations. The increasing cost-efficiency of genotyping platforms, in addition to the growing accessibility to large cohorts, all marked the beginning of GWA studies era (Morris et al., 2010). These have proved to be extremely successful. Perhaps one of the best early examples is a study carried out by the Wellcome Trust Case Control Consortium (WTCCC) on a cohort of 2,000 British ancestry individuals

for each of the common diseases chosen for the study, which identified three genes linked to type two diabetes, a region associated with obesity, four chromosome regions associated with type 1 diabetes, three regions associated with Crohn's disease and one gene, *PTPN2*, that predisposes individuals to both type 1 diabetes and Crohn's disease (Burton et al., 2007).

As previously mentioned, methods of investigating complex diseases were found successful in many instances across different diseases. Due to its well-defined mode of inheritance and early onset, type 2 diabetes (T2D) made it easy to collect large, extended pedigrees (Vaxillaire and Froguel, 2006). This made T2D at the forefront of diseases studied by different genetic analyses. In 2003, Reynisdottir et al., identified regions of suggestive linkage to T2D on chromosomes 5 and 10, later it was found that chromosome 10 harbours *TCF7L2* gene, which is one of the key genes leading to diabetes. As association analyses developed and emerged from candidate gene approach to more unbiased approaches, *PPARG* as well as *KCNJ11* were identified and are currently targets for anti-diabetes medications (Gloyn et al., 2003).

Moreover, linkage and GWA studies were also successful in untangling the genetic structures of quantitative traits such as immunoglobin E (IgE). IgE is a mediator of allergic inflammation, with often increased levels in individuals with asthma. Pedigree-based studies attributed a considerable amount of its variability to genetic factors with heritability levels ranging from 36% to 78% (Meyers et al., 1987; Jacobsen et al., 2001). Several linkage reports have been published, all showing a great disparity between populations, for example, an area on chromosomes four and 18 were found in Caucasian ancestries, another linkage was found on chromosome two in Afro-American populations. A linkage study performed on 200 families (1,171 family members) collected from a regional referral centre for pateints with asthma in the Netherlands, pointed at at a region on chromosome two with strong evidence for linkage (2q33) (Xu et al., 2000). This region was found to harbour two candiadte T cell activation genes *CTLA4* and *CD28.* Years later, a GWAS study on Taiwanses Han individuals was performed, and this identified a significant association with rs1181388 SNP in cytoband 2q33.2 within *CD28* gene (Lu et al., 2024). The same SNP was also reported in a study performed on the UKBB (Ghoussaini et al., 2021).

3.1.3 Association tests

The choice to deploy GWAS-type analyses depends on the nature of studied phenotype, the need to account for population structure and/or pedigree structure, as well as the presence of covariates to control for (Wang et al., 2019). GWA analysis methods can be classified into two groups: gene and SNP based investigations. Different approaches have been employed to account for population structure and different confounders and covariates such as, generalised mixed models (GMM) (Charles E. McCulloch & Shayle R. Searle, 2004), linear mixed models (LMM) (Loh et al., 2015), and variance components methods (VC) (Svishcheva et al., 2012); below is a comparison between those techniques (**Table 3.1**).

Bearing in mind data structure in all cohorts, particularly in the multicentre cohort, it was necessary to account for family structure in the analysis; additionally, neither of the cohorts had a particularly large sample, and m.3243A>G levels did not follow a normal distribution in either of the cohorts, a choice was made to opt against normalisation as will be thoroughly discussed in **Section 3.3.4**. All these factors narrowed the options down to generalised and LMM models. Considering that a META analysis on the retrieved GWAS summary statistics was planned, to ensure that the data was treated in the same way, and results were reported consistently across studies, I wanted to make sure that the same software was applied to analyse data from all cohorts.

In this chapter, the performance of REGENIE, SAIGE, and FaSTLMM were evaluated using the multicentre cohort, by comparing the lambda inflation factors obtained as a way to estimate their correction efficacy. Both REGENIE and SAIGE have been reported to be particularly efficient in biobank scale data (Schönherr et al., 2024). REGENIE and SAIGE are the default software used in UKBB and Genomics England (100kGP) cohorts, respectively.

FaSTLMM is a popular choice in LMM based associations and has a relatively low memory and time footprint compared to other LMM-reliant software (Lippert et al., 2011)

Table 3.1: Comparison between different GWAS methodologies. [Table compiled from Charles E. McCulloch & Shayle R. Searle, (2004); Liu et al., (2010); Loh et al., (2015); Svishcheva et al., (2012); Visscher et al., (2017); and Wang et al., (2019)]

APPROACH	WHEN TO USE	BENEFITS	DRAWBACKS	SOFTWARE
GENE-BASED ASSOCIATIONS	-When wanting to identify genes that play a part in a network, pathway manner. -Detect multiple variants within genes and sum up their effects in an overall score. -Accounts for between individual genetic heterogeneity.	-Reduces multiple testing problem by testing the genes in the genome rather than the millions of SNPs simultaneously. -Permutations are used to account for gene size and LD structures.	Permutations are computationally challenging to be applied on the whole genome level hence why this kind of approach particularly requires a genetic relationship matrix.	SAIGE-GENE
VARIANCE COMPONENTS	In case of unknown pedigree structures in the data, especially in the presence of enough genotyped SNPs. To account for any possible population structures even in seemingly homogeneous samples.	-When using the two-step approaches that measure relationship matrices in the first step and then the association between the genotypes and phenotype of interest in the second, have proved to be speedy and efficient in dealing with sequencing data of thousands of individuals. -Estimates the phenotypic variance attributed to these factors (heritability).	-Ascertainment biases are frequent, especially in non-normally distributed phenotypes. -It tends to require large sample sizes to perform accurately, in case of insufficient sizes, standard errors tend to be increased. -Better performance with quantitative traits.	GRAMMAR-Gamma
LINEAR MIXED MODELLING	Identifying associations while avoiding confounders as well as controlling for complex correlation structures.	-Compared to other methods that also correct for population structure and relatedness, LMM is one of the fastest, most computationally efficient methods.	-They assume that all variants are causal with a small effect size (Infinitesimal model). -High memory and time requirement that scales as the number of phenotypes increase. -Less power in ascertained case-control studies.	FaST-LMM, GCTA, GEMMA, EMMAX, BOLT-LMM
GENERALISED MIXED MODELS	When there are multiple covariates such as, age, sex, and array type that need to be corrected for in the analysis in addition to population structure.	-No assumption of normality in the phenotype data is needed. - Accommodates different kinds of data; linear regressions can be used with continuous data whereas logistic regressions with binary/ categorical data. -In unbalanced case-control studies, Firth regression uses penalization that reduces bias and provides more reliable statistical inferences -Can be used with multiple traits at a time which is useful in analyses that investigate pleiotropy.	-They assume that variants have an additive relationship with the phenotype. -More sensitive to population structure and relatedness compared to other methods.	REGENIE, SAIGE, GMMAT

3.2. Methods

3. 2.1 FaSTLMM

One of the biggest advantages of LMM is its ability to incorporate both fixed effects and random effects into the model in addition to accounting for population stratification and relatedness. Random effects can be clinical or different environmental factors such as: age, sex, BMI, or exposure (Dandine-Roulland and Perdry, 2016). To account for population stratification, it is a common practice to include 10 to 20 genetic principal components along with the fixed effects; this can be modelled as follows (Lippert et al., 2011):

$$\prod_{i=1}^{n} [\![X^{(i)}\mathrm{b} + Z^{(i)}g + \epsilon^{(i)}]\!]$$

Where:

$X^{(i)}$: the fixed effects for the ith individual

b : fixed effect weights

$Z^{(i)}$: SNP data for the ith individual

$\epsilon^{(i)}$: measurement error for the ith individual

In a linear mixed model, identified differences between populations receive more correction, and contributions identified due to relatedness are reduced, preventing by that the usage of redundant data that reflects correlation structure (Yang et al., 2014a).

LMM have a high memory and time footprint that scales drastically as the sample size increases; a way around this has been through the development of an improved version called factored spectrally transformed LMM (FaST-LMM). Basic LMMs typically use either realized relationship matrix (RRM), or identities by descent (IBD) which use a subset of the markers in a step prior to association testing to elucidate relationships and confounders within the data. On the other hand, FaST-LMM utilises spectral decomposition used in PC analysis, that can be performed on all markers, removing by that the "cubic computation" per SNP that made LMMs computationally inefficient and time consuming with large samples (Lippert et al., 2011b). Another pitfall of LMMs that has been efficiently controlled in software like FaSTLMM and GCTA, is the loss of power whenever candidate markers are included in the RRMs/IBD step (Yang et al., 2014a). This is due to over fitting of markers in the model; to avoid this, a new approach was developed by Listgarten et al., (2012) where they exclude candidate markers (low P

values) when estimating genetic similarities using ~8,000 equally spaced markers. In my analysis, all options offered by the software were left to their default providing a quality-controlled genotyping/sequencing file and a corresponding RRM file, different number of PCs were used as covariates which is further discussed in **Section 3.3.2**.

A discovery T2D GWAS performed on an extended Emirati family (n=178), utilised FaSTLMM software after transforming the data into binary format, and testing different inheritance patterns (additive, dominant, recessive) (Al Safar et al., 2013). FaSTLMM was the software of choice as it provides correction for the underlying family structure between the tested individuals. This, for the first time, identified novel loci associated with T2D in an Arab family; authors highlighted the need for replication on a larger cohort to generalise these findings. Another GWAS utilised FaSTLMM on 196 British Caucasian families, investigating nuclear associated loci with N-acyl ethanolamine (NAE), and ceramide (CER), both biomarkers for coronary artery disease and T2D (McGurk, Keavney and Nicolaou, 2019).

3.2.2 REGENIE

REGENIE was designed to allow multiple sophisticated regressions to be carried out in two steps. These ensure the confounding factors are accounted for, and that there is no "proximal contamination", which is the inflation of false positives due to uncorrected for familial relationships and/or population structures. As outlined in Mbatchou et al., (2021), step 1 level 0 aims to reduce dimensionality – using ridge regression, SNPs are grouped into blocks where *J* ridge regression predictors are then identified, this provides a rough estimate of the number of candidate markers in each block which can also be considered as the chosen markers out of an LD block. For analyses in this thesis, a block of 1000 SNPs was utilised for J ridge regression. A second ridge regression, that is ideal when dealing with multicollinearity (when two independent variables are highly correlated), is then carried out in step 1 level 1, this time to merge the markers/predictors all into one that is decomposed into 23 chromosomes, by that allowing the implementation of LOCO (Leave One Chromosome Out) that ensures none of the identified associations are due to LD patterns in genomic regions (cross contamination). Step 2 can utilise either a linear regression, in case of a continuous phenotype, or a logistic regression if it is a binary/

categorical trait, as will be discussed in **Section 3.2.2** both regressions were tested but a linear regression was the one of choice.

Results from step1 levels 0 and 1 can be saved and utilised for other phenotypes, reducing the used memory, and increasing analysis speed. Throughout this thesis however, only one phenotype was tested at a time.

P values are based on a likelihood ratio test (LRT), and in case P values are below the specified threshold (typically 0.05), Firth correction (FIRTH, 1993) is used as a way to remove bias associated with standard maximum likelihood estimates, which can arise in small sample sizes due to the limited amount of information available and may lead to a mis-specified distribution. The addition of a penalty term to the log likelihood leads to more accurate results. REGENIE also uses saddle point approximation (SPA) as a way to provide a more accurate estimate of the underlying data distribution (Daniels, 1954). Where instead of relying on only the first two cumulants (mean and the variance) in data, it uses all entire cumulant generating function, which is a sequence of numbers that describe a distribution. The underlying formulae for each step in REGENIE are outside the scope of this thesis, (Mbatchou et al., 2021a) is a comprehensive resource for those seeking further insight.

In this study, –firth and –pThresh of 0.01 (for mtDNA GWAS) and 0.05 (for nuclear GWAS) was used to indicate the p value threshold below which a firth regression should be applied.REGENIE has been used on several large-scale cohorts, such as the UKBB. It adjusts for confounding factors and is preferred for its speed, and efficient memory usage as analysis time does not scale-up with sample size as the case is with FaSTLMM, for example. More detail about the usage of REGENIE software is in **Section 3.3.4**.

3.2.3 SAIGE

SAIGE was designed to tackle the problem of type 1 errors that typically arise from using linear mixed models on binary and unbalanced case-control data (Zhou et al., 2018a). The decreasing costs of genotyping as well as sequencing means that the number of databases and cohorts will be growing, and a small number of controls compared to cases in a database setting is often a commonly observed pattern. SAIGE stands for Scalable and Accurate Implementation of GEeneralised mixed model, and it runs in two steps – step 1 requires raw genotyping data and uses Gaussian restricted maximum

likelihood (REML) to fit the null logistic model (Zhou et al., 2018a). This is used to estimate the parameters (variance components) of a model without relying on any predictors, creating a baseline model that will allow an accurate comparison upon the addition of any predictors, whether they are fixed or random (Gilmour, Thompson and Cullis, 1995). Step 1 is when genetic relatedness and population patterns are identified, accounted for and fitted into the model. To decrease storage space and computational cost, SAIGE uses an iteration of the common generalised relationship matrix (GRM) tool for modelling called PCG, principal component analysis on genotypes. GRM compares the genomes between pairs of individuals however, PCG, just like PC analysis, identifies the largest sources of variation in a whole dataset in the form of principal components that are then used as covariates in the analysis (Zhang and Pan, 2015). To reduce memory storage and cost, SAIGE stores the genotype data in binary format, and calculates the PCG only once instead of storing, something that is essential when dealing with large samples. Step 2 utilises a faster, more developed version of SPA called fastSPA (Dey et al., 2017), to test association between SNPs and the phenotype of interest, which is particularly efficient in correcting any inflation that may be caused by imbalanced case-controls. LOCO is also an available option for SAIGE, that ensures no cross contamination. When using SAIGE (v0.35.8.3), across all three cohorts, LOCO option was employed, with the remaining options left to default. The logistic regression model in SAIGE for both binary and continuous traits can be presented in the following:

$$\text{logit}(\mu_i) = X_i\alpha + G_i\beta + b_i + \epsilon_i$$

$$Y_i = X_i\alpha + G_i\beta + b_i + \epsilon_i$$

Where:

logit μ_i: is the probability of a binary outcome for the i-th observation

Xi : covariates for the i-th observation

α : fixed effects coefficients associated with covariates in X_i

G_i: genomic data of the i-th observation used to account for any confounders

β : effect size of G_i

b_i: random effects and intercepts for each i-th observation

Y_i: continious outcome for the i-th observation

ϵ_i: measurement error/ penalty

In studies comparing GWAS software, SAIGE was found to be much more time efficient with large-scale cohorts compared to LMM based software such as FaSTLMM and BOLT-LMM however, as reported in (Yang et al., 2014b) it is extremely computationally challenging on large scale data. Additionally, SAIGE utilises dense GRM, and this puts analysis at risk of proximal contamination, which was also demonstrated in (Yang et al., 2014b). All factors that reduced its usage in large scale analyses.

3.2.4 Binary trait vs continuous trait GWASs

Phenotypes are divided into two classes: continuous, and categorical. Both have been used in GWA studies however, continuous traits are often preferred and considered to be more powerful (Bush and Moore, 2012). Given they provide the statistical test with more information, it ensures that an effect is identified and not lost as a type II error (Bush and Moore, 2012). Nonetheless, a continuous trait is not a perquisite for a successful GWAS. Generalised linear models utilising SPA such as, REGENIE and SAIGE, have been proposed to be one of the best methods to implement on binary traits as they not only ensure that all confounders are accounted for, but the restricted binary-mean structure (from 0 to 1) is modelled accordingly (Gurinovich et al., 2022; Yang et al., 2014). Binary trait associations have some advantage over continuous traits – particularly when the distribution of data is skewed, as it can be robust to ascertainment, and it offers faster computation (Jiang, Mbatchou and McPeek, 2015).

As a way to test whether the association would yield better performance if m.3243A>G variant levels are coded as a binary trait, the cut_number function in R was used to group data into three quantiles. Data points that fell within the first quantile were coded as 0s and that belonging to quantile three were coded as 1s (**Figure 3.2**), analyses were performed with three different software and results are in the sections to follow.

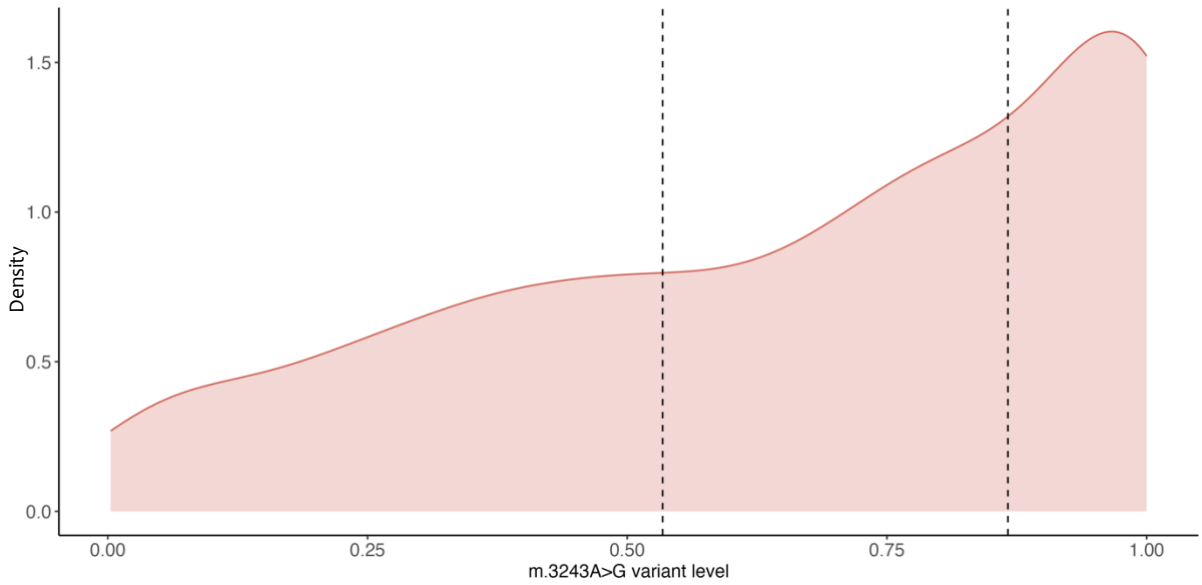

Figure 3.2: Distribution of age-adjusted m.3243A>G levels in the multicentre cohort. Plot shows the distribution of variant levels as well as the first 3 quantiles (n=408).

3.2.5 Principal component analysis (PCA)

As outlined in **Chapter Two**, PCA is a dimensionality reduction procedure, which aims to explain the variability in data by reducing the number of variables while preserving as much information as possible (Lever, Krzywinski and Altman, 2017). To interpret results, 1000 genomes project reference data which contains individuals from five genetically distinct populations: African, European, East Asian, admixed American, and South-Asian, was merged with data of each cohort. PLINK V1.9 (Purcell et al., 2007) (http://pngu.mgh.harvard.edu/purcell/plink/) was used to compute eigenvalues and eigenvectors.

This information used to account for population structure, which will be discussed in the sections to follow. For ease, the comparisons and analysis outlined are all carried out on the multicentre cohort.

3.3. Results

3.3.1 Nuclear principal component analysis

Analysis on the multicentre data, 100kGP, and the UKBB, each combined with 1000 genomes reference data was performed using PLINK V1.9. Eigenvalue, and eigenvector

data from all three cohorts was retrieved. **Figure 3.3** presents results from 100kGP (n=164) and the multicentre cohort (n=408). Collaborators from the university of Exeter performed PC analysis on the UKBB data (n=147) their results, however, are not presented in this thesis.

As expected, based on patient recruitment centres, the majority of m.3243A>G carrier individuals fall within the European reference cluster, as seen in **Figure 3.3**. Individuals who fell outside of the European population cluster (outliers), were identified and were either excluded from the analysis or accounted for by including PCs as covariates. Scree plots representing the percentage of variance explained by the first 20 PCs (**Figure 3.4**) show that the percentage of variance for the first three PCs varies drastically between the two cohorts. In 100kGP, PC1 accounts for ~53% of variability whereas in the multicentre cohort it is only ~8%. This, as well as the greater number of identified outliers in 100kGP, can be attributed to the greater population variability and diversity of data, compared to the multicentre cohort that harbours a greater number of pedigree data.

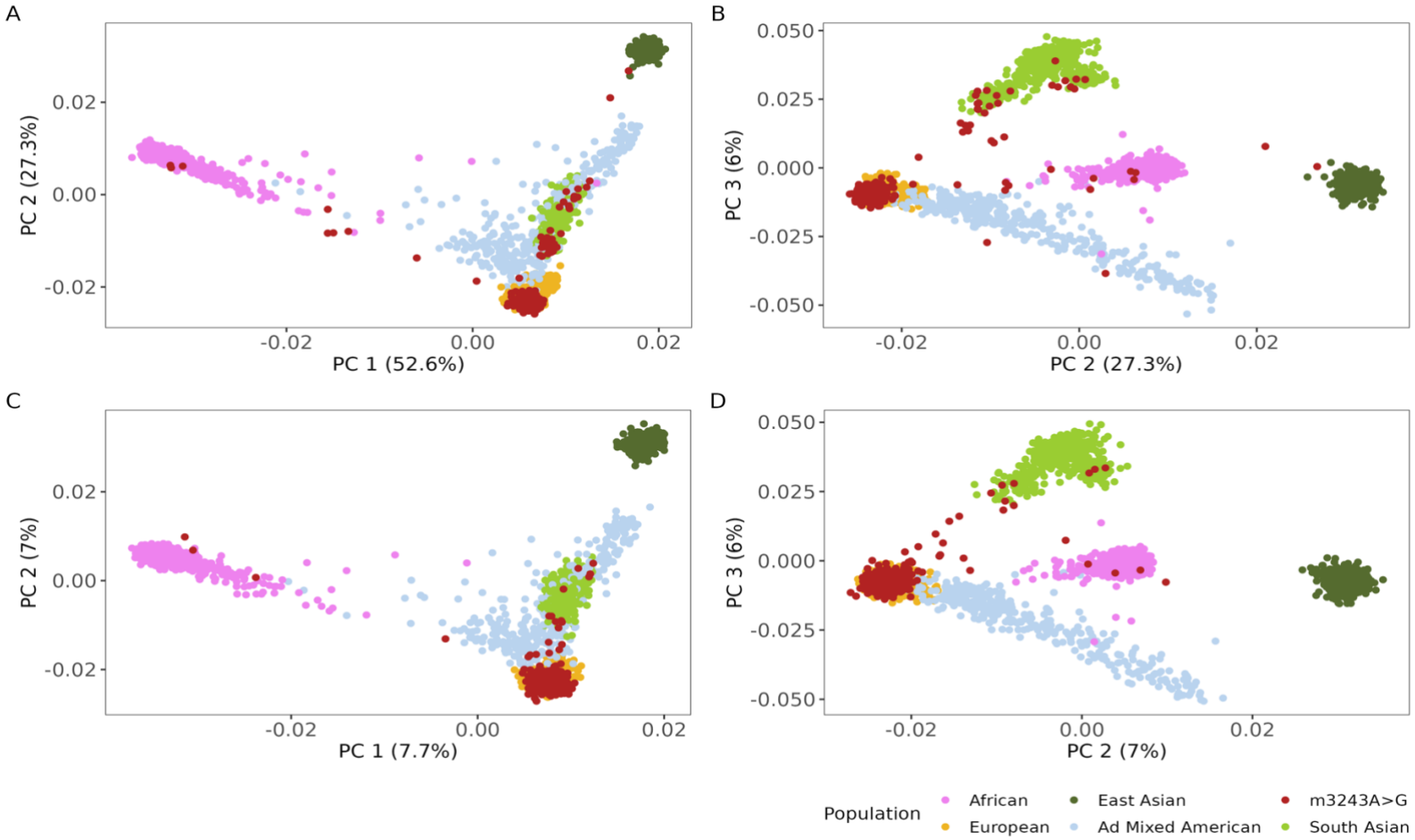

Figure 3.3: Principal component analysis biplots. (**A+B**) Results retrieved from 100kGP data, shows that the majority of m.3243A>G carriers (red) lay within the European population cluster. 24 individuals were excluded from analysis as they fell outside the European cluster. (**C+D**) Results from the multicentre cohort, total number of outliers identified is 28. Our Exeter collaborators identified 4 outliers (plots not shown).

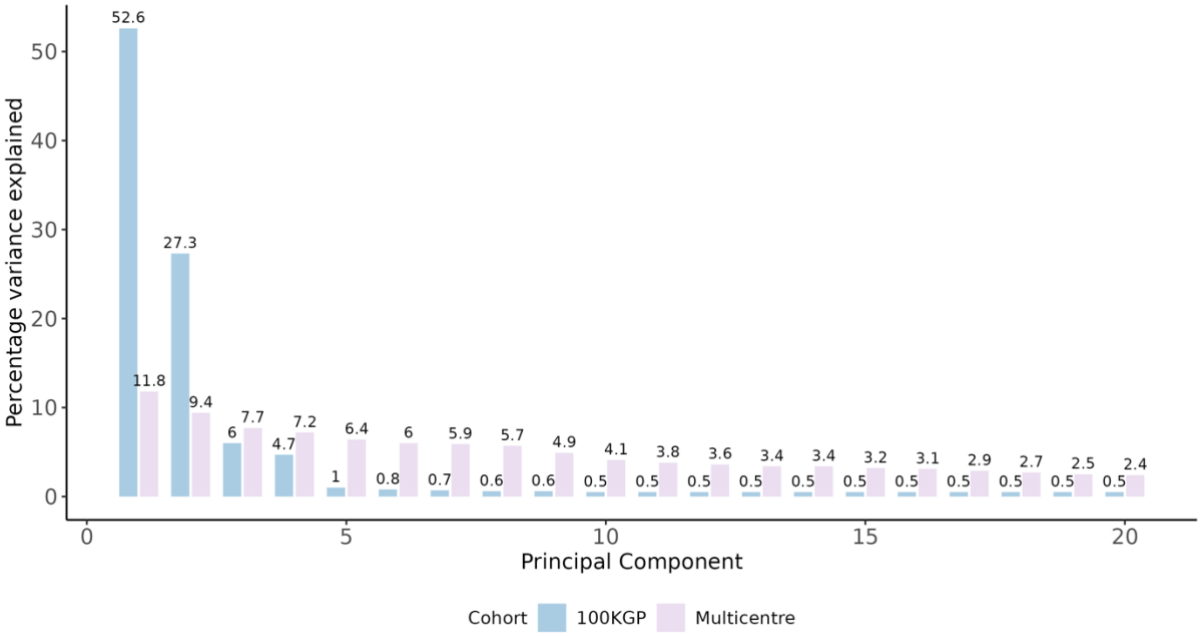

Figure 3.4: Scree plot depicting percentage of variance explained by nucPCs in each of the cohorts. The elbow pattern in 100kGP can be observed after the 2nd component, where levels decrease dramatically compared to the multicentre cohort where levels stabilise after the 10th PC.

3.3.2 Evaluating analysis software (FaSTLMM vs SAIGE vs REGENIE)

Considering the population structure as well as the presence of pedigree data in the multicentre cohort, the choice of software depended on which was most effective at correcting for these confounding factors. To do so, due to ease of access, data from the multicentre cohort was used to test the performance of three software. As presented in the table above, both SAIGE and REGENIE rely on generalised mixed models whereas FaSTLMM uses linear mixed models.

3.3.2.1 Does including principal components adequalty account for population stratification?

The incorporation of principal components as covariates is a common way to adjust for population structure, however, as discussed above, most software are designed to account for this using their own underlying methods. To test whether the addition of PCs, in addition to the methods they use to account for stratification, would present a sufficient correction (lambda ≤1.1), analyses were performed including 10 PCs as covariates (**Figure 3.5**). Inflation factors calculated for results from FaSTLMM and SAIGE GWASs, are below the accepted 1.1, which shows a sufficient correction for population stratification. Quantile-quantile (QQ) plots are used to compare the distribution of observed p-values from the association test with the expected distribution of p-values under the null hypothesis (no association). In **Figures 3.5-A and B**, this is aligned to the red, diagonal line, but shows a slight deviation at the tail; which reflects the SNPs with an increased association -log10(P value), that mirrors the observed skyscrapers within the Manhattan plot. On the other hand, REGENIE (**Figures 3.5-C**), has an over inflated lambda, which points at under-correction of population structure consequently, the QQ plot shows no alignment with the expected line.

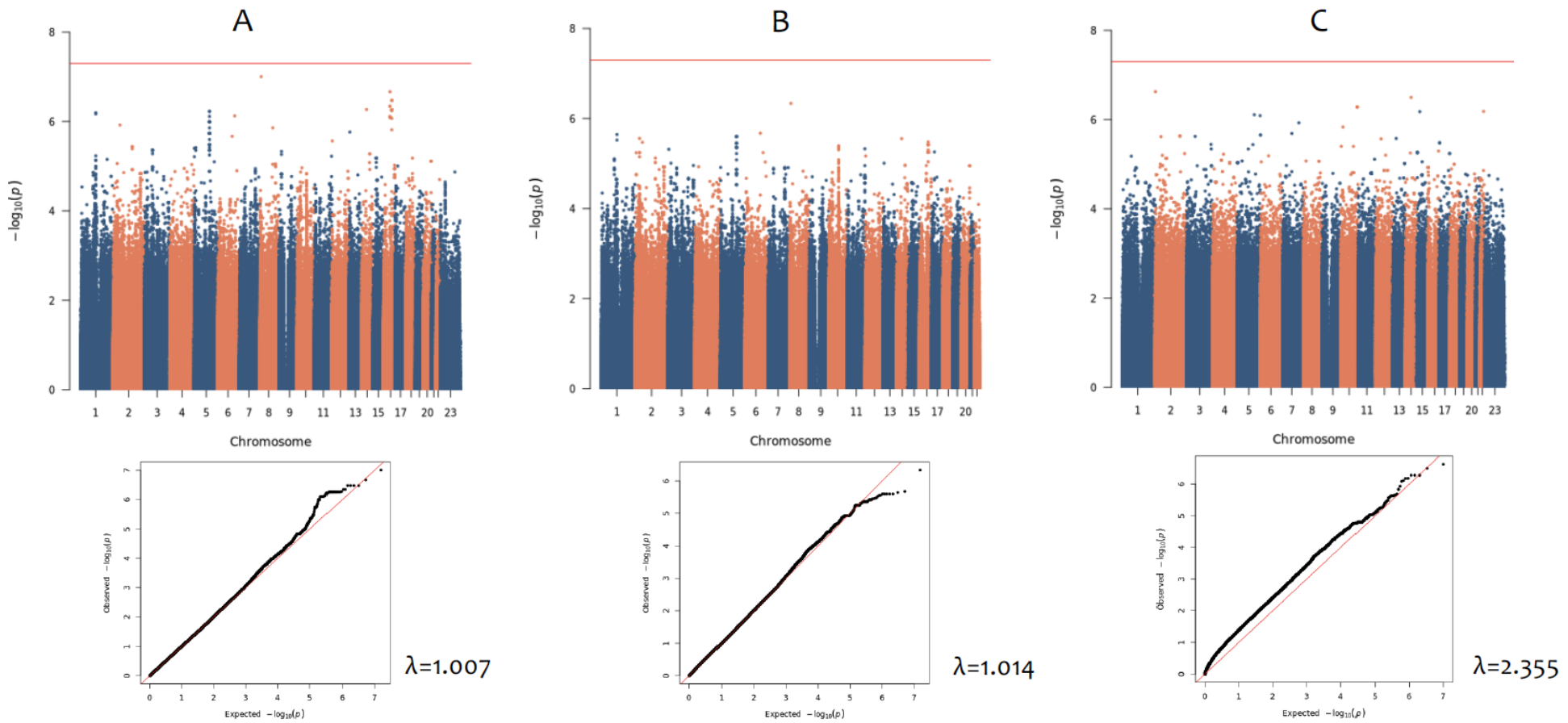

Figure 3.5: Multicentre GWAS results using first 10 nucPCs as covariates. (**A**) Analysis carried out on the full multicentre cohort (n=408) using FaSTLMM (**B**) SAIGE (**C**) REGENIE.

3.3.2.2 Given the non-normal distribution of the studied phenotype, would analyses yield similar results if modelled as a binary trait?

To test whether binary phenotype data would be a preferable input, multicentre m.3243A>G level data were converted into a binary format (as described in **Section 3.2.4**) (high (>0.8664982) and low (<0.534116982) levels) and used in GWAS (**Figure 3.6**). FaSTLMM and REGENIE showed the greatest difference upon using a binary phenotype, with an increase in inflation factors that indicated under correction. Given that FaSTLMM previously showed enough correction using 10 PCs, indicated that the increased inflation factor may be due to a reduced analysis power, especially that binary trait GWASs often require larger sample sizes compared to continuous trait GWASs (Visscher et al., 2017a). In a binary trait analysis, each individual provides less information about the underlying genetic variation compared to a continuous trait, where a range of values can better inform the analysis.

The observed under-correction built the reason to return to using a continuous trait whilst investigating into methods to account for PS.

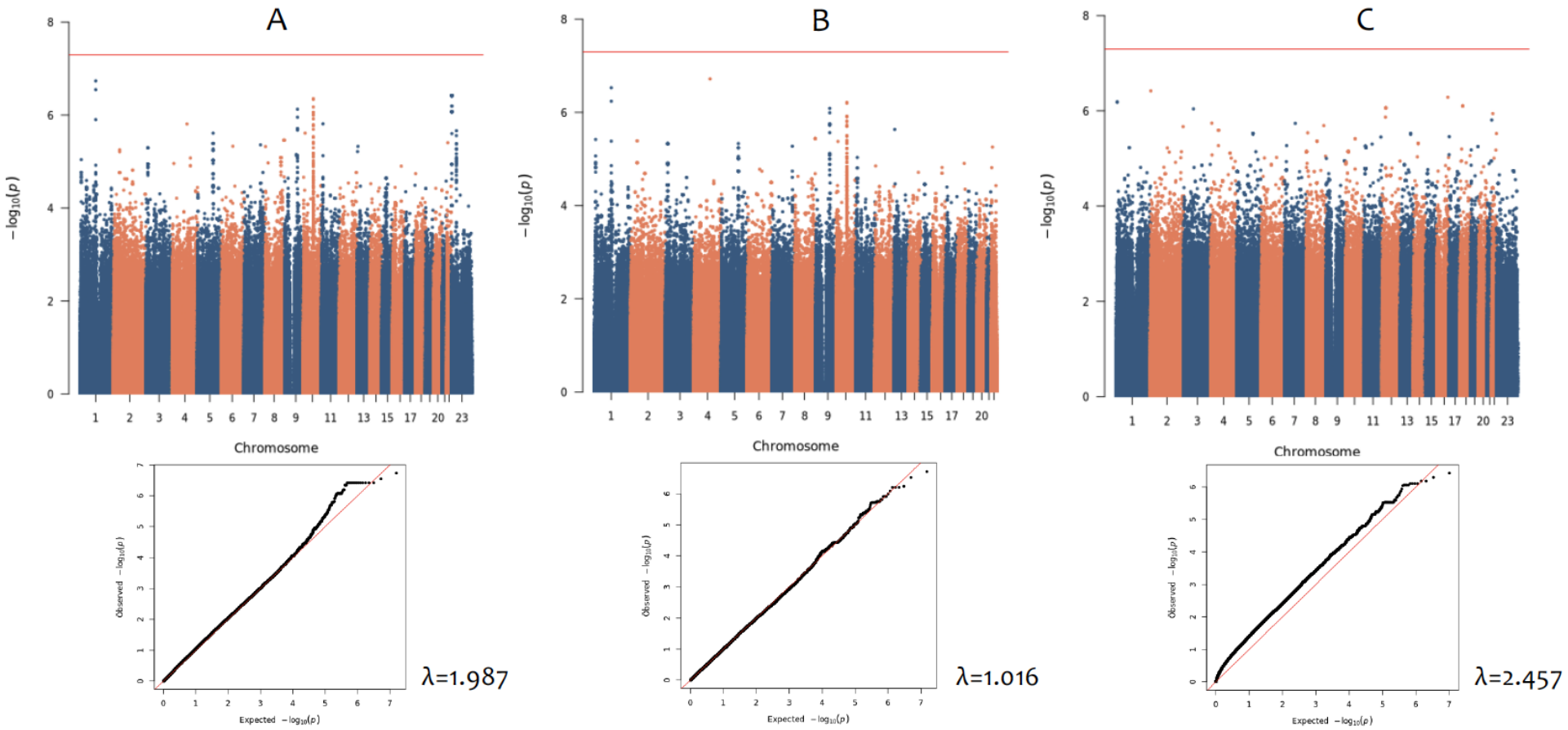

Figure 3.6: Multicentre GWAS results with m.3243A>G coded as a binary trait with 10 PCs. Figures depict Manhattan and a QQ plots retrieved from running the analysis on 408 individuals with a binary trait where m.3243A>G variant allele levels were converted into values of zeros and ones using (**A**) FaSTLMM. (**B**) SAIGE. (**C**) REGENIE.

3.3.2.3 Does excluding principal component outliers account for population stratification and lead to robust results?

Including PCs are a way of adjusting for PS however, as shown in **Section 3.3.2.1,** software like REGENIE yielded an over-inflated lambda. As a way to find an ultimate analysis design that would yield equally sufficient correction for PSs using different software, analyses were performed with a continuous phenotype without PCs as covariates yet excluding PCA outliers (n=24) (**Section 3.3.3.1**), which left 384 individuals for analysis (**Figure 3.7**). For the REGENIE software, the inflation factor decreased from 2.355 **(Figure 3.5)** to 1.042, indicating that the exclusion of outlier samples was a better suited analysis design for this software. The improved PS correction was observed in the Manhattan plot, where a Manhattan skyscraper pattern appeared as compared to the noise observed in design one. Both FaSTLMM and SAIGE showed a performance that was as good as with using covariates (**Figure 3.5-A and B**) where lambda inflation factors had almost no difference (FaSTLMM:1.007 to 1.008, SAIGE: 1.014 to 1.013).

To summarise, based on lambda inflation factors and the amount of random statistical noise in Manhattan plots, SAIGE had an overall best performance with data. Lambda inflation factors were within the accepted limits of ~ 1 – 1.10 in all designs, and QQ plots showed no extreme deviations from the null hypothesis. Manhattan plots had no noise and the majority of data, although non-significant, seemed to align into 'skyscraper' patterns across the chromosomes - which indicates robustness and absence of false positives. FaSTLMM showed good performance as well, however, a decision was made to opt against it due to its discontinuation of maintenance since 2019. On the other hand, REGENIE did not provide enough correction in the case of binary trait nor in the case of including PCs as covariates, with overly inflated lambda values. However, it presented a sufficient correction in the case of excluding PC outliers. REGENIE was the software of choice for our collaborators from Exeter, given that I was planning to perform a META analysis combining GWAS results from the UKBB and 100kGP, to avoid any discrepancy, if possible, I preferred to be consistent and use the same software on both cohorts. REGENIE showed a sufficient correction for confounding factors on the multicentre cohort when removing outlier samples, as a step before making the decision to retain REGENIE for my analyses, software performance had to be tested on 100kGP data. **Table 3.2** shows a summary of lambda inflation factors retrieved from all three designs.

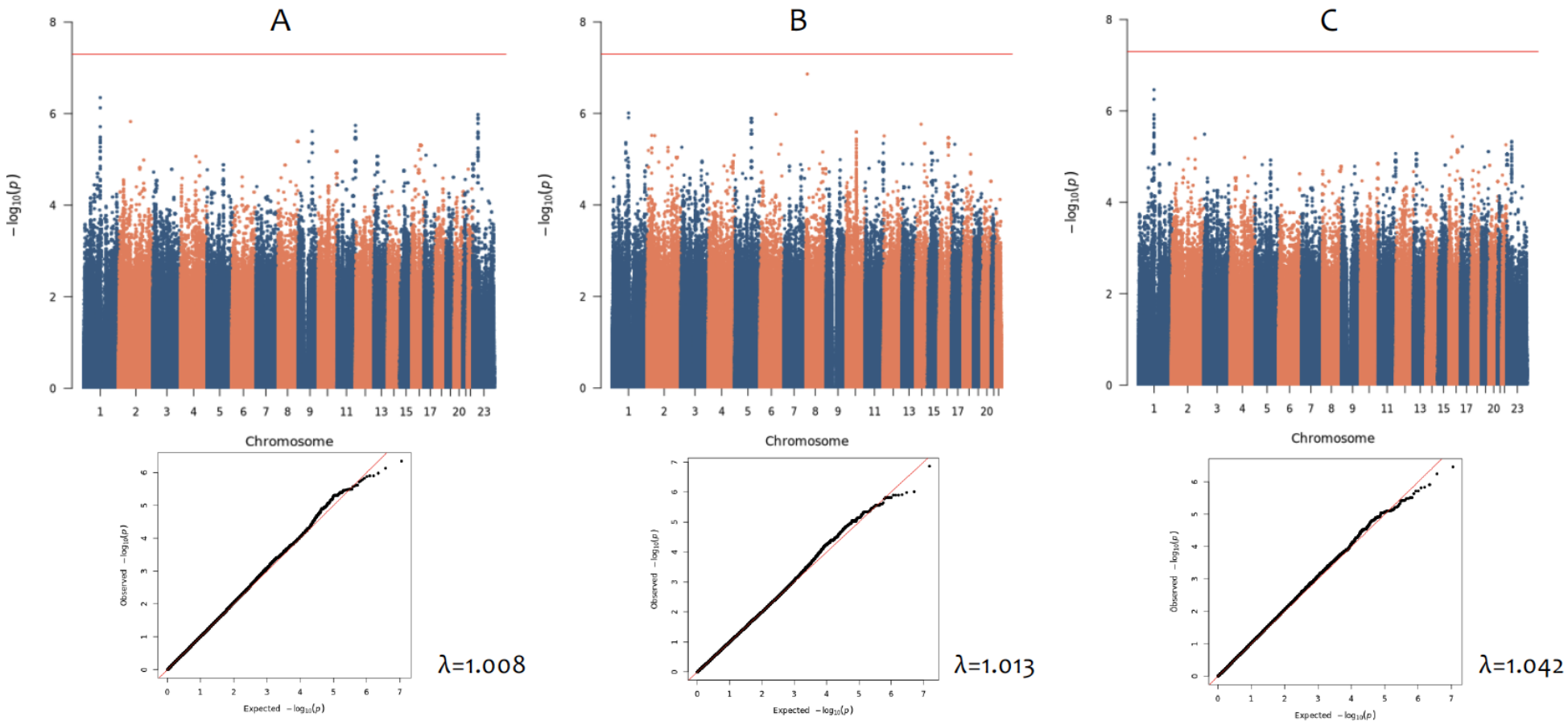

Figure 3.7: Multicentre GWAS results excluding PCA European population outliers with no covariates. (**A**) Principal component analysis outliers were excluded from the analysis (n=24) leaving 384 individuals. These results are retrieved from running FaSTLMM. (**B**) SAIGE. (**C**) REGENIE.

Table 3.2: Summary table presenting the lambda inflation factors retrieved from the three evaluated analysis designs.

	FaSTLMM	*SAIGE*	*REGENIE*
Design 1: Including PCS as covariates	1.007	1.014	2.355
Design 2: Excluding PCA outliers and no covariates	1.008	1.013	1.042
Design 3: With a binary trait	1.987	1.016	2.457

3.3.2.4 Does the chosen method also perform well in 100kGP data?

For its speed, efficiency in correcting for confounding founders, mixed model framework that allows the incorporation of both fixed and random effects, and robustness with large sample sizes, REGENIE was demonstrated to be the method of choice in large cohort studies such as the UKBB, Biobank Japan, Taiwan Biobank, and FinnGen (Bovonratwet et al., 2023; Chen et al., 2023). This applied for our Exeter collaborators who performed analysis in UKBB. A REGENIE GWAS test was carried out on 100kGP data using the third analysis design (after excluding PCA outliers) (**Figure 3.8**). REGENIE results on 100kGP show no noise in the Manhattan plot, and an acceptable inflation factor implying a good model fit, and correction for any possible confounders. Considering the good performance in both the multicentre cohort and 100kGP, the software proved to be a suitable choice for analysing the studied phenotype.

Figure 3.9 shows beta values retrieved from analysis design 3 using SAIGE and REGENIE. This positive correlation indicates that the direction of effect retrieved from both software is mostly the same, and the increase in BETAs is proportional. Therefore, although SAIGE seems to be the better choice as it shows consistent results when using both continuous and binary data, REGENIE is an acceptable alternative. As mentioned, to ensure the consistent usage of the same software across all three cohorts, this was the software retained for the remainder of this project.

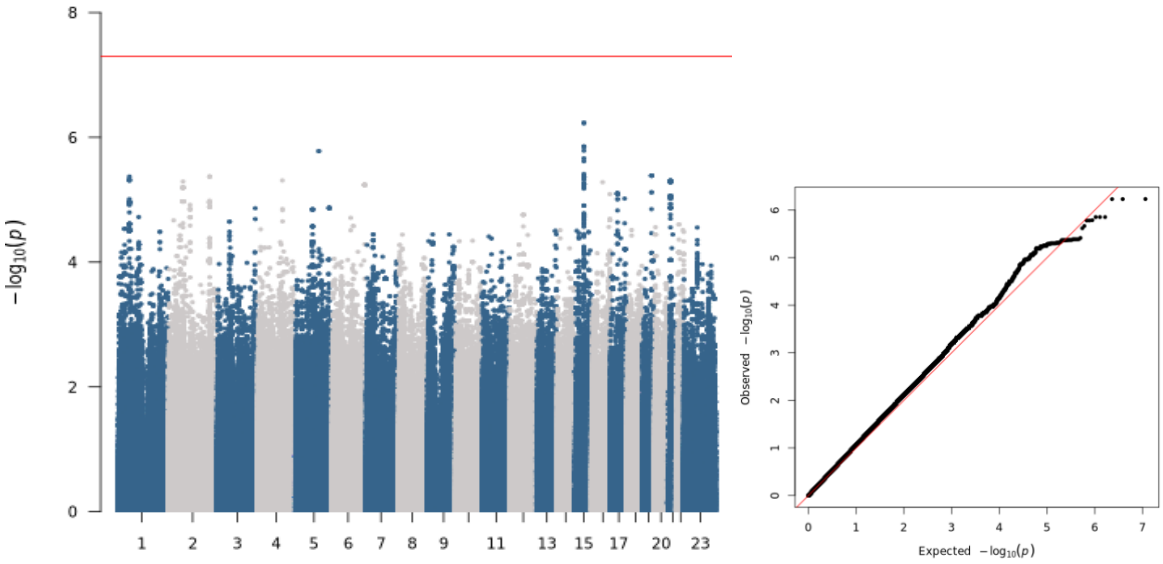

Figure 3.8: Results of REGENIE GWAS on 100kGP data. Manhattan plot shows a peak approaching significance on chromosome 15. QQ plot only deviates at the tail, and the lambda inflation factor is below 1.10 (1.094).

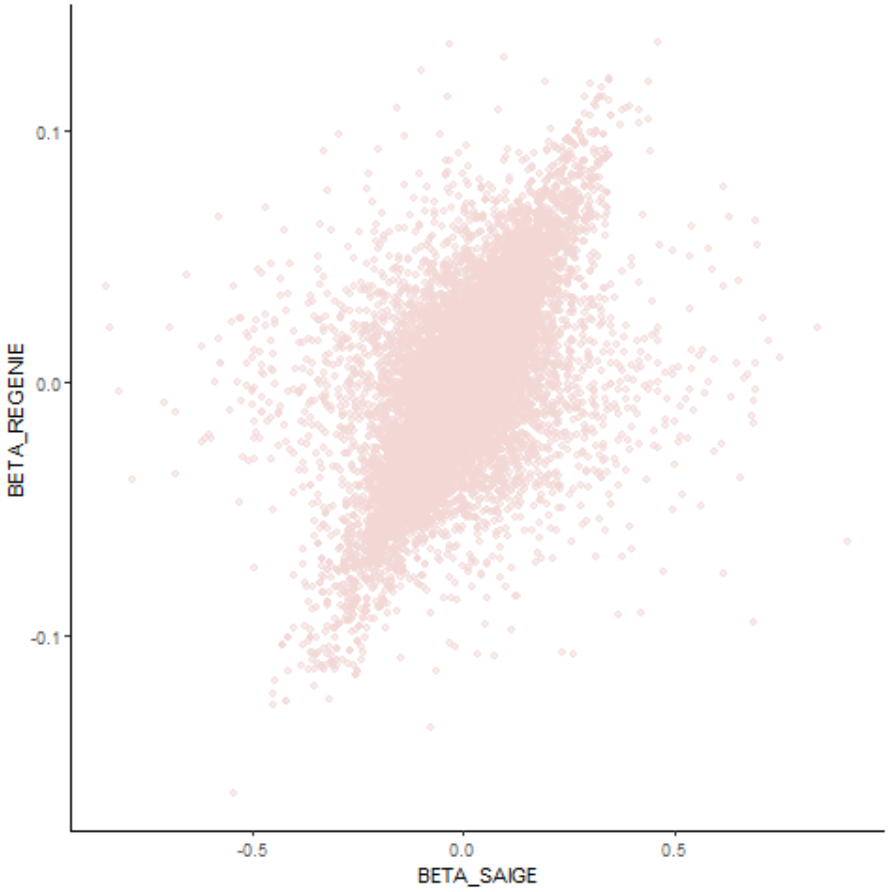

Figure 3.9: BETAs from running analysis with REGENIE vs SAIGE. Results retrieved from running analysis including PCs as covariates after excluding European population PCA outliers on the multicentre cohort.

3.4 Discussion

The comparison of multiple software on the ascertained, multicentre cohort demonstrated that both generalised mixed models and LMMs were efficient at accounting for GWAS confounders. Specifically, SAIGE showed the most consistent results in every analysis design. REGENIE on the other hand, provided an acceptable lambda in the analysis design that excluded outlier samples.

I decided to use REGENIE on all datasets to ensure consistency, particularly that it was the software of choice by our Exeter collaborators on the UKBB cohort, and I intended to combine the UKBB and 100kGP data using a META analysis.

The difference in software performance may be attributed to the distribution of data, as normally distributed data is easier to model making it less error prone. Something that I have tested (using inverse normalisation) however, decided not to use with either of the software; as the GWAS yielded peaks at completely different locations compared to before normalisation, indicating that the data were distorted. Normalising the data removes its natural variability by forcing it to follow a pattern against its 'nature' and losing its biological meaning. Several reports in literature support this and advise against normalising phenotype data ahead of GWA analysis, particularly that LMMs and GMMs are robust to different phenotypic distributions (Yang et al., 2010b; Zhou and Stephens, 2014).

Up to date, REGENIE is the only software that has been translated into a Nextflow pipeline (Schönherr et al., 2024b), which is a powerful and versatile workflow management system designed to streamline and enhance the execution of complex computational workflows (Di Tommaso et al., 2017). This is key for biobank-scale analysis, including GWASs. By automating the pre and post analysis procedures, the whole process becomes much more reproducible, and time-efficient, which is key for researchers! The fact REGENIE was the software of choice is an indication of its thoroughly tested efficiency. In support of this, Mbatchou et al., (2021b) compared the accuracy of effect sizes estimated by SAIGE and REGENIE software, degree of accounting for LD structure, analysis speed, and ability to analyse multiple traits at a time, on UKBB data. REGENIE with its underlying ridge regression and LOCO function

(discussed in **Section 3.2.2**) proved to be more efficient. Having said so, developers of REGENIE acknowledge that in studies of small sizes with high levels of relatedness, the efficacy of REGENIE decreases as it becomes overly conservative and is thus, not recommended for such cases (Zhou et al., 2018b).

PC analyses demonstrated the presence of population structure in both cohorts and confirmed the necessity to account for this. The exclusion of outliers, rather than including PCs as covariates, presented with overall lower inflation factors (for all three software).

Before excluding the usage of PCs as analyses covariates, I tested whether a better correction could be achieved upon the inclusion of additional PCs; this was done by performing analyses with 15 and then 20 PCs. Given the observed small percentage variance explained by each PC and its mild decrease after the assigned elbow at the 10^{th} PC (**Figure 3.4**), their addition led to an inflation factor well below 1, indicating over correction. The addition of further PCs was later found to be isolating the individual pedigrees, and so were dependent on pedigree size. Given that all used software account for family structure, this was driving the observed over-corrected inflation factors. These factors deemed the exclusion of outlier samples as the more appropriate solution in this case.

Large cohorts present several challenges for conducting GWAS analyses, including: relatedness in samples, imbalance in case-controls, as well as the sheer scale of the data and its computational footprint. Both SAIGE and REGENIE have proved to be successful in large studies, in the UKBB one-third of the individuals are third degree (cousins) or even closer relatives (Bycroft et al., 2018b), and REGENIE was the method of choice in multiple studies on UKBB data (Zhu et al., 2023; Schönherr et al., 2024b) as it uses GRM to efficiently account for that. It is also known that 66% of probands were recruited with family members in 100KGP (Hocking et al., 2023) and REGENIE was also used in multiple studies (Jadhav et al., 2023; Zheng et al., 2023).

Having had tested the performance of different software on the studied phenotype, and establishing the most efficient method of adjusting for the underlying confounding confounders, led the way into performing the analysis on the remaining cohorts used in this project.

Chapter 4. GWAS and follow up analysis results

This chapter will present GWAS results obtained from all three cohorts using the methodology described in **Chapter Three**, specifically, employing REGENIE software after excluding PCA outliers. Following the introduction of each analysis, the chapter will also discuss and present results from subsequent analyses, including META, fine mapping, and SNP-based heritability estimates.

4.1. Introduction

4.1.1 GWA analysis

Ever since the first publication of a genome wide association study in 2005 (Klein et al., 2005a), 72,014 genotype-disease phenotype associations with a $P < 5 \times 10^{-8}$ significance have been made (The GWAS Catalogue - June 2024) (Sollis et al., 2023), demonstrating great success at the identification of novel disease associated genes (Tam et al., 2019).

With the aim of understanding disease variation, a team from Harvard and Queensland universities used inverse variance methods on data from an insurance company called Aetna and investigated 560 diseases in 44 million individuals in the US (Lakhani et al., 2019). Out of that sample, 56,396 were twins and another 724,513 were siblings. On average, they concluded that 60% of disease variance can be attributed to genetic factors (nature), whereas the remaining 40% are explained by environmental factors (nurture). In a quest towards a better understanding of the genetics behind disease, in the past 15 years, GWA studies, on many occasions, have realised their promise and identified disease associated loci. Crohn's disease, type 2 diabetes, cardiovascular diseases in addition to schizophrenia, have all been areas in which GWAS led to significant progress (Abdellaoui et al., 2023). As more national cohorts and resources such as The UK BioBank emerged, it was estimated that in the past five years, the average sample size used in GWAS publications have more than tripled (Abdellaoui et al., 2023). This, along with the increase in WGS data, genotyping array sizes, imputation technologies and software tools that have improved GWAS, has naturally led to an increased ability to identify more variants in both diseases and complex traits.

In the context of this work, up to date, our Newcastle based multi-centre cohort presents a uniquely large cohort of m.3243A>G carriers. However, compared to other diseases and traits, a GWAS on ~408 individuals is underpowered to detect variants with a small effect size. To increase sample size, two large publicly available cohorts: UKBB and 100kGP (Genomics England) were included. Based on population carrier rate of 140~250 in 100,000 (Manwaring et al., 2007a), we estimated to identify approximately ~280-500 (out of 200,000) and ~86-153 (out of 61,000) additional m.3243A>G carriers from each cohort, respectively.

4.1.2 META analysis

As discussed in **Chapter Two**, META analysis allows the aggregation of results from multiple independent studies, and there are multiple methods to choose from to conduct a META. The aim is always to reduce false positives and increase statistical power to detect the most modest effects from a GWA study, or in case there are any findings, it can also be used to assess their consistency across studies. In fact, it is considered as a routine part of GWAS (Begum et al., 2012). The largest type 2 diabetes META was conducted in 2018, combining 32 studies (Mahajan et al., 2018). This included 74,124 cases and 824,006 controls from European genetic ancestries. The study performed a GWAS on combined data and identified 231 significant loci (Mahajan et al., 2018; DeForest and Majithia, 2022). Another meta-analysis examined the association between mtDNA copy number and cardiovascular diseases (CVD). This combined five studies that in total analysed 8,252 cases and 20,904 controls, results indicated that variations in mtDNA copy number could serve as potential biomarkers for predicting the risk and prognosis of CVDs (Yue et al., 2018). Compared to the prevalence of T2D (6059 cases per 100,000) (Khan et al., 2019) and other common polygenic disorders, the prevalence of m.3243A>G carriers ranges from 140 to 250 in 100,000 (Manwaring et al., 2007a). The disparity in prevalence dictated the sample size available to perform this analysis, and that is how META appeared to be a logical solution – given the impracticality of combining raw data to perform one large GWAS, META allowed the combination of multiple study results, by that overcoming the small detection power stemming from small, individual study sample sizes.

There are many caveats in a META analysis. Key caveats to consider include: publication bias; where studies with a positive result are more often published, consequently leading to an overestimation of the effect sizes (LeLorier et al., 1997). Selection bias is also a large caveat where studies considered should be similar in respect to criteria such as, population studied, sample size, study design and objective (Marsoni et al., 1990). Where significantly larger sample sizes will inevitably have a stronger influence on the results, and a high 'heterogeneity' in combined samples can reduce the reliability of the pooled effect size, making it difficult to draw clear conclusions (Walker, Hernandez and Kattan, 2008).

4.1.3 Fine mapping

GWAS and META analysis have been successful in identifying thousands of loci associated with various diseases (Visscher et al., 2017b). However, linkage disequilibrium (LD) patterns add a layer of complexity when it comes to refining these results and pinpointing directly at the associated loci. Multiple additional factors influence the performance of fine mapping, things such as: sample size, SNP density, and the number of causal SNPs in a region and their effect size (Schaid, Chen and Larson, 2018). Typically, the greater the SNP density and sample size is, the easier it is to accurately determine LD structure; hence, the greater is the power to elucidate causal variants at a higher resolution.

Fine-mapping software compute and output posterior inclusion probabilities (PIP), which is a way to quantify the probability of a certain variant being the one leading the association signal, by that reflecting the degree of uncertainty caused by LD (Cui et al., 2024). Similar to GWAS and META, there are different methods to choose from to conduct fine mapping (discussed in sections to follow). The accuracy of such analysis in general, is heavily dependent on analysis calibration – where the model used for retrieving genetic effects and the prior for genetic architecture is correctly modelled (Ulirsch, 2022). In an accurate fine-mapping setting, a PIP of > 90%, should indicate that 9 out of 10 variants are truly causal (Cui et al., 2024). Which placed emphasis on the importance of choosing the right software for these data.

4.1.4 SNP heritability estimates

In 2014, a META study performed by the Psychiatric Genomics Consortium identified 108 independent loci associated with schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). However, there was a huge disparity between the expected percentage of heritability compared to what was obtained from SNP based heritability analysis (64% compared to 3%). Once again, drawing the attention on the issue of missing heritability, in which estimates retrieved from classic heritability designs, are significantly higher than those explained by genetic variants (Mayhew and Meyre, 2017). Classic heritability estimates use family, or twin data to estimate genetic similarity between individuals based on their relatedness. The output of that informs us about the degree genetics contributes to the phenotype, it however, does not provide insight about the actual underlying architecture (Zhu & Zhou, 2020).

GWASs are best at detecting common variants with low effect size, having said so, in the case of missing heritability scenarios, just like the extremely polygenic schizophrenia, despite the large sample sizes, it is likely that there were more variants that were not picked up by the GWAS as significant, due to different reasons such as, their extremely small effect sizes, or the possibility that some SNPs were not genotyped due to their low frequencies (Visscher, Hill and Wray, 2008). This means that the 108 identified loci encapsulate only a small fraction of risk variants, that in total cover a small proportion of the expected heritability. Using variance components methods, Pickett et al., (2019) and colleagues used family pedigrees and estimated that the heritability of m.3243A>G is ~72%. To compare this with SNP-based heritability, GWAS and META analyses were performed with the aim of identifying associations with the variability of m.3243A>G levels, the output of these analyses provided the necessary inputs for SNP-based heritability calculations.

4.1.5 Significance thresholds

To identify potential associations, it is necessary to determine a significance threshold. Bonferroni correction stands behind the generally accepted genomic significance threshold of $p \leq 5 \times 10^{-8}$, which aims to correct for type 1 errors that arise from multiple testing that is characteristic in a GWAS, where typically more than a million SNPs are used. Despite the growing criticism, as it is thought to be very stringent leading to an increase in false negatives (Kaler & Purcell, 2019), this remains to be the generally accepted GWAS significance threshold that was employed in this chapter (Uffelmann et al., 2021). Bonferroni corrections assume that genetic data follow an independent nature which, given LD, we know is not true. Such assumptions are one of the reasons that make this correction over conservative (Johnson et al., 2010). An alternative for this has been an extension of the false discovery rate (FDR), q value, which is the expected proportion of false positives obtained when calling that feature (SNP) significant (Storey, 2002). A p value instead, is a measure of the false negatives when calling a specific feature significant. Some suggest that q values provide a more direct measure of the significant findings themselves as compared to those around it, which makes reporting q values more practical (Storey and Tibshirani, 2003). Nonetheless, p values remain to be the significance measures routinely reported in GWA studies (Reed et al., 2015; Aguilar et al., 2019).

4.2. Methods

4.2.1 Studied cohorts

As was outlined in **Chapter Two**, this project utilises data from m.3243A>G carriers and the identified obligate carriers from three different cohorts. GWAS conducted after the exclusion of non-European individuals identified from nuclear DNA principal component analysis, was found to yield the best correction for confounding factors (outlined in **Section 3.3.1**). The total number of outliers detected in the multicentre cohort was 24 (of 408 samples), 28 (of 164) in 100kGP data, and 4 (of 147) in the UKBB. As a result, the remaining sample sizes for the GWA studies were 384, 136, and

143 across the three cohorts, respectively. Data in each cohort underwent GWA analyses, and subsequently, using GWAS summary statistics, META analysis was performed. Due to the clinically ascertained nature of the data in the multicentre cohort, which resulted in a significantly different distribution of variant allele levels compared to the two population cohorts (see **Section 3.5**), I decided to exclude this cohort from both the META and fine mapping analyses. SNP heritability estimates however, were performed on META analysis output that combined 100kGP and the UKBB data, and separately, on the multicentre cohort data.

4.2.2 GWAS analysis

As was discussed in GWAS optimisation chapter (**Chapter Three**), REGENIE was the software of choice to conduct analysis on all three cohorts. As concluded from that same chapter, excluding PCA outliers resulted in a more optimal GWAS performance compared to including PCs as covariates. The phenotype used in all analyses was age-adjusted m.3243A>G levels (**Section 2.2.3**). Ahead of running the analyses, data QC was performed as outlined in **Section 2.4.3.B.**

4.2.3 META

When there is access to genotyping/ sequencing data from different study populations, then data can be directly combined, and mega-analysis are performed. In this case however, the export of such data was not possible due to data privacy regulations and GWAS summary statistics from 100kGP and the UKBB were combined via a META analysis instead. Together, this yielded a sample size of 279. GWAMA (v2.2.2) was the software of choice, where GWAS summary statistics were the input and --quantitative flag was used, while leaving the rest to default. It is worth noting that a fixed effects META is the default in GWAMA. In the context of this study, GWAMA was the software of choice due to several features as it: (1) accounts for in-between-study variation and population structure, and (2) makes analysis possible irrespective of the used array, by aligning data on the same reference strand which uses data from the Hap-Map and 1000 Genomes Project (Mägi and Morris, 2010b). To

decide on whether to perform a fixed, or random effect META, a between-study heterogeneity estimate is often useful. That is because a fixed effects META analysis assumes homogeneous allelic effects between studies, in case there is an increased I^2 estimate, then this assumption is inaccurate, and a random effects META should be used. This approach, first outlined in 1986 by DerSimonian & Laird, allows effect sizes (per SNP) to be different across studies, and instead, it utilises a generalised weighing method that considers the characteristics of each individual study, as well as the between-study heterogeneity.

4.2.4 Between study heterogeneity estimates (I^2)

As was discussed, between-study heterogeneity is a measure necessary to determine which method is most suited for conducting a META analysis. There are different causes for between-study heterogeneity, particularly when studies are performed by different teams where (1) different QC measures are applied, and (2) analyses were performed on different populations using different data analysis methods. The most classic measure for heterogeneity is Cochran's Q - this is obtained by summing the square of differences between each individual study and the pooled effect across studies at each tested SNP (Deeks JJ, Higgins JPT and Altman DG, 2023).

Like a chi-squared distribution, it assesses the degree of difference between individual study effects and the pooled effect across studies (Mägi and Morris, 2010b). However, it depends on the number of studies considered and thus, has a low power with small studies (Gavaghan, Moore and McQuay, 2000).

Conversely, in 2002, Higgins & Thompson proposed the widely popular I^2 measure which reflects the degree of variation due to heterogeneity (rather than chance or sampling errors) regardless of the number of studies (Mägi and Morris, 2010b). This is calculated using the formula below, and is a summary statistic provided by GWAMA software as a way to test whether the chosen (fixed effects META) approach is suitable or not:

$I^2 = \left(\frac{Q-df}{Q}\right) * 100\%$ where: Q: the chi-squared statistic

df: degrees of freedom

4.2.5 Power analysis

Estimation of power in study designs is an important step to determine the reliability, as well as the reproducibility of study findings (Kumle, Võ and Draschkow, 2021). To determine the sample size needed to achieve a desired power, or to calculate the power of a study with a known sample size, an accurate specification of the used study design is essential. What is trickier is the ability to determine the closest underlying biological model of the study. It is often that we assume a certain biological model that is still unknown in practice (Lettre, Lange and Hirschhorn, 2007). Most power calculation tools offer recessive, dominant, and additive genetic models. For this analysis, an R package called GENPWR (v1.0.4) was utilised (Moore, Jacobson and Fingerlin, 2019). A mis-specified model can lead to a loss of power, GENPWR allows robust calculations even under model misspecifications by increasing the degrees of freedom from the generally accepted 1 to 2 (Moore, Jacobson and Fingerlin, 2019). For these calculations an additive model was assumed, where the effect of having two risk alleles is twice as much of having one allele. This is also the model of choice in most GWAS analysis - mostly due to the fact this model assumes independence between alleles where the effect of each allele is independent of the presence of another allele (Zavala et al., 2011).

Effect sizes are essential in interpreting GWAS results, they reflect the strength of correlation between an identified SNP and the disease/phenotype of interest (Politi et al., 2023). The most common measure of effect size is Cohn's d where he classed effect sizes as follows: small (d = 0.2), medium (d = 0.5), and big (d ≥ 0.8) (Cohen, 1988). An identified association may be significant, yet with a trivial effect size, which if to quote Cohn's words: "The *primary product of a research inquiry is one or more measures of effect size, not P values*" (Cohen, 1992). Such findings, however, may be of a cumulative/ additive effect, which is the expected outcome of the GWA studies in this project. Large sample sizes are needed to detect variants with a small effect size, as well as variants with a low minor allele frequency (MAF), that is an estimate of the prevalence of the less common allele in a particular population (Politi et al., 2023).

4.2.6 Fine mapping analysis

Most available fine mapping software require original genotype-phenotype data to perform the analysis. In case of conducting the analysis after META, it is impractical and, in some cases, impossible to do so. Other software, such as, CAVIAR and PAINTOR rely on summary statistics however, their methods assume a random maximum number of causal variants in a locus, in CAVIAR for example, the default is six (Hormozdiari et al., 2014). Another caveat to consider is the way effect sizes are modelled. An explicit assumption that is made by methods such as CAVIAR, is the normal distribution of effect sizes in each locus, which is not always true (Hormozdiari et al., 2014). The output of almost all fine mapping methods is a list of SNPs with posterior inclusion probability (PIP), where a PIP approaching one indicates causality.

In Bayesian methods, this is calculated using the formula below, which requires a flat/ prior probability. In this case, the prior probability is that all SNPs are causal:

$$P(S_j causal \,|data) = \frac{\left(P(data|S_j\ causal) *\ P(S_j causal)\right)}{\sum_k P(data|S_k\ causal) *\ P(S_k causal)}$$

Where $P(S_j causal\ |data)$ is the probability of SNP j being causal given the data, and $P(data|S_j\ causal)$ is the probability of the data given that SNP j is causal, and $\sum k$ being the sum of all possible causal configurations. In case only one SNP out of k SNPs was allowed to be causal, then there would be only k possible models/ configurations. Whereas, if more variants are allowed, the possible configurations increase exponentially. This is when strategies such as, limiting the number of causal variants in a locus, or shotgun stochastic search (SSS) algorithms (explained below) step in.

A more recent software with a similar Bayesian statistical model, but different computational algorithm is FINEMAP (Benner et al., 2016). Just like the previously mentioned software, it requires summary statistics and SNP correlation data, that can be either from reference data such as, HapMap (Thorisson et al., 2005), 1000 Genomes Project (Auton et al., 2015), or using GWAS Z scores ($Z =$

$Effect\ size(\beta)/Standarad\ error\ (SE)$); the latter is preferable as it ensures that no data are lost. FINEMAP investigates the most likely configurations in each association area, without setting a maximum number of potential causal variants in each locus. Additionally, it makes use of per SNP effect sizes retrieved from GWAS summary statistics (Benner et al., 2016). Another common output of fine mapping is credible sets, which is a set of variants that contain the causal variant with a 95% probability; this output is achieved by adding variants with the highest PIPs until the total is equal to 0.95. The smaller the credible set is the better, as it provides higher resolution (Schaid, Chen and Larson, 2018).

FINEMAP is computationally more efficient due to the fact it relies on shotgun stochastic search (SSS) algorithms, which is a search approach used in regression models. It outperforms similar traditional methods such as Markov Chain Monte Carlo method (MCMC) in that it records multiple candidate models in parallel at each single iteration, compared to sequentially moving from one model to another which does not fully exploit the data (Hans, Dobra and West, 2007). The flags employed in running the software were –sss to indicate the shotgun stochastic search, as well as –dataset 1 since the input was a single META analysis. Additionally, to set the maximum number of casual SNPs, --n-causal-snps was used, once using five and another time using one.

4.2.7 SNP based heritability estimates

To choose a software for this analysis, options were narrowed down to those that can use summary statistics as an input; since retrieving individual-level genotypes was impossible due to data privacy, computational, and transfer restrictions.

There are nine heritability models, and each one has a different way of describing how much heritability each SNP is expected to contribute (Tang, Wang and Zhang, 2022). Ahead of running the analysis, it is key to choose a model; Bayesian variable selection regression (BVSR) was the first model used in heritability estimates (Zhu & Zhou, 2020). This assumes that only a small portion of SNPs contribute to the phenotype, and it assumes a point normal distribution of effect sizes across SNPs,

where only a fraction of SNPs will have an effect size (ES) of one, with the rest denoted as zeros (Guan and Stephens, 2011). A linear mixed model is one of the most common models used for heritability estimates, where all SNPs have a non-zero ES estimate and rather follow a normal distribution, this is applied in software such as GEMMA (Zhou and Stephens, 2012) and GCTA (Yang et al., 2011).

More recently, different software have been developed with an attempt to create the most biologically accurate model. Unlike the aforementioned models, linkage disequilibrium adjusted kinships (LDAK), which is the model underlying the software of choice, SumHer (Speed and Balding, 2019), considers both LD and MAF data in estimating the per SNP ES. Given the underlying population structure in the data, considering LD when measuring heritability ensures more accurate results. Analysis required two input files, the (GWAS/META) summary statistics file, and a tagging file, found online (https://dougspeed.com/calculate-taggings/); which allows the software to estimate the expected heritability compared to that in a reference dataset (~1.07 million common SNPs from 1000 GENOME project phase 3 study)(Auton et al., 2015).

4.3 Results

4.3.1 Power analysis

Power calculations showed that there was 95%, 66%, and 60% power to detect an association ($p < 10^{-8}$ / $\alpha \approx 0.05$) to variants with MAF ≥ 0.05 and ES ≥ 0.6, in the multicentre cohort, 100kGP, and UKBB cohorts, respectively. By combining 100kGP and UKBB cohorts (n=279) via a META analysis; power is increased to 87% assuming the same, medium effect size of 0.6 and a large minor allele frequency (>0.05) (**Figure 4.1**). In this analysis, there is sufficient power to detect common variants with medium to large effect sizes (Cohen's d: ES ≥ 0.5) however, power decreases drastically when it comes to small effect sizes. Where to detect variants with MAF ≥ 0.05 and ES ≥ 0.3 it decreases from the values mentioned above to 68%, 38% and 32% in the multicentre cohort, 100kGP, and UKBB cohorts, respectively.

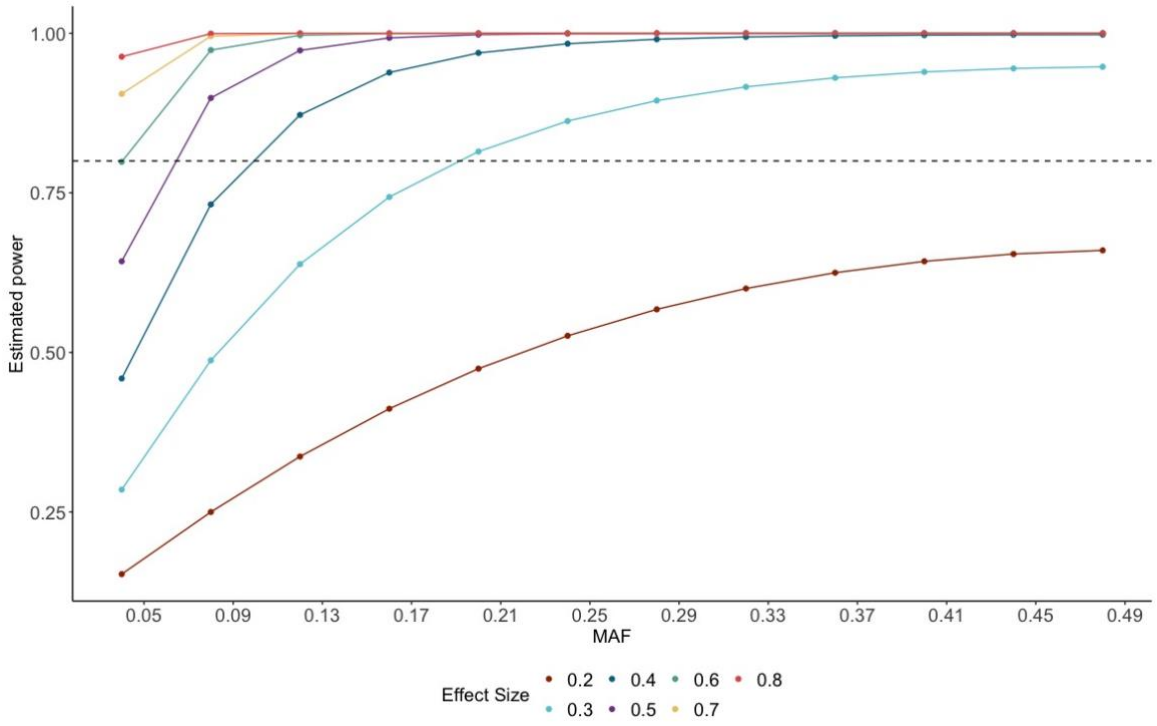

Figure 4.1: Power calculations using GENPWR. Plot reflects the power to detect variants with different MAFs and ESs using 279 samples. Dotted line indicates 80% power.

4.3.2 m.3243A>G variant allele levels

As previously mentioned, GWAS was performed on age-adjusted blood m.3243A>G levels as the phenotype (**Figure 4.2**). As detailed in **Chapter Two**, the multicentre cohort is a clinically ascertained cohort of m.3243A>G carrier individuals. Age-corrected variant levels fall within the same range of 0~1 across all three cohorts. However, the clinical ascertainment for individuals with mitochondrial disease diagnoses in the multicentre cohort explains the increased number of individuals with high variant levels (typically severe disease) in the multicentre cohort (blue in **Figure 4.2**).

Additionally, the 100kGP cohort (pink) shows a larger number of individuals with low variant allele levels compared to the UKBB. This discrepancy is attributed to the fact that the 100kGP identified 60 individuals as obligate carriers; before age correction, those were assigned a minimum allele frequency of 0.01, on the other hand, only three obligate carriers were identified in the UKBB. This discrepancy can also be explained by an overall lower number of individuals retrieved from the UKBB.

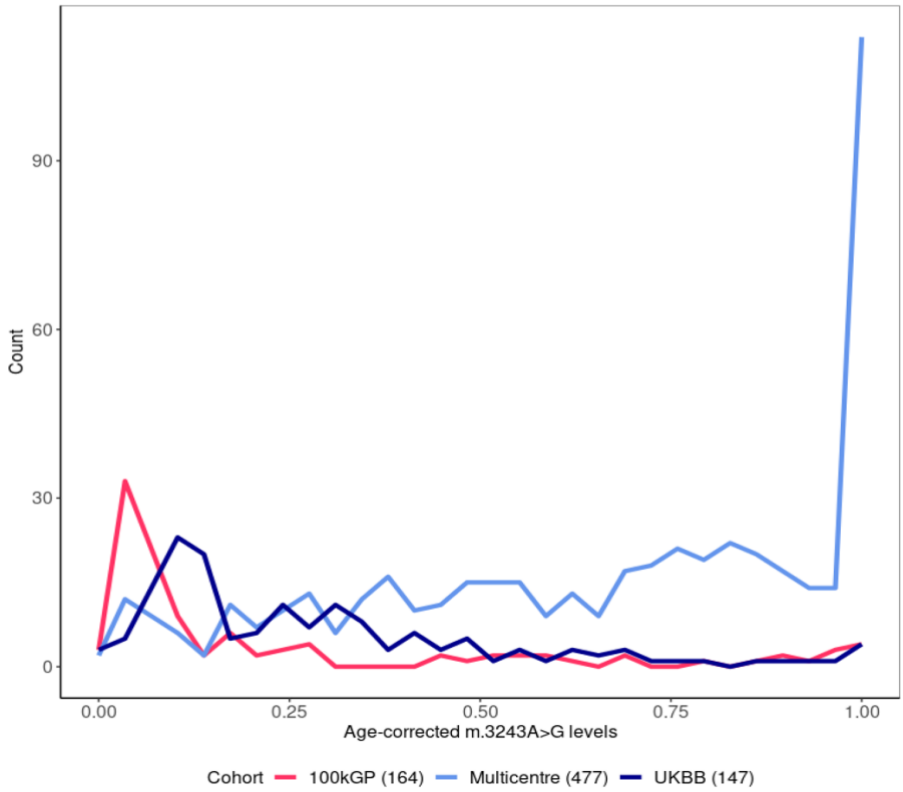

Figure 4.2: Age-corrected m.3243A>G variant levels in all three cohorts. Due to ascertainment in the multicentre cohort (blue), variant levels are skewed towards the higher levels compared to the other two cohorts. Individual counts with small variant levels in 100kGP (pink) are larger than those in the UKBB (navy), and that reflects the greater number of obligate carriers identified in the 100kGP cohort.

4.3.3 GWAS results

None of the three performed studies yielded a significant peak (**Figure 4.3**). However, peaks above the suggestive significance threshold of 5.3 were found in all three cohorts and they were as follows: in the multicentre cohort, a peak on chromosome one (lead SNP; 1:114542914A>T; -log(Pval) = 6.4); 100kGP with a peak on chromosome 15 (15:62868505G>A; -log(Pval) = 6.2), and the UKBB on chromosome eight (8:128594837c>c; -log(Pval) = 5.6). QQ plots show a good alignment with the red, expected line, with no deviations at the origin, indicating an appropriate correction for confounding factors, and a slight deviation at the tail, which reflects the absence of SNPs above the genomic significance threshold. Lambda inflation factors, which are calculated by dividing the median of the first quantile in the chi-square distribution over 0.456, that reflects the median of the distribution under the null

hypothesis (that is in the absence of inflation), are all below 1.1; both of which are indicatives of adequate population correction, and appropriate choice of software. **Table 4.1** shows a list of coding genes surrounding the GWAS peaks of suggestive significance in each of the cohorts.

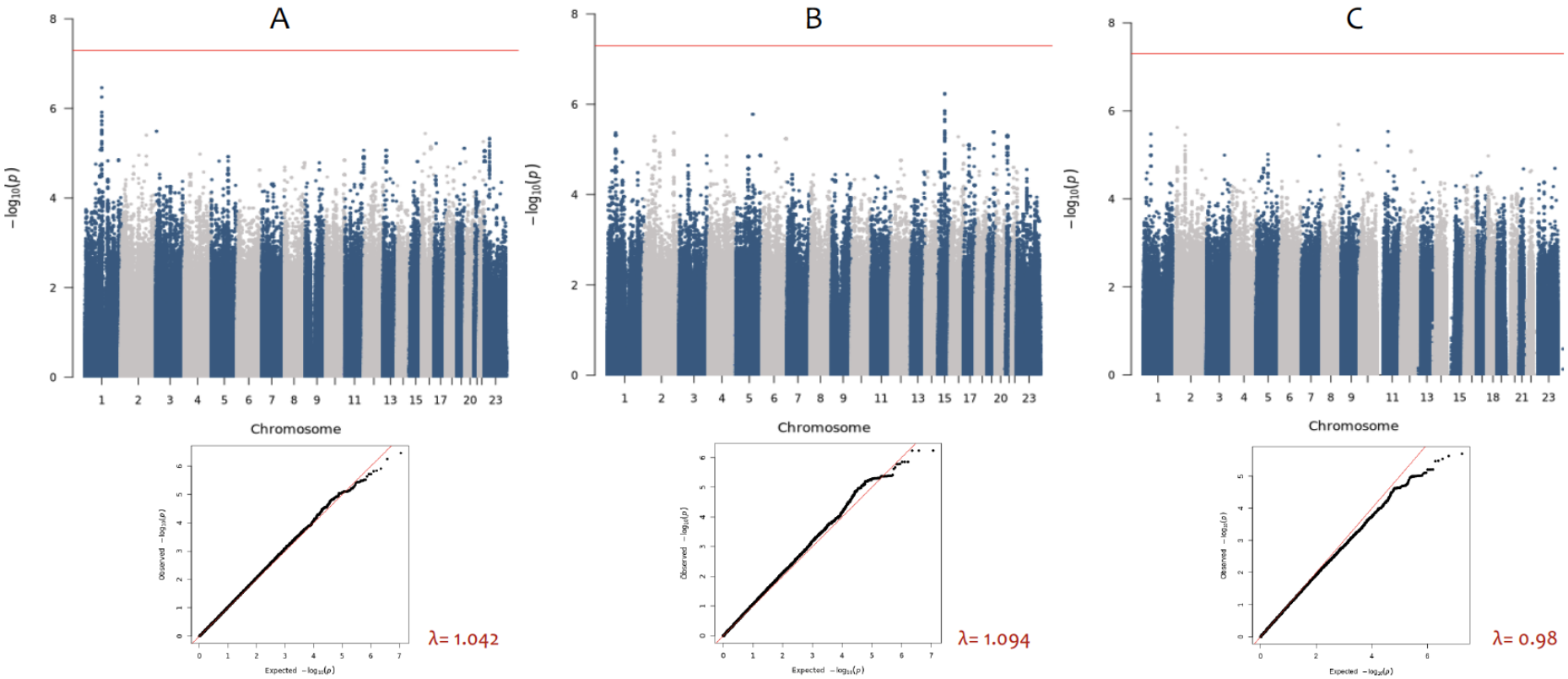

Figure 4.3: GWAS Manhattan, QQ plots and lambda inflation factors retrieved from different cohorts using REGENIE software. Results from: **A** the multicentre cohort, **B** 100kGP and **C** UKBB. None of the peaks reached the genomic significance threshold of 7.3 ($\log10(5x\mathbf{10^{-8}})$).

Table 4.1: Coding genes surrounding GWAS peaks in each of the studies. Table shows genes that are ±0.5 Mb from GWASs' highest peaks along with some of their protein functions. [Peaks were viewed using LocusZoom and gene functions retrieved from GeneCards (Stelzer et al., 2016)].

Study	*Chromosome*	*Surrounding genes*	*Gene Functions*
Multicentre cohort	1	*HIPK1*	The encoded protein is a Homeodomain Interacting Protein Kinase 1. It plays a part in gene expression and regulation of TP53 pathways.
		OLFML3	Encodes a scaffold protein that plays an essential role in dorsoventral patterning during early development.
		SYT6	Involved in Ca^{+2} exocytosis, and vascular trafficking.
		AP4B1	Component of the adapter protein complex 4. Involved in vesicle formation and cargo selection.
		BCL2L15	Regulation of apoptosis by parathyroid hormone-related protein.
		DCLRE1B	Central role in telomere protection, DNA maintenance, and repair.
		PTPN22	Involved in T-cell receptor signalling pathways. Mutations in this gene are associated with type 2 diabetes, and rheumatoid arthritis.
100kGP	15	*TLN2*	Cytoskeletal protein that plays a significant role in actin filament formation, plays an important role in cell adhesion.
		LACTB	Mitochondrial serine protease that regulates mitochondrial lipid metabolism. Associated diseases are gastroenteritis and lung abscess.
		CA12	Carbonic anhydrases which catalyzes the reversible hydration of CO_2. Involved in cellular processes such as, respiration, and formation of cerebrospinal fluid, saliva and gastric acid.
UKBB	8	*POU5F1B*	Recently found to be encoding for a DNA-binding transcription factor that plays a part in carcinogenesis and eye-development.
		MYC	Encodes a transcription factor that activates growth-related genes.
	2	*TRIB2*	Involved in tyrosine kinase activity, in addition to transferring phosphorus containing groups.

4.3.4 GWAS within linkage peaks in the multicentre cohort

Linkage analysis provided preliminary insights in the multicentre cohort; as outlined in **Chapter Three**, a region on chromosome one was identified with a LOD score > 3.3. Subsequent GWAS on the same dataset revealed a peak indicative of significance, also on chromosome one (**Figure 4.3 -A**). To compare the results, chromosome one data retrieved from both analyses were overlayed (**Figure 4.4**). The position of the peaks was visually widely disparate, linkage peaks, compared to those identified by GWA analyses, typically cover a greater genomic distance. The two peaks are separated by almost 80 megabases (Mb), specifically, the identified linkage peak spanned from 1:179,262,018-224,482,984, with a peak at 1:194,262,018, in contrast to the GWAS-identified peak (1:114,419,489-114,572,891), peaking at 1:114,542,914. The highest GWAS peak (lowest p value) within the linkage peak was at position 1:223,176,275, ~28 Mb away from the identified linkage peak.

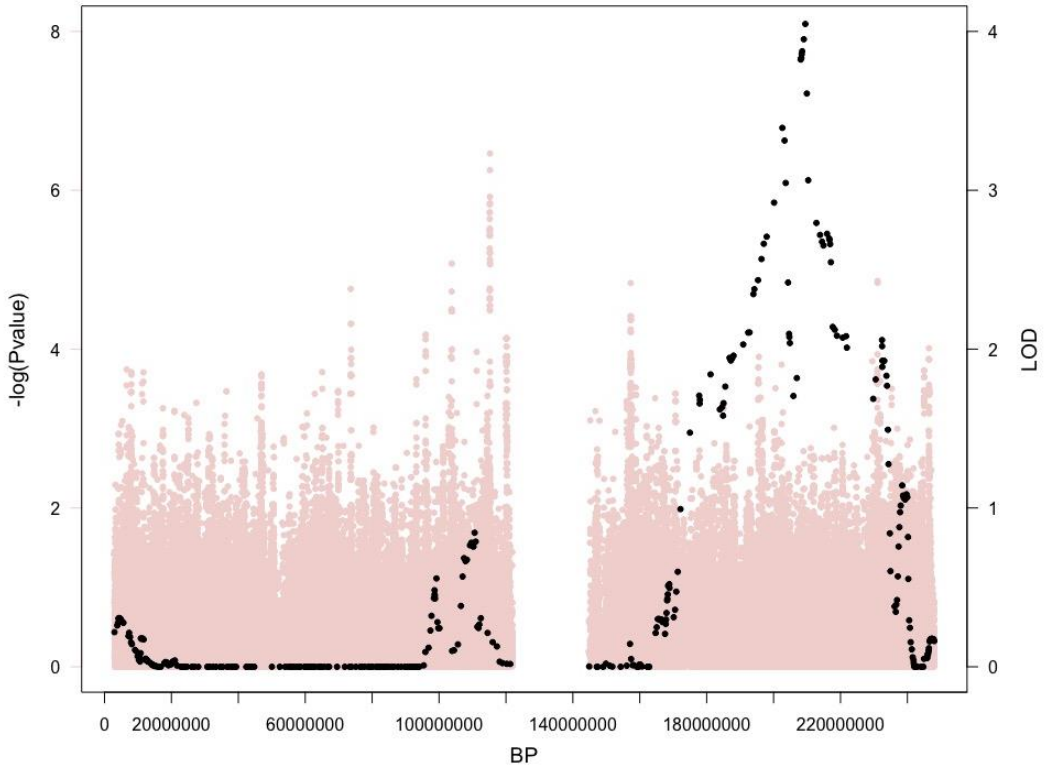

Figure 4.4: Linkage analysis LOD scores and GWAS P-values on chromosome 1 (GRCh37). X-axis shows base pair positions on chromosome one and the two y-axes show linkage LOD scores (black) and GWAS -log(P-value) (pink).

4.3.5 META

To increase the power to detect an effect, I decided to combine GWAS results from the UKBB and 100KGP cohorts in a META analysis. A decision was made not to include results from the multicentre cohort because the distribution of m.3243A>G levels was very different in this cohort (**Figure 4.5**).

Genotyping within the 100kGP cohort was performed using WGS whereas the UKBB cohort was genotyped using an array. However, both underwent the same QC steps, and analysis was performed using the same software, using the same study design. Nonetheless, data remain to be from two different cohorts, each with its own nuances. As mentioned in **Section 2.1.3**, data in the UKBB belong to individuals aged between 40 and 69, who do not represent the general population; with evidence suggesting the presence of bias towards 'healthy volunteer' within the cohort (Fry et al., 2017). In contrast, the data used in 100kGP belonged to the "rare disease" subset of the cohort, which were ascertained through individuals presenting with a rare, likely genetic disorder, and also included close relatives of these individuals (Greene et al., 2023). Both cohorts, however, displayed similar distributions of m.3243A>G levels and likely represent levels within the general, non-ascertained population.

Whilst acknowledging these differences, the average I^2 calculated by GWAMA software, was 15.10% (SD = 0.258), which according to Deeks JJ et al., (2023) is a value that indicates this heterogeneity "might not be important". As outlined in Begum et al., (2012), a random effect meta-analysis is not always the right choice for combining heterogeneous studies; particularly because an accurate estimate of heterogeneity can only be achieved with large numbers of studies. Due to this, and the negligible between-study heterogeneity, I chose to implement a fixed effects META analysis using GWAMA software. The retrieved results (**Figure 4.5**) demonstrated an effective inter- as well as intra-study correction with a lambda value of 1.029 and a QQ plot that deviated from the expected line only at the tail, reflecting a small number of SNPs with elevated -log10(p.values).

A peak was approaching significance on chromosome eight. This was led by rs1512802 (8:5882269G>C; -log10(Pvalue) = 6.9) that is in high LD with another group of SNPs in the same region. rs1512802 falls between two non-coding pseudogenes (**Figure 4.6**).

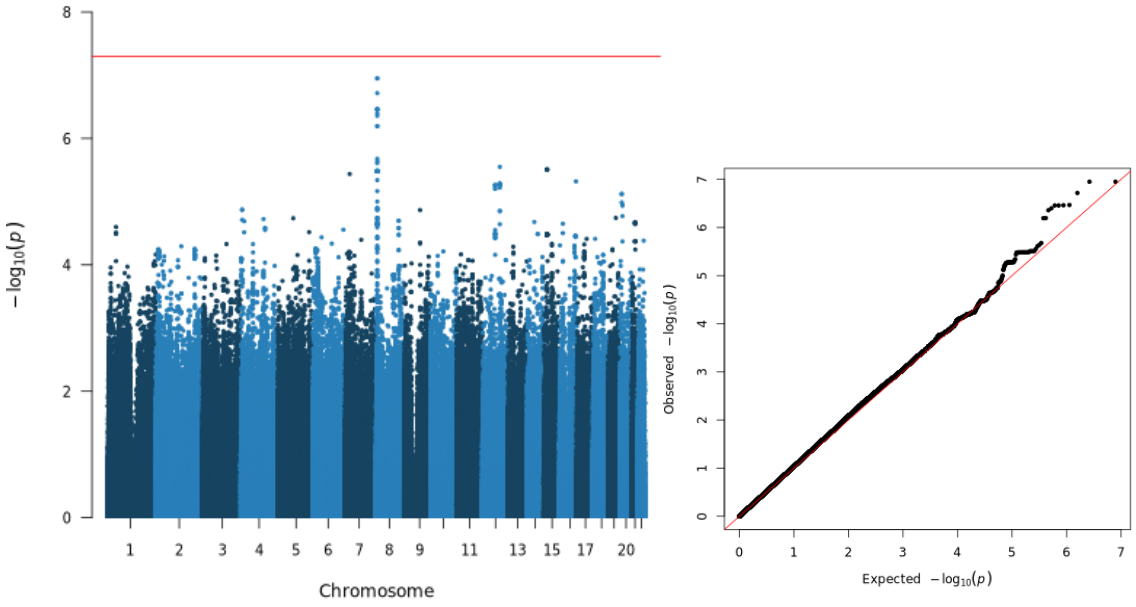

Figure 4.5: Manhattan and qq plot from 100kGP and UKBB fixed effects META. Peak on chromosome eight is approaching -log10(Pvalue) significance of 7.3 (indicated by the red line). QQ plot is well aligned to the expected, null line and lambda inflation factor is 1.029 which is within the permissible range of 1-1.10.

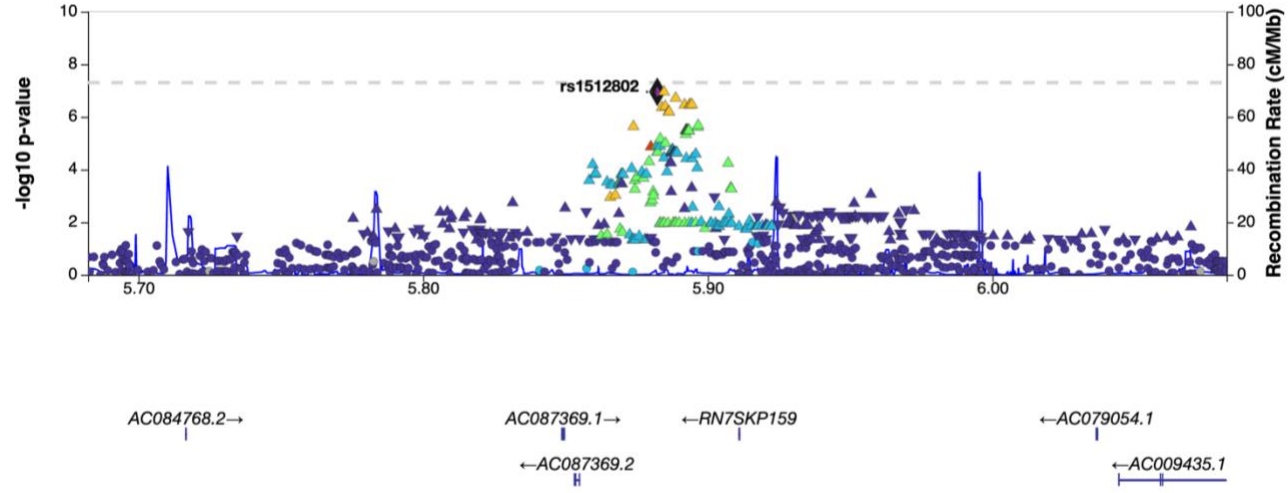

Figure 4.6: LocusZoom view of META association peak on chromosome eight. The lead SNP, rs1512802, is an intergenic variant mapping between *RN7SKP159 and AC087369.1* pseudogenes.

META lead SNP was then used for a few downstream analyses. Using the gene ranking option in Open Target Genetics (Ghoussaini et al., 2021), *MCPH1* gene that is 524,323 bp away from the lead SNP was the nearest coding gene. This encodes for a DNA damage response protein, microcephalin 1 which is involved in mitosis where it maintains the inhibitory phosphorylation of cyclin-dependent kinase 1, particularily important in neurogenesis. Associated diseases with mutations in this gene are the autosomal recessive microcephaly 1. Based on expression data from within the NIH website, this gene has increased expression in brain tissue, which reflects its function in neurogenesis, as was previously mentioned.
GTEx portal is a comprehensive public resource that harbours gene expression data within multiple different tissues and consequently, expression quantitative trait loci (eQTLs), that reflect the significant associations with the expression of nearby genes. Search within the portal revealed no significant association between rs1512802 and the expression of of any gene within any tissue. HiC software using blood tissue as a proxy was used to investigate and visualise chromatic interaction data between rs1512802 and its surrounding genomic regions (Belton et al., 2012). Which as well, lacked any detectable interactions.

4.3.6 Fine mapping

META analysis results combining 100kGP and UKBB data were followed-up with fine mapping using FINEMAP software (Benner et al., 2016). This was performed on a region of ~2Mbs (5768569 – 5952421) encompassing 785 SNPs surrounding and within the peak identified by META analysis on chromosome eight (**Figure 4.5**). Results are shown in **Table 4.2**. As discussed in **Section 2.5.4**, SNPs with the greatest -log10(p.val) do not always coincide with those that have the highest PIP value, which indicates that the SNPs are most likely to be causal in fine mapping. In other words, it is incorrect to assume that the GWAS/META lead SNP is the causal. As a check, PIP and -log10(p.val) of SNPs identified in the first fine mapping credible set (total of 47 SNPs) were plotted (**Figure 4.7**). In Bayesian statistics, a credible set is a range of values (a set of SNPs in this context), within which the true parameter value lies, the credibility of a set diminishes as it descends (one being the largest), indicating a smaller probability of that set encompassing the true, causal SNP. Results show that in this case, the META lead SNP (rs1512802 / 8:5882269G>C) is in LD ($0.6 < r^2 < 0.8$) and in the same credible

set with those identified with the highest PIP value (=1). However, 8:5882269G>C itself was not identified as causal.

The above was done using the parameter --n-causal-snps 5 that sets the maximum number of causal SNPs to 5. I have then tried the analysis using --n-causal-snps 1, and none of the SNPs presented a PIP = 1 i.e., analysis could not identify one SNP with a certainty of one and PIP value was distributed across 37 SNPs in the same credible set (compared to 47 SNPs in the previous design). None except for SNP 8:5882269G>C was shared between the two obtained credible sets. Also, SNP 8:5882269G>C had the highest PIP value of 0.1416, which can be explained by model constraint, where it was limited to assigning only one SNP as potentially causal, therefore resorting to the SNP with the smallest p value.

Table 4.1: Fine mapping summary statistics from the top five SNPs, alongside the lead META SNP. Beta, SE and -log10(p) values are all retrieved from the input file given for fine mapping (META summary statistics). PIP is the measure of probability, and Z is the values position relative to the mean (in SD), which utilises the provided META summary statistics.

rsid	ß	SE	Z	PIP	-Log10p
8:5782714C>T	0.032386	0.015481	2.09198	1	0.692
8:5778926C>G	-0.014495	0.026001	-0.557479	1	0.721
8:5783389G>C	-0.013499	0.013482	-1.00126	1	0.499
8:5776324C>G	-0.025597	0.016259	-1.57433	1	0.103
8:5886877A>G	0.061576	0.015239	4.04069	1	4.504
8:5882269G>C	0.076383	0.014387	5.30917	3.57898e-10	6.860

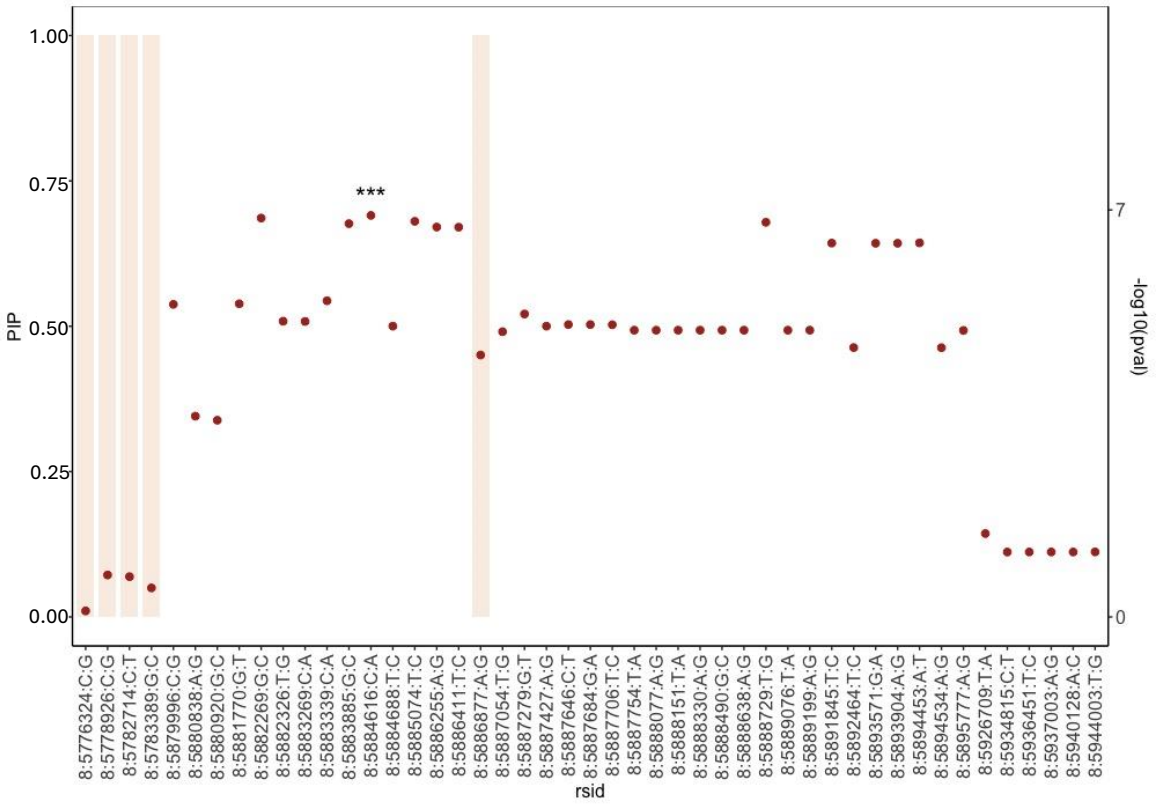

Figure 4.7: PIP and -log(10pval) of 47 SNPs in the first credible set obtained from fine mapping. Bars represent PIP values after multiplying by 10, whereas the red points are -log10(pvalues) of SNPs identified to be in the first fine mapping credible set. SNP 8:5882269G>C exhibits the greatest -log10(p.val) indicated by the stars, whereas 5 bars have equal PIP values, reflecting the possible causality of these SNPs.

4.3.7 Heritability estimates

As discussed in previous chapters, 72% of m.3243A>G level variability was estimated to be explained by additive nuclear factors (Pickett et al., 2019). To elucidate how much of heritability was explained by the variants obtained from META analysis combining 100kGP and UKBB data, a SNP based heritability estimate was performed; this yielded a total value of (h^2 = 17.717, SD = 4.757). Meaning that ~17% of m.3243A>G level difference is explained by nucDNA factors. SNP heritability estimates were also performed on the multicentre cohort, which includes the pedigrees used in the initial heritability estimate by Pickett et al., (2019); interestingly, this yielded a value of (h^2 = 12.210, SD = 2.962) which is much lower to the estimated heritability value of 0.72 (standard error = 0.26).

4.4. Discussion

In the early 2000s, the hope was that GWASs will have the power to identify, for example, 10 common risk loci, each of which will explain 1/10 of h^2, by that unravelling disease architecture as well as the molecular causes of disease. However, complex, polygenic diseases, in particular, proved to have much more complexity, leaving researchers with more unanswered questions than anticipated. None of the GWA studies presented here yielded a significant association; this may be attributed to an insufficient sample size and can be based on the performed power analysis, which did not show sufficient detection power for variants with MAF < 0.05 and ES < 0.6 (**Section 4.3.1**). m.3243A>G heterogeneity could be modulated by multiple variants with a much smaller effect size and/or, rare variants with large effect size. This GWAS would have missed rare variants as data were filtered for MAF > 0.05, additionally, the multicentre cohort and UKBB data are derived from a genotyping chip, followed by imputation, thus, some SNPs might be missing from the data. This explains the growing trend towards using WGS in GWA analysis (Selvaraj et al., 2022). A far bigger sample size would be needed to have enough power to detect small effect sizes.

Additionally, when it came to the multicentre cohort analysis, one limitation might have been the lack of correction for between-cohort differences. For example, 110 individuals were recruited from Exeter Centre of Excellence for Diabetes Research, which might have led to differences in m.3243A>G levels between this portion of samples compared to those recruited from specialised, mitochondrial disease clinics (Newcastle, Italy, Germany). In case differences were significant, true associations would be missed, potentially leading to a null result or reduced effect sizes in the pooled analysis. This can occur when one cohort exerts a disproportionate influence, thereby 'averaging out' significant findings that might be present in other cohorts. A way around this would be by performing an ANOVA test; to quantify the differences between cohorts and determine whether the observed variation is statistically significant, something that would necessitate the use of cohorts as covariates in the analysis. Follow-up PC analysis might also help determine if individuals from different cohorts cluster together or separately based on genetic data, which would help determine additional outlier samples (if present) based on batch effects.

When it comes to the two population-based cohorts, results were not concordant; given the small sample sizes, this might as well be a lack of detection power, and that the signals appearing were rather analysis noise. The fact that each of the cohorts is ascertained in its own way, should also be acknowledged – 100kGP selects for individuals with suspected rare genetic disease whereas the UKBB cohort consists of individuals between 40-69 years of age, who have a reduced incidence of all-cause mortality compared to the general population (Bycroft et al., 2018a). These differences may influence both m.3243A>G levels and the genetic factors that determine these levels.

The fact that the UKBB data belongs to individuals from an older age group, and that m.3243A>G has an annual, blood level decline rate of ~2% (Rajasimha, Chinnery and Samuels, 2008), also explains why the number of retrieved samples was less than that expected (mentioned at the beginning of the chapter), where with a population carrier rate of 140~250 in 100,000 (Manwaring et al., 2007a), we estimated to identify approximately ~280-500 (out of 200,000) yet the total retrieved number of samples was 147.

The identification of further samples from 100kGP and the UKBB meant there was a bigger sample to analyse, and an increased power to detect "causal" variants. However, as previously mentioned (**Section 4.3.5**), due to the different distribution of phenotype levels across the three cohorts, a decision was made to exclude the multicentre cohort from the META (384 samples). This is mainly because the GWASs were trying to identify nuclear variants driving high levels of m.3243A>G (multicentre cohort) and those driving low to medium levels of m.3243A>G (UKBB and 100kGP). This left 279 samples cleared for analysis, together resulting a peak approaching significance on chromosome eight.

Fine mapping identified five SNPs from the first credible set with values equal to one. These did not correspond to the SNPs with lowest p values from the META analysis, in fact, the SNP with the lowest p value had a PIP value well below one; indicating the uncertainty in its causality. PIP is a better measure of causality in fine mapping analysis than p value, or Z values; this is due to the fact that a PIP provides a probability measure that a variant is truly causal, whilst incorporating the effects of other variants and LD patterns, which are not accounted for when for example, calculating p values (Maller et al., 2012).

SNPs with PIP values of one were in LD with the lead META SNP, and had big, insignificant p values which translated into very small -log10(p values) as shown in **Table 4.2** above. Interestingly, when changing the software's default and asking for the top number of causal SNPs to be only one, although not equal to one, the lead META SNP (8:5882269G>C) appeared to have the largest PIP value in the obtained credible set. In fine mapping, it is generally recommended to allow for the possibility of multiple causal SNPs. This approach aligns better with the polygenic nature of complex traits and provides a more nuanced understanding of the genetic landscape. Limiting the model to one causal SNP might be overly simplistic and could result in miscalibration of analysis leading to misleading conclusions by focusing on the SNP with the smallest p-value, which may not truly be the causal variant (Burgess, 2022; Kanai et al., 2022).

The obtained SNP heritability of 17.717% for the META analysis, and 12.210% for the multicentre cohort, compared to the estimated classic heritability of 72%, shows a similar pattern to that seen for other complex traits (**Section 4.1.4**). There are different explanations for the issue of missing heritability: GWASs detect common variants with small effect sizes whereas it might be that rare variants, which typically have a much bigger effect size, are accounting for a percentage of that heritability; additionally, most available genotyping panels do not tag copy number variants, and it is possible that they might have a biologically important function that is missed (Cirulli and Goldstein, 2010; Craddock, Hurles and Cardin, 2010). A further explanation for this might be the fact that the classic heritability estimates utilised mother's m.3243A>G heteroplasmy data as covariates, by that enriching the residuals for the tested factors (Pickett et al., 2019a). Which may account for some of the observed discrepancies in heritability. **Figure 4.4** showed a discrepancy as well, where linkage and association peaks landed on different positions within chromosome one (separated by ~80 MBs), this may be attributed to several reasons most notably, the inherent difference of the methods. This dictates different genetic resolutions, as well as variant types detected. Both methods, however, provide valuable, yet distinct insights into the genetic architecture of the trait, and their combined interpretation can offer a more comprehensive understanding of the genetic factors involved.

Understanding the contributions of these variants, alongside the need for far bigger sample sizes, will be necessary for us to achieve a full understanding of the nuclear factors that play a role in determining m.3243A>G heteroplasmy.

Chapter 5. Mitochondrial DNA GWAS (miWAS)

5.1. Introduction

As outlined in **Section 1.2.6** mtDNA sequence variations can be maternally inherited and can also arise throughout our lifetime as somatic *de novo* mutations (Lawless et al., 2020).

Inherited mtDNA variation are associated with several common diseases, including Parkinson's disease (PD). Using association analysis, it was demonstrated by Hudson et al., (2013) that the m.2158T>C variant (mapping to J1b haplogroup) in mitochondrial *MTRNR2* gene, may provide a protective mechanism in PD, authors suggest that the variant could alter the synthesis of Humanin, the neurotoxicity suppressor protein encoded by this gene. A comprehensive study by Yonova-Doing et al., (2021) looked at mtDNA associations with 877 different complex diseases within 358,916 British ancestry individuals from the UK biobank. This identified 260 novel associations such as, m.8655C>T and type 2 diabetes, m.14766T>C and increased levels of aspartate aminotransferase (AST) which is a biomarker of various liver diseases including hepatitis and cirrhosis. A study analysed 38,638 individuals with 11 diseases, and 17,483 healthy controls, and suggested that common mtDNA variants may fill in the 'missing heritability' of several complex diseases (Hudson et al., 2014c). This identified a set of common mtDNA variants that were found to have an impact on several diseases. Such as the haplogroup U5a marker, m.14793 within *MTCYB* gene, and *MTCO3* m.9477 variant, a marker of haplogroup U5, both associated with an increased risk of Schizophrenia, Parkinson's disease (Huerta et al., 2005), ulcerative colitis, and multiple sclerosis. On the other hand, m.10398 variant within the *MTND3* gene, a marker of both haplogroups J and K, was found to be linked to a reduced risk of several diseases such as Parkinson's disease, multiple sclerosis and ischemic stroke (van der Walt et al., 2003; Hudson et al., 2014; Tzeng, 2022).

Studies using *trans*mitochondrial cybrid cells have shown that OXPHOS function varies under the influence of different haplogroups (Gómez-Durán et al., 2010a). Hudson et al., (2007) have investigated the variable clinical presentation of Leber hereditary optic neuropathy (LHON) disease within carriers of different haplogroups; where carriers of the variant

m.11778G>A with haplogroup H background seem to have milder presentation of disease symptoms, and those with haplogroup J background were at a higher risk of visual failure (Carelli et al., 2006; Yu-Wai-Man et al., 2009). This raises the possibility that haplogroup H exerts a protective mechanism, whereas haplogroup J aggravates m.11778A>G-related disease (Carelli et al., 2006; Yu-Wai-Man ct al., 2009). Other studies have explained the tissue specificity of certain mtDNA variants, such as m.60T>C within liver and kidney by replication advantage, especially when found within or close to mtDNA replication regulation sites. These variants were found in non-related individuals exclusively within the specified tissues and in this context, their observation fit no particular haplogroup pattern (Samuels et al., 2013).

Using GWA studies Gupta et al., (2023) identified a length variant at the m.302A>AC position that is associated with mitochondrial copy number (mtCN) variation. The longer the segment surrounding m.302 is, the smaller the copy number. Where certain nucDNA genes such as *TFAM* and *POLG2* were found to act in *trans* on the mitochondrial RNA polymerase (mtRNAP) switching off the replication machinery, by that modulating the efficiency of mtDNA replication. In contrast, haplogroup markers explained less than 0.5% of mtCN variability.

Unlike nuclear DNA, there is no recombination within the mtDNA which is due to its circular nature and lack of recombination machinery (Saville, Kohli and Anderson, 1998). This, however, was a point of debate in the early 2000s, when Awadalla et al., (1999) suggested evidence for mtDNA recombination. Plenty of groups after failed to replicate their findings, deeming the absence of recombination in mtDNA as the most agreed upon theory in the mitochondrial community; and attributing the conclusions of Awadalla and colleagues as an artefact due to an inappropriate LD measure (Elson et al., 2001a).

As a result of the absence of protective histones and the presence of ROS, some areas in the mtDNA are mutation hotspots that are ideal for the study of migration patterns. Additionally, due to the lack of recombination in mtDNA, all variants are inherited together as a single haplotype (Elson et al., 2001a), making it challenging to identify the specific variant responsible for an observed association.

Given disease associations with mtDNA variation, it is plausible that the background sequence variation of mtDNA could affect m.3243A>G levels. In case an effect is identified, it

could either be in *cis*, such as in the case of haplogroup markers and LHON disease (Carelli et al., 2006), or in *trans* by involving nucDNA QTLs that influence mtDNA regulation (Gupta et al., 2023). An alternative mechanism that could regulate m.3243A>G heteroplasmy is a bioenergetic alteration, caused by OXPHOS modulating variants. Such as the inherited basal differences in OXPHOS capacity reported by Gómez-Durán et al., (2010b), where haplogroup H markers were suggested to have a delayed and reduced OXPHOS complex one assembly, slowing the bioenergetics of a cell. Disrupted OXPHOS may lead to a selective advantage for mtDNA molecules carrying compensatory variants that partially restore OXPHOS function (Khrapko and Turnbull, 2014). This could result in an increase in heteroplasmy for those compensatory variants.

To answer this question, I conducted a mtDNA GWA analysis (miWAS) with the aim of determining whether mtDNA sequence variation is associated with m.3243A>G levels

A study in the French population carried out by Pierron and colleagues (2008), used control-region sequencing and RFLP survey of mtDNAs to determine haplogroups in m.3243A>G carriers. They report a statistically significant underrepresentation of m.3243A>G variant in carriers of haplogroup J from the French population. As an explanation, they hypothesised that the combination of m.3243A>G and haplogroup J could be lethal (potentially related to non-synonymous cyt *b* variation) and termed this hypothesis the 'haplogroup J paradox', hypothesising that this could result in negative selection of m.3243A>G on a J background. In their study, they have also observed m.3243A>G variant on different haplogroup backgrounds, suggesting that 3243 position is a mutational hotspot in European populations.

Therefore, I also decided to investigate this within our population, with the aim of elucidating whether any of the haplogroups manifest a similar, increased selection against m.3243A>G in the European populations of Britain.

5.2 Methods

5.2.1 Data

To investigate this, SNP genotyping and sequencing data of individuals with blood m.3243A>G levels ≥ 1% were used from three cohorts: the multicentre cohort, 100kGP (Genomics England), and the UKBB. In the multicentre cohort (408 individuals) had 222 mtDNA SNPs genotyped using UK Biobank Axiom® Array at a genotyping rate of 0.997 per SNP. No individuals were excluded for missingness (threshold 2%). After filtering using a MAF ≥ 0.01, and MAC ≥ 20, 53 SNPs were taken forward for analysis.

MtDNA sequencing data of 136 individuals within Main Programme Genomics England Data Release v12.0, and genotyping data of 143 individuals in the UKBB were used. After filtering for the same MAF and MAC thresholds, 69 SNPs were left in 100kGP data. For comparison reasons, particularly for META analysis, the same subset identified in 100kGP was used in UKBB data, where out of 69, 49 SNPs were matched.

In analyses comparing haplogroup distributions across m.3243A>G carrier populations, since the families were of different sizes, only one individual was kept and that is to avoid any bias. Total left was 268, 102, and 30,046, in the multicentre cohort, the carrier portion of 100kGP, and the non-carrier portion of 100kGP, respectively.

5.2.2 Haplogroup estimation

Haplocheck is a software used to detect contamination patterns in both whole genome as well as targeted mitochondrial sequencing studies (Weissensteiner et al., 2021). This uses Haplogrep2 and is based on sequence data from Phylotree V17 (van Oven and Kayser, 2009b) as a reference to identify the major and the minor haplogroups in data (Weissensteiner et al., 2016c). This was used as part of the quality control pipeline all cohorts underwent, which enabled the identification of participants' haplogroup information.

5.2.3 mtDNA principal component analysis (PCA)

Being one of the biggest confounders of GWASs, including miWAS, population structure (PS) had to be accounted for. A 2010 study compared the utility and efficacy of PS correction by using mitochondrial DNA PCs (mtPCs), mitochondrial haplogroup data, or nuclear DNA PCs (nucPCs) in miWAS. It was found that using haplogroups was inferior to using mtPCs; where analysis carried out using mtPCs yielded significantly lower mitochondrial genomic inflation factors (mtGIF) ($p = 0.022$) compared to using haplogroups (Biffi et al., 2010). The addition of nucPCs to mtPC adjusted analysis led to no significant difference in mtGIF ($p = 0.41$). It is possible that true mtDNA associations were missed in previously carried out miWAS studies which corrected for PS using nucPCs, because the addition of unnecessary nucPCs causes an increase in the degrees of freedom and therefore, a higher p value (Miller et al., 2019). To conduct mtDNA PC analysis, PLINK v1.9 was used to calculate eigenvectors and eigenvalues of mtDNA in both the multicentre cohort and the 100kGP cohort. As a great complement to nucDNA PCA studies, mtDNA was used to look deeper into the genetic ancestries of participants in both cohorts by overlaying results from mtDNA PC analysis over the previously performed nucDNA PC analysis (outlined in **Section 3.3.1**).

Ahead of carrying out the miWAS, data were checked to belong to individuals of European nuclear ancestries, which is the same portion of individuals used in the nucDNA GWAS. As outlined in **Section 3.3.1**, this was done by excluding nucPCA outlier samples to prevent confounding factors in the GWAS. This step reduces the likelihood of type 1 errors and minimizes the risk of obtaining spurious associations due to differences in population-specific LD patterns rather than real genetic links. This resulted in cohort sizes of 384 (out of 408), 136 (out of 164), and 143 (out of 147) for the multicentre, 100kGP, and UKBB cohorts, respectively. As outlined in Miller et al., (2019), in comparison to nucPCs, mtPCs are better at capturing intrapopulation variation, which can also be referred to as population substructures; therefore, by using mtPCs as covariates, correction for intra-European, mitochondrial genetic variations was ensured.

5.2.4 mtDNA GWA analysis optimisation

As discussed in **Section 2.1.1**, the multicentre cohort contains data from related individuals belonging to 95 families, so this also had to be accounted for in the analysis design. The suitability of linear mixed effects modelling as well as linear regression models in such cases was thoroughly outlined in **Sections 2.5.1** and **2.5.2**. A comparison between association software performance on mtDNA data, as well as the effect of inclusion and exclusion of PCs is demonstrated in the results section below. REGENIE was also the software of choice for mtDNA GWASs. In step 1, to fit the regression model, pruned nucDNA data was used particularly to account for cryptic relatedness within the multicentre cohort and 100kGP. In step 2, mtDNA data consisting of 49, 69, and 53 SNPs from the multicentre cohort, 100kGP, and the UKBB, respectively, were tested for association. Both steps used mtPCs as covariates to account for PS, and age-adjusted m.3243A>G levels as the tested phenotype.

5.2.5 Converting mapping data between genome builds.

As outlined in **Section 2.4.3.C**, genomic builds were different across cohorts. To avoid SNP mismatching, lifting over of mtDNA SNP coordinates was necessary ahead of the META analysis. This was performed using the web based UCSC lifting over tool https://genome.ucsc.edu/cgi-bin/hgLiftOver (Kent et al., 2002; Kuhn, Haussler and Kent, 2013).

5.2.6 META analysis

As outlined in **Sections 4.2.3 and 4.2.4**, there are various statistical methods for performing META analysis; which method to employ largely depends on between-study heterogeneity (I^2). In this case, only two cohorts were combined, and thus, as advised in Dettori R. et al., (2022) a decision was made to carry on with a fixed-effects META analysis using PLINK v1.9. This relies on inverse variance weighting (Willer, Li and Abecasis, 2010), where combined effect sizes (ESs) are used. ESs reported to be statistically more powerful than using the combined z scores; because combined effect sizes have the benefit of incorporating the precision of each individual study i.e: studies with smaller variance and more precise

estimates contribute more to the combined effect size and vise versa. The underlying statistical formula to calculate the combined effect size is as follows:

$$ES_c = \frac{\sum_{i=1}^{k} w_i . ES_i}{\sum_{i=1}^{k} w_i}$$

where:
ES_C : combined effect size
ES_i : effect size of i-th study
w_i : weight assigned to the i-th study based on inverse variance

The weights are calculated using the inverse of the variance of effect size in each study through the following: $w_i = \frac{1}{ES_i}$

5.2.7 Significance threshold

Bonferroni correction was used to derive the generally accepted GWAS significance threshold of 5 x 10^{-8}. Therefore, I decided to use the same correction method to determine a significance threshold for this mtDNA GWA study; the -log10(pvalue) significance threshold for the multicentre cohort was 3.03, and for the two population cohorts was 2.99. The significance threshold plotted in Manhattan plots was rounded to 3.00.

5.3. Results

5.3.1 PCA analysis

To view population structure in both cohorts, PC1 and PC2 then PC2 and PC3 eigenvalues retrieved from mtDNA PC analysis were plotted against each other (**Figure 5.1**). 95.3% and 88% of individuals in the multicentre cohort and Genomics England carry one of the nine European ancestry haplogroups: H, I, J, K, T, U, V, W, and X. This reflects the fact that in both cohorts, patient data/ samples were recruited from centres across the UK and the broader European continent. Although rare, individuals with admixed American, African, as well as East and South Asian ancestry haplogroups were present, showing that m.3243A>G is not an

exclusively European variant as was once thought (J.Morten, Poulton and Sykes, 1995). Mitochondrial principal components were successful at capturing the variance into distinct haplogroups, as observed, PC1 captures the biggest portion of variance in the multicentre as well as 100kGP data (~ 12% and ~11%, respectively).

The scatter plot also shows a degree of overlap between some haplogroup clusters, reflecting a shared genetic component. These groups however, could be further separated by plotting additional PCs, as they would provide more dimensions of variance to a two dimensional plot that is depicting multidimensional data (**Figure 5.1**).

To enhance our understanding of the genetic ancestry of subjects involved in this analysis and to investigate the concurrence of ancestry data between the two genomes, mtDNA haplogroups and nucPCs retrieved from analysis carried out combining 1000 genomes reference data, as described in **Chapter Three**, were overlaid. This revealed some individuals with mixed genetic ancestry patterns such as those in **Figure 5.2 – C**, where three individuals carrying the matrilineal African mtDNA haplogroup L (magenta) clustered over the South Asian nucDNA ancestry. This discordance between the maternal (mitochondrial) and nuclear ancestry, reflects a diversity resultant from an African admixture that occurred at some point in their maternal lineage. However, the majority of individuals' mitochondrial DNA ancestries coincide with their nucDNA ancestry, and their haplogroups lie over the expected nucPC cluster such as **Figure 5.2 – D** where a bright yellow cluster of Asian haplogroup M carriers lays over the South Asian nucPC cloud represented by diamonds, as well as the dense carrier group of European haplogroups clustered over the European nucPC cloud (**Figure 5.2**).

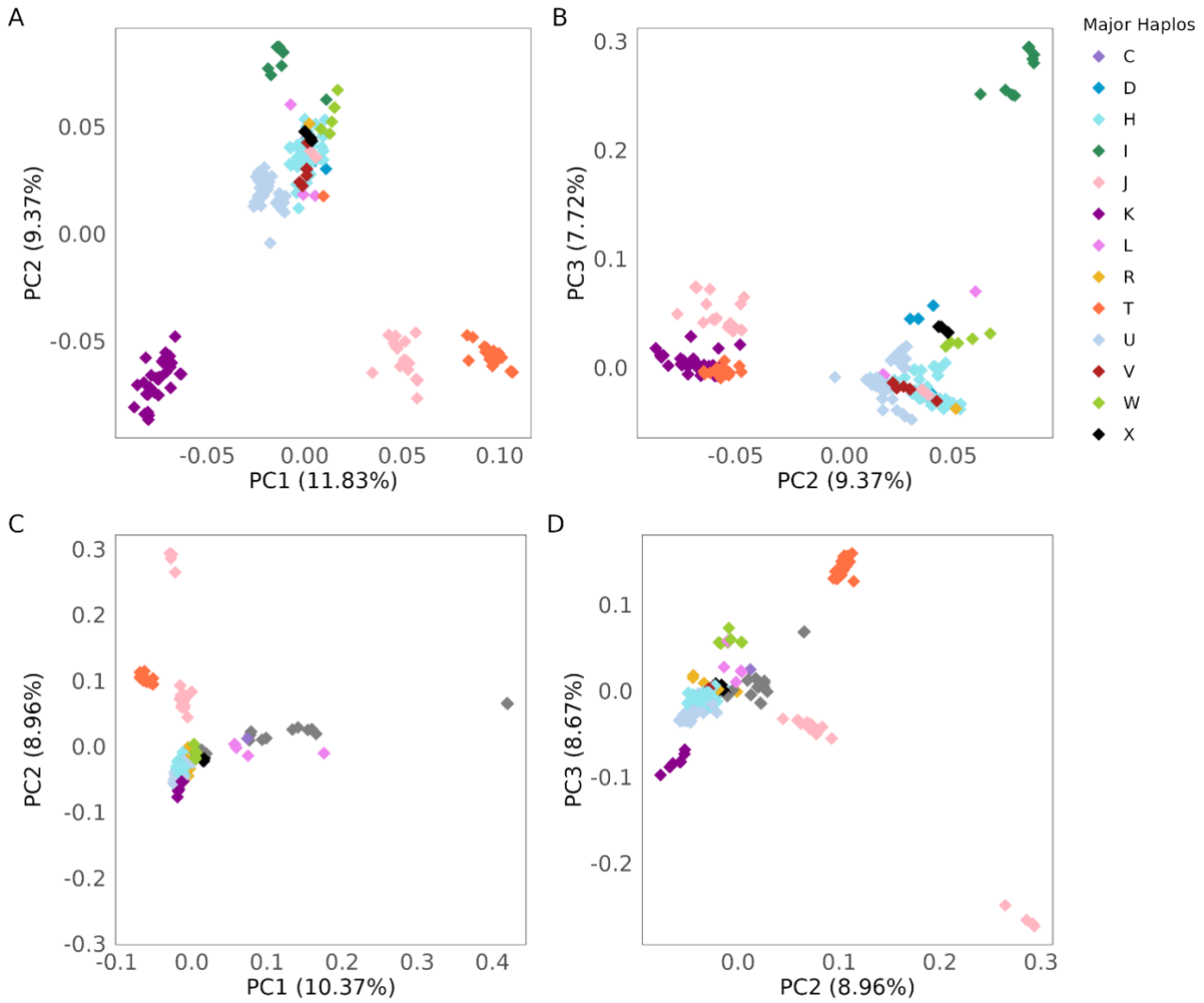

Figure 5.1: mtDNA PCA on both cohorts. (A+B) depict plots of PC1 and PC2 plotted against each other on the multicentre cohort whereas (C+D) are those in Genomics England.

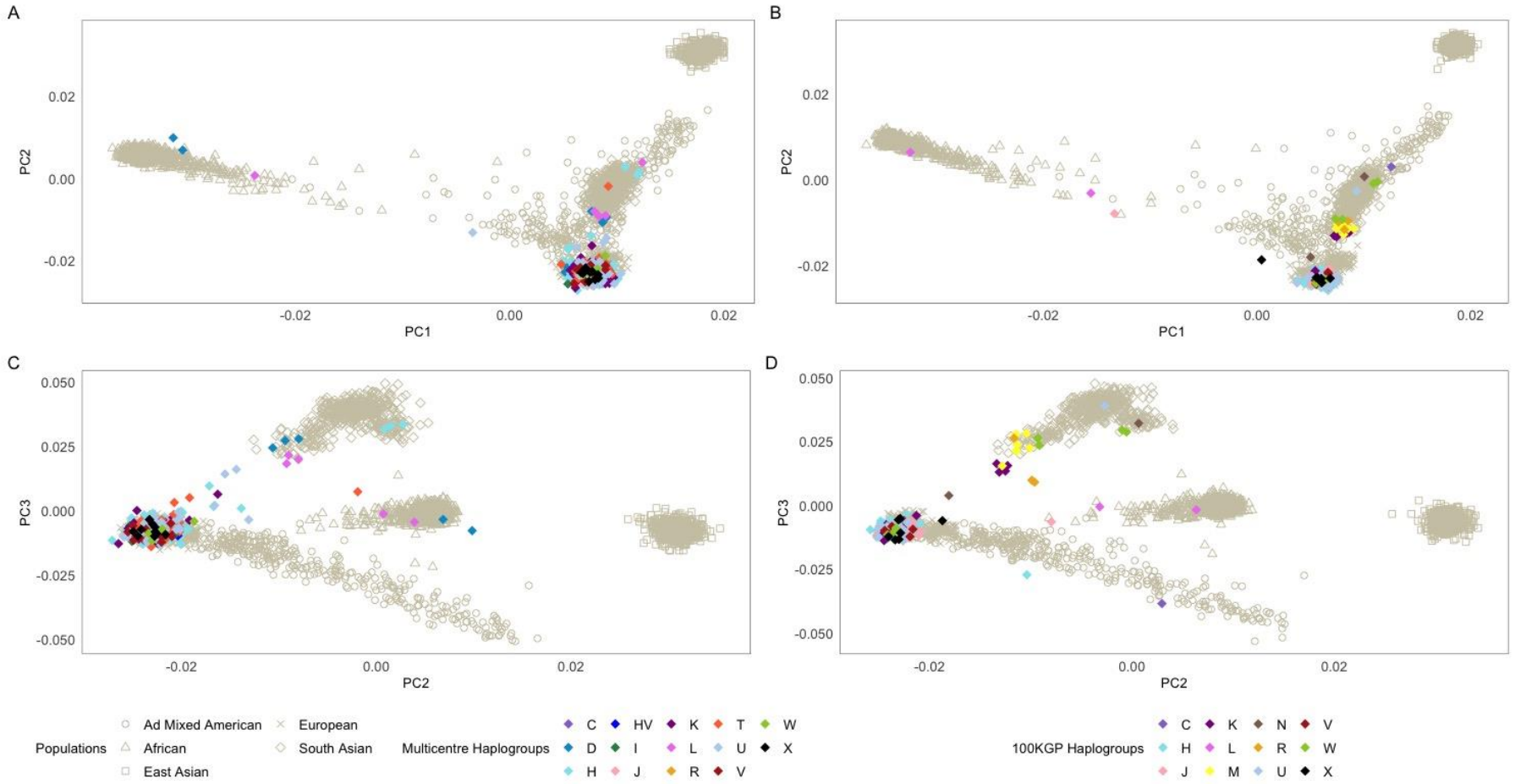

Figure 5.2: nDNA PCA and haplogroup plots. (A) Multicentre cohort nPC1 plotted against nPC2, (C) and nPC2 plotted against nPC3. (B+D) 100kGP 3 principal components plotted against each other. Filled, coloured diamonds represent individuals from the m.3243A>G cohorts; colours represent the different mtDNA haplogroups. Beige shapes represent individuals from the 1000 genomes project reference data (GRCh 38) who belong to five genetically distinct populations; African, European, East Asian, AD mixed American, and South-Asian. Most individuals within the m.3243A>G cohorts belong to the European ancestry population; European haplogroups (I, J, K, R, T, U, V, W, X) cluster over the European nPC cluster. Some individuals with African haplogroup L, and European haplogroup H co-locate with the south Asian nPC cluster (C), reflecting a mixed, African/ European maternal and Asian nuclear ancestry.

5.3.2 Frequency of m.3243A>G across different haplogroups

As a replication of the study on the French population (Pierron et al., 2008), the frequency of haplogroups across 100kGP, the UKBB, and the multicentre cohort was investigated in comparison to: 1) European population haplogroup reference data from from EUpedia , 2) the m.3243A>G non-carrier portion of Genomics England (100kGP) (for 100kGP data only), and 3) haplogroup reference data from EUpedia specifically for England (Capelli et al., 2003; Bowden et al., 2008; Winney et al., 2012; Hay, 2018). Haplogroups with fewer than ten entries in the multicentre cohort were counted as "other" in all three cohorts; those that remained were mainly European (H, J, K, T, and U).

Haplogroup frequencies within the multicentre and UKBB m.3243A>G cohorts were broadly similar to the estimated frequencies within the European and English populations.

Haplogroup H was the most common haplogroup across all cohorts (**Figure 5.3**), which is expected given it is the most common haplogroup in European populations.

In **Figure 5.3**, confidence intervals (CIs) reflect the higher variability in England compared to Europe, which is the results of a smaller sample size (2,333 vs. 27,341).

The frequency of haplogroup K was significantly higher in the multicentre cohort compared to Europe and England (**Table 5.1**; $p<0.0001$). However, this was not the case in the UKBB (**Figure 5.3**).

As a replication for the comparison between the UKBB and multicentre cohort to the English/ European general populations, m.3243A>G carrier and non-carrier portion of the 100kGP cohort were investigated separately (**Figure 5.4**). Results also show that the distribution of haplogroups does not differ between the carrier and non-carrier portion of 100kGP. The elevated frequency of haplogroup K seen in the multicentre cohort compared to non-carrier populations, was not replicated within 100kGP cohort either.

A significant difference in haplogroup distributions is observed between the European and the non-carrier portion of 100kGP data (**Table 5.1**, $p < 0.0003$), but not between the European and English populations. This reflects the fact that the non-carrier portion of 100kGP was selected for having a rare disease and thus, are not a reflection of the general, non-selected populations. Using the non-carrier portion of 100kGP as a comparison was instead a way to

compare the frequency of haplogroups between m.3243A>G carrier individuals and individuals of rare disease within the population.

Although not significant, there is a greater number of m.3243A>G carrier individuals with haplogroup J in both the multicentre and the carrier portion of Genomics England, compared to non-carrier populations, the opposite trend is seen for UKBB's m.3243A>G carriers (**Figure 5.3** and **Figure 5.4**). Therefore, these results do not replicate the previously reported association with haplogroup J in the French population (Pierron et al., 2008), as there are no signs of haplogroup J being underrepresented in m.3243A>G carrier populations.

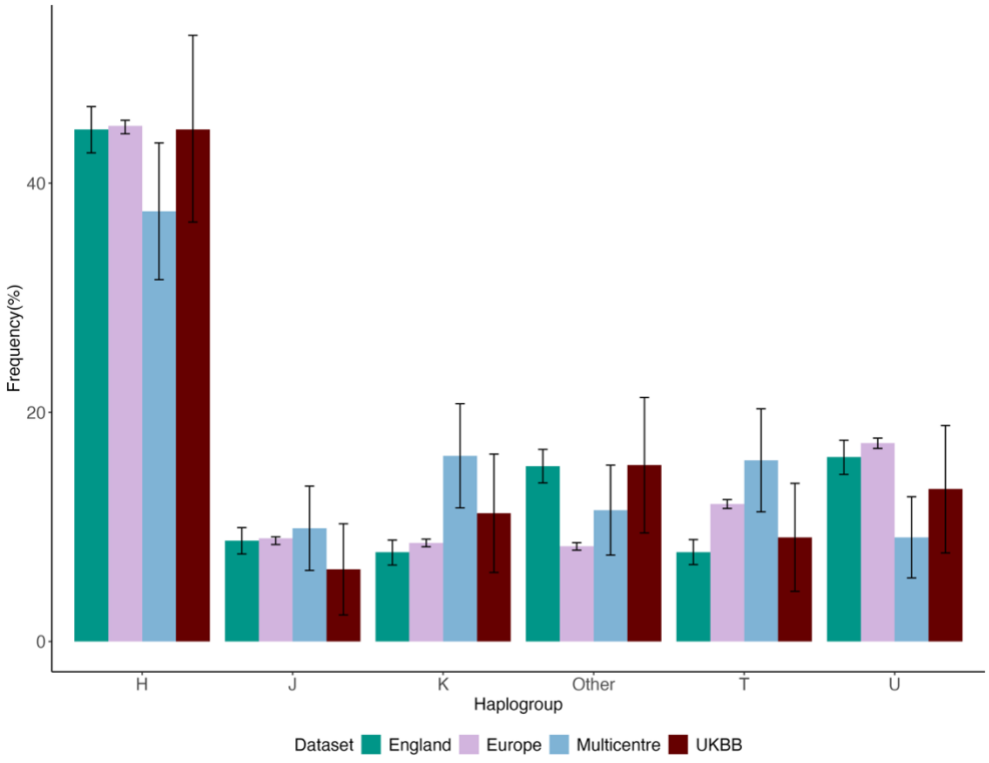

Figure 5.3: Comparison of European haplogroup frequencies between cohorts of m.3243A>G carriers and estimates of whole population frequency. Figure presents the percentage frequency of haplogroups and their confidence intervals across two cohorts of m.3243A>G carriers; the multicentre cohort (blue; one individual per family, total n=268 individuals) and UKBB (burgundy; n=143), compared to frequencies in the European population (violet; n=27341), as well as England (2333; n=green; both retrieved from EUpedia).

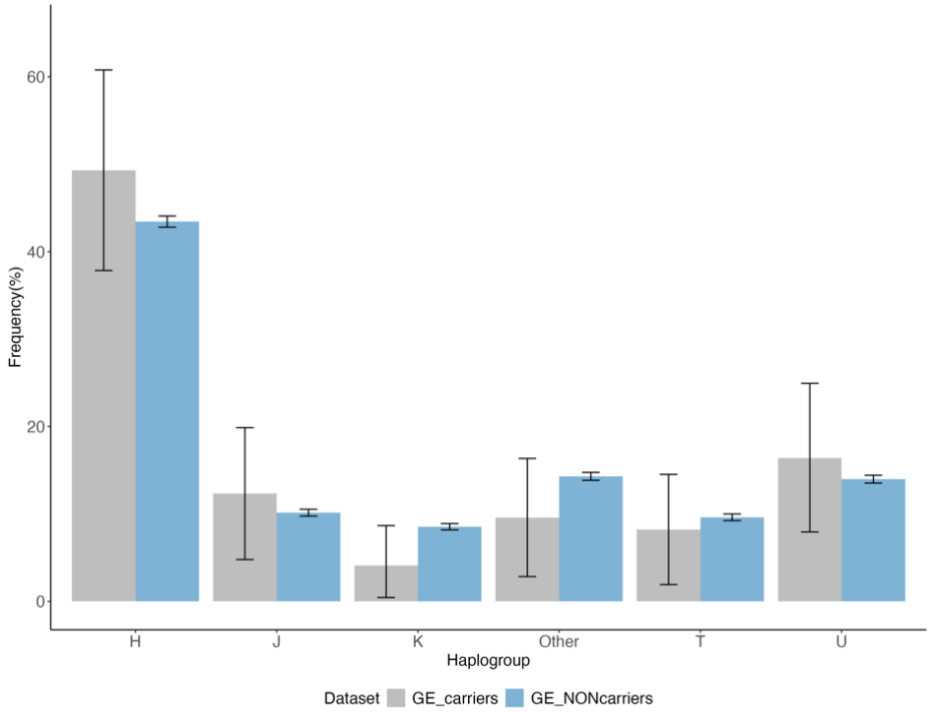

Figure 5.4: Haplogroup frequencies in 100kGP (GE) cohort m.3243A>G carrier and non-carrier individuals. Frequencies and confidence intervals retrieved while considering one individual per family, all from exclusively European nuclear genetic ancestry as per PCA analysis.

Haplogroup	Cohort	Multicentre	GE carriers	GE non-carriers	England	UKBB
H	Multicentre	NS				
	GE carriers	NS	NS			
	GE non-carriers	NS	NS	NS		
	England	NS	NS	NS	NS	
	Europe	NS	NS	NS	NS	NS
J	Multicentre	NS				
	GE carriers	NS	NS			
	GE non-carriers	NS	NS	NS		
	England	NS	NS	NS	NS	
	Europe	NS	NS	<0.00001 (1.14)	NS	NS
T	Multicentre	NS				
	GE carriers	NS	NS			
	GE non-carriers	NS	NS	NS		
	England	NS	NS	NS	NS	
	Europe	NS	NS	<0.00001 (1.8)	NS	NS
U	Multicentre	NS				
	GE carriers	NS	NS			
	GE non-carriers	NS	NS	NS		
	England	NS	NS	NS	NS	
	Europe	NS	NS	<0.00001 (3.33)	NS	NS
K	Multicentre	NS				
	GE carriers	NS	NS			
	GE non-carriers	<0.00002 (7.67)	NS	NS		
	England	<0.00001 (8.4)	NS	NS	NS	
	Europe	<0.00001 (0.8)	NS	0.000017 (7.6)	NS	NS
Other	Multicentre	NS				
	GE carriers	NS	NS			
	GE non-carriers	NS	NS	NS		
	England	NS	NS	NS	NS	
	Europe	NS	NS	<0.00001 (5.99)	NS	NS

Table 5.1: Summary statistics for the comparison of haplogroup frequencies between m.3243A>G cohorts and population estimates. Summary of p values retrieved from chi2 tests, significance threshold for 150 tests was 0.0003 (0.05/150). Percentage differences for significant comparisons are presented in parenthesis, non-significant test values are denoted by NS.

5.3.3 Distribution of m.3243A>G levels across different haplogroups

Having determined the frequency of haplogroups across m.3243A>G carrier individuals, I then asked whether m.3243A>G distribution differed between the most common haplogroups found in the multicentre and 100kGP data. Within each haplogroup, there was a wide distribution of m.3243A>G levels (**Figure 5.5**). In each of the cohorts, m.3243A>G levels were not associated with any particular haplogroup. Although not significant (chi2 test: p=0.186, p = 0.4) the median of m.3423A>G variant allele levels in both the multicentre cohort and 100kGP was the highest in samples harbouring haplogroup J. Levels of m.3243A>G are higher in the multicentre cohort compared to 100kGP as discussed in **Section 4.3.2**.

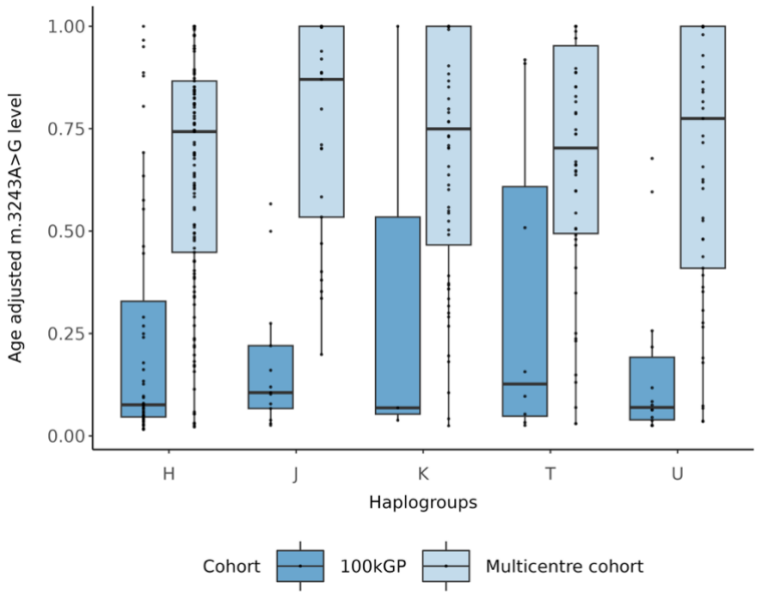

Figure 5.5: Haplogroups and distribution of m.3243A>G in the multicentre and Genomics England data. X-axis presents the haplogroups of individuals from the 100kGP cohort (blue) (n=134) and the multicentre cohort (light blue) (n= 384), plotted against their age-adjusted blood level of the m.3243A>G variant. Each dot represents an individual. Boxes show the interquartile range (IQR), which contains the middle 50% of the data (median line 2nd quartile). The bottom and top edges of the box indicate the first (25%) and third quartiles (75%), respectively. Whiskers extend to the smallest and largest values within 1.5 times the IQR.

5.3.4 mtDNA GWA analysis optimisation

Having determined that m.3243A>G levels show no association with haplogroup background, I decided to perform a mtDNA GWAS to investigate associations between individual mitochondrial SNPs and m.3243A>G levels.

To determine the optimal analysis method for a mtDNA GWAs, I first considered whether to include mtPCs as covariates. The percentage of variance explained (pve) in cohorts by the first 20 mtPCs was first visualised with a scree plot (**Figure 5.6**). The first ten principal components explained 69.04% of variability in the multicentre cohort, 69.7% in 100kGP, and 73.6% in UKBB. To test the degree of correction for population stratification (PS), mtDNA data in the multicentre cohort were used to perform GWAS analysis: (1) without including PCs, and then (2) using the first five, (3) followed by ten PCs as covariates.

The effect of the inclusion of PCs was observed through visual inspection of QQ plots and mtDNA genomic inflation factors (mtGIF) (**Figure 5.7**). Although not ideal, possibly owing to the limited number of SNPs, the mtGIF became closer to one, decreasing from 3.23 to 0.586 upon the addition of the first five PCs and then from 0.586 to 0.572 upon the addition of 10 PCs (**Figure 5.7**). Although the addition of PCs corrected for population structure and decreased the lambda from 3.23, values obtained after the addition of PCs were substantially less than one, which indicates over-correction, and raises concern over the potential of false negatives in the analyses. Referring to the scree plot below (**Figure 5.6**), using more than 10 PCs would have captured more variance, potentially yielding a larger, more acceptable inflation factor. The analyses, however, were still performed on 10 PCs as a way to avoid over correction as I hypothesized that adding more PCs would have captured the underlying family structure in the data, rather than PS. Referring to **Section 3.2.2**, REGENIE is designed to adjust and account for family structures by using the relationship matrix data.

After comparing the efficacy of linear mixed modelling and linear regression utilising software in **Chapter Three**, FaSTLMM was excluded, and the performance of two software (SAIGE vs REGENIE) was tested on mtDNA. Looking at the degree of deviation from the expected/ null line, as well as the inflation factors, SAIGE had an inflation factor closest to one, and thus is performing best with this data. However, as discussed in **Chapter Four**, to ensure consistency of results ahead of performing a META analysis, I chose to use REGENIE

to be consistent with the analysis performed in Exeter on the UKBB (Mbatchou et al., 2021c) (**Figure 5.8**).

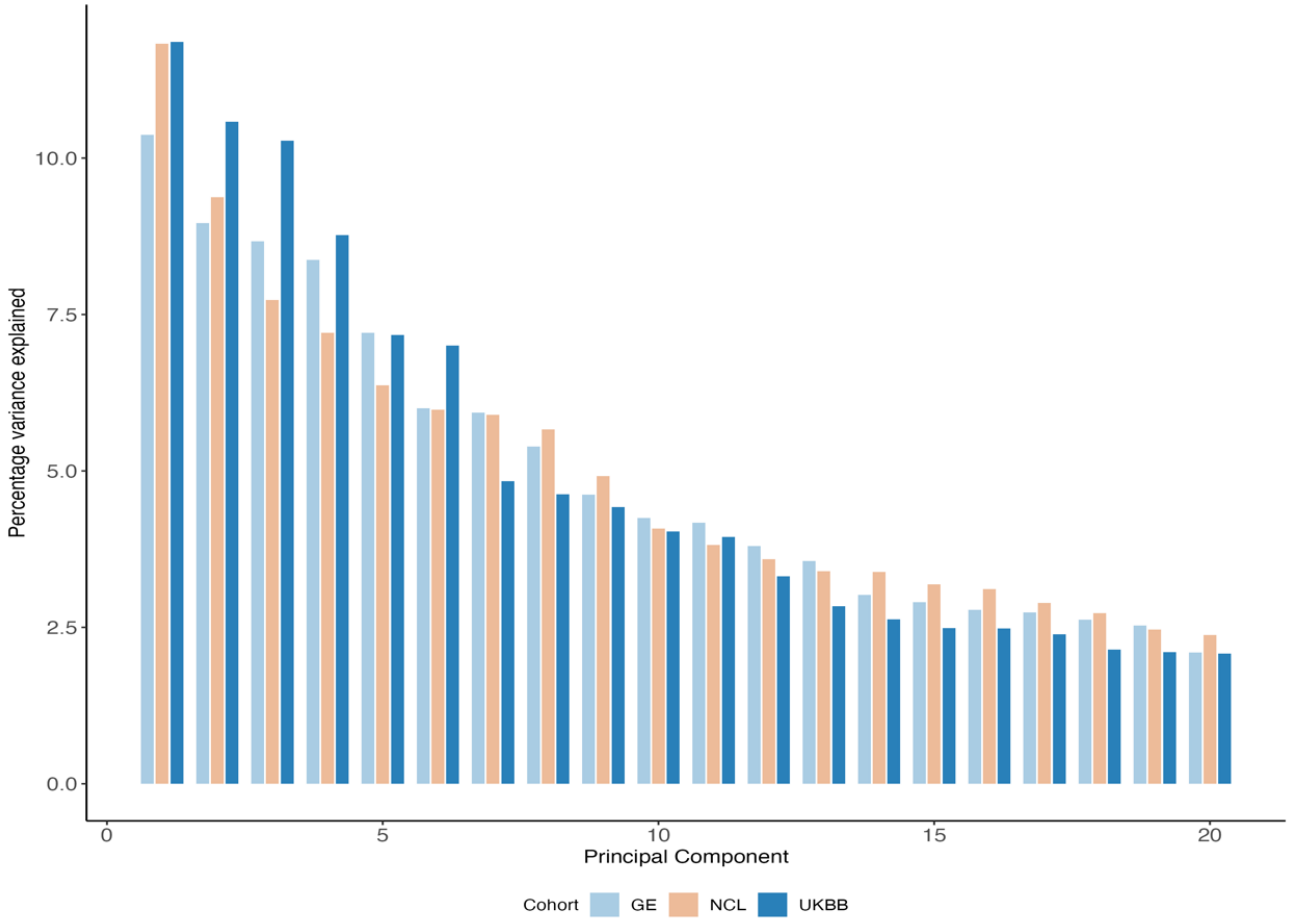

Figure 5.6: Scree plot depicting percentage of variance explained by mtPCs in each of the three cohorts.

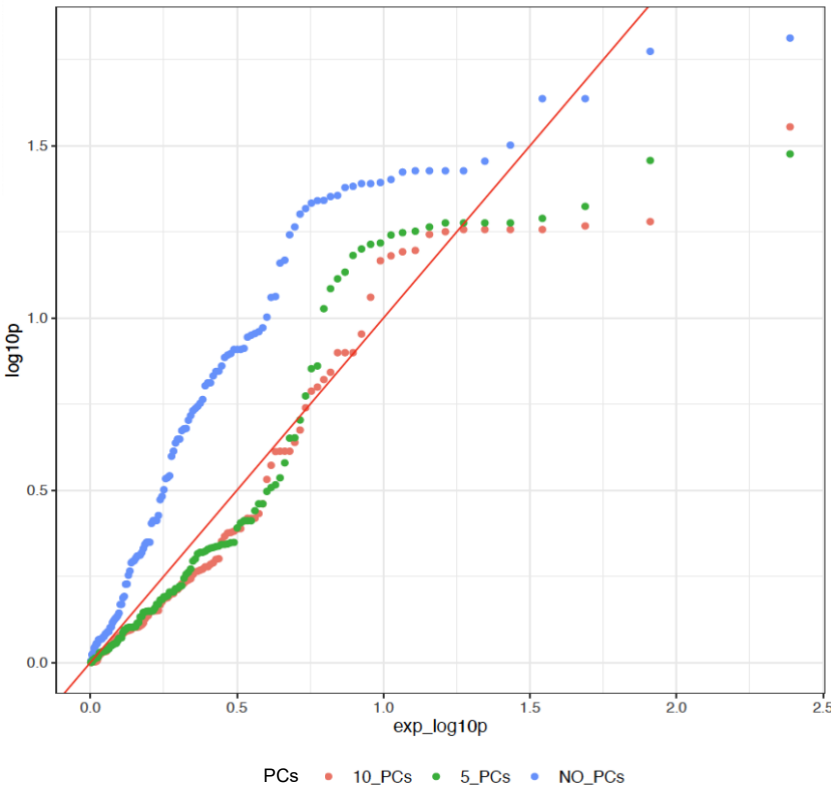

Figure 5.7: QQ plots from miWAS on the multicentre cohort with 5 then 10 mtPCs and without mtPCs using REGENIE. Results in pink were retrieved after performing the analysis with the addition of 10 mtPCs ($\lambda = 0.572$). Those in green were retrieved after the addition of 5 PCs ($\lambda = 0.582$). Compared to the results without covariates (blue, $\lambda = 3.23$), the QQ plot does not deviate from the origin of the (red) expected -log10(Pvals) line, which is the indicator of unaccounted for PS.

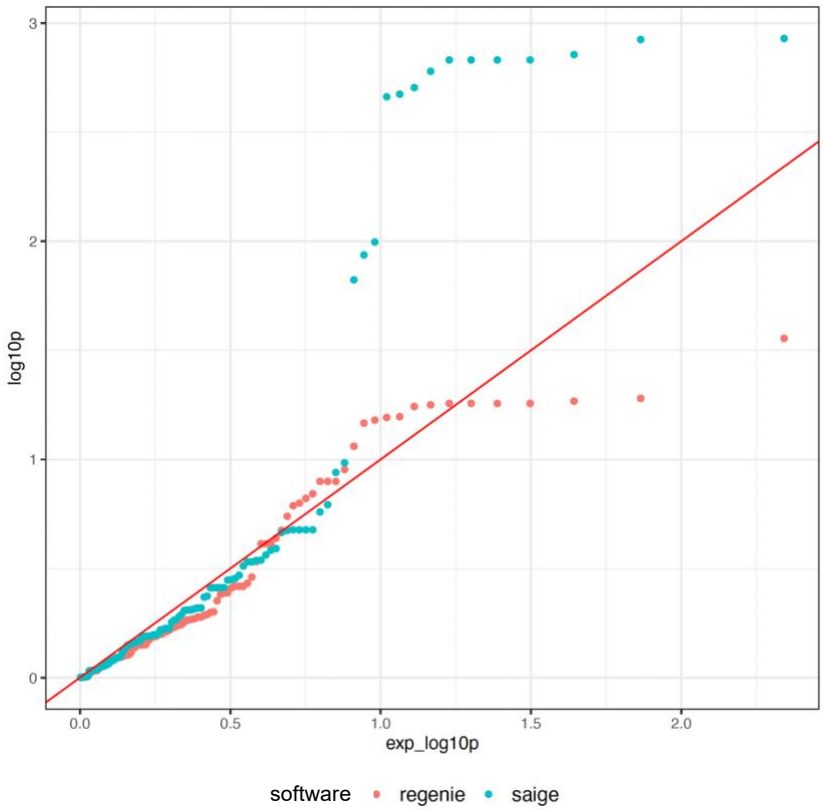

Figure 5.8: QQ plots retrieved from miWAS on the multicentre cohort with 10 mtDNA PCs using different software. Results in pink were retrieved after performing the analysis using SAIGE (λ = 0.67, in blue)) and REGENIE (λ = 0.57, in pink).

5.3.5 mtDNA GWAS

Sequencing and genotyping data of 53, 49, and 69 SNPs from the multicentre cohort, 100kGP, and the UKBB, respectively, was used to carry out an association analysis (miWAS); using REGENIE software, including the first 10 mtPCs as covariates and age-corrected m.3243A>G levels as the phenotype. In 100kGP data (**Figure 5.9-B**), one SNP (m.16356T>C, -log10(p.val) = 3.5), which is within the mitochondrial control region, and can be found on 14 separate subclades, was above the significance threshold of 3.00. On the other hand, none of the SNPs were significant in the multicentre cohort nor the UKBB. All inflation factors were well below 1.1, these low values indicate an over correction for PS, leading to under inflation of p values (false negative results). QQ plots present deviations from the null hypothesis (red line). This is known to generally reflect several things such as, cryptic relatedness, population structure, genotyping errors, and small sample sizes (Clayton et al., 2005). Additionally, in this case, the small number of tested SNPs can make the inflation factor a less stable measure. Thus, the observed increased p value in 100kGP data, at this stage, might be a false positive, particularly as it was not replicated in the UKBB data whose phenotype distribution is similar.

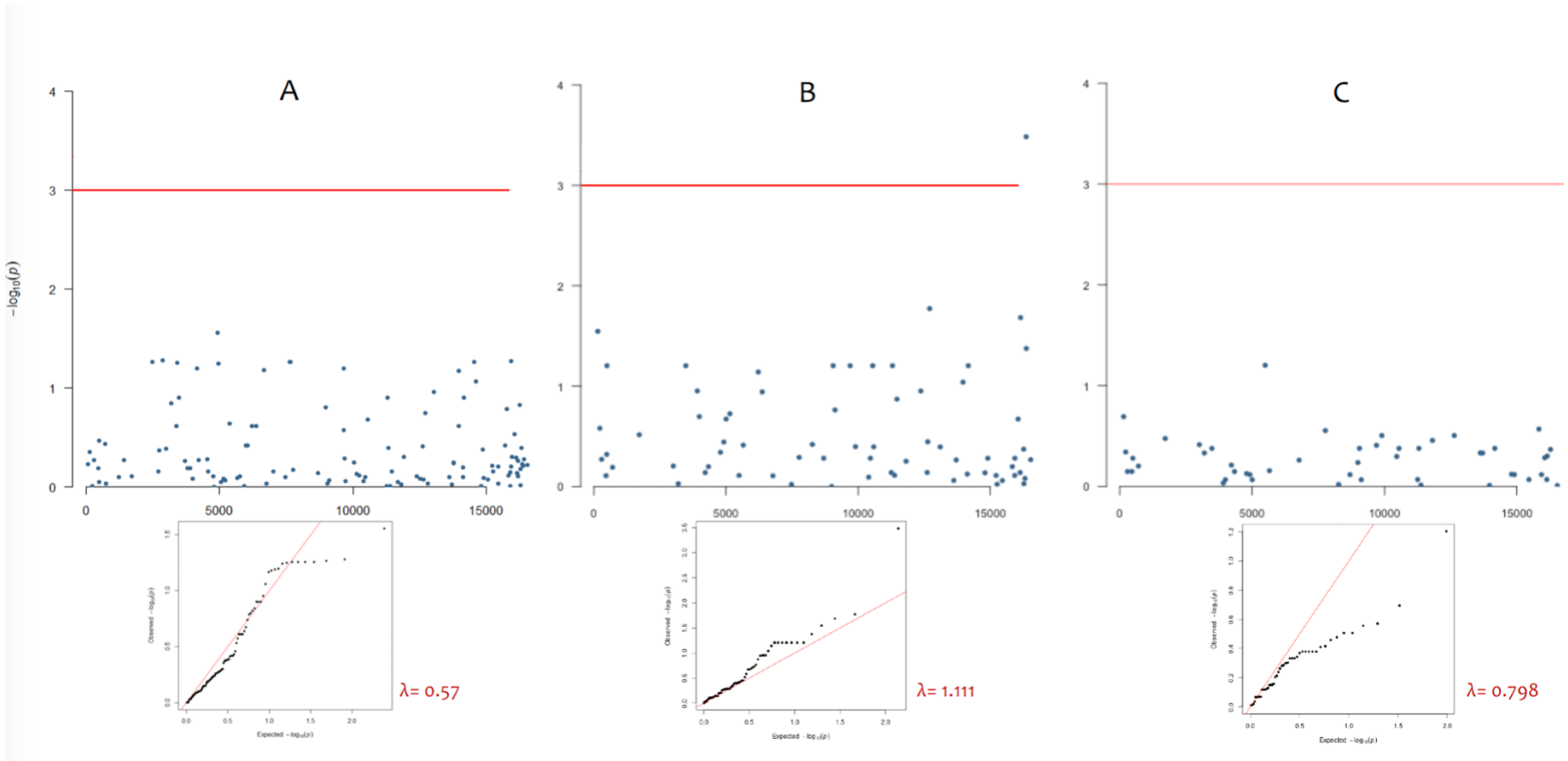

Figure 5.9: mtDNA association analysis using REGENIE. Manhattan and quantile-quantile plots retrieved from the analysis on (A) the multicentre cohort, (B) 100kGP, and (C) the UKBB. One SNP in 100kGP data (B) was above the genomic significance threshold of 3.1. All three lambda inflation factors are not exceeding 1.1, QQ plots show deviation from the null hypothesis (red line).

5.3.6 mtDNA META

META analysis combined summary statistics retrieved from both 100kGP and the UKBB. As was outlined in **Chapter Four**, due to ascertainment and the different distribution of the phenotype where levels in the multicentre cohort are higher compared to those in the public cohorts, data of the multicentre cohort was not included. None of the tested 49 SNPs were significant, including m.16356T>C identified from the GWAS on 100kGP data. Variant m.16145G>A had the lowest p value, and had a positive direction of effect in both cohorts but was well below the META significance threshold. Lambda inflation factor was correspondent to the obtained QQ plot that showed a deviation from the red-expected line, which is likely due to the small number of tested SNPs and as mentioned in **Section 5.1** the distinct LD pattern in mtDNA however, mtGIF was within the accepted range (1.06).

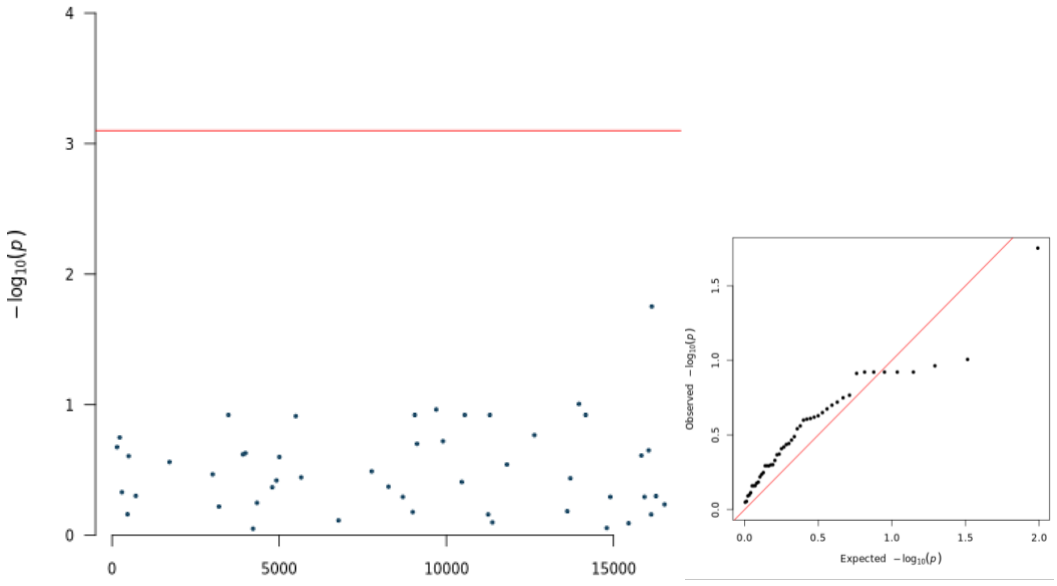

Figure 5.10: META analysis results. None of the mitochondrial SNPs were approaching the significance threshold, additionally m.16356T>C SNP observed in 100kGP did not appear to be amongst the SNPs with the lowest META p values, with SNP m.16145G>A replacing it (-log10(pval)=1.7). QQ plot deviated from the expected line, potentially reflecting the limited number of tested SNPs.

5.4 Discussion

GWA studies have primarily focused on the nuclear genome, and it is due to the characteristics of the mtDNA, such as the maternal inheritance and associated LD patterns, and lack of sufficient mtDNA genotyping/ sequencing data that made it challenging to conduct large-scale GWASs using mtDNA . Additionally, the fact that the majority of studied phenotypes are not mitochondrial related, made the mtDNA a rarer target for GWAS.

There are no universal guidelines for this branch of analysis, nonetheless, the general principles of GWAS good practice such as, uniform data quality control thresholds, phenotype standardisation, maintenance of homogeneity within participants populations, have all been meticulously followed in this study.

One SNP was above the significance threshold in 100kGP however, the fact that it was not replicated in the UKBB data, and that it was well below the significance threshold in the META analysis, indicated that it was likely to have been a false positive due to LD patterns.

Mitochondrial GWASs in complex diseases have been usually underpowered and rarely replicated, in-part due to the differences in SNP frequencies (haplogroup markers) between populations over short geographical distances, which makes it more difficult to collect a homogenous, large enough study cohort, and makes the analysis more susceptible to type 1 errors. A larger sample size than that for nuclear GWASs is needed to achieve sufficient power to detect associations to rarer haplogroups and haplogroup markers, more so given the hypermutability of mtDNA which adds a further layer of complexity (Samuels et al., 2006). Additionally, the inconsistencies in choosing a mtDNA reference sequence impair the comparability and generalizability of mtDNA GWAS studies (Ferreira and Rodriguez, 2024). However, just as the case with nuclear GWASs, small, homogenous study cohorts would be powered enough to detect associations to variants with large effect sizes. The fact that in this study we found none, indicates that if present, the mtDNA variants potentially underlying m.3243A>G heterogeneity are not with a large effect size. And thus, the notion that mtDNA variants or haplogroups are not associated with m.3243A>G levels, cannot be rejected. A larger sample size would not only enable the detection of small effect sizes, but

the inclusion of rarer variants into the analyses, which may be associated with m.3243A>G levels but were not tested due to sample size restrictions that dictated analysis power.

Comparisons of haplogroup distributions between m.3243A>G carrier and non-carrier populations showed that the frequency of haplogroup K in the multicentre cohort was significantly higher compared to both the English as well as the European population. Up to date (June 2024), an increased frequency of haplogroup K in individuals with m.3243A>G variant have not been previously reported. However, haplogroup K has been previously associated with various complex diseases. For example, a study has investigated the distribution of haplogroups between a cohort of 620 Italian, idiopathic Parkinson's disease (PD) patients, and two control groups from a matched genetic ancestry. They reported a significantly decreased frequency of haplogroup K in PD patients, suggesting a decreased risk of PD on haplogroup K backgrounds in Italians (Ghezzi et al., 2005). Another investigation performed on the Australian Blue Mountains Hearing Study, noted an increased prevalence of age-related hearing loss in individuals with haplogroup K background compared to other backgrounds, linking it to a potentially reduced mitochondrial function (Manwaring et al., 2007b).

The fact that this pattern was not consistent across all variant carrier populations (the UKBB and carriers in 100kGP), may be due to the fact that the multicentre cohort is clinically ascertained, with individuals having higher levels of m.3243A>G, and thus are associated with different heritable factors. As shown in **Figure 5.5**, individuals with haplogroup K background show no significantly increased levels of m.3243A>G however, haplogroup K background is the second most common after haplogroup H in the multicentre cohort. This may suggest an overrepresentation of a particular haplogroup K subclade in carriers of high variant levels. To elucidate this, further analysis with finer haplogroup classifications would be needed.

There have been several reports of haplogroups associated with mitochondrial function, copy number, and dynamics. The first of such examples is that conducted by Suissa et al., (2009) where they explained the observed increase of mitochondrial copy numbers on haplogroup J backgrounds. A variant that marks haplogroup J, m.295C>T, was suggested to

be causing an increased binding of TFAM, and the capacity of transcribing a region that is associated with RNA priming of mtDNA replication. Leading to significant increase of mtDNA copy numbers, with no difference in mtDNA transcript levels. This was followed by several other reports of decreased or enhanced mtDNA replication efficacy on different mitochondrial haplogroup backgrounds (Gómez-Durán et al., 2010c; Kenney et al., 2014; Gupta et al., 2023), all attributing this disparity to the effect of haplogroup marking SNPs that regulate mitochondrial functions by acting in trans with nuclear genes.

And thus, a possible explanation for this increased frequency of haplogroup K could be similar to those findings mentioned above, where a certain marker within a haplogroup K subclade preferentially replicates mtDNA molecules carrying the m.3243A>G variant and thus predisposes individuals to having elevated levels of this variant. Using another clinically ascertained cohort will be a more accurate replication to this observation.

Contrary to the study on the French population (Pierron et al., 2008), there was no evidence for haplogroup J underrepresentation in m.3243A>G carriers in the studied populations instead, haplogroup J was found to be one of the most common. And thus, this study does not support the reported 'haplogroup J paradox' observed in the French population, as haplogroup J shows no selection against m.3243A>G variant. However, haplogroup subclades in this project were not investigated, meaning that the finer genetic ancestry could not be specified, nor could we make the assumption that patients recruited in centres across the UK were exclusively of a British genetic ancestry.

As mentioned in the introduction, there is a growing number of studies that has identified mtDNA variants in association to many complex diseases such m.7028T>C and alleles m.14766C>T and an increased risk of cardiomyopathy (Fernández-Caggiano et al., 2012), m.8655C>T and type 2 diabetes, (Poulton, 2002; Yonova-Doing et al., 2021). Based on this, it was valid to hypothesise the potential of observing an overlapping association to variants reported in related phenotypes to 3243 patients – but none were found.

Chapter 6. General discussion

At the beginning of the thesis, the aims of this project were outlined and below I will summarise the results obtained from each individual chapter. This will then be followed by projects' strengths and limitations, and finally a discussion of the implications and future directions of this project.

6.1. Summary of the results

6.1.1 Data collection

To perform GWAS, the conducted power analysis estimated that with the available amount of data from m.3243A>G ascertained multicentre cohort (384 samples), there was a 95% power to detect variants with a MAF ≥ 0.05 and medium effect size of ≥ 0.6. To further increase power that would enable the detection of variants with a smaller effect size, a larger sample size is necessary (Cantor, Lange and Sinsheimer, 2010). This was the rationale behind employing heteroplasmy calling pipelines on 100kGP and UKBB data, which consequently identified 164 m.3243A>G samples from 100kGP and 147 additional samples from the UKBB.

6.1.2 Chapter 3: GWAS analysis optimisation

Data from the multicentre cohort were used to evaluate the performance of GWAS using different software, and various analysis designs (frameworks). After assessing lambda inflation factors, which reflect the degree of correction for confounding factors, excluding PCA European population-outlier samples from the analysis seemed to provide sufficient correction for data. REGENIE software was the most suitable choice given the factors thoroughly discussed in **Section 3.3.2**.

6.1.3 Chapter 4: GWAS and follow-up analysis

GWA analyses were carried out separately on each one of the three cohorts. None of the studies presented a result significant on the genome wide scale however, top SNPs identified in 100kGP as well as the multicentre cohort were above the suggestive significance threshold of 5.3. In the multicentre cohort SNP (1:114542914A>T) had a -log(Pval) value of 6.4; 100kGP's top SNP was on chromosome 15 (15:62868505G>A; -log(Pval)=6.2), and the UKBB on chromosome eight (8:128594837c>A -log(Pval)= 5.6). The cohorts had 95%, 66%, and 60% power to detect SNPs with MAF ≥5% and ES ≥ 0.6 at a genome-wide significant results, respectively.

The 100kGP and UKBB data were combined via a fixed effects **META analysis**, which identified an association peak on chromosome eight with a top SNP rsID1512802 (8:5882269G>C) having a -log10(p.val) of 6.9. This was identified as an intergenic variant falling between two non-coding pseudogenes.

Fine mapping analysis revealed that the SNPs with the lowest p values, did not coincide with SNPs in the first credible set with the greatest PIPs. In fact, the lead SNP identified by META had a PIP far below one (3.57898e-10) however, it was in LD with those in the first credible set ($0.6 < r^2 < 0.8$). Fine mapping estimated five variants to have a PIP = 1, which is the highest estimate of potential causality. These variants reside around pseudogenes, and replication of analysis on a larger sample size is needed for drawing conclusions, particularly given that the initial META peak was only approaching significance (see **Section 2.1** for further commentary). Following a series of GWAS and META analysis, **SNP based heritability analyses** were performed. Results indicated that approximately, 12.210% (SD = 2.962) and 17.717% (SD = 4.757) of heritability may be explained by the nuclear factors captured by GWAS in the multicentre GWAS and the meta-analysis, respectively.

6.1.4 Chapter 5: Mitochondrial DNA GWAS (miWAS) and differential haplogroup distribution

The only significant miWAS result was in the 100kGP m.3243A>G cohort where SNP m.16356T>C had a -log10(p.val) = 3.5; which is a haplogroup U marker. However, association with this SNP was not present in the META analysis combining 100kGP and UKBB data, indicating a potential false positive result.

Although not significant, haplogroups J was found to be more common in m.3243A>G carriers in the multicentre cohort as well the 100kGP data, which contrary to the French study (Pierron et al., 2008), indicate no underrepresentation of haplogroup J in m.3243A>G carrier individuals. This pattern, however, was not observed in the UKBB data. Additionally, frequency of haplogroup K was significantly greater in the multicentre than the general population of Europe and England ($p < 0.00001$), this pattern was not consistent across all variant carrier populations (the UKBB and carriers in 100kGP), and may be due to the fact that the multicentre cohort is clinically ascertained, with individuals having higher levels of m.3243A>G, and thus are being modulated using different heritable factors. A replication, clinically ascertained cohort of m.3243A>G carriers would be needed to test whether this observation persists.

6.2 Strengths and limitations

The main limitation of this study has been sample size, which is a common challenge in rare diseases. Although identifying m.3243A>G carrier data from the UKBB and Genomics England has contributed an additional 279 samples, neither of the individual GWASs nor the META analysis combining UKBB and 100kGP data had sufficient power to detect variants with small effect sizes. The performed META analysis had 87% power to detect variants with MAF ≥ 0.2 and ES > 0.6 at a genome wide significance. Small cohort studies, such as the GWAS on age-related macular degeneration disease, which was conducted on 96 cases and 50 controls, was successful at identifying an association with a large effect, intronic, and common variant

in the complement factor H gene (CFH) (Klein et al., 2005b). However, the fact that no such variants were identified in this study suggests that the variants underlying m.3243A>G heterogeneity are either rare or have smaller effect sizes, both of which would require a larger sample size to be detected.

Using ATAC sequencing methods on nine variant carriers, Walker et al., (2020) observed a significant, and rapid selection against m.3243A>G in T cells compared to other blood cell types, this observation was also confirmed by Franklin et al., (2023). However, bulk blood heteroplasmy measures average heteroplasmy levels across all blood cell types and blood variant heteroplasmy levels (with or without age correction) are equally good predictors of disease burden, when compared to estimates from muscle tissue (Grady et al., 2018). Nonetheless, age-correction formulae are needed to obtain an estimated measure of heteroplasmy at birth. All the presented age-correction formulae and blood heteroplasmy decline rates are estimates, and there is none that is 100% definitive (Rajasimha, Chinnery and Samuels, 2008). Franco et al., (2022) suggested that the age correction formula used in this project is under correcting for individuals with high m.3243A>G variant levels, and over correcting for those with low levels while being sufficient for those with medium levels. This stems from the observation that mutation levels in individuals with high variant levels tend to have a more aggressive decline than the proposed 2% (Rajasimha, Chinnery and Samuels, 2008), whereas those with low levels stabilise, or in some cases, increase over time, indicating a dichotomous pattern that cannot be described by a single decline rate (Franco et al., 2022b). Having said so, unless better methods to estimate heteroplasmy are developed (**Section 6.3.3** below), age-corrected measures from blood remain to be the source that was able to provide this project with the greatest number of samples.

In addition to identifying m.3243A>G carrier samples, this project utilised family tracing to identify obligate carrier individuals. The minimum detection threshold set for the m.3243A>G calling pipeline was 1%, and this was the minimal value assigned to obligate carriers before taking them forwards for age-correction. Ideally, allele frequencies of obligate carriers obtained directly from the variant calling pipeline should have been used for age-correction.

Nonetheless, we estimate that the difference between age-corrected levels from the pipeline and those based on the arbitrary age-corrected values would be minimal.

Efforts have been made with the attempt to explain the variability of m.3243A>G levels, genetic bottlenecks during inheritance (Hauswirth and Laipis, 1982b; Cruchaga et al., 2014), selection (Khrapko and Turnbull, 2014; Franco et al., 2022b), and random genetic drift (Wilson et al., 2016), have been the leading explanatory theories. However, Pickett et al., (2019) used family pedigrees and estimated that ~72% of the observed heterogeneity can be explained by heritable, genetic factors which is what this study endeavoured to discover.

Up to date, this is the first GWAS study that investigates the underlying causes of the pathogenic m.3243A>G heteroplasmy variability. Although none of the GWASs yielded a significant result, this study is a proof of concept, with applicable methods that should be further employed on larger m.324A>G carrier cohorts.

Additionally, the devised heteroplasmy calling pipelines, and data QC procedures implemented in this study would facilitate a faster identification of m.3243A>G carrier samples in further cohorts.

6.3 Implications and further directions

6.3.1 A more complex underlying structure

As mentioned previously, GWAS have been successful in identifying loci associated with a plethora of complex, rare as well as common diseases. However, unless there are large sample sizes, the detection of associations with low frequency variants ($0.5\% \leq MAF \leq 5\%$), and rare variants ($MAF < 0.5\%$) has not been a strong suit of GWAS (Lee et al., 2014a). The reason for this is multi-fold: (1) GWAS rely on the concept of LD, which facilitates the detection of casual variants, and rare variants are less likely to be found in LD with surrounding variants, which makes it a challenge for GWAS. (2) Rare variants tend to have large effect sizes, which may be detected with a relatively small sample size GWAS however, if the effect size of the rare variant is small or moderate, this will be left undetected.

Additionally, leaving rare variants unfiltered in the data ahead of a GWAS, whilst knowing that it is underpowered to detect them, leads to type 1 errors. Hence why rare variants were filtered out in this study. (3) A large sample size is often crucial to achieve a sufficient detection power, and the detection of rare variant associations would require even a larger sample size – this is because rare variants tend to have a small population frequency and analysis detection power decreases as allele frequency decreases (Asimit and Zeggini, 2010).

Momozawa & Mizukami, (2021) estimated that to detect genome-wide significant associations ($p \leq 5x10^{-8}$) with 80% power, a sample size of 100,000 individuals is needed for variants with a MAF of 0.01 and a complete correlation (ES of 1). For variants with an effect size of 0.1, a sample size of one million individuals is required. Additionally, they suggest that, because rare variants tend not to be in LD, more tests are needed for their identification and thus, an even lower significance threshold should be used, which further decreases the power (Lee et al., 2014b).

Gene or region aggregation tests could be a method worth implementing on the data used in this project, as they help overcome the issue of missing associations with rare variants. Instead of analysing individual variants, they assess the combined impact of various genetic variants within a specific gene or region, increasing the ability to detect associations when multiple variants contribute to a particular trait (Asimit et al., 2012). For example, an association between Alzheimer's disease and the *PLD3* gene was identified using a gene-burden test, which yielded a p-value of 1.4×10^{-11}. In contrast, no single variant within the gene reached a significant p value. Numerous rare variants within the *PLD3* were found to be shared among affected individuals, but their significance was limited by their very low MAF. Consequently, the gene-based test, which aggregated these rare variants, offered greater statistical power (Cruchaga et al., 2014).

6.3.2 The rise of large sequencing datasets

The availability of large datasets provides a greater detection power for analyses such as GWAS. Large WGS datasets, would enable the identification of both m.3243A>G variant carriers (via variant calling pipelines), as well as rare nuclear variants (**Section 6.2.1** above).

There are growing efforts put into both constructing large WGS patient- as well as population-based datasets and making them more accessible to researchers worldwide. For example, in 2019, the world economic forum led the pilot project: Breaking Barriers to Health Data, with genomic institutions in Australia, Canada, the UK, and US to '*create a model to share rare disease data across borders in federated data systems*', updated in 2020, it was a proof of concept that balanced the need for data access with privacy and security concerns (Thorogood, 2020).

Most recent example of available national datasets is the American 'All of US' research program, that is aiming to recruit and sequence data of more than half a million individuals across the US, from diverse, underrepresented nuclear ancestries (Bick et al., 2024). International, collaborative datasets have been limited by data sharing restrictions however, it is a constantly growing and improving area. Datasets such as The International Cancer Genome Consortium (ICGC) (Zhang et al., 2011), The Parkinson's Progression Markers Initiative (PPMI) (Marek et al., 2011), and The International Multiple Sclerosis Genetics Consortium (IMSGC) (Booth et al., 2009) are the most notable.

The recent publication of WGS data of additional 300,000 individuals' data by the UK Biobank (Li et al., 2023), is currently being analysed by colleagues, and so is a work in progress that is further expanding this project by identifying additional carrier samples.

Meanwhile, the use of animal models, such as mice, provides an idea of the possible underlying mechanisms in humans. For example, most of our understanding about germline selection and bottlenecks stemmed from initial research conducted on mouse embryos (Jenuth et al., 1996b; Cree et al., 2008b; Sharpley et al., 2012).

Due to their short reproductive cycle, large litters, and their genetic similarity to humans (mammals which are likely to have similar mechanisms), several methods have been developed to edit mouse mitochondrial genomes: such as mitochondrial-targeted zinc-finger nucleases (mtZFNs) (Gammage et al., 2014), or transcription activator-like effector nucleases (TALENs) (Bacman et al., 2013). However, the novel Double-strand break-induced DNA Deaminase Cytosine Base Editors (DdCBEs) method has provided higher precision and lower off-target effects (Mok et al., 2020). This new technology will enable the creation of further mitochondrial disease mouse models, as well as offering the potential to develop mitochondrial disease modifying cures based on mtDNA base editing (Silva-Pinheiro and Minczuk, 2022).

6.3.3 Improved heteroplasmy estimates

Blood has been routinely used for diagnosis and disease burden estimates, due to both its accuracy as well as practically. However, research is providing evidence on the incompletely accurate picture provided by bulk blood measurements (Franco et al., 2022b). This necessitates an improved measurement approach, or ideally, measurement from muscle tissue whose variant allele levels is known to remain stable over time (and so no age corrections would be needed) (Stewart and Chinnery, 2021b). This, however, is much more invasive, expensive and impractical for routine measurements and sample collections. Given access to this tissue is no longer a prerequisite for diagnosing m.3243A>G-related mitochondrial disease, adds an additional layer of complexity.

6.4 Final Conclusion

This project lays the groundwork for further exploration into the intricate relationship between genetic variation (involving both the nuclear and mitochondrial genomes) and the m.3243A>G variant, and how these influence variability in mtDNA heteroplasmy. The identification of m.3243A>G heteroplasmy associated factors would give us crucial insight into the timing this 'interplay' happens based on gene expression patterns. Whether for example, expression is enriched in early stages of oogenesis, in embryo development stages, or postnatally.

Findings from such investigations would not only further our knowledge about the mechanisms underlying m.3243A>G variant level heterogeneity, but it would provide insight into the aetiology underlying other, rarer mitochondrial pathogenic variants, such as m.8344A>G. This would ultimately improve our understanding of mitochondrial DNA disease, which holds the promise of improving patient outcomes and the lives of those that suffer from mitochondrial DNA disease.

References

Abdellaoui, A., Yengo, L., Verweij, K.J.H. and Visscher, P.M., 2023. 15 years of GWAS discovery: Realizing the promise. The American Journal of Human Genetics, 110(2), pp.179–194. https://doi.org/10.1016/j.ajhg.2022.12.011.

Abecasis, G.R., Cherny, S.S., Cookson, W.O. and Cardon, L.R., 2002. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nature Genetics, 30(1), pp.97–101. https://doi.org/10.1038/ng786.

Aguilar, I., Legarra, A., Cardoso, F., Masuda, Y., Lourenco, D. and Misztal, I., 2019. Frequentist p-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle. Genetics Selection Evolution, 51(1), p.28. https://doi.org/10.1186/s12711-019-0469-3.

Ahmed, S.T., Taylor, R.W., Turnbull, D.M., Lawless, C. and Pickett, S.J., 2022. Quantifying phenotype and genotype distributions in single muscle fibres from patients carrying the pathogenic mtDNA variant m.3243A>G. medRxiv. [online] https://doi.org/10.1101/2022.04.04.22272484.

Allkanjari, K. and Baldock, R.A., 2021. Beyond base excision repair: an evolving picture of mitochondrial DNA repair. Bioscience Reports, 41(10). https://doi.org/10.1042/BSR20211320.

Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H.L., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J.H., Staden, R. and Young, I.G., 1981a. Sequence and organization of the human mitochondrial genome. Nature, 290(5806), pp.457–465. https://doi.org/10.1038/290457a0.

Asimit, J. and Zeggini, E., 2010. Rare Variant Association Analysis Methods for Complex Traits. Annual Review of Genetics, 44(1), pp.293–308. https://doi.org/10.1146/annurev-genet-102209-163421.

Asimit, J.L., Day-Williams, A.G., Morris, A.P. and Zeggini, E., 2012. ARIEL and AMELIA: Testing for an Accumulation of Rare Variants Using Next-Generation Sequencing Data. Human Heredity, 73(2), pp.84–94. https://doi.org/10.1159/000336982.

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., et al. 2015a. A global reference for human genetic variation. Nature, 526(7571), pp.68–74. https://doi.org/10.1038/nature15393.

Awadalla, P., Eyre-Walker, A. and Smith, J.M., 1999. Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science, 286(5449), pp.2524–2525. https://doi.org/10.1126/science.286.5449.2524.

Bacman, S.R., Williams, S.L., Pinto, M., Peralta, S. and Moraes, C.T., 2013. Specific elimination of mutant mitochondrial genomes in patient-derived cells by mitoTALENs. Nature Medicine, 19(9), pp.1111–1113. https://doi.org/10.1038/nm.3261.

Barbeira, A.N., Bonazzola, R., Gamazon, E.R., Liang, Y., Park, Y., Kim-Hellmuth, S., Wang, G., Jiang, Z., Zhou, D., Hormozdiari, F., Liu, B., Rao, A., Hamel, A.R., Pividori, M.D., Aguet, F., Bastarache, L., Jordan, D.M., Verbanck, M., Do, R., Stephens, M., Ardlie, K., McCarthy, M., Montgomery, S.B., Segrè, A. V., Brown, C.D., Lappalainen, T., Wen, X. and Im, H.K., 2021. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. Genome Biology, 22(1), p.49. https://doi.org/10.1186/s13059-020-02252-4.

Battle, S.L., Puiu, D., Verlouw, J., Broer, L., Boerwinkle, E., Taylor, K.D., Rotter, J.I., Rich, S.S., Grove, M.L., Pankratz, N., Fetterman, J.L., Liu, C. and Arking, D.E., 2022. A bioinformatics pipeline for estimating mitochondrial DNA copy number and heteroplasmy levels from whole genome sequencing data. NAR Genomics and Bioinformatics, 4(2). https://doi.org/10.1093/nargab/lqac034.

Begum, F., Ghosh, D., Tseng, G.C. and Feingold, E., 2012. Comprehensive literature review and statistical considerations for GWAS meta-analysis. Nucleic Acids Research, https://doi.org/10.1093/nar/gkr1255.

Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods. 2012;58(3):268-276. doi:10.1016/j.ymeth.2012.05.001

Benner, C., Spencer, C.C.A., Havulinna, A.S., Salomaa, V., Ripatti, S. and Pirinen, M., 2016a. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics, 32(10), pp.1493–1501. https://doi.org/10.1093/bioinformatics/btw018.

Bernardino Gomes, T.M., Ng, Y.S., Pickett, S.J., Turnbull, D.M. and Vincent, A.E., 2021. Mitochondrial DNA disorders: from pathogenic variants to preventing transmission. Human Molecular Genetics, 30(R2), pp.R245–R253. https://doi.org/10.1093/hmg/ddab156.

Bick, A.G., Metcalf, G.A., Mayo, K.R., Lichtenstein, L., Rura, S., Carroll, R.J., Musick, A., Linder, J.E., Jordan, I.K., Nagar, S.D., et.al., 2024. Genomic data in the All of Us Research Program. Nature, 627(8003), pp.340–346. https://doi.org/10.1038/s41586-023-06957-x.

Biffi, A., Anderson, C.D., Nalls, M.A., Rahman, R., Sonni, A., Cortellini, L., Rost, N.S., Matarin, M., Hernandez, D.G., Plourde, A., de Bakker, P.I.W., Ross, O.A., Greenberg, S.M., Furie, K.L., Meschia, J.F., Singleton, A.B., Saxena, R. and Rosand, J., 2010. Principal-Component Analysis for Assessment of Population Stratification in Mitochondrial Medical Genetics. American Journal of Human Genetics, 86(6), pp.904–917. https://doi.org/10.1016/j.ajhg.2010.05.005.

Blok, R.B., Gook, D.A., Thorburn, D.R. and Dahl, H.-H.M., 1997. Skewed Segregation of the mtDNA nt 8993 (TrG) Mutation in Human Oocytes. The American Journal of Human Genetics, 60(6), pp.1495–1501. https://doi.org/10.1086/515453.

Bodmer, W.F., 1986. Human genetics: The molecular challenge. Cold Spring Harb. Symp. Quant. Biol, 51, pp.1–13.

Boggan, R., 2022. Using genetic linkage analysis to identify nuclear genetic modifiers of the pathogenic mtDNA variation m.3243A>G. Newcastle university.

Boggan, R.M., Ng, Y.S., Franklin, I.G., Alston, C.L., Blakely, E.L., Büchner, B., Bugiardini, E., Colclough, K., Feeney, C., Hanna, M.G., Hattersley, A.T., Klopstock, T., Kornblum, C., Mancuso, M., Patel, K.A., Pitceathly, R.D.S., Pizzamiglio, C., Prokisch, H., Schäfer, J., Schaefer, A.M., Shepherd, M.H., Thaele, A., Thomas, R.H., Turnbull, D.M., Woodward, C.E., Gorman, G.S., McFarland, R., Taylor, R.W., Cordell, H.J. and Pickett, S.J., 2022a. Defining the nuclear genetic architecture of a common maternally inherited mitochondrial disorder. medRxiv. [online] https://doi.org/10.1101/2022.11.18.22282450.

Booth, D.R., Heard, R.N., Stewart, G.J., Goris, A., Dobosi, R., Dubois, B., Lorentzen, Å.R., Celius, E.G., Harbo, H.F., Spurkland, A., Olsson, T., Kockum, I., Link, J., Hillert, J., Ban, M., Baker, A., Sawcer, S., Compston, A., Mihalova, T., Strange, R., Hawkins, C., Ingram, G., Robertson, N.P., De Jager, P.L., Hafler, D.A., Barcellos, L.F., Ivinson, A.J., Pericak-Vance, M., Oksenberg, J.R., Hauser, S.L., McCauley, J.L., Sexton, D. and Haines, J., 2009. The expanding genetic overlap between multiple sclerosis and type I diabetes. Genes & Immunity, 10(1), pp.11–14. https://doi.org/10.1038/gene.2008.83.

Bovonratwet, P., Kulm, S., Kolin, D.A., Song, J., Morse, K.W., Cunningham, M.E., Albert, T.J., Sandhu, H.S., Kim, H.J., Iyer, S., Elemento, O. and Qureshi, S.A., 2023. Identification of Novel Genetic Markers for the Risk of Spinal Pathologies. Journal of Bone and Joint Surgery, 105(11), pp.830–838. https://doi.org/10.2106/JBJS.22.00872.

Bowden, G.R., Balaresque, P., King, T.E., Hansen, Z., Lee, A.C., Pergl-Wilson, G., Hurley, E., Roberts, S.J., Waite, P., Jesch, J., Jones, A.L., Thomas, M.G., Harding, S.E. and Jobling, M.A., 2008. Excavating Past Population Structures by Surname-Based

Sampling: The Genetic Legacy of the Vikings in Northwest England. Molecular Biology and Evolution, 25(2), pp.301–309. https://doi.org/10.1093/molbev/msm255.

Burgess, D.J., 2022. Fine-mapping causal variants — why finding 'the one' can be futile. Nature Reviews Genetics, 23(5), pp.261–261. https://doi.org/10.1038/s41576-022-00484-7.

Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., Samani, et.al., 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447(7145), pp.661–678. https://doi.org/10.1038/nature05911.

Bush, W.S. and Moore, J.H., 2012. Chapter 11: Genome-Wide Association Studies. PLoS Computational Biology, 8(12), p.e1002822. https://doi.org/10.1371/journal.pcbi.1002822.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P. and Marchini, J., 2018a. The UK Biobank resource with deep phenotyping and genomic data. Nature, 562(7726), pp.203–209. https://doi.org/10.1038/s41586-018-0579-z.

Calvo, S.E., Clauser, K.R. and Mootha, V.K., 2016. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. Nucleic Acids Research, 44(D1), pp.D1251–D1257. https://doi.org/10.1093/nar/gkv1003.

Cann, R.L., Stoneking, M. and Wilson, A.C., 1987. Mitochondrial DNA and human evolution. Nature, 325(6099), pp.31–36. https://doi.org/10.1038/325031a0.

Cannon, S., Hall, T., Hawkes, G., Colclough, K., Boggan, R.M., Wright, C.F., Pickett, S.J., Hattersley, A.T., Weedon, M.N. and Patel, K.A., 2023. Large-scale blood mitochondrial genome-wide study provides novel insights into mitochondrial disease-related traits. medRxiv. [online] https://doi.org/10.1101/2023.06.12.23291273.

Canter, J.A., Kallianpur, A.R., Parl, F.F. and Millikan, R.C., 2005. Mitochondrial DNA G10398A Polymorphism and Invasive Breast Cancer in African-American Women. Cancer Research, 65(17), pp.8028–8033. https://doi.org/10.1158/0008-5472.CAN-05-1428.

Cantor, R.M., Lange, K. and Sinsheimer, J.S., 2010. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. The American Journal of Human Genetics, 86(1), pp.6–22. https://doi.org/10.1016/j.ajhg.2009.11.017.

Capelli, C., Redhead, N., Abernethy, J.K., Gratrix, F., Wilson, J.F., Moen, T., Hervig, T., Richards, M., Stumpf, M.P.H., Underhill, P.A., Bradshaw, P., Shaha, A., Thomas, M.G., Bradman, N. and Goldstein, D.B., 2003. A Y Chromosome Census of the British Isles. Current Biology, 13(11), pp.979–984. https://doi.org/10.1016/S0960-9822(03)00373-7.

Carelli, V., Achilli, A., Valentino, M.L., Rengo, C., Semino, O., Pala, M., Olivieri, A., Mattiazzi, M., Pallotti, F., Carrara, F., Zeviani, M., Leuzzi, V., Carducci, C., Valle, G., Simionati, B., Mendieta, L., Salomao, S., Belfort, R., Sadun, A.A. and Torroni, A., 2006. Haplogroup Effects and Recombination of Mitochondrial DNA: Novel Clues from the Analysis of Leber Hereditary Optic Neuropathy Pedigrees. The American Journal of Human Genetics, 78(4), pp.564–574. https://doi.org/10.1086/501236.

Carelli, V., d'Adamo, P., Valentino, M.L., La Morgia, C., Ross-Cisneros, F.N., Caporali, L., Maresca, A., Loguercio Polosa, P., Barboni, P., De Negri, A., et al., 2016. Parsing the differences in affected with LHON: genetic versus environmental triggers of disease conversion. Brain, 139(3), pp.e17–e17. https://doi.org/10.1093/brain/awv339.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M. and Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4(1), p.7. https://doi.org/10.1186/s13742-015-0047-8.

Chapman, J., Ng, Y.S. and Nicholls, T.J., 2020. The Maintenance of Mitochondrial DNA Integrity and Dynamics by Mitochondrial Membranes. Life, 10(9), p.164. https://doi.org/10.3390/life10090164.

Charles E. McCulloch and Shayle R. Searle, n.d. Generalized_Linear_and_Mixed_Models.

Chen, C.-Y., Chen, T.-T., Feng, Y.-C.A., Yu, M., Lin, S.-C., Longchamps, R.J., Wang, S.-H., Hsu, Y.-H., Yang, H.-I., Kuo, P.-H., Daly, M.J., Chen, W.J., Huang, H., Ge, T. and Lin, Y.-F., 2023. Analysis across Taiwan Biobank, Biobank Japan, and UK Biobank identifies hundreds of novel loci for 36 quantitative traits. Cell Genomics, 3(12), p.100436. https://doi.org/10.1016/j.xgen.2023.100436.

Chen, H., Chomyn, A. and Chan, D.C., 2005. Disruption of Fusion Results in Mitochondrial Heterogeneity and Dysfunction. Journal of Biological Chemistry, 280(28), pp.26185–26192. https://doi.org/10.1074/jbc.M503062200.

Chen, Y., Torroni, A., Excoffier L, Santachiara-Benerecetti AS and Wallace, D., 1995. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. Am J Hum Genet, pp.133–149.

Chick, J.M., Munger, S.C., Simecek, P., Huttlin, E.L., Choi, K., Gatti, D.M., Raghupathy, N., Svenson, K.L., Churchill, G.A. and Gygi, S.P., 2016. Defining the consequences of genetic variation on a proteome-wide scale. Nature, 534(7608), pp.500–505. https://doi.org/10.1038/nature18270.

Chinnery, P., 1997. Molecular pathology of MELAS and MERRF. The relationship between mutation load and clinical phenotypes. Brain, 120(10), pp.1713–1721. https://doi.org/10.1093/brain/120.10.1713.

Chinnery, P.F., Craven, L., Mitalipov, S., Stewart, J.B., Herbert, M. and Turnbull, D.M., 2014. The Challenges of Mitochondrial Replacement. PLoS Genetics, 10(4), p.e1004315. https://doi.org/10.1371/journal.pgen.1004315.

Chinnery, P.F., Zwijnenburg, P.J., Walker, M., Howell, N., Taylor, R.W., Lightowlers, R.N., Bindoff, L. and Turnbull, D.M., 1999. Nonrandom tissue distribution of mutant mtDNA. American journal of medical genetics, 85(5), pp.498–501.

Cirulli, E.T. and Goldstein, D.B., 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature Reviews Genetics, 11(6), pp.415–425. https://doi.org/10.1038/nrg2779.

Clayton, D.G., Walker, N.M., Smyth, D.J., Pask, R., Cooper, J.D., Maier, L.M., Smink, L.J., Lam, A.C., Ovington, N.R., Stevens, H.E., Nutland, S., Howson, J.M.M., Faham, M., Moorhead, M., Jones, H.B., Falkowski, M., Hardenbol, P., Willis, T.D. and Todd, J.A., 2005. Population structure, differential bias and genomic control in a large-scale, case-control association study. Nature Genetics, 37(11), pp.1243–1246. https://doi.org/10.1038/ng1653.

Clausen, L., Okarmus, J., Voutsinos, V. et al. 2024. PRKN-linked familial Parkinson's disease: cellular and molecular mechanisms of disease-linked variants. Cell. Mol. Life Sci. 81, 223. https://doi.org/10.1007/s00018-024-05262-8

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences. Routledge. https://doi.org/10.4324/9780203771587.

Cohen, J., 1992. Things I have learned (so far). In: Methodological issues & strategies in clinical research. Washington: American Psychological Association. pp.315–333. https://doi.org/10.1037/10109-028.

Cole, S.R. and Hernán, M.A., 2002. Fallibility in estimating direct effects. International Journal of Epidemiology, 31(1), pp.163–165. https://doi.org/10.1093/ije/31.1.163.

Collier, J.J., Oláhová, M., McWilliams, T.G. and Taylor, R.W., 2023. Mitochondrial signalling and homeostasis: from cell biology to neurological disease. Trends in Neurosciences, 46(2), pp.137–152. https://doi.org/10.1016/j.tins.2022.12.001.

Craddock, N., Hurles, M.E. and Cardin, N., 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature, 464(7289), pp.713–720. https://doi.org/10.1038/nature08979.

Craven, L., Murphy, J., Turnbull, D.M., Taylor, R.W., Gorman, G.S. and McFarland, R., 2018. Scientific and Ethical Issues in Mitochondrial Donation. The New Bioethics, 24(1), pp.57–73. https://doi.org/10.1080/20502877.2018.1440725.

Craven, L., Tuppen, H.A., Greggains, G.D., Harbottle, S.J., Murphy, J.L., Cree, L.M., Murdoch, A.P., Chinnery, P.F., Taylor, R.W., Lightowlers, R.N., Herbert, M. and Turnbull, D.M., 2010. Pronuclear transfer in human embryos to prevent transmission of mitochondrial DNA disease. Nature, 465(7294), pp.82–85. https://doi.org/10.1038/nature08958.

Cree, L.M., Samuels, D.C., de Sousa Lopes, S.C., Rajasimha, H.K., Wonnapinij, P., Mann, J.R., Dahl, H.-H.M. and Chinnery, P.F., 2008b. A reduction of mitochondrial DNA molecules during embryogenesis explains the rapid segregation of genotypes. Nature Genetics, 40(2), pp.249–254. https://doi.org/10.1038/ng.2007.63.

Cruchaga, C., Karch, C.M., Jin, S.C., Benitez, B.A., Cai, Y., Guerreiro, R., Harari, O., Norton, J., Budde, J., Bertelsen, S., Jeng, A.T., Cooper, B., Skorupa, T., et.al., 2014. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. Nature, 505(7484), pp.550–554. https://doi.org/10.1038/nature12825.

Cui, R., Elzur, R.A., Kanai, M., Ulirsch, J.C., Weissbrod, O., Daly, M.J., Neale, B.M., Fan, Z. and Finucane, H.K., 2024. Improving fine-mapping by modeling infinitesimal effects. Nature Genetics, 56(1), pp.162–169. https://doi.org/10.1038/s41588-023-01597-3.

Dandine-Roulland, C. and Perdry, H., 2016. The Use of the Linear Mixed Model in Human Genetics. Human Heredity, 80(4), pp.196–206. https://doi.org/10.1159/000447634.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G. and Durbin, R., 2011. The

variant call format and VCFtools. Bioinformatics, 27(15), pp.2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M. and Li, H., 2021. Twelve years of SAMtools and BCFtools. GigaScience, 10(2). https://doi.org/10.1093/gigascience/giab008.

Daniels, H.E., 1954. Saddlepoint Approximations in Statistics. The Annals of Mathematical Statistics, 25(4), pp.631–650. https://doi.org/10.1214/aoms/1177728652.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., Mcgue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W.G., Swaroop, A., Scott, L.J., Cucca, F., Kronenberg, F., Boehnke, M., Abecasis, G.R. and Fuchsberger, C., 2016. 1 2 8 4 VOLUME 48 | NUMBER 10 | OCTOBER. Nature GeNetics. https://doi.org/10.1038/ng.3656.

Davis, A.F. and Clayton, D.A., 1996. In situ localization of mitochondrial DNA replication in intact mammalian cells. The Journal of cell biology, 135(4), pp.883–893. https://doi.org/10.1083/jcb.135.4.883.

Deeks JJ, Higgins JPT and Altman DG, 2023. Cochrane Handbook for Systematic Reviews of Interventions version 6.4 . In: Cochrane Handbook for Systematic Reviews of Interventions, 6.4.

DeForest, N. and Majithia, A.R., 2022. Genetics of Type 2 Diabetes: Implications from Large-Scale Studies. Current Diabetes Reports, 22(5), pp.227–235. https://doi.org/10.1007/s11892-022-01462-3.

DerSimonian, R. and Laird, N., 1986. Meta-analysis in clinical trials. Controlled Clinical Trials, 7(3), pp.177–188. https://doi.org/10.1016/0197-2456(86)90046-2.

Devlin, B. and Roeder, K., 1999. Genomic Control for Association Studies. Biometrics, 55(4), pp.997–1004. https://doi.org/10.1111/j.0006-341X.1999.00997.x.

Dey, R., Schmidt, E.M., Abecasis, G.R. and Lee, S., 2017. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. The American Journal of Human Genetics, 101(1), pp.37–49. https://doi.org/10.1016/j.ajhg.2017.05.014.

Diaz, F., 2002. Human mitochondrial DNA with large deletions repopulates organelles faster than full-length genomes under relaxed copy number control. Nucleic Acids Research, 30(21), pp.4626–4633. https://doi.org/10.1093/nar/gkf602.

Dierckxsens, N., Mardulyn, P. and Smits, G., 2020. Unraveling heteroplasmy patterns with NOVOPlasty. NAR Genomics and Bioinformatics, 2(1). https://doi.org/10.1093/nargab/lqz011.

DiMauro, S., Schon, E.A., Carelli, V. and Hirano, M., 2013. The clinical maze of mitochondrial neurology. Nature Reviews Neurology, 9(8), pp.429–444. https://doi.org/10.1038/nrneurol.2013.126.

D'Souza, A.R. and Minczuk, M., 2018. Mitochondrial transcription and translation: overview. Essays in Biochemistry, 62(3), pp.309–320. https://doi.org/10.1042/EBC20170102.

Dudbridge, F. and Newcombe, P.J., 2015. Accuracy of Gene Scores when Pruning Markers by Linkage Disequilibrium. Human Heredity, 80(4), pp.178–186. https://doi.org/10.1159/000446581.

Elliott, H.R., Samuels, D.C., Eden, J.A., Relton, C.L. and Chinnery, P.F., 2008. Pathogenic Mitochondrial DNA Mutations Are Common in the General Population. The American Journal of Human Genetics, 83(2), pp.254–260. https://doi.org/10.1016/j.ajhg.2008.07.004.

Elson, J.L., Andrews, R.M., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M. and Howell, N., 2001a. Analysis of European mtDNAs for Recombination. Am. J. Hum. Genet, .

Elson, J.L., Samuels, D.C., Turnbull, D.M. and Chinnery, P.F., 2001b. Random Intracellular Drift Explains the Clonal Expansion of Mitochondrial DNA Mutations with Age. The

American Journal of Human Genetics, 68(3), pp.802–806. https://doi.org/10.1086/318801.

Falkenberg, M., Gaspari, M., Rantanen, A., Trifunovic, A., Larsson, N.-G. and Gustafsson, C.M., 2002. Mitochondrial transcription factors B1 and B2 activate transcription of human mtDNA. Nature Genetics, 31(3), pp.289–294. https://doi.org/10.1038/ng909.

Falkenberg, M. and Gustafsson, C.M., 2020. Mammalian mitochondrial DNA replication and mechanisms of deletion formation. Critical Reviews in Biochemistry and Molecular Biology, 55(6), pp.509–524. https://doi.org/10.1080/10409238.2020.1818684.

Falkenberg, M., Larsson, N.-G. and Gustafsson, C.M., 2007. DNA Replication and Transcription in Mammalian Mitochondria. Annual Review of Biochemistry, 76(1), pp.679–699. https://doi.org/10.1146/annurev.biochem.76.060305.152028.

Falkenberg, M., Larsson, N.-G. and Gustafsson, C.M., 2024. Replication and Transcription of Human Mitochondrial DNA. Annual Review of Biochemistry. https://doi.org/10.1146/annurev-biochem-052621-092014.

Fannjiang, Y., Cheng, W.-C., Lee, S.J., Qi, B., Pevsner, J., McCaffery, J.M., Hill, R.B., Basañez, G. and Hardwick, J.M., 2004. Mitochondrial fission proteins regulate programmed cell death in yeast. Genes & Development, 18(22), pp.2785–2797. https://doi.org/10.1101/gad.1247904.

Ferreira, T. and Rodriguez, S., 2024. Mitochondrial DNA: Inherent Complexities Relevant to Genetic Analyses. Genes, 15(5), p.617. https://doi.org/10.3390/genes15050617.

Fernandes, H. and Zhang, P., 2014. Overview of Molecular Diagnostics in Clinical Pathology. In: Pathobiology of Human Disease. Elsevier. pp.3287–3303. https://doi.org/10.1016/B978-0-12-386456-7.06306-1.

Fernández-Caggiano, M., Barallobre-Barreiro, J., Rego-Pérez, I., Crespo-Leiro, M.G., Paniagua, M.J., Grillé, Z., Blanco, F.J. and Doménech, N., 2012. Mitochondrial Haplogroups H and J: Risk and Protective Factors for Ischemic Cardiomyopathy. PLoS ONE, 7(8), p.e44128. https://doi.org/10.1371/journal.pone.0044128.

FIRTH, D., 1993. Bias reduction of maximum likelihood estimates. Biometrika, 80(1), pp.27–38. https://doi.org/10.1093/biomet/80.1.27.

Fisher, R.P. and Clayton, D.A., 1988. Purification and Characterization of Human Mitochondrial Transcription Factor 1. Molecular and Cellular Biology, 8(8), pp.3496–3509. https://doi.org/10.1128/mcb.8.8.3496-3509.1988.

Fitzgerald, M.-C., O'Halloran, P.J., Connolly, N.M.C. and Murphy, B.M., 2022. Targeting the apoptosis pathway to treat tumours of the paediatric nervous system. Cell Death & Disease, 13(5), p.460. https://doi.org/10.1038/s41419-022-04900-y.

Fox, J., 2002. Linear mixed models: Appendix to emphAn R and S-PLUS Companion to Applied Regression. Relation, .

Franco, M., Pickett, S.J., Fleischmann, Z., Khrapko, M., Cote-L'Heureux, A., Aidlen, D., Stein, D., Markuzon, N., Popadin, K., Braverman, M., Woods, D.C., Tilly, J.L., Turnbull, D.M. and Khrapko, K., 2022b. Dynamics of the most common pathogenic mtDNA variant m.3243A > G demonstrate frequency-dependency in blood and positive selection in the germline. Human Molecular Genetics, 31(23), pp.4075–4086. https://doi.org/10.1093/hmg/ddac149.

Franklin, I.G., Milne, P., Childs, J., Boggan, R.M., Barrow, I., Lawless, C., Gorman, G.S., Ng, Y.S., Collin, M., Russell, O.M. and Pickett, S.J., 2023. T cell differentiation drives the negative selection of pathogenic mitochondrial DNA variants. Life Science Alliance, 6(11), p.e202302271. https://doi.org/10.26508/lsa.202302271.

Fry, A., Littlejohns, T.J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R. and Allen, N.E., 2017. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. American Journal of Epidemiology, 186(9), pp.1026–1034. https://doi.org/10.1093/aje/kwx246.

Gałecki, A. and Burzykowski, T., 2013. Linear Mixed-Effects Model. pp.245–273. https://doi.org/10.1007/978-1-4614-3900-4_13.

Gammage, P.A., Rorbach, J., Vincent, A.I., Rebar, E.J. and Minczuk, M., 2014. Mitochondrially targeted <scp>ZFN</scp> s for selective degradation of pathogenic mitochondrial genomes bearing large-scale deletions or point mutations. EMBO Molecular Medicine, 6(4), pp.458–466. https://doi.org/10.1002/emmm.201303672.

Garg, S., Zimorski, V. and Martin, W.F., 2016. Endosymbiotic Theory. In: R.M. Kliman, ed. Encyclopedia of Evolutionary Biology. [online] Oxford: Academic Press. pp.511–517. https://doi.org/https://doi.org/10.1016/B978-0-12-800049-6.00191-8.

Gavaghan, D.J., Moore, A.R. and McQuay, H.J., 2000. An evaluation of homogeneity tests in meta-analyses in pain using simulations of individual patient data. Pain, 85(3), pp.415–424. https://doi.org/10.1016/S0304-3959(99)00302-4.

Ghezzi, D., Marelli, C., Achilli, A., Goldwurm, S., Pezzoli, G., Barone, P., Pellecchia, M.T., Stanzione, P., Brusa, L., Bentivoglio, A.R., Bonuccelli, U., Petrozzi, L., Abbruzzese, G., Marchese, R., Cortelli, P., Grimaldi, D., Martinelli, P., Ferrarese, C., Garavaglia, B., Sangiorgi, S., Carelli, V., Torroni, A., Albanese, A. and Zeviani, M., 2005. Mitochondrial DNA haplogroup K is associated with a lower risk of Parkinson's disease in Italians. European Journal of Human Genetics, 13(6), pp.748–752. https://doi.org/10.1038/sj.ejhg.5201425.

Ghoussaini, M., Mountjoy, E., Carmona, M., Peat, G., Schmidt, E.M., Hercules, A., Fumis, L., Miranda, A., Carvalho-Silva, D., Buniello, A., et al., 2021. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. Nucleic Acids Research, 49(D1), pp.D1311–D1320. https://doi.org/10.1093/nar/gkaa840.

Gilmour, A.R., Thompson, R. and Cullis, B.R., 1995. Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. [online] Available at: <https://about.jstor.org/terms>.

Glass, G. V, 1976. Primary, Secondary, and Meta-Analysis of Research. Educational Researcher, 5(10), pp.3–8. https://doi.org/10.3102/0013189X005010003.

Gloyn, A.L., Weedon, M.N., Owen, K.R., Turner, M.J., Knight, B.A., Hitman, G., Walker, M., Levy, J.C., Sampson, M., Halford, S., McCarthy, M.I., Hattersley, A.T. and Frayling, T.M., 2003. Large-Scale Association Studies of Variants in Genes Encoding the Pancreatic β-Cell KATP Channel Subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) Confirm That the KCNJ11 E23K Variant Is Associated With Type 2 Diabetes. Diabetes, 52(2), pp.568–572. https://doi.org/10.2337/diabetes.52.2.568.

Glynos, A., Bozhilova, L. V., Frison, M., Burr, S., Stewart, J.B. and Chinnery, P.F., 2023. High-throughput single-cell analysis reveals progressive mitochondrial DNA mosaicism throughout life. Science Advances, 9(43). https://doi.org/10.1126/sciadv.adi4038.

Gómez-Durán, A., Pacheu-Grau, D., López-Gallardo, E., Díez-Sánchez, C., Montoya, J., López-Pérez, M.J. and Ruiz-Pesini, E., 2010a. Unmasking the causes of multifactorial disorders: OXPHOS differences between mitochondrial haplogroups. Human Molecular Genetics, 19(17), pp.3343–3353. https://doi.org/10.1093/hmg/ddq246.

Gómez-Durán, A., Pacheu-Grau, D., López-Gallardo, E., Díez-Sánchez, C., Montoya, J., López-Pérez, M.J. and Ruiz-Pesini, E., 2010b. Unmasking the causes of multifactorial disorders: OXPHOS differences between mitochondrial haplogroups. Human Molecular Genetics, 19(17), pp.3343–3353. https://doi.org/10.1093/hmg/ddq246.

Gorman, G.S., McFarland, R., Stewart, J., Feeney, C. and Turnbull, D.M., 2018. Mitochondrial donation: from test tube to clinic. The Lancet, 392(10154), pp.1191–1192. https://doi.org/10.1016/S0140-6736(18)31868-3.

Goto, Y., Nonaka, I. and Horai, S., 1990. A mutation in the tRNALeu(UUR) gene associated with the MELAS subgroup of mitochondrial encephalomyopathies. Nature, 348(6302), pp.651–653. https://doi.org/10.1038/348651a0.

Grady, J.P., Campbell, G., Ratnaike, T., Blakely, E.L., Falkous, G., Nesbitt, V., Schaefer, A.M., McNally, R.J., Gorman, G.S., Taylor, R.W., Turnbull, D.M. and McFarland, R., 2014. Disease progression in patients with single, large-scale mitochondrial DNA deletions. Brain, 137(2), pp.323–334. https://doi.org/10.1093/brain/awt321.

Grady, J.P., Pickett, S.J., Ng, Y.S., Alston, C.L., Blakely, E.L., Hardy, S.A., Feeney, C.L., Bright, A.A., Schaefer, A.M., Gorman, G.S., Mcnally, R.J., Taylor, R.W., Turnbull, D.M. and Mcfarland, R., 2018. mtDNA heteroplasmy level and copy number indicate disease burden in m.3243A>G mitochondrial disease. https://doi.org/10.15252/emmm.201708262.

Green, S.B., 1991. How Many Subjects Does It Take To Do A Regression Analysis. Multivariate Behavioral Research, 26(3), pp.499–510. https://doi.org/10.1207/s15327906mbr2603_7.

Greene, D., Pirri, D., Frudd, K., Sackey, E., Al-Owain, M., Giese, A.P.J., Ramzan, K., Riaz, S., Yamanaka, I., Boeckx, N., Thys, C., Gelb, B.D., Brennan, P., Hartill, V., Harvengt, J., Kosho, T., Mansour, S., Masuno, M., Ohata, T., Stewart, H., Taibah, K., Turner, C.L.S., Imtiaz, F., Riazuddin, S., Morisaki, T., Ostergaard, P., Loeys, B.L., Morisaki, H., Ahmed, Z.M., Birdsey, G.M., Freson, K., Mumford, A. and Turro, E., 2023. Genetic association analysis of 77,539 genomes reveals rare disease etiologies. Nature Medicine, 29(3), pp.679–688. https://doi.org/10.1038/s41591-023-02211-z.

Guan, Y. and Stephens, M., 2011. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. The Annals of Applied Statistics, 5(3). https://doi.org/10.1214/11-AOAS455.

Guha, P., Srivastava, S.K., Bhattacharjee, S. and Chaudhuri, T.K., 2013. Human migration, diversity and disease association: a convergent role of established and emerging DNA markers. Frontiers in Genetics, 4. https://doi.org/10.3389/fgene.2013.00155.

Gupta, R., Kanai, M., Durham, T.J., Tsuo, K., McCoy, J.G., Kotrys, A. V., Zhou, W., Chinnery, P.F., Karczewski, K.J., Calvo, S.E., Neale, B.M. and Mootha, V.K., 2023. Nuclear genetic control of mtDNA copy number and heteroplasmy in humans. Nature, 620(7975), pp.839–848. https://doi.org/10.1038/s41586-023-06426-5.

Gurinovich, A., Li, M., Leshchyk, A., Bae, H., Song, Z., Arbeev, K.G., Nygaard, M., Feitosa, M.F., Perls, T.T. and Sebastiani, P., 2022. Evaluation of GENESIS, SAIGE, REGENIE

and fastGWA-GLMM for genome-wide association studies of binary traits in correlated data. Frontiers in Genetics, 13. https://doi.org/10.3389/fgene.2022.897210.

Haas, R.H., 2019. biology Editorial Mitochondrial Dysfunction in Aging and Diseases of Aging. [online] https://doi.org/10.3390/biology8020048.

Hägg, S., Jylhävä, J., Wang, Y., Czene, K. and Grassmann, F., 2021. Deciphering the genetic and epidemiological landscape of mitochondrial DNA abundance. Human Genetics, [online] 140(6), pp.849–861. https://doi.org/10.1007/s00439-020-02249-w.

Hall, A.R., Burke, N., Dongworth, R.K. and Hausenloy, D.J., 2014. Mitochondrial fusion and fission proteins: novel therapeutic targets for combating cardiovascular disease. British Journal of Pharmacology, 171(8), pp.1890–1906. https://doi.org/10.1111/bph.12516.

Hans, C., Dobra, A. and West, M., 2007. Shotgun Stochastic Search for "Large p " Regression. Journal of the American Statistical Association, 102(478), pp.507–516. https://doi.org/10.1198/016214507000000121.

Hatefi, Y., 1985. THE MITOCHONDRIAL ELECTRON TRANSPORT AND OXIDATIVE PHOSPHORYLATION SYSTEM. Annual Review of Biochemistry, 54(1), pp.1015–1069. https://doi.org/10.1146/annurev.bi.54.070185.005055.

Hauswirth, W.W. and Laipis, P.J., 1982a. Mitochondrial DNA polymorphism in a maternal lineage of Holstein cows. Proceedings of the National Academy of Sciences, 79(15), pp.4686–4690. https://doi.org/10.1073/pnas.79.15.4686.

Hauswirth, W.W. and Laipis, P.J., 1982b. Mitochondrial DNA polymorphism in a maternal lineage of Holstein cows. Proceedings of the National Academy of Sciences, 79(15), pp.4686–4690. https://doi.org/10.1073/pnas.79.15.4686.

Hay, M., 2018. European mtDNA haplogroups frequencies by country. https://www.eupedia.com/europe/european_mtdna_haplogroups_frequency.shtml .

Heitzer, E., Auinger, L. and Speicher, M.R., 2020. Cell-Free DNA and Apoptosis: How Dead Cells Inform About the Living. Trends in Molecular Medicine, 26(5), pp.519–528. https://doi.org/10.1016/j.molmed.2020.01.012.

Hellwege, J.N., Keaton, J.M., Giri, A., Gao, X., Velez Edwards, D.R. and Edwards, T.L., 2017. Population Stratification in Genetic Association Studies. Current Protocols in Human Genetics, 95(1). https://doi.org/10.1002/cphg.48.

Hernansanz-Agustín, P. and Enríquez, J.A., 2021. Generation of Reactive Oxygen Species by Mitochondria. Antioxidants, 10(3), p.415. https://doi.org/10.3390/antiox10030415.

Hertel, J.K.H., Johansson, S., Midthjell, K., Nygård, O., Njølstad, P.R. and Molven, A., 2013. Type 2 diabetes genes – Present status and data from Norwegian studies. Norsk Epidemiologi, 23(1). https://doi.org/10.5324/nje.v23i1.1597.

Higgins, J.P.T. and Thompson, S.G., 2002. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine, 21(11), pp.1539–1558. https://doi.org/10.1002/sim.1186.

Hirschhorn, J.N. and Daly, M.J., 2005. Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics, 6(2), pp.95–108. https://doi.org/10.1038/nrg1521.

Hocking, L.J., Andrews, C., Armstrong, C., Ansari, M., Baty, D., Berg, J., Bradley, T., Clark, C., Diamond, A., Doherty, J., Lampe, A., McGowan, R., Moore, D.J., O'Sullivan, D., Purvis, A., Santoyo-Lopez, J., Westwood, P., Abbott, M., Williams, N., Aitman, T.J., Biankin, A. V., Cooke, S.L., Humphrey, W.I., Martin, S., Meynert, A., Murphy, F., Nourse, C., Semple, C.A., Williams, N., Dean, J., Foley, P., Robertson, L., Ross, A., Williamson, K., Berg, J., Goudie, D., McWilliam, C., Fitzpatrick, D., Fletcher, E., Jackson, A., Lam, W., Porteous, M., Barr, K., Bradshaw, N., Davidson, R., Gardiner, C., Gorrie, J., Hague, R., Hamilton, M., Joss, S., Kinning, E., Longman, C., Martin, N., McGowan, R., Paterson, J., Pilz, D., Snadden, L., Tobias, E., Wedderburn, S., Whiteford, M., Aitman, T.J. and Miedzybrodzka, Z., 2023. Genome sequencing with gene panel-based analysis for rare inherited conditions in a publicly funded

healthcare system: implications for future testing. European Journal of Human Genetics, 31(2), pp.231–238. https://doi.org/10.1038/s41431-022-01226-3.

Holmes, M. V., Ala-Korpela, M. and Smith, G.D., 2017. Mendelian randomization in cardiometabolic disease: challenges in evaluating causality. Nature Reviews Cardiology, 14(10), pp.577–590. https://doi.org/10.1038/nrcardio.2017.78.

Hong, Y.S., Battle, S.L., Shi, W., Puiu, D., Pillalamarri, V., Xie, J., Pankratz, N., Lake, N.J., Lek, M., Rotter, J.I., Rich, S.S., Kooperberg, C., Reiner, A.P., Auer, P.L., Heard-Costa, N., Liu, C., Lai, M., Murabito, J.M., Levy, D., Grove, M.L., Alonso, A., Gibbs, R., Dugan-Perez, S., Gondek, L.P., Guallar, E. and Arking, D.E., 2023. Deleterious heteroplasmic mitochondrial mutations are associated with an increased risk of overall and cancer-specific mortality. Nature Communications, 14(1), p.6113. https://doi.org/10.1038/s41467-023-41785-7.

Hoppins, S., Edlich, F., Cleland, M.M., Banerjee, S., McCaffery, J.M., Youle, R.J. and Nunnari, J., 2011. The Soluble Form of Bax Regulates Mitochondrial Fusion via MFN2 Homotypic Complexes. Molecular Cell, 41(2), pp.150–160. https://doi.org/10.1016/j.molcel.2010.11.030.

Horan, M.P. and Cooper, D.N., 2014. The emergence of the mitochondrial genome as a partial regulator of nuclear function is providing new insights into the genetic mechanisms underlying age-related complex disease. Human Genetics, 133(4), pp.435–458. https://doi.org/10.1007/s00439-013-1402-4.

Horan, M.P., Gemmell, N.J. and Wolff, J.N., 2013a. From evolutionary bystander to master manipulator: the emerging roles for the mitochondrial genome as a modulator of nuclear gene expression. European Journal of Human Genetics, 21(12), pp.1335–1337. https://doi.org/10.1038/ejhg.2013.75.

Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B. and Eskin, E., 2014. Identifying Causal Variants at Loci with Multiple Signals of Association. Genetics, 198(2), pp.497–508. https://doi.org/10.1534/genetics.114.167908.

Howell, N., Halvorson, S., Kubacka, I., McCullough, D.A., Bindoff, L.A. and Turnbull, D.M., 1992. Mitochondrial gene segregation in mammals: is the bottleneck always narrow? Human Genetics, 90(1–2). https://doi.org/10.1007/BF00210753.

Hudson, G., Carelli, V., Spruijt, L., Gerards, M., Mowbray, C., Achilli, A., Pyle, A., Elson, J., Howell, N., La Morgia, C., Valentino, M.L., Huoponen, K., Savontaus, M.L., Nikoskelainen, E., Sadun, A.A., Salomao, S.R., Belfort, R., Griffiths, P., Man, P.Y.W., De Coo, R.F.M., Horvath, R., Zeviani, M., Smeets, H.J.T., Torroni, A. and Chinnery, P.F., 2007. Clinical expression of leber hereditary optic neuropathy is affected by the mitochondrial DNA-haplogroup background. American Journal of Human Genetics, 81(2), pp.228–233. https://doi.org/10.1086/519394.

Hudson, G., Gomez-Duran, A., Wilson, I.J. and Chinnery, P.F., 2014a. Recent Mitochondrial DNA Mutations Increase the Risk of Developing Common Late-Onset Human Diseases. PLoS Genetics, 10(5), p.e1004369. https://doi.org/10.1371/journal.pgen.1004369.

Hudson, G., Keers, S., Yu, P., Man, W., Griffiths, P., Huoponen, K., Savontaus, M.-L., Nikoskelainen, E., Zeviani, M., Carrara, F., Horvath, R., Karcagi, V., Smeets, H.J.M. and Chinnery, P.F., 2005. Identification of an X-Chromosomal Locus and Haplotype Modulating the Phenotype of a Mitochondrial DNA Disorder. Am. J. Hum. Genet, .

Hudson, G., Nalls, M., Evans, J.R., Breen, D.P., Winder-Rhodes, S., Morrison, K.E., Morris, H.R., Williams-Gray, C.H., Barker, R.A., Singleton, A.B., Hardy, J., Wood, N.E., Burn, D.J. and Chinnery, P.F., 2013a. Two-stage association study and meta-analysis of mitochondrial DNA variants in Parkinson disease. Neurology, 80(22), pp.2042–2048. https://doi.org/10.1212/WNL.0b013e318294b434.

Hudson, G., Takeda, Y. and Herbert, M., 2019. Reversion after replacement of mitochondrial DNA. Nature, 574(7778), pp.E8–E11. https://doi.org/10.1038/s41586-019-1623-3.

Huerta, C., Castro, M.G., Coto, E., Blázquez, M., Ribacoba, R., Guisasola, L.M., Salvador, C., Martínez, C., Lahoz, C.H. and Alvarez, V., 2005. Mitochondrial DNA polymorphisms and risk of Parkinson's disease in Spanish population. Journal of the Neurological Sciences, 236(1–2), pp.49–54. https://doi.org/10.1016/j.jns.2005.04.016.

Hutchin, T. and Cortopassi, G., 1995. A mitochondrial DNA clone is associated with increased risk for Alzheimer disease. Proceedings of the National Academy of Sciences, 92(15), pp.6892–6895. https://doi.org/10.1073/pnas.92.15.6892.

Hwang, C.-H., Lee, N.-K. and Paik, H.-D., 2022. The Anti-Cancer Potential of Heat-Killed Lactobacillus brevis KU15176 upon AGS Cell Lines through Intrinsic Apoptosis Pathway. International Journal of Molecular Sciences, 23(8), p.4073. https://doi.org/10.3390/ijms23084073.

Idaghdour, Y. and Hodgkinson, A., 2017. Integrated genomic analysis of mitochondrial RNA processing in human cancers. Genome Medicine, 9(1), p.36. https://doi.org/10.1186/s13073-017-0426-0.

Inak, G., Rybak-Wolf, A., Lisowski, P., Pentimalli, T.M., Jüttner, R., Glažar, P., Uppal, K., Bottani, E., Brunetti, D., Secker, C., Zink, A., Meierhofer, D., Henke, M.-T., Dey, et al., 2021. Defective metabolic programming impairs early neuronal morphogenesis in neural cultures and an organoid model of Leigh syndrome. Nature Communications, 12(1), p.1929. https://doi.org/10.1038/s41467-021-22117-z.

Jacobsen, H.P., Herskind, A.M., Nielsen, B.W. and Husby, S., 2001. IgE in unselected like-sexed monozygotic and dizygotic twins at birth and at 6 to 9 years of age: High but dissimilar genetic influence on IgE levels. Journal of Allergy and Clinical Immunology, 107(4), pp.659–663. https://doi.org/10.1067/mai.2001.113565.

Jadhav, B., Garg, P., van Vugt, J.J.F.A., Garikano, K.I., Gagliardi, D., Lee, W., Martin-Trujillo, A., Gies, S.L., Barbosa, M., Jain, M., Houlden, H., Paten, B., Genomics England Research Consortium, Project MinE ALS Sequencing Consortium, Veldink, J., Tucci, A. and Sharp, A.J., 2023. A GCC repeat expansion in AFF3 is a significant cause of

intellectual disability. medRxiv : the preprint server for health sciences. [online] https://doi.org/10.1101/2023.05.03.23289461.

Janssen, G.M.C., Maassen, J.A. and van den Ouweland, J.M.W., 1999. The Diabetes-associated 3243 Mutation in the Mitochondrial tRNALeu(UUR) Gene Causes Severe Mitochondrial Dysfunction without a Strong Decrease in Protein Synthesis Rate. Journal of Biological Chemistry, 274(42), pp.29744–29748. https://doi.org/10.1074/jbc.274.42.29744.

Jenuth, J.P., Peterson, A.C., Fu, K. and Shoubridge, E.A., 1996a. Random genetic drift in the female germline explains the rapid segregation of mammalian mitochondrial DNA. Nature Genetics, 14(2), pp.146–151. https://doi.org/10.1038/ng1096-146.

Jiang, D., Mbatchou, J. and McPeek, M.S., 2015. Retrospective Association Analysis of Binary Traits: Overcoming Some Limitations of the Additive Polygenic Model. Human Heredity, 80(4), pp.187–195. https://doi.org/10.1159/000446957.

J.Morten, K., Poulton, J. and Sykes, B., 1995. Multiple independent occurrence of the 3243 mutation in mitochondrial tRNA leuUUR in patients with the MELAS phenotype. Human Molecular Genetics, 4(9), pp.1689–1691. https://doi.org/10.1093/hmg/4.9.1689.

Johnson, R.C., Nelson, G.W., Troyer, J.L., Lautenberger, J.A., Kessing, B.D., Winkler, C.A. and O'Brien, S.J., 2010. Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics, 11(1), p.724. https://doi.org/10.1186/1471-2164-11-724.

Joiret, M., Mahachie John, J.M., Gusareva, E.S. and Van Steen, K., 2019. Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. BioData Mining, [online] 12(1), p.11. https://doi.org/10.1186/s13040-019-0199-7.

Kakarla, R., Hur, J., Kim, Y.J., Kim, J. and Chwae, Y.-J., 2020. Apoptotic cell-derived exosomes: messages from dying cells. Experimental & Molecular Medicine, 52(1), pp.1–6. https://doi.org/10.1038/s12276-019-0362-8.

Kanai, M., Elzur, R., Zhou, W., Daly, M.J., Finucane, H.K., Zhou, W., Kanai, M., Wu, K.-H.H., Rasheed, H., Tsuo, K., et al., Biobank of the Americas, Biobank Japan Project, BioMe, BioVU, CanPath - Ontario Health Study, China Kadoorie Biobank Collaborative Group, Colorado Center for Personalized Medicine, deCODE Genetics, Estonian Biobank, F., Generation Scotland, Genes & Health Research Team, LifeLines, Mass General Brigham Biobank, Michigan Genomics Initiative, National Biobank of Korea, Penn Medicine BioBank, Qatar Biobank, The Qskin Sun and Health Study, Taiwan Biobank, The Hunt Study, Ucla Atlas Community Health Initiative, Uganda Genome Resource, Uk Biobank, Martin, A.R., Willer, C.J., Daly, M.J. and Neale, B.M., 2022. Meta-analysis fine-mapping is often miscalibrated at single-variant resolution. Cell Genomics, 2(12), p.100210. https://doi.org/10.1016/j.xgen.2022.100210.

Kang, E., Wu, J., Gutierrez, N.M., Koski, A., Tippner-Hedges, R., Agaronyan, K., Platero-Luengo, A., Martinez-Redondo, P., Ma, H., Lee, Y., Hayama, T., Van Dyken, C., Wang, X., Luo, S., Ahmed, R., Li, Y., Ji, D., Kayali, R., Cinnioglu, C., Olson, S., Jensen, J., Battaglia, D., Lee, D., Wu, D., Huang, T., Wolf, D.P., Temiakov, D., Belmonte, J.C.I., Amato, P. and Mitalipov, S., 2016. Mitochondrial replacement in human oocytes carrying pathogenic mitochondrial DNA mutations. Nature, 540(7632), pp.270–275. https://doi.org/10.1038/nature20592.

Karchin, R., Diekhans, M., Kelly, L., Thomas, D.J., Pieper, U., Eswar, N., Haussler, D. and Sali, A., 2005. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics, 21(12), pp.2814–2820. https://doi.org/10.1093/bioinformatics/bti442.

Kazuno, A., Munakata, K., Nagai, T., Shimozono, S., Tanaka, M., Yoneda, M., Kato, N., Miyawaki, A. and Kato, T., 2006. Identification of Mitochondrial DNA

Polymorphisms That Alter Mitochondrial Matrix pH and Intracellular Calcium Dynamics. PLoS Genetics, 2(8), p.e128. https://doi.org/10.1371/journal.pgen.0020128.

Keele, G.R., Quach, B.C., Israel, J.W., Chappell, G.A., Lewis, L., Safi, A., Simon, J.M., Cotney, P., Crawford, G.E., Valdar, W., Rusyn, I. and Furey, T.S., 2020. Integrative QTL analysis of gene expression and chromatin accessibility identifies multi-tissue patterns of genetic regulation. PLOS Genetics, 16(1), p.e1008537. https://doi.org/10.1371/journal.pgen.1008537.

Kenney, M.C., Chwa, M., Atilano, S.R., Falatoonzadeh, P., Ramirez, C., Malik, D., Tarek, M., del Carpio, J.C., Nesburn, A.B., Boyer, D.S., Kuppermann, B.D., Vawter, M.P., Jazwinski, S.M., Miceli, M. V., Wallace, D.C. and Udar, N., 2014. Molecular and bioenergetic differences between cells with African versus European inherited mitochondrial DNA haplogroups: Implications for population susceptibility to diseases. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 1842(2), pp.208–219. https://doi.org/10.1016/j.bbadis.2013.10.016.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, and D., 2002. The Human Genome Browser at UCSC. Genome Research, 12(6), pp.996–1006. https://doi.org/10.1101/gr.229102.

Khan, M.A.B., Hashim, M.J., King, J.K., Govender, R.D., Mustafa, H. and Al Kaabi, J., 2019. Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends. Journal of Epidemiology and Global Health, 10(1), p.107. https://doi.org/10.2991/jegh.k.191028.001.

Khawaja, A., Itoh, Y., Remes, C., Spåhr, H., Yukhnovets, O., Höfig, H., Amunts, A. and Rorbach, J., 2020. Distinct pre-initiation steps in human mitochondrial translation. Nature Communications, 11(1), p.2932. https://doi.org/10.1038/s41467-020-16503-2.

Khrapko, K. and Turnbull, D., 2014. Mitochondrial DNA Mutations in Aging. pp.29–62. https://doi.org/10.1016/B978-0-12-394625-6.00002-7.

King, M.P. and Attardi, G., 1989a. Human Cells Lacking mtDNA: Repopulation with Exogenous Mitochondria by Complementation. Science, 246(4929), pp.500–503. https://doi.org/10.1126/science.2814477.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.-Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., Bracken, M.B., Ferris, F.L., Ott, J., Barnstable, C. and Hoh, J., 2005a. Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science, 308(5720), pp.385–389. https://doi.org/10.1126/science.1109557.

Kmietowicz, Z., 2015a. UK becomes first country to allow mitochondrial donation. BMJ, 350(feb25 14), pp.h1103–h1103. https://doi.org/10.1136/bmj.h1103.

Kotrys, A. V., Durham, T.J., Guo, X.A., Vantaku, V.R., Parangi, S. and Mootha, V.K., 2024. Single-cell analysis reveals context-dependent, cell-level selection of mtDNA. Nature, 629(8011), pp.458–466. https://doi.org/10.1038/s41586-024-07332-0.

Kuhn, R.M., Haussler, D. and Kent, W.J., 2013. The UCSC genome browser and associated tools. Briefings in Bioinformatics, 14(2), pp.144–161. https://doi.org/10.1093/bib/bbs038.

Kumle, L., Võ, M.L.-H. and Draschkow, D., 2021. Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. Behavior Research Methods, 53(6), pp.2528–2543. https://doi.org/10.3758/s13428-021-01546-0.

Lakhani, C.M., Tierney, B.T., Manrai, A.K., Yang, J., Visscher, P.M. and Patel, C.J., 2019. Repurposing large health insurance claims data to estimate genetic and environmental contributions in 560 phenotypes. Nature Genetics, 51(2), pp.327–334. https://doi.org/10.1038/s41588-018-0313-7.

Lane, N., 2017. Serial endosymbiosis or singular event at the origin of eukaryotes? Journal of Theoretical Biology, 434, pp.58–67. https://doi.org/10.1016/j.jtbi.2017.04.031.

Laricchia, K.M., Lake, N.J., Watts, N.A., Shand, M., Haessly, A., Gauthier, L., Benjamin, D., Banks, E., Soto, J., Garimella, K., Emery, J., Rehm, H.L., MacArthur, D.G., Tiao, G.,

Lek, M., Mootha, V.K. and Calvo, S.E., 2022. Mitochondrial DNA variation across 56,434 individuals in gnomAD. Genome Research, 32(3), pp.569–582. https://doi.org/10.1101/gr.276013.121.

Lawless, C., Greaves, L., Reeve, A.K., Turnbull, D.M. and Vincent, A.E., 2020. The rise and rise of mitochondrial DNA mutations. Open Biology, 10(5). https://doi.org/10.1098/rsob.200061.

Lazcano, A. and Peretó, J., 2017. On the origin of mitosing cells: A historical appraisal of Lynn Margulis endosymbiotic theory. Journal of Theoretical Biology, 434, pp.80–87. https://doi.org/10.1016/j.jtbi.2017.06.036.

Lean, I.J., Rabiee, A.R., Duffield, T.F. and Dohoo, I.R., 2009. Invited review: Use of meta-analysis in animal health and reproduction: Methods and applications. Journal of Dairy Science, 92(8), pp.3545–3565. https://doi.org/10.3168/jds.2009-2140.

Lee, S., Abecasis, G.R., Boehnke, M. and Lin, X., 2014a. Rare-Variant Association Analysis: Study Designs and Statistical Tests. The American Journal of Human Genetics, 95(1), pp.5–23. https://doi.org/10.1016/j.ajhg.2014.06.009.

Lee, S.R. and Han, J., 2017. Mitochondrial Nucleoid: Shield and Switch of the Mitochondrial Genome. Oxidative Medicine and Cellular Longevity, 2017, pp.1–15. https://doi.org/10.1155/2017/8060949.

LeLorier, J., Grégoire, G., Benhaddad, A., Lapierre, J. and Derderian, F., 1997. Discrepancies between Meta-Analyses and Subsequent Large Randomized, Controlled Trials. New England Journal of Medicine, 337(8), pp.536–542. https://doi.org/10.1056/NEJM199708213370806.

Lettre, G., Lange, C. and Hirschhorn, J.N., 2007. Genetic model testing and statistical power in population-based association studies of quantitative traits. Genetic Epidemiology, 31(4), pp.358–362. https://doi.org/10.1002/gepi.20217.

Lever, J., Krzywinski, M. and Altman, N., 2017. Points of Significance: Principal component analysis. Nature Methods, https://doi.org/10.1038/nmeth.4346.

Lewontin, R.C. and Kojima, K., 1960. The Evolutionary Dynamics of Complex Polymorphisms. Evolution, 14(4), p.458. https://doi.org/10.2307/2405995.

Li, D., Liang, C., Zhang, T., Marley, J.L., Zou, W., Lian, M. and Ji, D., 2022. Pathogenic mitochondrial DNA 3243A>G mutation: From genetics to phenotype. Frontiers in Genetics, 13. https://doi.org/10.3389/fgene.2022.951185.

Li, H., Slone, J., Fei, L. and Huang, T., 2019. cells Mitochondrial DNA Variants and Common Diseases: A Mathematical Model for the Diversity of Age-Related mtDNA Mutations. [online] https://doi.org/10.3390/cells8060608.

Li, H., Uittenbogaard, M., Hao, L. and Chiaramello, A., 2021. Clinical Insights into Mitochondrial Neurodevelopmental and Neurodegenerative Disorders: Their Biosignatures from Mass Spectrometry-Based Metabolomics. Metabolites, 11(4), p.233. https://doi.org/10.3390/metabo11040233.

Li, S., Carss, K.J., Halldorsson, B. V, Cortes, A. and Consortium, U.K.B.W.-G.S., 2023. Whole-genome sequencing of half-a-million UK Biobank participants. medRxiv. [online] https://doi.org/10.1101/2023.12.06.23299426.

Li, Y.I., van de Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y. and Pritchard, J.K., 2016. RNA splicing is a primary link between genetic variation and disease. Science, 352(6285), pp.600–604. https://doi.org/10.1126/science.aad9417.

Lightowlers, R.N., Taylor, R.W. and Turnbull, D.M., 2015. Mutations causing mitochondrial disease: What is new and what challenges remain? Science, 349(6255), pp.1494–1499. https://doi.org/10.1126/science.aac7516.

Lim, S.E., Longley, M.J. and Copeland, W.C., 1999. The Mitochondrial p55 Accessory Subunit of Human DNA Polymerase γ Enhances DNA Binding, Promotes Processive DNA Synthesis, and Confers N-Ethylmaleimide Resistance. Journal of Biological Chemistry, 274(53), pp.38197–38203. https://doi.org/10.1074/jbc.274.53.38197.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I. and Heckerman, D., 2011a. FaST linear mixed models for genome-wide association studies. Nature Methods, 8(10), pp.833–835. https://doi.org/10.1038/nmeth.1681.

Listgarten, J., Lippert, C., Kadie, C.M., Davidson, R.I., Eskin, E. and Heckerman, D., 2012. Improved linear mixed models for genome-wide association studies. Nature Methods, 9(6), pp.525–526. https://doi.org/10.1038/nmeth.2037.

Liu, H., Zhen, C., Xie, J., Luo, Z., Zeng, L., Zhao, G., Lu, S., Zhuang, H., Fan, H., Li, X., Liu, Z., Lin, S., Jiang, H., Chen, Y., Cheng, J., Cao, Z., Dai, K., Shi, J., Wang, Z., Hu, Y., Meng, T., Zhou, C., Han, Z., Huang, H., Zhou, Q., He, P. and Feng, D., 2024. TFAM is an autophagy receptor that limits inflammation by binding to cytoplasmic mitochondrial DNA. Nature Cell Biology. https://doi.org/10.1038/s41556-024-01419-6.

Liu, J.Z., McRae, A.F., Nyholt, D.R., Medland, S.E., Wray, N.R., Brown, K.M., Hayward, N.K., Montgomery, G.W., Visscher, P.M., Martin, N.G., Macgregor, S., Mann, G.J., Kefford, R.F., Hopper, J.L., Aitken, J.F., Giles, G.G. and Armstrong, B.K., 2010. A versatile gene-based test for genome-wide association studies. American Journal of Human Genetics, 87(1), pp.139–145. https://doi.org/10.1016/j.ajhg.2010.06.009.

Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., Patterson, N. and Price, A.L., 2015. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nature Genetics, 47(3), pp.284–290. https://doi.org/10.1038/ng.3190.

Longino H.E., 2013. Studying Human Behavior: How Scientists Investigate Aggression and Sexuality. University of Chicago Press.

Lopez Sanchez, M.I.G., Mercer, T.R., Davies, S.M.K., Shearwood, A.-M.J., Nygård, K.K.A., Richman, T.R., Mattick, J.S., Rackham, O. and Filipovska, A., 2011. RNA processing in human mitochondria. Cell Cycle, 10(17), pp.2904–2916. https://doi.org/10.4161/cc.10.17.17060.

Lu, H.-F., Chou, C.-H., Lin, Y.-J., Uchiyama, S., Terao, C., Wang, Y.-W., Yang, J.-S., Liu, T.-Y., Wong, H.S.-C., Chen, S.C.-C. and Tsai, F.-J., 2024. The enome-wide ssociation tudy of erum IgE evels emonstrated a hared enetic ackground in llergic iseases. Clinical Immunology, 260, p.109897. https://doi.org/10.1016/j.clim.2024.109897.

Lücking, C.B., Dürr, A., Bonifati, V., Vaughan, J., De Michele, G., Gasser, T., Harhangi, B.S., Meco, G., Denèfle, P., Wood, N.W., Agid, Y., Brice, A., Nicholl, D., Breteler, M.M.B., Oostra, B.A., Marconi, R., Filla, A., Bonnet, A.M., Broussolle, E., Pollak, M., Rascol, O., and Arnould, I. 2000. Association between early-onset Parkinson's disease and mutations in the parkin gene. New England Journal of Medicine, 342(21), pp.1560-1567

Ma, H. and O'Farrell, P.H., 2016. Selfish drive can trump function when animal mitochondrial genomes compete. Nature Genetics, 48(7), pp.798–802. https://doi.org/10.1038/ng.3587.

Maeda, K., Kawai, H., Sanada, M., Terashima, T., Ogawa, N., Idehara, R., Makiishi, T., Yasuda, H., Sato, S., Hoshi, K., Yahikozawa, H., Nishi, K., Itoh, Y., Ogasawara, K., Tomita, K., Indo, H.P. and Majima, H.J., 2016. Clinical Phenotype and Segregation of Mitochondrial 3243A>G Mutation in 2 Pairs of Monozygotic Twins. JAMA Neurology, 73(8), p.990. https://doi.org/10.1001/jamaneurol.2016.0886.

Mägi, R. and Morris, A.P., 2010a. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics, 11(1), p.288. https://doi.org/10.1186/1471-2105-11-288.

Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J., Steinthorsdottir, V., Scott, R.A., Grarup, N., et al., 2018. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nature Genetics, 50(11), pp.1505–1513. https://doi.org/10.1038/s41588-018-0241-6.

Maher, B., 2008. Personal genomes: The case of the missing heritability. Nature, 456(7218), pp.18–21. https://doi.org/10.1038/456018a.

Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J.M.M., Auton, A., Myers, S., Morris, A., Pirinen, M., Brown, M.A., Burton, P.R., Caulfield, M.J., Compston, A., Farrall, M., Hall, A.S., Hattersley, A.T., Hill, A.V.S., Mathew, C.G., Pembrey, M., Satsangi, J., Stratton, M.R., Worthington, J., Craddock, N., Hurles, M., Ouwehand, W., Parkes, M., Rahman, N., Duncanson, A., Todd, J.A., Kwiatkowski, D.P., Samani, N.J., Gough, S.C.L., McCarthy, M.I., Deloukas, P. and Donnelly, P., 2012. Bayesian refinement of association signals for 14 loci in 3 common diseases. Nature Genetics, 44(12), pp.1294–1301. https://doi.org/10.1038/ng.2435.

Manoli, I., Alesci, S. and Chrousos, G.P., 2007. Mitochondria. Encyclopedia of Stress, pp.754–761. https://doi.org/10.1016/B978-012373947-6.00559-6.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F.C., McCarroll, S.A. and Visscher, P.M., 2009b. Finding the missing heritability of complex diseases. Nature, 461(7265), pp.747–753. https://doi.org/10.1038/nature08494.

Manwaring, N., Jones, M.M., Wang, J.J., Rochtchina, E., Howard, C., Mitchell, P. and Sue, C.M., 2007a. Population prevalence of the MELAS A3243G mutation. Mitochondrion, 7(3), pp.230–233. https://doi.org/10.1016/j.mito.2006.12.004.

March, R.E., 1999. Gene mapping by linkage and association analysis. Molecular Biotechnology, [online] 13(2), pp.113–122. https://doi.org/10.1385/MB:13:2:113.

Marees, A.T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C. and Derks, E.M., 2018. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. International Journal of Methods in Psychiatric Research, 27(2). https://doi.org/10.1002/mpr.1608.

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Coffey, C., Kieburtz, K., Flagg, E., Chowdhury, S., Poewe, W., et al., 2011. The Parkinson Progression Marker Initiative (PPMI). Progress in Neurobiology, 95(4), pp.629–635. https://doi.org/10.1016/j.pneurobio.2011.09.005.

Marsoni, S., Torri, W., Taiana, A., Gambino, A., Grilli, R., Liati, P., Franzosi, M.G., Pistotti, V., Parazzini, F., Focarile, F. and Liberati, A., 1990. Critical review of the quality and development of randomized clinical trials (RCTs) and their influence on the treatment of advanced epithelial ovarian cancer. Annals of Oncology, 1(5), pp.343–350. https://doi.org/10.1093/oxfordjournals.annonc.a057772.

Martin, W. and Müller, M., 1998. The hydrogen hypothesis for the first eukaryote. Nature, 392(6671), pp.37–41. https://doi.org/10.1038/32096.

Martínez-Reyes, I. and Chandel, N.S., 2020. Mitochondrial TCA cycle metabolites control physiology and disease. Nature Communications, 11(1), p.102. https://doi.org/10.1038/s41467-019-13668-3.

Matthews, L.J. and Turkheimer, E., 2022. Three legs of the missing heritability problem. Studies in History and Philosophy of Science, 93, pp.183–191. https://doi.org/10.1016/j.shpsa.2022.04.004.

Mayhew, A.J. and Meyre, D., 2017. Assessing the Heritability of Complex Traits in Humans: Methodological Challenges and Opportunities. Current Genomics, 18(4). https://doi.org/10.2174/1389202918666170307161450.

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J.A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., Habegger, L., Ferreira, M., Baras, A., Reid, J., Abecasis, G., Maxwell, E. and Marchini, J., 2021a. Computationally efficient whole-genome regression for quantitative and binary traits. Nature Genetics, 53(7), pp.1097–1103. https://doi.org/10.1038/s41588-021-00870-7.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A. and Hirschhorn, J.N., 2008. Genome-wide association studies for complex traits:

consensus, uncertainty and challenges. Nature Reviews Genetics, 9(5), pp.356–369. https://doi.org/10.1038/nrg2344.

McFarland, R., Clark, K.M., Morris, A.A.M., Taylor, R.W., Macphail, S., Lightowlers, R.N. and Turnbull, D.M., 2002. Multiple neonatal deaths due to a homoplasmic mitochondrial DNA mutation. Nature Genetics, 30(2), pp.145–146. https://doi.org/10.1038/ng819.

McGurk, K., Keavney, B. and Nicolaou, A., 2019. 104 Heritability and family-based GWAS analyses of the circulating ceramide, endocannabinoid, and N-acyl ethanolamide lipidome. In: Stable IHD/Prevention/Hypertension/Lipids. BMJ Publishing Group Ltd and British Cardiovascular Society. p.A86.1-A86. https://doi.org/10.1136/heartjnl-2019-BCS.101.

McShane, M., Hammans, S., Sweeney, M., Holt, I., Beattie, T., Brett, E. and Harding, A., 1991. Pearson syndrome and mitochondrial encephalomyopathy in a patient with a deletion of mtDNA. Am J Hum Genet, 48(1), pp.39–42.

Merriwether, D.A., Clark, A.G., Ballinger, S.W., Schurr, T.G., Soodyall, H., Jenkins, T., Sherry, S.T. and Wallace, D.C., 1991. The structure of human mitochondrial DNA variation. Journal of Molecular Evolution, 33(6), pp.543–555. https://doi.org/10.1007/BF02102807.

Meyers, D., Beaty, T., Freidhoff, L. and Marsh, D., 1987. Inheritance of total serum IgE (basal levels) in man. Am J Hum Genet., 41(1), pp.51–62.

Milenkovic, D., Matic, S., Kuhl, I., Ruzzenente, B., Freyer, C., Jemt, E., Park, C.B., Falkenberg, M. and Larsson, N.-G., 2013. TWINKLE is an essential mitochondrial helicase required for synthesis of nascent D-loop strands and complete mtDNA replication. Human Molecular Genetics, 22(10), pp.1983–1993. https://doi.org/10.1093/hmg/ddt051.

Miller, B., Arpawong, T.E., Jiao, H., Kim, S.J., Yen, K., Mehta, H.M., Wan, J., Carpten, J.C. and Cohen, P., 2019. Comparing the utility of mitochondrial and nuclear DNA to

adjust for genetic ancestry in association studies. Cells, 8(4). https://doi.org/10.3390/cells8040306.

Mitchell, P., 1961. Coupling of Phosphorylation to Electron and Hydrogen Transfer by a Chemi-Osmotic type of Mechanism. Nature, 191(4784), pp.144–148. https://doi.org/10.1038/191144a0.

Mizrahi-Man, O., Woehrmann, M.H., Webster, T.A., Gollub, J., Bivol, A., Keeble, S.M., Aull, K.H., Mittal, A., Roter, A.H., Wong, B.A. and Schmidt, J.P., 2022. Novel genotyping algorithms for rare variants significantly improve the accuracy of Applied BiosystemsTM AxiomTM array genotyping calls: Retrospective evaluation of UK Biobank array data. PLOS ONE, 17(11), p.e0277680. https://doi.org/10.1371/journal.pone.0277680.

MN, W., L, J., JW, H., KS, R., J, T., AT, H. and CF, W., 2021. Use of SNP chips to detect rare pathogenic variants: retrospective, population based diagnostic evaluation. BMJ, p.n214. https://doi.org/10.1136/bmj.n214.

Mok, B.Y., de Moraes, M.H., Zeng, J., Bosch, D.E., Kotrys, A. V., Raguram, A., Hsu, F., Radey, M.C., Peterson, S.B., Mootha, V.K., Mougous, J.D. and Liu, D.R., 2020. A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. Nature, 583(7817), pp.631–637. https://doi.org/10.1038/s41586-020-2477-4.

Momozawa, Y. and Mizukami, K., 2021. Unique roles of rare variants in the genetics of complex diseases in humans. Journal of Human Genetics, 66(1), pp.11–23. https://doi.org/10.1038/s10038-020-00845-2.

Moore, C.M., Jacobson, S.A. and Fingerlin, T.E., 2019. Power and Sample Size Calculations for Genetic Association Studies in the Presence of Genetic Model Misspecification. Human Heredity, 84(6), pp.256–271. https://doi.org/10.1159/000508558.

Morris, A.P., Lindgren, C.M., Zeggini, E., Timpson, N.J., Frayling, T.M., Hattersley, A.T. and McCarthy, M.I., 2010. A powerful approach to sub-phenotype analysis in population-

based genetic association studies. Genetic Epidemiology, 34(4), pp.335–343. https://doi.org/10.1002/gepi.20486.

Mozdy, A.D., McCaffery, J.M. and Shaw, J.M., 2000. Dnm1p Gtpase-Mediated Mitochondrial Fission Is a Multi-Step Process Requiring the Novel Integral Membrane Component Fis1p. The Journal of Cell Biology, 151(2), pp.367–380. https://doi.org/10.1083/jcb.151.2.367.

Mühlenhoff, U., Braymer, J.J., Christ, S., Rietzschel, N., Uzarska, M.A., Weiler, B.D. and Lill, R., 2020. Glutaredoxins and iron-sulfur protein biogenesis at the interface of redox biology and iron metabolism. Biological Chemistry, 401(12), pp.1407–1428. https://doi.org/10.1515/hsz-2020-0237.

Mulder, R.H., Neumann, A., Cecil, C.A.M., Walton, E., Houtepen, L.C., Simpkin, A.J., Rijlaarsdam, J., Heijmans, B.T., Gaunt, T.R., Felix, J.F., Jaddoe, V.W. V, Bakermans-Kranenburg, M.J., Tiemeier, H., Relton, C.L., van IJzendoorn, M.H. and Suderman, M., 2021. Epigenome-wide change and variation in DNA methylation in childhood: trajectories from birth to late adolescence. Human Molecular Genetics, 30(1), pp.119–134. https://doi.org/10.1093/hmg/ddaa280.

Muller, S. and Radic, M., 2016. Oxidation and mitochondrial origin of NET DNA in the pathogenesis of lupus. Nature Medicine, 22(2), pp.126–127. https://doi.org/10.1038/nm.4044.

Murphy, A.E., Schilder, B.M. and Skene, N.G., 2021. MungeSumstats: a Bioconductor package for the standardization and quality control of many GWAS summary statistics. Bioinformatics, 37(23), pp.4593–4596. https://doi.org/10.1093/bioinformatics/btab665.

Nagata, S., 2018. Apoptosis and Clearance of Apoptotic Cells. Annual Review of Immunology, 36(1), pp.489–517. https://doi.org/10.1146/annurev-immunol-042617-053010.

Nandakumar, P., Tian, C., O'connell, J., Team, R., Hinds, D., Paterson, A.D. and Sondheimer, N., 2021. Nuclear genome-wide associations with mitochondrial heteroplasmy. [online] Sci. Adv, Available at: <http://advances.sciencemag.org/>.

Nesbitt, V. and McFarland, R., 2011. Phenotypic spectrum of m.3243A>G mitochondrial DNA mutation in children. Archives of Disease in Childhood, [online] 96(Suppl 1), pp.A28–A28. https://doi.org/10.1136/adc.2011.212563.57.

Nesbitt, V., Pitceathly, R.D.S., Turnbull, D.M., Taylor, R.W., Sweeney, M.G., Mudanohwo, E.E., Rahman, S., Hanna, M.G. and McFarland, R., 2013. The UK MRC Mitochondrial Disease Patient Cohort Study: clinical phenotypes associated with the m.3243A>G mutation--implications for diagnosis and management. Journal of Neurology, Neurosurgery & Psychiatry, 84(8), pp.936–938. https://doi.org/10.1136/jnnp-2012-303528.

Nettle D, 2019. Modelling and visualizing data using R:A practical introduction .

Nolfi-Donegan, D., Braganza, A. and Shiva, S., 2020. Mitochondrial electron transport chain: Oxidative phosphorylation, oxidant production, and methods of measurement. Redox Biology, 37, p.101674. https://doi.org/10.1016/j.redox.2020.101674.

Nyholt, D.R., 2000. All LODs Are Not Created Equal**A Microsoft Excel spreadsheet, for performing easy calculations of P values for the LOD scores described in this review, is available on request from the author. The American Journal of Human Genetics, 67(2), pp.282–288. https://doi.org/10.1086/303029.

Obeng, E., 2021. Apoptosis (programmed cell death) and its signals - A review. Brazilian Journal of Biology, 81(4), pp.1133–1143. https://doi.org/10.1590/1519-6984.228437.

Oborník, M., 2019. In the beginning was the word: How terminology drives our understanding of endosymbiotic organelles. Microbial Cell, 6(2), pp.134–141. https://doi.org/10.15698/mic2019.02.669.

Ojala, D., Montoya, J. and Attardi, G., 1981. tRNA punctuation model of RNA processing in human mitochondria. Nature, 290(5806), pp.470–474. https://doi.org/10.1038/290470a0.

Olivo, P.D., Van de Walle, M.J., Laipis, P.J. and Hauswirth, W.W., 1983. Nucleotide sequence evidence for rapid genotypic shifts in the bovine mitochondrial DNA D-loop. Nature, 306(5941), pp.400–402. https://doi.org/10.1038/306400a0.

Ott, J., Wang, J. and Leal, S.M., 2015a. Genetic linkage analysis in the age of whole-genome sequencing. Nature Reviews Genetics, 16(5), pp.275–284. https://doi.org/10.1038/nrg3908.

Otten, A.B.C., Sallevelt, S.C.E.H., Carling, P.J., Dreesen, J.C.F.M., Drüsedau, M., Spierts, S., Paulussen, A.D.C., de Die-Smulders, C.E.M., Herbert, M., Chinnery, P.F., Samuels, D.C., Lindsey, P. and Smeets, H.J.M., 2018. Mutation-specific effects in germline transmission of pathogenic mtDNA variants. Human Reproduction, 33(7), pp.1331–1341. https://doi.org/10.1093/humrep/dey114.

van den Ouweland, J.M.W., Lemkes, H.H.P.J., Ruitenbeek, W., Sandkuijl, L.A., de Vijlder, M.F., Struyvenberg, P.A.A., van de Kamp, J.J.P. and Maassen, J.A., 1992. Mutation in mitochondrial tRNALeu(UUR) gene in a large pedigree with maternally transmitted type II diabetes mellitus and deafness. Nature Genetics, 1(5), pp.368–371. https://doi.org/10.1038/ng0892-368.

van Oven, M. and Kayser, M., 2009a. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Human Mutation, 30(2), pp.E386–E394. https://doi.org/10.1002/humu.20921.

Palanichamy, M. gounder, Sun, C., Agrawal, S., Bandelt, H.-J., Kong, Q.-P., Khan, F., Wang, C.-Y., Chaudhuri, T.K., Palla, V. and Zhang, Y.-P., 2004. Phylogeny of Mitochondrial DNA Macrohaplogroup N in India, Based on Complete Sequencing: Implications for the Peopling of South Asia. The American Journal of Human Genetics, 75(6), pp.966–978. https://doi.org/10.1086/425871.

Pallotti, F., Binelli, G., Fabbri, R., Valentino, M.L., Vicenti, R., Macciocca, M., Cevoli, S., Baruzzi, A., DiMauro, S. and Carelli, V., 2014. A Wide Range of 3243A>G/tRNALeu(UUR) (MELAS) Mutation Loads May Segregate in Offspring through the Female Germline Bottleneck. PLoS ONE, 9(5), p.e96663. https://doi.org/10.1371/journal.pone.0096663.

Park, J.H., 2003. Evolved Disease-Avoidance Processes and Contemporary Anti-Social Behavior. Journal of Nonverbal Behavior, 27(2), pp.65–87. https://doi.org/10.1023/A:1023910408854.

Patterson, N., Price, A.L. and Reich, D., 2006. Population Structure and Eigenanalysis. PLoS Genetics, 2(12), p.e190. https://doi.org/10.1371/journal.pgen.0020190.

Penrose, L.S., 1952. "The general purpose sib-pair linkage test. Annals of eugenics, 17(1), pp.120–124.

Petronek, M.S., Spitz, D.R. and Allen, B.G., 2021. Iron–Sulfur Cluster Biogenesis as a Critical Target in Cancer. Antioxidants, 10(9), p.1458. https://doi.org/10.3390/antiox10091458.

Pezet, M.G., Gomez-Duran, A., Klimm, F., Aryaman, J., Burr, S., Wei, W., Saitou, M., Prudent, J. and Chinnery, P.F., 2021. Oxygen tension modulates the mitochondrial genetic bottleneck and influences the segregation of a heteroplasmic mtDNA variant in vitro. Communications Biology, 4(1), p.584. https://doi.org/10.1038/s42003-021-02069-2.

Pickett, S.J., Blain, A., Ng, Y.S., Wilson, I.J., Taylor, R.W., McFarland, R., Turnbull, D.M. and Gorman, G.S., 2019a. Mitochondrial Donation — Which Women Could Benefit? New England Journal of Medicine, 380(20), pp.1971–1972. https://doi.org/10.1056/nejmc1808565.

Pickett, S.J., Grady, J.P., Ng, Y.S., Gorman, G.S., Schaefer, A.M., Wilson, I.J., Cordell, H.J., Turnbull, D.M., Taylor, R.W. and McFarland, R., 2018. Phenotypic heterogeneity in

m.3243A>G mitochondrial disease: The role of nuclear factors. Annals of Clinical and Translational Neurology, 5(3), pp.333–345. https://doi.org/10.1002/acn3.532.

Pierron, D., Rocher, C., Amati-Bonneau, P., Reynier, P., Martin-Négrier, M.L., Allouche, S., Batandier, C., de Camaret, B., Godinot, C., Rotig, A., Feldmann, D., Bellanne-Chantelot, C., Arveiler, B., Pennarun, E., Rossignol, R., Crouzet, M., Murail, P., Thoraval, D. and Letellier, T., 2008. New evidence of a mitochondrial genetic background paradox: Impact of the J haplogroup on the A3243G mutation. BMC Medical Genetics, 9. https://doi.org/10.1186/1471-2350-9-41.

Politi, C., Roumeliotis, S., Tripepi, G. and Spoto, B., 2023. Sample Size Calculation in Genetic Association Studies: A Practical Approach. Life, 13(1), p.235. https://doi.org/10.3390/life13010235.

Posit team, 2023. RStudio: Integrated Development Environment for R. Posit Software.

Poulton, J., 2002. Type 2 diabetes is associated with a common mitochondrial variant: evidence from a population-based case-control study. Human Molecular Genetics, 11(13), pp.1581–1583. https://doi.org/10.1093/hmg/11.13.1581.

Powledge, T.M., 2003. Human genome project completed. Genome Biology, 4(1). https://doi.org/10.1186/gb-spotlight-20030415-01.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics, 38(8), pp.904–909. https://doi.org/10.1038/ng1847.

Price, A.L., Zaitlen, N.A., Reich, D. and Patterson, N., 2010a. New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics, 11(7), pp.459–463. https://doi.org/10.1038/nrg2813.

Protasoni, M. and Zeviani, M., 2021. Mitochondrial Structure and Bioenergetics in Normal and Disease Conditions. International Journal of Molecular Sciences, 22(2), p.586. https://doi.org/10.3390/ijms22020586.

Pulst, S.M., 1999. Genetic Linkage Analysis. Archives of Neurology, 56(6), p.667. https://doi.org/10.1001/archneur.56.6.667.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J. and Sham, P.C., 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics, 81(3), pp.559–575. https://doi.org/10.1086/519795.

Rahman, S., Poulton, J., Marchington, D. and Suomalainen, A., 2001. Decrease of 3243 A→G mtDNA Mutation from Blood in MELAS Syndrome: A Longitudinal Study. The American Journal of Human Genetics, 68(1), pp.238–240. https://doi.org/10.1086/316930.

Rajasimha, H.K., Chinnery, P.F. and Samuels, D.C., 2008. Selection against Pathogenic mtDNA Mutations in a Stem Cell Population Leads to the Loss of the 3243A→G Mutation in Blood. The American Journal of Human Genetics, 82(2), pp.333–343. https://doi.org/10.1016/j.ajhg.2007.10.007.

Ramchandani, D., Berisa, M., Tavarez, D.A., Li, Z., Miele, M., Bai, Y., Lee, S.B., Ban, Y., Dephoure, N., Hendrickson, R.C., Cloonan, S.M., Gao, D., Cross, J.R., Vahdat, L.T. and Mittal, V., 2021. Copper depletion modulates mitochondrial oxidative phosphorylation to impair triple negative breast cancer metastasis. Nature Communications, 12(1), p.7311. https://doi.org/10.1038/s41467-021-27559-z.

Rath, S., Sharma, R., Gupta, R., Ast, T., Chan, C., Durham, T.J., Goodman, R.P., Grabarek, Z., Haas, M.E., Hung, W.H.W., Joshi, P.R., Jourdain, A.A., Kim, S.H., Kotrys, A. V., Lam, S.S., McCoy, J.G., Meisel, J.D., Miranda, M., Panda, A., Patgiri, A., Rogers, R., Sadre, S., Shah, H., Skinner, O.S., To, T.L., Walker, M.A., Wang, H., Ward, P.S., Wengrod, J., Yuan, C.C., Calvo, S.E. and Mootha, V.K., 2021. MitoCarta3.0: An updated mitochondrial proteome now with sub-organelle localization and pathway annotations. Nucleic Acids Research, 49(D1), pp.D1541–D1547. https://doi.org/10.1093/nar/gkaa1011.

Rebelo, A.P., Dillon, L.M. and Moraes, C.T., 2011. Mitochondrial DNA transcription regulation and nucleoid organization. Journal of Inherited Metabolic Disease, 34(4), pp.941–951. https://doi.org/10.1007/s10545-011-9330-8.

Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M.P. and Foulkes, A.S., 2015. A guide to genome-wide association analysis and post-analytic interrogation. Statistics in Medicine, 34(28), pp.3769–3792. https://doi.org/10.1002/sim.6605.

Reynisdottir, I., Thorleifsson, G., Benediktsson, R., Sigurdsson, G., Emilsson, V., Einarsdottir, A.S., Hjorleifsdottir, E.E., Orlygsdottir, G.Th., Bjornsdottir, G.T., Saemundsdottir, J., Halldorsson, S., Hrafnkelsdottir, S., Sigurjonsdottir, S.B., Steinsdottir, S., Martin, M., Kochan, J.P., Rhees, B.K., Grant, S.F.A., Frigge, M.L., Kong, A., Gudnason, V., Stefansson, K. and Gulcher, J.R., 2003. Localization of a Susceptibility Gene for Type 2 Diabetes to Chromosome 5q34–q35.2. The American Journal of Human Genetics, 73(2), pp.323–335. https://doi.org/10.1086/377139.

Richards, M., Macaulay, V., Torroni, A. and Bandelt, H.-J., 2002. In Search of Geographical Patterns in European Mitochondrial DNA. The American Journal of Human Genetics, 71(5), pp.1168–1174. https://doi.org/10.1086/342930.

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S.R.F., Wilkie, A.O.M., McVean, G. and Lunter, G., 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature Genetics, 46(8), pp.912–918. https://doi.org/10.1038/ng.3036.

Risch, N. and Merikangas, K., 1996. The future of genetic studies of complex human diseases. Science, https://doi.org/10.1126/science.273.5281.1516.

Robinow, C. and Kellenberger, E., 1994. The bacterial nucleoid revisited. Microbiological Reviews, 58(2), pp.211–232. https://doi.org/10.1128/mr.58.2.211-232.1994.

ROSSIGNOL, R., FAUSTIN, B., ROCHER, C., MALGAT, M., MAZAT, J.-P. and LETELLIER, T., 2003. Mitochondrial threshold effects. Biochemical Journal, 370(3), pp.751–762. https://doi.org/10.1042/bj20021594.

Rossignol, R., Malgat, M., Mazat, J.-P. and Letellier, T., 1999. Threshold Effect and Tissue Specificity. Journal of Biological Chemistry, 274(47), pp.33426–33432. https://doi.org/10.1074/jbc.274.47.33426.

Ruiz-Pesini, E. and Wallace, D.C., 2006. Evidence for adaptive selection acting on the tRNA and rRNA genes of human mitochondrial DNA. Human Mutation, 27(11), pp.1072–1081. https://doi.org/10.1002/humu.20378.

Al Safar, H.S., Cordell, H.J., Jafer, O., Anderson, D., Jamieson, S.E., Fakiola, M., Khazanehdari, K., Tay, G.K. and Blackwell, J.M., 2013. A Genome-Wide Search for Type 2 Diabetes Susceptibility Genes in an Extended Arab Family. Annals of Human Genetics, 77(6), pp.488–503. https://doi.org/10.1111/ahg.12036.

Sagan, L., 1967. On the origin of mitosing cells. Journal of Theoretical Biology, 14(3), pp.225-IN6. https://doi.org/10.1016/0022-5193(67)90079-3.

Salomon, M.P., Li, W.L.S., Edlund, C.K., Morrison, J., Fortini, B.K., Win, A.K., Conti, D. V., Thomas, D.C., Duggan, D., Buchanan, D.D., Jenkins, M.A., Hopper, J.L., Gallinger, S., Le Marchand, L., Newcomb, P.A., Casey, G. and Marjoram, P., 2016. GWASeq: targeted re-sequencing follow up to GWAS. BMC Genomics, 17(1), p.176. https://doi.org/10.1186/s12864-016-2459-y.

Samuels, D.C., Carothers, A.D., Horton, R. and Chinnery, P.F., 2006. The Power to Detect Disease Associations with Mitochondrial DNA Haplogroups. The American Journal of Human Genetics, 78(4), pp.713–720. https://doi.org/10.1086/502682.

Samuels, D.C., Li, C., Li, B., Song, Z., Torstenson, E., Boyd Clay, H., Rokas, A., Thornton-Wells, T.A., Moore, J.H., Hughes, T.M., Hoffman, R.D., Haines, J.L., Murdock, D.G., Mortlock, D.P. and Williams, S.M., 2013. Recurrent Tissue-Specific mtDNA Mutations Are Common in Humans. PLoS Genetics, 9(11). https://doi.org/10.1371/journal.pgen.1003929.

Sasarman, F., Antonicka, H. and Shoubridge, E.A., 2008. The A3243G tRNALeu(UUR) MELAS mutation causes amino acid misincorporation and a combined respiratory

chain assembly defect partially suppressed by overexpression of EFTu and EFG2. Human Molecular Genetics, 17(23), pp.3697–3707. https://doi.org/10.1093/hmg/ddn265.

Saville, B.J., Kohli, Y. and Anderson, J.B., 1998. mtDNA recombination in a natural population. Proceedings of the National Academy of Sciences, 95(3), pp.1331–1335. https://doi.org/10.1073/pnas.95.3.1331.

Schaid, D.J., Chen, W. and Larson, N.B., 2018. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics, 19(8), pp.491–504. https://doi.org/10.1038/s41576-018-0016-z.

Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014. Biological insights from 108 schizophrenia-associated genetic loci. Nature, 511(7510), pp.421–427. https://doi.org/10.1038/nature13595.

Schönherr, S., Schachtl-Riess, J.F., Di Maio, S., Filosi, M., Mark, M., Lamina, C., Fuchsberger, C., Kronenberg, F. and Forer, L., 2024a. Performing highly parallelized and reproducible GWAS analysis on biobank-scale data. NAR Genomics and Bioinformatics, 6(1). https://doi.org/10.1093/nargab/lqae015.

Scott, I. and Youle, R.J., 2010. Mitochondrial fission and fusion. Essays in Biochemistry, 47, pp.85–98. https://doi.org/10.1042/bse0470085.

Selvaraj, M.S., Li, X., Li, Z., Pampana, A., Zhang, D.Y., Park, J., Aslibekyan, S., Bis, J.C., Brody, J.A., Cade, B.E., et al,., 2022. Whole genome sequence analysis of blood lipid levels in >66,000 individuals. Nature Communications, 13(1), p.5995. https://doi.org/10.1038/s41467-022-33510-7.

Sharpley, M.S., Marciniak, C., Eckel-Mahan, K., McManus, M., Crimi, M., Waymire, K., Lin, C.S., Masubuchi, S., Friend, N., Koike, M., Chalkia, D., MacGregor, G., Sassone-Corsi, P. and Wallace, D.C., 2012. Heteroplasmy of Mouse mtDNA Is Genetically Unstable and Results in Altered Behavior and Cognition. Cell, 151(2), pp.333–343. https://doi.org/10.1016/j.cell.2012.09.004.

Shoffner, J.M., Lott, M.T., Lezza, A.M.S., Seibel, P., Ballinger, S.W. and Wallace, D.C., 1990a. Myoclonic epilepsy and ragged-red fiber disease (MERRF) is associated with a mitochondrial DNA tRNALys mutation. Cell, 61(6), pp.931–937. https://doi.org/10.1016/0092-8674(90)90059-N.

De Silva, D., Tu, Y.-T., Amunts, A., Fontanesi, F. and Barrientos, A., 2015. Mitochondrial ribosome assembly in health and disease. Cell Cycle, 14(14), pp.2226–2250. https://doi.org/10.1080/15384101.2015.1053672.

Silva-Pinheiro, P. and Minczuk, M., 2022. The potential of mitochondrial genome engineering. Nature Reviews Genetics, 23(4), pp.199–214. https://doi.org/10.1038/s41576-021-00432-x.

Sinnwell, J.P., Therneau, T.M. and Schaid, D.J., 2014. The kinship2 R Package for Pedigree Data. Human Heredity, [online] 78(2), pp.91–93. https://doi.org/10.1159/000363105.

Slatkin, M., 2008b. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics, 9(6), pp.477–485. https://doi.org/10.1038/nrg2361.

Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., Ibrahim, A., Ji, Y., John, S., Lewis, E., MacArthur, J.A.L., McMahon, A., Osumi-Sutherland, D., Panoutsopoulou, K., Pendlington, Z., Ramachandran, S., Stefancsik, R., Stewart, J., Whetzel, P., Wilson, R., Hindorff, L., Cunningham, F., Lambert, S.A., Inouye, M., Parkinson, H. and Harris, L.W., 2023. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Research, 51(D1), pp.D977–D985. https://doi.org/10.1093/nar/gkac1010.

Song, Y., Wang, W., Wang, B. et al. The Protective Mechanism of TFAM on Mitochondrial DNA and its Role in Neurodegenerative Diseases. Mol Neurobiol 61, 4381–4390 (2024). https://doi.org/10.1007/s12035-023-03841-7

Spain, S.L. and Barrett, J.C., 2015. Strategies for fine-mapping complex traits. Human Molecular Genetics, 24(R1), pp.R111–R119. https://doi.org/10.1093/hmg/ddv260.

Speed, D. and Balding, D.J., 2019. SumHer better estimates the SNP heritability of complex traits from summary statistics. Nature Genetics, 51(2), pp.277–284. https://doi.org/10.1038/s41588-018-0279-5.

Speed, D., Holmes, J. and Balding, D.J., 2020. Evaluating and improving heritability models using summary statistics. Nature Genetics, 52(4), pp.458–462. https://doi.org/10.1038/s41588-020-0600-y.

Spinelli, J.B. and Haigis, M.C., 2018. The multifaceted contributions of mitochondria to cellular metabolism. Nature Cell Biology. [online] https://doi.org/10.1038/s41556-018-0124-1.

Spyropoulos, A., Manford, M., Horvath, R., Alston, C.L., Yu-Wai-Man, P., He, L., Taylor, R.W. and Chinnery, P.F., 2013. Near-Identical Segregation of mtDNA Heteroplasmy in Blood, Muscle, Urinary Epithelium, and Hair Follicles in Twins With Optic Atrophy, Ptosis, and Intractable Epilepsy. JAMA Neurology. https://doi.org/10.1001/jamaneurol.2013.4111.

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T.I., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M. and Lancet, D., 2016. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Current Protocols in Bioinformatics, 54(1). https://doi.org/10.1002/cpbi.5.

Stewart, J.B. and Chinnery, P.F., 2015. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. Nature Reviews Genetics, 16(9), pp.530–542. https://doi.org/10.1038/nrg3966.

Stewart, J.B. and Chinnery, P.F., 2021a. Extreme heterogeneity of human mitochondrial DNA from organelles to populations. Nature Reviews Genetics, 22(2), pp.106–118. https://doi.org/10.1038/s41576-020-00284-x.

Storey, J.D., 2002. A Direct Approach to False Discovery Rates. Journal of the Royal Statistical Society Series B: Statistical Methodology, 64(3), pp.479–498. https://doi.org/10.1111/1467-9868.00346.

Storey, J.D. and Tibshirani, R., 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences, 100(16), pp.9440–9445. https://doi.org/10.1073/pnas.1530509100.

Stranger, B.E., Stahl, E.A. and Raj, T., 2011. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics, [online] 187(2), pp.367–383. https://doi.org/10.1534/genetics.110.120907.

Sturtevant, A.H., 1913. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. Journal of Experimental Zoology, 14(1), pp.43–59. https://doi.org/10.1002/jez.1400140104.

Suissa, S., Wang, Z., Poole, J., Wittkopp, S., Feder, J., Shutt, T.E., Wallace, D.C., Shadel, G.S. and Mishmar, D., 2009. Ancient mtDNA Genetic Variants Modulate mtDNA Transcription and Replication. PLoS Genetics, 5(5), p.e1000474. https://doi.org/10.1371/journal.pgen.1000474.

Sun, L. and Dimitromanolakis, A., 2012. Identifying Cryptic Relationships. pp.47–57. https://doi.org/10.1007/978-1-61779-555-8_4.

Supinski, G.S., Schroder, E.A. and Callahan, L.A., 2020. Mitochondria and Critical Illness. Chest, 157(2), pp.310–322. https://doi.org/10.1016/j.chest.2019.08.2182.

Svishcheva, G.R., Axenovich, T.I., Belonogova, N.M., Van Duijn, C.M. and Aulchenko, Y.S., 2012. Rapid variance components-based method for whole-genome association analysis. Nature Genetics, 44(10), pp.1166–1170. https://doi.org/10.1038/ng.2410.

Tachibana, M., Sparman, M., Sritanaudomchai, H., Ma, H., Clepper, L., Woodward, J., Li, Y., Ramsey, C., Kolotushkina, O. and Mitalipov, S., 2009. Mitochondrial gene replacement in primate offspring and embryonic stem cells. Nature, 461(7262), pp.367–372. https://doi.org/10.1038/nature08368.

Tachibana, M., Amato, P., Sparman, M., Woodward, J., Sanchis, D.M., Ma, H., Gutierrez, N.M., Tippner-Hedges, R., Kang, E., Lee, H.-S., Ramsey, C., Masterson, K., Battaglia, D., Lee, D., Wu, D., Jensen, J., Patton, P., Gokhale, S., Stouffer, R. and Mitalipov, S., 2013. Towards germline gene therapy of inherited mitochondrial diseases. Nature, 493(7434), pp.627–631. https://doi.org/10.1038/nature11647.

Takeda, Y., Hyslop, L., Choudhary, M., Robertson, F., Pyle, A., Wilson, I., Santibanez-Koref, M., Turnbull, D., Herbert, M. and Hudson, G., 2023. Feasibility and impact of haplogroup matching for mitochondrial replacement treatment. EMBO reports, 24(10). https://doi.org/10.15252/embr.202154540.

Tang, D.-L., Zhou, X., Li, X., Zhao, L. and Liu, F., 2006. Variation of mitochondrial gene and the association with type 2 diabetes mellitus in a Chinese population. Diabetes Research and Clinical Practice, 73(1), pp.77–82. https://doi.org/10.1016/j.diabres.2005.12.001.

Tang, M., Wang, T. and Zhang, X., 2022. A review of SNP heritability estimation methods. Briefings in Bioinformatics, 23(3). https://doi.org/10.1093/bib/bbac067.

Taylor E.W., Xu J., Jabs E.W. and Meyers D.A., 1997. Linkage analysis of genetic disorders. Methods in molecular biology (Clifton, N.J.), pp.11–25.

Taylor, R.W. and Turnbull, D.M., 2005. Mitochondrial DNA mutations in human disease. Nature Reviews Genetics, 6(5), pp.389–402. https://doi.org/10.1038/nrg1606.

Tetsuka, S., Ogawa, T., Hashimoto, R. and Kato, H., 2021. Clinical features, pathogenesis, and management of stroke-like episodes due to MELAS. Metabolic Brain Disease, 36(8), pp.2181–2193. https://doi.org/10.1007/s11011-021-00772-x.

Thömmes, P., Farr, C.L., Marton, R.F., Kaguni, L.S. and Cotterill, S., 1995. Mitochondrial Single-stranded DNA-binding Protein from Drosophila Embryos. Journal of Biological Chemistry, 270(36), pp.21137–21143. https://doi.org/10.1074/jbc.270.36.21137.

Thompson, K., Stroud, D.A., Thorburn, D.R. and Taylor, R.W., 2023. Investigation of oxidative phosphorylation activity and complex composition in mitochondrial disease. pp.127–139. https://doi.org/10.1016/B978-0-12-821751-1.00008-7.

Thorisson, G.A., Smith, A. V., Krishnan, L. and Stein, L.D., 2005. The International HapMap Project Web site: Figure 1. Genome Research, 15(11), pp.1592–1593. https://doi.org/10.1101/gr.4413105.

Thorogood, A., 2020. International data sharing and rare disease: the importance of ethics and patient involvement. Rare Diseases.

Tieu, K. and Imm, J., 2014. Mitochondrial dynamics as a potential therapeutic target for Parkinson's disease? Advances in Clinical Neuroscience & Rehabilitation. https://doi.org/10.47795/RQGP4036.

Tifoun, N., De las Heras, J.M., Guillaume, A., Bouleau, S., Mignotte, B. and Le Floch, N., 2021. Insights into the Roles of the Sideroflexins/SLC56 Family in Iron Homeostasis and Iron-Sulfur Biogenesis. Biomedicines, 9(2), p.103. https://doi.org/10.3390/biomedicines9020103.

Tiranti, V., 1997. Identification of the gene encoding the human mitochondrial RNA polymerase (h-mtRPOL) by cyberscreening of the Expressed Sequence Tags database. Human Molecular Genetics, 6(4), pp.615–625. https://doi.org/10.1093/hmg/6.4.615.

Tzeng, I., 2022. Role of mitochondria DNA A10398G polymorphism on development of Parkinson's disease: A PRISMA-compliant meta-analysis. Journal of Clinical Laboratory Analysis, 36(3). https://doi.org/10.1002/jcla.24274.

Uffelmann, E., Huang, Q.Q., Munung, N.S., de Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T. and Posthuma, D., 2021. Genome-wide association studies. Nature Reviews Methods Primers, 1(1), p.59. https://doi.org/10.1038/s43586-021-00056-9.

Ulirsch, J.C., 2022. Identification and Interpretation of Causal Genetic Variants Underlying Human Phenotypes. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences. Harvard University.

Vaillant-Beuchot, L., Mary, A., Pardossi-Piquard, R., Bourgeois, A., Lauritzen, I., Eysert, F., Kinoshita, P.F., Cazareth, J., Badot, C., Fragaki, K., Bussiere, R., Martin, C., Mary, R., Bauer, C., Pagnotta, S., Paquis-Flucklinger, V., Buée-Scherrer, V., Buée, L., Lacas-Gervais, S., Checler, F. and Chami, M., 2021. Accumulation of amyloid precursor protein C-terminal fragments triggers mitochondrial structure, function, and mitophagy defects in Alzheimer's disease models and human brains. Acta Neuropathologica, 141(1), pp.39–65. https://doi.org/10.1007/s00401-020-02234-7.

van der Walt, J.M., Nicodemus, K.K., Martin, E.R., Scott, W.K., Nance, M.A., Watts, R.L., Hubble, J.P., Haines, J.L., Koller, W.C., Lyons, K., Pahwa, R., Stern, M.B., Colcher, A., Hiner, B.C., Jankovic, J., Ondo, W.G., Allen Jr., F.H., Goetz, C.G., Small, G.W., Mastaglia, F., Stajich, J.M., McLaurin, A.C., Middleton, L.T., Scott, B.L., Schmechel, D.E., Pericak-Vance, M.A. and Vance, J.M., 2003. Mitochondrial Polymorphisms Significantly Reduce the Risk of Parkinson Disease. The American Journal of Human Genetics, 72(4), pp.804–811. https://doi.org/10.1086/373937.

Vaxillaire, M. and Froguel, P., 2006. Genetic Basis of Maturity-Onset Diabetes of the Young. Endocrinology and Metabolism Clinics of North America, [online] 35(2), pp.371–384. https://doi.org/https://doi.org/10.1016/j.ecl.2006.02.009.

Vercellino, I. and Sazanov, L.A., 2022. The assembly, regulation and function of the mitochondrial respiratory chain. Nature Reviews Molecular Cell Biology, 23(2), pp.141–161. https://doi.org/10.1038/s41580-021-00415-0.

Vianello, C., Cocetta, V., Caicci, F., Boldrin, F., Montopoli, M., Martinuzzi, A., Carelli, V. and Giacomello, M., 2020. Interaction Between Mitochondrial DNA Variants and Mitochondria/Endoplasmic Reticulum Contact Sites: A Perspective Review. DNA and Cell Biology, 39(8), pp.1431–1443. https://doi.org/10.1089/dna.2020.5614.

Vincent, A.E., Rosa, H.S., Pabis, K., Lawless, C., Chen, C., Grünewald, A., Rygiel, K.A., Rocha, M.C., Reeve, A.K., Falkous, G., Perissi, V., White, K., Davey, T., Petrof, B.J., Sayer, A.A., Cooper, C., Deehan, D., Taylor, R.W., Turnbull, D.M. and Picard, M., 2018. Subcellular origin of mitochondrial DNA deletions in human skeletal muscle. Annals of Neurology, 84(2), pp.289–301. https://doi.org/10.1002/ana.25288.

Viscomi, C. and Zeviani, M., 2017. MtDNA-maintenance defects: syndromes and genes. Journal of Inherited Metabolic Disease, 40(4), pp.587–599. https://doi.org/10.1007/s10545-017-0027-5.

Visscher, P.M., Brown, M.A., McCarthy, M.I. and Yang, J., 2012. Five years of GWAS discovery. American Journal of Human Genetics, https://doi.org/10.1016/j.ajhg.2011.11.029.

Visscher, P.M., Hill, W.G. and Wray, N.R., 2008. Heritability in the genomics era — concepts and misconceptions. Nature Reviews Genetics, 9(4), pp.255–266. https://doi.org/10.1038/nrg2322.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A. and Yang, J., 2017a. 10 Years of GWAS Discovery: Biology, Function, and Translation. The American Journal of Human Genetics, 101(1), pp.5–22. https://doi.org/10.1016/j.ajhg.2017.06.005.

WALKER, E., HERNANDEZ, A. V. and KATTAN, M.W., 2008. Meta-analysis: Its strengths and limitations. Cleveland Clinic Journal of Medicine, 75(6), pp.431–439. https://doi.org/10.3949/ccjm.75.6.431.

Walker, M.A., Lareau, C.A., Ludwig, L.S., Karaa, A., Sankaran, V.G., Regev, A. and Mootha, V.K., 2020. Purifying Selection against Pathogenic Mitochondrial DNA in Human T Cells. New England Journal of Medicine, 383(16), pp.1556–1563. https://doi.org/10.1056/NEJMoa2001265.

Wallace, D.C., 1986. Mitotic segregation of mitochondrial dnas in human cell hybrids and expression of chloramphenicol resistance. Somatic Cell and Molecular Genetics, 12(1), pp.41–49. https://doi.org/10.1007/BF01560726.

Wallace, D.C., 1992. Mitochondrial Genetics: A Paradigm for Aging and Degenerative Diseases? Science, 256(5057), pp.628–632. https://doi.org/10.1126/science.1533953.

Wallace, D.C., 2012. Mitochondria and cancer. Nature Reviews Cancer, 12(10), pp.685–698. https://doi.org/10.1038/nrc3365.

Wallace, D.C., 2015. Mitochondrial DNA Variation in Human Radiation and Disease. Cell, 163(1), pp.33–38. https://doi.org/10.1016/j.cell.2015.08.067.

Wallace, D.C., Singh, G., Lott, M.T., Hodge, J.A., Schurr, T.G., Lezza, A.M.S., Elsas, L.J. and Nikoskelainen, E.K., 1988. Mitochondrial DNA Mutation Associated with Leber's Hereditary Optic Neuropathy. Science, 242(4884), pp.1427–1430. https://doi.org/10.1126/science.3201231.

Wang, D., Taniyama, M., Suzuki, Y., Katagiri, T. and Ban, Y., 2001. Association of the mitochondrial DNA 5178A/C polymorphism with maternal inheritance and onset of type 2 diabetes in Japanese patients. Experimental and Clinical Endocrinology & Diabetes, 109(07), pp.361–364. https://doi.org/10.1055/s-2001-17407.

Wang, F., Zhang, D., Zhang, D., Li, P. and Gao, Y., 2021. Mitochondrial Protein Translation: Emerging Roles and Clinical Significance in Disease. Frontiers in Cell and Developmental Biology, 9. https://doi.org/10.3389/fcell.2021.675465.

Wang, M.H., Cordell, H.J. and Van Steen, K., 2019a. Statistical methods for genome-wide association studies. Seminars in Cancer Biology, https://doi.org/10.1016/j.semcancer.2018.04.008.

Wanrooij, S. and Falkenberg, M., 2010. The human mitochondrial replication fork in health and disease. Biochimica et Biophysica Acta (BBA) - Bioenergetics, 1797(8), pp.1378–1388. https://doi.org/10.1016/j.bbabio.2010.04.015.

Wanrooij, S., Fusté, J.M., Farge, G., Shi, Y., Gustafsson, C.M. and Falkenberg, M., 2008. Human mitochondrial RNA polymerase primes lagging-strand DNA synthesis in vitro. Proceedings of the National Academy of Sciences, 105(32), pp.11122–11127. https://doi.org/10.1073/pnas.0805399105.

Wei, W. and Chinnery, P.F., 2020. Inheritance of mitochondrial DNA in humans: implications for rare and common diseases. Journal of Internal Medicine, 287(6), pp.634–644. https://doi.org/10.1111/joim.13047.

Wei, W., Gomez-Duran, A., Hudson, G. and Chinnery, P.F., 2017. Background sequence characteristics influence the occurrence and severity of disease-causing mtDNA mutations. PLOS Genetics, 13(12), p.e1007126. https://doi.org/10.1371/journal.pgen.1007126.

Wei, W., Tuna, S., Keogh, M.J., Smith, K.R., Aitman, T.J., Beales, P.L., Bennett, D.L., Gale, D.P., Bitner-Glindzicz, M.A.K., Black, G.C., et al,. 2019a. Germline selection shapes human mitochondrial DNA diversity. Science, 364(6442). https://doi.org/10.1126/science.aau6520.

Weissensteiner, H., Forer, L., Fendt, L., Kheirkhah, A., Salas, A., Kronenberg, F. and Schoenherr, S., 2021. Contamination detection in sequencing studies using the mitochondrial phylogeny. Genome Research, 31(2), pp.309–316. https://doi.org/10.1101/GR.256545.119.

Weissensteiner, H., Forer, L., Fuchsberger, C., Schöpf, B., Kloss-Brandstätter, A., Specht, G., Kronenberg, F. and Schönherr, S., 2016a. mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. Nucleic Acids Research, 44(W1), pp.W64–W69. https://doi.org/10.1093/nar/gkw247.

Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.J., Kronenberg, F., Salas, A. and Schönherr, S., 2016b. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Research, 44(W1), pp.W58–W63. https://doi.org/10.1093/nar/gkw233.

Whittaker, R.G., Blackwood, J.K., Alston, C.L., Blakely, E.L., Elson, J.L., McFarland, R., Chinnery, P.F., Turnbull, D.M. and Taylor, R.W., 2009. URINE HETEROPLASMY IS THE BEST PREDICTOR OF CLINICAL OUTCOME IN THE m.3243A>G mtDNA MUTATION. Neurology, [online] 72(6), p.568. https://doi.org/10.1212/01.wnl.0000342121.91336.4d.

Willer, C.J., Li, Y. and Abecasis, G.R., 2010. METAL: Fast and efficient meta-analysis of genomewide association scans. Bioinformatics, 26(17), pp.2190–2191. https://doi.org/10.1093/bioinformatics/btq340.

Wilson, A. and Cann, R., 1992. The recent African genesis of humans. Sci Am, pp.68–73.

Wilson, D.F., 2017. Oxidative phosphorylation: regulation and role in cellular and tissue metabolism. The Journal of Physiology, 595(23), pp.7023–7038. https://doi.org/10.1113/JP273839.

Wilson, I.J., Carling, P.J., Alston, C.L., Floros, V.I., Pyle, A., Hudson, G., Sallevelt, S.C.E.H., Lamperti, C., Carelli, V., Bindoff, L.A., Samuels, D.C., Wonnapinij, P., Zeviani, M., Taylor, R.W., Smeets, H.J.M., Horvath, R. and Chinnery, P.F., 2016. Mitochondrial DNA sequence characteristics modulate the size of the genetic bottleneck. Human Molecular Genetics, 25(5), pp.1031–1041. https://doi.org/10.1093/hmg/ddv626.

Winney, B., Boumertit, A., Day, T., Davison, D., Echeta, C., Evseeva, I., Hutnik, K., Leslie, S., Nicodemus, K., Royrvik, E.C., Tonks, S., Yang, X., Cheshire, J., Longley, P., Mateos, P., Groom, A., Relton, C., Bishop, D.T., Black, K., Northwood, E., Parkinson, L., Frayling, T.M., Steele, A., Sampson, J.R., King, T., Dixon, R., Middleton, D., Jennings, B., Bowden, R., Donnelly, P. and Bodmer, W., 2012. People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population. European Journal of Human Genetics, 20(2), pp.203–210. https://doi.org/10.1038/ejhg.2011.127.

Winter, B., 2013. Linear models and linear mixed effects models in R with linguistic applications.arXiv.

Xu, J., Postma, D.S., Howard, T.D., Koppelman, G.H., Zheng, S.L., Stine, O.C., Bleecker, E.R. and Meyers, D.A., 2000. Major Genes Regulating Total Serum Immunoglobulin E Levels in Families with Asthma. The American Journal of Human Genetics, 67(5), pp.1163–1173. https://doi.org/10.1086/321190.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Goddard, M.E. and Visscher, P.M., 2010a. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics, 42(7), pp.565–569. https://doi.org/10.1038/ng.608.

Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M., 2011. GCTA: A Tool for Genome-wide Complex Trait Analysis. The American Journal of Human Genetics, 88(1), pp.76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.

Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M. and Price, A.L., 2014a. Advantages and pitfalls in the application of mixed-model association methods. Nature Genetics, 46(2), pp.100–106. https://doi.org/10.1038/ng.2876.

Yasukawa, T., 2001. Wobble modification defect in tRNA disturbs codon-anticodon interaction in a mitochondrial disease. The EMBO Journal, 20(17), pp.4794–4802. https://doi.org/10.1093/emboj/20.17.4794.

Yellen, G., 2018. Fueling thought: Management of glycolysis and oxidative phosphorylation in neuronal metabolism. Journal of Cell Biology, 217(7), pp.2235–2246. https://doi.org/10.1083/jcb.201803152.

Yonova-Doing, E., Calabrese, C., Gomez-Duran, A., Schon, K., Wei, W., Karthikeyan, S., Chinnery, P.F. and Howson, J.M.M., 2021. An atlas of mitochondrial DNA genotype–phenotype associations in the UK Biobank. Nature Genetics, 53(7), pp.982–993. https://doi.org/10.1038/s41588-021-00868-1.

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S. and Buckler, E.S., 2006. A unified mixed-model method for association mapping that accounts for multiple

levels of relatedness. Nature Genetics, 38(2), pp.203–208. https://doi.org/10.1038/ng1702.

Yue, P., Jing, S., Liu, L., Ma, F., Zhang, Y., Wang, C., Duan, H., Zhou, K., Hua, Y., Wu, G. and Li, Y., 2018. Association between mitochondrial DNA copy number and cardiovascular disease: Current evidence based on a systematic review and meta-analysis. PLOS ONE, 13(11), p.e0206003. https://doi.org/10.1371/journal.pone.0206003.

Yu-Wai-Man, P., Griffiths, P.G., Hudson, G. and Chinnery, P.F., 2009. Inherited mitochondrial optic neuropathies. Journal of Medical Genetics, https://doi.org/10.1136/jmg.2007.054270.

Zavala, C., Serao, N., Villamil, M.B., Caetano-Anolles, G. and Rodriguez-Zas, S.L., 2011. Additive and multiplicative genome-wide association models identify genes associated with growth. In: 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW). IEEE. pp.975–977. https://doi.org/10.1109/BIBMW.2011.6112527.

Zeviani, M. and Donato, S. Di, 2004. Mitochondrial disorders. Brain, [online] 127, pp.2153–2172. https://doi.org/10.1093/brain/awh259.

Zhang, J., Baran, J., Cros, A., Guberman, J.M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., Wong-Erasmus, M., Yao, L. and Kasprzyk, A., 2011. International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. Database, 2011(0), pp.bar026–bar026. https://doi.org/10.1093/database/bar026.

Zhang H, Burr SP, Chinnery PF. The mitochondrial DNA genetic bottleneck: inheritance and beyond. Essays Biochem. 2018 Jul 20;62(3):225-234. doi: 10.1042/EBC20170096.

Zhang, Y. and Pan, W., 2015. Principal Component Regression and Linear Mixed Model in Association Analysis of Structured Samples: Competitors or Complements? Genetic Epidemiology, 39(3), pp.149–155. https://doi.org/10.1002/gepi.21879.

Zheng, S.L., Henry, A., Cannie, D., Lee, M., Miller, D., McGurk, K.A., Bond, I., Xu, X., Issa, H., Francis, C., De Marvao, A., et al,., n.d. Genome-wide association analysis reveals insights into the molecular etiology underlying dilated cardiomyopathy. HERMES Consortium † , Folkert W Asselbergs, [online] 36. https://doi.org/10.1101/2023.09.28.23295408.

Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., Bastarache, L.A., Wei, W.Q., Denny, J.C., Lin, M., Hveem, K., Kang, H.M., Abecasis, G.R., Willer, C.J. and Lee, S., 2018a. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature Genetics, 50(9), pp.1335–1341. https://doi.org/10.1038/s41588-018-0184-y.

Zhou, X. and Stephens, M., 2012. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics, 44(7), pp.821–824. https://doi.org/10.1038/ng.2310.

Zhu, H. and Zhou, X., 2020b. Statistical methods for SNP heritability estimation and partition: A review. Computational and Structural Biotechnology Journal, 18, pp.1557–1568. https://doi.org/10.1016/j.csbj.2020.06.011.

Zhu, N., LeDuc, C.A., Fennoy, I., Laferrère, B., Doege, C.A., Shen, Y., Chung, W.K. and Leibel, R.L., 2023. Rare predicted loss of function alleles in Bassoon (BSN) are associated with obesity. npj Genomic Medicine, 8(1), p.33. https://doi.org/10.1038/s41525-023-00376-7.

Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R. and Lander, E.S., 2014. Searching for missing heritability: Designing rare variant association studies. Proceedings of the National Academy of Sciences, 111(4). https://doi.org/10.1073/pnas.1322563111.