Lancaster University

MSc by research Biomedical Sciences

Making Mitochondrial Haplogroup and DNA Sequence Predictions from Low-Density Genotyping Data

Author
Emma Lucy Eve Drummond, BA

Supervisors
Dr. David J. Clancy
Prof. Joanne Knight

13th July 2021

The thesis is my own work, and has not been submitted in substantially the same form for the award of a higher degree elsewhere.

Emma L.E. Drummond

Dedicated to my husband, Neil, my children, David and Clare, and my parents, Virginia and Alan.

Acknowledgements

I owe a great deal to my supervisors, Dave Clancy and Jo Knight, for their patient support for longer than expected, due to a global pandemic and many months of home-schooling. The research flowed more easily than the writing and their feedback enabled me to write a document which does justice to the project's achievements.

My chief cheerleader, Neil Drummond, cleared dozens of weekends to entertain and exercise our children, allowing me many hours in the shed fortified with litres of tea.

Contents

List of Tables

List of Figures

Abstract

Author: Emma Lucy Eve DRUMMOND, BA
Supervisors: Dr. David J. CLANCY and Prof. Joanne KNIGHT
For the degree of: MSc in Research Biomedical Sciences,
Lancaster University
13th July 2021

The mitochondrial genome (mtDNA) is inherited differently and mutates more frequently than the genetic material residing in the cells' nucleus. Whilst the genome of the mtDNA is small, at only 16.5 kilobases, it contains key components of the metabolic chain, and must communicate in a precise and timely way with the genes in the nuclear genome and sense the minute-to-minute needs of its host cell. MtDNA is an underexplored place to search for health-related variants.

Unlike the time-consuming and expensive methods of whole genome sequencing, genotyping examines certain positions in the genome allowing imputation of the other variants typically linked to these positions. Current methods, which use nuclear genome data to model their predictions, do not tailor imputation to take advantage of the different inheritance patterns of the mtDNA.

I present a novel method, using an open-source library of fully sequenced mtDNA samples with manually assigned haplogroups, to take genotyping data and predict the other variants present in the sample's mtDNA sequence, a two-stage method referred to as *in silico* genotyping and barcode matching. The method has been assessed for performance on a test data set to explore inconsistencies across the mitochondrial genome and the human mtDNA phylogeny. The first use of *in silico* genotyping and barcode matching is presented; extending the use of UKBiobank's data [22].

The UKBiobank represents data which is not only rich in detail but also covers a large population of individuals aged between 51 and 84 in 2021. The phenotypic data is health-focussed, including general health records, which is being augmented by new diagnoses or events in the participants' medical history.

Extensive use is being made of the data in UKBiobank with the exception of the mitochondrial DNA (mtDNA). The scale of the phenotypic data collected by the UKBiobank is proving a valuable resource, values all the more because of the difficulty and expense of its collection. Making further use of phenotyping by extending potential associations into the mtDNA is vital, and likely to offer substantial rewards.

Using the method described below to transform genotypes into predicted mtDNA sequence opens the doors for mitochondrial variation to be put to considerable use too.

The introduction presents evidence that: (a) the mitochondrion is essential for cell and organism function, (b) mtDNA can harbour variations associating with phenotypes, and (c) the current methods of mtDNA imputation can be improved upon.

The method presented mimics any genotyping microarray to produce a library of data transformed to appear as if it had been genotyped by the physical array. The effectiveness and accuracy of this transformation have been investigated and the results are presented.

Finally, the transformed library is used to predict the UKBiobank participant data to greatly extend a data set with huge reserves of potential especially for mitochondrial data.

My development of *in silico* genotyping and barcode matching has allowed me to make weighted prediction for test samples, guessing their haplogroups and the variants they carry. Whilst I admit to the significant potential to improve algorithms, the overall accuracy of these predictions is at a level high enough to search for links between UKBiobank samples and their phenotypic data in a GWAS-style search.

Chapter 1

Introduction

1.1 Mitochondria

1.1.1 Introduction

The idealised mitochondrion, familiar to many, is a small, lozenge-shaped cell organelle, filled with a pleated inner membrane separating two nested compartments. The cells' "powerhouse" according to very many sources, including Andersson et al. [12], the mitochondrion is an ubiquitous source of energy. This thumb-nail sketch fails to capture the complexities of a vital organelle balanced in an "endosymbiotic relationship" [88] with a host and containing a partial genome.

This review shall cover the historical study of mitochondria from their discovery to the publication of their genome. Then a closer examination of their structure and functions, followed by an examination of the relevant methods of mitochondrial genetic data analysis.

1.1.2 Discovery and Early Advances

In 1841, the mitochondrion was noted as a thready, grainy object by Henle in his paper with Müller, [80] and named by [15], coining the term "Mitochondrion" using the Greek for thread and grain, "mitos" and "chondros", respectively.

Two years later, Michaelis noted that Janus Green B, a redox method of cell stain, preferentially stained the mitochondria in cell slides [77]. The use of Janus Green B quickly became the predominant route for visualising mitochondria. The method was especially valued as the dye is "taken up only by metabolically active mitochondria in which the inside-negative membrane potential is intact" [8]. Using Janus Green B differentiated cell contents from healthy, metabolising mitochondria and continues to this day to aid colourimetric investigations into the health of mitochondria [8].

However, Ernster and Schatz [36] observe that mitochondria's responsiveness to a redox stain offered very significant clue as to the inner workings of the organelle, but a clue which went unnoticed. Despite Michaelis' dual interests in the mitochondrion and "biological redox processes", his observation did not prompt a theoretical bridge between the stain's chemistry and the function of the mitochondrion [36].

As early as 1890, Richard Altmann had imagined mitochondria were intracellular organisms in "Die elementarorganismen und ihre beziehungen zu den zellen". This tome is recognized as a transformative early histological description of mitochondria, or in his terminology, "bioblasts" (life germs), which he believed were autonomous elementary organisms responsible for metabolic and genetic functions [10]. This idea was reiterated by Mereschowsky in 1910, buoyed by his work on the symbiotic lichens, he proposed that mitochondria and chloroplasts originated through similar marriages [61]. Many decades would pass before it was realised these were indeed prescient ideas.

In 1912, Kingsbury tells us the mitochondria are "a structural expression of the reducing substances concerned in cellular respiration" [60], a "foresighted conclusion" according to Ernster and Schatz who note that, despite the lack of knowledge of the chemical processes, links between structure and function were proposed [36].

Fusion and fission of mitochondria was witnessed and recorded very early in the exploration of mitochondrial structure. Several mitochondria formed a tube-like structure to cleave again were captured in time-series sketches. Lewis and Lewis comment on the how active these structures were; moving, fusing and dividing. They include a careful time series drawings of the changing mitochondria [70].

Bensley and Hoerr [16] published a technique in 1934 which allowed the harvesting of mitochondria from tissues, although it took another 14 years of development by Hogeboom et al. [49] before the collected mitochondria were thought to be "well preserved" [36].

As with many biological systems, structure and function are intricately related so this observational, non-chemical approach remained for several decades. Ernster [36] discuss the rewards for researchers exploring the chemistry of the organelle, with a burst of activity in revealing the chemical pathways. The citric acid cycle, elucidated by Krebs in 1937, was investigated by Kalckar using kidney mitochondria to show oxidative phosphorylation steps.

The late 1940s saw the evidence that the citric acid cycle, oxidative phosphorylation and fatty acid oxidation all firmly sited in the mitochondria, and soon after the respiration chain was located to a mitochondrial membrane, although which membrane was yet to be discovered.

Eventually, over half a century after the first observation, Lazarow and Cooperstein finally associated the chemistry of Janus Green B staining to how the organelle's functions are linked to its structure. They observed that the action of this stain was actually the different chemical environments of the cell lumen and the contents of the mitochondria. The differential staining of the mitochondria could be removed using cyanide, and would return on the removal of the cyanide, confirming a difference between the redox environment of the cytoplasm and the mitochondria [68].

Improvements in techniques and visualisation through the 1950s were revealing the morphology of the MT and both Sjostrand and Palade put forth models explaining the images seen. The septa model championed by Sjostrand [93] suggested that these invaginations sectioned the matrix, leaving the central space entirely compartmentalised. Palade disagreed, considering the matrix a single space, with deep, parallel grooves, like a ruler's pressure on the side of a water balloon, increasing surface area of the inner membrane without dividing the central space [83].

The late 1950s saw considerable successes exploring the components of the respiratory chain by Kelin and King, and Hatefi at al., [36]. On this basis, Azzone and Ernster began the development of the chemiosmotic hypothesis, with the proton pumping across the membrane powered by the oxidative phosphorylation steps [14]. This theory was further fleshed out by Mitchell in 1966 [78].

Once the critical role of the mitochondrial function in the cell had been realised the health implications of faulty mitochondria became apparent. In 1962, Luft published his observations on the first example of mitochondrial disease; non-thyroidal hyper-metabolism [72]. This paper tells us that these poorly functioning mitochondria were due to an uncoupling of the phosphorylation, which allowed metabolism to run at a high pace without gain. This condition mimicked the respiratory changes seen in patients administered dinitrophenol, known to uncouple the chain of reaction in the mitochondria and prevent ATP production. The careful description of the patient's symptoms highlights how a mitochondrion was not just a mere battery. The organelle must respond to the energy draw of its host cell, scaling up or back according to requirements.

The intimacy between studies of mitochondrion's shape and its function is discussed by Frey and Manella, remarking on how the two parallel one another, and discovery of one is limited by techniques in the other. Once sample preparation or microscopy developments are made, biochemical details can further advance too [39].

Further developments in methods in the mid 1960s allowed researchers to separate the inner and outer membrane, elucidating the significant differences between the two. Two papers, Levy et al. and Schnaitman et al. , forged a technique exploiting the membrane composition differences, largely in cholesterol concentrations, to fractionally separate the two membranes for further analysis [69, 89]. This permitted a better understanding and characterisation of the many differences in content and behaviour of the two lipid bilayers, according to Ernster [36]. Cell membranes are two-dimensional fluids, and as such, substances adhered to either lipid bilayer, or traversing the membrane, will move in the surface. The membrane complexes' concentrations would equilibrate unless the proteins were bound to a structural frame or limited by a boundary.

Beyond the excellent review by Ernster and Schatz, the year 1981 saw two other significant publications. Anderson et al. published the first complete genome, that of the human mitochondrion, which was welcomed as a great step forward from the detailed restriction mapping [11]. In addition, Margulis (1981) reignited the idea that mitochondria were derived from free-living precursors and proposed other domesticated "endosymbionts" and the "serial endosymbiosis theory" [76]. Although initially meeting with resistance, this idea has gained more and more traction.

The observations of common sequences and the creation of phylogenies have confirmed the emergence of the mitochondrial line from a bacterial ancestor. Molecular clock techniques suggest dates of about 1.6 billion years have passed since an α-proteobacteria became a permanent fixture in a cell with which they still have a very close genetic homology especially among the protein complex units of the respiratory chain [12, 62].

1.1.3 Structure and Function

Frey and Mannella provide a description of the general structure of mitochondria; as "an organelle bound by two membranes – outer and inner – which enclose a dense matrix that includes enzymes of intermediary metabolism and multiple copies of a genome that encodes for a few inner-membrane proteins and the RNAs needed for their translation" [39].

Ryan and Hoogenraad [87] broadly divide the mitochondrion into four distinct reaction regions; the outer membrane, the inter-membrane space, the inner membrane and the matrix. As the membranes are physically separated, and then act to separate the two internal spaces, the chemistry and contents have been allowed to diverge and hone their specialisms.

Outer Membrane

The mitochondrial outer membrane is the surface in contact with the cell. It is penetrable to all but large molecules and holds no transmembrane potential.

The mitochondrion must communicate bi-directionally with its host cells. The energy production capacity and current usage need to be close to a balance as storage of ATP is not an option. Any chemical needs of the mitochondrion, for example replacement of damaged proteins, must be sent for as the vast majority of genes which code for mitochondrial proteins are hosted in the nucleus.

Whilst details of these communication systems are being revealed in papers such as [34], the complexity and the importance of the communication issue cannot be understated. As it is the surface exposed to the cytosol, the outer membrane is the site of membrane-tethered proteins which do functions from initiating fusion with other mitochondria or inducing apoptosis of the host cell [62].

The mitochondrial outer membrane houses general porin proteins, which form non-specific channels for small molecules to diffuse through, but carries no membrane potential as ions are free to move. The outer membrane is also the site of many translocases, which move certain, larger molecules at a more controlled rate. The amino acid chains in the cytosol which are precursors of mature proteins destined for the mitochondrion are captured by translocase proteins and fed through the membrane to the intermembrane space [62].

A very significant protein in the mitochondrial outer membrane is the voltage-dependant anion channel, or VDAC. Tan and Columbini write that VDAC complexes stud the outer membrane, which suggests an importance in the functioning of the mitochondrion [96]. Das and Steenberg say the VDAC is the principal route for the movement of ions and metabolites into the mitochondrial interior [29].

The pores the VDAC creates are limited in size but leave the MOM open to the entry of chemicals such as the substrates of the respiratory chain from the cytosol [106].

Nucleus-made amino-acid chains targeting the MT are able to pass through the pores in the MOM but remain in the intermembrane space. They are prevented from either proceeding to the matrix or returning the cytosol by the creation of permanent folds. The intermembrane space is the site of the oxidative folding of proteins which have entered the mitochondrion. Folding by the bonding of disulphide bridges on pairs of cysteine residues in the protein chains, prevents their slipping back through the relatively large pores of the VDAC, trapping them in the mitochondrion's intermembrane space.

Furthermore, VDAC proteins have been found to play a part in the formation of a huge complex which forms a direct importation pore to the matrix by binding the two membranes together [24]. According to Kuhlbradt, this complex appears to be formed around the incoming nascent protein chain [62].

In relative abundance and sitting in the outer membrane allows the targetting of mitochondria through the targetting of the VDAC. Shoshan notes that the VDAC is a perfect marker of mitochondrial membranes being both positioned on the surface and numerous [92].

Intermembrane Space

The intermembrane space is semi-permeable, with the permeability largely dependant on molecule size, so concentrations of smaller molecules are equal to those in the cytosol. Proteins produced for this compartment and created in the cytosol are drawn in through porins and prevented from escaping through the creation of permanent folds made by disulphide bridges between cysteine residues.

A common component in the intermembrane space is a protein called Cytochrome C. In healthy mitochondria, the Cytochrome C works hard to transfer electrons in the redox chain (see 1.1) and is a vital part of respiration. However, should the Cytochrome C escape the mitochondrion, we see "apoptotic induction upon its release into the cytosol" [87]. Spillage of Cytochrome C is a useful marker of defective or damaged mitochondria and induces apoptosis.

A significant portion of the intermembrane space is termed "cristae". These are hard to generalise as they form different structures according to the metabolic needs of the host cell, with an appropriate amount of surface area for the complexes [52]. They act to increase the surface area of the inner membrane, but also have some major differences from the remainder of the inner membrane and intermembrane space.

The intermembrane space is kept narrow by the mitochondrial contact site and cristae-organising system (MICOS) and mitochondrial intermembrane-space bridging (MIB) complexes of proteins. These are built from membrane-bound proteins in both the inner and outer membranes and act to pin the membranes together.

Huynen et al. found that the concentration of the MICOS complex correlates well with the volume of the intermembrane space classified as cristae [52], which itself has been noted to correlate well with the host cell's ATP usage. It is that a variety of crista structural differences are seen according to Frey and Mannella who tell us of observation on brown adipose tissue where the crista are plate-like, stacked along the mitochondrion [39]. Ernster asserts one rule; that the amount of space reserved for cristae correlates positively with respiratory rates of the cell [36].

The importance of the membrane structure and stabilisation is highlighted as this MICOS complex breaks up as cells age. The inner membrane looses contact with the outer membrane, folds in the inner membrane are lost and the mitochondrion falls into a state of dysfunction. Kuhlbrandt et al. paint a vivid story of a descent into cell death as the complexes which hold the membranes in shape fail to bind, the structures inside the mitochondria disintegrate and respiration ability drops. The matrix forms discrete pockets in the mitochondria and the outer membrane fails to enclose the mitochondria contents, triggering apoptosis of the cell [62].

The varied structures of the cristae have commonalities which point to their function. Narrow necks to the invaginations concentrate their contents and create a distinct second environment in the intermembrane space. As protons are pumped out of the matrix, the electrochemical proton gradient created is enhanced by the partial sealing of the cristae which retains the protons to allow faster recycling [62].

The Electron-Transport Chain

The cristae is both a specialised region of the intermembrane space and form a specialised subset of the inner membrane because the cristae membrane are the site of the large, multi-unit complexes which form the electron transport chain, passing the electron gleaned from metabolism to the similarly cristae-sited ATP synthase complex [62]. Kuhlbrandt et al. also include one of the best diagrams of this chain, see diagram 1.1.

Conversion of fuels such as sugars into energy accessible to the processes in the cell is vital. Adenosine triphosphate is a energy-primed molecule which enables the further priming of enzymatic processes that would otherwise not proceed quickly enough or at all in the conditions of the cell. Conformational changes made by the transfer of the phosphate group from the ATP to the enzyme couple a energetically unfavourable reaction to one which is favourable; the hydrolysis of the phosphate bond leaving adenosine diphosphate (ADP) and an inorganic phosphate ion. Its ubiquity makes ATP the most popular energy currency across cell types [9].

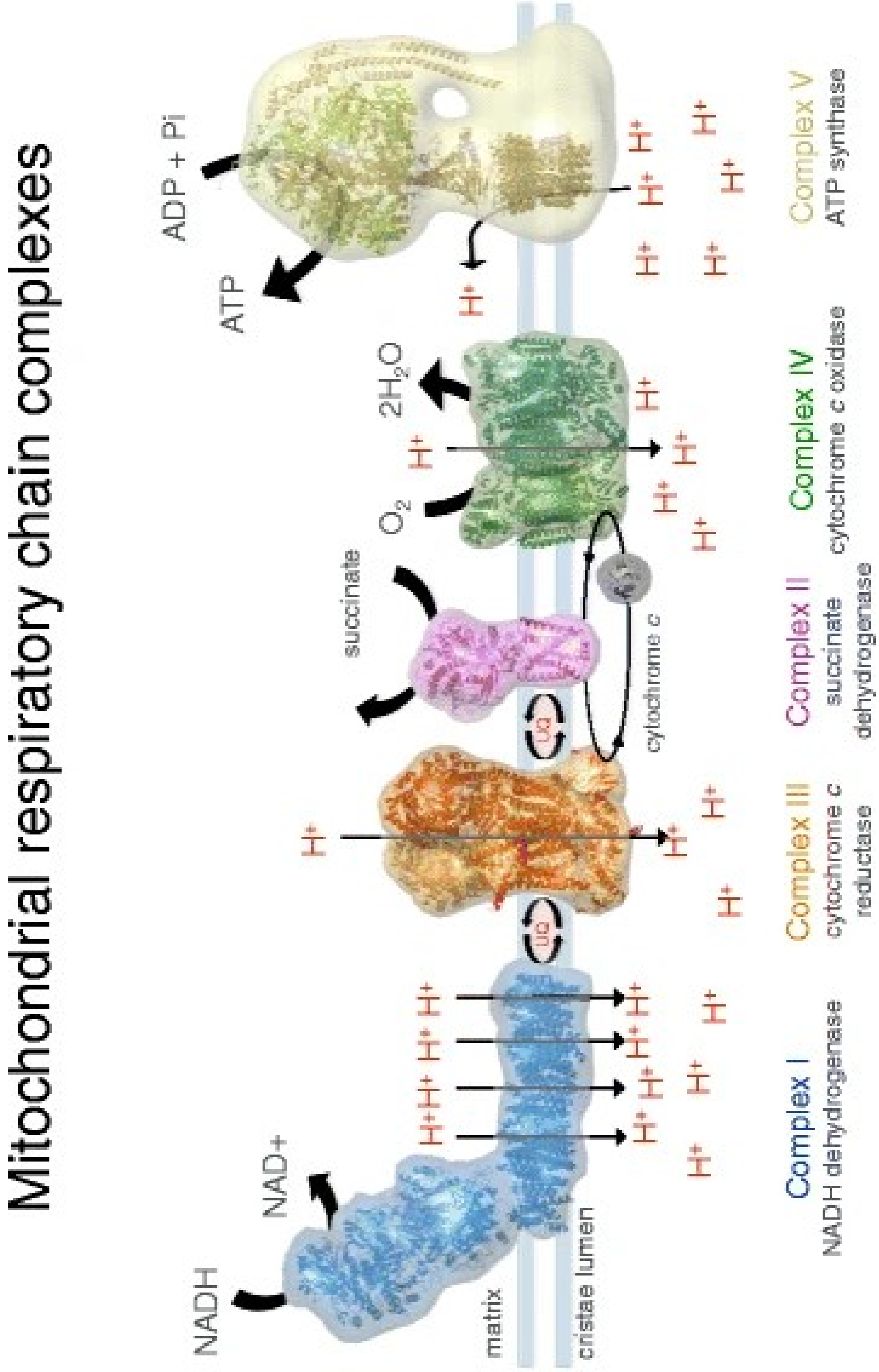


Figure 1.1: The electron-transfer chain from Kuhlbrandt et al. (2015) [62]

Mitochondria are not the only cell sites where ATP is produced. The first stages of metabolism, anaerobic glycolysis occurs in the cytosol. However, little ATP is harvested in the journey from glucose to pyruvate. Alberts et al. explain how this very partial metabolism is substantially improved upon as the pyruvate is further oxidised, within the mitochondria. This secondary metabolism increases ATP yield 15-fold and takes the pyruvate, with the addition of O_2 to CO_2 and H_2O [9].

The reactions in the cytosol prime the glucose using the hydrolysis of two ATP molecules, but yield four and two NADH molecules, primed to transport electrons. The resulting pyruvate is pumped into the mitochondria to react further in the citric acid cycle.

Oxidative phosphorylation (OXPHOS) is the main process housed in the mitochondria and from where the mitochondria get their reputation for being like batteries. Frey and Mannella [39] place the mitochondrion at the very centre of metabolism in eukaryotic life, and much of prokaryotic life too.

The harvesting of ATP from the step-wise transfer of high energy electrons is in two parts. Electron transfers through a chain of complexes seated in the inner mitochondrial membrane allow the membrane proteins to pump H^+ ions (protons) out of the matrix into the crista lumen.

Frey and Manella provide a good general description, taking us through chemiosmosis, where controlled oxidation is harnessed to the pumping of protons out of the matrix into the cristae lumen. The ATP synthase uses the energetically favourable flow of protons back into the matrix to power the conversion of ADP to ATP with the addition of a free phosphate ion [39]. Helpfully, Kuhlbrandt et al. quantify this activity of this recycling process, estimating that perhaps 50kg a day were made in an average human but hinting at an even higher volume for anyone engaged in more intensive activity for extended periods [62].

Kuhlbrandt et al. explain it well in their prose:

> "The proton gradient across the cristae membrane is generated by three large membrane protein complexes of the respiratory chain in the cristae, known as complex I (NADH/ubiquinone oxidoreductase), III (cytochrome c reductase) and IV (cytochrome c oxidase). Complex I feeds electrons from the soluble carrier molecule NADH into the respiratory chain and transfers them to a quinol in the membrane. The energy released in the electron transfer reaction is utilized for pumping four protons from the matrix into the crista lumen. Complex III takes the electrons from the reduced quinol and transfers them to the small, soluble electron carrier protein cytochrome c, pumping one proton in the process. Finally, complex IV transfers the electrons from cytochrome c to molecular oxygen and contributes to the proton gradient by using up four protons per consumed oxygen molecule to make water. Complex II (succinate dehydrogenase) transfers electrons from succinate directly to quinol and does not contribute to the proton gradient"[62]

To further concentrate the efforts of the electron transfer chain, the chain of large complexes in the cristae membranes also assemble into supercomplexes sometimes called respirosomes. This is where the free floating Cytochrome C molecule becomes so useful shuttling electrons around the active sites of this supercomplex and is also retained in the crista lumen [62].

The mechanistic qualities of the ATP synthase are nodded to by Kuhlbrandt, amongst others, calling it an "ancient nanomachine". The ATP synthase spins like a turn-stile, pushed by the protons keen to correct their concentration and charge imbalances across the inner membrane. The imbalances are exploited to power ATP's production [62]. The ATP synthase complex looks and behaves like a turbine, with a "rotor ring" and central shaft forcing the torque produced by the flow of protons to create conformation changes to the active site [62].

The positioning of the turbine-like units in the inner membrane surface was initially thought to be random but dimers which were found in paired rows in the fractured membranes of frozen specimens. Observed first in yeast, these arrays were seen in each species examined in this way. The complex arrays were consistently positioned on the curved surfaces of the cristae, prompting debates as to the direction of causation: were the complexes trapped by the membrane curve, or causing the deformation by aligning? Davies et al. ended the debate with their report that they had found that it was the association of dimer row forming the shape of the membrane [30].

Although not lethal in yeast, evidence from strains lacking the ability to dimerise suggests measurable disadvantages visible in phenotypes. These mutant strain grow 60% slower and there is also a 50% drop in the potential across mitochondrial inner membrane [17].

Inner Membrane

Intuitively, the crista lumen feels different to the rest of the inter-membrane space because of the local concentration changes but the cristae junctions also have an effect on the inner membrane. The cristae junctions leave the cristae-based proteins unable to diffuse freely with the rest of the inner membrane, allowing two distinct membrane environments to exist.

The inner membrane is very different to the outer membrane as its low permeability is vital to the function of the mitochondrion. Kuhlbrandt et al. describe it as a "tight diffusion barrier to all ions and molecules" punctured only by very specific ion or molecule transporter proteins [62].

Alongside the membrane transport proteins, the inner membrane is the home of five respiration complexes forming a reaction chain. Complexes I-IV use the oxidation of metabolites to pump protons from the matrix through to the intermembrane space. Complex V (ATP synthase) tethers the return flow of proteins back into the matrix to creating ATP. The stores the chemical energy created during metabolism and allows the energy to be used throughout the cell. These complexes are bound in complexes in the inner membrane, and they are largely trapped within the cristae membrane.

Membrane-bound complexes have created a problem for mitochondrial protein replacement. Many mitochondrial genes have been abducted and sheltered in the nuclear genome. Proteins bound for the mitochondria are drawn through pores in the outer membrane, as discussed above, however this creates a dilemma in the production of membrane-bound hydrophobic complex units, such as those in the centres of the respiration complexes. Those of the hydrophobic, membrane-bound complex centres are the very few genes still housed in the mitochondrial genome, alongside those which code for the tRNAs to allow translation. The mitochondrial ribosomes, which all sit conveniently in the inner membrane, create the vital, non-transportable, hydrophobic proteins which are immediately integrated in the phospholipid membrane [62].

Matrix

The matrix is noted as being of a high pH and very high protein content, both bound by the tight seal of the inner membrane. The proteins include the mitochondrial ribosomes and their tRNA for translating the mtDNA molecules, and the synthesis and metabolism of the mitochondrial work load. The protein content needed for all of the activity is close to the crystallisation point [62].

Whilst the complexes of the electron transport chain are integrated into the inner membrane, their substrates, the metabolites, are resident in the matrix. As are the by-products, the reactive oxygen species (ROS). Approximately 90% of a cell's ROS are found in the mitochondrial matrix [66].

Mitochondrial DNA is stored, translated, transcribed and replicated in the matrix. First thought to be free-floating, the DNA was located, by Kukat et al., in nucleoids [64, 63]. Compaction of nuclear DNA is often used as partial protection from chemical damage, and we see this in the nucleoids, however this storage also performs a second function. The nucleoids contain a single DNA molecule and are adsorbed on the inside of the inner membrane. This physical linking ensure that mitochondrial fission produced two mitochondria both replete with DNA molecules.

Iborra, Kimura and Cook investigated the physical position of the DNA in the mitochondrial matrix of a stable human cell line where YFP was expressed in the MT. They found the mtDNA to be in packaged units, each containing between six and ten mtDNA molecules, and tethered to the inner membrane by kinesin . There are several reasons suggested for the close packing of several genome copies; to improve translation efficiency through proximity; to coordinate transcription; or genome replication. Dissolving the anchorage of the nucleoid to the inner membrane was shown to increase the number of mitochondria devoid of mtDNA after fission, highlighting the benefits of the kinesin tethering [53].

Despite the compaction into nucleoids, mtDNA remains exposed to conditions which are not optimal for the integrity of the genetic information. This is thought to be one of the main evolutionary drivers for the movement of genes from the mtDNA into the nuclear genome and is certainly reflected in a much higher mutation rate than the DNA located in the nucleus.

1.1.4 Origins and Endosymbiosis

Phylogenetic studies have also found evidence that all extant mitochondria branch from a single endosymbiotic event, suggesting that it only happened once, or we only have evidence of a single event in the genetic record available today. All branches of this tree are rooted at the same event [12].

The host cell offered the proto-mitochondrion shelter in return for a more complete metabolism of nutrients, wringing out more energy and creating an huge advantage for the mitochondriated cell. Whilst oxidative respiration evolved as oxygen concentrations rose in the atmosphere, there is some debate as to the truth of the assumption that mitochondrial precursors merely sheltered in the cell to survive [12].

Once the mitochondriated cell lines dominated, refining of the relationship was needed. Avoidance of the inevitable mutational meltdown, predicted by Muller, through incrementally acquired mutations was required. In order to survive and then thrive, mitochondrial genes were acquired by the nucleus to be stored on the chromosomes, protected from the mutagenic conditions in the mitochondrial matrix and to be refreshed by access to selection pressures from sexual inheritance and recombination [79].

The movement of genes integral to the mitochondrial functions incrementally tied the pair together and with the increase dependence of the partners on each other [34].

1.1.5 Mitochondrial Genome

In 1981, Anderson et al. [11] published the mitochondrial genome's annotated sequence with predictions of the genome's contents. Some 18 years later, with the improvement of techniques and accuracy, Andrews et al. found and corrected the original sequence in ten places, called the rCRS (revised Cambridge reference sequence) [13].

The mtDNA is certainly insubstantial compared to the DNA held in the nucleus but a comparison of length leaves us mislead to the different densities of information (gene encoding to non-gene-encoding) and to the different average values of that information.

	Nuclear DNA	Mitochondrial DNA (rCRS)
Genome Length (bp)	approximately 3 billion	16,569
DNA encoding genes (%)	approximately 1%	92.7%
Protein-coding Genes	20,000 - 25,000	13
Non-protein-coding Genes	17,000 - 20,000	24

Table 1.1: Basic comparison of nuclear and mitochondrial genomes

Gonçalves et al. talks about the mtDNA encoding for some 13 components of oxidative phosphorylation complexes, from a total of approximately 110 units [43]. Shokolenko et al. include a detailed map in their publication which gives a good sense of how closely packed the intronless genes are on the mtDNA [91].

A template genome prompted the quest to explore the "magic circle" where the focus of mitochondrial research fell on its genome. Chinnery and Turnbull [26] reflects that we should see more emphasis on the interactions between the two genomes, nuclear-mitochondrial communication or "mitonuclear ecology" according to Hill [48].

Regarding the mitochondrial genome, Van Oven observe the uses it has for several fields plotting the history of life, species or even individuals [98]. Whilst variations of arrangement and storage of the mtDNA do exist, Ladoukakis et al. [65] tell us the normal format for animal mtDNA is "a circular, compact molecule about 17 KB with little variation in size, containing 13 protein-coding genes, 2 rRNA and 22 tRNA genes". This rule is nearly unanimous among life with left-right symmetry. The genes remaining on the mtDNA are without introns and hold translation differences from the rules in the nucleus [65]. Retention of these genes points at the language differences between the nucleus and the mitochondria, and, most intriguingly, different from their proto-mitochondrial ancestors [44].

The natural extrapolation of this process would be to see many species with mitochondria devoid of any genetic material. However, only one, Amoebophrya ceratii; a dinoflagellate, does not possess a separate genome [56]. This suggests the existence of one or more evolutionary pressures for the maintenance of a partial genome in mitochondria. Researchers have explored the remaining sequence for clues held within the attenuated mitochondrial genomes.

A second set of genes retained on the human mtDNA enable the translation of the different coding in mitochondria. Anderson et al. [11] explains the differences The UGA tRNA codes for tryptophan and does not initiate chain termination, although AGA and AGG do trigger chain termination rather than chain elongation with an arginine residue. The AUA tRNA carries methionine not isoleucine. These rule exceptions are the tRNAs encoded on the mtDNA.

Escaping mutational meltdown pressured the movement of mitochondrial genes to shelter in the host nucleus, with mitochondria having retained only an estimated 1% with the remainder

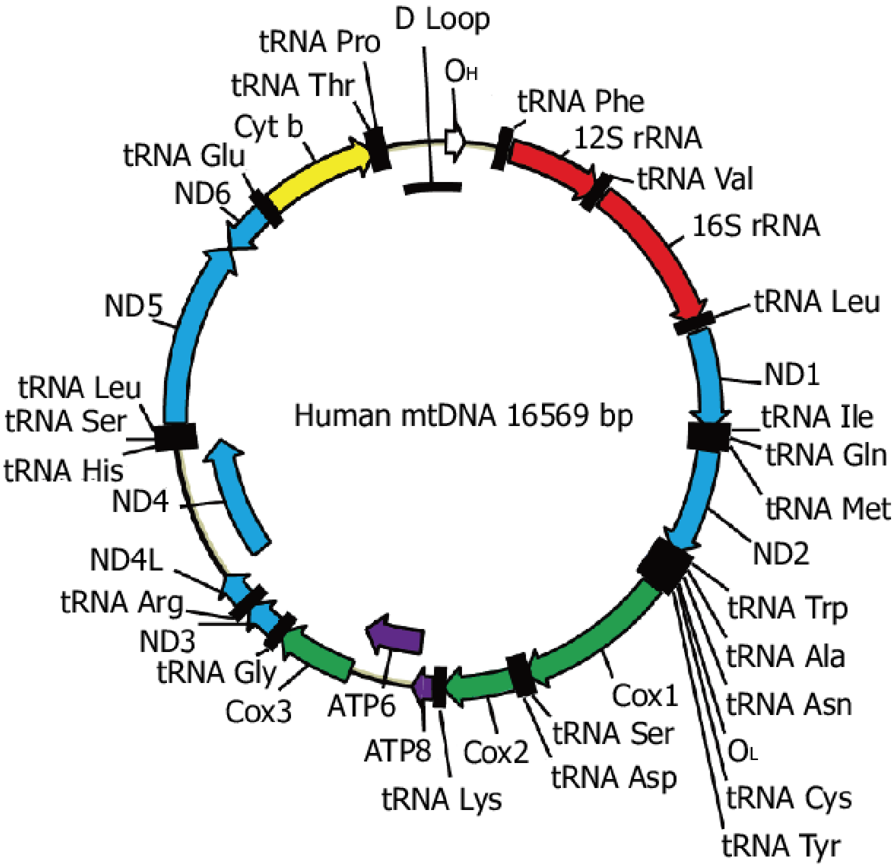


Figure 1.2: Map of the mtDNA, from Shokolenko et al. [91] The arrow direction signifies the direction of transcription of each gene, falling either on the heavy strand or light strand of the mtDNA. Blue arrows are genes encoding subunits of NADH dehydrogenase (complex I). Yellow genes encode Cytochrome B (complex III). Green encodes for subunits of the Cytochrome Oxidase (complex IV) and purple for subunits of ATPase (complex V). Red genes are the rRNA regions and black represents the tRNA encoding regions.

stored in the nucleus, and a general reduction of DNA in the mtDNA. [62]. Anderson et al. were the first to publish their mtDNA sequence and comment on the lack of non-coding DNA even padding between neighbouring genes [11].

1.1.6 Mitochondrial Genetics

The presence of a genetic material sequestered in the mitochondrion was reported as "intramitochondrial fibres with DNA characteristics" in 1963 by Nass and Nass [81]. Making the link between phenotypes and the genetic material in the mitochondria took the observation that the petite yeast strains all suffered from significant alterations in mtDNA base composition [36].

However, these yeast strains contained mitochondria missing large genome segments and were too blunt an instrument for examining finer details of the mtDNA. Soon after, searches for new lines found some which offered more refined deletions. Rather than the loss of huge portions of the genome, these mutants were missing single components, allowing their classification as intrinsic to the mitochondrial genome not translated from mRNA imported from the cytoplasm [36].

Inspired by chasing the inheritance patterns in the mitochondria causing phenotypes, the mid-1970s marked the point when a physical map of restriction enzyme sites could be built [36]. Just a few years later, in April of 1981, Andersson et al., published the first human mitochondrial genome sequence, alongside considerable exploration of what the sequence encoded and a great deal of theoretical work on the consequences of what they had found. Along with the historical review by Ernster and Schatz published the same year, these papers reflect well-observed comment both of which have withstood time remarkably [11, 36].

Many writers have noted how little mtDNA is wasted on non-coding regions, with the vast majority being transcribed and translated. However, there is a region called the d-loop. Anderson et al. tell us they have found no open reading frames of a detectable size in this 1.1kb region. They go on to say that only very significant RNA splicing could offer the small possibility that the d-loop region encodes for protein. [11]. This region, named the displacement region by Doda et al., contains a third piece of DNA, forms complex secondary structures and is much more variable in sequence than the remainder of the genome [33]. These are hallmarks of a control region, which is pointed out by Anderson et al. [11], telling us that the mtDNA transcription process is controlled entirely from a small number of promoters, making up part of the non-coding, control region in the d-loop.

The d-loop region is much more variable and non-coding but shows associations with conditions such as polycystic ovarian syndrome (PCOS). Chinese participants with a certain set of variants (one SNP and two short deletions) in their d-loop region were found to be at reduced risk of PCOS [31].

The usefulness of the mtDNA sequence is explained by Templeton et al. telling us that, largely because of the increased copy number, useful mtDNA sequence can be gathered from even very compromised samples. Identifying mtDNA signatures, lost from the nuclear DNA, can still be found. [97].

Mitochondrial DNA holds true to many of the basic rules of genetics, which are explored in nuclear DNA systems, such as using the same bases and base pairings. However differences in the storage and inheritance of mtDNA have huge implications on many significant factors, both internal to the mitochondrion and in the relationship of mitochondria and host.

The common view that the mtDNA is non-recombining, maternally inherited, and has high mutation rate is reported by both Ladoukakis et al, and Burr, Pezet and Chinnery [65, 20]. For anyone versed in Mendelian genetics, several, significant consequences of these rules should unfold. The expectations of the mtDNA, based on the nuclear rules of inheritance, may fail to capture what we see. Mitochondrially speaking, offspring will be near copies of female parents, with no mixing of her parents and the addition of a higher-than-expected number of novel mutations. On deeper examination, these rules will have much wider and long term consequences, which I shall outline below.

Maternal Transmission

Despite a large number of mitochondria in the sperm, the mitochondria for each offspring are of the same type as the maternal parent. They reside in the ovum and propagate as the offspring grows so that mitochondria proliferate largely according to the metabolic needs of the tissue. This mtDNA copy number varies from 100 per cell to a hundred times that [20].

Ensuring this uni-parental inheritance are a wide variety of mechanisms. Ladoukakis at al offer us a set of examples; mammalian sperm mitochondria are tagged with ubiquitin for later destruction; the mitochondria of Drosophila do not make it into mature sperm; Japanese rice fish and *Caenorhabitis elegans* target and destroy the male's mitochondria. Having such a range of techniques and a "ubiquity of the transmission mode among organisms" suggests that a) uni-parental inheritance of mitochondria is vital, and b) pressures exists to cheat the system have been resisted with new mechanisms [65].

A need to have a single mtDNA type throughout the organism, called homoplasmy, appears to be vital for both the organism and its offspring. Methods for ensuring not just uniparental inheritance, but also inheritance of a purified, homoplasmic population, act through pre- and post-fertilization bottlenecks. A reduction of mitochondrial number occurs during oogenesis and then a second period of attrition occurs just after fertilisation. During this period, cell division proceeds at a very high rate but mitochondrial replication is suppressed [58].

These seemingly disadvantageous actions to severely limit the genetic diversity are common and enforced to avoid heteroplasmy. "Heteroplasmy has been involved in mitochondrial diseases" [65]. Whilst this is a consistent observation in severe mitochondrial disease, the authors do say that heteroplasmy is hard to detect, especially at low levels and when mitochondrial type may vary widely across tissue type due to development. It is also easy to imagine that homoplasmic embryos with defective mitochondria failing to thrive prenatally, acting like a homozygous lethal variant on the nuclear genome.

Lack of Recombination

A second rule of mtDNA inheritance also aids the use of this genome as a phylogeny tracker; lack of recombination. Nuclear chromosomes form pairs to mix their DNA during meiosis to create gametes. Like-for-like regions are swapped, separating variation formerly on the same chromosome, perhaps linked for many generations. Recombination muddies the water when looking at nuclear phylogenetics as the physical links between genetic regions are often broken. For mtDNA, very little evidence of recombination exists leaving clean inheritance routes to be plotted along phylogenies and linked genetic regions remaining linked through generations.

Until recent the detection of recombination in human mtDNA was beyond the limits of technology because of the rarity of the recombined molecule. The evidence for recombination would rely on having a sequencing technology with a lower error rate than the concentration of recombined genomes, a point we have only recently reached [65].

However, Hagstrom et al. looked for evidence of recombination in mice mitochondria and found enough evidence to rule out that possibility and that animal mtDNA could be thought of non-recombining. However, despite the lack of genetic evidence, mitochondria retain the enzymes to enable the process leaving the door open to the possibility, if not finding direct evidence [46].

Further circumstantial evidence is explored by Ladoukakis et al. who suggest that recombination would explain why the mtDNA is not full of dangerous mutations randomly accumulated, a process embodied in Muller's ratchet. The mtDNA should have collapsed long ago but still goes strong after approximately 2 billion years of bombardment. Ladoukakis et al. attribute this to recombination [79, 65].

A non-recombining genome also builds a level of linkage disequilibrium hiding loci from the reach of natural selection [40].

Heteroplasmy

When mitochondria fail to keep up with the demands of the cell, we see dysfunction. Wallace et al. list and order the organs and tissues most at risk; the neurons of the central nervous system and the eyes being most at risk, with cardiac muscle, skeletal muscles, kidney tissue, and tissues of the endocrine system and hepatic system following [103]. Ladoukakis et al. observes that no defining pattern of mitochondrial disease exists, which make their diagnosis challenging, but heteroplasmy is seen in many cases, and its extent correlates with dysfunction [65].

Whilst Payne et al. tell us that a low level of heteroplasmy is found in almost all humans, Chinnery comments that "patients carrying potentially deleterious heteroplasmic mtDNA mutations, the severity of disease symptom tends to correlate with the level of heteroplasmy, and if a certain biochemical threshold is released reached the individual will develop pathogenic phenotypes" [84, 25, 20].

Several routes lead to a heteroplasmic mix of mtDNA and we see that both the cell and the mitochondria attempt to suppress these. However, Ladoukakis et al. tell us that, with a lack of control of the replication of the mtDNA, a deficient mtDNA molecule can outcompete a complete mtDNA in the mitochondria. The deficiency of the mtDNA molecule is masked by the complete molecule's performance until a mitochondrion or cell is homoplasmic for the defective version. Even without a relative replication advantage, defective molecules are a hazard with the risk of a mitochondrial division or cell division leaving the defective molecule responsible for energy production [65].

Whilst a normal discussion regarding genetic diversity celebrates variety as a sign of health and survival, this is not the aim for mitochondria. Every cell needs a functioning set of mitochondria and heteroplasmy masks mtDNA containing deleterious mutations. Once we consider pairs of defective molecules, masking each others' faults, the likelihood of deficient cells multiplies [65]. As long as a cell contains both molecules, function is maintained. However, at each cell division and mitochondrial fission a cell line may result with significant problems. We see this manifested in the behaviours of mitochondrial diseases caused by mtDNA. Progression, sporadic, very varied penetrance and chimeric individuals all point to cell inheritance of new or heteroplasmy-masked deleterious mutations through a type of cell-line genetic drift.

Detecting heteroplasmy is a significant challenge. Sampling may miss affected tissues, detecting low concentrations of mutants in the mtDNA pool is a challenge due to the inherent noise involved in traditional Sanger sequencing. Sequencing-related challenges look to be largely overcome as the next generation sequencing can not only be trusted to pick up the rare mtDNA differences, but can also be used to exclude heteroplasmy if no evidence is seen in sequencing reads [65].

A significant route to mtDNA purification and heteroplasmy loss is through a period of attrition orchestrated very early in development. The newly fertilised ovum has been seen to hold hundreds of thousands of copies of the mtDNA. Whilst corroborated by PCR-based experiments and computer simulations, the same magnitude of change has not been replicated in the same model system so remains a theory. [20].

There are other opportunities for selection; population, individual, cellular and subcellular. At all of these levels, theoretical replication selectivity could be put into practice. Burr, Pezet and Chinnery admit that the actions of the purifying routes miss problems evidenced by the high level of heteroplasmies in common population. Burr et al. continue by pointing out that many of the signature markers of haplogroup membership are also pathogenic variants and avoiding selection pressures to remain common [20].

High Mutation Rate

Compared to the nuclear DNA, the increased mutation rate of mtDNA can be observed when looking at the large numbers of samples in libraries of mtDNA sequence, such as GenBank. Johnson and Johnson tell us of an increase in error rate of several orders of magnitude compared to the nucleus [57]. Lagouge and Larsson agree and mention a cause; the reactive oxygen species abundant in the matrix where the mtDNA is stored [66].

Burr et al. also cite two further reasons; mtDNA is continuously replicated regardless of their point in the cell cycle, and the mtDNA has fewer route to the repair of all types of damage compared to the nucleus [20].

As in the nuclear genome, we see the rate varies across the molecule. The 1.1kb-long d-loop mutates at about ten times the rate of the coding region [50, 82]. The story is further complicated with wide raging differences between narrow hot spots of multiple mutations and stable regions where no individuals show changes. The hot spots are of note from a mapping perspective, as Van Oven and Keyser tell us. Mutation which reverts mtDNA to atavistic sequence, appearing as if no change has occurred, complicate the building of a family tree [100].

The variation seen in the mtDNA can be put to work on a population scale to map the development of the species, Lawless et al. point out that mutation is likely to have a pathological effect because most of the mtDNA is coding region, and there are no introns. They continue by pointing out that the mutations accumulated through an individual's life hamper tissue performance and cause some of the effects of ageing [67].

Implications

Whilst Ladoukakis et al. remind us "No "rule" about mtDNA remains unbroken"[65], the differences we see from the nuclear DNA and the mtDNA's very existence, creating "mitonuclear ecology", have significant and wide-ranging implications [48].

As mitochondria became domesticated, a process of co-evolution began. Any genes lost from the mitochondrial genome must have been accounted for by a mirrored changed in the nuclear genome. Any change in one genome which is accommodated by the other, obliges their dual inheritance. The two genomes have been "co-evolving synergistically" and inheriting co-adapted genomes is the only route to a fully functioning individual [65].

For full exploration of theoretical consequences of two, differently inherited genomes, Hill is hard to better. He refers to "mt genes" as those genes required by the mitochondrion which reside on the mtDNA, as opposed to the "N-mt gene" which are vital for the mitochondrion by are now residing in the nuclear genome.

A review paper, entitled Mitonuclear Ecology and published in 2015, hammers home the importance of inheriting co-adapted genome pairs; "the interdependencies of genes made it essential that sets of genes be inherited together. A growing body of research, however, suggests that the most significant co-adapted gene complexes for eukaryotes are N-mt and mt genes" [48].

The driver of this co-specificity is that the level of intimacy that the mitochondria and cell have, one which extends to the genes and their products and actions. ATP production must be responsive, reliable, and well-matched, and failure on any of these counts will be severely punished by natural selection. Higher levels of reactive oxygen species and power cuts in stressed tissues will result in dysfunction. Any individual inheriting a not perfectly compatible genome pair risks substantial problems particularly in muscle and neuronal tissue, and in male fertility and ageing, according to Ladoukakis at al. The paper continues by pointing out that mitonuclear interactions are mostly unstudied but very likely to be a rich source of causes of dysfunction in, particularly, stressed, aged, or very metabolically active cells and tissues. They note this should be investigated especially as we explore cloning and three-parent embryos [65].

The serious implications of incompatibility are thought to be behind the checking systems, such as the genetic bottlenecks imposed during oogenesis and conception. Burr, Pezet and Chinnery write of the importance of the bottleneck theory in germ line cells.[20]. We hear that this "antagonism", prompts pressure from the nucleus to maintain the strictly matrilineal inheritance, kidnapping more genes from the mtDNA to give the nucleus more independence. This is counteracted by the mitochondrion's pressure to increases its own essential roles for the cell by performing vital roles [65].

Hill's reviews include many journeys into the consequences of co-adaptation and paths the evolution of which may be explained as routes to avoiding or weeding out incompatible pairings. Much is made of gender asymmetry. This is well explained by Frank and Hurst [38] who tell us that, for mtDNA, there is an asymmetry about the selection pressures. They tell us that selection pressures on mitonuclear incompatibilities will be much greater in females. Therefore, genes which are deleterious in males but neutral or advantageous in females will remain sheltered from selection. This theory, the Mother's Curse, explains why we get differences between how males and females age including some very specific risks like that of heart disease, although causal evidence is not yet abundant. The effects of gendered selection asymmetry suggest the search for the effects of mito-nuclear incompatibilities begins with conditions with a increased severity in men [38].

Heart disease shows significantly different outcomes for men and women. Inouye et al. used UKBiobank participants and established nuclear genetic risk scores for each person. Combined with lifestyle risk factors, such as obesity or smoking status, subgroups based on risk could be made. At each level of genetic risk, males with the fewest lifestyle risk factors had the same level of cardiac incidents as the females with the highest number of lifestyle factors [21, 54].

The links from sexually dimorphic health outcomes, such as heart disease, are still largely theoretical. However, the likelihood of issues in communication between the nucleus and its mitochondria is high. Eisenberg-Bord discusses the need of mitochondria to respond to their cell's current needs but also admit that mitochondria depend on a great deal from the nucleus to allow that response. This paper insists that molecules and signals must pass in both directions to constantly regulate all processes to match the needs of the other party [34]. For potential methods of message transduction, Ladoukais reminds us of the many short-chain proteins which are encoded on the mtDNA but are not directly linked to oxidative phosphorylation [65].

Haskett sums up this complex relationship writing about co-adaptation and evolution, and the rapid to-and-fro of information, hinting at the both the intimacy and importance of the associations between each cell and their mitochondria [47]. Once considering the message delays, misunderstandings, over-compensations which would be prompted by host-mitochondrial pairing which were not co-adapted or mutations in either genes, we have a recipe for health problems in cells lacking perfect, rapid, clear communication.

1.1.7 Mitochondrial Pathologies

With so much potential for faulty mtDNA, a range of conditions are caused by genetic variations of the DNA held by mitochondria and they illustrate the functions or inheritance patterns of the mitochondria. The affected tissues are largely tissues with a high metabolic demand, either for short periods or continuously.

Leigh's syndrome can be inherited via the nuclear- or the mitochondrial genome. It is caused by a variation in one of the seventy-five subunits of complexes I-IV of the oxidative phosphorylation chain. Leigh's syndrome is usually noticed before the first birthday of the child carrying the variation as the child fails to maintain normal growth or neurological development after approximately 6 months of being largely asymptomatic. Decline is sporadic and may see a little development between crises, but is life-limiting [32].

NARP (Neurogenic weakening, ataxia and retinitis pigmentosa) is also linked to a complex of the OXPHOS chain. Mutations affecting ATPase or complex V result in severe symptoms the return of protons pumped across the inner membrane is prevented, the pathway is overloaded and oxidative stress poisons the mitochondria [32].

Sexual dimorphism is seen in mtDNA pathologies. Leber's hereditary optic neuropathy (LHON) is caused by one of three mutations on the mtDNA and only seen in a subset of carriers, with men being more likely to become symptomatic. LHON tends to be diagnosed in a sufferer's 20s or 30s, with sight loss in one then the other eye [32].

Mitochondrial conditions are highly variable; carrying the mutation is not a guarantee of ill health. One mtDNA mutation is the cause of two conditions: MELAS (mitochondrial encephalomyopathy, lactic acidosis and stroke-like episodes) and MIDD (Maternally Inherited Diabetes and Deafness). Asymptomatic carriers of this single mutation are common and the pattern of disease in sufferers is decided by the level and layout of heteroplasmy, such as rare cases of cardiomyopathy [32].

MERRF (myoclonic epilepsy with ragged red fibres) sufferers present with ataxia (unsteadiness) common with the large muscular weakness frequent with mtDNA pathologies, seizures and memory loss. Interestingly, this mutation also encourages the laying down of adipose tissue on the back of the neck and upper torso [32, 41].

DiMauro and Schon build a table showing the common human conditions caused by mitochondrial genetic variants to highlight the common patterns and symptoms amongst these conditions [32].

1.2 Data Analysis Methods

1.2.1 Introduction

Mendelian traits are inherited cleanly and in a statistically predictable manner. With the advent of genome maps, we know the locus that contains the genotypic information predicting the phenotype of the individual. However, even as these traits and conditions were being reported, it was seen that these were special cases at the far end of a continuum. Many, many more traits fell on a scale from extremely heritable through to entirely environmental.

Looking into the genome for explanations of human conditions has become easier since the human genome project published a draft genome to begin to map out gene regions and other functional elements. This draft was expected to solve mysteries and mark the beginning of the end of conditions with a heritability. Unfortunately, geneticists were faced with picture even more complex than they imagined [101].

However, several branches of analysis were enabled by having an overview of the human genome. Genome-wide association (GWA) studies rose in popularity as both the data gathering became cheaper and easier, and the investment in computing resources grew [73]. Since the first success in 2005 reported associations for age-related macular degeneration, GWA studies have found nuclear genome regions linked to tens of thousands of traits, revolutionising the field [95].

An immediate drive towards increasing GWAS participant numbers began to allow more subtle relationships to be discovered. This is, in part down, to the conditions being investigated which have complex associations with reams of genes. The common disease-common variant hypothesis suggests that the genetic drivers of the most common conditions are a plethora of common variants all with a small effect on the likelihood or severity of the condition [90].

In addition, the statistical method involves multiple testing. Using a 95% statistical test on millions of loci means tens of thousands of false positives just because researchers are checking so many loci. As the numbers included in the studies increase, more, subtle associations are found. Very high level of statistical significance are required to exclude chance, compensation for testing millions of loci for each GWA study. Despite this, Tam et al. tell us that, in 2019, there does not appear to be any levelling off of discovery rates for any traits [95].

In mid December 2020, the GWAS Catalogue list 4809 of peer-reviewed studies which have found 227,262 statistically significant regions in the human genome which explain levels of trait heredity not previously explained through simple Mendelian genetics [73]. Study sample numbers are growing to enable complex relationships or rare variations to be characterised. More is being added to this library to refine models and improve predictions.

As GWA studies gained data and momentum, associations were found all through the nuclear genome but methods used were not optimal for the genome in the mitochondria. A look at the GWAS catalogue's diagram of their data reveals that mitochondrial data is not even represented. There remain just two associations on the mtDNA reported in the catalogue at time of writing.

1.2.2 Imputation

Marchini and Howie wrote a review article on the process in 2010 defining imputation as "the term used to describe the process of predicting or imputing genotypes that are not directly assayed in a sample of individuals". Tam et al. also discuss the method explaining that imputation effectively increases the number of variants which can be association tested beyond the variants for which there is conclusive genotyping data [95]. This is achieved through the use of an extensive panel of reference haplotypes with the same population profiles as the samples to be imputed, in fact, this panel may be a subsection of the study itself to ensure population parity [102, 75, 19].

Imputation takes known data points about the sequence of interest and employs information about that same region in other individuals, long range effects and recombination rates to build a series of probabilities for the unknown bases [75, 45, 51].

Methods of imputation of genotyping are have advanced and enable analyses into the genetic patterns across the nuclear genome. This can be achieved with enough accuracy to perform genome-wide association studies, where phenotypes can be associated with versions of regions of the genome. A comparison of the variants possessed by carriers of the phenotype compared to non-carrier is made, revealing protective or causative variants [102].

Genetic sequence can be viewed as a one-dimensional chain of nodes where each node may be in one of four states, A, C, T or G. Further than this, it is not created afresh for each individual but inherited, changes are step-wise allowing clades to be built and DNA carries information so much of the sequence is far from random. This adds up to a great deal of predictive power.

Using the known data points a map of likelihoods for interpolation can be made; something which, in English, would sound like;" `Given there is a Guanine at locus 1 and a Thymine at locus 3, our prediction at locus 2 looks like a 92% likelihood of an Adenine`".

Imputation ensures that the probability estimates reflect the distribution suggested by the known genotypes. This is important when using the imputed data for further analysis, such as for GWA studies [74]. Using regions of common sequence differences, the likelihood of other genetic differences in the individual can be estimated, with high accuracy [71].

Methods which impute genotypes rely on the known genotypes being di-allelic assessments, where two SNP types are searched for by the microarray. The methods also assume that the data is for autosomal DNA, with two copies of each chromosomal region. Code development has improved to include exploration of the sex chromosomes in males. As males have only one X chromosome, they are referred to as hemizygous.

Hemizygosity is also applicable to the mtDNA, as it too has no paired copy. Whilst the best method would be tailor the processing to differentiate between males and females, the imputation methods currently available attempt to overcome the issue by making the males homozygous at every position on their X chromosomes [74].

Imputation of the nuclear genome is built around the inheritance patterns in the DNA, and what can be extrapolated from the genotyping data. The inheritance patterns of mtDNA differs significantly from the systems seen in the nuclear genome with; (a) maternal inheritance, (b) no (visible level) of recombination, and (c) an elevated mutation rate. These differences render imputation acceptable, but not efficient, for use on mtDNA.

XXX Imputation of mtDNA can be approximated using the imputation methods built for nuclear DNA, and is used as such, however, the idea of linkage breaks

1.2.3 Mitochondrial DNA Data

Mitochondrial DNA has long been approached differently. The differences we see with this piece of DNA have lead researchers to use haplogroup structures to search for associations.

Haplogroups are named groups of mtDNA sequences which share common variations. As the mtDNA does not recombine, and variants are inherited together, it can be assumed that the level of genetic similarity between two sequences is directly correlated to their level of relatedness.

The sympathetic inheritance patterns of the mtDNA allow the phylogenetic tree to be traces back to human roots and a single African origin [65]. Organising the information emerging from comparisons of mtDNA lead to two tools; haplogroups and a tree rooted at a theoretical, female, shared, human ancestor, Mitochondrial Eve [20].

At time of writing, there are 4947 haplogroups. These groups can be gathered onto the nodes of a much simpler tree, see figure 1.3, which forms a clear branching structure.

Since Mitochondrial Eve, the mtDNA has been accumulating small variations throughout the mtDNA. These may not be physically close, as seen in the associated variants of the nuclear GWA studies, but travel together undivided by recombination events. The sequence of changes is used to follow the mtDNA's descent down branches [100].

An obvious way of categorising groupings developed in haplogroups; discrete, defined groups of common variations, which, when found together, decide the membership of the haplogroup. The history of the human species is far more complex that a cascading divergence from this single mother and also the variants, which began as heteroplasmic, private mutations before homoplasmy and fixation in the group [20].

Other extinct branches certainly existed long ago although failed to leave descendants and traces in today's collections of genomes. However, the enduring image of a matrilineal line aids the understanding on the motivations for the project in this thesis as we have yet to find a human mtDNA which cannot be placed within this family tree rooted at Eve.

Building a tree of mtDNA sequences allows both the grouping of similar sequences and the differentiation of sequences on few variations. The sharing of variants suggests kinship and a common ancestry, although, in the rapidly changing mitochondrial genome, this assumption cannot always be made. Naming and defining branches in the mtDNA phylogeny was a natural move and categorises the sequences we see in Africa and through the diaspora which migrated across the planet.

Some defining variants are thought to hold advantages or adaptations beneficial for life outside of Africa, with the emergence and blooming of mutations appearing with large migrations [20]. Whilst the common sense of this appeals, the idea must be approached with caution. The changes in variant frequencies may be more to do with the effects of a narrow genetic bottleneck than decisively beneficial mtDNA variants enabling human survival in new climates [28, 86, 20, 35].

Once researchers became aware that the mtDNA sequences they saw formed a phylogenetic tree, sequence similarities were used to group the sequences. Naming these haplogroups involved finer and finer precision as differences in sequence became visible; through restriction enzyme analysis and sequencing [23].

The gain of information about the relationships of the groups through assigning structured, defined haplogroups is of benefit to exploring accuracy. This, in turn, enables forensic profiling of material with the most famous case being King's investigation into human remains confirmed to be those of Richard III [59], where a continuous line of matrilineal inheritance could be traced from Richard III to present descendants via his sister. The descendants and the potentially royal remains shared a common set of rare variants chosen to be a decisive as possible.

Whilst this special case gripping for the general public, there are benefits of mitochondrial genetics for both population and medical usage according to Rock [85]. The linking of phenotypes to certain macrobranches, branches and groups in examples such as Gomez-Duran is similar, in effect, to the GWAS approach using linkage disequilibrium. Gomez-Duran reflects on the links between haplogroups in the macrobranch J and their increased body temperature. Individuals

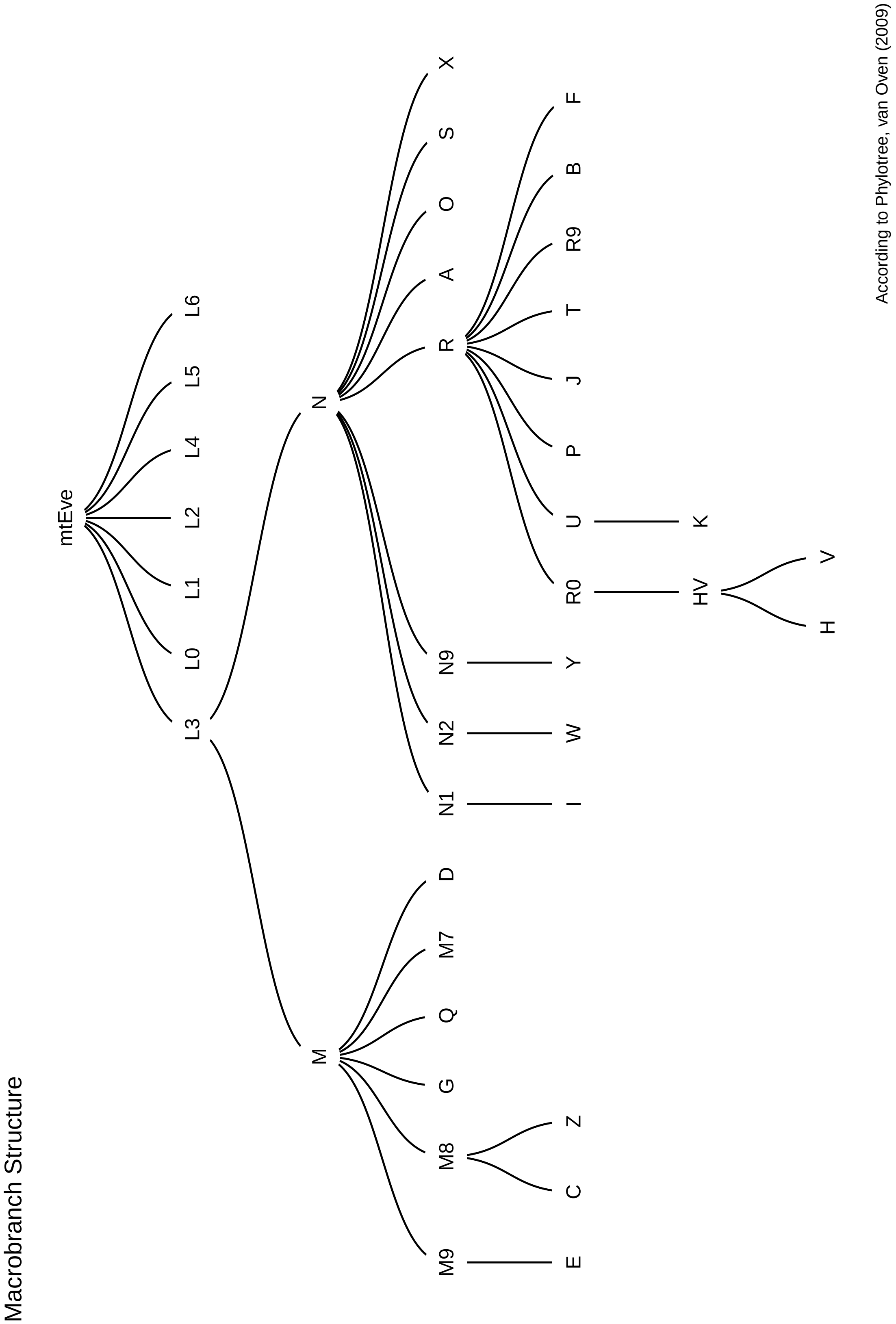


Figure 1.3: Macrobranch structure in the 4947 haplogroups in the phylogenetic tree of the human mtDNA, according to [100]. This merely represents relationships between the macrobranches. The number of genetic differences between branches varies greatly and is not captured in this diagram.

within these groups share a decrease in their metabolic efficiency and may be related to their propensity to developing a mitochondrially inherited form of sight-loss through neuropathy in the optic nerve [42].

1.2.4 UKBiobank

The UKBiobank project was focussed on scale from its start in 2007. An impressive 438,427 participants were recruited, phenotyped and genotyped for over 800,000 markers across the genome. There is a huge body of data on the health and lifestyle of the participants, many of whom return for follow-up and have agreed to link their medical records [22].

Whilst the cost of full genomic sequence has dropped significantly in the years since the human genome project, it still remains beyond the reach of projects looking for elusive, complex genetic traits [4]. In 2007, when the UKBiobank began recruitment, the price per genome was $10,000,000 and has dropped to $1,000 in the 13 years since. This reduction in price still makes the genome sequencing costs for the UKBiobank's participants nearly half a billion dollars at the current rate.

Large numbers were needed to have the statistical power required for accurate GWAS. When tests are binary for a single locus, in the case of Mendelian traits, numbers of samples can be small and remain conclusive. However, once traits are merely indicative of an increased likelihood of a phenotype or are required in combination with other loci, large-scale genetic characterisation projects are required.

UKBiobank have genotyped their participants in 265 positions in the mitochondrial genome; at a rate nearly a hundred times higher than the nuclear genome, with 0.03% of the positions for 0.0005% of the DNA sequence. Imputation of the mtDNA should be better, despite the increased mutation rate of the mitochondrial genome, although association on the mtDNA are not registering on places such as the GWAS catalogues [73].

1.2.5 Competing Programs to Process MtDNA

Haplogrep2 is commonly used to automate the classification of mtDNA sequences to haplogroups, according to their variants. Matching patterns of this type is very amenable to computational methods to improve accuracy and speed [104].

Although mtDNA is short and does not lose data relationships through the mixing action of recombination, Haplogrep2 must contend with a very high and variable mutation rate, with hot spots so prone to change that their information is excluded. Because we have data on a continuum from rare and decisive to common or degenerate, Haplogrep2 uses weightings to ensure it balances the data values. Generally, the weights are inversely related to their number of appearances in Phylotree, [104, 100]

A major drawback to the use of Haplogrep2 is simply that it is designed to accommodate small mtDNA batches with non-systematic missingness, such as badly decomposed DNA. Information about a mtDNA sequence is submitted, again as differences from the rCRS, and information about regions which are missing. From this a best guess at a haplogroup is produced with lists of variants which agree with the assignment and those which disagree.

Haplogrep2 does accommodate the frequency of the variants, weighting their reliability and decisiveness with their commonness in the entire haplogroup tree. The haplogroups are also varied in their frequencies, which Haplogrep2 does not account for, as this is not expected to be of relevance for the situations of its implementation [104].

Gonçalves et al. make headway taking SNP microarray data about their experimental participants, imputing more variants and then using Haplogrep2 to find haplogroup macrobranches. This paper specifically explores phenotypes around schizophrenia, but the approach is generalisable. From the macrobranch, frequency testing can show which macrobranches are over- or under-represented in phenotype groups. Whilst this approach does allow focussing to the macrobranch level, the variation within the macrobranches cannot be resolved [43].

Whilst some variations characteristic of a branch or haplogroup cause a measurable phenotype, many more simply pepper the mtDNA and enable haplogrouping assignment. These variants were gathered together in a resource called Phylotree, which has become the central tree to which new mtDNA sequences are placed. Rock notes that the existence of Phylotree, and its universal uptake as a standard, has enabled software developments which improved haplogrouping processes [85]

This development was aided by having an accepted method of information coding, effectively relating all mtDNA sequences to a reference sequence, called the rCRS or revised Cambridge Reference Sequence. The coding used to efficiently record mtDNA by deciding on a mapping system with base pairs from 1 to 16,569, and referring only to differences from that reference sequence. This has enabled the very concise sample information which can be compared easily [100].

Having more and more data on which to draw and a convention as to how to code the information has allowed strides forward in understanding of how the mtDNA behaves, but is not completed work. Whilst undoubtedly allowing the improvement of understanding and detail available on the mtDNA, there is a problem with the haplogroup assignments, particularly when the sequence is incomplete. Furthermore, the library of data used to classify sequences is growing but its limitations will limit accuracy [100, 85].

These haplogrouping tools, underpinned by growing number of available mtDNA sequences, still have their limits. As they assign haplogroups using defining mutations littered over the MT genome, missing data is a huge challenge. Rock found that missing a single variant resulted in an inaccurate haplogroup assignment, generally further up up the tree. Rock warns of relying on the use of partial data as this may have caused incorrect haplogroup assignment [85].

Here lies the problem with Haplogrep2, and other tools such as Phy-mer, Mito-tool and MitoIMP; missing data and partial sequence can be accommodated but are expected, through DNA degradation, a random process. The samples' sequences is expected to be partially lost, not genotyped.

Fan et al. admits that with degraded DNA distant haplogroups can become equally weighted best guesses using Mito-tool. Without enough data, the samples must be quality checked and hand assigned[37]. Manual inspection for half a million UKBiobank participants is not ideal.

MitoIMP, developed by Ishiya, Mizuno, Wang and Ueda offers another route to mtDNA imputation of low coverage DNA. Again, however, the focus is on problematic, degraded mtDNA, not the regular structure emerging from a SNP microarray. The use of their code is discussed as being applicable to this type of data, but the test panel used have validated their code at 10-90%, much higher coverage than the 1.6% covered by UKBiobank's arrays. The accuracy level they quote regarding "haplogroups", reflects assignments to the correct "macro-haplogroup linages" and [55].

Ishiya et al. do mention the challenge of a higher level of diversity in the sub-Saharan Africa lineages, namely L macrobranches. This is an effect expected to be universal as the complexity seen is made harder to study by a lower coverage by sample number [55].

In their test panel, the method described by Ishiya et al, does bring additional value to their samples, with a high level of accuracy, which would yield more signal to noise in a GWA study. However, the use of this framework in microarray sample data seems out of reach.

The implications to an attempt to assign haplogroups to the 450,000 participants just on the genotyping of a few hundred base pairs is clear; using current techniques will result in biases for the individuals, at the population level, and for any attempt to extract variant frequencies for GWA studies.

The use of biased data for a mtDNA GWA study would mask any signals. GWA studies look for a halo of over-represented variants around hits and, whilst this would look different in the mtDNA with a lack of recombination, the biases introduced through best-guessing hundreds of samples in particular directions would hide these halos.

1.3 Relevance to this Project

The mitochondrion makes multicellular life possible through ATP production and is implicated in pathways such as apoptosis. Mito-nuclear communication must be clear, consise and timely. Any weak links are punished severely by pathologies in tissues starved of energy when stressed, or energy gluts wasted. However, the mitochondria continue to carry a portion of their vital genes on a genome exposed to mutagenic conditions but sheltered from improving recombination.

Given the vital functions the mitochondria must perform efficiently and responsively and the genetic issues surrounding the mtDNA, the mtDNA is an excellent place to search for variation influencing to human health and ageing.

UKBiobank (UKB) has collected data on nearly half a million people and continues to augment the database with updates on the participants' health. Since recruitment, the health changes of these older individuals have recorded their journeys from middle-age and into later life. The participants have also being genotyped which allows researchers to find associations between rates of disease with gene regions.

A portion of the information is lost when using the imputation method developed on the nuclear genome to process genotyped mtDNA. A second route which uses the few known bases to guess a haplogroup also fails to include some assumptions in its priors, leaving space to improve on this method too.

The aim of this project is to develop an improved imputation algorithm for low coverage mtDNA, mitochondrial DNA variants and the haplogroup tree. This would enable exploration of the UKBiobank's participants mtDNA variants.

Preceding this step, the library of data must undergo a process called "*in silico* genotyping", as if the library samples also underwent the UKBiobank's genotyping. In silico genotyping must be extensively validated for its accuracy before implementation to extrapolate the UKBiobank's valuable samples from genotyping data into predicted sequence.

The human mitochondrial sequences in GenBank (`https://www.ncbi.nlm.nih.gov/genbank/`) are used as the source of statistics on global haplogroup frequencies, variation and haplogroup. This data set is used to build and inform the structure of Phylotree, the central relatedness tree for all human mtDNA samples [100].

Mitomap contains a webpage analysing the haplogroup frequencies for the complete genome sequences stored in GenBank. The relevant human mitochondrial sequences have been collated with the details of their search being:

```
ddbj_embl_genbank[filter] AND txid9606[orgn:noexp] AND
complete-genome[title] AND mitochondrion[filter]
```

MitoMap have processed this data and list each sample against the sample's haplogroup and the places that the genetic sequence differs from the rCRS, stored at https://www.mitomap.org/foswiki/bin/view/MITOMAP/GBFreqInfo. It is from this website the training data is taken, and referred to throughout as the "MitoMap Haplogroup library" [18].

Chapter 2

Methodology

Please note: a **glossary** and the **coding** can be found in the appendix 4.5.

All functions were designed, written and tested by the author, and are included in the appendices 4.5. The functions make use of the packages: *Stringr*, (specifically; `str_detect`, `n_char`, `str_sub`) [2], *Ggplot2* [105] and *Ggraph* for the graphical plots and tree diagrams [1].

I have developed a novel method to transform a library of complete mtDNA sequences to appear as if genotyped, referred to as *in silico* genotyping. The *in silico* genotyped library can be used to impute experimental samples, which have been genotyped, using pattern matching.

The method has been developed with a view to its first application on an experimental dataset: the UKBiobank [21]. A diagram of the steps is shown in figure 2.2.

My library of complete mtDNA sequences has been sourced from GenBank [6]. Validation of the process is provided by the examination of predictions made on a set of test data. The test data are randomly drawn from MitoMap Haplogroup library, with the remaining MitoMap Haplogroup library samples providing the training data. See figure 2.1 for a map.

The structures of the MitoMap Haplogroup library data and the UKBiobank experimental data are quite different. However, the recording systems offer a route to potential equivalence, due to their both being recorded using differences from a common reference sequence, the revised Cambridge Reference Sequence (rCRS) [13]. *In silico* genotyping translates the MitoMap Haplogroup library data into a format equivalent to that of the UKBiobank experimental data.

Both the *in silico* genotyped MitoMap Haplogroup library sample data and UKBiobank experimental sample data consist of a list containing only zeros, ones and full stops. A concatenation of a sample's list into a single string is called a barcode. Barcodes enable easier match searches including the use of regular expression matching. Using regular expression match searching is more computationally intensive but is essential to find the compatible barcodes. Regular expression searching makes allowances for the `neither` calls, so these loci can match both `reference` and `alternative` calls.

I have created four accuracy tests which validate the translation and ensure the equivalence of the two data sets. Now the two data sets are compatible, library-training matches can be found for each of the test samples. The set of matching samples are collapsed into two predictions;

UKBiobank experimental sample's (a) haplogroup(s) prediction and, (b) variant list prediction. Both predictions have corresponding weightings to represent the confidence with which each expected haplogroup or expected variant is predicted.

The process of making the two data structures completely align is complex to automate and requires validation before use. This is accomplished by segmenting the training data and retaining a subset which is used as a test set. The success of this testing is explored by answering;

1. What proportion of the library-test samples can be found library-training matches and, hence, predictions?
2. What proportion of the predictions predict the expected haplogroup?
3. What is the distance of each haplogroup in the prediction from the expected haplogroup?
4. How many expected information units are shared by all of the haplogroups in the prediction?
5. What proportion of the variant list is correct?

2.1 Data Sources and Structures

2.1.1 MitoMap Haplogroup Library

MitoMap contains a library of mtDNA sequences from human mitochondrial haplogroups from all over the world. All of the MitoMap Haplogroup library sequences have been published with manually defined haplogroups using the details of their sequence and Haplogrep2 [104]. The MitoMap Haplogroup library is not perfect either in scope or the quality of its information but offers a large training data set from which to form predictions.

The data were downloaded from MitoMap's frequency information page (`https://www.mitomap.org/MITOMAP/GBFreqInfo`) on 26th of July 2019 [5].

The format of data storage at MitoMap Haplogroup library is to record each sample's; (a) unique ID code, (b) haplogroup, and (c) a string containing a complete list of the mtDNA sequence differences from the rCRS. Each difference from the rCRS, called a variant, in the form of a concatenation of short strings built from the base pair address (locus) and the nature of the difference from the rCRS (sequence change). The rCRS has been located to the H2a2a1 haplogroup [11].

Table 2.1 has five example samples from the MitoMap Haplogroup library used as the training data in this project. These samples are from closely related haplogroups and hold high level of similar variants from the rCRS sequence. The coding of these variants marks deletions, such as the "514d.CA" present in all the of examples, with a "d.". Four of the five examples include the same insertion, "309CCT". Insertions are discernable from a single-nucleotide polymorphism by carrying more than one letter after the numbered locus.

Descriptive Statistics

The information MitoMap Haplogroup holds about each sequence allows some additional variables to be calculated. When making use of the MitoMap Haplogroup library to make predictions, it is salient to know things such as;

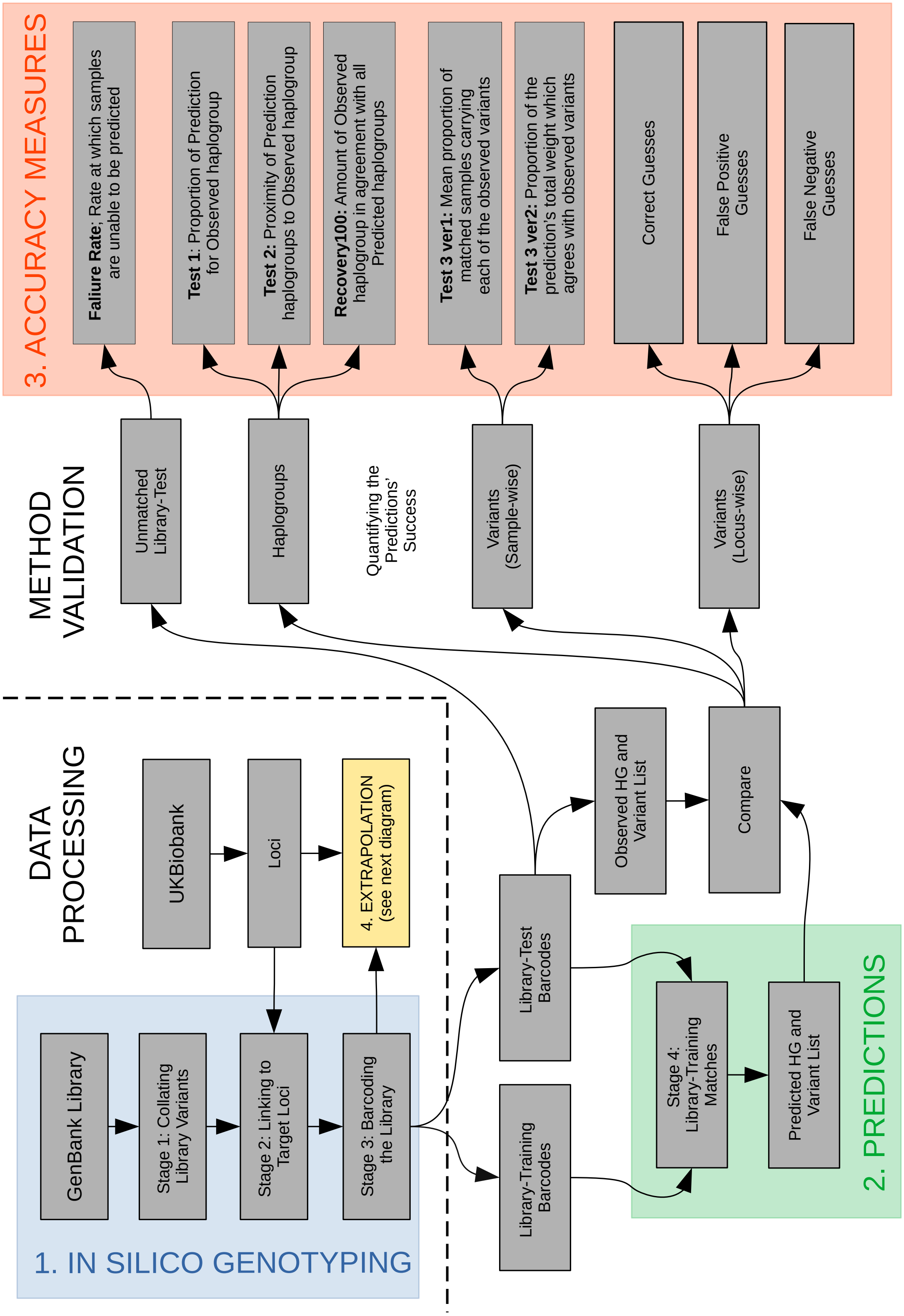

1. IN SILICO GENOTYPING
GenBank Library
Stage 1: Collating Library Variants
Stage 2: Linking to Target Loci
Stage 3: Barcoding the Library
DATA PROCESSING
UKBiobank
Loci
4. EXTRAPOLATION (see next diagram)
Library-Training Barcodes
Library-Test Barcodes
2. PREDICTIONS
Stage 4: Library-Training Matches
Predicted HG and Variant List
Observed HG and Variant List
Compare
METHOD VALIDATION
Unmatched Library-Test
Haplogroups
Quantifying the Predictions' Success
Variants (Sample-wise)
Variants (Locus-wise)
3. ACCURACY MEASURES
Faliure Rate; Rate at which samples are unable to be predicted
Test 1: Proportion of Prediction for Observed haplogroup
Test 2: Proximity of Prediction haplogroups to Observed haplogroup
Recovery100: Amount of Observed haplogroup in agreement with all Predicted haplogroups
Test 3 ver1: Mean proportion of matched samples carrying each of the observed variants
Test 3 ver2: Proportion of the prediction's total weight which agrees with observed variants
Correct Guesses
False Positive Guesses
False Negative Guesses

Figure 2.1: Pathway from UKBiobank's loci to MitoMap Haplogroup Library's barcodes

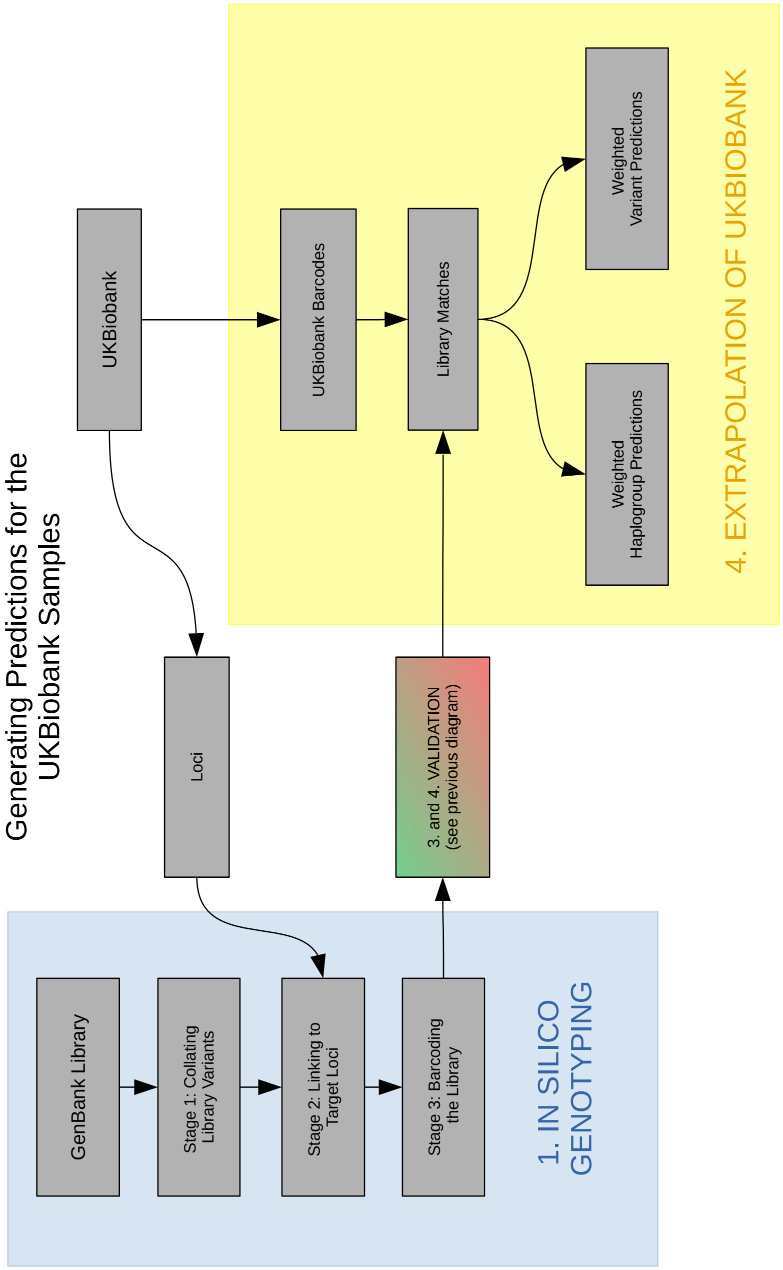

Generating Predictions for the UKBiobank Samples
1. IN SILICO GENOTYPING
GenBank Library
Stage 1: Collating Library Variants
Stage 2: Linking to Target Loci
Stage 3: Barcoding the Library
Loci
UKBiobank
3. and 4. VALIDATION (see previous diagram)
4. EXTRAPOLATION OF UKBIOBANK
UKBiobank Barcodes
Library Matches
Weighted Haplogroup Predictions
Weighted Variant Predictions

Figure 2.2: Pathway from MitoMap Haplogroup Library's barcodes to extrapolation of UKBiobank data

GenBank ID	Haplogroup	Variants
AF346971.1	A2b1	73G, 146C, 153G, 228A, 235G, 263G, 315CC, **514d.CA**, 663G, 750G, 1438G, 1736G, 2706G, 4248C, 4769G, 4824G, 7028T, 8027A, 8794T, 8860G, 10609C, 11365C, 11719A, 12007A, 12705T, 14766T, 15326G, 15731A, 16111T, 16223T, 16265G, 16290T, 16319A, 16362C
AF382010.2	A2q	4T, 73G, 153G, 235G, 263G, **309CCT**, 310C, 437T, **514d.CA**, 663G, 750G, 1438G, 1736G, 2706G, 4248C, 4769G, 4824G, 5480G, 7028T, 8027A, 8794T, 8860G, 11719A, 12007A, 12705T, 13448T, 14766T, 15205T, 15326G, 16111T, 16209C, 16223T, 16290T, 16319A, 16362C, 16519C
AP008265.1	A5a	73G, 235G, 263G, **309CCT**, 310C, **514d.CA**, 663G, 750G, 1438G, 1736G, 2156AA, 2706G, 4248C, 4655A, 4769G, 4824G, 7028T, 8563G, 8794T, 8860G, 11536T, 11647T, 11719A, 12705T, 14766T, 15326G, 16187T, 16223T, 16290T, 16319A
AP008276.1	A5a1a1	73G, 235G, 263G, **309CCT**, 310C, **514d.CA**, 663G, 750G, 1438G, 1736G, 2156AA, 2706G, 4248C, 4655A, 4769G, 4824G, 5773A, 7028T, 8563G, 8794T, 8860G, 10801A, 11536T, 11647T, 11719A, 12705T, 12880C, 14766T, 14944T, 15326G, 16187T, 16223T, 16290T, 16319A
AP008290.1	A5a1a1a	73G, 235G, 263G, **309CCT**, 310C, **514d.CA**, 663G, 750G, 1438G, 1736G, 2156AA, 2706G, 4248C, 4655A, 4769G, 4824G, 5460A, 5773A, 7028T, 7492T, 8563G, 8794T, 8860G, 10801A, 11536T, 11647T, 11719A, 12705T, 12880C, 13221G, 13225A, 14766T, 14944T, 15326G, 16187T, 16223T, 16290T, 16319A

Table 2.1: Example of the data structure from the MitoMap Haplogroup library. The variants highlighted in bold are examples of indels (insertions or deletions). "514d.CA" is present in all the of examples and marked as a deletion with a "d.". Four of the five examples include the same insertion, "309CCT". Insertions are identifiable as they have a list of letters with a length greater than one.

- **Number of samples also in that haplogroup**; a count of library samples with the same string,
- **Macrobranch membership**; denoted by the first part of the haplogroup string,
- **Variant number**; a count of the distinct variants in the variant string,
- **Phylogenetic tree distance from the rCRS**; explained below.

Phylotree provided the macrobranch node structure seen in dendrogram 1.3. Finding a measure of the phylogenetic tree distance on a longer scale than the macrobranch involved a proxy measure of distance. The rCRS is located to the H macrobranch node on the tree so the number of jumps needed to get from that node to the sample's macrobranch was used. By counting the nodes from the sample's macrobranch node to the rCRS node, a measure of distance was found which incorporated the tree's shape and could relate to the variant number.

The distances of each macrobranch are shown in figure 3.2.

2.1.2 UKBiobank

In silico genotyping was developed with the aim of rendering the UKBiobank experimental data set accessible to association studies by interpolating further using the known data. This is enabled by the re-formatting of the MitoMap Haplogroup Library data to mimic the structure of the UKBiobank experimental data.

The source file of the UKBiobank data format is preceded by a range of metadata, and arranged in an array of the positions as rows and the samples each form a column of results, one encoded answer for each loci. See figure 2.3 for an example.

The primary intended use of the data collected about the mitochondrial DNA on the array was to assign the sample an approximate haplogroup. This fact is reflected in the choices of locus and alternative SNPs which, when used in combination would help to narrow the range of possible haplogroups to which sample may belong.

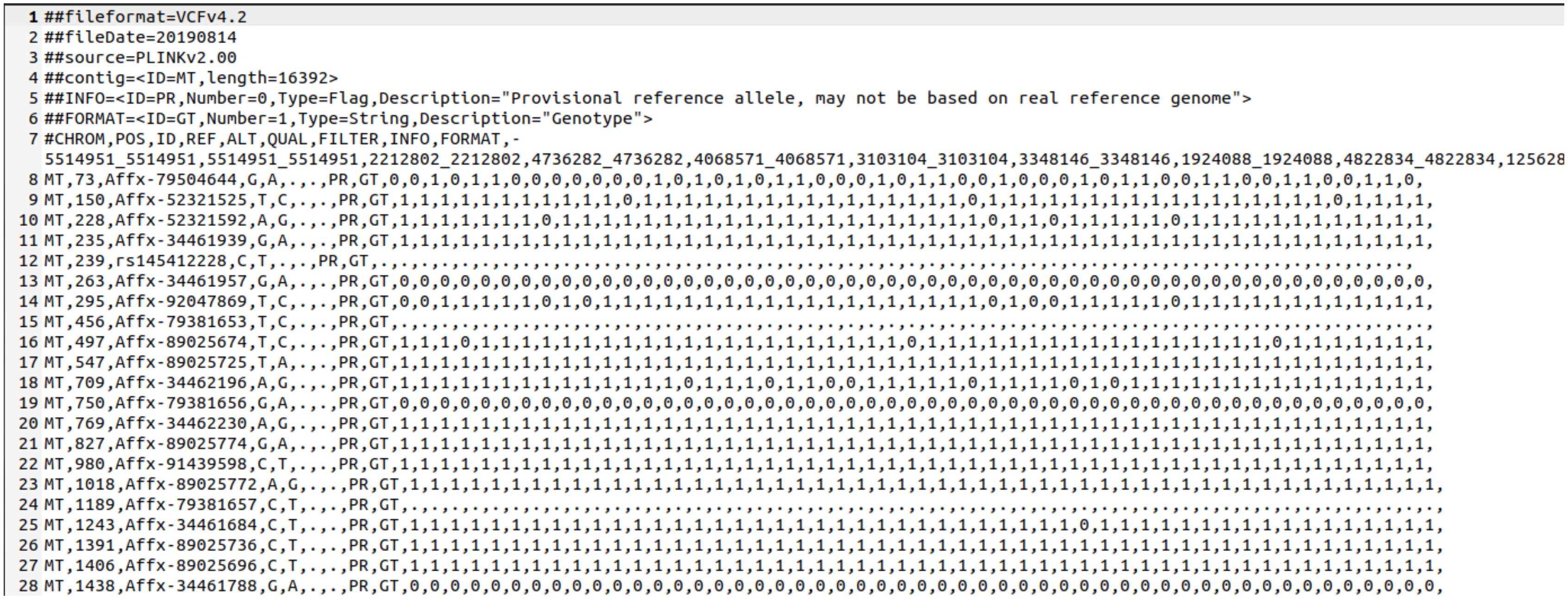
Descriptive Statistics

In order to mimic the UKBiobank's array data using the library data, identical loci must be used.

The UKBiobank has collected sample genotype data, having checked a subset of 265 positions (1.6%) in the 16.5kb mitochondrial genome. Each position is checked for just two variants which are referred to as `reference` (identical to the rCRS) and `alternative` (different in a single, specified way to the rCRS). Each sample is recorded as answers at these positions and no others. The data structure of UKBiobank is a variant call format (VCF) array resulting from a microarray collecting a binary call on specified positions in the genome.

In addition, the data includes the use of a full stop (".") for when a definitive call is not possible, called `neither`. Such a failure may arise because:

1. the DNA was unable to produce a result (sample-level failure),
2. the microarray failed at this locus (microarray-level failure),
3. or, the sample genuinely shows neither the `reference` variant nor the `alternative` variant at this position as a third variant was found at the locus, or close enough to mask the true SNP call.



```
##fileformat=VCFv4.2
##fileDate=20190814
##source=PLINKv2.00
##contig=<ID=MT,length=16392>
##INFO=<ID=PR,Number=0,Type=Flag,Description="Provisional reference allele, may not be based on real reference genome">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM,POS,ID,REF,ALT,QUAL,FILTER,INFO,FORMAT,-
5514951_5514951,5514951_5514951,2212802_2212802,4736282_4736282,4068571_4068571,3103104_3103104,3348146_3348146,1924088_1924088,4822834_4822834,125628
MT,73,Affx-79504644,G,A,.,.,.,PR,GT,0,0,1,0,1,1,0,0,0,0,0,0,1,0,1,0,1,0,1,1,0,0,0,1,0,1,1,0,0,1,0,0,0,1,0,1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,
MT,150,Affx-52321525,T,C,.,.,.,PR,GT,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,
MT,228,Affx-52321592,A,G,.,.,.,PR,GT,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,0,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,
MT,235,Affx-34461939,G,A,.,.,.,PR,GT,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
MT,239,rs145412228,C,T,.,.,.,PR,GT,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,
MT,263,Affx-34461957,G,A,.,.,.,PR,GT,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
MT,295,Affx-92047869,T,C,.,.,.,PR,GT,0,0,1,1,1,1,1,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,0,0,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,
MT,456,Affx-79381653,T,C,.,.,.,PR,GT,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,
MT,497,Affx-89025674,T,C,.,.,.,PR,GT,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,
MT,547,Affx-89025725,T,A,.,.,.,PR,GT,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
MT,709,Affx-34462196,A,G,.,.,.,PR,GT,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,0,1,1,0,0,1,1,1,1,1,0,1,1,1,1,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
MT,750,Affx-79381656,G,A,.,.,.,PR,GT,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
MT,769,Affx-34462230,A,G,.,.,.,PR,GT,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
MT,827,Affx-89025774,G,A,.,.,.,PR,GT,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
MT,980,Affx-91439598,C,T,.,.,.,PR,GT,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
MT,1018,Affx-89025772,A,G,.,.,.,PR,GT,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
MT,1189,Affx-79381657,C,T,.,.,.,PR,GT,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,.,
MT,1243,Affx-34461684,C,T,.,.,.,PR,GT,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
MT,1391,Affx-89025736,C,T,.,.,.,PR,GT,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
MT,1406,Affx-89025696,C,T,.,.,.,PR,GT,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
MT,1438,Affx-34461788,G,A,.,.,.,PR,GT,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
```

Figure 2.3: Example of genotyping data stored as a VCF file

Quality Control

I undertook quality control on the UKBiobank data with the removal of loci and samples with high levels of missingness.

The *neither* call rate for each loci was found through counting the percentage of UKBiobank samples getting a full stop for that locus. Loci with a percentage over 85% were removed.

To avoid the inclusion of loci which are entirely homogenous in the MitoMap Haplogroup library, I cross-referenced the list of potential target loci with the list of loci which are entirely homogenous in the MitoMap Haplogroup library samples. Only target loci which fell on heterogeneous bases were used.

Next, the *neither* call rate for each sample was found through counting the UKBiobank target loci getting a full stop. Samples with a count of over 50 were removed.

2.2 In Silico Genotyping of Library Data

Because the structures of the MitoMap Haplogroup library data and UKBiobank experimental data differ, the library data will be processed to create an equivalent format; a process called *in silico genotyping*.

I have written a series of functions to perform *in silico* genotyping on the MitoMap Haplogroup library data, translating the data into a format equivalent to as if each sample had been probed by the microarray used by UKBiobank on their participants.

The microarray used by UKBiobank contained hundreds of thousands of probes designed to find a set of variants, including several hundreds variants on the mtDNA. To mimic the mtDNA data, the subset of variants with loci on the mtDNA shall be the *target loci*. It is this subset that the *in silico* genotyping shall emulate.

Firstly, an exhaustive list of every variant in every MitoMap Haplogroup library sample is collated. Next, variants local to the target loci are assessed for how the presence of the variant would effect how the probe would bond, which alters the decision made at the locus. Thirdly, one target locus at a time, each MitoMap Haplogroup library sequence is searched for variants local to the locus. The result is a emulation of the result had that mtDNA sequence been probed by that microarray at that locus.

By working through the target loci, each MitoMap Haplogroup library sample is given a set of results, one for each target locus, which can be used in further processing. The results are represented by ones, zeros and full stops, just as the UKBiobank data which it is mimicking.

The pattern of zeros, ones and full stops is concatenated into a single string, for cataloguing, called a “barcode”. Barcodes are made for each of the MitoMap Haplogroup library samples and each of the UKBiobank experimental samples.

Due to the equivalence of the *in silico* genotyped MitoMap Haplogroup library data format and the format of the UKBiobank experimental data, matches can be made between samples from opposing data sources.

2.2.1 Stage One: Collating Library Variants

The first stage of *in silico genotyping* extracts a list of all the variants (differences from the rCRS) present in library samples. The frequencies of these variants vary from single occurrences, as private mutations, through to being present in most samples, as commonly held variants not present in the rCRS.

The resulting array contains details about each variant; mtDNA base-pair address for the start point, and the bases which differ. For example, output at this stage includes these variants: ”3708G”, ”12170A”, ”16069CG”, ”14614G”, ”5972T”, ”2971T” ”8605T”, ”10083T”, ”15049T”, ”446C”, ”15191A”, ”14933C”, ”12745CTAG”, ”15575A”, ”13969T”, ”8510d.A”, ”14161d.AG”, ”16285d.A”, ”11T” and ”12756C”.

Variant Finder Function
Make a unique, central list of all variations present by combining all the lists of variants for each sample.

Array Maker Function, using Matrix Process Fixed Function
Split each variant using boundaries between numbers and letters (using my function to split “information units”). For example, “132CCT” becomes “132” and “CCT”.

Knowing the wider range effects of variants is vital to incorporate into the target loci barcoding decisions. For each target locus, all of the relevant variants present in the library must be gathered. This simply equates to finding all the variants with a start locus within the range of the probe. This is enabled using the data returned by Array Maker.

Loci Variant Analysis

The locus to which any variant is linked can be found using the number at the start of the string describing the variant. For instance, “73C” describes a single nucleotide difference from the rCRS at 73 base pairs.

Two measures are used to map out the broader structures seen in the mtDNA sequences by aggregating the variations in the library samples linked to each locus;

1. **Homogeneity** is a binary grouping where bases are either;
   (a) **Heterogeneous**: There are instances of variation at this locus in one or more of the library samples,
   (b) **Homogenous**: There are no instances of variations at this locus in any of the library samples,
2. **Variation**: The frequency of variations linked to this locus in the library samples.

Despite there being a relationship between these two measures, as homogeneity is a variation of 0, both numbers were need for further investigation to encompass the complexity of the mtDNA.

Until this point, the data processing has been without reference to the specific loci to be targeted. Any further data processing will be specific to the UKBiobank loci (“target loci”), but could also be used to assess the performance of other microarray loci sets.

2.2.2 Stage Two: Linking to Target Loci

The MitoMap Haplogroup library variants are now in a format where we can decide for each of the target loci:

1. Which variants hold information relevant to the target locus?
2. Does the variant classify the variant possessor as reference, alternate or equal to neither at this locus?

The next section takes the target loci and results in a look-up table where all of the variations relevant to each target locus are recorded and assigned as either representing a call of (a) a base the same as the reference template (rCRS), (b) a base the same as the alternative specified by the UKB, or (c) a base equal to neither of those.

Locus	rCRS or reference	alternative Variants	neither Variants
123	123T	123C	123G, 120A, 124.dT, 127AAT
456	456G	456A	455T, 459C

Table 2.2: Example of output of Linking to Loci

Figure 2.4 shows the process of assigning a locus. Categorising variants reduces to three rules:

1. Reference For this locus, the MitoMap Haplogroup library samples hold no relevant variants.
2. Alternative The MitoMap Haplogroup library samples holds only the specified alternative SNP, marked yellow.
3. Neither The MitoMap Haplogroup library samples holds a variant from the neither list. This call vetoes either a reference call or an alternative call.

The process of variant-locus assignment is completed by two functions;

Find RAN Function
Categorise the variant's base as the same as **Alternate** or **Neither**.

Pre-Barcoding Function
Create a list of **Alternate** library variants connected to each of the target loci. Create a list of **Neither** library variants connected to each of the target loci.

The data frame returned has a complete list of all the library variants which would produce a sample's assignment as neither ("."), and the variant required to produce an assignment as alternative, if found alone in the range surrounding the target locus in question. Each target locus has a row containing the two lists.

From this point, the variant list of each of the library samples can simply be searched for alternate or neither calls at each locus in the target list.

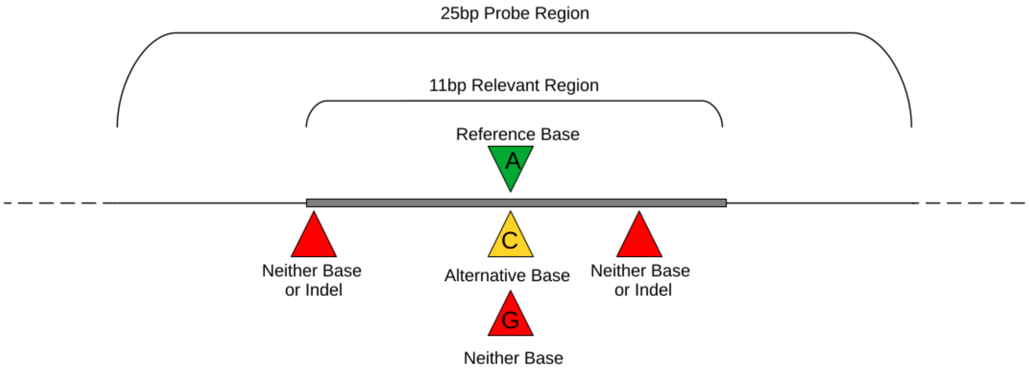


Figure 2.4: Example of locus assignment

Comparing Variant Frequencies

In order to search for problems, a comparison is made of the call rates for the *in silico genotyping* of the MitoMap Haplogroup library with the *in vitro genotyping* of the UKBiobank experimental data. Any loci with issues should have altered call rates.

2.2.3 Stage Three: Barcoding

As the variants are now classified, the samples' variant lists can be analysed to obtain the samples' barcodes.

Using the look-up table produced in the previous step (2.2), each library sample is searched for relevant variants for each target locus. The default call for each locus is `reference` so, if no relevant variants for the locus were found, the call would remain as a default `reference`. Referring to figure 2.4, this would represent a sample holding only the green variant but would be coded by an omission of data reflecting no variations which differe from the rCRS. A sample holding just the yellow variant, a call of `alternative` is made. As the `reference` variant and the `alternative` variant occupy the same position on the genome, they are mutually exclusive.

Variants which are local to the locus can be held at the same time as variants on the probed locus. When finding any combination including a variant marked on diagram 2.4 as a red triangle, the call of `neither` is made, masking any signals from the probed locus. The masking effect of the off-target signal jamming variants must be coded for in later stages by relaxing the rules of a regular expression match search, involving matching any locus with a `reference`, `alternative` or `neither`. The search criteria must work bi-directionally for `neither` calls in either the search bait or the search pool, discussed in greater detail in table 2.3.

Using the examples in table 2.2, if a sample only carries 123C, that locus will be assigned `alternative`. If it carries any of the variants listed in the `neither` variants column for that locus, an assignment of `neither` will result.

The results of *in silico* genotyping samples should reflect how the sample would appear genotyped on UKBiobank's main microarray. By comparing the alternative and `neither` variant list for each of the loci and making a call, the sample's locus assignments build into a list. This list is concentrated into a single string 240 characters long and added to the sample's information as a "barcode".

Building the answers into a single string encoding the 240 bits of information, referred to as a barcode, was preferred to creating a VCF-style file. At the next stage, the barcodes are to be matched to one another and this was easy to do using a regular expression search, which accommodated the `neither` calls' match non-specificity.

Using the finalised list of loci and their relevant variants, each MitoMap Haplogroup library sample can be given a code for each of the 240 target loci.

Barcoding 1 Function
A sub-routine of Barcoding 2, this accepts three lists; (a) the sample's variants, and the (b) `alternative` variants and the (c) `neither` variants at a single position. This function returns a list of two elements which represent the lengths of the lists of intersects between (a) and (b), and (a) and (c). This quantifies how many of the sample's variants match those quoted as `alternative` or `neither`.

Logically, each sample should score as below:

- `alternative` 0, `neither` 0 There are no variants within range of the locus. This is a reference call resulting in a zero.
- `alternative` 1, `neither` 0 There is only the single, specified `alternative` variant in range of the locus. This is an `alternative` call resulting in a 1.
- `alternative` number irrelevant, `neither` $>$ 0 There is one or more of the `neither` list present, which over-rides any other call, and a call of `neither` is made. This results in a "." being added to the barcode in this position.

Each of the library samples has now received a barcode which should reflect exactly how that sample would have been assessed using the UKBiobank's microarray. A link between the library information and the UKBiobank data has been forged, allowing predictions to be made.

Regular Expression Searching

Regular expression searches allow searching of text far beyond looking for the exact string. Agreed symbolic representations of options let the search include much more information than the known specific letters or numbers [99].

I have included "bait" and "pool" to aid description of the direction of the searching. The "bait" is the barcode of the sample for which matches are being sought. *In silico* genotyping extracts and builds barcodes for the library samples. When a sample's barcode is used to find compatible matches, it is referred to as the search "bait". The barcodes which are being search through are called the the "pool". These original barcode string must be adapted to perform in these roles, in fact, two barcode formats are used "bait" and "pool" to enable the accommodation of `neither` calls.

The extension enabled by regular expression searching is used to include the `neither` calls in the barcodes. Barcodes used to find matches, called bait and used in the regular expression search, must be converted to enable the matching behaviour in table 2.3. To produce a searchable version of the library-training samples, the "." must be replaced with a "2", referred to as the “pool”. To convert the library-test barcodes into an appropriate “bait”, several changes must be made; (a) `reference` “0”s changed to “[02]”, (b) `alternative` “1”s changed to “[12]”, and (c) `neither` “.”s changed to “[012]”. With these adaptations, the pool of library-training samples can be searched for a set of matches to each bait from the library-test set.

The options held in the square brackets are the digits which would trigger a match at that locus.

	Pool	Pool	Pool
	`reference`	`alternative`	`neither`
	0	1	2
Bait `reference` [02]	Match		Match
Bait `alternative` [12]		Match	Match
Bait `neither` [012]	Match	Match	Match

Table 2.3: Regular Expression search outcomes for a single locus in the bait and pool barcodes

Normal string matching searches, which look for exact matches, cannot be used here as `neither` calls act to reduce the specificity at that locus. Regular expression searches accommodate indecision or wobble modelling how an additional mutation in the range of the locus could jam the signal of `reference` or `alternative`.

Regular expression searching requires considerable computation time but is necessary to accommodate the structure of the two data sets, which both include `neither` as a possibility.

2.2.4 Stage Four: Forming Predictions

Creating predictions from the MitoMap Haplogroup library samples with barcodes which agree with our sample has four stages.

1. Take the “bait” type barcode to find any “pool” type barcodes which match, account for the unknown bases in both the bait and the pool as in table 2.3,
2. Gather all the matching samples in the Library,
   (a) Gather the haplogroups of these samples into a single list and pass to “Full to freqs” function which returns an unique list of haplogroups and a list of their relative frequencies,
   (b) Gather the variants of these samples into a single list and pass to “Full to freqs” function which returns an unique list of variants and a list of their relative frequencies.
3. Record the new columns to the MitoMap Haplogroup Library data frame.

The MitoMap Haplogroup library-test predictions will appear as table 2.4.

Sample ID	HG	HG Prediction	HG Weights	Variant Prediction	Variant Weights
ID123456	H1a	H1a,H1,H1a1	0.6, 0.1, 0.3	123A, 456T, 789G, 1011C	1, 1, 0.9, 0.4

Table 2.4: An illustrative example of the appearance of a prediction

As each matching sample can only have one haplogroup but many variants, the weights in the haplogroup list add up to 1, but the variant list weights may be much higher.

Full to Freqs Function
This function takes a full list of items (including repeats) and returns a "unique" list and a list of the frequencies of each item in the unique list. Both lists should contain the same number of items and correspond directly.

2.3 Validation with Test Data Set

2.3.1 Source of the Test Data

The MitoMap Haplogroup library consists of 45,888 samples. When producing predictions for the UKBiobank experimental samples, this entire MitoMap Haplogroup library data set can be used. However, before this is completed, the validity and accuracy of this method of imputing from the barcode must be investigated.

In order to perform this, the MitoMap Haplogroup library is split into 80% "library-training data" and 20% "library-test data". The division of the library into 20% library-test and 80% library-training set was based on recommendation from machine learning data splits [94]. The MitoMap Haplogroup library sequences are in alphabetical order so a random number generator was used to choose 10,000 (21.8%) indices to prevent a skew in the haplogroups selected.

Using a random choice of samples, 10,000 samples are withdrawn from the MitoMap Haplogroup library to form the library-test data set. The remaining 35,888 samples are used to form predictions for each of the MitoMap Haplogroup library-test samples. This will be performed by; (a) taking each MitoMap Haplogroup library-test sample in turn, (b) using its barcode as bait, (c) finding a set of regular-expression-search matches in the MitoMap Haplogroup library-training data, and (d) forming weighted predictions of haplogroups and variants.

2.3.2 Failure Rates

Some 7.5% of the MitoMap Haplogroup library-test samples fail to find matches despite using a regular-expression-based search with the barcode converted to a regular expression matching as in table 2.3. These samples are referred to "Library-test-fails". The "Library-test-passes" are samples which do find a regular expression match in the MitoMap Haplogroup library-training data.

Being unable to make predictions for a subset of the global population requires further investigation as excluding these samples will lead to biases in any population-focused genome-wide association studies.

In an effort to avoid over-tuning, sample match frequencies of zero are used to diagnose problems, not failures to make accurate predictions.

Three approaches to exploring this; (a) finding the failure rate in the experimental data set to confirm the problem, (b) calculating a theoretical expected failure rate, and (c) finding correctable patterns.

Calculation of an Expected Failure Rate

Dividing the MitoMap Haplogroup library to form a test set and a training set to perform the validation introduces problems specific to the validation process. The structure of the library, effectively many small groups gathered together, makes this not ideal.

Using a combination of the matching sample number and frequency of each barcode in the library, an expected failure rate for each barcode can be obtained. By estimating a failure rate when we remove 10,000 samples to use as a test data set, it can be assumed that a failure rate in that region is ostensibly attributable to the act of dividing the library data set. This issue will be avoided when predicting the UKBiobank samples as the entire library will be utilised for this task. A failure rate which exceed the expected rate would point to problems in the *in silico* genotyping.

For each barcode, A, in the library, it has m library matches including the match to itself. Each sample has a $p(chosen)$ chance of being selected for the test batch.

$$p(\text{chosen}) = \frac{N_{\text{samples}}}{N_{\text{total}}} \tag{2.1}$$

For the barcode to fail, all of its matches must also fall into the library-test set.

An approximation for the probability that a barcode would fail is:

$$p(\text{fail}_A) \approx (p(\text{chosen}))^{m-1} \tag{2.2}$$

We know that there are f samples with m matches with a chance of being selected of $p(\text{chosen})$. Each sample has a chance of having all of its potential matches removed to the MitoMap Haplogroup library-test set, which plummets as the number of matches increases. The cumulative sum gives E, an estimate of failing samples

$$E = (f_1 * p(\text{fail}_1)) + \ldots (f_x * p(\text{fail}_x)) \tag{2.3}$$

This was extended to cover a 10% library-test set and a 40% library-test set by changing $p(\text{chosen})$.

By comparing these predictions of the library-test samples with the observed haplogroup and observed variants, the accuracy of these predictions can be measured with the three tests explained below.

Plotting Match Number

In order to find the number of potential matches each sample had, each sample's bait barcode was used to search the entire list of the MitoMap Haplogroup library's pool barcodes. Each sample in the library was found matches in the entire library, although each bait barcode found its pool barcode so this was corrected for by subtracting 1. Each MitoMap Haplogroup library sample has a value for the number of potential matches it had.

To create replicates, the library was passed through five cycles of the following algorithm: (a) randomly select 20% of library for library-test, (b) find library-test-fail samples, and (c) find the potential match number for these library-test fail samples.

Library-test-fails should have very low numbers of potential matches, unless their failure is caused by an additional problem in the *in silico* genotyping.

Comparing Passing and Failing Samples

The library-test set fell into two groups of samples called library-test-pass and library-test-fail, respectively. To collate an `alternative` call rates for each of the target loci in the library-test-fail samples the barcode of each sample in the library-test-fail group is split and the indices of each target locus receiving a `alternative` call are found. The number of `alternative` calls for each target locus are collated.

The `alternative` call rates are collated in the same way for the library-test-pass samples. The rates of `alternative` call for each locus can be compared to reveal any loci over- or under-represented in the library-test-fail group. A difference in `alternative` call rate between the groups would suggest the assignment of a variant on that locus was being differently assigned in the library-test-fail group, triggering that failure.

2.3.3 Quantifying Predictions' Success

Two tests form the basis of quantifying the distance between the predictions and the observed haplogroups. Tests 1 and 2 are measures of the haplogroup predictions made and will be presented in relation to the phylogenetic tree of haplogroups to highlight the spacial patterns in the characterisation.

Haplogroup Prediction Accuracy Tests

Each test sample is given a list of likely haplogroups, weighted by the number of matches in that haplogroup. With n_{matches} as the number of matches this sample's barcode found, and $n_{\text{correctHG}}$ as the number of the matches which the same haplogroup as the observed haplogroup.

$$\text{Weight} = \frac{n_{\text{correct}}}{n_{\text{matches}}} \tag{2.4}$$

Because the MitoMap Haplogroup library-test samples also have an observed haplogroup, a value was obtained for the proportion of the prediction which pointed to the correct haplogroup.

As illustrated in figure 2.5, the predicted haplogroup list is searched for the observed haplogroup. If present, the test 1 score is the weight which that haplogroup carried in the prediction. If the observed haplogroup is not present, the score is 0.

This score has a range from 0–1.

To gain insight into this width of haplogroup prediction accuracy, the haplogroup prediction is assessed using the information encoded in the haplogroup names. For example, "H1a1" is a sub-branch of "H1a" so having some idea of how close the wrong guesses were can help augment the binary method of test 1.

Figure 2.5 illustrates the process. The observed haplogroup is divided into information units. "H1a" would split into "H", "1", and "a". These units are rebuilt in increments, giving "H", "H1" and "H1a". The weights of the predicted haplogroups which begin with the "H" are added up to give a total for that unit, and so on until each unit has a score. Incremental sectioning the strings prevents H2a looking closely related to D2a and exploits the decreasing importance of each unit of information in the string.

The list of scores are returned with a length equivalent to the number of information units in the observed haplogroup.

This list is used in three ways:

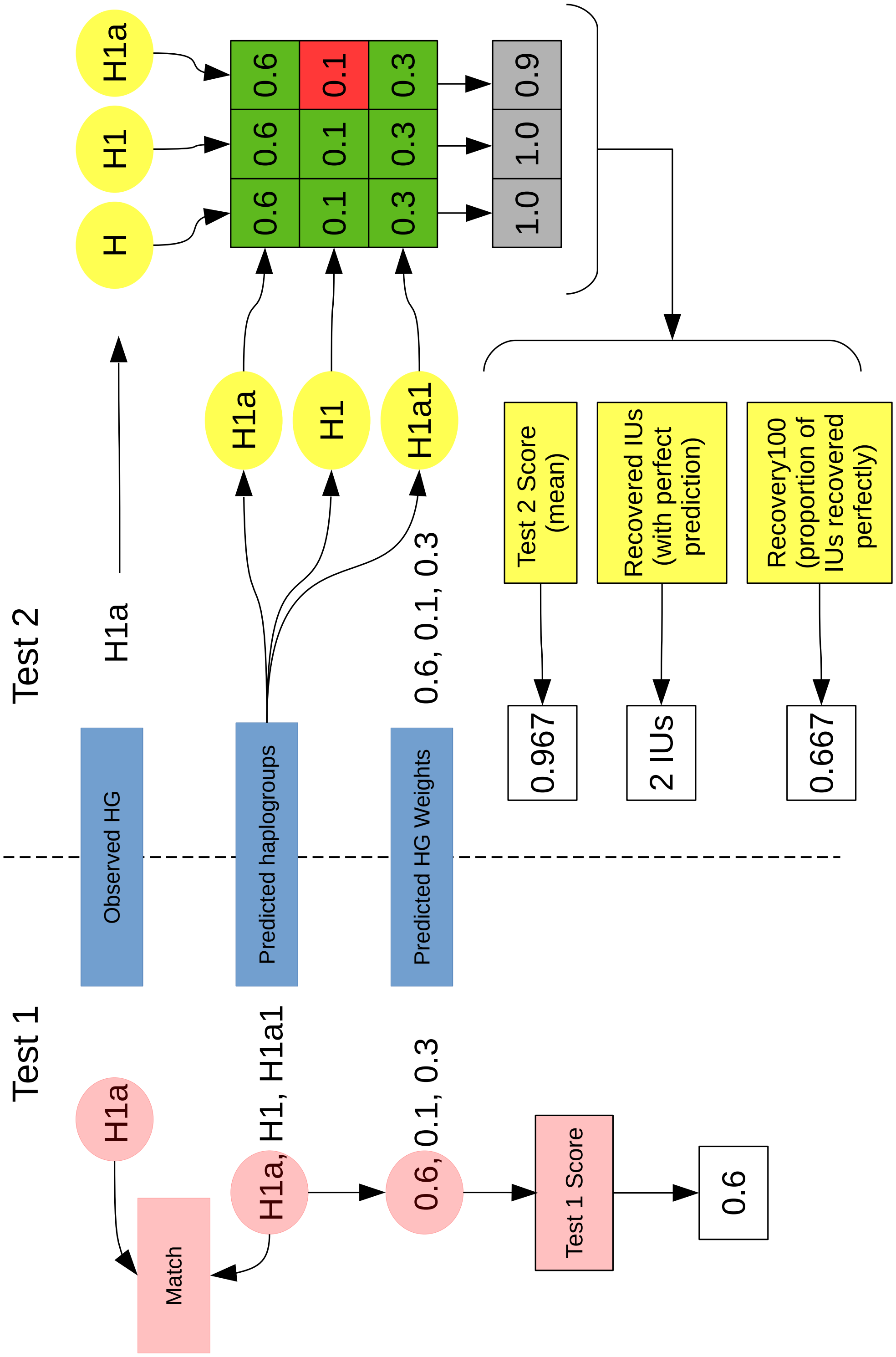


Figure 2.5: Haplogroup Accuracy Tests, tests 1 and 2

1. **Test 2 Score** Mean of the scores in list,
2. **Recovered IUs** The number of IUs predicted with perfect accuracy, i.e. a score of 1 because every matching sample possessed the expected IU.
3. **Recovery100** Proportion of expected IUs which were perfectly predicted.

Haplogroups vary in the number of information units they carry. When this factor is included in the Test 2 analysis, a measure of how many observed information units were recovered by all of the predicted haplogroups.

Variant Prediction Accuracy Testing

Test 3 is a measure of the accuracy of the variants expected compared to the variants observed in the samples.

The steps of this method are to;

1. Gather all the samples with pool barcodes compatible with the bait barcode,
2. Make a single, full, potentially very repetitive list of all the variants present in the group of matching samples,
3. For each of the expected variants, find a count of the times it appears in the full, repetitive list.
4. Convert each count into a rate, by dividing by number of samples, for a list of scores between 0 and 1. This represents the proportion of the compatible pool samples which contain each expected variant.
5. Compare the observed variant list and the predicted variant list as demonstrated in 2.6, collate sample scores for correct (C), false positive and false negative.
6. Create a single, non-repetitive list of variants which appear either in the sample's observed variant list or the predicted variant list. The length of this list (T) represents the sample's upper bound for correct scores.
7. Test 3 score is derived as: $Test3 = \frac{C}{T}$ and will fall between $0 - 1$.

As each variant appearing in the either the observed variant list or the predicted variant list scores the sample a point, the length of the combined list represents a maximum for the Test 3 score. The single point may be assigned entirely to correct, false positive or false negative, or split, but the entire point is assigned each time.

2.3.4 Macrobranch-Based Analysis

When representing mtDNA sequences using solely the 240 target loci, there is a systematic error introduced as the signature variation of some groups will be included while other groups will have their signature variation excluded.

If haplogroups are characterised with a varied level of accuracy because of which data are included, the quality of their predictions will vary too. Gathering the MitoMap Haplogroup library-test predictions by the sample's observed macrobranch will highlight the inconsistent treatment.

Poor characterisation by the target loci will result in inaccurate or broad predictions. These are apparent in both the haplogroup and variant predictions and are explored through the tests outlined above.

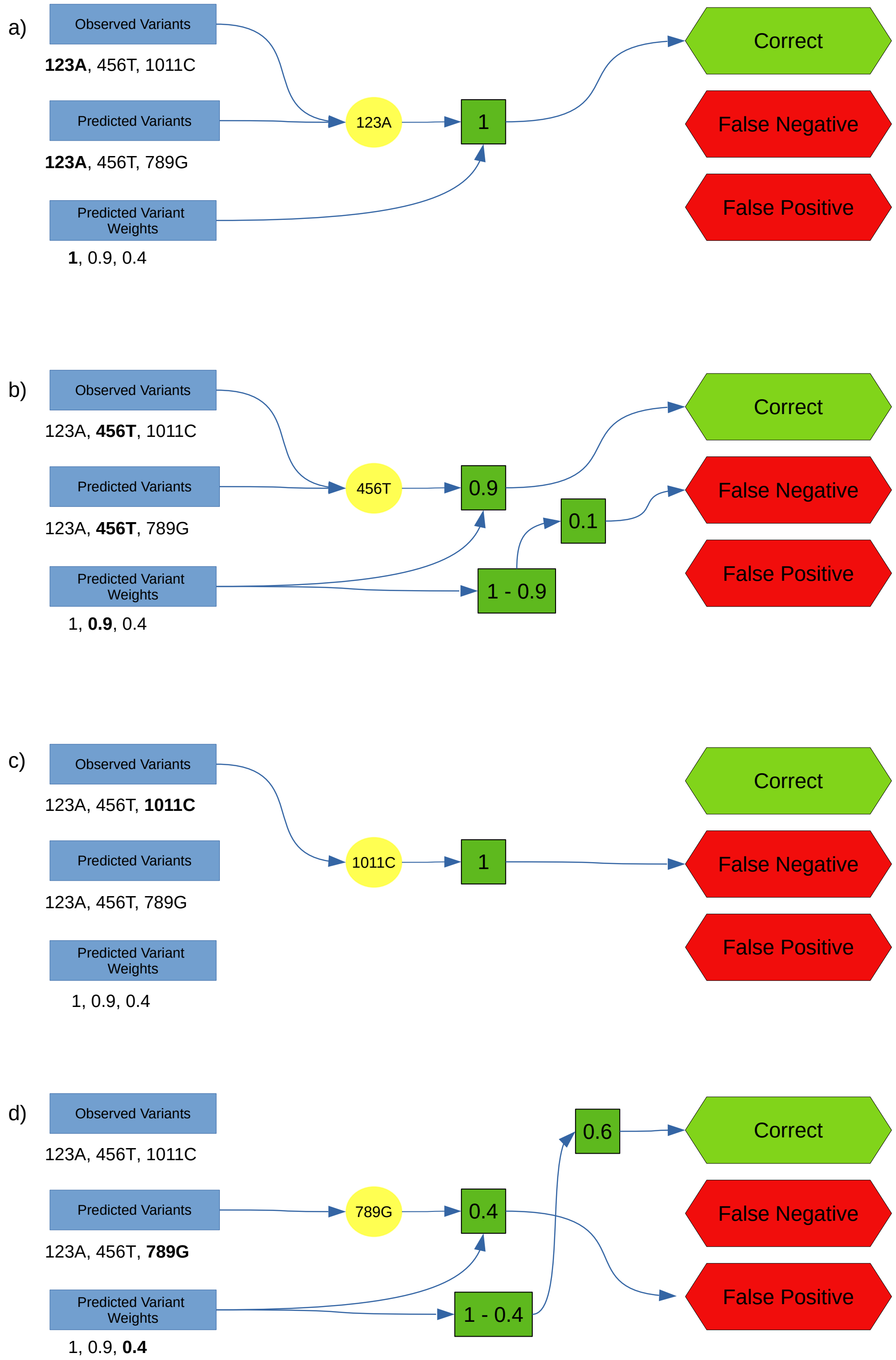


Figure 2.6: A demonstration of the allocation of weightings for a fictional library-test sample with three observed variants (two shared with the prediction and one false negative). The prediction contains two variants which are found in the observed list, and a false positive. The four diagrams show how the prediction weightings are assigned when assessing variant predictions when a) prediction confidence is 100% and correct, b) prediction confidence is less than 100%, c) observed variant is not present in the prediction, and d) predicted variant is not in the observed variant list.

2.3.5 Genome-Wide Variant Prediction Analysis

Variant predictions are also analysed for genome-wide correlation. The mtDNA divides into coding/non-coding, target loci/non-target loci, and homogenous regions/varying regions. These base types are checked to ensure that their prediction accuracy and error direction (false positive or false negative) are consistent.

In order to gain a context for the variant analysis, each variant is assessed for a number of relevant variables:

1. **Position**; to allow mapping,
2. **Protein coding status**; to compare relative success between the two, differently behaving environments,
3. **Variant Type**; SNP, insertion or deletion,
4. **Local homogeneity**; to explore the effect on prediction accuracy of DNA in relation to low variability within the library samples, and
5. **Local rate of base variability**; to explore the effect on prediction accuracy of DNA in relation to high variability within the library samples.

To enable a measure of how likely the variant was to be a private mutation, and therefore unpredictable, local homogeneity was estimated. Whilst the prediction of homogeneous loci is a pointless exercise, having an idea of the level of local homogeneity for each base pair is a useful variable. This was created using 21 base pairs made up of the locus; the ten base pairs downstream and the ten base pairs upstream.

The addition of a measure of variability was made to capture how the predictions performed when the base pair was a mutation warm or hot spot. This would lead to several variants linking to the same locus, more options for that base pair and, potentially, a lower accuracy.

There are four measures of each predicted base:

1. **Total**: The cumulated weights of each time the variant is present in an observed or expected variant list in the library-test samples,
2. **Correct**: The cumulative weights of all the correct predictions for this variant. This comes from the *correct weightings* and *1 - incorrect weightings*. These two routes to correctness represent how a correctly predicted variant, with a weight of 0.75, is only predicted by 75% of the compatible samples, so wrongly predicted absent 25% of the time.
3. **False Positive**: The cumulative weights of incorrect guesses, predicted but not present, and
4. **False Negative**: The cumulative weights of incorrect guesses, present but not predicted.

Diagram 2.6 illustrates the assigning of weightings.

2.4 Generating Predictions for the UKBiobank Samples

The UKBiobank samples' barcodes can be used, one-by-one, as bait gathering matching library samples. Taking each UKBiobank sample and performing a regular expression search for matches in the library, I can make a prediction of both haplogroup and variants based on these matches.

The UKBiobank sample's barcode is a text string which is used in the Grep search function of the library samples' barcodes, finding a list of matching samples. The barcode's text string may include some target loci to which it was impossible to assign a "0" or a "1". These calls of `neither` are included in the barcode as ".", which allows regular-expression-compatible matching to anything.

Having gathered a number of library samples which produced the same barcode as the UKBiobank samples "bait" barcode, two predictions are formed; the UKBiobank sample's (a) haplogroup(s) and (b) variants which differ from the rCRS.

Firstly, to create a haplogroup prediction for the UKBiobank sample, the group of matching library samples will have a set of haplogroups. This forms, what is hoped to be, a very repetitive list, where many of the matching library samples agree on the haplogroup predicted for the UKBiobank sample. This list of haplogroups is reduced to an "unique" list, one without repetition. Each member of this unique list is given a weight; the proportion of matching library samples with this exact haplogroup. The level of repetition for each haplogroup or variant is used to give weight or confidence to the predictions built.

To produce the prediction of a UKBiobank's sample variant list, a similar process is repeated with the variants listed for each of the matching library samples. A full list is collated of all the variants present in the list of library samples with a matching barcode. This is reduced to a list of unique variants and their associated weight, again a proportion of samples in which this variant appears.

2.4.1 Finding the UKBiobank Coverage Rate

Moving from a theoretical failure rate to explain the performance in the library-test data to the performance of the experimental data, will confirm or remove the suspicions that problems remain in the *in silico* genotyping processing.

The UKBiobank experimental data barcodes are employed as bait to find the number of matches in the full set of library samples. Any significant numbers of barcodes failing to find any matches in the library samples would point to certain variants triggering certain loci to be assigned differently, that the algorithm does not reflect the physical microarray's assignment.

Chapter 3

Results

3.1 Data Sources and Structures

3.1.1 MitoMap Haplogroup Library

MitoMap's mtDNA library of sequences are used for *in silico* genotyping and training data as they offer an all-important link from barcodes to both haplogroup assignments and potential variants. Each sample is listed as a set of mtDNA differences from the template rCRS and a haplogroup membership.

Descriptive Statistics

The MitoMap Haplogroup library data set contain 45,882 samples with variant and one of the 4947 haplogroups assigned to each sample. The haplogroups can be represented by several samples, ranging from 531 single-sample haplogroups through to a group of 600 samples all assigned the same haplogroup string. The mean number is nine samples and the median is 4 samples per haplogroup.

Haplogroup names fall into 41 macrobranches, which form a dendrogram rooted at "mtEve", see figure 1.3 which uses a structure based on the relationships defined by Phylotree [100]. These macrobranches contain between 8 and 6167 samples, with a mean of 1092 samples per macrobranch. The macrobranches contain 118 haplogroups on average, ranging between 2 and 934. The number of subgroups a macrobranch contains correlates well with its sample number (Pearson's correlation coefficient of 0.95), see figure 3.1.

The genetic sequence of the samples in the MitoMap Haplogroup library data set is recorded as lists of variants, places where the sample differs from the rCRS sequence. The number of differences from the rCRS sequence varies from 1 to 108. The average number of variants from the rCRS is 36 but shows variation between groups according to their distance from the rCRS.

The names of the haplogroups and macrobranches fail to communicate the relationships within the structure so an approximate measure of genetic distance would be a proxy for the dendrogram. Figure 3.2 is coloured to highlight the approximate distances from the rCRS, measured by counting the nodes between the sample and the H node, where the rCRS sits.

Figure 3.3 shows the number of nodes from the rCRS in the dendrogram 3.2 and the number of variants recorded for the sample, which are differences from the rCRS template sequence in Macrobranch H. The two measures have a strong correlation, except at level 7. This level

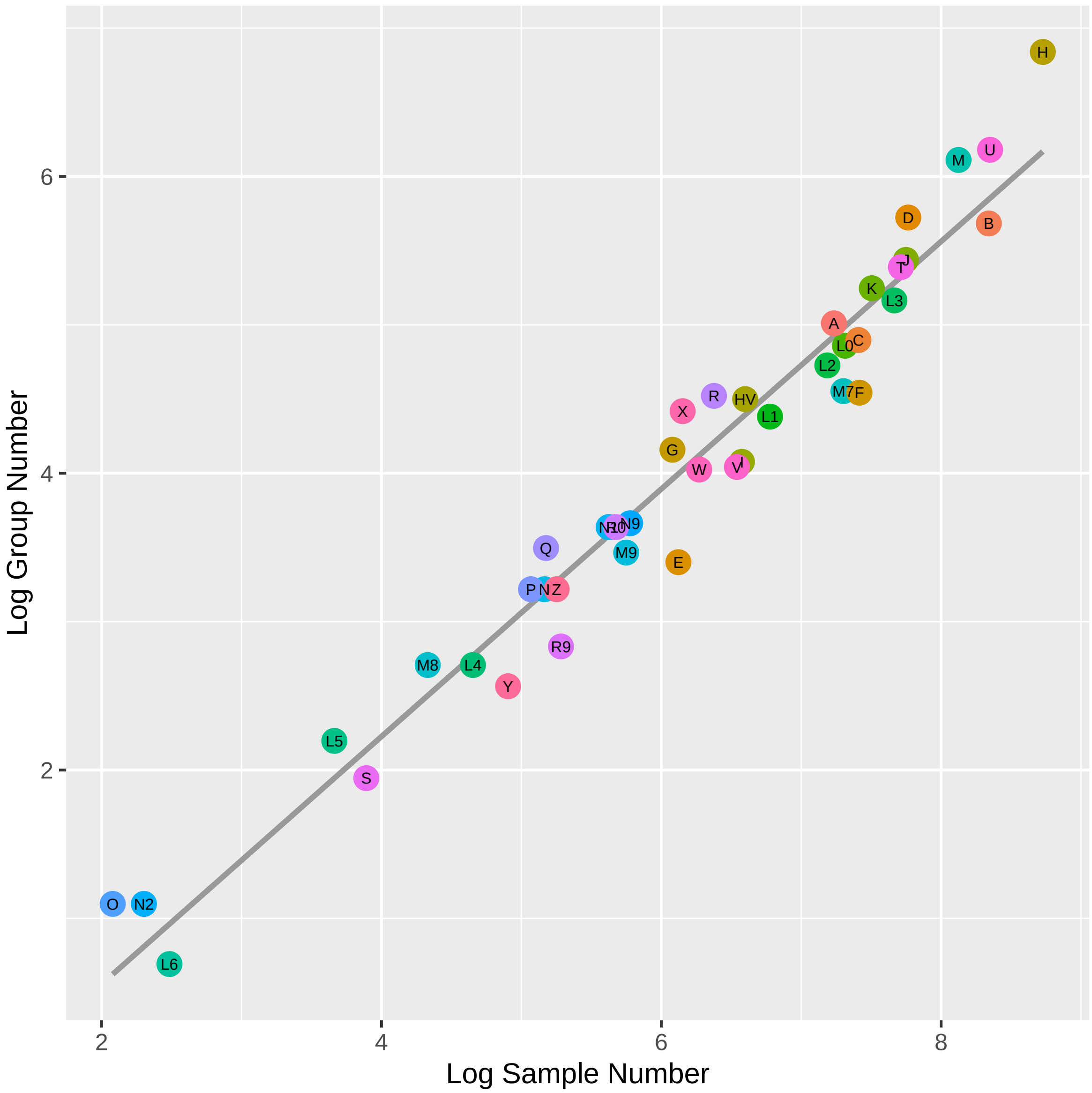


Figure 3.1: Comparison of two descriptive statistics of the MitoMap Haplogroup Library's macrobranches; the number of haplogroups in the MitoMap Haplogroup Library which belong to the macrobranch, and the number of samples in the MitoMap Haplogroup Library which belong to the macrobranch. A log-log to avoid over-plotting at point of origin

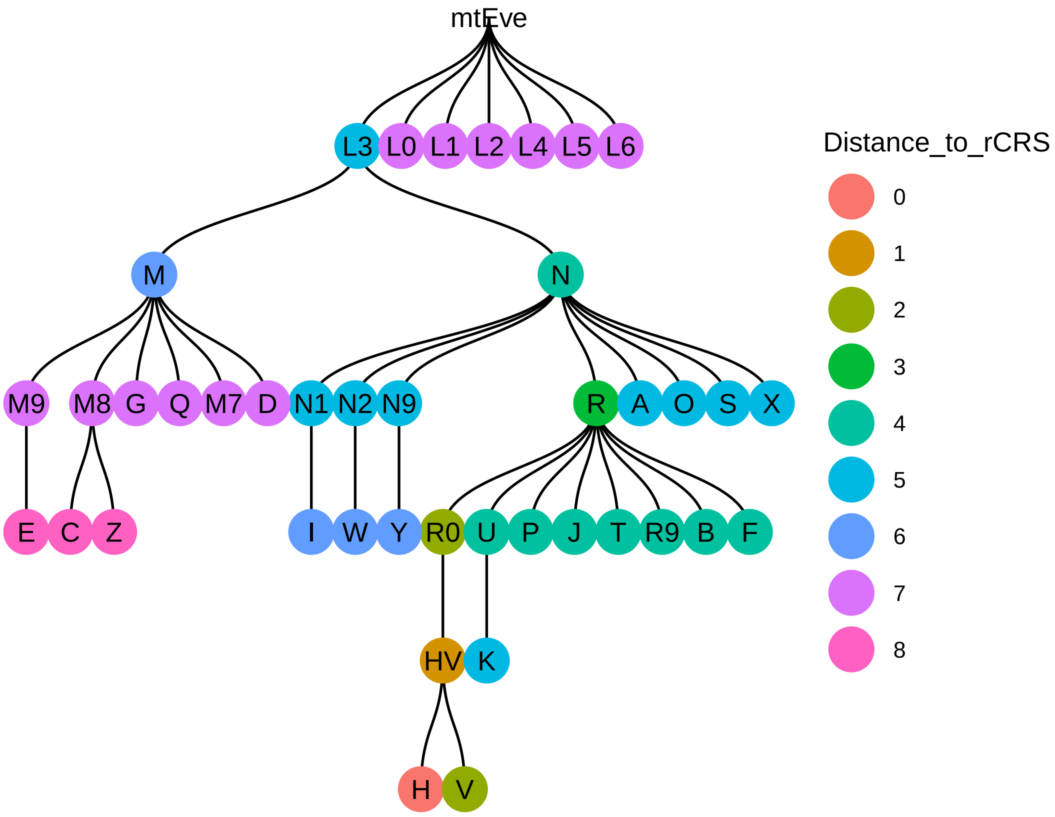


Figure 3.2: Macrobranch nodes coloured by the number of nodes from the rCRS in Macrobranch H for each Macrobranch

includes samples located to the macrobranches M7, M8, M9, G, Q and D (coloured blue) and the L macrobranches L0, L1, L2, L4, L5 and L6 (coloured green). The samples showing unexpectedly high levels of variant number are purely those coloured green. The Pearson correlation coefficient for all the samples is 0.737, but is improved greatly on the exclusion of the green points to 0.869.

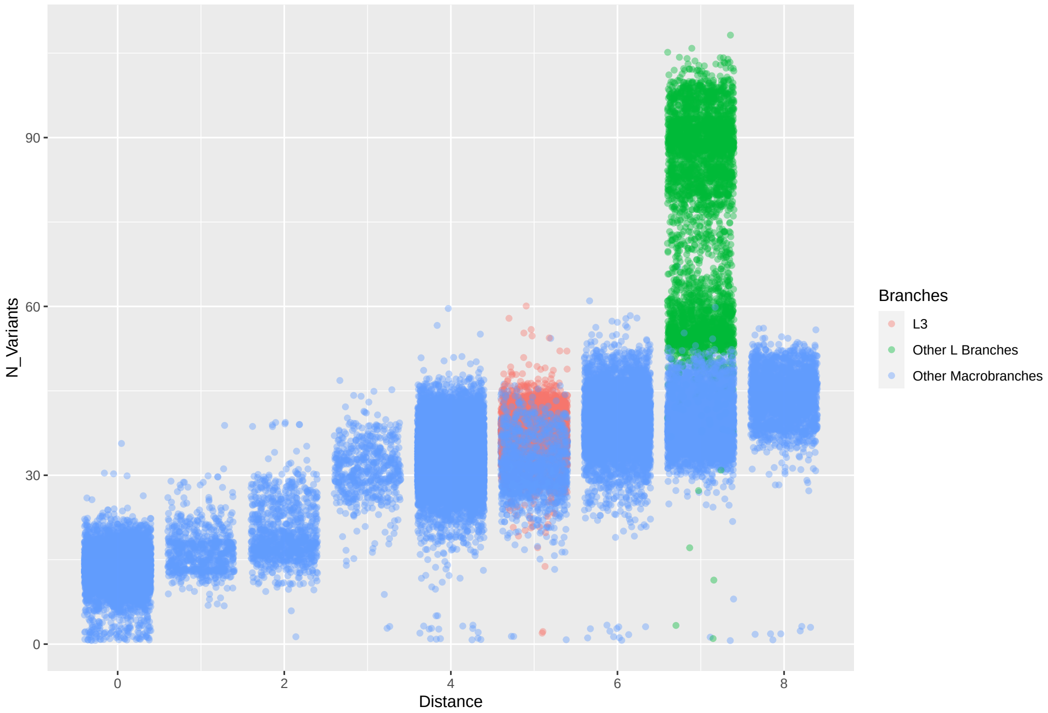


Figure 3.3: Jitter plot of MitoMap Haplogroup Library samples' macrobranch's distance to the rCRS (x-axis) and the samples' number of variants, an independent measure of distance to the rCRS

Despite the samples in macrobranches eight nodes from the rCRS being further according to the tree's structure, they hold fewer variants than the samples marked in green on figure 3.3. The green samples are not a change in behaviour which continues to a distance of eight, but a single set of anomalous results. The tree of macrobranches fails to describe the library samples' variation well, particularly at a node distance of 7. Node 7 samples are further investigated below.

Taking just the L macrobranches, figure 3.4 highlights how diverse these groups are, and that several of macrobranches contain two or more subgroups, with distinctive variant numbers.

Figure 3.4 also reveals some of the complexity hidden within the L macrobranches. L3 is the progenitor of all the other branches, reflected by low variant numbers, but there appears to be further structures in the other L branches. L0 has a single mode, L1 and L2 have several bands. L4, L5 and L6 are woefully underpopulated given their huge number of variants, leading to an expectation of their very poor prediction using this data set.

The number of variants per sample reveals complexity in the simple dendrograms, and the positions of the variants the samples possess uncovers further details about the genome under investigation. There are samples at the very low bounds of each column in 3.3 and hold very low in variant numbers compared to the others at the same distance from the rCRS. Using thresholds of greater than 2 nodes distance and fewer than 7 variants, 44 MitoMap Haplogroup library

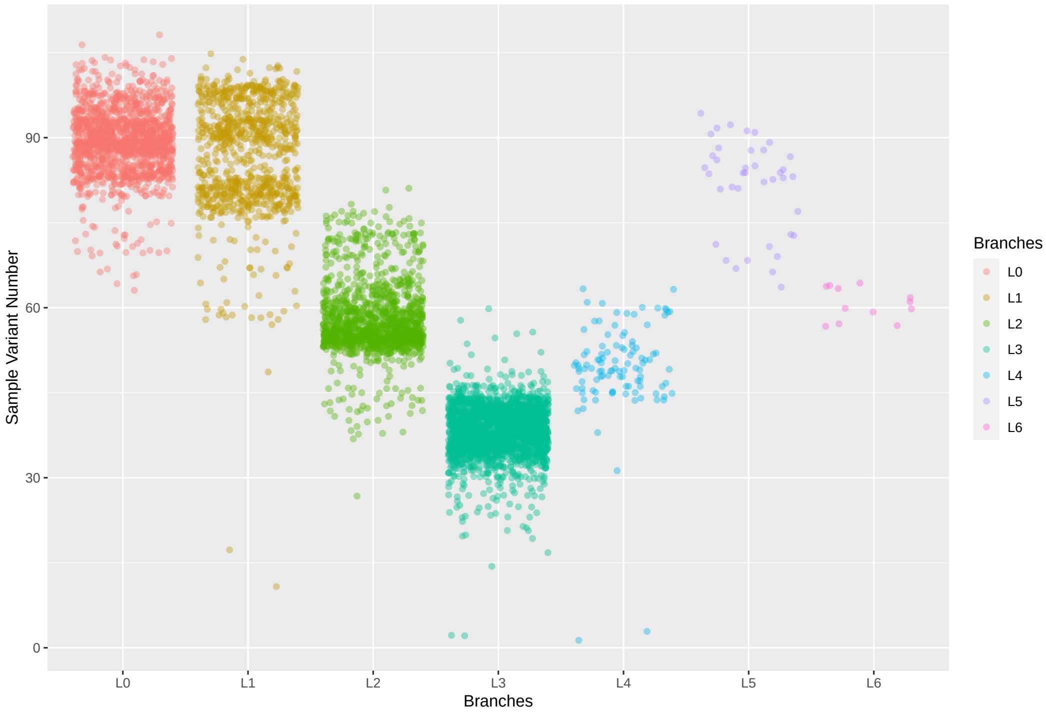


Figure 3.4: Jitter plot of variant Numbers in the MitoMap Haplogroup Library, against approximate genetic distance from rCRS for each L Macrobranch

samples are found, which fall into 33 different haplogroups. Taking each of these haplogroups in turn, all MitoMap Haplogroup library samples in the haplogroup can be compared. In every case, the variant numbers for the samples are above 20 except for the anomalous sample(s) in the group, which is fewer than 4. This confirms that the samples are anomalous in their haplogroup and should have been removed.

Even a conservative estimate finds that 8 of the MitoMap Haplogroup library samples with anomalously low variant numbers have been selected for the library-test set. However, more significant problems arise due to the samples which remained to form the library-training set. These bad samples appear in the predictions of 1357 of the library-test samples. A brief exploration of the where the effected samples appear in plots of the test scores shows the problems to be limited to a small number of macrobranches.

3.1.2 UKBiobank

Descriptive Statistics

The UKBiobank has typed over 490,000 individuals at 265 positions on the mtDNA, equating to 1.6% of bases in the mitochondrial genome being genotyped. Each locus on the mtDNA is recorded as `reference` (identical to the `reference` sequence), `alternate` (a single, specified `alternative` to the reference base) or `neither` (the base is neither `reference` nor `alternate`). The resulting array contains a very partial representation of each sample's genome representing a very variable amount of information.

Quality Control

The rates of neither calls for the 265 UKBiobank loci were found to fall into three strata at approximately 90% (22 loci), 10% (85 loci) and 0% (158 loci). A cut-off of 85% was used to cleanly excise the 22 worst loci.

A subset of samples were found to have *neither* calls in all of the 85 loci in the 10% stratum. Removing all samples with a score of *neither* calls greater than 50 excised this subset almost exclusively. Using this threshold also removed a further 72 samples with *neither* calls of up to 161, suggestive of generally poor performance.

Homogeneity in the MitoMap Haplogroup library was also checked for. Three loci were only finding reference-assigned variants in the entire MitoMap Haplogroup library and these loci have been excluded from the list. Having begun with 265 loci typed over the two microarrays, the final number used for barcoding was 240, see appendix 4.1.

The final UKBiobank experimental data set held 240 loci's data for 438,355 individuals, once samples with high level of missingness were removed from the data set.

3.2 In Silico Genotyping of MitoMap Haplogroup Library Data

3.2.1 Stage One: Collating MitoMap Haplogroup Library Variants

Loci Variant Analysis

The number of listed variants linked to loci varies from 0 to 21. These may be the three types of SNPs which differ from the reference base in that locus of the rCRS, but also any deletions or insertions located to this locus.

The variant number per sample is indicative of the sample's distance from the rCRS but variants also vary in their library population frequency. The number of appearances a variant makes in the library was counted and, on average, variants are present in 145 (0.03%) samples in the library, 3159 (27.8%) variants are found just once, and one variant (15326G) is found in 45,448 samples (93%).

Over 47% (7808) of the 16,569 bases have no variants present in the library, so are entirely homogenous in every sample. Of the 8761 variant-bearing bases, 78% (6835) hold only one variant and over 99% hold 3 variants or fewer. As for the maximum, one base in the notoriously rapidly changing D-loop, at 309bp, holds 21 different variants all located to that locus. The 21 variations are "309CCT", "309CCCT", "309CCNT", "309T", "309CC", "309CCTY", "309d.CT", "309CCCCT", "309CCCNT", "309CCN", "309CCCN", "309CCTT", "309CCC-CCT", "309CGCT", "309CNT", "309CNNN", "309CNN", "309CCCY", "309CCCCCTCCCCT", "309CCNN" and "309CCY". The nature of these 21 variations suggest DNA polymerase copying issues, either in vivo resulting in true additions to the sequence or in vitro, as the sequencing reaction stutters preventing clear results. A call of "N" represents the knowledge that there is no gap in the sequence but no further information about the base is confident. The "Y" represents either a "C" or "T" base, with equal likelihood.

The mtDNA is 92.3% coding DNA. Coding mtDNA is 48.9% homogeneous, so approximately half of all coding bases are identical in all human mtDNA sequenced and stored in MitoMap. In contrast, only 19.4% of non-coding mtDNA bases are homogeneous, over 80% vary in one or more individuals in the library.

From the 45,882 samples in the library, a total of 11,368 distinct variants were found. A combination of SNPs, insertions and deletions allows a list of as many as 21 variants addressed to a single base pair locus. Deletions make up 258 (2.3%) of the variants and insertions another 445 (3.9%). The remaining 10,665 (93.8%) variants are single nucleotide polymorphisms (SNPs) and much easier to categorise. The frequency with which these three library variants appear varies enormously, with a bias favouring the SNPs. The library holds records of 1,643,871 variants in its samples. Of these, deletions are 24,820 (1.5%), insertions number 38,769 (2.4%) and SNPs the remaining 1,580,282 (96.1%).

Variants which are deletions or insertions create challenge for *in silico* genotyping. Deletions range in length from 1–14 base pairs and insertions range from 1–20 base pairs. Taking this into account, plus a moderate effect range around the locus, the number of SNPs, deletions and inserts with influence over the UKBiobank target loci can be found. Of the 11,368 library variants, 1786 (15.7%) are in range of at least one of the UKBiobank's target loci. Deletions account for 52 (2.9%) variants within range, insertions for 61 (3.4%) and SNPs for the remaining 1673 (96.7%).

3.2.2 Stage Two: Linking to Loci

The target loci are the data points from the UKBiobank microarray which have been retained after the quality control. They represent 240 base pairs targeted by the microarray which need to also be the target of the *in silico* genotyping of the MitoMap Haplogroup library samples.

A chain of steps laid out in the previous chapter produces a table which links the two data sets by; (a) finding the variants relevant to each locus, and (b) classifying the variants as `reference`, `alternative` or `neither` compared to the rCRS sequence at that locus.

In silico genotyping links all target loci with the MitoMap Haplogroup library variant or variants which represent non-`reference` calls; either `alternative` or `neither`, (see table 2.2). Every target locus received a single `alternative`-linked variant bar one at 547 bp. For this SNP, the `alternative` base choice (an A to T change) on the microarray is not present in the library or the UKBiobank samples, some half a million participants. There are, however, samples in both data sets showing a `neither` call (an A to G change) suggesting there might have been a design error in the choice of `alternative` base.

For `neither`-linked MitoMap Haplogroup library variants; a total of 153 loci received none, 69 loci received a single variant, 14 received a pair, one locus received three, two loci received four, and one locus received nine `neither`-linked variants. This means that 153 of the 240 target loci (63.8%) will not be assigned a `neither` call in any of the MitoMap Haplogroup library barcodes, under current conditions.

The coding of the variants prevents the same locus appearing twice in a variant list. Despite there being a list of `alternative` and `neither` options at some loci, it should not be too complex to decide on the locus outcome based on any variant list. The process can be illustrated with an example; the locus at 16,193bp, which records one `alternative` (16193T) and nine `neither`-linked variants (16193CC, 16193CCC, 16192CT, 16193CCCC, 16193CTC, 16192CY, 16193A, 16193CTT, 16192C). Just as expected, although the list of relevant variants is long, the samples cannot hold more than one. Of the 45,882 library samples, 44,740 contain no variants relevant to this locus (i.e. are `reference`), 610 samples hold only the `alternative` variant, and the remaining 532 samples hold one of the neither variants. None of the samples hold more than one neither variant, or both the alternative and a neither.

All UKBiobank loci were expected to find at least one relevant alternative variants because of; (a) the large size of the library, (b) the library's representation of a global population, and (c) the choices of loci to include on the UKBiobank's microarray. Even UKBiobank's attempts to find rare, fatal variants in their population, should appear in the published data gathered in the MitoMap Haplogroup library data because their publication is a matter of clinical interest so life-limiting variants will be present in samples gathered by MitoMap, if not in the older population UKBiobank participants.

Comparing Variant Frequencies

The rate at which a target locus receives a call of alternative is variable. This locus call rates of each target locus should correlate between equivalent data sets. Correlation would be a strong indicator that *in silico* genotyping mimics the microarray's behaviour well. Some differences would be expected due to the data sets' populations, but these should be low in absolute terms.

There are 15 loci where the UKBiobank microarray has found no samples with a call of alternative (235, 547, 3395, 3460, 3946, 3949, 4160, 6386, 8344, 12950, 14094, 14178, 14318, 14550, 14552). This is not reflected in the samples from the MitoMap Haplogroup library, where there are samples holding alternative variants in all cases bar 547.

The average magnitude difference between the rates is 2.7% over the 240 target loci, with 11.3% of the loci sitting at more than 5% rate difference, and 4.5% of the loci beyond 10% difference.

Figure 3.5 compares the rate of alternative call for each of the 240 loci in the two data sets. A call of 0% represents a 1:1 ratio, and the red thresholds show a rate of ± 10%. All but ten loci lie close to a 1:1 ratio, with nine loci being relatively enriched in the library and a single locus being enriched in the UKBiobank data.

Plotting the direct comparison, as in figure 3.6, reveals the loci with the smallest difference in rate are those at the extremes of rate, either overwhelmingly reference or overwhelmingly alternative calls. This suggests that the current threshold, again in red, may not be optimal at finding the comparatively enriched loci.

The ten problem loci which have an alternative call rate of more than 10% may be due to differences in the sample population. This cannot be accounted for without knowing the haplogroup of the participants in the UKBiobank (which would render this exercise null), however, if the alternative rate calls are actually symptomatic of mis-assignments of variants during the data processing stages, their assignment directions must be corrected.

3.2.3 Stage Three: Barcoding

In the MitoMap Haplogroup library, 7276 unique barcodes are found. Each barcode has a mean of 6.3 (0.013%) samples, with 4,423 (61%) barcodes containing just a single sample and one barcode encompassing 1129 (2.3%) samples.

The barcode groups contain 1.7 haplogroups on average, with 6058 (83%) barcodes containing a single haplogroup. The barcode group containing the most haplogroups had 189 (3.8%) of 4947 haplogroups represented in the library.

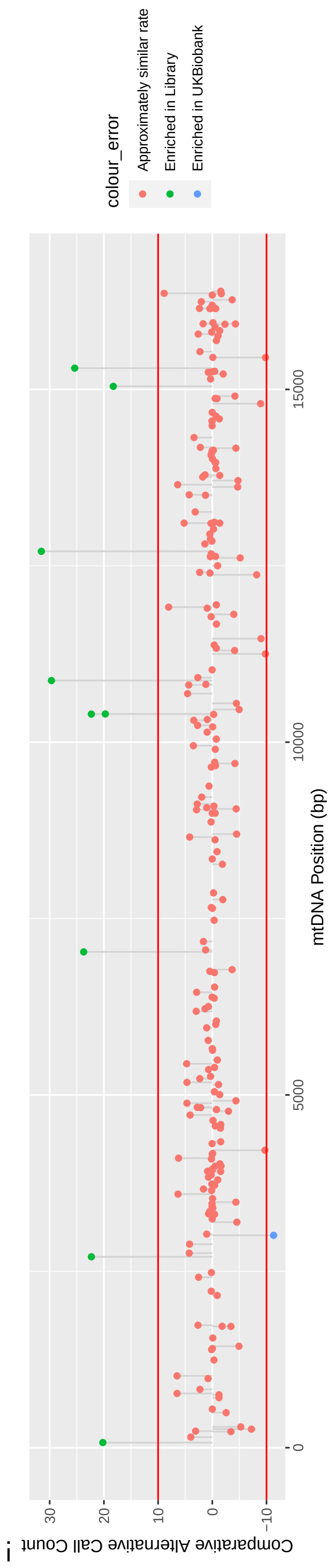


Figure 3.5: Rate difference of `alternative` Call between the UKBiobank and the MitoMap Haplogroup Library samples (%), coloured by biases in enrichment. The red lines represent a difference of rate of 10%.

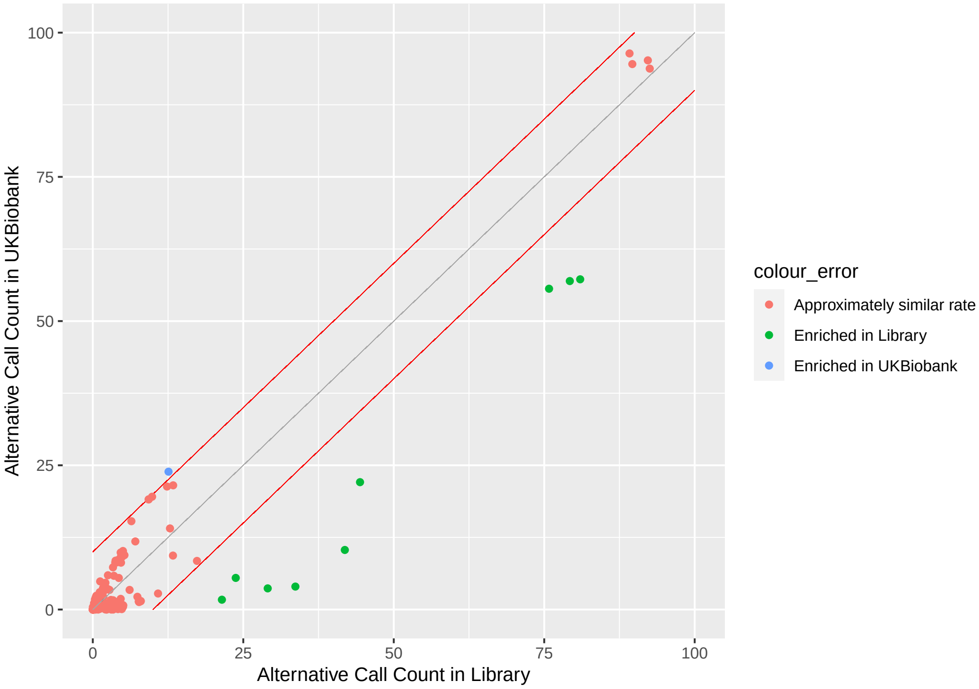


Figure 3.6: Comparison of `alternative` Call rates of the UKBiobank and the MitoMap Haplogroup Library barcodes (%). The red lines represent a difference of rate of 10%

3.2.4 Stage Four: Forming Predictions

The MitoMap Haplogroup library-test set were used as bait in a regular expression search for compatible matches in the MitoMap Haplogroup library-training set. Nearly 10.2% (1017) of the library-test samples found no compatible matches in library-training set. The mean match number was 196.5 and the median match number was 57. A total of 358 (3.6%) samples had the maximum number of matches of a 1188, discussed below.

Each member of the library-test set has a group of compatible library-training samples from which a set of haplogroups can be extracted. This is condensed into a list of predicted haplogroups and a corresponding list of weights, which sum to one as the test sample may hold only one haplogroup.

Variant predictions are formed in a similar way by collating the compatible library-training samples' variant lists and reducing this central list down to set of possible variants and a second corresponding list of weights. This may sum to much higher than one as library-test samples can hold many variants.

The 385 library-test samples getting 1188 library-training matches mentioned above were found to form a good cluster with each of the observed haplogroups starting with "B4a1a". They were expected to have very poor predictions built from a diverse groups of matches. Their library-training matches also formed a cluster, with all matches beginning "B4a1" except one from haplogroup "R0+16189". This was reflected in the good test 2 scores for the library-test samples (mean 0.891, compared to 0.696 overall), good test 3 scores (mean 0.669 compared to 0.369, and 0.615 compared to 0.349, respectively), but Recovery100 scores of 0 due to the single "R0+16189".

3.3 Validation with a Test Data Set

3.3.1 Source of the Test Data

Sample-by-sample, each of the 10,000 library-test barcodes are used as a “bait” for the regular expression search in the 38,882 library-training, “pool” samples. The matches returned are combined into predictions for the library-test sample.

The regular expression searching using the bait failed to find any matches in 1017 (10.2%) of the library-test samples' pool of barcodes.

The test set used is a randomly chosen subset of the library. Dividing the 45,882 library samples into a “library-test” set of 10,000 (20.5%) and the remainder, referred to as the “library-training”, will be used to form predictions for the test samples.

3.3.2 Failure Rates

As a portion of the library-test samples fail to find any matches in the library-training set, having an estimate of an expected level of matching failure is a useful place to begin. If the failure rate is approximately what we expected, then we have evidence that the method is working well in the vast majority of cases.

Calculation of an Expected Failure Rate

Whilst the library data set is large, barcode match groups can be small. If all of a barcode match group happen to be selected for the library-test group, those samples will fail to find any matches in the library-training set.

In order to ascertain if the failure rate found was due to the division of the library data, several measurements were made. By repeating the selection of 10,000 (21.8%) library-test samples and their attempted matching with the remaining library-training samples, the variability of the failure rate could be found. A total of 15 cycles were completed as three training-test ratios (5,000 (10.9%), 10,000 (21.8%) and 20,000 (43.6%)) were each processed five times. Whilst the five-fold replication showed a tight distribution, the three training-test ratios showed the observed data followed the curve predicted by theory.

Using approximations I have calculated an estimate of how many MitoMap Haplogroup library-test samples are expected to fail just because their match partners are all in the test set too. This line is plotted in red in figure 3.7. The cyan points mark the repeated sampling-matching of MitoMap Haplogroup at the three test set sizes.

Plotting Match Number

Each MitoMap Haplogroup library sample has a number of potential matches in the entire library.

All of the 1017 MitoMap Haplogroup library-test-fail samples had a potential match set sized smaller than 3, with 88.7% of the library-test-fails having a potential match number of just a single sample. 10.4% had two potential matches in the library, and 0.9% having three matches.

All five cycles of (a) random sample selection, (b) finding the library-test-fails, and (c) ascertaining the library-test-fail samples' number of potential matches in the entire library, found library-test-fails only had match numbers of four or fewer. In addition, all five curves are very

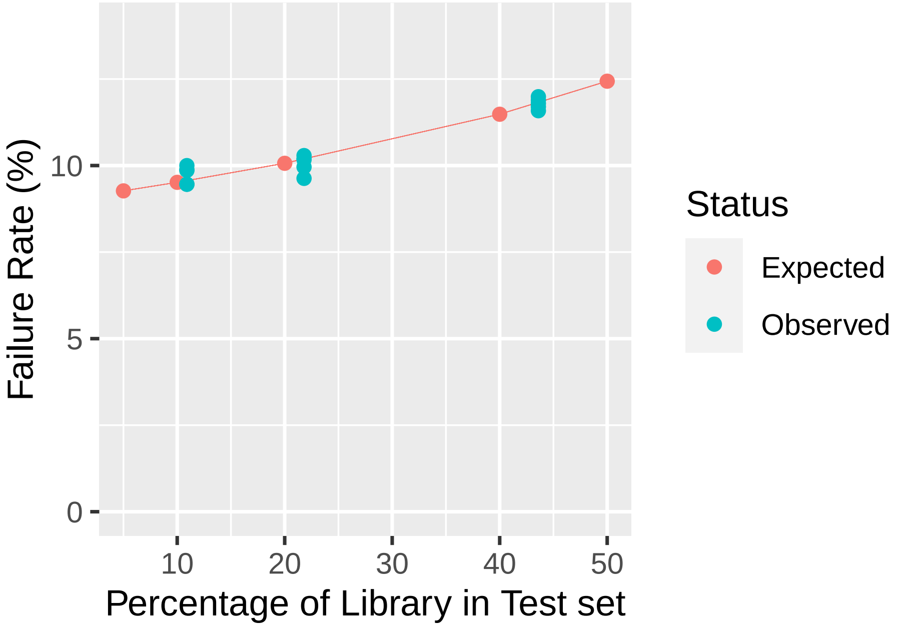


Figure 3.7: Comparing the expected failure rates from theory and the observed failure rates. Sample matching failure rates were measured five times at each of the three different library-test split sizes in blue. Theoretical failure rates are in red and represent the failure rate caused solely by the effects of splitting the data leaving unmatchable samples in the test set

much in agreement of the total at around 1000 fails and proportion; 89.7% have a single match, 8.9% have two matches, 1.2% have three library matches and 0.2% have four library matches.

The numbers of library-fails can also be gathered into their macrobranches. Looking at the failure rate gathered by macrobranch, figure 3.9 shows the failure rate of library-test samples collated by macrobranch. The red line marks the overall failure rate and marks a line where the larger groups seem to stabilise from a group size of 300 or so. The distinct outlier is macrobranch O, with a failure rate of 100%. There are 8 samples which fall into the O macrobranch, one of which was selected for the Library-test set. A closer inspection of the patterns in the eight samples from O macrobranch shows that the sample which was selected for the library-test has three additional calls of `alternative` compared to the four other samples with the same haplogroup string. Whilst comparing the barcodes of samples with the same haplogroup strings risks over-tuning, it should be attempted in future work. This macrobranch is explored at greater depth in the appendix 4.5.

Comparing Passes and Fails

Making a comparison of the library-test-fails and the library-test-passes should reveal if any particular loci have problems. In order to check, a comparison of the rates of `alternative` calls in the library-test-fails and library-test-passes at each of the 240 loci. The rates show an extremely high correlation, with a coefficient of 0.998 and no outliers prompting further investigation.

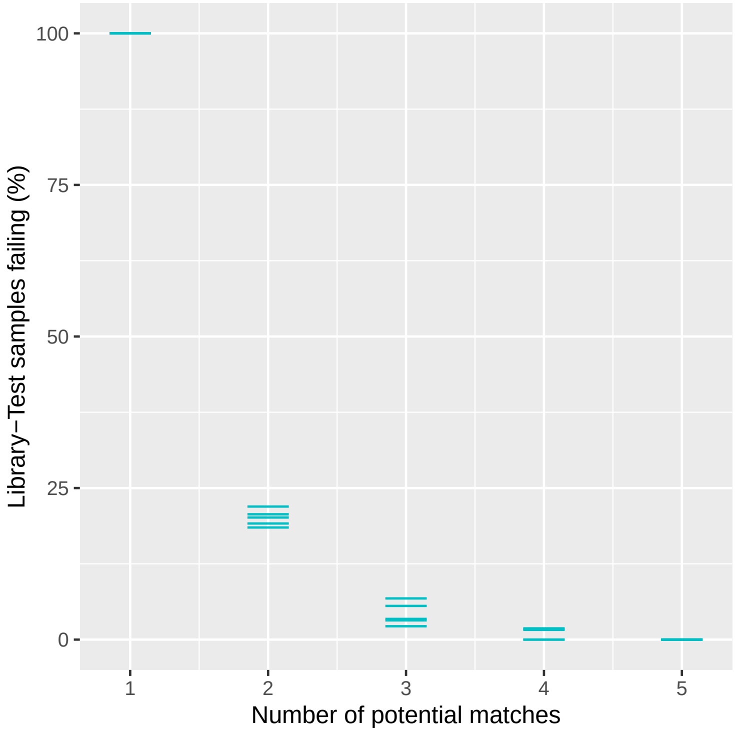


Figure 3.8: MitoMap Haplogroup Library-fail sample numbers grouped by the number of matches in the entire library. The cyan bars show five, randomly selected replicates.

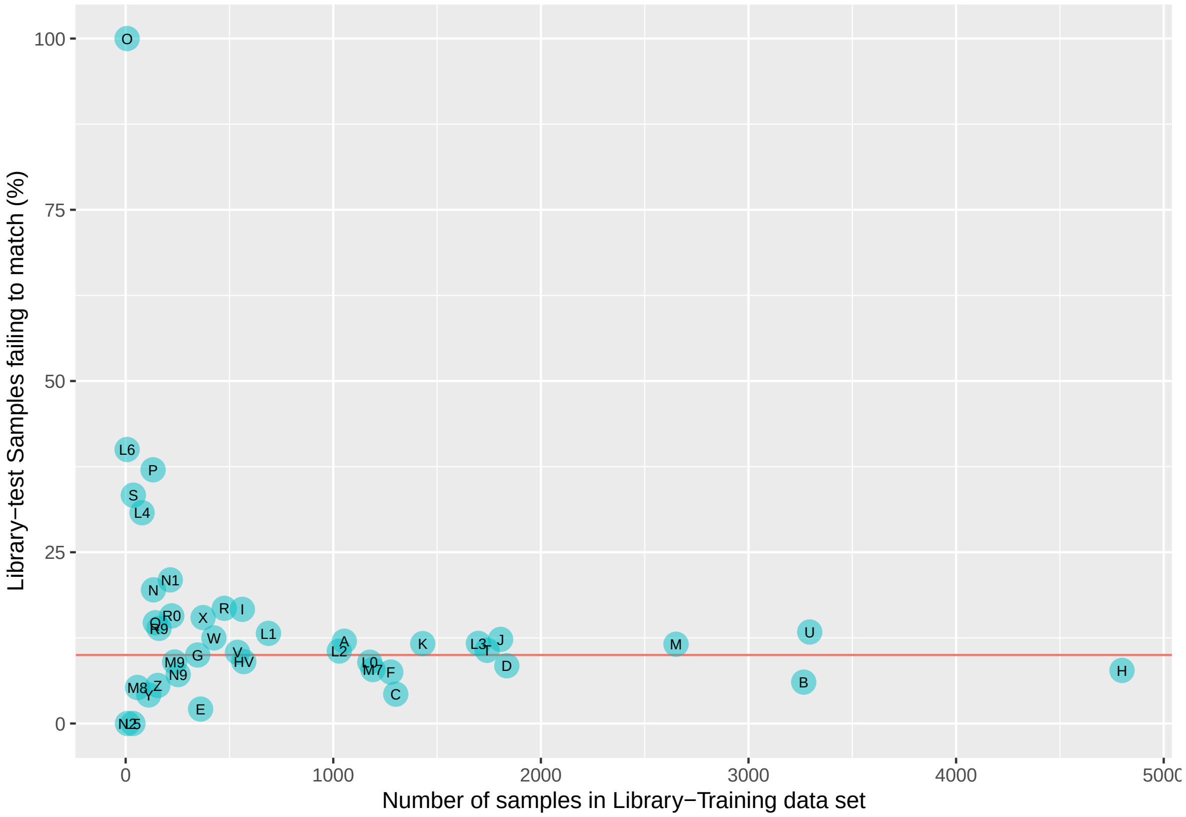


Figure 3.9: Library-Test sample match failure rate against sample number in Library-Training per macrobranch

3.3.3 Quantifying the Predictions' Success

There are many ways of comparing sample's observed haplogroup or variant list with the weighted haplogroup or variant predictions. A measure of correctness depends on the intended use of the data so measuring accuracy has been done in four different ways, see 3.1.

	Description	Mean	Std. Dev.	Histogram
Test 1	Proportion of prediction pointing to observed HG	0.32	0.37	3.10
Test 2	String similarity of predicted HGs to observed HG	0.70	0.32	3.11
Recovery100	Proportion of observed IUs guessed perfectly	0.52	0.36	3.13
Test 3	Proportion of potential score assigned as correct	0.85	0.17	3.15

Table 3.1: The four prediction accuracy tests; details, descriptive statistics and plots

Haplogroup Prediction Accuracy Tests

A score of zero could be received by library-test-fail samples, just through having no predictions, and through a library-test-pass sample having an incorrect prediction. I have attempted to discriminate between these two cases using a black bar of length 1017 to cover the library-test-fail samples. The histograms are also all at equivalent scales to allow better comparison by the reader.

The figure 3.10 shows the distribution of test 1 scores for the 10,000 library-test samples. It can be seen that the test 1 scores are generally low, presumably largely down to the stringency of the test. The proportion of a prediction pointing to the observed haplogroup depends on the frequency of observed haplogroup and the number of haplogroups in the prediction. The samples receiving a maximum score of 1 are likely to be narrow predictions and popular haplogroups. The number of samples receiving a test 1 score of 0 is nearly equally split between the unpredicted 1017 library-test-fail samples, and the library-test-pass samples which received a prediction which did not include their observed haplogroup.

Both the observed haplogroup and the predicted haplogroups can be split into their information units and incrementally built back together. The comparison of incremental information units was used to measure prediction accuracy in test 2 by comparing the information units in the observed haplogroup with the haplogroups of the matching library samples from which the prediction is made. See 2.5 for an example and diagram.

Figure 3.11 shows the distribution of test 2 scores among the library-test samples. The mean score for test 2 is 0.70 with 2257 (22.6%) getting a score of 1. Only 12 (0.12%) samples received predictions and got a test 2 score of 0.

The number of observed incremental information units which are predicted with perfect accuracy is seen in 3.12. Again I have tried to allow the discrimination between unpredicted samples (Library-test fails, in red) and incorrectly predicted samples (in black). As the number of information units in the observed haplogroups varies from 1 to 12, moving from an absolute number to a proportion of units predicted with perfect accuracy (Recovery100) was also used.

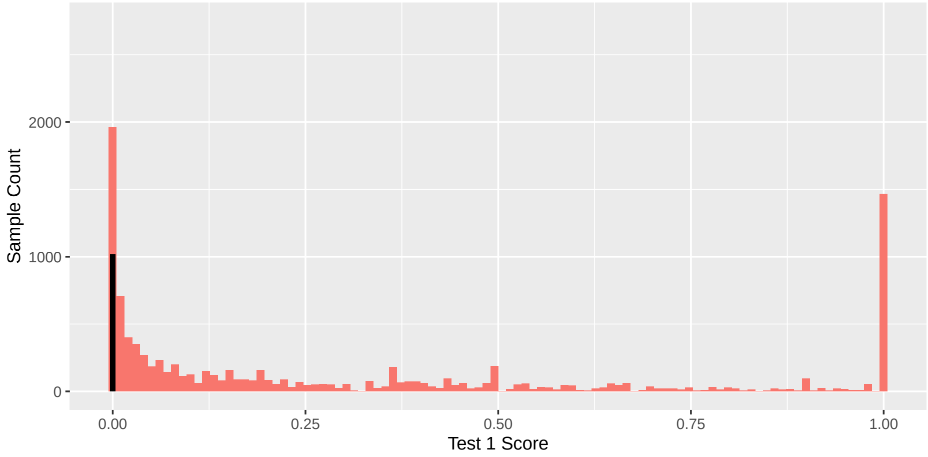


Figure 3.10: Distribution of Test 1 Results in Library-Test samples. The black bar on the far-left represents the scores of 0 which are the library-test-fails, with the remaining red bar representing predictions which have found matches and still scored a zero.

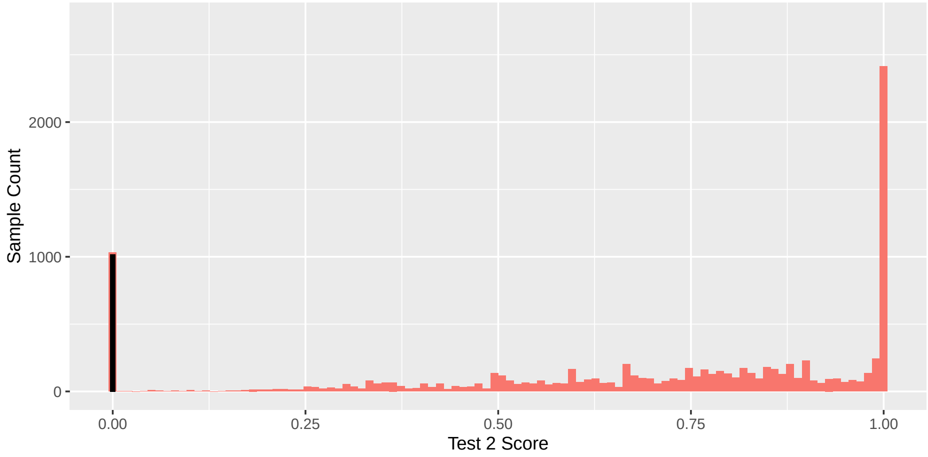


Figure 3.11: Distribution of Test 2 Results in Library-Test samples. The black bar on the far-left represents the scores of 0 which are the library-test-fails, with the remaining red bar representing predictions which have found matches and still scored a zero.

Figure 3.12 shows the spread of the Recovery100 scores. A sample with six information units in its haplogroup string will have between zero and six information units predicted perfectly by all haplogroups in the prediction. The last information unit of the expected haplogroup also present in all of the haplogroups in the sample's prediction is recorded. Figure 3.12 shows this number, grouped by macrobranch and differentiated between the scores of zero caused by bad predictions and zeros caused by a failure to find matches (in red). For example, for a score of 3, all predicted haplogroups have to contain the first three information units which match those of the observed haplogroup.

Figure 3.12 shows the number of information units guessed correctly by all the predicted haplogroups, groups by macrobranch. The failed samples, which would always score a 0, are coloured in cyan.

Rather than absolute numbers of units, a proportion of the information units which the predictions recover gives a better picture of the performance over the macrobranches. This is called Recovery100, the distribution of which can be seen in 3.13. Whilst the shape of this plot is notable for being nearly symmetrical and patterned, this is merely a reflection of the option available to the data; the denominator (a discrete number between 1 and 10) and the numerator (a discrete number between 0 and the denominator).

To gain an overall estimate of the prediction of information units, the mean Recovery100 score is 0.52. This means that the predictions recover 52.2% of the information units perfectly. However, some macrobranch groups show many library-test-pass samples which do not get perfect predictions of their first information unit, meaning their predictions are spread outside of the macrobranch. Because of haplogroup naming inconsistencies, this could not be a significant failure as it may represent matches in the parental macrobranch and it should be investigated further.

Having a sample-wise picture of accuracy from the five tests is important, however, once the samples are gathered into their respective observed macrobranches, a macrobranch-wise picture of accuracy is revealed.

The test scores based on the haplogroup prediction accuracy are compared in figure 3.14. The general behaviour of each test seems to remain constant for all of the macrobranches.

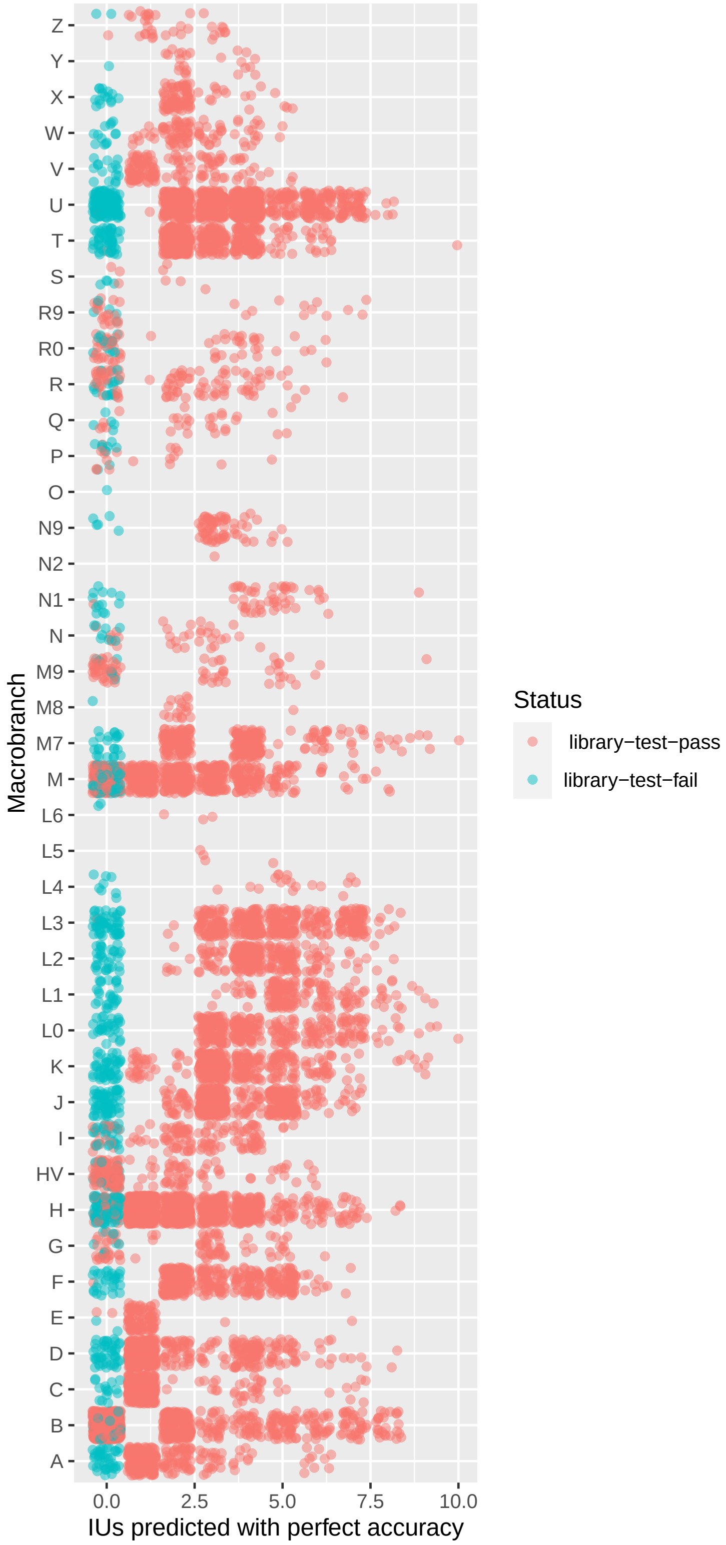


Figure 3.12: Jitter plot of perfect IU guesses by macrobranch in library-test samples, with failed samples in cyan.

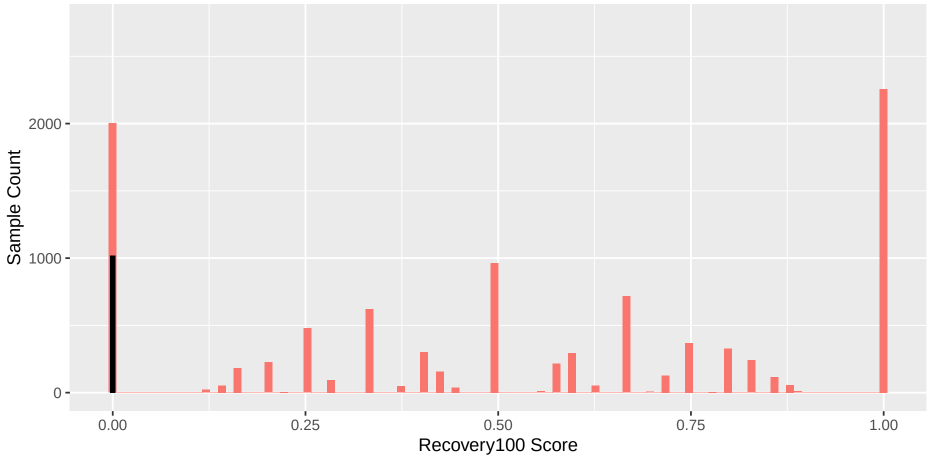


Figure 3.13: Distribution of Recovery100 Results in Library-Test samples. The black bar on the far-left represents the scores of 0 which are the library-test-fails, with the remaining red bar representing predictions which have found matches and still scored a zero.

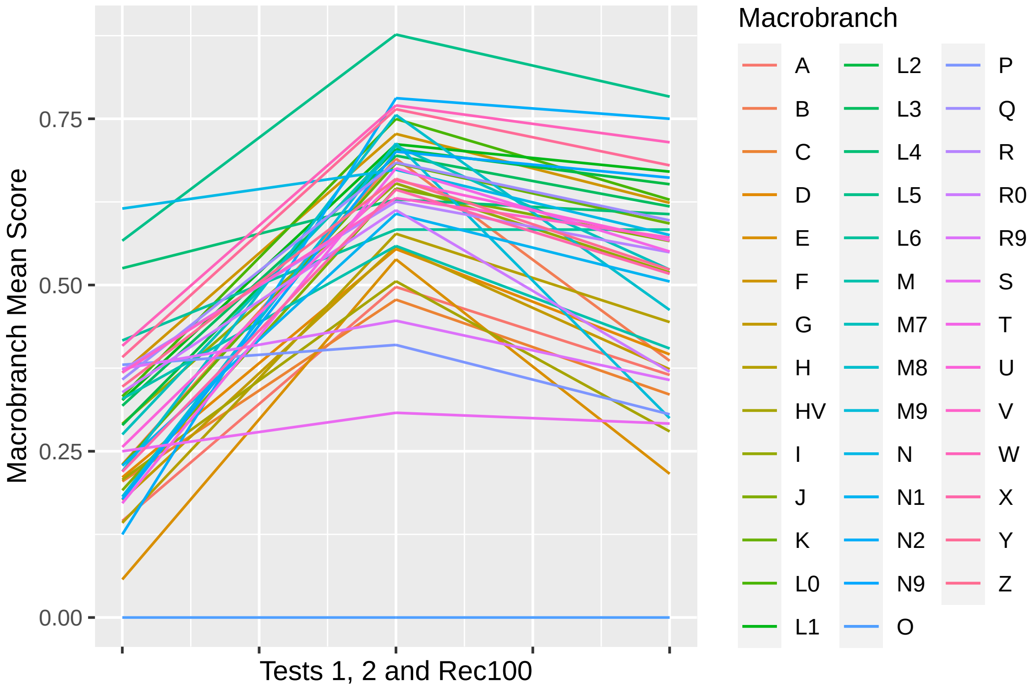


Figure 3.14: Mean Scores for three haplogroup tests gathered by Macrobranch, comparing the behaviour of tests 1,2 and Recovery100. Lines between data points illustrate the relationship between the scores of each macrobranch

Variant Prediction Accuracy Tests

The predictions made for each sample's variants are explored using test 3. Each sample has a list of variants mentioned in either the observed variant list or the predicted variant list. The number of unique variants in these lists represents a maximum correct score. Test 3 represents the proportion of this maximum score assigned as correct and is a good first measure of the accuracy of the variants prediction for a test sample.

The distribution of test 3 scores can be seen in 3.15. This too uses a black bar to show the number of library-test samples which scored a 0 because no matches were found in the library-training set.

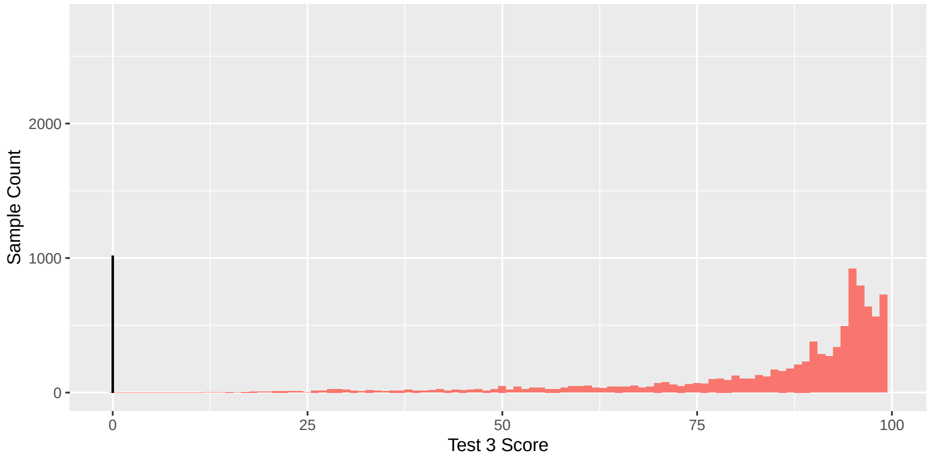


Figure 3.15: Distribution of Test 3 Results in Library-Test samples. Test 3 measures the proportion of the sample's maximum score recovered as correctly guessed.

Gathering the test samples by their macrobranch allows a single mean to be plotted. In figure 3.16 the mean of test 3 scores for all samples on each macrobranch is compared to the tree distance of that macrobranch from the rCRS. The plot shows two possible stories, marked with lines. The red line is suggested by a lower limit of mean accuracy which decreases as the tree distance from the rCRS increases. Alternatively, had H been lower, the data may also suggest that macrobranches L3, L4, L5 and S are outliers from the generally level trend about the mean of 72, marked with the grey bands.

The library-test-pass samples have a mean false positive rate of 9.9% and a mean false negative rate of 10.6%. Figure 3.3.3 shows how the false positive-to-negative rates compare. The low-sitting outliers from 3.3.3 have differing behaviours; (a) P and S have higher than average false positives, (b) L1 has higher than average false negative, (c) L4-6 have high false positive and very high false negatives, and (d) N and Q have slightly raised levels of both.

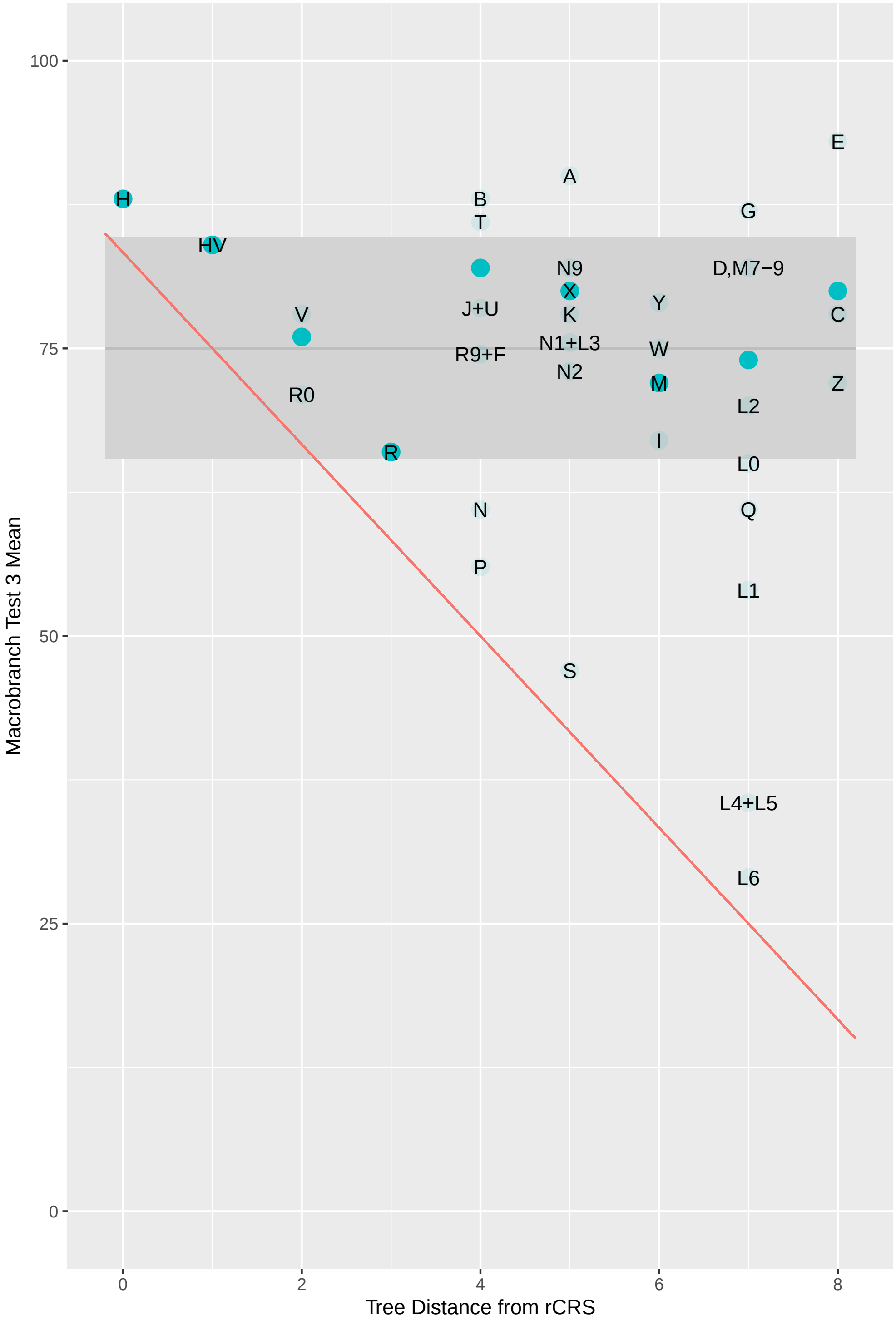


Figure 3.16: Comparison of macrobranch test 3 mean score plotted against the tree distance from the rCRS. Overlapping branch data points, such as D, M7, M8 and M9, are labelled as a single point. The red line highlights a lower bound seen in the data, with a negative correlation between test 3 mean and distance from rCRS. The grey band highlights a second pattern in the data; the data do not negatively correlate, but outliers make it appear so. Cyan points represent the average test 3 score for each distance from the rCRS, highlighting the effect of low sample number on prediction.

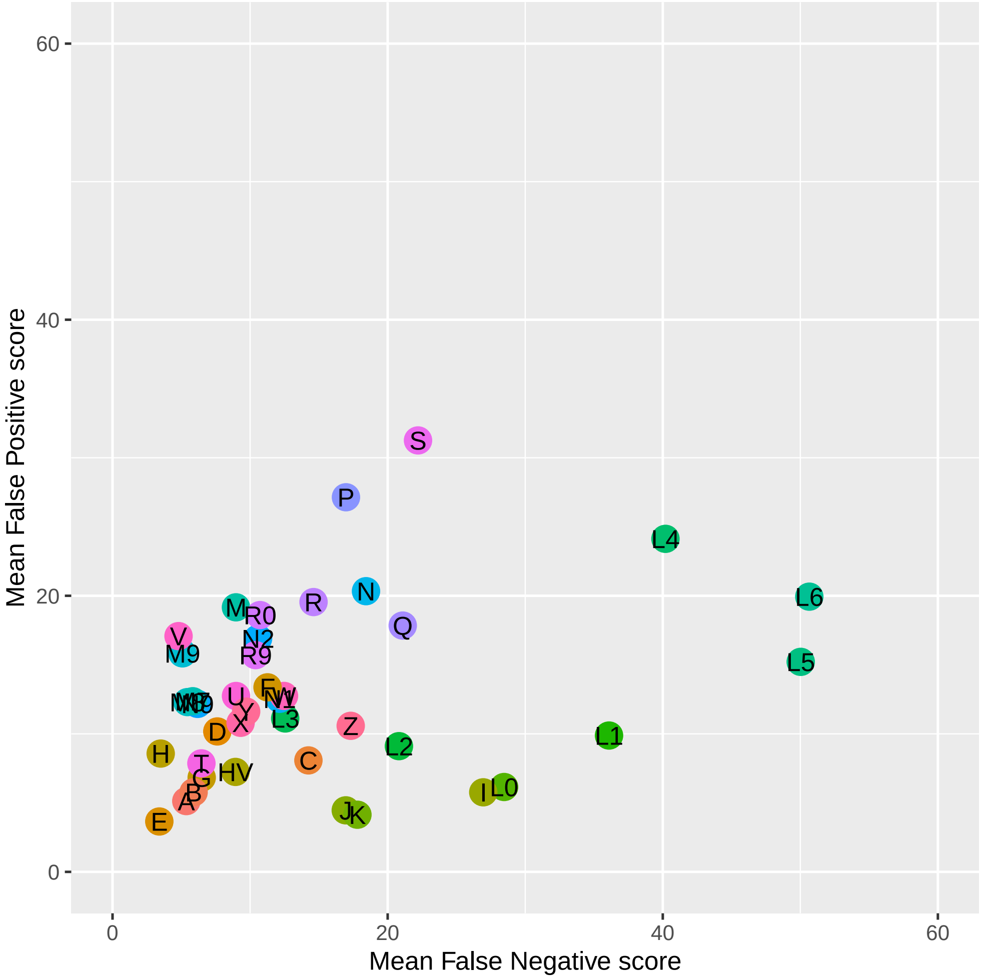


Figure 3.17: Comparison of macrobranch error mean scores plotted against one another.

3.3.4 Genome-wide Variant Prediction Analysis

The variant predictions made for the library-test samples were for variants covering the mtDNA. Each variant prediction list is compared to the observed variants in the sample, expected versus observed, and scored for correctness, false positives and false negatives. Whilst this has been investigated sample-wise using tests 3 versions 1 and 2, these scores can also be examined in relation to the mitochondrial genome.

In order to explore how the predictions perform genome-wide, the correct, false positives and false negatives are gathered by variant. Each variant appears in variant lists, both observed and predicted variant lists, with an attached weight. These are scored as in figure 2.6.

Assessing the variant prediction consistency across the mtDNA is vital to ensure that the predictions are not introducing errors. A good method of predicting mtDNA for a large batch of samples, in order to perform genome-wide association studies, introduced little or no error. Genetic regions, base types or positions which are predicted less accurately are suggestive that improvements should be made before

Having the mtDNA portion of the UKBiobank experimental data in an accessible format to undergo genome-wide association studies requires the predictions to be unbiased over the genome. To explore this requires the variant predictions to be assessed locus-by-locus, rather than sample-by-sample as before.

Viewing the prediction accuracy of the variants in the library samples mapped onto the mtDNA reveals many patterns. There are two measures of the variability seen across the mtDNA variation in the library; (a) homogeneity, and, (b) variants per locus. A locus is deemed homogenous here when there is complete consensus of base for every member of MitoMap library data. I admit that this definition is extreme as the library contain nearly 50,000 sequences and will be enriched for unique or notable samples. These two measures of variability are related, with homogeneity being zero variants per locus. Both are needed to capture regions where human mtDNA sequences are all in agreement and regions of repeated change.

The content of the mtDNA is mapped in figure 3.18. Genes are in cyan, the D-loop region is marked on the map in red and the start points of each gene are marked with black bars. Non-coding DNA is marked with a purple bar from top to bottom of the figure.

Over the entire mtDNA, there are an average of 0.68 variants per locus. This drops a little to 0.62 variants per coding locus and the number of variants leaps to 1.52 on non-coding loci.

Largely, three region types exist in the mtDNA. Over 92% of the bases are in regions which are transcribed into RNA, either tRNA or mRNA to be translated into protein. The non-coding bases fall into 11 very short regions between genes averaging 9.1 bases long, and a single D-loop region of 1123bp which is known to be important in DNA polymerase activity initiation, copying of the genome.

The level of local homogeneity is in black. A measure of the region local to the base represents the proportion of the binned bases which are homogeneous. The bin size used for this figure is 21 base pairs.

The variation per locus is at the top of the figure, coloured navy. There are three hot spots of variation in the mtDNA outside of the D-loop; two have multiple loci with high variant/locus rates nestled well within narrow bands of non-coding DNA (at 5900bp and 8300bp approximately). A third is distinct as an outlier in the region of extended low variation/locus rate, base pair 955 holding 7 variants and resident in the rRNA small subunit gene.

Using the entire library, we can see that 7,792bps (47.1%) of the bases hold no known

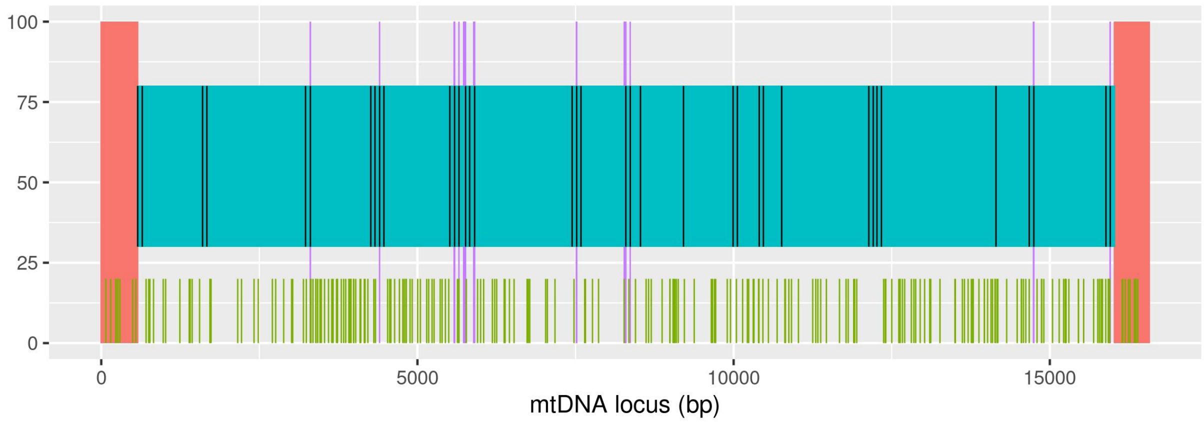


Figure 3.18: A map of the mtDNA. Genes are in cyan, the D-loop region is marked on the map in red and the start points of each gene are marked with black bars. Non-coding DNA is marked with a purple bar. UKBiobank's target loci are marked on in green.

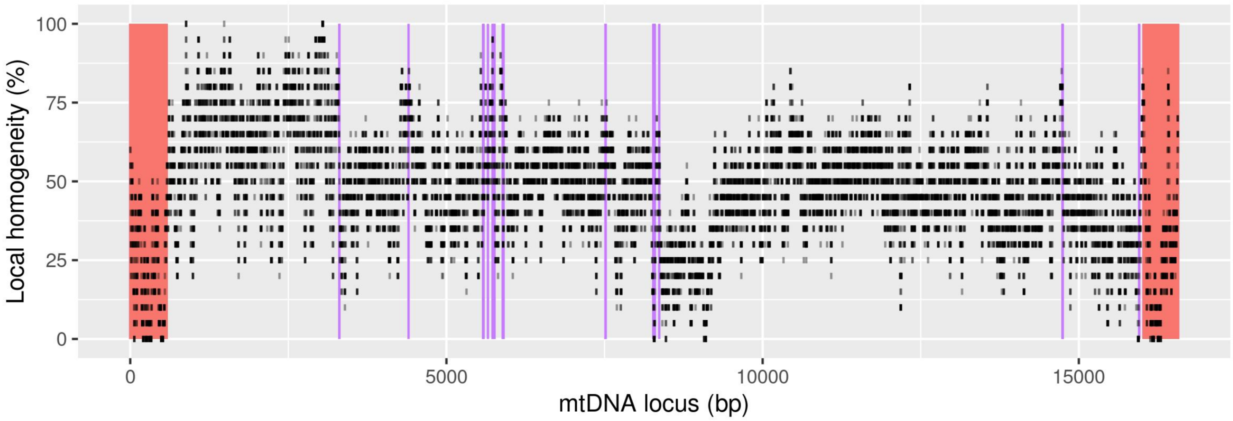


Figure 3.19: A map of the mtDNA showing homogeneity (in black). The D-loop region is marked on the map in red and the non-coding DNA is marked with a purple bar from top to bottom of the figure.

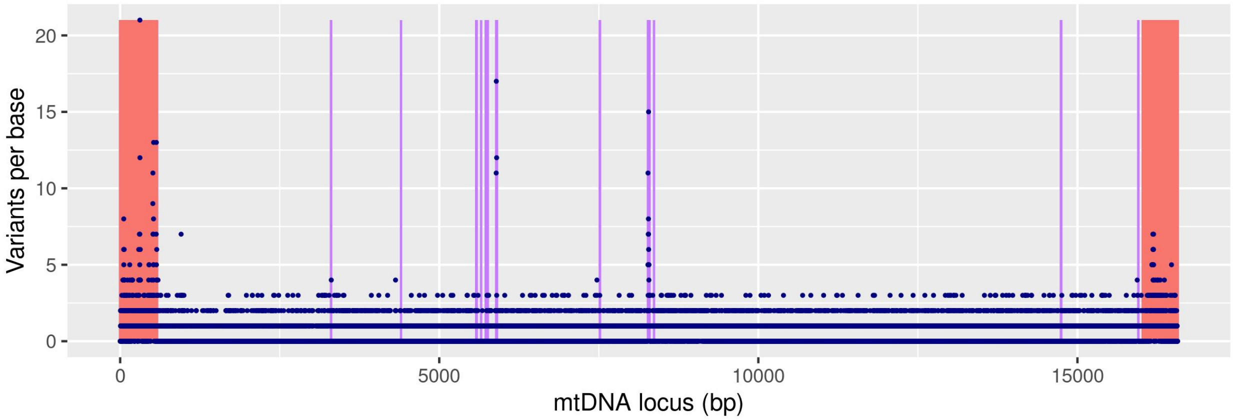


Figure 3.20: A map of the mtDNA showing variation levels (in navy). The D-loop region is marked on the map in red and the non-coding DNA is marked with a purple bar from top to bottom of the figure.

variations, making these bases always equal to the base found in the rCRS in every participant. The remainder of the mtDNA bases are predicted through the 240 variants represented in the barcode. The variant data is stored as differences from the rCRS, as are predictions. The homogeneous bases are not predicted and do not appear in any further analysis.

Further variants are excluded as are not predicted using the library-test set. False negative variants which were observed but not predicted will have a score and be included in the assessment.

The variant predictions for each library-test samples can be viewed across the mitochondrial genome with each locus getting accuracy scores accumulate from all the library-test samples' predictions. This process is complicated by the prediction weightings but can be untangled as in the diagram 2.6.

Each variant gathers weights for correct guesses, false positive guesses and false negative guesses. The total is worked out and then the percentages of weights which fall into the correct, false positive and false negative.

As there may be as many as 21 variants linked to each locus, finding the locus address for each variant allows plotting of patterns across the mtDNA.

Where errors are made, their direction is recorded. When a variant was observed but not expected a false negative error is made. When a variant was expected but not observed a false positive error is made. Dividing genomic regions and comparing rates is shown in table 3.3.

The levels of homogeneity and the numbers of variants local to the predicted variants were different in coding or non-coding bases. However, the accuracy of the variant predictions were consistent in coding and non-coding regions, for correct, false positive and false negative scores.

Spearman's correlation coefficients of the local homogeneity and variation both show no or negligible correlation with the variants' scores for correct, false positive or false negative.

	Local Homogeneity	Variants per base
Correct	0.0013	-0.0126
False Positive	-0.0044	0.0272
False Negative	0.0002	0.0037

Table 3.2: Pearson's correlation coefficients for rates of correct, false positive and false negative predictions against homogeneity and variation level

In order to pinpoint any variables which do correlate with the accuracy or error direction, a breakdown is provided in table 3.3.

Rates of correct, false negative and false positive remain stable when dividing the loci into coding and non-coding, and also in the SNPs. The indel variants have a moderate decrease in performance but also a lower average weight. The average total weight of the SNPs is 532.5, insertions 331.8 and deletions 342.8.

Dividing the variants by those with low total weights and others solved the problem. Improving the general level of performance to 95.3% with a drop in both false positive and false negative rates. Figure 3.21 concurs with the finding which that the majority of the worst performing variants can be found to have lowest rates of representation in the predictions (coloured cyan).

On the removal of the 843 variants with a total weight of less than 4 the overall accuracy of the predictions increase from 90.6% to 95.3%. These low weight variants have a mean accuracy of 38.6%.

	N. Bases	N. Variants	Correct	False Positive	False Negative
All variants (incl. not predicted)	8,777	11,368	80.6%	2.5%	5.9%
All predicted	8,054	10,113	90.6%	2.8%	6.6%
Coding	7,178	8,457	90.6%	2.8%	6.6%
Non-Coding	876	1,656	90.6%	2.9%	6.5%
SNPs	7,942	9,527	90.8%	2.8%	6.5%
Insertions	188	210	**86.3%**	**4.2%**	**9.5%**
Deletions	226	376	**88.4%**	**3.0%**	**8.6%**
Total Weight < 4	822	843	**38.6%**	**11.8%**	**49.5%**
Total Weight ≥ 4	7576	9270	**95.3%**	**2.0%**	**2.7%**

Table 3.3: Comparing prediction accuracy and error direction for each base, separated by locus category; coding, variant type, and low prediction rate

	Total Weight ≥ 4	Total Weight < 4
SNP	92.1%	7.9%
Insertions	84.6%	15.4%
Deletions	84.8%	15.2%

Table 3.4: Comparing the SNP and indel representation in the variants with very low total weights

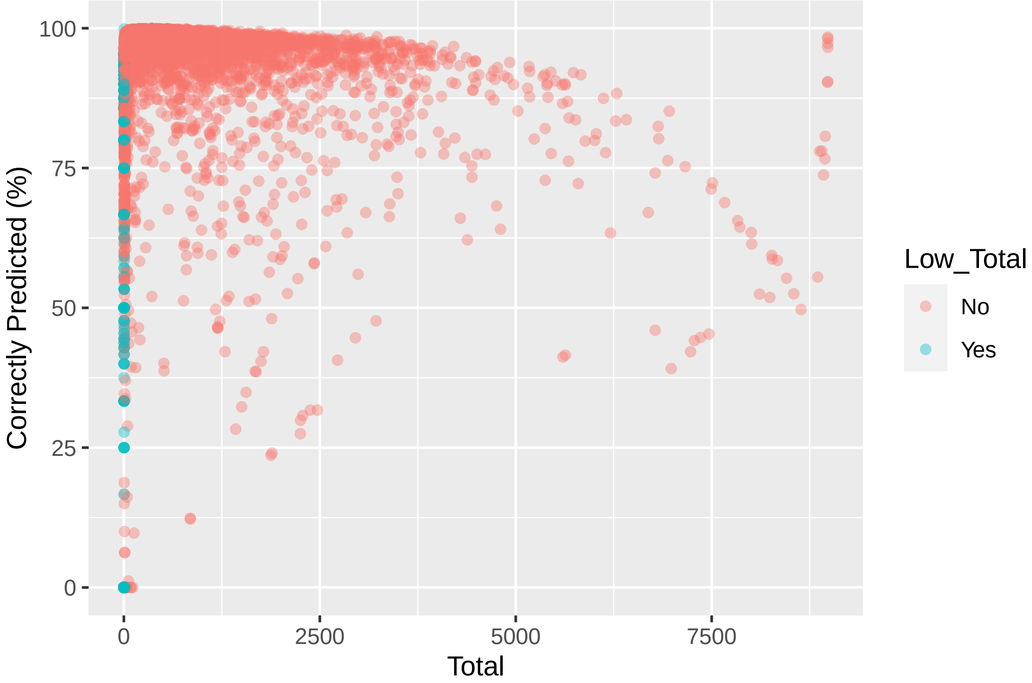


Figure 3.21: Variant prediction rate plotted against percent of predictions which are correct, coloured to highlight variants with totals of less than 4

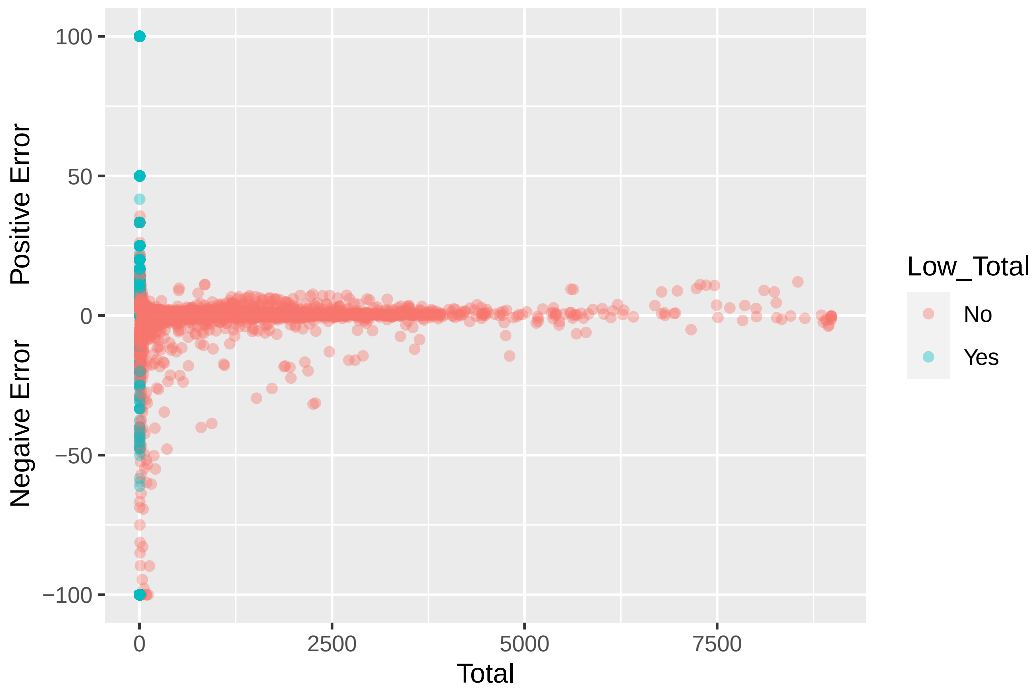


Figure 3.22: Variant prediction rate against amount of error being false negative, coloured to highlight variants with totals of less than 4

Plot 3.22 shows the bias in the direction of error of the variants, again with the low weight data highlighted in cyan. It can be seen that removing these cyan points excludes the majority of the false positives and a proportion of the false negatives. Further exploration of the effect of error is seen in table 3.5, where the behaviour of the loci with extreme errors is seen to differ from loci with a mixture of error type.

The variants with extreme errors can be divided between the variants where the error is only false negative and the variants where the error is only false positive. These groups behave very differently. The FN variants have a much smaller mean total and have no correct guesses. This is suggestive of their being private mutations, which are outside of the abilities of this project. The FP variants still show a lower mean total than the overall rate but also a higher mean correct score. The false positive error is a minor portion of a generally excellently predicted variant.

A short investigation into improvements made simply by removing the variants with a total weight of less than 4 is included in table 3.6. The removal of the low weight variants excluded 89.8% of the entirely false negative data, with a skew towards those in coding regions. The variants with an entirely false positive error were noted for their high accuracy, and 79.5% are retained.

Figure 3.23 shows a histogram of the percentage of the total weight which is correct. There are 577 variants with an accuracy of less than 50%. Removing all variants with a weight of less than 4 will exclude 463 of these, just over 80%, coloured cyan. However, removing the variants with a weight of less than 4 also removes better performing variants. The number of variants with a weight of less than 4 is 843, so 380 good variants are removed with poor ones.

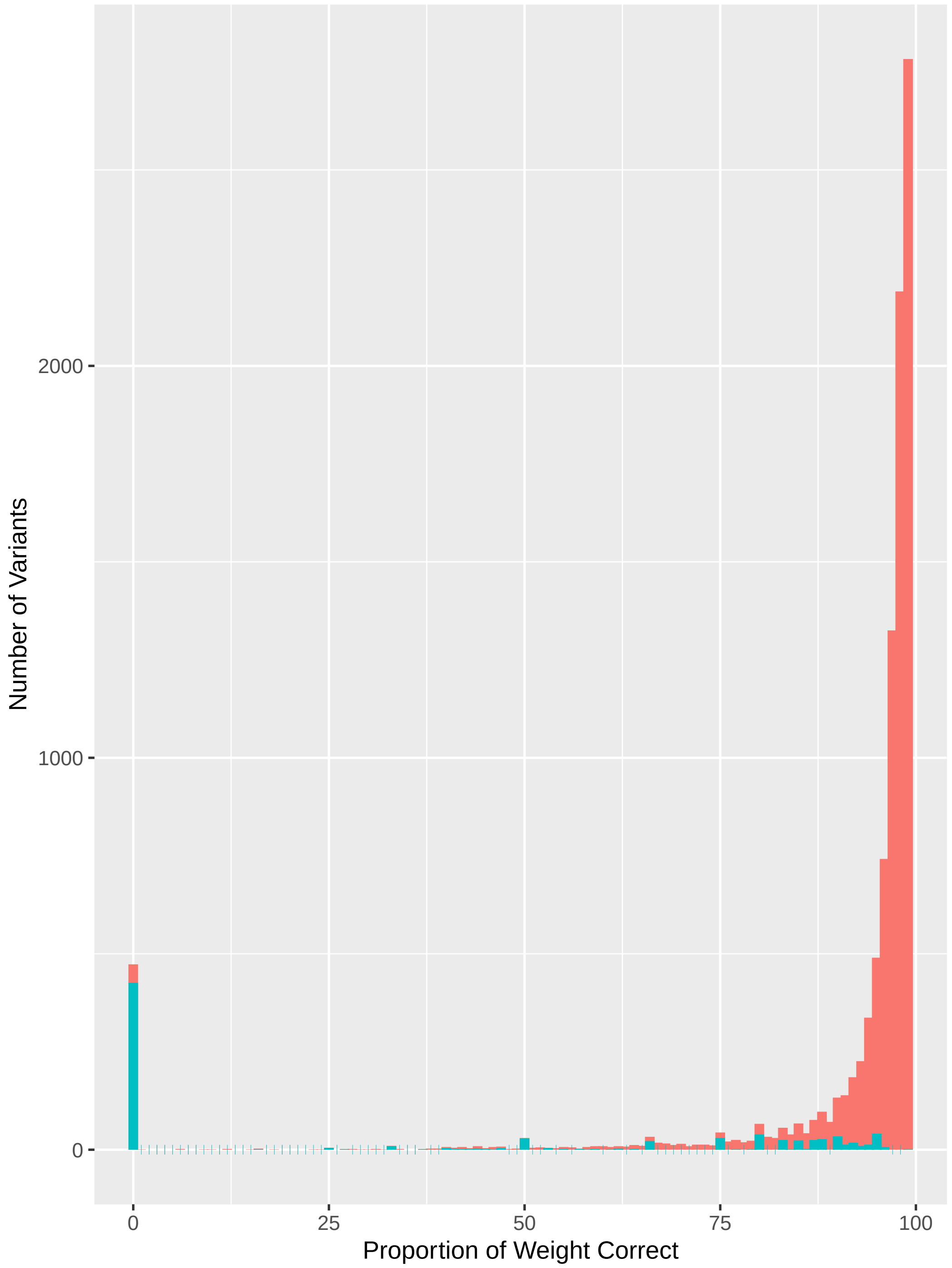


Figure 3.23: Distribution of correctly predicted proportion of weight for all variants. Variants with a total weight of less than 4 are coloured cyan. The y-axis is extended to allow the short cyan bars to be seen.

	N. Variants	Mean Total	Mean Correct	Mean Coding
Overall	10,113	521.3	90.6%	83.6%
Mixed Error	7,567	649.7	78.4%	84.4%
Only FN Error	435	**2.5**	**0%**	84.4%
Only FP Error	3,366	**105.3**	**96.0%**	82.3%
Error > 50%	577	196.1	6.6%	85.0%

Table 3.5: Comparing errors made to base category variables; frequency, predicted rate, correct rate, coding

	N. Variants	Mean Total	Mean Correct	Mean Coding
Overall	9,270	568.5	95.3%	83.7%
Mixed Error	6,250	789.5	94.5%	84.3%
Only FN Error	45	**13.5**	**0%**	**93.3%**
Only FP Error	2,975	118.9	98.5%	82.4%
Error > 50%	114	987.4	21.1%	90.4%

Table 3.6: Comparing errors made to base category variables after the removal of the variants with a total weight of less than 4; frequency, predicted rate, correct rate, coding

3.4 Generating Predictions for the UKBiobank Samples

Using the target set of 240 mtDNA loci, the UKBiobank samples produce 78,720 unique barcodes. The mean number of samples per barcode is 5.6. One barcode was represented by 24,773 samples, 5.6% of all the UKBiobank samples. There were 60,998 barcodes which were only found once, some 77.5%. The number of unique barcodes is high, especially when compared to the more heterogeneous MitoMap Haplogroup library, which yields 7276 barcodes.

3.4.1 Finding the UKBiobank Barcode Coverage Rate

The UKBiobank samples have been processed to take the target loci data only and build a barcode from the data for each sample. The barcodes have been translated into "bait" using the same rules as with the library barcodes to enable grep-style searching of the entire library samples' barcodes. A high coverage by the library barcodes of the list of the UKBiobank's barcodes should indicate that the *in silico* genotyping is mimicking the in vitro genotyping.

It is to be expected that some library barcodes will not be present in the UKBiobank as the library includes an exhaustive list of examples from across the globe. However, this same exhaustive list is expected to cover all the mitochondrial haplogroups found within the UKBiobank population. Having a measure of the library's coverage would indicate how well the *in silico* genotyping pathway is working.

The library returns 7,276 different barcodes and the UKBiobank returns 78,720. Because of the microarray method, some microarray loci in some participants have been recorded as unknown. There are several genuine reasons this may be the case, but experimental error is one which inevitably affects barcodes. The unknown loci are allowed to find "1" or "0" or another unknown. This results in more barcode hits, and a watering down of the prediction, but these are unavoidable. Some 57.8% of the UKBiobank samples have no `neither` calls in their barcodes, and 85% have one or fewer. The mean is 2.3 and the maximum is 50, which was our QC limit.

A comparison of samples which fail to find MitoMap Haplogroup library matches (UKB-Fail) and those which find MitoMap Haplogroup library matches (UKB-Pass) is in table 3.7. This gives a general picture of algorithm success.

	UKB Samples (post-QC)	UKB-Pass Samples	UKB-Fail Samples
Sample Number	438,355	381,786 (87.1%)	56,659 (12.9%)
Barcode Number	78,720	55,351 (70.3%)	23,409 **(29.7%)**
Samples/Barcode	5.6	6.9	**2.4**
Singleton Barcodes	60,998 (77.6%)	42,102 (76.1%)	18,896 **(80.7%)**
Neither Calls/Barcode	2.3	2.3	**2.1**

Table 3.7: The frequencies of samples, barcodes, singleton barcodes, samples per barcode and `neither` calls per barcode for the UKBiobank samples

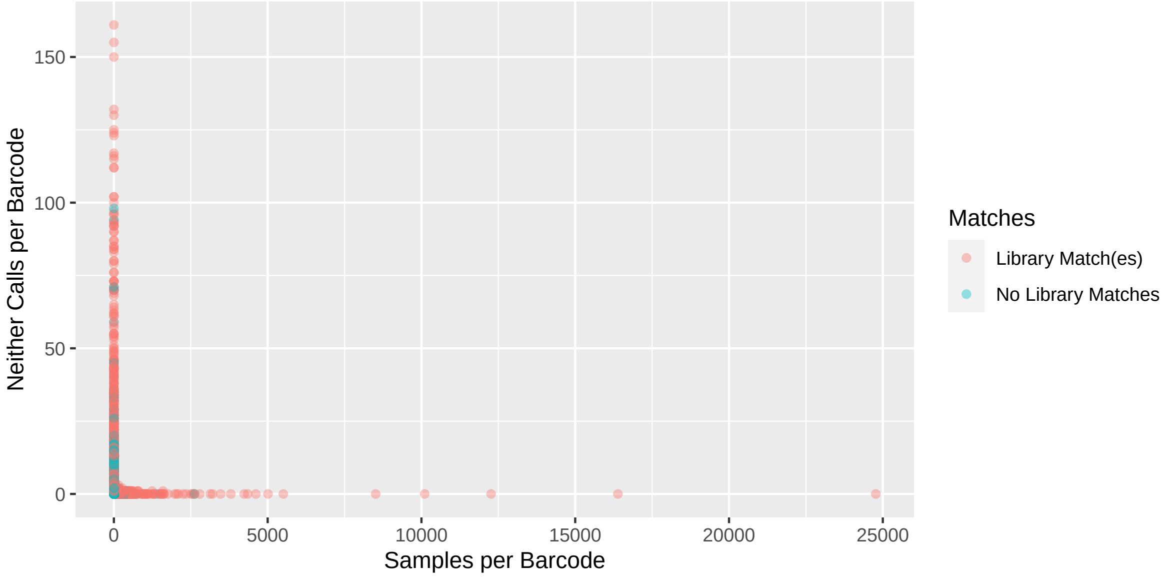


Figure 3.24: The sample frequency of each UKBiobank barcode against its frequency of `neither` calls. Points are coloured red for barcodes which found one or more matches in the MitoMap Haplogroup library barcodes, and cyan are barcodes which found no matches

Comparisons between the `alternative` call rates for each loci show a high level of correlation between; (a) UKB-pass and UKB-fail samples (0.954), (b) UKB-pass and library samples (0.926), and (c) UKB-fail and library samples (0.935).

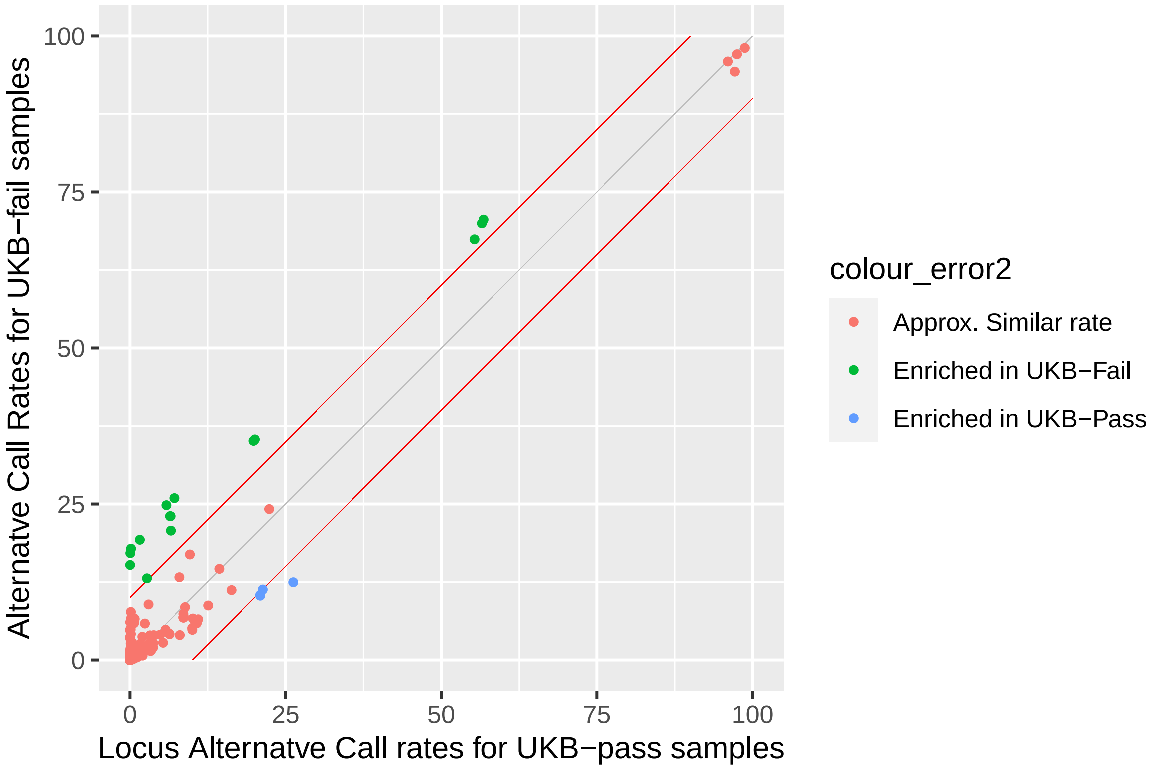


Figure 3.25: Alternative Call rates comparing UKB-pass and UKB-fail samples. The red lines represent a locus call rate difference of 10%.

Chapter 4

Discussion

Genotyped DNA, including mtDNA, is interpolated from the genotyped SNP data in a process called imputation. This process is firmly based on the inheritance rules of the nuclear DNA linkage, but with mtDNA interpolating using linkage is not really possible; the small size and high mutation rate of mtDNA has resulted in thousands of different haplotypes based on mutations from the original as well as many reversion mutations.

Genome-wide association studies (GWAS) find phenotypes linked to genomic regions using variant frequency differences between phenotypic groups. The use of variant frequencies demands any interpolation used in a GWAS must be as accurate as possible as errors will hide true signals or create signals where there are none. A process which answers the need for a tailored approach to imputation of the mtDNA would enable GWAS searches to extend to the mitochondrial genome.

Haplogrep2 is the algorithm used to attempt the extrapolation from the UKBiobank's loci data currently. This code is designed to take small batches of poor quality, patchy data and assign each sample the most likely haplogroup. What became obvious early on was that, once the data was on a significantly larger scale using the most likely haplogroup was not accurate enough. Genetically diverse groups of samples can appear identical on a microarray so predictions of haplogroup membership and variant list need to be much more nuanced and statistically weighted.

I have presented a novel method of interpolating mtDNA using a library subset of mtDNA sequences from MitoMap as training data. The library of training data had to appear as if it had simply been SNP genotyped using a specific set of SNPs. I developed *in silico* genotyping to perform this step which can be easily tailored to different SNP sets. Any sample which undergoes *in silico* genotyping receives a list of coded answers, one at each of the loci. The answer list is collapsed into a single string, referred to as a barcode.

Pattern matching was used to enable predictions to be made which was used first on a test data set, to validate the method, and then on the large body of experimental data from UKBiobank.

The development of *in silico* genotyping and barcode matching was focussed on predicting the haplogroups and variants of UKBiobank's samples, with a knowledge that it would enable the same benefits to any data set with mtDNA genotypes linked to phenotype data.

Whilst the work recorded here was a considerable effort, there are, admittedly, many places where, with more data or iterative cycles, further progress could be made. These are mentioned at appropriate points in boxes marked "Further Work".

4.1 Data Sources and Structures

Bridging the gap between mtDNA genotype data and a library of complete genome sequence was expected to be complex, especially as perfect accuracy was the target. Several aspects of the data did offer considerable challenge but having the common template rCRS sequence was extremely helpful.

The default, economic storage of mtDNA sequence data relates any sequence information to a template sequence, called the revised Cambridge reference sequence (rCRS). The rCRS is a slightly adapted version of the first mtDNA sequence compiled of a white British individual [13]. Any other mtDNA sequence can then be simply recorded as a list of differences from the template sequence. Having an agreed short-hand way of expressing SNPs, insertions and deletions of mtDNA enables a common route to comparing sequences, such as a library of known sequences to any microarray with loci on the mtDNA.

4.1.1 MitoMap Haplogroup Library

A large bank of mtDNA training data was essential for making good predictions from the microarray data and this was found in the library of mtDNA sequences stored in Mitomap. The MitoMap Haplogroup library is built from published genetic sequences from all over the globe [6] and includes 45,882 human mtDNA sequences.

Having such a resource has allowed large-scale developments such as Phylotree, which crystallises the data in the MitoMap Haplogroup library into a branching tree of related sequences using the inheritance patterns of mtDNA variants. By grouping sequences with common variants and, therefore common genealogical descent, wide- and narrow scale relationships can overlay and augment the basic data.

Groups of very similar mtDNA sequences are called haplogroups. Haplogroups can be considered an approximate, short-hand way of describing a mtDNA sequence. For every haplogroup, there are a set of differences (variants) from the rCRS which are markers for that group. Every mtDNA sequence can be placed in the phylogenetic tree using the variants it carries. Any sequence may also carry private mutations which are recent and particular to that mtDNA, but an assigned haplogroup should allow the extrapolation to a large set of expected variants, and vice versa.

Descriptive Statistics

The library of training data was sourced from MitoMap and contained 45,882 complete, or near complete, mtDNA sequences were accessible and stored as differences from the same rCRS sequence and with assigned haplogroups.

The mtDNA sequences found in the library form a network of nearly 5,000 distinct haplogroups. The MitoMap Haplogroup library of samples holds a significant body of information but it is not without limitations which were encountered at several points on the journey to interpolate mtDNA.

With a median sample membership of the haplogroups at four samples, this represents a large batch of data cut into very small subsets. In fact, many groups are represented by just one sequence. Logically, a haplogroup must contain more than one member to be considered divergent from its parental group rather than a parental group member with additional private mutations, however numerous the mutations. Qualification for a new group should be the fixation of at least one mutation common to the group members but additional to the variants seen in the parent group. New group status has not been applied consistently over time or over the phylogenetic tree. Moreover, having a single example of mtDNA sequence in the MitoMap Haplogroup library representing all the members of that haplogroup may force the rapid expansion of the tree. If the mtDNA example contains even a single private mutation, any other member of the same haplogroup may be given a new haplogroup, when the two should be grouped. Defining a haplogroup using a single sequence prevents the distinguishing of variants specific to all haplogroup members from private mutations not fixed in the sequences located at that node. It is quite possible that this is simply not known, or how far on the journey to fixed variants the once private variants have travelled.

Whilst I see many advantages to researchers of the haplogroup system, I have reservations about its use. With an expanding data base of mtDNA sequences the urge to 'place a flag' in new groups is strong when we have novel sequence, but this leads to a thicket of branches which occludes the structures of the family tree. We may be in a problematic hinterland, where we have too many mtDNA sequences to place them comfortably in vague data structure, but too few to decide which variants represent important, defining information.

Haplogroups are a convenient method of grouping mtDNA sequences but their names do not fully capture the relationships between the groups. Larger data structures, called macrobranches, group the haplogroups by genetic similarities. Macrobranch membership is recorded by the first unit of information of the haplogroup string, although, as usual, this system has not been consistently applied. However, once the macrobranch data structure augments the haplogroups names, the whole tree of human mtDNA can be visualised. For this purpose, macrobranches work well on the medium scale.

Macrobranch membership also varies from between 8 members in the O branch up to 6167 in macrobranch H. The vast number of haplogroups and a need to visualise patterns in the 5000 data points across the phylogenetic tree forced me to summarise data into larger sets. With data falling into 41 macrobranches, this structure is used to explore some of the accuracy data later but admittedly fails to provide enough detail in places. Using case studies of several, representative macrobranches was considered but the amount of analysis needed for doing this, say, three-fold was beyond the scope of time and word count[1].

Geography drove the macrobranch names with early researchers assuming a geographical distance or visual difference would indicate the proximity would reflect the relatedness of peoples. As the evidence accumulated, a tree structure became increasingly complex and rooted in Africa. Founder effects narrowed diversity replaced only by newly accumulated genetic changes relying on the passing of time.

The number of samples and the number of subgroups in the macrobranches correlate worryingly well. Many reasons would cause the numbers of certain groups to increase or decrease, moving their numbers away from correlation, but there is little evidence of this in this data. Having such a strong correlation between the two factors looks worryingly like a bias in what gets included in the library. The MitoMap Haplogroup library is a cited source of haplogroup frequency information, and used as such for this project. To use an analogy, studying the contents of a town's library contents to inform a prediction of the town's residents' bookshelves would provide a list of likely titles. However, the comparative frequencies of the books would be misunderstood.

With such a strong correlation between haplogroup number and sample number, macro-

[1] There is a short exploration of the distinctly unrepresentative O macrobranch in the appendices which was prompted by a very poor set of results.

branches have a consistent number of samples per haplogroup. Knowing human populations have passed through many bottlenecks and exploded in number on reaching new continents, having such consistency in frequency seems unlikely, even before bringing in any natural selection pressures. If all sequenced mtDNA was published and therefore included in Mitomap, some haplogroups would be more common than others, and represented more often in the library. If, however, people largely publish novel sequences, these will be over-represented as more groups are logged but not augmented with additional data. As a consequence, weightings for both types of predictions will be better than not having weightings but not truly comparable. Such a strong relationship here is a worry as it is suggestive of another effect; publication bias towards novel findings. This has implications for the assumptions made later.

Complete reliance on the haplogroup strings to define mtDNA sequence relationships is not wise. Further information, beyond that held in the haplogroup string, is required to build longer distance relationships. This prevents the use of haplogroup naming strings being a proxy to mtDNA sequence at a tree-wide scale.

Larger-scale relationships show how macrobranches link to one another, with most relationships being undocumented in haplogroups' strings. Macrobranches emerge from other macrobranches, such as macrobranch N being the parental groups for 8 other macrobranches. Several of these also begin with an N, such as N1 and N2, but others do not, including R, which is itself another prolific parental node. Overly conservative predictions, where a lack of data results in a prediction further up the tree, are inevitable. The relationships made explicit in the haplogroup strings, such as R0's descendance from R, will be scored much more favourably than predictions which span an implicit relationship, such as U from R.

Future Work Compare prediction success score of macrobranches which descend from parent groups with common prefixes, with macrobranches which have novel prefixes.

Without a consistent and systematic renaming of all mitochondrial haplogroups this route to classification will fall short in expressing the sample matches' accuracy or precision. Predictions made for the UKBiobank samples will have to also list the samples' probable variants to account for this.

Future Work Rebuild and rename the haplogroups, potentially re-rooting mtDNA data to mtEve as the template sequence. Viral and bacterial phylogenies may hold better formats, even for private usage to aid analyses. Look into blockchain or dewey-decimal systems for ways of defining relationships whilst allowing infinite flexibility to add new data.

The MitoMap Haplogroup library is Eurocentric, a drawback which is being slowly but actively targetted [27]. The rCRS, the default template sequence, sits on the node marked H. This is understandable for historical reasons but makes estimating genetic distances challenging, especially when using haplogroup string names as a proxy. From this we might assume that the number of variants possessed by a sample which differ from the rCRS, would generally increase with macrobranch leaps from the rCRS made.

When making predictions, errors are made. As mtDNA data has significant structures, such as correlations between variants and strongly varying levels of homogeneity, it was important to ensure error rate and direction (either false negative or false positive) were consistent over variables such as variant number or genome region. Correlations or patterns in the errors would point to ways to improve the *in silico* genotyping or matching processes.

To aid the searching for correctable patterns in the errors, an additional variable was derived which approximated the distance of each macrobranch from the rCRS's macrobranch. This was a simple count of nodes between the two macrobranches called distance. The first use of the distance number was to highlight how the L branches differed from the other macrobranches. Their distance number was smaller than some macrobranches but they held many more differences from the rCRS sequence. There was close, straight-line correlation for the other macrobranches, including African macrobranch L3, from which all non-L macrobranches emerged, but the other L macrobranch distance numbers were too low. This suggests a greater divergence of the L branches than is suggested by in the haplogroup nomenclature, and a distance number in the mid teens.

The MitoMap Haplogroup library of mtDNA sequences does contain a global population but the representation levels require improvement. A case could be made that, in order to capture the vastly increased genetic diversity found in human haplogroups from Africa, enrichment of these groups is needed, beyond the numbers needed to capture the genetic diversity outside of Africa. It was fair to assume that taking samples from superficially diverse peoples from around the globe would provide representation, however the MitoMap Haplogroup library proves how much of human genetic diversity remains largely invisible.

The implications of eurocentricity in the reference library were seen as inconsistencies in the predictions' accuracies. Test samples from the L-branches differ are least well represented in the MitoMap Haplogroup library and this is reflected in poorer scores in the tests of the haplogroup and the variant predictions.

Using the node distance showed up a group of samples with comparatively low numbers of variants. These are a set of samples in the library which are not entire sequence, comparing the variant lists of the samples which share their haplogroup assignment in Mitomap. Quality control should have removed these sequences as, even though rare, they will affect predictions. It appears that 8 of the bad samples, anomalously low in variants, are in the library-test set, and the remaining ones effect the predictions made for 13.6% of the library-test samples. This is covered in appendix 4.5.

Future Work Library QC; A very brief look at the number of sequences excluded using a stringent low variant number threshold for each distance level, suggests about 46 (0.1%) would be removed. An improvement would look at macrobranches for samples with an anomalously low number of variants. Once identified, the samples with low variants numbers can be removed from the matching library-training sample lists before recreating the predictions.

More generally, repeating the harvesting of MitoMap data and then re-processing predictions was outside of the time frame. New sequence is arriving in MitoMap weekly as the cost of whole genome sequencing drops and projects find it affordable and publish their results. Updating the library would certainly improve performance especially if the inclusion of large scale whole genome sequencing is starting to bear fruit.

4.1.2 UKBiobank

Studies of the phenotypes in ageing humans and the DNA's many and varied changes associated with them were made possible with the massive data collection undertaken by the UKBiobank [22]. To enable the scale needed to find these variants, UKBiobank genotyped their participants using SNP data and not sequencing the complete genomes. The mtDNA was typed at a higher density than the nuclear genome but not at a rate to capture all the details of the rapidly mutating mitochondrial genome.

Making further use of the UKBiobank's mitochondrial information offers considerable opportunities. Reasons abound for looking at the genome of an intracellular symbiont upon which the host relies for the vast majority of its ATP. Mitochondria also provide triggers for programmed cell death and must cooperate with their host for energy production and efficiency.

UKBiobank has a focus on the phenotypes of ageing. Their recruitment of only participants of middle age and later prevents research into many conditions, however, whilst severely life-limiting mitochondrial diseases are removed from the research population, UKBiobank's data is very fertile ground for the milder, and commoner, mitochondrial dysfunctions of ageing.

For the extended use of the UKBiobank's mtDNA data, we needed a route to interpolate from the known data to predict the other variations associated with the pattern of variants we see using the microarray. The process described in the methods section has allowed a conversion from 240 genotyped data points to a predicted set of haplogroups and variants. The results of the conversion were validated through the use of a test data set and have been extensively measured for accuracy.

Descriptive Statistics

UKBiobank used two microarrays to gain generate SNP data for all its approximately half a million participants. The microarrays contained loci on both the nuclear and the mitochondrial genomes. Genotyping was the obvious choice for a huge project of the type where success was determined by having enough power to resolve differences between groups of participants. Sequencing whole genomes was not within the financial reach as the costs of sequencing during the project's inception being more then 10,000 times the cost according to `genome.gov/sequencingcosts`. This has and will continue to move towards making sequencing within the reach of more projects. In fact, the Vangard project, a collaboration with Iceland's deENCODE and the UK's Sanger Centre, is excepted to publish the whole genome data for the UKBiobank's participants in 2023 [3].

The mtDNA data are 265 specific loci genotyped on each individual. Most of the addresses of these loci are chosen to gain the most information about the mitochondrial genome they are probing, although a subset were added to look for known pathogenic variants in the UKBiobank population.

The microarray from which these barcode's loci are taken was designed primarily with haplogrouping analysis in mind and includes 180 decisive SNPs on the chip. This still leaves upwards of 60 loci aimed at finding known variants strongly linked to health outcomes. These may be rare, even too rare for the test or library data. Using these loci is inadvisable for two reasons; (a) variations which cause significant early onset disease are likely to be recently acquired, private variants, and (b) these loci are likely to be homogeneous in the library of unaffected individuals so offering only additional computation challenge with no gain of information.

Further than just studying a small subset of the loci, only two variations on each locus could be definitively found using a microarray. The microarray would say if the participant held the same base as the template rCRS sequence (a `reference` call). Alternatively, the microarray would make an `alternative` call if the participant held a single, pre-specified base at this locus. Finally, it could be unable to make a decision and return a call of `neither`. A call of `neither` represents a variety of options, including a failure to decide. Given this issue, the `neither` call had to match `reference`, `alternative` and `neither` calls at that locus.

Differing from the rCRS, some `alternative` options looked to be very rare, or even absent from the samples held in Mitomap. This was expected as some `reference`/`alternative` pairs were chosen to probe for known disease-causing variants, known to be rare in healthy, ageing population, such as the individuals living long enough to be samples for UKBiobank. In addition, simply testing ten times the number of recruits to UKBiobank increased the chances of finding rarer genotypes.

The initial inspiration for this project was to investigate the limits of the UKBiobank's data points and what they can tell us about the remainder of the mtDNA sequence. Approximately 50% of the bases in the human mtDNA we have to date are common to all sequences, however those bases which are not homogenous hold information. A group of identical mtDNA sequences can be assumed to share inheritance. Comparing similar sequences with overlaps, we can assume a close relationship and a largely common inheritance but the discord tells us something about the relationship between the two.

Once many sequences are compared, the differences provides information about their relationships. Small-scale structures such as haplogroups will appear as groups of very similar sequences, with a list of defining variants built from the consensus of the group. Deviations from these haplogroup-defining consensus lists can now be found and recorded as private mutations. On a larger scale, the direction of a relationship between two sequences cannot be determined until more sequences are compared. Which sequence is the ancestor is impossible to decide until more data is added and the inheritance patterns of the variants can be followed.

As certain variants appear together, the correlations between variants create the challenges for the microarray designers. When two variations invariably appear together, only one of the pair should be included. On the other hand, correlations enable prediction making, as certain combinations of data are highly predictive of certain other data points. Other methods take known data points in the mtDNA and return a best guess at the haplogroup, weighting according to variant appearance in the phylogenetic tree. Once the most likely haplogroup has been assigned, it could be used to predict the presence of other variants common to the haplogroup members [104]. Haplogrep2 is a well used and excellent method of automating haplogroup assignment and, even when submitting partial data, Haplogrep2 may well be the best choice. However, once large quantities of genotyping data are submitted, a best guess is no longer suitable. Once you have many samples and are hoping to explore the frequencies of the haplogroups in subsets, many identical best guesses create errors and noise. The clustering is simply due to the data which is missing. Haplogrep2 expects sample data randomly missing portions, and this is not the case for microarray data.

In attempting to recover some of the nuance in predicting haplogroups and variant list from just a little known data, some previously undiscovered association signals can be lifted from the noise. Finding all library-training samples with compatible barcode patterns allows a better picture of frequency to be built into the predictions.

Quality Control

UKBiobank's microarray checked 265 loci of the mtDNA and looked for two SNPs at each locus. Despite this being a tiny proportion of the information, some loci chosen contributed very little information. As this process will compare each of the experimental samples and every member of the MitoMap Haplogroup library, each data point requires additional time investment. If little or no information is gleaned, the data point should be excluded. With the search task in mind and the fact that each additional locus added to the barcodes increases the scale of searching, loci with no information were explored for exclusion.

The performance of a locus could be estimated with the number of `neither` calls made. There were three well-defined bands of loci, at 90%, 10% and 0%, allowing a threshold of 80% to decisively exclude the 22 very poor loci at 90%. `Neither` calls are triggered by issues other than a failure of the locus, such as a large number of truly `neither` variants, but the rates of these was negligible compared to the number introduced by incomplete testing.

When genotyping by microarray, samples which fail to produce a signal at a locus appear identical to those which do not have either the reference base or the alternative base at the locus. Both are coded in the data as a ".".

Future Work Can some neither calls be assumed to be true fails, whilst some are more likely to be correct calls of a third base at that position? Dividing the neither calls would help make predictions more specific as this fourth call could make the barcodes more specific.

An obvious subset of data was included in the set with a mismatching set of loci creating the trimodal distribution of `neither` calls in the loci and enabling the data to be sectioned easily and decisively. Having established which were the very poorly performing loci, these loci were used to divide the samples.

The larger set of experimental data from the UKBiobank were the combination of two data sets. Each set was typed on a different microarray and were of different sizes. Approximately 10% of the participants were typed on the first microarray. Some 22 loci were dropped and 85 were added, before typing the remaining 90%. The subsets had the effect of forming three bands of loci missingness. Removing the 10% subset of participants was important as these now had a further increase of missingness having lost 22 loci on which they previously held data.

With each additional data point, especially one which is under-performing, the barcodes become more complex. There are three options for each locus (`reference`, `alternative`, and `neither`) so each additional locus makes three times as many more potential barcodes. For example, one locus would have three options ("0", "1" and ".") and two loci would have nine options ("00", "01", "0.", "10", "11", "1.", ".0", ".1" and ".."). There are a potential 3^n barcodes, where n is the number of target loci data points. The search becomes more onerous as both the bait is longer and the search pool is larger. However, automating the searching is easy and including these samples and loci would not have been beyond the time available. As loci combinations are inherited together, many potential barcodes do not appear in the data and many combinations are simply not found. Hence the 240 target loci do not return 3.2×10^{114} barcodes but fewer than 80,000.

The central worry I had about processing both sets of data with the same loci was one of analysis. It was very likely that the two data sets would have differences in performance, both overall and in specific situations. Having this additional sources of statistical noise may have masked other issues with my code and prevented effective trouble-shooting.

> **Future Work** Can inaccurate MitoMap Haplogroup library barcodes be found simply because their combination is too far from the set expected from the phylogenetic tree? Can the UKBiobank's set of barcodes be used to exclude incorrect locus calling in *in silico* genotyping? Will this over-fit the library to the UKBiobank?

The UKBiobank project's data were an amalgam of nearly half a million participants built from two subsets genotyped on different microarrays. Choosing to exclude loci using a missingness threshold was simple in the case of the UKBiobank as there was a distinct group of loci with very high missingness, however exploring loci usefulness measurements would be useful when situations were more marginal. There is certainly a case for including all data and using brute force to churn through all the barcodes, however, as the UKBiobank is actually two groups, this introduced an additional, significant variable when exploring accuracy. Tackling the two data sets separately removed this confounding factor and offered a second proof-of-concept data set. Simply excluding the smaller batch of participants, typed on the earlier microarray, made further analyses simpler and offered a second data set which could be used later.

> **Future Work** Some of the loci add very little information to the predictions. Exploring the amount of information held by each of the loci would allow a better threshold to be drawn according to computer resources and time. Shannon entropy is a measure of the information held in a data point would be a good place to start.

> **Future Work** The pilot loci and nearly 50,000 samples could be used as a second proof-of-concept data set. Will the prediction using this set of loci differ greatly in accuracy and which groups are well-defined? Was the microarray improved by the redesign?

The substantial, systematic loss of data seen in the UKBiobank genetic information will produce bias by obscuring differences in certain groups. The loci chosen will inequitably differentiate certain branches and groups, with the result being preferential assignments for well-characterised areas of the phylogenetic tree, including more accurate predictions.

The effective consequence is the unevenness of detail in the predictions, with some groups seen at a lower resolution than others and receiving a broader, more varied prediction. However, the inclusion of the weights in the predictions, which made calculations and data processing significantly more onerous, can mitigate some of these effects. Methods of comparing prediction accuracy and precision are required.

Exploration and quantification of these biases help to ameliorate their effects when making predictions of the UKBiobank's samples. Bias in these predictions is really important to avoid if the predictions are to be used for further analysis of genome-wide associations. Tests to compare prediction accuracy and precision are required. The results of the tests will give a picture of how successfully the set of loci chosen cover the salient genetic information to allow their extrapolation to other variants.

My methods rest on the fact that all samples from the same haplogroup will look alike when reduced to the same subset of loci, however small, that make it a subset.

Future Work Use these tests to improve the choices of loci, ultimately creating a set of ideal loci for barcode numbers of, say, 100, 250, 500 and 1000. Is there an ideal loci data set which maximises uniformity of coverage? When using different numbers of loci, does the included group change? Which loci are associated (or anti-associated) with one another, adding little extra when included together? This could be taken to the extremes, where an expensive, lengthy barcode codes for all variation observed in the library, or the cheap, brief barcode merely divides the samples in to two groups. The balance of these two pressures could be explored using an iterative process where a barcode length is provided and a genetic algorithm searches for the optimal loci for the best barcodes.

Whilst the UKBiobank's loci list is not ideal, using the information to interpolate further offers reliable insights. With the final list of both loci and participants settled upon, processing could continue.

4.2 In Silico Genotyping of Library Data

Modelling the microarray chip's signal is vital to accurately recreate the signal at each data point for each possible variant. Collating a full list of variants which appear in the MitoMap Haplogroup library forms a starting point from which to decide which signals would be generated from each variant. In cases where combinations of variants within range of a locus in the same mtDNA sequence, the coding must account for a hierarchy of variants overriding one another and variants' reach and distance from the locus.

Compared to the sparsely mutated nuclear genome, there is a need for additional care as mtDNA has a higher mutation rate. A single mtDNA sequence could have several variants within range of a locus and ensuring each locus can be decided when faced with any combination of local variants is vital.

4.2.1 Stage One: Collating Library Variants

The process's success rests on ensuring identical mtDNA molecules processed in different pipelines would create the same signals, producing identical barcodes whether through *in vitro*- or *in silico* genotyping. Perfectly modelled chemistry would accurately categorise each of the 240 loci – accounting for the combination of local variants and their effect range of each variant – to emulate the microarray signals produced by the mtDNA.

Each variant has an effect range. The variant becomes relevant when the effect range includes a target locus analysed by the microarray. The relationship between the target locus and the variant decides the variant's influence direction; does possessing the variant prompt a `alternative` call, or a `neither` call?

Loci Variant Analysis

The first step on this journey is to collate the variants in the MitoMap Haplogroup library. The relationship between each of the 11,368 variants and each of the 240 target locus is decided. Most relationship pairs can be ignored as they are out of range, but once in range, the call direction (either `alternative` or `neither`) must be decided upon. With the mtDNA being so dense with variation, some variants are relevant to more than one target locus. The variant may appear on `alternative` list for one locus, and the `neither` list of one or more others.

The processing of the variants in the MitoMap Haplogroup library reveals a great deal about the mtDNA's structures, which is covered as a prelude to the section on the analysis of the variant predictions in section 4.3.3.

4.2.2 Stage Two: Linking to Loci

For the purposes of a genotyping microarray, target loci are single base pairs but, to obtain specificity, a length of DNA called a probe, is employed. In order to differentiate between two SNP possibilities, both versions of the probe are designed to complement the region flanking the SNP. The flanking sequence forces the probe to anneal to a single position in the DNA with such specificity that the base in the central locus can be distinguished. Having heterogeneity in a mtDNA probe's region beyond the expected SNPs, should have a significant effect on the calls made.

To ensure an even DNA-DNA annealing temperature, each probe is generally 25 base pairs long, and contains the SNP approximately centrally, predominantly as the 13th nucleotide residue. Two probes for each locus are used to ascertain which nucleotide sits at their SNP target locus; `reference` or `alternative`. SNP microarrays are designed for less varied nuclear genomes so face a much more heterogeneous environment when employed on the mtDNA.

Accommodating and modelling the effects of other variants in the probe region was complex. The effects of the variants within the probe range are likely to be attenuated by distance from the probed SNP. Several sets of target loci are within the probe regions of each other suggesting that longer range effects have been accounted for. From experience designing PCR primers, I anticipate that SNPs will have a shorter effect range than indels, and the effect range of indels will be modified by their length.

In an effort to base my model on evidence, I looked to the loci call rates from UKBiobank. Assuming the worse case scenario where all calls of `neither` were caused by additional variants within the range of the probe, this was a having a small effect. The mean `neither` call rate was 0.3%, with a range of between 0.02% and 4.75%.

The functions to search for relevant library variants for each locus were built knowing that the effect of additional variants was relatively small. The functions could undoubtedly be improved after exploring variables such as the correct loci range to assess and the influence reach of each type of variant.

I have included in the appendices table 4.1 which specifies the variants which my functions found of relevance to each of the target loci. The functions also place the relevant variants into two categories; (a) variants which would represent an `alternative` call and, (b) variants which would represent a call of `neither`. These lists must be exhaustive and include all relevant variants in any of the MitoMap Haplogroup library samples. The list does not however yet tackle the problem of finding combinations of variants; this is dealt with by later functions.

Table 4.1 also lists the MitoMap Haplogroup library variants which would, if found in a sequence, induce a `neither` call at the locus. Finding one or more variants from this list forces a `neither` call, masking any calls made at target locus. It is this effect than makes the affect range of variants so vital to get right. Too broad a range would render calls largely `neither` masking any data, and making the range too narrow would ignore the effects of the extra variants.

Getting the effect ranges wrong has implications. The effects of assigning a locus with the `neither` are felt at the barcode matching stage, where too many `neither` calls would allow the barcode to match too many library barcodes but assigning a definitive answer, where a `neither` call should have been made would be worse. Being overly specific when the microarray cannot make a call, prevents the barcode finding any correct matches at all. Including `neither` calls enables the locus to match both `reference` and `alternative` calls in the MitoMap Haplogroup library but has a catch. The aim of the process is to make a narrow and accurate prediction from the data held at the 240 loci on each of the UKBiobank samples. Introducing `neither` calls is costly but unavoidable in the presence of local variants.

Ideally, the target loci would be chosen to minimise the probabilities of additional, local variants preventing the SNP's base being ascertained. With so little information about the loci selection process, I cannot be sure how much effort went into this but, remembering how few complete mtDNA sequences were in the MitoMap Haplogroup library when the UKBiobank were designing their array, it is likely that the target loci could be significantly further optimised.

The entire MitoMap Haplogroup library data set offers a great deal further exploration of the predictive power of variant combinations. Additional, local variants are only a problem if they are present in significant numbers and mask important data. Potentially, rare additional local variants may be entirely predictive of the SNP bases they are masking and a `neither` call could be extrapolated with high confidence to the SNP it could be assumed it was covering. Validation includes a series of tests which compare the predictions to the observed. The results of the tests will give a picture of how successfully the set of loci chosen cover the salient genetic information to allow their extrapolation to other variants.

Whilst the lack of influence local loci were having on each other was encouraging, this may have had a more complex reason; if locus A prevents the accurate calling of locus B, this is only an issue if the two SNPs are found in the same individual. If A and B are decisive SNPs for different branches, they do not appear together with a significant frequency. This must be included when investigating the information available at each locus, suggested as further work below.

Future Work How to modulate the effects of SNPs, insertions and deletions local to the target locus? This would involve cycles of iterative and methodical tuning to optimise the target loci's performance.

For the variant effect range, I decided that allowing variants within 5bps of the locus to influence the outcome seemed moderate. This was suggested by the distance between the target loci. Target loci regularly fell in the probe ranges of other target loci. One locus, at 16,148bp, had three others within 6bps. The number of other target loci in range of the probe was not correlated to the rate of `neither` calls. Should the effect range been longer, these target loci would have had increased levels of `neither` calls in the *in silico* genotyped MitoMap Haplogroup library to levels not seen in the *in vitro* genotyped UKBiobank data.

Future Work Explore UKBiobank data patterns symptomatic of pairs of variants triggering `neither` calls for each other. This would feed into the search for an optimal target loci set.

The variants seen in the library is nearly 94% SNPs and, once frequency is accounted for

SNPs make up over 96%. Indels are represented in fewer samples than SNPs but a policy for the indels was also important. Insertions and deletions make large changes to the DNA strand presented to the probe and vary considerably in their length. Deletions and insertions, regardless of length, were considered to be relevant to a target locus only if their base pair address was in the ±5 bp range. This could certainly be improved upon by incorporating an effect range to all indels, related to their length, and using the lengths of the deletions to guide their effect beyond the ±5 bps.

Now the MitoMap Haplogroup library variants were processed, the target loci were all assigned their relevant variants which all had a linked call of `alternative` or `neither`. With the creation of the exhaustive list for each target locus, calls could be made to find the MitoMap Haplogroup library call rate for each target locus.

Comparing Variant Frequencies

The MitoMap Haplogroup library and UKBiobank populations share aspects in common, such as a moderate Eurocentricity, but are very different in mtDNA source. Novel groups or interesting sequences worthy of publishing will be enriched, and exact repeats are less likely to be added, until large batches of sequence from large-scale projects are published. Conversely, the UKBiobank participant were recruited from people living in the UK at the time of the study. The UK, whilst not a homogeneous place, is not a globally representative population.

When comparing the call rates of the target loci between the *in silico* genotyped MitoMap Haplogroup library samples and the *in vitro* genotyped UKBiobank samples, finding a correlation between data sets would show things were working well, but too strong a correlation cannot be expected as the populations represented in the two data sets are different. For each of the 240 loci, the difference between call rates from the two data sets were compared to look for compatibility. Whilst this constitutes a good review for the data processing, too much adherence to getting the rates to match would be foolhardy as the two data sets differ in several ways.

The population differences between the library and UKBiobank mean that making `alternative` rate comparisons is fine but expecting the two to match perfectly is unwise. An imperfect matching rate of 95.5% of the 240 loci within 10% points of each other was deemed acceptable.

`Alternative` calls are more common for MitoMap Haplogroup library samples, with nine barcode target loci showing a significantly higher `alternative` call rate in MitoMap Haplogroup library samples. Only one locus showed a higher `alternative` call rate in UKBiobank barcodes. This could reflect that, although MitoMap Haplogroup has a level of Eurocentricity, that bias is more pronounced in the UKBiobank. It must be remembered that a call of `reference` is telling us that the locus is in agreement with the rCRS sequence from a white, British person. A more globally representative population would hold more `alternative` calls as it is less white and less British.

> **Future Work** Using the list of loci with enriched `alternative` calls in the MitoMap Haplogroup library samples, explore the haplogroups which commonly see `alternative` calls for these loci. Are these extinct groups? Are they likely to be under-represented in the UK population due to their typical haplogroup?

Otherwise, comparing rate locus call rate of the two data sets is convincing evidence there are no significant problems up to this point and barcodes can be built.

4.2.3 Stage Three: Barcoding

Having settled on a final list of loci, barcodes can be built for the UKBiobank experimental samples, including the information only from these target loci. The MitoMap Haplogroup library samples can also be assigned barcodes because variants and their effects on locus assignment have also been found.

Comparison checks at this stage would include gathering the frequencies of the barcodes in both sets to ensure the two data sets are still compatible, exploring to ensure that not only loci calls are in the correct proportion but also the correct combinations. MitoMap Haplogroup Library barcodes may well be missing from the UKBiobank array of barcodes because of the differences in population. UKBiobank barcodes have `neither` calls introduced because of experimental failures, such as poor sample preparation. Compounding the familiar population differences is the complexity of the microarray data. Collecting data using this method is fast, cheap and scalable but incorporates `neither` calls of its own. Poorly performing samples, probes, or batch effects all add `neither` bases and additional barcodes to the list.

The use of the `neither` calls to accommodate the experimental failure means that, whilst a test barcode may not find identical equivalent barcodes in the training barcode pool, it will still yield matches. The `neither` calls will make the group of MitoMap Haplogroup library matches less specific, obviously a disadvantage, but failing to accommodate the dirty data from the UKBiobank is far worse. If the UKBiobank's data was without experimental failures caused by bad samples or target loci, the appearance of a `neither` call could be used as further data. When the bait barcode held a `neither` call at locus X, a `neither` call at locus X was needed for any pool barcodes to match. However, because experimental failures were also represented by a call of `neither`, this assumption could not be made and `neither` had to match `reference`, `alternative` and `neither` calls.

The number of barcodes which the MitoMap Haplogroup library yields reflects the quality of UKBiobank's target loci choices. The choices UKBiobank made many years ago will impact on the information extraction limits and the quality of any predictions made. Choosing 240 unvarying, homogenous loci would result in all library samples with a barcode of "0000...".

> **Future Work** Can the barcode yield be used as a simple number to guide the optimisation of a loci set at a variety of set sizes? With groups of loci being predictive of each other, it would be unlikely that there would be one single optimal answer. A machine learning approach to finding maxima in the number of barcodes would offer a good start to exploring the problem.

4.2.4 Stage Four: Forming Predictions

Making predictions for library-test samples requires the barcode of the same to find at least one compatible sample using a regular expression search. For over 10% of the library-test samples no compatible matches were found and, therefore, no predictions could be formed. The mean match number was nearly 200 which could be a lot of information to compress into a prediction, but should contain substantial numbers of identical or very similar samples. Identical and very similar samples would add weight to the same haplogroups and variants, leading to a confident prediction.

At the upper end of the scale, there were 358 library-test samples which found nearly 1200 matches. A short investigation into this high-match-count group shows that the library-test samples are closely clustered, with the first five IU in common. On exploring their haplogroup predictions shows that the matches are also very well clustered. Each prediction is let down by the inclusion of a single outlier; a library-training sample with the haplogroup "R0+16189". The test scores reflect the large, tight cluster with a single outlier with:

- a lower than average score for test 1 due to the low proportion of the prediction pointing to exactly the correct haplogroup string.
- a high score for test 2 as all but one match fell into a list where all haplogroup string shared the first four

4.3 Validation with a Test Data Set

The method outlined seeks to take the sparse data confirmed by UKBiobank's microarray to impute further information about the UKBiobank participants' mtDNA, predicting the haplogroup set to which the participant belong and the range of variants that the participant's mtDNA sequence should hold. Both predictions are weighted to give an idea of the confidence with which the haplogroup or variant is predicted. The expected use of the inferred data is to find variants which associate with phenotypes, such as health outcomes. This relies on the *in silico* genotyping of a library data set to a level of accuracy that allows signals generated by predictions to be seen above statistical noise.

A superficial comparison of *in vitro*- and *in silico* genotyping results was made but major issues may have still existed. Performance was thoroughly explored and quantified before making any use of the inferred data as it was vital to explore errors made by; (a) varied ability to make predictions (removing some groups from further analysis), (b) varied accuracy (producing wrong predictions) and (c) varied precision (making low quality predictions).

Validation should cover:

1. How many samples find matches and receive predictions?
2. How many samples are placed in their correct haplogroup? What proportion of the haplogroup predictions are correct?
3. How broad are the predictions? What variable(s) correlate(s) with this breadth?
4. How much of the haplogroup string is reliable?
5. How well is the entire phylogenetic tree characterised?
6. Are most samples placed within the correct macrobranch of the tree?
7. How much of the variant prediction weight is correct?
8. How do the errors divide into false positives (over-prediction) and false negatives (under-predicting)?
9. How well is the mtDNA predicted generally?
10. How consistent are the predictions? Do some regions get over- or under-predicted?

Haplogroup and variant predictions performance can be explored sample-wise, gathered in haplogroups or macrobranches. Variant predictions can be further extended to explore the mtDNA regions through their prediction accuracy.

4.3.1 Source of the Test Data

The MitoMap Haplogroup library is a global repository of mtDNA sequences. This offers a bank of open-source data updated as additional sequences are published to use as data from which to make predictions. However, no additional data source exists for comparison. To create a test set of data, with haplogroups, full sequences and genotype-like data, the MitoMap Haplogroup library had to be the source of this test set. It is a common approach when training algorithms to divide the data set into a test data set and a training data set, with a ratio of about 1:5 commonly used [94].

Splitting the MitoMap Haplogroup library data set is an option which enables the vital validation of *in silico* genotyping, but it creates two issues:

1. The MitoMap Haplogroup library-test set have not been genotyped by the UKBiobank microarray but undergone the same *in silico* genotyping process as the MitoMap Haplogroup library-training set.
2. Removing a subset of the MitoMap Haplogroup library data to form a library-test set means some of the high number of haplogroups with few representative samples will be;
   (a) entirely unexplored because all samples fall into the library-training set, or,
   (b) entirely unpredictable because all of the samples fall into the library-test set.

With these caveats, the MitoMap Haplogroup library data set was divided according to rule of thumb that test data sets should be approximately 20% of the entire data set [94]. A pseudo-random number generator selected 10,000 samples using an unweighted distribution and this group of samples became the library-test set. The random selection was essential as the MitoMap Haplogroup library was approximately ordered by haplogroup due to alphabetisation. The remaining 35,882 samples became the library-training data set.

Each of the library-test samples' barcodes was used to find a set of matches in the library-training set. This step used regular expression matching which accommodated the `neither` call in both the library-test barcode and the library-training barcode pool in which the search took place. A record was made of how many matches the library-test sample barcode found and which library-training samples it matched.

Immediately the numbers of library-test samples failing to find any barcode matches was found to be a little over 10%. Whilst failures were expected, having an expected value for the number of fails is important. Having excess failures would be symptomatic of significant issues and prompt an investigation into the causes. The patterns in failing samples would be indicative of specific errors which could be resolved to improve the *in silico* genotyping method.

Whilst the investigation into the library-test samples which fail to find library-training matches finds evidence to support the theory that sub-sectioning small groups in the data could be the sole cause for the failures, the investigation began as I had expected a failure rate of close to zero and assumed that failures were caused by errors in coding. A failure rate of over 10% rang alarm bells and I set about investigating the loci for problems. I divided data into SNPs, deletions and insertions to see if an assumption or rule change would bring the failure rate to close to nil. I reanalysed and repeated the *in silico* genotyping but kept finding no patterns in the failures' loci choice, variant type, or haplogroup. I do think that there are many ways to improve the *in silico* genotyping but I was unable to find a general fail-triggering rule, except that of haplogroup rarity.

There is a danger of over-fitting by tuning the algorithm to predicting the library-test well, to the detriment of future data sets. If the failure rate is largely or entirely due to sub-setting, any attempts to prevent inevitable failure will only act to over-fit. Should we find a proportion of the failures are down to algorithmic issues, a route to their discovery is below.

Future Work Compare library-test-fails with other members of their haplogroup which were either library-test-passes or library-training. This should reveal loci which are incorrect in that barcode, which would allow the tracing back to the incorrectly assigned variant. Patterns in the variants would enable improvements in the coding. This is partially covered in the appendix section on Macrobranch O (4.5).

Investigating the sub-setting issue using overall frequencies was a better initial route towards excluding the failure rate as a symptom of poor performance.

The effect of the small subsets in the library data was implicated in causing the matching failure rate through three routes; (a) a theoretical failure rate in agreement with the experimental rate, (b) the very low potential match numbers of all the failing samples, and (c) very high correlation between the loci call rates of the library-test-fail and the library-test-pass samples. The (a), (b) and (c) are discussed further below.

Calculation of an Expected Failure Rate

When the library data are divided, it appears as if we remove 10,000 samples from a batch of 45,882. In reality, the MitoMap Haplogroup library represents members from thousands of subgroups and dividing the library into two risks removing all group members from the match pool. When the matching groups is small, all of the potential match partners of a sample may be selected for the library-test leaving these samples to fail to find matches.

Establishing that dividing the library is that cause of the failure requires an estimate of how many samples might be expected to fail, given the structure of the data, specifically the distribution of matching group sizes.

The proportion of MitoMap Haplogroup Library-test samples which failed was 10.2% but for a more thorough examination of the behaviour of the failure rate, more data was collected. Please note that the first library-test set remains the one used throughout the validation steps later. The replicates studied here are used solely for the explorations into the failure rate.

Firstly, three ratios of test to training data were measured. By using test batches of 5,000, 10,000 and 20,000, a curve of the behaviour could be plotted. Each time, the library-test samples were selected using the random number generator. These library-test samples were found matches in the remainder of the library, the library-training data set. Library-test samples which recorded no matches were counted as failures. This was replicated five times at each ratio.

Secondly, each of the 45,882 library samples were found matches in the remainder of the entire library. The match number received for samples in this case is a reflection of potential matches the sample had in the entire library. From the distribution of potential match number in the library, an expected number of failures was estimated at five ratios. By comparing the observed rate of failure with the expected rate, it can be seen how well the two agree. The rate of failure can be very accurately guessed when armed only with the frequencies of the barcodes. Rare MitoMap Haplogroup library barcodes fail to find matches more often.

Plotting Match Numbers

Each MitoMap Haplogroup library sample has a number of fellow library samples to which it will match, measured by the match number. Once the MitoMap Haplogroup library is divided and assuming the only reason for a sample to fail is because all of its matches are also in the library-test group, there should be a correlation between potential match number and the likelihood of the sample being among those sample failing to find any matches.

As the probability of selecting all the potential matches of a sample diminishes rapidly as the match number increases, all of the fails are seen in samples with four or fewer potential matches, using a library-test set of 20%.

If all of the library-test-failing samples have low potential match rates, it can be assumed that it is the match number alone which is causing a proportion of library-test samples to fail to find any matches among the library-training, which was confirmed to be the case. All library-test-fail samples had fewer than five potential matches in the entire library. This pattern was confirmed in the five replicates used for the theoretical failure rate in the section above. See figure 3.8, and note that having a single match in the library ensures failure as this is a match to itself.

All five replicates of the library division cycles only had failures with very low numbers of potential matches, and the ratios of potential match numbers are extremely close. These two pieces of evidence suggest that the failures are solely due to the effect of small groups.

It would be interesting to explore the structures of the matching groups. Does the system form a single complete network? Are there islands of samples which share no common barcodes?

Future Work Topology: What does the matching network look like? Does the inclusion of the `neither` calls prevent even small islands of unconnected barcodes forming? Should the entire set of barcodes represent a network with island only caused by incorrect locus calls? Are there hubs and spokes or more uniform linking?

Comparing Passes and Fails

Beyond their rarity, a search was made for other causes of barcodes to fail using the comparison between passing barcodes and failing barcodes. Problems in the process of *in silico* genotyping would be visible as differences in the locus call rate triggered by the mis-assignments of one or more of the variants. By checking the correlation between the locus rates problematic loci can be pinpointed for further investigation. However, none of the loci call rates disagreed. Each of the target loci was found a rate of `alternate` calls for the failing barcodes and the passing barcodes. Problematic target loci would have a non-correlating call rate, with shifted ratio of `reference:alternative`. None of the target loci differ in call rate from library-test-passes and library-test-fails. There was a strong call rate correlation with none of the target loci appearing to falter, which suggests that either there is a close agreement between the results of *in vitro*- and *in silico*.

I suspect that this correlation masks a range of errors which affect only a few barcodes for certain variant holders. Investigation into this at any depth was beyond the scope of this project.

Conclusion

I have presented strong evidence that the failure rate of the library-test set is largely, or entirely, a case of low group numbers. None of the loci look to be predictive of failure. Probability of failure is purely related to the sample's barcode having very few potential matches in the library-training set, tested over several replicates and sampling proportions. However, all three tests have compared library samples with other library samples. A truly conclusive comparison would involve comparing the *in silico* genotyping of the library samples and the *in vitro* genotyping of the UKBiobank.

From this point, all matches and predictions are made for the original test-training data split.

4.3.2 Quantifying the Predictions' Success

With a strong case that the failure rate of 10% was due to the structure of the data and sub-groupings of the MitoMap Haplogroup Library samples, deeper analysis of the content of the predictions for the 10,000 Library-test samples can be completed.

After investigation, several routes to a measure of sample prediction accuracy have dominated;(a) a measure of the probability of each haplogroup of receiving a correct haplogroup assignment using my model, (b) a measure of the homogeneity in the string names of the haplogroups gathered in each prediction, and (c) a measure of the homogeneity in the variants of the haplogroups gathered in each prediction. These are explained more fully below.

Five main scores are used to mark the success of the predictions, comparing expected haplogroup or variant lists with the observed haplogroup or variants. These are explained in table 3.1.

Validation of Haplogroup Predictions

Creating scores of the predictions had to be focussed on the future uses of the data. The common use of haplogroups or macrobranches as proxies for signature variants will require me to produce validated predictions of possible haplogroup lists. These haplogroup predictions are scored using two measures; (a) test 1 measures the proportion of the prediction which point to the correct haplogroup, and (b) test 2 measures the proximity of the predictions to the observed haplogroup.

Test 1 shows a distribution of results where 15% of library-test samples get the haplogroup right for the entire prediction, figure 3.10. The weight of the remaining samples, however, sits at very low end of the scale, with nearly 1000 samples sitting near 0, despite having found matches in the library-training set. A more extensive analysis into how these fall in the haplogroups and macrobranches would show where the poor performers were. The general poor overall test 1 score is symptomatic of how genetically similar the members of related haplogroups are, especially viewed just on the genotyped bases. This detail is lost to the 240 loci used by the microarray and so the proportion of exact haplogroup string matches drops to near zero.

Future Work Where do the samples which are wrongly assigned end up in the phylogenetic tree? What proportion of samples have the correct haplogroup as the haplogroup with the heaviest weight?

Test 1 numbers generate a picture of how the method works overall and can be used to explore coverage or consistency over the branches or macro-branches of the phylogenetic tree. However, this assessment is a blunt instrument. In theory, barcodes could vastly vary in their variant predictive power, whilst having the same proportion of correct haplogroup, according to which other haplogroups fall into the same barcode. A barcode might get the correct haplogroup 10% of the time, but those other guesses could range from close seconds to wildly inaccurate.

Having another test of the haplogroup predictions is important because, once predictions are made for the UKBiobank participants, researchers need to follow haplogroups in both directions. They may need to ask: how often is my haplogroup of interest predicted correctly? or which barcodes link to members of my haplogroups of interest and with what frequency?

Test 2, a measure of how closely related each predicted haplogroup is to the observed haplogroup, is a more nuanced measure. Test 2 has to work within a complex system of names, with the information held in the names of different values. Each haplogroup string will breakdown into a number of information units. This is a complex task as the rules of haplogroup string generation have some information built in. The first unit usually seats the haplogroup in its macrobranch except for every L branch and portions of each of the M, N and R branches. These are followed by a single-digit number. After this unit, the addition of a lower-case letter is followed by a digit, a lower-case letter, and so on. Finally, any group can be augmented by a short list of additional variations, quoted simply as the locus base pair. This acts to add new variations to groups without creating a new subgroup.

The hierarchical nature inherent in the information had to be captured in test 2. Having a "3" as a second information unit was not enough to suggest a close proximity to other "3"-bearing observed haplogroups. This was overcome simply by joining the units together in incremental stages. Matching demands that each correct information unit must also be preceded by the correct units.

The method created will work well for nearly all haplogroup strings. Breaking the units to rebuild them in stages allows the scoring of how many predicted haplogroups carry the first information unit, how many carry the first two information units. . . and so on. A situation where this may not perform fairly is with the augmented additional variations. These are added the string in order of base pair number, so if we compare "HG + 123 + 789" and "HG + 123 + 456 + 789" the algorithm will not credit the matching of the "789", as the preceding unit does not match. The code could and should be improved to prevent this.

A second issue would be present for samples close to the top of their macrobranch. These samples may have a different macrobranch assignment to some of the predicted haplogroups, but still be very closely related. Often, the haplogroup naming systems throw up as many difficulties as they solve.

Future Work Are the samples with the low test 2 scores from haplogroups near to discontinuities in the phylogenetic tree? If the true haplogroup is very high in a macrobranch, some of the haplogroups predicted will be of the parental macrobranch and look like very inaccurate guesses.

Test 2 scores sit generally higher up the scale in comparison to test 1, see figure 3.11. There are well over 20% of the samples on, or close to a top score of 1. The weight of the remaining samples sit in the upper half of the scale, and only very few samples get a score of 0 having found library-training matches.

The test 2 score is the average of a score for each of the information units of the observed haplogroup. This gives a series of interesting data for each match. As the first information unit matches as most common, this usually gives a score of 1 as all the predicted haplogroups agree at this unit. As each information unit is tested, the proportion matching the observed haplogroup drops as the criteria tighten. An averaging of the scores takes into account the number of information units in the observed haplogroup as matching ten levels of information is more challenging than one or two.

The route to scoring used for test 2 offers a second route to analysis; by deciding on an acceptable accuracy threshold, the number of accurate information units which were guessed at or above this threshold can be found. The threshold used in this instance is 100%, so a score of x means that every library-training sample found to match our library-test sample held all of same x information units as the sample being predicted. A score of 0 represents samples where the predictions cannot even agree on which macrobranch the sample should belong.

Recovery100, in figure 3.13, shows a very interesting distribution based largely on fractions. This is a ratio of the number of correct information units guessed correctly divided by the total number of information units in the observed string.

The route to the two integers is better illustrated in figure 3.12, where samples are plotted by their macrobranch. The red points are the samples which received a score of zero because their barcode found no matches in the library-training data set.

For the top scorers, there are samples in T, M7 and L0 where all the haplogroups in the predictions agree to ten information units. Having predictions with such strong agreement is much easier when there is a single match creating that prediction and there may well be a good deal of correlation between number of information units of a haplogroup and the number of members of that group.

Future Work Explore ways of adding a value to indicate confidence to the information unit recovery predictions to reflect the number of library-training samples which were included in the prediction.

Conversion of this information unit prediction score into a measure of information unit recovery allows the complexity of the observed haplogroup to be accounted for. A recovery score of 0.5 tells us that half of the information units in the observed haplogroup were predicted correctly by 100% of the predicted haplogroups. This allows a comparison between observed haplogroups with different numbers of information units.

An advantage of this additional use of the test 2 scores is that it offers a way of characterising the depth to which the predictions can be relied on. The motivation to using such a stringent threshold was to illustrate prediction performance whilst excluding any inaccurate guesses at that level. Researchers may require a level of accuracy up to a particular depth, which can be found for any haplogroup, branch or macrobranch in the mtDNA tree.

Future Work Having 4947 haplogroups to investigate for scores and accuracy was far too much for visualisation in phylogenetic form. Taking three large and well-spaced macrobranches for intense study would be very helpful. Visualising where each sample in each of the macrobranch was predicted would highlight relative haplogroup performance.

Test 2 scoring does not account for the macrobranch joins. If a sample is high in a macrobranch, it shares a great deal of genetic similarity with the parent macrobranch, although has inherited none of the naming string. This would reduce the test 2 score unevenly, particularly for samples near the macrobranch node.

Future Work Renaming the entire phylogenetic tree to include macrobranch joins. The current coronavirus outbreak has placed naming conventions of this cladistic structure in the news. Converting mtDNA haplogroups similar naming structures would improve the test 2 scoring.

Examining the three haplogroup scores together would reveal large scale patterns. This is plotted in 3.14. The behaviour of the three scores closely follows, with a lower test 1 score, high test 2 and wider spread for Recovery100. There is a general trend where most macrobranches receive their highest mean score for test 2, lower for recovery100 and lowest for test 3.

Validation of Variant Predictions

The intended use of the predictions was to extend the search for variants associated with the varied phenotypes collected by UKBiobank. Haplogroups do provide a good proxy for variant lists but have limitations. Most haplogroup strings cannot be traced along branches due to macrobranch name divisions, but variant combinations can be traced. Secondly, due to the increased mutation rate, some variants appear and disappear along the branches. Just because all members at a parent node carry a variant, it may not persist in all the members below that branch. While back-mutations (reversions) are evident in the tree, the majority of what look like variant loss, is actually due to the use of a reference mtDNA toward to extreme end of a branch. Losses are actually gains of variants to the rCRS reference. Using variant combinations highlights the incremental variant list changes from group to group, be that of variant gain, or loss.

Complications involved in using haplogroups lead to problems when looking for associations so predictions of variants were also made. A variant list is built from the list of observed variants and those predicted, with the, hopefully frequent, repeats removed. Each member of this list provides one point, which can be assigned as correct, false positive or false negative, either in its entirety or divided. The huge number of comparisons made this way can be used to assess and improve the entire algorithmic pipeline.

It is vital to note that a single point is assigned for all variants in the list, regardless of the weight of confidence for predicted variants. The level of decisiveness is more relevant here. Correct variants close to 1 or false positive variants with a weight close to 0, both produce a correct score of nearly 1. The poor predictions might be regarded as the scores of near 0.5, as the samples which gather in this prediction are too diverse to be decisive for the variant under investigation.

The prediction comparison data can be amalgamated in two ways; (a) across the samples to assess the performance of each haplogroup and, (b) across the mtDNA to assess the performance of each locus. Building up a picture of how the scores fall into correct, false positive and false negative is vital to find errors.

Starting with (a), test 3 scores are equal to the proportion of the maximum score which is assigned as correct. The macrobranch membership of each of the library-test sample is found, and a mean for each of the macrobranches as the average of all the test 3 scores of all the samples on the macrobranch. A perfect prediction will feature a series of full scores, 1s. Once there is a wider range of samples in the prediction, the distribution of scores for the variants is often bi-modally distributed, reflecting how some variants are not found in all of the prediction samples, and some of the variants guessed are incorrect. The overall level of decisiveness reflects how tightly the matched MitoMap library samples were and how genetically similar they were.

When gathering the test 3 scores by macrobranch, there appears to be an initial pattern where the tree distance negatively correlates with macrobranch test 3 mean. However, once the number of samples from which the macrobranch takes its mean is accounted for, a mean can be derived for the samples at each distance.

A number of variables interact to confuse here. The number of observed variants, the number of samples and the distance from the rCRS all correlate. Understandably, a greater distance from rCRS is reflected in a larger number of variants but, less understandably, a smaller number of samples at the macrobranch. Both low sample number and high average variant number act to reduce prediction accuracy and thus test 3 score.

The test 3 score reflects the proportion of samples in the prediction which also carry this variant but a significant detail is the direction in which the errors are made.

Reducing the data for each sample to the mean value represents a considerable loss of data about things like the spread or which variants are correct and which are predicted unreliably. Using the test 3 scores for a genome-focussed analysis, mentioned in (b) is the next step.

4.3.3 Genome-wide Variant Prediction Analysis

The accuracy of the variant predictions made for the MitoMap Haplogroup library-test samples using *in silico* genotyping and barcode matching has been examined sample-wise. To perform genome-wide association studies using the predictions made on the UKBiobank experimental samples, the variant predictions must not introduce bias or error. In order to show that this method is robust enough to make predictions from genotyping data without adding errors, the variant predictions made for the library-test samples must also be examined variant-wise and located on the mtDNA.

Just as there were relevant phylogenetic structures invisible using haplogroup strings, there are genomic structures which are invisible without placing the variants in their mtDNA context. The MitoMap Haplogroup library samples help again here. Each library sample has a list of observed variants that the mtDNA sequence possessed. The variants are named using the base pair of the rCRS sequence with which they are linked. Each SNP, deletion and insertion is encoded with the locus to which they link and the sequence difference found there. The lists of variants were collated and analysed to map the stability and coding status at each point of the mtDNA.

By analysing the variants present in the MitoMap Haplogroup library samples, the fluctuating levels of homogeneity and variation and the coding/non-coding regions can be mapped across the mtDNA, see figure 3.18.

Variant rate was measured per base, with a count made of the MitoMap Haplogroup library variants which were linked to that locus. This rate had a maximum of 21 differences, but has a mean of 0.68 variants per locus over the entire genome. Among non-homogeneous bases, the average was 1.26 variants per base. The navy line at the top of figure 3.18 plots the rate across the genome. This depiction highlights the strong intermittent signals in a generally low background rate of variants per base. A rate of 0 variants per locus would represent a homogeneous base, for which every example in the MitoMap Haplogroup library there was agreement with the rCRS. This, plotted in black point on figure 3.18, is a measure of local homogeneity to give a better picture that the binary scoring. The score is for the locus and the ten base pairs upstream and downstream. This would have a maximum of 21 homogeneous bases.

In order to allow for differentiation of loci at the extremes of the continuum from loci seemingly protected from change to mutational hotspots, both a measure for variant number and a measure of homogeneity were extracted. The variables were required to ensure variants predictions were performing consistently in these extremes of conditions.

There is a correlation between an mtDNA locus being in a coding region and both increased homogeneity and reduced variants per base. Broadly speaking, non-coding DNA showed a marked higher level of variants, with fewer invariant bases (bases common to all the sequences in the MitoMap Haplogroup library) and higher rates of variants per locus. Both long- and short-range effects on homogeneity are visible in the admittedly noisy data, with clear regions of lower average homogeneity (such as the D-loop and a region approximately centrally) and a prolonged region of high average homogeneity from 500 - 3000bp, approximately. Shorter range effects appear as low homogeneity in the regions of non-coding mtDNA, generally immediately preceded by high homogeneity.

There is some evidence of longer range effects in the rate of variants per base, with a dip in variation/locus rate from 500–2000bp approximately and a general rise in the D-loop, however variation rate acts much more locally.

It is on this mtDNA map, augmented with homogeneity and variation level, that we can analyse the performance of the MitoMap Haplogroup library-test variant predictions. Diagram 2.6 in the method illustrates where the three categories of scoring originate; correct, false positive and false negative.

By bringing together the scores for each predicted variants in the prediction for the library-test set, a genome-wide picture of accuracy was built.

Each variant in predictions produced for the library-test samples has;

1. a locus; the variant's position in the mtDNA
2. an overall weight total; the sum of all the predicted and expected weights.
3. a correct weight total; the sum of all the correctly predicted weights.
4. a false positive weight total; the sum of all weights where the variant was expected but not present.
5. a false negative weight total; the sum of all weights where the variant was present but not expected.

A study of these can reveal patterns in the predictions which may allow quality control when barcode matching is used to form predictions for the experimental data.

Fluctuations in local homogeneity and variants per locus scores both showed no correlation to correctness and no bias to negative or positive errors in table 3.2. Whilst bringing all the data down to just six non-significant numbers was heartening and evidence of a lack of prediction bias, much further investigation was needed.

By breaking the base pairs down in a variety of ways, patterns in scores are revealed, see table 3.3. The headline figures in the top row offer a picture of predictions with a good general accuracy, however these include a proportion of variants which made no appearances in the MitoMap Haplogroup library-test variant predictions or the observed variant list of these 10,000 samples. Once these variants are removed (see the second row of 3.3), we see an even healthier picture with an average accuracy of over 90%. A little less than 10% of the weights of the predictions were incorrect.

The coding/non-coding division of variants is particularly even with only one decimal place difference between the two treatments; a slight change in the direction of error. This is heartening as coding mtDNA must face different evolutionary pressures to the non-coding regions, and could differ greatly as a genomic environment. However, the two region types appears to be predicted with the same level of accuracy and the same distribution of error type.

The type of variant makes a difference to the level of accuracy. Of the variants, 94.2% are SNPs and the remainder are indels, which suffer from a moderate drop in accuracy and a swing towards an increased false negative error tendency. Indels are made up of insertions and deletions, both of which suffer from the reduction in accuracy.

The reduced performance of indel predictions can be explained using a third very strong predictor of prediction performance: total weight. Total weight is the sum of all the weights of all the appearances in the observed and predicted variant lists. Indels suffer from being overrepresented in the variants with an extremely low total weight, at about twice the rate compared to variants with a weight of above 4. Table 3.4 shows how the proportion of variants with very low total weights should be approximately 8%. Over 15% of both insertions and deletions held total weights of less than 4. The average total weight of indel variants is 338.9 but SNPs hold 532.5 on average.

The final two rows of table 3.3 show that rare variants are predicted with very low accuracy, and a swing towards false negative errors. Excluding the 843 variants with very low total weights improves the performance even further, up to 95.3% accuracy.

Figure 3.21 provides conclusive evidence that the exclusion of variants with very low weights is justified. The low weight data are coloured in cyan. Whilst these points do often fall below the 50% accuracy threshold, there are some higher. Some investigation is needed before a blanket exclusion of variants with low weights.

Apart from those at very close to the y-axis, the spread of correctness was usually in the 90s of percent. The proximity of the poor predictors and those variants with an extremely high accuracy makes separation a challenge.

Plot 3.23 shows exactly which scores are removed in removing the extremely low weight variants. The majority of the variants with a total weight of less than 4 have an accuracy of 0%. Tables 3.5 and 3.6 compare the accuracy statistics before and after the removal of the variants with a total weight of less than 4.

Also explored in tables 3.5 and 3.6 are the directions of error. Each variant's error, regardless of the magnitude, was categorised as being entirely false positive, entirely false negative or having a mixed error.

Figure 3.22 shows the weights plotted against the direction of the error. Generally, all of the variants with an extreme error, regardless of direction, have a low weight. However, we can see that there are red dots remaining in these extreme false negative groups. In fact, the two error directions behave quite differently. False positives are exclusively variants with very low weights. False negative remain even in variants with a weight of 2500, which cannot be completely removed with a threshold which does not also exclude many accurate variants.

Variants with a very low total weight also had a wide range of error, see figure 3.22. Even among the variants with a very low weight, patterns are evident. There are variants with purely positive or negative guesses, which are explored later. The variants with a significantly predominant negative error are found in variants with total weights of over 3000, unlike the positive biased variants which disappear at weights of less than 100.

It was important to find situation where a variant was not likely to be predicted accurately. For this, the next step was comparing the general picture with that of variants with extreme error, i.e. when an error was made, the error was found to be entirely false positive or false negative, see table

3.5. These 3801 variants (representing 37.6%) were still found to have been correctly predicted approximately 85% of the time, although their average total was greatly reduced marking these variants as among the rarer.

Variants with solely false negative errors were universally incorrect. Because of their rarity, it can be assumed that these variants are private mutations. Predicting private mutations based on previously observed combinations of variants will never work, as private mutations have not been observed before. Thankfully, variants of this type account for only 4% of the variants found in the library samples.

Conversely, variants possessing purely false positive error is a marker for variants which were very well predicted. These 3,366 variants (33.3% of the total) had an average of 96% correct calls. We might assumes that these variants are very close to being homogenous but their cumulative weights are not high enough.

When the variants with very low total weights are removed, we can see the effects on the performance comparing tables 3.5 and 3.6. Removing the small number of variants with a weight of less than 4 excludes more than 80% of the badly performing variants, and also discriminates between the inaccurate variants with just negative errors and the highly accurate variants with just positive errors. We see that 380 samples of acceptable accuracy are also excluded. This loss of 3.8% may be acceptable viewed overall but may itself introduce biases by excluding rare but well predicted variants.

Removing the variants with very low weights is beneficial, although further investigation into the specific effect on haplogroups or macrobranches is needed. The threshold of weight 4 is conservative and could be as high as 20. Optimising this would require more investigation.

Future Work Explore the error introduced by removing low weight variants and optimise the threshold.

Future Work Can prediction errors be anticipated because of other factors? Can the weightings be adjusted to ameliorate for these other factors?

4.3.4 Benchmarking: Comparing to Haplogrep

Before celebrating any success, a comparison to the predictions made using other means must be completed. Does *in silico* genotyping and barcode matching produce predictions which are more accurate with less bias than the current methods? The go-to method for mtDNA imputation is Haplogrep2, so benchmarking should involve a comparison with this code.

How well does Haplogrep2 do using a best-guess approach? Have I improved on this? Is Haplogrep2's level of variant accuracy acceptable to undertake the genome-wide association studies?

Haplogrep2 has been mentioned at several points as the current method of mtDNA interpolation. There are theoretical issues with extending beyond Haplogrep2's intended use, which was to define haplogroups for entire mtDNA sequences, or those with randomly missing portions, to assign microarray data. However, the results of Haplogrep2's interpolation of MitoMap Haplogroup library *in silico* genotype data must be compared to the results I present.

Future Work Exploration of Haplogrep2's mtDNA interpolation from just the data available from then UKBiobank microarray target loci. Ideally, each MitoMap Haplogroup library-test sample would receive scores for the same five tests I developed to allow method comparison.

4.4 Generating Predictions for the UKBiobank Samples

4.4.1 UKBiobank Coverage Rate

The process of taking the MitoMap Haplogroup library through to *in silico* genotyped barcodes of the target loci was a central aim of the project. The project tried to find matches for the barcodes from UKBiobank, build predictions from the compatible MitoMap Haplogroup library barcoded samples and use the predictions to search for variants which are linked to phenotypes. Finding the overlap between the two data sets points to extent of the success of modelling *in vitro* genotyping with *in silico* genotyping.

In order to find and correct problems, the next step is to compare the MitoMap Haplogroup library barcodes and the UKBiobank barcodes to explore; (a) how many UKBiobank barcodes which fail to find MitoMap Haplogroup library matches, (b) the comparative loci call rates and (c) the barcode frequencies in the two data sets.

The overlap between the UKBiobank barcodes and the barcodes from the MitoMap Haplogroup library is a useful indicator of the performance of the *in silico* genotyping of the MitoMap Haplogroup library. The mis-assignment of one or more target loci would create a mismatch between the two data sets revealed by a significant number of UKBiobank barcodes failing to find matches in the MitoMap Haplogroup library barcode list.

The complete, undivided library is used to search for matches to the UKBiobank barcodes as it provides examples of all known haplogroups now converted into their barcode equivalent using *in silico* genotyping. Near complete coverage is expected in this case, with barcodes from all known haplogroups making representative barcodes to cover all the expected haplogroups in the UKBiobank, and many more.

After performing regular expression searching accommodating the *neither* calls, nearly 13% of the UKBiobank samples received no matches at all. A search began for variables which correlated with UKBiobank's sample failures. Making a comparison between the barcodes (see table 3.7) reveals several differences between the UKBiobank samples which were able to find matching barcodes in the MitoMap Haplogroup library data (UKB-pass), and the samples which did not find matches (UKB-fail). Primarily, the numbers of samples per barcode were lower in the UKB-fail samples at 2.4 samples per barcode compared to 6.9 in the UKB-pass samples. This was, in part, due to a higher number of barcodes represented by single samples in the UKB-fails. The UKB-fail barcodes have fewer `neither` calls so are more specific. Both of these factors require further exploration.

It can be seen in figure 3.24 that two barcode variables disagree. Barcodes with a sample representation of over 5 have a `neither` call count close to 0. Barcodes with a `neither` call count higher than 5, have a sample representation close to 0. Logically, this is entirely to be expected. If there are a set of barcodes with 240 units of data, as loci are correlated with one another only a subset of barcodes will emerge from our UKBiobank data. Whilst some target loci have a small number of genuine `neither` calls, many of the call of this type will be due to experimental errors or failures. The majority of these errors will be randomly located on a target locus, making one barcode from a set appear different.

I suggest that the exploration into the UKBiobank sample failures might wish to treat these as separate populations and look for problematic loci specific to each group. It may be the case that grouping the populations together hides problematic loci.

Future Work Extend the search for problematic loci by dividing all the UKBiobank barcode into those with a `neither` call of greater than 5, and those with a sample number of greater than 5. Care should be taken with how to deal with the large group of barcodes which falls into both categories.

The reasons behind MitoMap Haplogroup library-test samples failing to find matches among library-training samples was complicated by the division of the library data set. The situation with the UKBiobank samples is not hindered by the issues caused by the data division so having rare failing barcodes from the UKBiobank data is more convincingly supportive for there being mtDNA sequences which are unrepresented in the MitoMap Haplogroup library.

The UKBiobank shows a level of Eurocentricity, which is mirrored albeit less extremely in the MitoMap Haplogroup library, but is of a much greater size.

Future Work Is there evidence of new mtDNA sequences in the UKBiobank? There is a subset of barcodes which are found with a very low frequency which have no matches in the library. Are these examples of mtDNA sequences which are undiscovered? This would only be revealed by sequencing the mtDNA of these participants but well worth the small financial investment if they were to yield novel human mtDNA sequences from people living in the well-sampled UK.

The targeting of the failing barcodes in UKBiobank for full sequencing is likely to find novel sequences, certainly selecting participants with rare barcodes will enrich for rarity. However, adding the results of this will increase the noted issue in the MitoMap; the sample frequencies in MitoMap will show a further disparity with the global frequencies.

Alternatively, there is also evidence that the UKB-fail barcodes are more specific, with a rate of 2.1 *neither* calls per barcode compared to 2.3 for the UKB-pass samples. This fact suggests that there were loci which were being called incorrectly, which would be enriched in the UKB-fail barcodes. Whilst the correlation was close (see figure 3.25), there are several outliers worth investigating.

Future Work Explore which loci have a different *alternative* call rate in the UKB-fail barcodes compared to the UKB-pass barcodes.

Future Work Further limits on performance are felt with the partial coverage of the library-test data. A sensible move would be to repeatedly resample, predict, analyse to build up data covering all the haplogroups in the library and give a robustness to the scores as fails trigger zeros which could improve in a different sampling cycle.

Future Work The UKBiobank's samples reduce to approximately 70,000 barcodes using the target loci. Predictions can be made for each of these loci and, with a little adaptation, scores for some of the tests can be produced. These would have to examine the breadth of the predictions rather than their distance from the correct haplogroup or variant list. However, the numbers could be used to; (a) illustrate the confidence we have in the predictions, and (b) give feedback on overall success which could be used to iteratively improve weightings, loci assignments. . . etc.

4.5 Conclusions

As DNA sequencing becomes more accurate and affordable, even for novel large-scale projects, powerful phenotypic data bound to SNP-genotyped genomes will lose value. Because only some projects will find funding to re-interrogate the genomes of their participants, potentially useful phenotypic data will remain locked behind mtDNA genotyped participants from past projects. This represents a significant waste of investment from funders, researchers and the people who participated.

A method of recovering additional information from the mtDNA without the cost of sequencing would need to accommodate the translation of SNPs and indels, and impute without adding errors. The rewards for a development would be to extend the use of genotyped mtDNA data and allow older data to be included in novel analyses.

The prediction training data is from a global mtDNA sequence resource called MitoMap. With a method of translating the expanding MitoMap Haplogroup library sequence data into a format which is compatible with microarray experimental data, a route to access past data is opened. The method presented in this document, *in silico* genotyping, was focussed on the microarray used by the UKBiobank but is easily repurposed for the mtDNA target list of any microarray. The second stage, barcode matching, forms predictions for each of the experimental samples from compatible MitoMap Haplogroup library samples.

Each UKBiobank participant can be given a much more detailed prediction by the incorporation of more information, through using mtDNA knowledge such as the mtDNA inheritance patterns and sequence population frequencies. These mtDNA-specific assumptions improve the interpolation as mtDNA is inherently more predictable although mutating at a much higher rate than the nuclear DNA.

In silico genotyping and barcode matching represent a significant amount of performance-validated work but at several points of code development, decisions about thresholds have been made. Attempts were made to base these decisions on knowledge of the system to be modelled, but they should be calibrated more scientifically. Small changes to some of these parameters could well have large effects on the predictions made.

Despite some guesswork in the threshold positions, method validation gave favourable results. Validation was performed on a subset of the MitoMap Haplogroup library. This is by no means the ideal data set on which to validate the process, but the only choice. I had full sequence, haplogroup and genotyping results, albeit for *in silico* genotyping. Bad assumptions made in the journey from full sequence to *in silico* genotype, could well boost prediction performance due to both groups of data having common mistakes. However, a data set bridging genotyping and full sequence does not yet exist, which is why *in silico* genotyping has been developed.

Analysis of the information units of the haplogroups in the prediction showed significant promise. Whilst it is a challenge to perfectly correctly predict every information unit of a haplogroup, perfect predictions can be made for the majority of IUs of the majority of library-test samples. This opens the door to the inclusion of more details in analyses than currently available.

Pleasingly, the variant predictions obtained a very high level of accuracy of over 95%. The weights showed overwhelmingly that the guesses were correct and the errors made were balanced between false positives and false negatives.

During the validation, it must be remembered that interpolation performance relies on two factors: the performance of the algorithm and the target loci choices made by experienced geneticists at UKBiobank. A route to testing just algorithm is offered through interpolation of the *in silico* genotyping data using Haplogrep2 The second set of results, accuracy testing using the same test as outline in this thesis, would be useful.

The MitoMap Haplogroup library is a large, publicly available resource but the nature of mtDNA grouping means the large dataset is subsectioned very finely. The limiting effects of low representation number were seen at three points. Rarer MitoMap Haplogroup library haplogroups and rarer variants both saw significantly inflated prediction errors. Whilst the removal of the rare barcodes and variants does improve the overall levels of prediction accuracy, it risks introducing different errors to the errors it corrects. The list of mtDNA in the MitoMap continues to grow and should become more globally representative, in frequency as well as sequence, as larger batches of mitochondrial genomes are fully sequenced.

When translating the experimental data, the rarer UKBiobank barcodes also ran into problems, notably an inflated failure rate. This suggests the existence of rare haplogroups which are not yet included in the MitoMap Haplogroup library, and the targetting of these samples for mtDNA sequencing would yield novel data. Whilst blaming failures on novel sequences alone is tempting, the optimisation of earlier steps would be a sensible route to pursue too.

There is evidence that the *in silico* genotyping and barcode matching provides prediction with enough accuracy to attempt to find variants which associate with health phenotypes, particularly in the absence of full mtDNA sequence data. For the next steps, phenotypes involving tissue ischaemia, particularly where the mitochondria of the cardiac muscle or central nervous system are placed under acute and extreme stress, should be initially prioritised for association studies. Beyond these conditions, there are countless human ageing phenotypes, from many branches of dysfunction, that would benefit from investigation using the mtDNA predictions produced using *in silico* genotyping and barcode matching.

Moves to use the prediction of the UKBiobank experimental data are beginning as they are put to work in attempts to perform genome-wide association studies on some of the health outcomes in the UKBiobank data. Uncovering genuine signals in the noise to reveal mtDNA variants which are associated with conditions such as heart disease, would be a fitting use for this imputed data.

A opportunity to fully assess the success the work undertaken looms as UKBiobank are expected to publish the results of the whole genome sequencing on the participants in 2023 [7]. This offers a route to checking the predictions enabled by *in silico* genotyping, should the coverage extend to the mtDNA, this offers an opportunity for true validation of results which I enthusiastically welcome.

Appendices

Glossary

Alternative General VCF file: A single specified variation from the reference sequence, rCRS. Microarray VCF file: A single specified base which varies from the reference sequence, rCRS.
Bait A translated barcode used to search a list for matches.
Barcode A string of ones, zeros and full stops. This is a concatenated representation of the answers at the target loci for the sample in question. A one represents an alternative variants, a zero is no change from the reference sequence and a full stop represents no known answer for that locus.
Coding Region of the mitochondrial genome which is transcribed into tRNA or mRNA.
Collapse The reduction of definable groups due to the reduction of detail available on samples because of a lack of information.
D-loop A non-coding, highly variable, control region of the mtDNA, between base pairs 16024 - 576.
Deletion Base or bases which are not present compared to the rCRS sequence.
Dendrogram A tree-like structure encoding relationships.
Expected The prediction of either probable haplogroups or variants.
False Negative A variant which is observed but not predicted by the expected variant list.
False Positive A variant which is expected but is not in the observed variant list.
Genotyping The checking of a single base of a sequence for two pre-specified bases; the base in the reference sequence or a single specified `alternative`.
Grep Search This type of searching uses the regular expression code to allow a search for barcodes with either a "1" or a "0" in that position of the barcode.
Homogeneity A base pair or region of genome which has no variants linked by locus address in the entire library.
Imputation Method of making a statistical guess on the unknown variation between the known sequence.
In silico genotyping A process of extracting data from specified loci from a full list of a sample's variants.
In vitro genotyping See genotyping.
Indel An insertion or deletion. A variant which is not a SNP.
Information Unit Haplogroup names are strings of letters, numbers and some punctuation marks, which can be cut into units. The start-end points of the Information Units are not regular but are formed by either a punctuation mark or a change from letter(s) to number(s), or from number(s) to letter(s).
Insertion Base or bases which are present but not expected compared to the rCRS sequence.

Library A data set gathered from MitoMap Haplogroup Containing 45,888 samples of the Haplogroup and a full list of where and how the sample differs from the rCRS.
Library-Test A subset of the library data found matches among the Library-training set and given predictions with which to compare with the observed haplogroup and variants.
Library-Test-Fail A sample from the Library, selected for the test set, but which fails to find matches in the training set.
Library-Test-Pass A sample from the Library, selected for the test set, which finds matches in the training set.
Library-Training A subset of the Library which is used to create predictions linked to barcodes.
Macrobranch A large subset of haplogroups which begin with the same information unit or units.
Match A Library sample possessing a barcode which, when converted to a "pool" barcode agrees with the "bait" barcode used in a regular expression search.
Microarray A small plate loaded with spots of short DNA pieces designed to anneal to sample DNA to detect specified sequence variations.
Missingness A measure of how much data is unknown.
mtEve Mitochondrial Eve. A theoretical female who carried the mtDNA from which all observed human mtDNA sequence have descended.
Non-grep Search This search ignores the regular expression wildcard meaning the Regular expression term "." will not find matches with "0" or "1".
Non-coding Region of the mitochondrial genome which is not transcribed into tRNA or mRNA.
Node Distance The number of macrobranch nodes between two positions on the phylogenetic tree.
Observed The actual haplogroup or variant of the Library-test sample.
Phylogeny The tree-like construction which can be built using the changes in the mtDNA and their relationships.
Pilot A separate group of data produced on a different microarray.
Pool A list of potential matches searched using the bait.
Prediction A condensation of the matches into a list of possibilities and a corresponding list of weights.
Private mutation Variation which is unique to the individual, not inherited from the parental mtDNA.
Probe A short DNA chain used to find its complementary sequence in mixed DNA.
rCRS A template sample, called revised Cambridge Reference Sequence, which acts as a comparison sequence. The Library samples' sequences are all specified as lists of differences from rCRS.
Probe A short DNA chain used to find its complementary sequence in mixed DNA.
SNP A single base change, not inserted or deleted but differing from the base expected compared to the rCRS sequence.
Single Nucleotide Polymorphism A single base change, not inserted or deleted but differing from the base expected compared to the rCRS sequence.
String A series of alphanumerics which are concatenated together.
Target This is a locus on the mtDNA which appears on a genotyping microarray.
Template A commonly used reference sequence, to which all other sequences are compared.
Test Data A subset of the library data set, randomly selected, which is used to test the predictions created with the training set.
UKB The UKBiobank is an organisation which collected health-related data and genotyping data.
UKB-Fail A UKBiobank sample with a barcode which did not find matches in the Library set.

UKB-Pass A UKBiobank sample with a barcode which found matches in the Library set.
Unique list A list with no repeated elements.
Variant A difference of genetic sequence from the rCRS. Specified by a string built from a concatenation of the start locus and the different base(s).
VCF Variant call format. A file type used to record sequence data as differences from a reference sequence.
Weight A confidence value for a prediction.

Target Loci

Table 4.1 is a list of the subset of UKBiobank loci which were used for the *in silico* genotyping. The table includes every relevant MitoMap Haplogroup library variants which trigger a call of `alternative` or `neither`.

	UKBiobank Microarray Information			Output of *in silico* genotyping stage 1	
	POS	**REF**	**ALT**	**Alternative Variants**	**Neither Variants**
1	73	A	G	73G	73C 73d.A
2	150	C	T	150T	150A 150d.C 150G 150CT
3	228	G	A	228A	228T
4	235	A	G	235G	none
5	263	A	G	263G	263d.A 263C
6	295	C	T	295T	295A 294TT
7	497	C	T	497T	none
8	547	A	**T** (incorrect, should have been G)	**none**	547G
9	709	G	A	709A	709C
10	750	A	G	750G	750C
11	769	G	A	769A	none
12	827	A	G	827G	none
13	980	T	C	980C	980G
14	1018	G	A	1018A	none
15	1243	T	C	1243C	none
16	1391	T	C	1391C	none
17	1406	T	C	1406C	none
18	1438	A	G	1438G	none
19	1555	A	G	1555G	none
20	1719	G	A	1719A	1718AA 1719GG

Table 4.1: Target Loci with `reference` and `alternative` bases. Including the relevant library variants and their assignments, continues below. Please note that line 8 contains an erroneous locus choice by UKBiobank, reflected in a complete lack of alternative variants in the MitoMap library

	UKB. Microarray Info.			Output of *in silico* genotyping stage 1	
	POS	**REF**	**ALT**	**Alternative Variants**	**Neither Variants**
21	1721	C	T	1721T	none
22	1736	A	G	1736G	1736T
23	2158	T	C	2158C	2157TA
24	2218	C	T	2218T	2218A
25	2416	T	C	2416C	none
26	2483	T	C	2483C	none
27	2706	A	G	2706G	2706C 2706T
28	2758	G	A	2758A	none
29	2885	T	C	2885C	2885d.TA
30	3010	G	A	3010A	none
31	3027	T	C	3027C	none
32	3197	T	C	3197C	none
33	3243	A	G	3243G	none
34	3308	T	G	3308G	3308A 3308C 3307AA 3308TC
35	3316	G	A	3316A	none
36	3348	A	G	3348G	none
37	3394	T	C	3394C	none
38	3395	A	G	3395G	3395C
39	3423	T	G	3423G	3423A 3423C
40	3460	G	A	3460A	none
41	3480	A	G	3480G	none
42	3531	G	A	3531A	3531T
43	3594	C	T	3594T	3594A
44	3645	T	C	3645C	3645A
45	3666	G	A	3666A	none
46	3720	A	G	3720G	3720T
47	3733	G	A	3733A	none
48	3736	G	A	3736A	none
49	3796	A	G	3796G	3796T
50	3834	G	A	3834A	3834C
51	3866	T	C	3866C	3866G
52	3915	G	A	3915A	none
53	3918	G	A	3918A	none
54	3936	C	T	3936T	3936G 3936A
55	3946	G	A	3946A	none
56	3949	T	C	3949C	none
57	3990	C	T	3990T	none
58	3992	C	T	3992T	none
59	4024	A	G	4024G	4024T
60	4093	A	G	4093G	none

	UKB. Microarray Info.			Output of *in silico* genotyping stage 1	
	POS	**REF**	**ALT**	**Alternative Variants**	**Neither Variants**
61	4104	A	G	4104G	none
62	4160	T	C	4160C	none
63	4171	C	A	4171A	4171T
64	4216	T	C	4216C	none
65	4310	A	G	4310G	none
66	4336	T	C	4336C	none
67	4529	A	T	4529T	4529G
68	4561	T	C	4561C	4561G
69	4580	G	A	4580A	none
70	4639	T	C	4639C	none
71	4715	A	G	4715G	none
72	4769	A	G	4769G	4769C 4769T
73	4793	A	G	4793G	none
74	4820	G	A	4820A	none
75	4824	A	G	4824G	none
76	4883	C	T	4883T	4883A
77	4917	A	G	4917G	none
78	5004	T	C	5004C	none
79	5046	G	A	5046A	none
80	5147	G	A	5147A	5147C
81	5178	C	A	5178A	5178T
82	5231	G	A	5231A	5231C
83	5263	C	T	5263T	none
84	5360	C	T	5360T	none
85	5390	A	G	5390G	none
86	5442	T	C	5442C	none
87	5495	T	C	5495C	none
88	5633	C	T	5633T	none
89	5656	A	G	5656G	none
90	5773	G	A	5773A	none
91	5951	A	G	5951G	none
92	5999	T	C	5999C	5999A
93	6047	A	G	6047G	none
94	6185	T	C	6185C	none
95	6221	T	C	6221C	6221A
96	6253	T	C	6253C	none
97	6371	C	T	6371T	none
98	6386	C	T	6386T	none
99	6455	C	T	6455T	none
100	6528	C	T	6528T	none

	UKB. Microarray Info.			Output of *in silico* genotyping stage 1	
	POS	**REF**	**ALT**	**Alternative Variants**	**Neither Variants**
101	6734	G	A	6734A	none
102	6752	A	G	6752G	6752C
103	6776	T	C	6776C	none
104	7028	C	T	7028T	7028G
105	7055	A	G	7055G	none
106	7175	T	C	7175C	none
107	7476	C	T	7476T	none
108	7645	T	C	7645C	7645A
109	7657	T	C	7657C	none
110	7768	A	G	7768G	none
111	7864	C	T	7864T	none
112	8269	G	A	8269A	none
113	8344	A	G	8344G	none
114	8448	T	C	8448C	none
115	8616	G	T	8616T	8616A
116	8655	C	T	8655T	none
117	8697	G	A	8697A	8697C
118	8869	A	G	8869G	8869C
119	8993	T	G	8993G	8993C
120	8994	G	A	8994A	8994T
121	9042	C	T	9042T	none
122	9055	G	A	9055A	9055T
123	9072	A	G	9072G	none
124	9093	A	G	9093G	9093C
125	9123	G	A	9123A	9123C
126	9221	A	G	9221G	none
127	9377	A	G	9377G	none
128	9647	T	C	9647C	9647A
129	9667	A	G	9667G	none
130	9698	T	C	9698C	none
131	9716	T	C	9716C	9716A
132	9899	T	C	9899C	none
133	9950	T	C	9950C	none
134	10044	A	G	10044G	none
135	10142	C	T	10142T	none
136	10217	A	G	10217G	none
137	10238	T	C	10238C	none
138	10310	G	A	10310A	10310C
139	10321	T	C	10321C	none
140	10394	C	T	10394T	10394A 10394G

	UKB. Microarray Info.			Output of *in silico* genotyping stage 1	
	POS	**REF**	**ALT**	**Alternative Variants**	**Neither Variants**
141	10398	A	G	10398G	10398T
142	10400	C	T	10400T	10400A
143	10463	T	C	10463C	10463A
144	10550	A	G	10550G	none
145	10688	G	A	10688A	none
146	10810	T	C	10810C	10810A
147	10819	A	G	10819G	none
148	10873	T	C	10873C	none
149	10915	T	C	10915C	none
150	11025	T	C	11025C	none
151	11251	A	G	11251G	none
152	11299	T	C	11299C	none
153	11332	C	T	11332T	11332G
154	11377	G	A	11377A	none
155	11467	A	G	11467G	none
156	11674	C	T	11674T	none
157	11778	G	A	11778A	none
158	11812	A	G	11812G	11812C
159	11899	T	C	11899C	none
160	11914	G	A	11914A	11914C
161	11947	A	G	11947G	none
162	12372	G	A	12372A	none
163	12397	A	G	12397G	none
164	12406	G	A	12406A	none
165	12501	G	A	12501A	12501C
166	12612	A	G	12612G	12612T
167	12630	G	A	12630A	none
168	12633	C	A	12633A	12633T
169	12669	C	T	12669T	none
170	12705	C	T	12705T	12705A
171	12810	A	G	12810G	none
172	12850	A	G	12850G	none
173	12879	T	C	12879C	none
174	12950	A	G	12950G	12950C
175	13020	T	C	13020C	13020A 13020G
176	13101	A	C	13101C	13101G
177	13104	A	G	13104G	none
178	13105	A	G	13105G	none
179	13117	A	G	13117G	none
180	13263	A	G	13263G	none

	UKB. Microarray Info.			Output of *in silico* genotyping stage 1	
	POS	**REF**	**ALT**	**Alternative Variants**	**Neither Variants**
181	13500	T	C	13500C	none
182	13506	C	T	13506T	none
183	13617	T	C	13617C	none
184	13650	C	T	13650T	none
185	13708	G	A	13708A	none
186	13759	G	A	13759A	none
187	13780	A	G	13780G	13780C
188	13789	T	C	13789C	none
189	13879	T	C	13879C	13879A
190	13965	T	C	13965C	none
191	13966	A	G	13966G	none
192	14016	G	A	14016A	none
193	14070	A	G	14070G	14070T
194	14094	T	C	14094C	none
195	14133	A	G	14133G	14133C
196	14139	A	G	14139G	none
197	14167	C	T	14167T	none
198	14178	T	C	14178C	none
199	14318	T	C	14318C	none
200	14484	T	C	14484C	none
201	14550	T	C	14550C	none
202	14552	A	G	14552G	none
203	14582	A	G	14582G	none
204	14620	C	T	14620T	none
205	14674	T	C	14674C	none
206	14798	T	C	14798C	none
207	14869	G	A	14869A	14869C
208	14872	C	T	14872T	14872A
209	14905	G	A	14905A	none
210	15043	G	A	15043A	none
211	15148	G	A	15148A	none
212	15218	A	G	15218G	15218C
213	15244	A	G	15244G	15244T
214	15250	C	T	15250T	none
215	15257	G	A	15257A	none
216	15301	G	A	15301A	none
217	15452	C	A	15452A	15452T
218	15454	T	C	15454C	none
219	15535	C	T	15535T	none
220	15693	T	C	15693C	none

	UKB. Microarray Info.			Output of *in silico* genotyping stage 1	
	POS	**REF**	**ALT**	**Alternative Variants**	**Neither Variants**
221	15758	A	G	15758G	none
222	15784	T	C	15784C	none
223	15812	G	A	15812A	15812C
224	15833	C	T	15833T	none
225	15884	G	A	15884A	15884C 15884T
226	15924	A	G	15924G	none
227	15928	G	A	15928A	15928T
228	15930	G	A	15930A	15930T
229	15946	C	T	15946T	15945CT
230	16144	T	C	16144C	16144A
231	16145	G	A	16145A	none
232	16148	C	T	16148T	16148G
233	16153	G	A	16153A	none
234	16193	C	T	16193T	16193CC 16193CCC 16192CT 16193CCCC 16193CTC 16192CY 16193A 16193CTT 16192CG
235	16243	T	C	16243C	16243A
236	16270	C	T	16270T	16270G 16270A
237	16337	C	T	16337T	none
238	16356	T	C	16356C	none
239	16362	T	C	16362C	16362A 16362G 16362d.T
240	16391	G	A	16391A	16391GG 16391T

Exploration: Macrobranch O

Macrobranch O is represented by just 8 samples in the MitoMap Haplogroup library. The table below contains the pattern of `alternative` calls in the barcodes of these samples. A single sample (numbered 1 and marked *) was chosen, at random, for the library-test set and it failed to find any matches to its barcode despite other haplogroup members.

Columns 1-8 represent the eight mtDNA samples of the MitoMap library which from the O macrobranch. Table 4.2 only compares a subset of target loci for which `alternative` calls were made for one our more of the samples in the O macrobranch.

		1	2	3	4	5	6	7	8	
	POS	**O***	**O**	**O**	**O**	**O**	**O1**	**O1a**	**O1a**	
1	73	ALT	ALT	ALT	ALT	ALT	ALT	ALT	ALT	Needed for O
2	150				**ALT**					**Possible problem**
5	263	ALT	ALT	ALT	ALT	ALT	ALT	ALT	ALT	Needed for O
10	750	ALT	ALT	ALT	ALT	ALT	ALT	ALT	ALT	Needed for O
18	1438	ALT	ALT	ALT	ALT	ALT	ALT	ALT	ALT	Needed for O
27	2706	ALT	ALT	ALT	ALT	ALT	ALT	ALT	ALT	Needed for O
72	4769	ALT	ALT	ALT	ALT	ALT	ALT	ALT	ALT	Needed for O
75	4824	**ALT**								**Possible problem**
80	5147						ALT			Needed for O1 only?
95	6221						ALT	ALT	ALT	Needed for O1
104	7028	ALT	ALT	ALT	ALT	ALT	ALT	ALT	ALT	Needed for O
170	12705	ALT	ALT	ALT	ALT	ALT	ALT	ALT	ALT	Needed for O
185	13708	**ALT**								**Possible problem**
226	15924							ALT	ALT	Needed for O1a
231	16145	**ALT**								**Possible problem**
239	16362						ALT	ALT	ALT	Needed for O1

Table 4.2: Table of the eight members of the O macrogroup to allow a comparison of their differing barcodes

The target loci discriminate the samples from Macrobranch O very well with eight loci (1, 5, 10, 18, 27, 72, 104 and 170) which are common to all eight samples. `Alternative` calls in these loci are present in all samples in the O macrobranch but also other samples, when not combined with additional `alternative` calls at other loci. Three haplogroup O samples (numbered 2,3 and 5) find 43 matches in macrobranches N, N1, O and S. These matches were found using the regular expression search tool but 41 of these matches were identical to the macrobranch O pattern. Sample 4 finds 17 matches in macrobranches N, O and S, with the O and S matches overlapping, but not matches in the N group. This identical barcode was also found in 15 MitoMap Haplogroup library samples.

Using solely the target loci, to be a member of haplogroup O1 or its subgroup, O1a, two further `alternative` calls are found at 95 and 239. To specify membership of haplogroup O1a, barcodes gain another `alternative` call at locus 226. Locus 80 appears to be only in the O1 sample which could represent a gain-loss in quick succession, a variation gained after the groups split, or a private mutation in this one sample. There is not enough data to decide.

Finally, comparing samples in macrobranch O also reveals four loci which are problematic, marked in bold. Sample 4 holds an additional `alternative` call at 150 which acts to prevent it matching with the other members of the haplogroup. Sample 1 hold three spurious loci (at 75, 185, 231). These prevent its matching any other MitoMap Haplogroup library samples and may represent a mistake in classification as haplogroup O, which should represent a more general classification as belonging to macrobranch O.

Having three spurious call in a single sample point to issues in the *in silico* genotyping process which need addressing. Comparing common macrobranch members or members of the same haplogroup offers a good way of finding problem loci, which would lead to problem variants. An analysis of this type would improve *in silico* genotyping but is beyond the reach of the current project.

Exploration: Samples with Low Variant Numbers

The MitoMap Haplogroup data set was found to contain a small number of samples with an anomalously low number of variants given their haplogroup. A brief exploration of the implications of their inclusion in the MitoMap Haplogroup library data set I used for the project is below.

Firstly, the 44 samples (0.1%) with a low variant number (LVN) were gathered using thresholds of fewer than 7 variants in samples supposed to have a distance of more than 2 nodes from the rCRS template sequence, see figure 4.1.

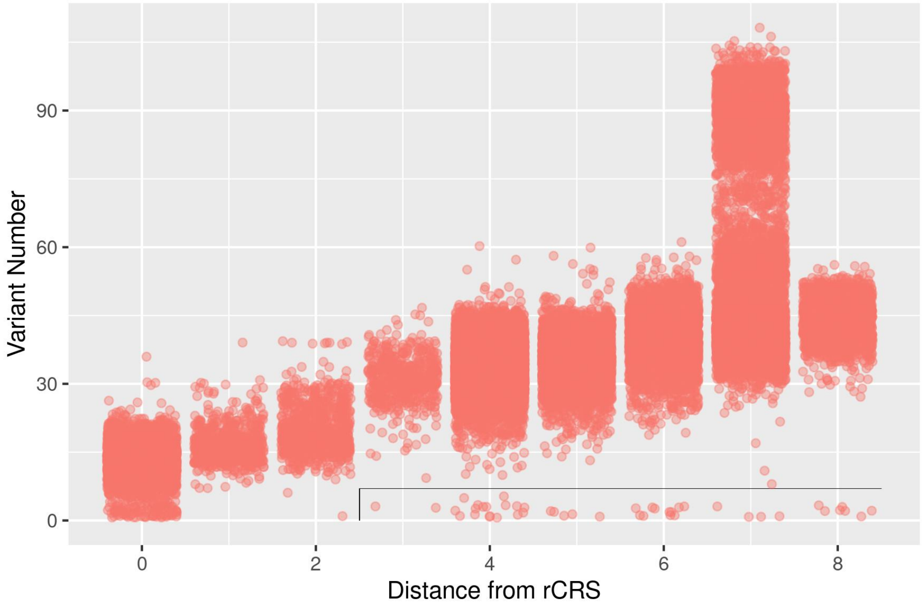


Figure 4.1: Thresholds used for selecting the MitoMap Haplogroup library samples with low variant numbers

Many of the LVN samples had haplogroup assignment in common with other samples. The variant numbers of samples in these haplogroups were compared and the anomalousness of the LVN samples was confirmed (see figure 4.2).

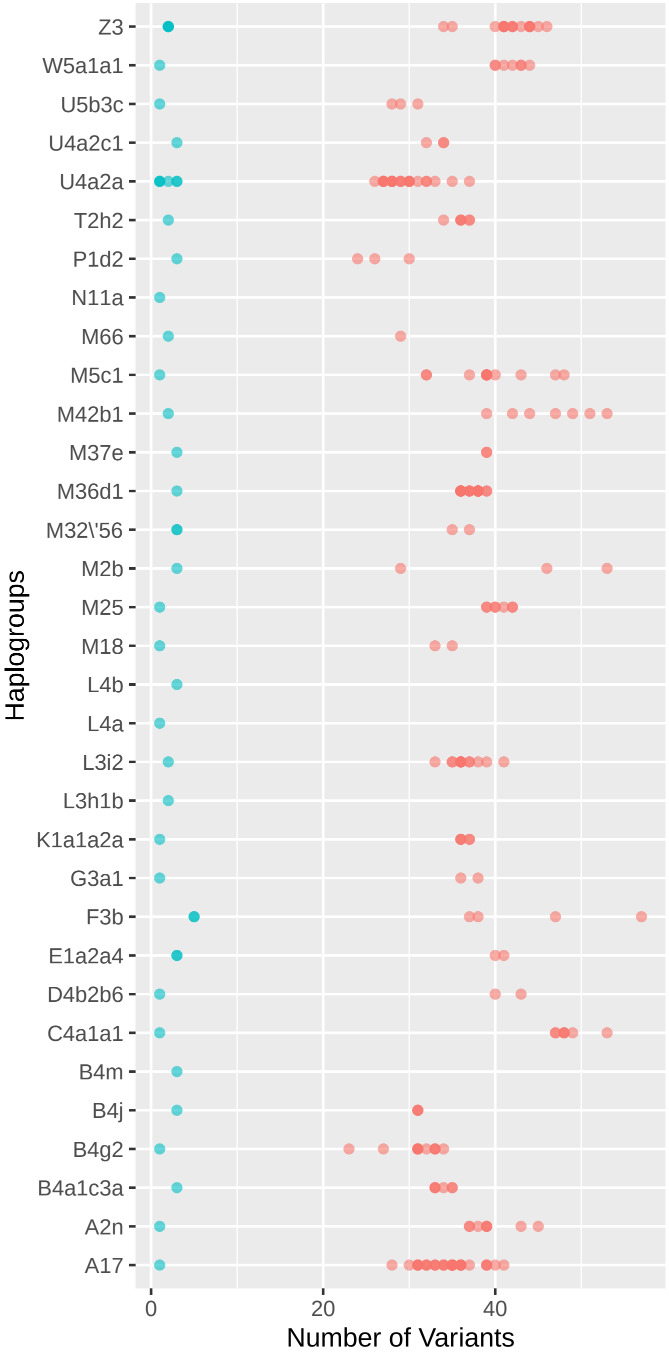


Figure 4.2: Comparison of all samples in haplogroups with anomalous variant numbers. Samples in cyan have anomalously low numbers of variants

Of the 46 LVN samples, 12 were used in the library-test set, receiving poor scores for all tests, which may pull down haplogroup and macrobranch average values but should not be too problematic. However, the 34 samples which remained in the MitoMap Haplogroup library-training data have influenced the predictions of 1357 samples. The scores for test 1 were depressed in samples with one or more LVN samples in their prediction. The other scores were, however, improved.

I cannot pursue this any further here but just removing the LVN samples from the MitoMap Haplogroup library before making the division into test and train is possible and should be followed by re-analysis.

Functions

Variant Finder

```
# takes a VCF-style table and creates a list of all variants

Variant_finder_VCF = function(t){
        pos =  t$POS
        ref = t$REF
        alt = t$ALT
        temp_l = c()
        for (x in 1:length(pos)){
                p = pos[[x]]
                r = ref[[x]]
                pr = paste(p,r, sep = "", collapse = "")
                temp_l = c(temp_l,pr)
                a = strsplit(alt[[x]], split = ",")
                a = a[[1]]
                for (y in a){
                        py = paste(p,y, sep = "", collapse = "")
                        temp_l = c(temp_l,py)
                        }
                }
        return(temp_l)
}

#this function accepts a column of stringed lists and
#returns a single, unique list of all found variants.

Variant_finder_Lists = function(c, split){
        l = c()
        for (x in c){
                x = strsplit(x, split = split)
                x = x[[1]]
                l = c(l, x)
                l = unique(l)
                }
        return(l)
}
```

Matrix Process Fixed

```
# these subroutines build to a function which takes a
# string and return a list of "information units".

matrix_coding = function(HG){
        library(stringr)
        l = nchar(HG)
        blocks = c()
        codes = c()
        for (n in 1:l){
                block = str_sub(HG,n,n)
                blocks = c(blocks, block)
        word_char = str_detect(block,"\\w")
        digit = str_detect(block,"\\d")
        code = 0
        if(word_char == TRUE){
                code = code + 1
                if(digit == TRUE){
                        code = code + 1
        }}
        codes = c(codes, code)
        }
        df = data.frame(blocks,codes,stringsAsFactors = FALSE)
        return(df)
}

# code to find the end point of the next chunk

return_pos = function(p,m,n,l){
        # looks at the status code for the start point
        b = m[n,2]
        # looks at the status code of the possible end point
        new_b = m[p,2]
        # compares the codes
        if (b != new_b){
                return(p-1)} else {
                        p = p + 1
                        if (p > l){
                                return(p-1)} else {
                                        return_pos(p,m,n,l)}}
}

matrix_process_fixed = function(HG){
        cl = c()
        # chunk list holder
        matrix_y = matrix_coding(HG)
        # creates matrix of decisions
        l = length(matrix_y[,1])
        # finds length of matrix, i.e. the number of pieces in the string
        x = 1
        z = 1
        while ((z <= l) == TRUE) {
                # until start point is longer than matrix
                z = return_pos(z, matrix_y, x, l)
                # finds end point of that block
                block = str_sub(HG,x,z)
                # cuts chunk
                cl = c(cl,block)
                # stores chunk
                x = z + 1
                z = x
        }
        cl2 = c()
        for (b in cl){
                if (str_detect(b,"\\w") == TRUE){cl2 = c(cl2,b)}
        }
        return(cl2)
}
```

Array Maker

```
# This function takes the list supplied by either
# variant_finders, and returns a df of list/locus/
# base(s)/end

array_maker = function(l){
        library(stringr)
        loci = c()
        bases = c()
        ends = c()
        for (x in l){
                x = matrix_process_fixed(x)
                loci = c(loci, as.numeric(x[[1]]))
                if (length(x) > 1){
                        bases = c(bases, x[[2]])
                        b = strsplit(x[[2]], split = "")
                        b = b[[1]]
                        ends = c(ends, (as.integer(x[[1]]) + length(b)-1))}
                                else {
                                        bases = c(bases, "none")
                                        ends = c(ends, as.integer(x[[1]]))}
                        }
        a = data.frame(l, loci, bases, ends, stringsAsFactors = FALSE)
        return(a)
}
```

Find RAN

```
# defining if the base is REF/ALT or Neither.

find_RAN = function(a,t){
        ran = c()
        for (x in 1:length(a[,1])){
                ind = which(t$POS == a[x,4])
                ref = t$REF
                ref = ref[[ind]]
                alt = t$ALT
                alt = alt[[ind]]
                if (a[x,5] == ref){add = "ref"} else {
                        if(a[x,5] == alt){add = "alt"}
                                else {add = "neither"}}
                ran = c(ran, add)
                }
        r = data.frame(a,ran, stringsAsFactors = FALSE)
        return(r)
}
```

Pre-Barcoding

```
# Gathers list of variants which are ALT or NEITHER
# for each of the target loci listed.

pre_barcoding = function(a, t){
        Alt_variants = c()
        Neither_variants = c()
        for (x in t$POS){
                ind = which(a$target_match == x)
                if (length(ind) == 0){v_a = "none"; v_n = "none"} else {
                        temp = a[c(ind),]
                        ind_a = which(temp$ran == "alt")
                        ind_n = which(temp$ran == "neither")
                        if (length(ind_a) == 0){v_a = "none"}
                                else {
                                        v_a = temp[c(ind_a),1]
                                        v_a = paste(v_a, sep = "␣",
                                                collapse = "␣")}
                        if (length(ind_n) == 0){v_n = "none"}
                                else {
                                        v_n = temp[c(ind_n),1]
                                        v_n = paste(v_n, sep = "␣",
                                                collapse = "␣")}}
        Alt_variants = c(Alt_variants, v_a)
        Neither_variants = c(Neither_variants, v_n)}
        df = data.frame(t,Alt_variants, Neither_variants,
                stringsAsFactors = FALSE)
        return(df)
}
```

Barcoding

```
# this first function will accept a list of samples
# variants, and compare it to a list of alt and neither
# variants, returning a list of number of intersect
# between two lists and the samples list.

barcoding_1 = function(sv,a,n){
        sv = strsplit(sv, split = "␣")
        sv = sv[[1]]
        a = strsplit(a, split = "␣")
        a = a[[1]]
        n = strsplit(n, split = "␣")
        n = n[[1]]
        x = intersect(sv,a)
        y = intersect(sv,n)
        r = c(length(x),length(y))
        return(r)
}

# deals with output of Barcoding_1; by retunrning
# fails for problems and 0/1/. appropriately.
# Returns a list of barcodes.

barcoding_2 = function(dl,t){
        library(stringr)
        r = c()
        for (c in 1:length(t[,1])){
                result = barcoding_1(b,t[c,4],t[c,5])
                add = "bob"
                n = result[[1]] + result[[2]]
                if (n > 1){add = "fail"} else {
                        if (n == 0){add = 0} else {
                                if (result[[1]] == 1){add = 1}
                                        else {if(result[[2]]==1){add="."}}}}
                if (add == "bob"){add = "fail2"}
                l = c(l,add)}
        l = paste(l, sep = "", collapse = "")
        r = c(r,l)
        return(r)
}
```

Extract Alternative and Neither

```
# I now need to process for lists for each target
# locus.

Extract_ALT_NEITHER = function(a,t){
        alt_variants = c()
        neither_variants = c()
        for (x in t$POS){
                ind = which(a$target_match == x)
                if (length(ind) < 1){alt = "none"; nei = "none"}
                        else {temp = a[c(ind),]
                                a_ind = which(temp$ran == "alt")
                if (length(a_ind) < 1){alt = "none"}
                else {
                        alt = temp[c(a_ind),1];
                        alt = paste(alt, sep = "␣", collapse = "␣")}
                n_ind = which(temp$ran == "neither")
        if (length(n_ind) < 1){nei = "none"}
                else {
                        nei = temp[c(n_ind),1];
                        nei = paste(nei, sep = "␣", collapse = "␣")}}
        alt_variants = c(alt_variants, alt)
        neither_variants = c(neither_variants, nei)}
        r = data.frame(t$POS,alt_variants,neither_variants,
                stringsAsFactors = FALSE)
        return(r)
}
```

Bibliography

[1] https://cran.r-project.org/web/packages/ggraph/index.html.

[2] https://cran.r-project.org/web/packages/stringr/index.html.

[3] https://www.ebi.ac.uk/gwas/.

[4] https://www.genome.gov/27541954/dna-sequencing-costs-data/.

[5] https://www.mitomap.org/mitomap.

[6] https://www.ncbi.nlm.nih.gov/genbank/.

[7] https://www.sanger.ac.uk/collaboration/uk-biobank-whole-genome-sequencing-project/.

[8] Faraz Ahmad, Widyan Alamoudi, Shafiul Haque, Mohammad Salahuddin, and Khaldoon Alsamman. Simple, reliable, and time-efficient colorimetric method for the assessment of mitochondrial function and toxicity. *Bosnian Journal of Basic Medical Sciences*, 18(4):367, 2018.

[9] B Alberts, D Bray, J Lewis, M Raff, K Roberts, and JD Watson. Molecular biology of the cell, 3rd addition. *New York: Garland Science*, 1994.

[10] Richard Altmann. *Die Elementarorganismen und ihre Beziehungen zu den Zellen*. Veit, 1894.

[11] Sharon Anderson, Alan T Bankier, Bart G Barrell, Maarten HL de Bruijn, Alan R Coulson, Jacques Drouin, Ian C Eperon, Donald P Nierlich, Bruce A Roe, Frederick Sanger, et al. Sequence and organization of the human mitochondrial genome. *Nature*, 290(5806):457–465, 1981.

[12] GE Andersson, Olof Karlberg, Björn Canbäck, and Charles G Kurland. On the origin of mitochondria: a genomics perspective. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 358(1429):165–179, 2003.

[13] Richard M Andrews, Iwona Kubacka, Patrick F Chinnery, Robert N Lightowlers, Douglass M Turnbull, and Neil Howell. Reanalysis and revision of the cambridge reference sequence for human mitochondrial dna. *Nature genetics*, 23(2):147–147, 1999.

[14] Giovanni Felice Azzone and Lars Ernster. Respiratory control and compartmentation of substrate level phosphorylation in liver mitochondria. *Journal of Biological Chemistry*, 236(5):1501–1509, 1961.

[15] Carl Benda. Ueber die spermatogenese der vertebraten und höherer evertebraten, ii. theil: Die histiogenese der spermien. *Arch Anat Physiol*, 73:393–398, 1898.

[16] Robert R Bensley and Normand L Hoerr. Studies on cell structure by the freezing-drying method vi. the preparation and properties of mitochondria. *The anatomical record*, 60(4):449–455, 1934.

[17] Carsten Bornhövd, Frank Vogel, Walter Neupert, and Andreas S Reichert. Mitochondrial membrane potential is dependent on the oligomeric state of f1f0-atp synthase supracomplexes. *Journal of Biological Chemistry*, 281(20):13990–13998, 2006.

[18] Marty C Brandon, Marie T Lott, Kevin Cuong Nguyen, Syawal Spolim, Shamkant B Navathe, Pierre Baldi, and Douglas C Wallace. Mitomap: a human mitochondrial genome database—2004 update. *Nucleic acids research*, 33(suppl_1):D611–D613, 2005.

[19] Brian L Browning and Sharon R Browning. Genotype imputation with millions of reference samples. *The American Journal of Human Genetics*, 98(1):116–126, 2016.

[20] Stephen P Burr, Mikael Pezet, and Patrick F Chinnery. Mitochondrial dna heteroplasmy and purifying selection in the mammalian female germ line. *Development, growth & differentiation*, 60(1):21–32, 2018.

[21] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. Genome-wide genetic data on˜ 500,000 uk biobank participants. *BioRxiv*, page 166298, 2017.

[22] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[23] Rebecca L Cann, Mark Stoneking, and Allan C Wilson. Mitochondrial dna and human evolution. *Nature*, 325(6099):31–36, 1987.

[24] Manon Carré, Nicolas André, Gérard Carles, Hélène Borghi, Laetitia Brichese, Claudette Briand, and Diane Braguer. Tubulin is an inherent component of mitochondrial membranes that interacts with the voltage-dependent anion channel. *Journal of Biological Chemistry*, 277(37):33664–33669, 2002.

[25] Patrick F Chinnery, Neil Howell, Robert N Lightowlers, and Douglass M Turnbull. Molecular pathology of melas and merrf. the relationship between mutation load and clinical phenotypes. *Brain: a journal of neurology*, 120(10):1713–1721, 1997.

[26] Patrick F Chinnery and Douglass M Turnbull. Epidemiology and treatment of mitochondrial disorders. *American journal of medical genetics*, 106(1):94–101, 2001.

[27] Ananyo Choudhury, Shaun Aron, Laura R Botigué, Dhriti Sengupta, Gerrit Botha, Taoufik Bensellak, Gordon Wells, Judit Kumuthini, Daniel Shriner, Yasmina J Fakim, et al. High-depth african genomes inform human migration and health. *Nature*, 586(7831):741–748, 2020.

[28] Pinar E Coskun, Eduardo Ruiz-Pesini, and Douglas C Wallace. Control region mtdna variants: longevity, climatic adaptation, and a forensic conundrum. *Proceedings of the National Academy of Sciences*, 100(5):2174–2176, 2003.

[29] Samarjit Das, Charles Steenbergen, and Elizabeth Murphy. Does the voltage dependent anion channel modulate cardiac ischemia–reperfusion injury? *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1818(6):1451–1456, 2012.

[30] Karen M Davies, Claudio Anselmi, Ilka Wittig, José D Faraldo-Gómez, and Werner Kühlbrandt. Structure of the yeast f1fo-atp synthase dimer and its role in shaping the mitochondrial cristae. *Proceedings of the National Academy of Sciences*, 109(34):13602–13607, 2012.

[31] Xiaohong Deng, Dongmei Ji, Xinyuan Li, Yuping Xu, Yu Cao, Weiwei Zou, Chunmei Liang, Jordan Lee Marley, Zhiguo Zhang, Zhaolian Wei, et al. Polymorphisms and haplotype of mitochondrial dna d-loop region are associated with polycystic ovary syndrome in a chinese population. *Mitochondrion*, 57:173–181, 2021.

[32] Salvatore DiMauro and Guido Davidzon. Mitochondrial dna and disease. *Annals of medicine*, 37(3):222–232, 2005.

[33] Jackie N Doda, Catharine T Wright, and David A Clayton. Elongation of displacement-loop strands in human and mouse mitochondrial dna is arrested near specific template sequences. *Proceedings of the National Academy of Sciences*, 78(10):6116–6120, 1981.

[34] Michal Eisenberg-Bord and Maya Schuldiner. Ground control to major tom: mitochondria–nucleus communication. *The FEBS journal*, 284(2):196–210, 2017.

[35] JL Elson, DM Turnbull, and Neil Howell. Comparative genomics and the evolution of human mitochondrial dna: assessing the effects of selection. *The American Journal of Human Genetics*, 74(2):229–238, 2004.

[36] Lars Ernster and Gottfried Schatz. Mitochondria: a historical review. *J Cell Biol*, 91(3):227s–255s, 1981.

[37] Long Fan and Yong-Gang Yao. Mitotool: a web server for the analysis and retrieval of human mitochondrial dna sequence variations. *Mitochondrion*, 11(2):351–356, 2011.

[38] SA Frank and LD Hurst. Mitochondria and male disease. *Nature*, 383(6597):224, 1996.

[39] Terrence G Frey and Carmen A Mannella. The internal structure of mitochondria. *Trends in biochemical sciences*, 25(7):319–324, 2000.

[40] Neil J Gemmell, Victoria J Metcalf, and Fred W Allendorf. Mother's curse: the effect of mtdna on individual fitness and population viability. *Trends in Ecology & Evolution*, 19(5):238–244, 2004.

[41] Robert Christopher Gilson and Sandra Osswald. Madelung lipomatosis presenting as a manifestation of myoclonic epilepsy with ragged red fibers (merrf) syndrome. *JAAD case reports*, 4(8):822, 2018.

[42] Aurora Gómez-Durán, David Pacheu-Grau, Íñigo Martínez-Romero, Ester López-Gallardo, Manuel J López-Pérez, Julio Montoya, and Eduardo Ruiz-Pesini. Oxidative phosphorylation differences between mitochondrial dna haplogroups modify the risk of leber's hereditary optic neuropathy. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1822(8):1216–1222, 2012.

[43] Vanessa F Gonçalves, Stephanie N Giamberardino, James J Crowley, Marquis P Vawter, Richa Saxena, Cynthia M Bulik, Zeynep Yilmaz, Christina M Hultman, Pamela Sklar, James L Kennedy, et al. Examining the role of common and rare mitochondrial variants in schizophrenia. *PloS one*, 13(1), 2018.

[44] Michael W Gray, B Franz Lang, Robert Cedergren, G Brian Golding, Claude Lemieux, David Sankoff, Monique Turmel, Nicolas Brossard, Eric Delage, Tim G Littlejohn, et al. Genome structure and gene content in protist mitochondrial dnas. *Nucleic acids research*, 26(4):865–878, 1998.

[45] Yongtao Guan and Matthew Stephens. Practical issues in imputation-based association mapping. *PLoS Genet*, 4(12):e1000279, 2008.

[46] Erik Hagström, Christoph Freyer, Brendan J Battersby, James B Stewart, and Nils-Göran Larsson. No recombination of mtdna after heteroplasmy for 50 generations in the mouse maternal germline. *Nucleic acids research*, 42(2):1111–1116, 2013.

[47] Dorothy R Haskett. Mitochondrial dna (mtdna). *Embryo Project Encyclopedia*, 2014.

[48] Geoffrey E Hill. Mitonuclear ecology. *Molecular Biology and Evolution*, 32(8):1917–1927, 2015.

[49] George H Hogeboom, Walter C Schneider, and George E Pallade. Cytochemical studies of mammalian tissues i. isolation of intact mitochondria from rat liver; some biochemical properties of mitochondria and submicroscopic particulate material. *Journal of Biological Chemistry*, 172(2):619–635, 1948.

[50] Neil Howell, Joanna L Elson, Corinna Howell, and Douglass M Turnbull. Relative rates of evolution in the coding and control regions of african mtdnas. *Molecular biology and evolution*, 24(10):2213–2221, 2007.

[51] Bryan N Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, 5(6):e1000529, 2009.

[52] Martijn A Huynen, Mareike Mühlmeister, Katherina Gotthardt, Sergio Guerrero-Castillo, and Ulrich Brandt. Evolution and structural organization of the mitochondrial contact site (micos) complex and the mitochondrial intermembrane space bridging (mib) complex. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1863(1):91–101, 2016.

[53] Francisco J Iborra, Hiroshi Kimura, and Peter R Cook. The functional organization of mitochondrial genomes in human cells. *BMC biology*, 2(1):9, 2004.

[54] Michael Inouye, Gad Abraham, Christopher P Nelson, Angela M Wood, Michael J Sweeting, Frank Dudbridge, Florence Y Lai, Stephen Kaptoge, Marta Brozynska, Tingting Wang, et al. Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *Journal of the American College of Cardiology*, 72(16):1883–1893, 2018.

[55] Koji Ishiya, Fuzuki Mizuno, Li Wang, and Shintaroh Ueda. Mitoimp: A computational framework for imputation of missing data in low-coverage human mitochondrial genome. *Bioinformatics and biology insights*, 13:1177932219873884, 2019.

[56] Uwe John, Yameng Lu, Sylke Wohlrab, Marco Groth, Jan Janouškovec, Gurjeet S Kohli, Felix C Mark, Ulf Bickmeyer, Sarah Farhat, Marius Felder, et al. An aerobic eukaryotic parasite with functional mitochondria that likely lacks a mitochondrial genome. *Science advances*, 5(4):eaav1110, 2019.

[57] Allison A Johnson and Kenneth A Johnson. Exonuclease proofreading by human mitochondrial dna polymerase. *Journal of Biological Chemistry*, 276(41):38097–38107, 2001.

[58] Kyunga Kim, Shlomit Kenigsberg, Andrea Jurisicova, and Yaakov Bentov. The role of mitochondria in oocyte and early embryo health. *OBM Genetics*, 3:1–1, 2019.

[59] Turi E King, Gloria Gonzalez Fortes, Patricia Balaresque, Mark G Thomas, David Balding, Pierpaolo Maisano Delser, Rita Neumann, Walther Parson, Michael Knapp, Susan Walsh, et al. Identification of the remains of king richard iii. *Nature communications*, 5(1):1–8, 2014.

[60] BF Kingsbury. Cytoplasmic fixation. *The Anatomical Record*, 6(2):39–52, 1912.

[61] Klaus V Kowallik and William F Martin. The origin of symbiogenesis: An annotated english translation of mereschkowsky's 1910 paper on the theory of two plasma lineages. *Biosystems*, 199:104281.

[62] Werner Kühlbrandt. Structure and function of mitochondrial membrane protein complexes. *BMC biology*, 13(1):89, 2015.

[63] Christian Kukat, Karen M Davies, Christian A Wurm, Henrik Spåhr, Nina A Bonekamp, Inge Kühl, Friederike Joos, Paola Loguercio Polosa, Chan Bae Park, Viktor Posse, et al. Cross-strand binding of tfam to a single mtdna molecule forms the mitochondrial nucleoid. *Proceedings of the National Academy of Sciences*, 112(36):11288–11293, 2015.

[64] Christian Kukat, Christian A Wurm, Henrik Spåhr, Maria Falkenberg, Nils-Göran Larsson, and Stefan Jakobs. Super-resolution microscopy reveals that mammalian mitochondrial nucleoids have a uniform size and frequently contain a single copy of mtdna. *Proceedings of the National Academy of Sciences*, 108(33):13534–13539, 2011.

[65] Emmanuel D Ladoukakis and Eleftherios Zouros. Evolution and inheritance of animal mitochondrial dna: rules and exceptions. *Journal of Biological Research-Thessaloniki*, 24(1):2, 2017.

[66] Maxime Lagouge and N-G Larsson. The role of mitochondrial dna mutations and free radicals in disease and ageing. *Journal of internal medicine*, 273(6):529–543, 2013.

[67] Conor Lawless, Laura Greaves, Amy K Reeve, Doug M Turnbull, and Amy E Vincent. The rise and rise of mitochondrial dna mutations. *Open biology*, 10(5):200061, 2020.

[68] Arnold Lazarow and SJ Cooperstein. Studies on the enzymatic basis for the janus green b staining reaction. *Journal of Histochemistry & Cytochemistry*, 1(4):234–241, 1953.

[69] M Levy, R Toury, and J Andre. Purification and enzymatic characterization of the external membrane of mitochondria. *Comptes rendus hebdomadaires des seances de l'Academie des sciences. Serie D: Sciences naturelles*, 263(22):1766, 1966.

[70] Margaret Reed Lewis and Warren Harmon Lewis. Mitochondria (and other cytoplasmic structures) in tissue cultures. *American Journal of Anatomy*, 17(3):339–401, 1915.

[71] Yun Li, Cristen J Willer, Jun Ding, Paul Scheet, and Gonçalo R Abecasis. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, 34(8):816–834, 2010.

[72] Rolf Luft, Denis Ikkos, Genaro Palmieri, Lars Ernster, Björn Afzelius, et al. A case of severe hypermetabolism of nonthyroid origin with a defect in the maintenance of mitochondrial respiratory control: a correlated clinical, biochemical, and morphological study. *The Journal of clinical investigation*, 41(9):1776–1804, 1962.

[73] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, et al. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, 45(D1):D896–D901, 2016.

[74] Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511, 2010.

[75] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906, 2007.

[76] Lynn Margulis. Symbiosis in cell evolution: Life and its environment on the early earth. 1981.

[77] Leonor Michaelis. Die vitale färbung, eine darstellungsmethode der zellgranula. *Archiv für mikroskopische Anatomie*, 55(1):558–575, 1899.

[78] Peter Mitchell. Chemiosmotic coupling in oxidative and photosynthetic phosphorylation, 1966.

[79] Hermann Joseph Muller. The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 1(1):2–9, 1964.

[80] Johannes Müller and Jacob Henle. *Systematische beschreibung der Plagiostomen*, volume 3. Veit, 1841.

[81] Margit MK Nass and Sylvan Nass. Intramitochondrial fibers with dna characteristics i. fixation and electron staining reactions. *Journal of Cell Biology*, 19(3):593–611, 1963.

[82] Brigitte Pakendorf and Mark Stoneking. Mitochondrial dna and human evolution. *Annu. Rev. Genomics Hum. Genet.*, 6:165–183, 2005.

[83] George E Palade. An electron microscope study of the mitochondrial structure. *Journal of Histochemistry & Cytochemistry*, 1(4):188–211, 1953.

[84] Brendan AI Payne, Ian J Wilson, Patrick Yu-Wai-Man, Jonathan Coxhead, David Deehan, Rita Horvath, Robert W Taylor, David C Samuels, Mauro Santibanez-Koref, and Patrick F Chinnery. Universal heteroplasmy of human mitochondrial dna. *Human molecular genetics*, 22(2):384–390, 2013.

[85] Alexander W Röck, Arne Dür, Mannis Van Oven, and Walther Parson. Concept for estimating mitochondrial dna haplogroups using a maximum likelihood approach (emma). *Forensic Science International: Genetics*, 7(6):601–609, 2013.

[86] Eduardo Ruiz-Pesini, Dan Mishmar, Martin Brandon, Vincent Procaccio, and Douglas C Wallace. Effects of purifying and adaptive selection on regional variation in human mtdna. *Science*, 303(5655):223–226, 2004.

[87] Michael T Ryan and Nicholas J Hoogenraad. Mitochondrial-nuclear communications. *Annu. Rev. Biochem.*, 76:701–722, 2007.

[88] Lynn Sagan. On the origin of mitosing cells. *Journal of Theoretical Biology*, 14(3):225 – IN6, 1967.

[89] Carl Schnaitman and John W Greenawalt. Enzymatic properties of the inner and outer membranes of rat liver mitochondria. *The Journal of cell biology*, 38(1):158–175, 1968.

[90] Nicholas J Schork, Sarah S Murray, Kelly A Frazer, and Eric J Topol. Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development*, 19(3):212–219, 2009.

[91] Inna Shokolenko, Glenn Wilson, and Mikhail Alexeyev. Aging: A mitochondrial dna perspective, critical analysis and an update. *World journal of experimental medicine*, 4:46–57, 11 2014.

[92] V Shoshan-Barmatz, A Israelson, D Brdiczka, , and SS Sheu. The voltage-dependent anion channel (vdac): function in intracellular signalling, cell life and cell death. *Current pharmaceutical design*, 12(18):2249–2270, 2006.

[93] Fritiof S Sjöstrand. Electron microscopy of mitochondria and cytoplasmic double membranes: ultra-structure of rod-shaped mitochondria. *Nature*, 171(4340):30–31, 1953.

[94] David Spiegelhalter. *The art of statistics: learning from data*. Penguin UK, 2019.

[95] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484, 2019.

[96] Wenzhi Tan and Marco Colombini. Vdac closure increases calcium ion flux. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1768(10):2510–2515, 2007.

[97] Jennifer EL Templeton, Paul M Brotherton, Bastien Llamas, Julien Soubrier, Wolfgang Haak, Alan Cooper, and Jeremy J Austin. Dna capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification. *Investigative genetics*, 4(1):1–13, 2013.

[98] Joelle M Van Der Walt, Kristin K Nicodemus, Eden R Martin, William K Scott, Martha A Nance, Ray L Watts, Jean P Hubble, Jonathan L Haines, William C Koller, Kelly Lyons, et al. Mitochondrial polymorphisms significantly reduce the risk of parkinson disease. *The American Journal of Human Genetics*, 72(4):804–811, 2003.

[99] Jan Van Leeuwen. *Handbook of theoretical computer science (vol. A) algorithms and complexity*. Mit Press, 1991.

[100] Mannis Van Oven and Manfred Kayser. Updated comprehensive phylogenetic tree of global human mitochondrial dna variation. *Human mutation*, 30(2):E386–E394, 2009.

[101] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[102] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[103] Douglas C Wallace. Mitochondrial defects in cardiomyopathy and neuromuscular disease. *American heart journal*, 139(2):s70–s85, 2000.

[104] Hansi Weissensteiner, Dominic Pacher, Anita Kloss-Brandstätter, Lukas Forer, Günther Specht, Hans-Jürgen Bandelt, Florian Kronenberg, Antonio Salas, and Sebastian Schönherr. Haplogrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic acids research*, 44(W1):W58–W63, 2016.

[105] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[106] Naoufal Zamzami and Guido Kroemer. The mitochondrion in apoptosis: how pandora's box opens. *Nature reviews Molecular cell biology*, 2(1):67–71, 2001.