Massively parallel sequencing of enriched mitochondrial DNA in patients with clinical suspicion of mitochondrial disease

Ibrahim Akkouh

Master's thesis

Department of Molecular Biosciences

UNIVERSITY OF OSLO

27.05.2015

År: 2015

Tittel: Massively parallel sequencing of enriched mitochondrial DNA in patients with clinical suspicion of mitochondrial disease

Forfatter: Ibrahim Akkouh

http://www.duo.uio.no/

Trykk: Reprosentralen, Universitetet i Oslo

Acknowledgements

Many people have contributed to this work, both directly and indirectly, and I want to show my gratitude to all of them, including the three study participants and their parents. I am especially grateful to my supervisor, Professor Eirik Frengen, for always being available and ready to help, and for continuously giving advice during the entire writing process. I am also thankful to Doriana Misceo, Tuva Barøy, and Christeen Pedurupillay for their help and guidance, both practically and theoretically. Tuva Barøy was instrumental in the identification of the *SURF1* variant. Asbjørn Holmgren deserves special thanks for doing all the hard computational work, always working behind the scenes. I also want to thank Nadia Skauli, Thilini Gamage, Johan Robert Helle, and Senait Samuel for their general support and good company. Siri Flåm and Marit Sletten were very helpful with all technical and practical matters relating to the lab, and they also deserve special thanks. Lastly, I want to show my gratitude to Gregor Duncan Gilfillan of the Norwegian Sequencing Center for helping me choose the right sequencing platform.

Abstract

Mitochondrial disorders constitute a clinically and genetically heterogeneous group of diseases that arise as a result of dysfunction of the mitochondrial respiratory chain. They can be caused by mutations in either the nuclear or mitochondrial DNA. In this study, three patients with clinical suspicion of mitochondrial disease were investigated by whole exome sequencing (WES), identifying the homozygous splice site mutation *SURF1* c.106+1G>C in patient 1. This mutation, which was demonstrated to cause skipping of exon 2 in *SURF1*, was concluded to be the genetic cause of the Leigh syndrome like phenotype in the patient. Patients 2 and 3 were further investigated by massively parallel sequencing (MSP) of mtDNA from fibroblasts using a PCR free protocol for the preparation of mtDNA enriched samples. Sequencing resulted in uniform and ultra-deep coverage (>40,000) of the entire mtDNA genome, demonstrating that the PCR free approach is the method of choice when analyzing mitochondrial DNA by MSP. In parallel, analysis of mtDNA from muscle from patient 3 revealed the novel and nearly homoplastic deletion m.6211_8681del, which was concluded to be the cause of disease in the patient. This deletion was also identified in mtDNA from the patient's blood and skin tissues, but it was present in too low levels of heteroplasmy to be detected by MSP.

Table of contents

1 Introduction

1.1 The scope of medical genetics

Medical genetics as we know it today had its beginning around 1900 when A. E. Garrod published his work on alkaptonuria [1]. Although many genetic disorders were known and described before that, for example Huntington's disease, Garrod was the first to publish an account of a case of recessive inheritance in humans and to suggest a molecular basis of the disease. During the following hundred years, medical genetics grew from a small field of research concerned with a few rare hereditary disorders to a specialty of medicine that involves the diagnosis and management of many diseases, both rare and common [2]. This development did not, however, start to accelerate before the last decade of the 20th century, with the ongoing Human Genome Project contributing significantly to the progress. Now that the human genome is mapped and a reference sequence is available, several genes can be studied as an entity, rather than one gene at a time. Medical genetics has thus become an important part of the broader field of genomic medicine.

In contrast to genetics, which is the study of specific genes and their effects, genomics is the study of the functions and interactions of all the genes in the genome, including interactions with environmental factors [3]. Genomic medicine has therefore a broader goal than medical genetics. It aims at explaining and treating not only those relatively rare conditions that follow one of the Mendelian patterns seen in single-gene disorders, but also common conditions such as schizophrenia, heart disease, diabetes, bipolar disorder, and Parkinson's disease. These complex disorders, which are also called "multifactorial" because of the many factors that contribute to them, affect almost two of every three individuals during their lifetime [4].

The main task of medical genetics is essentially to explain the causal relationship between genotype and phenotype on the one hand, and to give counseling and manage the phenotype on the other. Its starting point is therefore the human genome. Since the human nuclear genome and the mitochondrial genome are different in several important ways, some of which will be discussed later, these two will be treated separately in this introductory part.

1.2 The nuclear genome

The six billion DNA base pairs that make up the nuclear genome in somatic cells are divided between 23 chromosome pairs and twined around octamers of histone proteins, forming a dense network of DNA and protein called chromatin. It is only during cell division that chromatin is condensed to such a degree that it appears as visible chromosomes under the microscope, otherwise it is variably distributed throughout the cell nucleus. Chromosomes differ both in size and gene content, and even though genes are located throughout any given chromosome, they tend to cluster in specific regions. A consequence of this is that the clinical severity of a chromosomal abnormality depends not only on the type of the defect, but also on where in the genome it is located.

The estimated number of genes contained in the genome was drastically reduced with the completion of the initial draft sequence of the human genome in 2001, and the number has continued to fall ever since. In fact, current estimates indicate that the number may be as low as 19,000 protein-coding genes [5]. This means that protein-coding regions make up only one percent, or approximately 30 MB, of the entire human genome, divided between ~180,000 exons [6]. While most of the remaining regions may have other biochemical and regulatory functions [7] and can therefore not be considered unimportant or superfluous, the vast majority, about 85 percent, of the mutations with large effects on disease-related phenotypes are located in the exons [8]. Most of these mutations are so-called rare variants that – in contrast to common variants or polymorphisms – have a minor allele frequency (MAF) of less than one percent in the population [9].

1.2.1 Classification of variants

If we compare the DNA sequence of any two individuals, we find that they are 99.9 percent identical. Still it is the small variable fraction that contributes to the genetically determined differences among humans. The mutations responsible for this fraction are mostly polymorphisms, while only a small part of them are rare variants. Although the "common disease-common variant" hypothesis has been popular for a long time, the correlation between the frequency of a variant and the effect it has on human health is not such that common diseases are *only* caused by common variants [10]. Rare and generally deleterious

variants may actually stand for much of the genetic component of the common multifactorial diseases [11].

Variants can be classified into three categories based on their size: numerical chromosomal variants, which alter the number of chromosomes in the cell; structural chromosomal variants, which affect the structure of individual chromosomes; and genetic variants, which alter individual genes [12]. The most common type of numerical chromosomal aberrations is aneuploidy; a deviation from the normal number of 46 chromosomes. This abnormality, which is usually caused by meiotic nondisjunction[1], is rarely compatible with life [13]. One of the exceptions is trisomy 21, which is seen in ~95 percent of all cases of Down syndrome [14].

Structural chromosomal aberrations – whether deletions, duplications, translocations, insertions, or inversions – are formed by chromosomal breakage or unequal crossing-over [15]. A range of these abnormalities can result in disease, but then they are usually unbalanced rearrangements. This means that the complete chromosome set is no longer present, so that genetic information is either added or missing. Deletions, duplications, and insertions are all unbalanced rearrangements, and they can lead to several severe disorders, like Prader-Willi and Angelman syndromes [16]. In contrast, balanced rearrangements, where the full complement of DNA material is retained, are generally not associated with disease.

Genetic variants are usually classified as deleterious or normal based on their phenotypic effect, or, if this is not known, as variants of unknown significance (VUS). Variants that may result in premature truncation of the gene product – such as deletions, insertions, non-sense and frame shift variants, or variants that alter the splice site – are generally regarded as deleterious [17]. But for missense mutations, where one amino acid is substituted for another, it is often difficult, or even impossible, to predict the effect they will have on the phenotype. The reason for this is partly that the physicochemical differences between any two amino acids may or may not be so great as to have an impact on protein function. However, there are certain clues that do indicate whether a given missense variant disrupts the function or not. One of these is the variant's position within the protein. If, for example, the variant is located within a functional domain or if it affects an amino acid that has already been shown to contain a pathogenic mutation, then it is likely to be pathogenic [18]. Another important clue is the conservation between species, as mutations occurring in positions that are conserved

[1] Nondisjunction is the failure of sister chromatids to separate properly during cell division.

throughout evolution are likely to be deleterious [19]. “Sorting intolerant from tolerant” (SIFT) and “polymorphism phenotyping” (PolyPhen) are two bioinformatics tools that use evolutionary conservation as a basis for predicting whether non-synonymous changes are likely to disrupt protein function or not [18].

1.2.2 Mendelian inheritance

In contrast to multifactorial diseases, the underlying genetic causes of many Mendelian disorders are well understood. This is because the phenotypic effects of the variants underlying Mendelian diseases are less affected by environmental factors than complex disorders. In addition, these effects are often distinct and very severe, allowing for their early and accurate identification [20]. But the fact that the phenotypic consequences are largely unaffected by the environment also means that they follow inheritance patterns that accord with the rules of probability. Patterns of inheritance basically fall into one of three groups: autosomal, X-linked[2], and mitochondrial. Here we will briefly mention autosomal and X-linked inheritance, while mitochondrial inheritance, which differs from Mendelian inheritance in many important respects, will be discussed later.

The inheritance patterns of single-gene (or Mendelian) disorders depend mainly on two factors: whether the phenotype is recessive or dominant, and whether the causative gene variant is located on an autosome or a sex chromosome. A phenotype is recessive when it is manifest if the underlying genotype is homozygous or compound heterozygous – that is, if both alleles are mutated, either by harboring the same mutation or different mutations. If, on the other hand, one mutated allele causes the phenotype to be expressed, the phenotype is considered dominant.

The other important factor determining the pattern of inheritance is the location of the gene variant causing the phenotype. The clinical phenotype of an autosomal mutation is the same in both males and females. But for X-linked mutations, that is not always the case. Since males normally have only one X-chromosome, they can neither be homozygous nor heterozygous for a mutated X-linked allele, only hemizygous. In consequence, males are generally much more vulnerable to X-linked recessive disorders then females, who are usually affected only if they have two mutated alleles. When it comes to X-linked dominant

[2] Because Y-linked disorders are very few in number (less than 10 Y-linked diseases are described in the OMIM database), Y-linked inheritance is not included in this categorization.

disorders, the situation is even more so as some of these disorders are lethal in males only. Approximately ten different X-linked dominant male-lethal disorders have been described in the literature so far [21].

Online Mendelian Inheritance in Man (OMIM; www.omim.org) is a comprehensive online database of all known Mendelian disorders and known causative genetic variants. To date, more than 4000 phenotypes with known genetic basis are described in OMIM, and new descriptions are continuously added [22].

1.3 Dissecting the molecular basis of single-gene disorders

Diagnosis of a genetic disease is usually established based on a combination of clinical manifestations, family history, and biochemical and genetic tests, such as karyotyping or array-based comparative genomic hybridization (aCGH), or the selection of candidate genes for testing [23]. Over the last few years, however, whole exome sequencing (WES) has proved to be a powerful and cost-effective tool for dissecting the genetic basis of disease, greatly improving both diagnosis and genetic counseling [24]. When whole exome sequencing is performed to identify the genetic cause of a disease, the whole process of identifying the genetic cause often proceeds in three steps: whole exome sequencing, data filtering of the called variants, and verification by Sanger sequencing. What follows is a brief description of the principles behind each of these three steps.

1.3.1 Whole exome sequencing

Since most rare and deleterious variants are located in the protein-coding regions, which constitute only 1 percent of the entire human genome, whole exome sequencing (WES) is a more rapid and cost-effective method for routinely identification of rare and novel variants than most genome-wide approaches [6]. In general, different WES approaches follow the same four-step process regardless of the capture method and sequencing platform used. These major steps are: library preparation, hybridization to capture arrays, massively parallel sequencing of eluted fragments, and bioinformatics [25, 26]. As outlined in Figure 1, genomic DNA is first randomly shared to construct a library, the fragments of which are flanked with adapters to allow sequencing. Next, the library is enriched for exons by hybridization to biotinylated DNA or RNA baits in the presence of blocking oligonucleotides complementary

to the adapters. The hybridized fragments are then captured (for example by streptavidin coated magnetic beads), washed, and amplified. The library is finally sequenced by massively parallel sequencing, followed by mapping and variant calling [26].

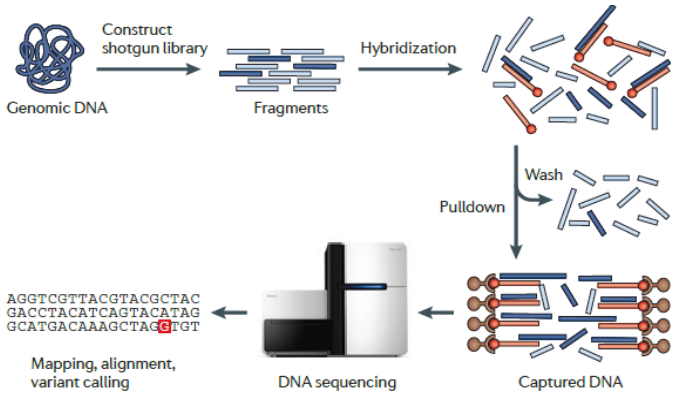


Figure 1 Workflow of exome sequencing. The figure shows the major steps of WES as explained in the text. A more detailed description of the computational part of the WES process is found in 2.2.2. Dark blue fragments: exons; orange fragments: biotinylated DNA or RNA baits. Adapters and the blocking nucleotides are not shown (from Bamshed *et al.* [26]).

1.3.2 Data filtering

On average, whole exome sequencing identifies at least 20,000 single nucleotide variants (SNVs) per individual [26]. A significant challenge of using exome sequencing to find disease-causing alleles is therefore to pick out the pathogenic variants from the background of benign variants and sequencing errors. Typical filtering steps that are applied to filter out non-causative variants are summarized in Figure 2, which is modified from Stitziel *et al.* [27]. What follows is a brief discussion of the different steps along with key assumptions made at each of them.

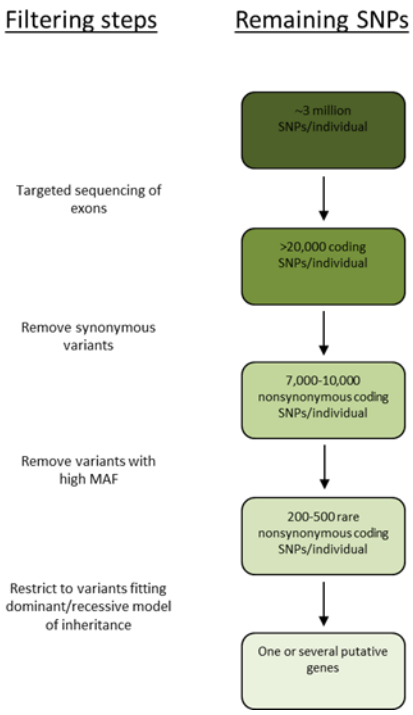


Figure 2 Typical filtering steps applied to exome sequencing projects to reduce the number of candidates for disease-causing variants.

About 3 million SNVs per genome are discovered using whole-genome sequencing [28]. Assuming that causal variants are coding and ignoring regulatory elements and other non-coding variants outside of exon definitions, targeted sequencing of exons excludes more than 99% of the detected variants. By focusing only on non-synonymous (amino acid-altering) changes, up to 75% of the roughly 20,000 exon variants can further be removed [29]. The assumption here is that causative variants will change the amino acid sequence, ignoring rare cases of functional synonymous changes [30]. But not all protein-altering variants are considered. Often, only those are retained that are predicted to be deleterious to protein function; either because a truncated protein product results, as is the case with nonsense, splice site, and frame shift mutations; or because the mutation interrupts protein function in other ways, as is the case with deleterious missense mutations. SIFT and PolyPhen identify missense mutations that are likely to be deleterious based on the level of conservation of the mutated nucleotide and amino acid. Under the assumption that pathogenic changes are rare in the healthy population, non-synonymous variants with a minor allele frequency (MAF) of more than 1% are usually filtered out. This step currently reduces the list of putative causal variants to about 200 to 500 [27].

To further restrict the search, a recessive or dominant model of disease is imposed when one of them is more likely than the other as indicated by the pedigree or the disease itself. If the pedigree suggests a recessive mode of inheritance, a putative causal variant must either be present in a homozygous or compound heterozygote state. On the other hand, a dominant model requires affected individuals to be heterozygous for the same variant [31].

This short overview is not meant to be exhaustive or to outline the exact sequential application of the different filtering steps; it is only meant to illustrate the effectiveness with which simple filtering can drastically reduce the number of potential disease-causing variants. The goal is to reduce the list to a single putative variant or gene that can further be analyzed and verified by functional studies.

1.3.3 Sanger sequencing

Because exome sequencing does have some technical limitations and errors may therefore occur, like poor coverage or inaccurate calling, Sanger sequencing is often required to verify the findings one is left with after filtering. Developed by Sanger and his colleagues in 1977, Sanger sequencing was the most widely used sequencing method before massively parallel sequencing became commercially available, and it is still widely used today, although in a modified manner. In their original publication, Sanger *et al.* [32] used four separate reactions for sequencing, each reaction containing a DNA template, a primer, DNA polymerase, reaction buffer, radioactively labeled deoxynucleotides (dNTPs), and one of the four chain-terminating dideoxynucleotides (ddNTPs). When they initiated DNA synthesis, extension products of different lengths terminated with a ddNTP at the 3' end would result. They would then employ gel electrophoresis to separate the extension products by size at a resolution of one base pair, so that the sequence could be read fairly easy, from bottom to top, by comparing the relative positions of the different bands among the four lanes (Figure 3).

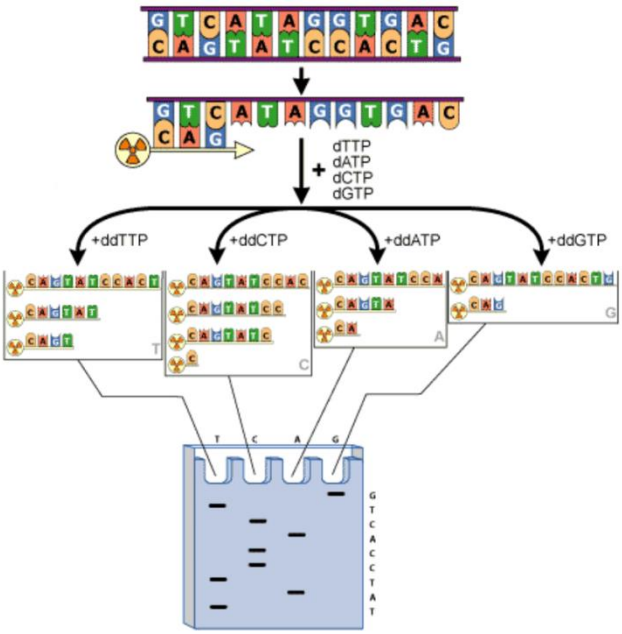


Figure 3 A simplified overview of the traditional Sanger method.

Cycle sequencing is a modified version of the traditional Sanger method that is widely used today for automated long read DNA sequencing. The principle is essentially the same but the method differs in at least two important respects. Firstly, instead of radioactively labeled dNTPs, cycle sequencing uses ddNTPs tagged with different fluorescent dyes, each of which emits a unique wavelength when excited by light. Thus, all the four labeled ddNTPs (dye terminators) are combined in one reaction that is subjected to cycles of annealing, extension, and denaturation in a thermal cycler [33].

Secondly, the DNA sequencing products are not separated by gel electrophoresis but by capillary electrophoresis in which a denaturing liquid polymer is used. During capillary electrophoresis, the negatively charged extension fragments are forced into the capillaries through a high voltage charge that is applied to the buffered sequencing reaction. Shortly before reaching the positive electrode, the fluorescently labeled extension products, which are separated by size based on their total charge, move through a laser beam that causes the dyes on the fragments to fluoresce. The emitted fluorescence signals are then detected by an optical detection device and converted into digital raw data. Finally, a data analysis software processes the raw data and displays it as an electropherogram – a sequence of peaks in four colors, each representing one of the four bases. The main advantage of using capillary

electrophoresis for separation is its flexibility and the ease with which it can be incorporated into automated instruments [34].

1.4 Mitochondrial genetics

Although the mitochondrial genome is extremely small compared to the nuclear genome, mutations in the mitochondrial genome is an important cause of inherited disease [35]. But despite the small size, mitochondrial genetics is complicated by several features that distinguish it from nuclear genetics and that present unique experimental and clinical challenges. Some of these features and challenges as they relate to inheritance patterns, sequencing, and diagnosis will be discussed in this second part of the introduction. But what follows first are some basic facts about mitochondrial genetics.

1.4.1 The mitochondrial genome

Mitochondria are found in all nucleated cells and function primarily as generators of cellular ATP by oxidative phosphorylation (OXPHOS). Most cell types contain hundreds of mitochondria, each harboring several copies of mitochondrial DNA (mtDNA) [36]. The mitochondrial genome consists of a circular double-stranded DNA molecule of 16.6 kb, which encodes 13 mRNAs that are all translated within the organelle, plus 22 tRNAs and 2 rRNAs. The 13 polypeptides encoded are protein subunits of the electron transport chain complexes I, III, IV, and V, while the tRNAs and rRNAs are necessary components of the mitochondrial translation machinery (Figure 4). But the vast majority of respiratory chain subunits, translation machinery components, and almost all of the factors involved in mtDNA replication, transcription, translation, and repair are encoded by the nuclear genome [37]. From a genetic point of view, the respiratory chain is thus unique, as it is formed by two separate but complementary genetic systems, the mitochondrial and the nuclear genomes (Figure 5).

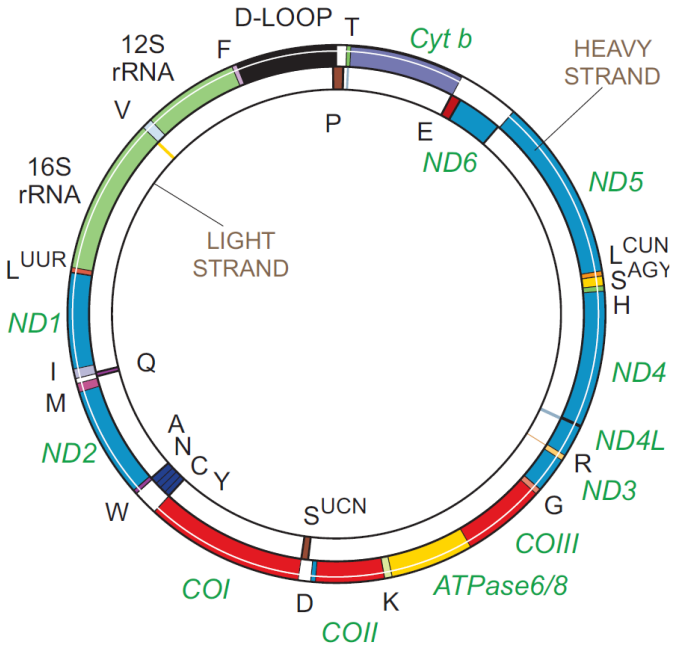


Figure 4 Structure of human mitochondrial DNA. The 13 protein-coding genes found in the mitochondrial genome are *ND1-ND6* (blue), which encode the subunits of respiratory chain complex I; *COI-COIII* (red) which encode the subunits of complex IV; *ATPase6* and *8* (yellow) which encode subunits 6 and 8 of complex V; and *Cyt b* (purple) which encodes cytochrome b of complex III. The 22 tRNA genes are indicated by the single letter code of the corresponding amino acids, while the two rRNA genes, 12S and 16S, are marked in green. The D-loop, sometimes referred to as the control region, is a stretch of non-coding mtDNA. It is the most polymorphic region of the human mitochondrial genome (from Zeviani *et al.* [38]).

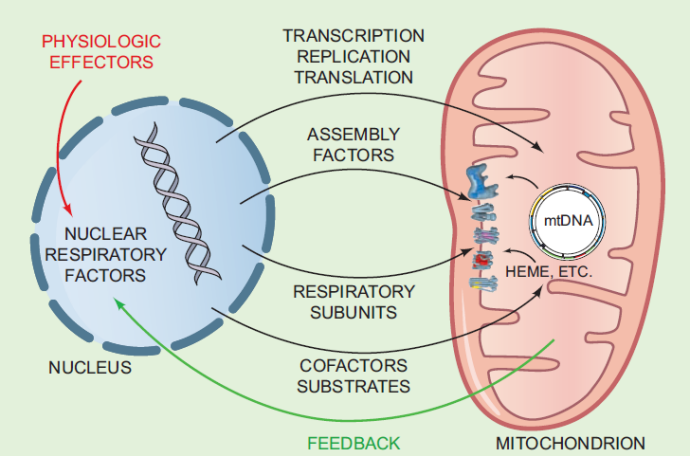


Figure 5 Nucleus-mitochondrion interactions. The nucleus encodes for most of the protein subunits of the respiratory chain complexes. In addition, nucleus-encoded factors act in the assembly and turnover of the respiratory chain, in the synthesis and transport of cofactors and substrates, and in the maintenance and expression of mtDNA (from Zeviani *et al.* [38]).

Both the small size and the fact that it is covalently closed circular are very significant structural properties that mtDNA shares with bacterial plasmids. A third property that it has in common with bacterial plasmids is that the mitochondrial genome is not twined around proteins and packaged as chromatin. In this respect, it differs from both the nuclear genome and bacterial chromosomal DNA (as opposed to bacterial plasmids). Although the latter is not intertwined with proteins in the same way as nuclear chromosomes, it is nevertheless constantly interacting with and even wound around multiple proteins that help to form and maintain the supercoiled state [39]. Since mitochondrial DNA and bacterial plasmids share several important aspects in their structural organizations, a simple bacterial plasmid isolation protocol can be utilized for enrichment of mtDNA (see 2.5).

1.4.2 Mitochondrial inheritance

Diseases that result from mitochondrial mutations show a pattern of inheritance that differs radically from Mendelian inheritance. This distinctive inheritance pattern is mainly because of three unique features of mitochondria: vegetative segregation, homoplasmy and heteroplasmy, and maternal inheritance.

Alleles of nuclear genes do not segregate during mitosis, which means that, for example, heterozygous cells produce heterozygous daughters. The reason for this is that nuclear genomes are said to be “stringent” genomes in which a) each chromosome replicates exactly once during interphase, and b) both daughter cells receive one copy of all the 46 chromosomes [40]. In contrast, alleles of mitochondrial genes undergo vegetative segregation, meaning that they segregate during mitosis as well as meiosis [41]. The basic reason underlying vegetative segregation is that both replication and partitioning of mitochondrial genomes are so-called relaxed processes. At cell division, the many copies of mtDNA in each of the mitochondria in the cell replicate and sort randomly between the daughter mitochondria. In addition, the daughter mitochondria themselves are distributed randomly between the two daughter cells [41].

The second unique feature of mitochondria is homoplasmy and heteroplasmy. Homoplasmy means that all copies of the mitochondrial DNA in a cell, a tissue, or an organism are identical; heteroplasmy means that there exists a mixture of two or more mitochondrial genotypes [35]. Three factors contribute to mtDNA diversity. Firstly, ~25% of healthy individuals inherit a mixture of normal and variant mtDNA; secondly, mtDNA has an

estimated mutation rate of ~5-15 times that of the nuclear genome, partly because of proximity to the electron transport chain that is the major intracellular source of oxidative free radicals, and partly because of limited mtDNA protection and repair mechanisms; and thirdly, the level of variant mtDNA can change during life through relaxed replication and partitioning [42]. The presence of heteroplasmy is of particular interest when we consider mtDNA variants that lead to disease, as the disease phenotype is expressed only when the percentage of mutant molecules exceeds a critical threshold level (Figure 6). Depending on the type of mutation and the tissue type, this level is usually in the range of 70-90% [43].

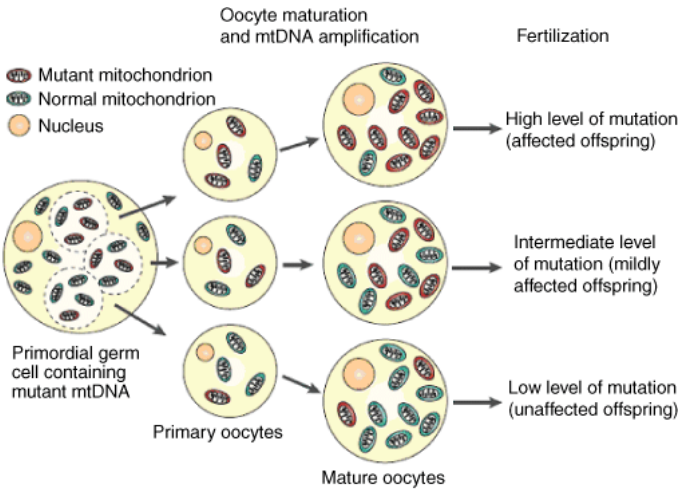


Figure 6 The effect of heteroplasmy on disease (from Longo *et al.* [44]).

Maternal inheritance is the last feature underlying the distinct inheritance pattern of diseases that result from mutations in the mitochondrial genome. Although paternal inheritance of mtDNA in humans has been reported before [45], this is the exception rather than the rule [46]. The widely accepted view is that in the cells of most animals mtDNA is inherited solely from the oocyte from which the animal develops [47]. In mammals, paternal mitochondria derived from sperm generally enter the cytoplasm of the oocyte where they shortly co-exist with an excess of maternal mitochondria [48]. However, the paternal mitochondria and their DNA are subsequently eliminated and not transmitted to the offspring. Two hypotheses have been proposed to explain the underlying mechanism of maternal inheritance of mtDNA in mammals. The "simple dilution model" suggests that the paternal mtDNA, because it is present at a much lower copy number, is simply diluted away by the excess of maternal

mtDNA and is therefore not detectable in the offspring [49]. According to the "active degradation model", paternal mitochondria are selectively degraded by ubiquitination to actively prevent transmission to the next generation [50].

1.5 Sequencing the mitochondrial genome

Even though sequencing of the whole mitochondrial genome is routinely practiced, it nevertheless presents with several challenges that must be overcome in order to produce high quality sequences. One of these challenges is the presence of nuclear DNA sequences of mitochondrial origin (NUMTs), which are the result of mtDNA transfers to the nucleus (see 4.3.3). If using PCR to amplify mtDNA, primers must be carefully designed to avoid co-amplification of NUMTs as they pose a real problem for accurate sequence interpretation [51]. A second challenge is the universal presence of low levels of heteroplasmy in human mitochondrial DNA [42]. In order to detect low-level heteroplasmic mtDNA variants, it is necessary to use sequencing methods that achieve very great depth of coverage per base position. A third challenge is the use of PCR amplification itself. Despite the high number of mitochondrial genomes in each cell, mtDNA only constitutes a small portion of total cellular DNA. For this reason, it is necessary to enrich samples for mtDNA before sequencing. PCR amplification is the most efficient and most commonly used method for enrichment, but may produce artifacts since many cycles are needed for sufficient enrichment, which again can lead to biased results [52].

To overcome the third challenge, Quispe-Tintaya *et al.* [52] designed a simple, fast, and cost-effective method for preparation of samples that are enriched for mtDNA. This method, which we applied in our study, consists of two steps: extraction of mitochondrial DNA using a standard kit for isolation of bacterial plasmids followed by additional purification using solid-phase reversible immobilization (SPRI) paramagnetic beads. The principles underlying these two techniques are described in 2.5.

1.6 Mitochondrial disorders

Mitochondrial disorders constitute a clinically heterogeneous group of diseases that arise as a result of dysfunction of the mitochondrial respiratory chain. They can be caused by mutations in either the nuclear or mitochondrial DNA. Due to the complexity of mitochondrial genetics and the fact that mitochondria are vital components of all nucleated cells, the clinical

manifestations of mtDNA disorders are extremely diverse (Table 1). They range from lesions that affect only one organ or tissue, such as the optic nerve in Leber's hereditary optic neuropathy (LHON), to lesions that are more widespread including encephalomyopathies, myopathies, and cardiopathies, to complex multisystem syndromes that may present at any age [38]. While many affected individuals display a number of clinical features that fall into a discrete syndrome, such as chronic progressive external ophthalmoplegia (CPEO) and myoclonic epilepsy with ragged red fibers (MERRF), some individuals present with a considerable variety of clinical symptoms that do not fit into one specific category. One illustration of this is the overlapping spectrum of disease phenotypes that result from mutations in the nuclear *POLG* gene, which is the most common nuclear gene causing mitochondrial disorders [53, 54].

Phenotypic Expression of Mitochondrial Diseases

Neurologic Manifestations	Systemic Manifestations
Neuromuscular	**Heart**
Ophthalmoplegia	Cardiomyopathy
Myopathy	Cardiac conduction defects
Exercise intolerance	**Endocrine**
Peripheral sensory-motor neuropathy	Diabetes
CNS	Exocrine pancreas dysfunction
Myelopathy	Hypoparathyroidism
Headache	Multiple endocrinopathy
Stroke	Short stature
Seizures	**Blood**
Dementia	Pancytopenia
Movement Disorders	Sideroblastic anemia
Ataxia	**Mesenchyme**
Dystonia	Hepatopathy
Parkinsonism	Nephropathy
Myoclonus	Intestinal pseudo-obstruction
Eye	**Metabolism**
Blindness	Metabolic acidosis
Optic neuropathy	Nausea and vomiting
Pigmentary retinopathy	
Cataract	
Ear	
Sensorineural deafness	

Table 1 Phenotypic expression of mitochondrial diseases. One way to classify the extremely diverse phenotypic expressions of mitochondrial disorders is to split them into the two categories of neurological and systemic manifestations (from Zeviani *et al.* [38]).

1.6.1 Mitochondrial disorders of nuclear origin

A major feature that distinguishes mitochondrial disorders of nuclear origin from those of mtDNA origin is the mode of inheritance, as the former group follows a Mendelian pattern of inheritance while the latter does not. Even in cases where mitochondrial syndromes characterized by the presence of mtDNA abnormalities are inherited as Mendelian traits, the underlying genetic cause is usually the existence of mutations in nuclear genes affecting the integrity of the mitochondrial genome [55]. Although more than 90% of the mitochondrial proteins related to the respiratory chain are encoded by the nuclear genome, the number of mitochondrial disorders shown to be caused by defects in nuclear genes is still small [55]. Mutations have been identified in only about 150 of the estimated 1,500 nuclear genes associated with mitochondria [54].

A classification has been proposed that organizes these disorders into four categories based on the target of the respective gene defect [56]. The first category consists of the nuclear genes responsible for mtDNA stability. These are all related to replication of mtDNA, by encoding enzymes or factors that are either part of the mitochondrial replication machinery or supply the building blocks of mtDNA. The second category includes all genes encoding structural components of the respiratory chain complexes. The third group consists of nuclear genes involved in expression, assembly, and turnover of the OXPHOS complexes, usually complexes III and IV. In infancy, the most frequent manifestation of complex IV deficiency is Leigh syndrome [57]. The last category consists of the disorders caused by defects in nuclear genes that are indirectly related to OXPHOS. One such disorder is X-linked sideroblastic anemia[3] and ataxia caused by mutations in the iron mitochondrial exporter ABC7, which controls the generation of cytosolic iron-sulfur proteins [38, 55].

1.6.2 Mitochondrial disorders of mtDNA origin

In contrast to mitochondrial disorders of nuclear origin, those of mtDNA origin are all directly related to the electron transport chain. The reason for this is that each of the 37 genes contained in the mitochondrial genome either encodes a protein subunit of the respiratory chain or a component of the mitochondrial translational machinery needed to make the

[3] Sideroblastic anemia is a form of anemia in which the bone marrow produces ringed sideroblasts rather than healthy erythrocytes. Sideroblasts are atypical, abnormal nucleated erythroblasts (precursors to mature erythrocytes) with granules of iron accumulated in perinuclear mitochondria. In sideroblastic anemia, the body has iron available but cannot incorporate it into hemoglobin.

subunits. These disorders, with all their complexity with regard to inheritance and clinical manifestations, are often classified into two major categories: (i) disorders due to large-scale rearrangements of mtDNA, and (ii) disorders due to mtDNA point mutations [58].

Most large-scale rearrangements of mtDNA are sporadic. They can either be partial deletions of mtDNA or, less frequently, partial duplications. Both types are usually heteroplasmic as they exist together with normal mtDNA. Large-scale rearrangements are mainly associated with Kearn-Sayre syndrome (KSS), progressive external ophthalmoplegia (PEO), and Pearson's syndrome [38, 58] (see 4.6). On the other hand, mtDNA point mutations are typically maternally inherited. Characteristic of these mutations is that they affect highly conserved nucleotides/amino acids or that they cause loss of function of the protein product. When these mutations are heteroplasmic, which is usually but not always the case, the severity of the disease tends to correlate with the degree of heteroplasmy. This is for example seen in neurogenic weakness, ataxia, and retinitis pigmentosa (NARP), a syndrome mainly associated with the mutation m.8993T>G in *MT-ATP6*. In contrast to many heteroplasmic point mutations, the clinical expression of disorders caused by homoplasmic mutations is often stereotypical and restricted to one tissue, as is the case in LHON. Over 95% of LHON patients worldwide carry one of three mtDNA point mutations: m.11778G>A (*ND4*), m.14484T>C (*ND6*), or m.3460G>A (*ND1*), all of which affect genes encoding complex I subunits of the respiratory chain [59].

1.6.3 Leigh syndrome

Leigh syndrome (LS) is a severe neurodegenerative disease that typically arises in infancy, affecting 1 in 40,000 newborns [60, 61]. This condition is characterized by progressive loss of mental and motor abilities; characteristic MRI findings in the basal ganglia, thalamus, cerebellum, or brain stem; high lactate levels in blood or cerebrospinal fluid; and signs and symptoms of lesions in brain stem or basal ganglia, such as ataxia, dystonia, optic atrophy, hypotonia, respiratory failure, chorea, and swallowing or feeding difficulties [57]. LS is genetically heterogeneous, and so the inheritance pattern of the disorder varies accordingly. Thus, it is maternally inherited when caused by mutations in mtDNA genes, such as *MT-ATP6* encoding subunit 6 of ATP synthase (complex V). LS can also be inherited in a Mendelian manner; either X-linked due to a deficient pyruvate dehydrogenase complex

caused by mutations in *PDHA1*, or autosomal recessive due to mutations in nuclear genes encoding respiratory chain subunits or assembly proteins [57].

Cytochrome c oxidase (COX) deficiency is one of the most common biochemical abnormalities found in LS patients [62]. COX is the terminal component (complex IV) of the respiratory chain and catalyzes the transfer of electrons from reduced cytochrome c to oxygen (Figure 7). The complex is composed of 13 subunits, 10 of which are encoded by nuclear genes and three by mtDNA genes. Assembly of the subunits is controlled by at least six nuclear genes [63]. Loss-of-function mutations of one of these "COX assembly genes", *SURF1*, are the best documented autosomal cause of LS worldwide [62, 64]. To date, more than 70 mutations in *SURF1* have been reported in patients with LS [57].

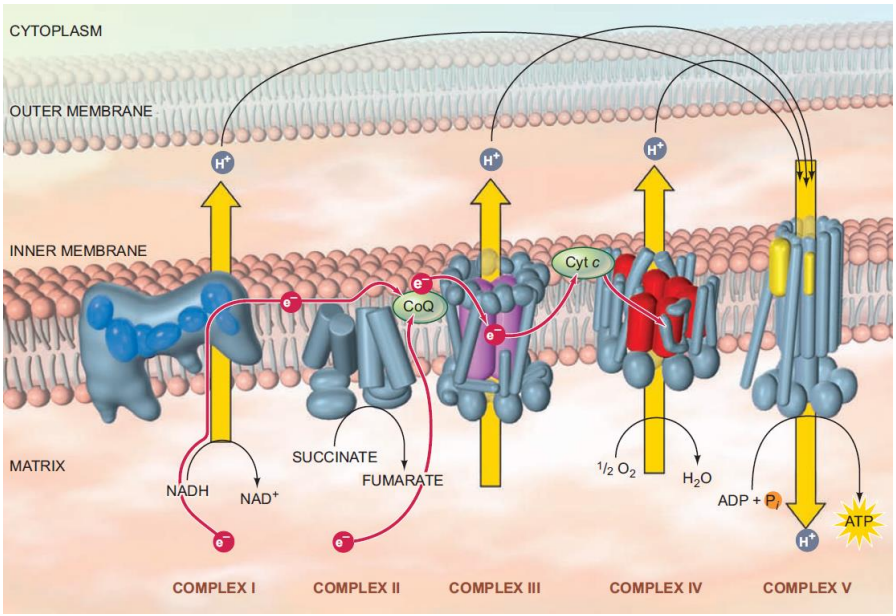


Figure 7 Oxidative phosphorylation and the respiratory chain. Oxidative phosphorylation is the last stage of cellular respiration. The energy released by oxidation of organic molecules is conserved in the reduced electron carriers NADH and $FADH_2$ (not shown). Electrons from these carriers are then transferred by complexes I and II to the electron carrier ubiquinone (or coenzyme Q, CoQ) before they are carried on to cytochrome c by complex III. Finally, cytochrome c oxidase (COX, complex IV) catalyzes the transfer of electrons from the reduced cytochrome c to oxygen, which is the last electron acceptor in the respiratory chain. In the course of electron transfer, the large amount of energy released is used to translocate protons across the mitochondrial inner membrane from inside out, thus generating an electrochemical gradient across the membrane. The energy stored in this gradient drives the synthesis of ATP by ATP synthase (complex V) (from Zeviani *et al.* [38]).

1.6.4 Kearn-Sayre syndrome

Kearns-Sayre syndrome (KSS), first described in 1958, is a rare multisystem disorder which belongs to the group of mitochondrial cytopathies[4] [65]. The diagnosis of KSS is made on the basis of three obligatory characteristics: onset of symptoms before 20 years of age, progressive external ophthalmoplegia (PEO), and pigmentary retinopathy. In addition, KSS patients have at least one of the following symptoms: cardiac conduction defects, cerebellar ataxia, or abnormally high levels of protein (>100 mg/dL) in the cerebrospinal fluid [66, 67]. The molecular basis underlying the disease is sporadic large-scale rearrangements of the mitochondrial DNA, usually deletions varying in location, size, and percentage. The most common deletion, found in approximately one third of KSS patients, is m.8470_13446del4977. Extending over 4,977 base pairs, this deletion affects twelve mitochondrial genes [66, 68].

1.7 Clinical description of patients

Three patients with clinical suspicion of mitochondrial disease were studied in the present work (Table 2). Patient 1 is a 9 year old boy born to healthy consanguineous Pakistani parents. The delivery was normal. He has two brothers of 13 and 15 years of age who are both healthy. At the age of two, he was short of stature and delayed in his motoric development, but growth hormone levels were normal. At four years of age, the patient presented with additional clinical manifestations, including ataxia, microcephaly, general hirsutism, general dystrophy, elevated protein and pyruvate levels in cerebrospinal fluid (CSF), increased lactate levels in blood and CSF, and strabismus. Examination of muscle tissue carried out at the Nijmegen Centre for Mitochondrial Disorders (NCMD) revealed a decreased complex II and IV activity but no mtDNA mutations were found. Nor did the muscle biopsy show any signs of ragged-red fibers or any other morphological changes of the mitochondria. He slowly developed symptoms of retinopathy and progressive external ophthalmoplegia (PEO), such as impaired eye movement, drooping eyelids (ptosis), and salt-and-pepper fundus. Other clinical findings were demyelinating neuropathy, cognitive failures, and MRI images showing signal changes in deep grey matter and the brain stem. The patient does not suffer from cardiopathy. The clinical manifestations are suggestive of Leigh syndrome (LS).

[4] Mitochondrial cytopathies represent a heterogeneous group of multisystem disorders which preferentially affect the muscle and nervous systems.

Patient 2 is a Norwegian boy born in 2009 to healthy non-consanguineous Norwegian parents. His 7 years old brother and 11 years old sister are both healthy. No family history of ataxia or neurological disease is known. At 14 months of age, he was 9 months behind in his motoric development. Later he also developed progressive ataxia with intention tremor. MR imaging revealed no cerebellar atrophy but he had elevated tau protein levels (302 ng/L; reference: 110-177.5 ng/L) in CSF. Lactate and pyruvate levels were both normal. Electromyography detected changes in the electrical muscle activity (myopathy), and at 4.5 years of age, he showed signs of partial PEO. No firm diagnosis has yet been established.

Patient 3 is a Turkish boy born in 1999 to consanguineous healthy parents. He has a brother with epilepsy and a healthy sister. At the age of 9, the main symptoms were poor growth, exercise intolerance, and general fatigue. Chromosome analysis, blood test, and head MRI were all normal. But later, signs of retinopathy, ataxia, and PEO appeared. In addition, serum levels of lactate were elevated (6.3 mmol/L; reference: 0.5-2.2 mmol/L). He also had abnormally high levels of protein (587 mg/dL; reference: 15-60 mg/dL) in CSF. At 15 years of age, retinitis pigmentosa (RP) was confirmed by electroretinography (ERG). A second head MRI revealed cerebral leukodystrophy and lesions in the brain stem, cerebellum, and basal ganglia. The clinical picture is indicative of a KSS diagnosis.

Table 2 Clinical presentation and genetic mutations in three patients with suspected mitochondrial disease. The genetic defects in patients 1 and 3 were identified in the current study.

Patient	Born	Sex	Ethnicity	CS	Neuromuscular	Metabolic	Brain MRI lesions	Others	Mutation	Clinical diagnosis
1	2006	M	Pakistani	Yes	Ataxia Myopathy PEO RP	Elevated lactate and pyruvate levels	Basal ganglia Brain stem Cerebellum	-	*SURF1* c.106+1G>C	LS
2	2009	M	Norwegian	No	Aataxia Myopathy	-	Cerebral cortex?	Increased tau in CSF	-	?
3	1999	M	Turkish	Yes	Aataxia PN RP Leukodystrophy	Elevated lactate	Basal ganglia Brain stem White matter Cerebellum	-	m.6211_8681del	KSS

M, male; CS, consanguineous family; PEO, progressive external ophthalmoplegia; RP, retinitis pigmentosa; PN, progressive neuropathy; KSS, Kearns-Sayre syndrome; LS, Leigh syndrome.

1.8 Aim of thesis

The main aim of this thesis was to identify the causative genetic variants in three patients suspected to have a mitochondrial disease. Since the majority of respiratory chain subunits, translation machinery components, and almost all of the factors involved in mtDNA replication, transcription, translation, and repair are encoded by nuclear genes, the nuclear genomes of the three patients were first analyzed to identify the genetic variants most likely to be responsible for the clinical phenotypes. Thereafter, the mitochondrial genomes were analyzed for the same purpose. A sub-aim of this thesis was therefore to evaluate the suitability of using massively parallel sequencing for the detection of disease-causing mutations in the mitochondrial genome.

2 Methods

2.1 Blood sampling and DNA/RNA extraction

Blood from the three patients and their healthy parents was sampled in K/EDTA blood collection vials and DNA was extracted using the QIAsymphony DSP DNA Kit on the QIAsymphony SP instrument (Qiagen GmbH) according to manufacturer's recommendations. For RNA extraction, blood from all three families was collected in PAXgene Blood RNA tubes and RNA was extracted with the PAXgene Blood RNA Kit (PreAnalytiX GmbH) as described in the manufacturer's protocol. Quantification and purity measurement of the DNA and RNA samples were carried out with the NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific).

2.2 Whole exome sequencing

Whole exome sequencing (WES) of DNA from all subjects was performed to identify potentially disease-causing variants. The samples were prepared by Asbjørn Holmgren at the Department of Medical Genetics, Oslo University Hospital, and delivered to the Norwegian Sequencing Centre (NSC) for sequencing on the Illumina HiSeq 2000 platform.

2.2.1 Library preparation and exome capture

Individual libraries were prepared for each DNA sample by randomly shearing the DNA with a Covaris ultrasonicator (Covaris) producing fragments of approximately 150bp. Using the SureSelect XT Library Prep Kit (Agilent Technologies), the fragments were end repaired, 3' end adenylated, ligated with paired-end adaptors, and amplified according to the manufacturer's instructions. Hybridization with biotynilated RNA baits and exome capture using streptavidin-coated magnetic beads were then performed using the same kit. Finally, the amplified exome captured library was quantified with a Qubit 2.0 Fluorometer (Life Technologies) and the fragment size distribution was measured in an Agilent 2100 Bioanalyzer instrument using an Agilent High Sensitivity DNA Chip (Agilent Technologies).

2.2.2 Sequencing, alignment, and variant calling

The captured library was sequenced on an Illumina HiSeq 2000 platform at the NSC. Reads that did not pass Illumina's standard filter were removed prior to alignment, while the remaining reads were aligned to the reference human genome (GRCh37/hg19) using Burrows-Wheeler Aligner tool (http://bwa.sourceforge.net/). PCR duplicates were removed. Approximately 99% of exome targets defined by Agilent were covered, with an average of 95x coverage per targeted base for each of the samples. Variant calling was performed using the Genome Analysis Toolkit (GATK) (https://www.broadinstitute.org/gatk/) and variants were annotated with SnpEff (http://snpeff.sourceforge.net/) and VEP (http://www.ensembl.org/info/docs/tools/vep/). VCFtools (http://vcftools.sourceforge.net/) was used to manipulate variant calling format (VCF) files and the Integrative Genomic Viewer (http://www.broadinstitute.org/igv/) was used to visualize the data.

2.2.3 Analysis and filtering of called variants

The variant calling format (VCF) files generated were analyzed and filtered with FILTUS (http://folk.uio.no/magnusv/FILTUS/). The following genes and variants were filtered out: genes that generate false positive signals in exome sequencing [69]; variants seen more than two times in 268 exome sequences available in-house; variants with a Phred quality score less than 30 (meaning that the accuracy of the base call is less than 99.9%[5]; see 3.3.4); variants with minor allele frequency (MAF) greater than 0.01 according to the 1000 Genomes project (http://1000genomes.org/), dbSNP138 (http://ncbi.nlm.nih.gov/SNP) and Exome Variant Server (http://evs.gs.washington.edu/EVS); and variants predicted to be benign/tolerated by SIFT (http://sift.jcvi.org/) and PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/). Only protein-coding variants were retained, with a special focus on small deletions/insertions and missense, nonsense, frameshift, and splice site mutations. Based on the family history of the patients, mutations with recessive homozygous mode of inheritance were chosen for further examination.

[5] See: http://www.phrap.com/phred/

2.3 Verification by Sanger sequencing

To verify the *SURF1* mutation detected in patient 1 by whole exome sequencing, we Sanger sequenced a 455bp DNA region containing the position of this point mutation. This we did for the patient and his parents using the DNA extracted with the QIAsymphony DSP DNA Kit.

2.3.1 Primer design, PCR amplification and purification

Primers for PCR amplification were design using the web-based primer design tool Primer3 (http://bioinfo.ut.ee/primer3/). A DNA sequence of approximately 600bp containing the mutation site was pasted into the software and following parameters were chosen: product size range 450-600bp; primer length 18-23bp; melting temperature (T_m) 57-70°C; GC content 30-70%; and max T_m difference between forward and reverse primers 5°C. The DNA sequence was also labeled for the software to avoid the common SNPs found in *SURF1* as defined by Ensebl (http://www.ensembl.org/). In addition, the designed primers were checked for sequence similarities in other DNA regions using BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat) and In-Silico PCR (http://genome.ucsc.edu/cgi-bin/hgPcr). Finally, the primers were tailed with universal M13 sequences and ordered by Eurofins Genomics (Ebersberg, Germany) (Table 3).

Name	Sequence (5'- 3')	Primer size (bp)	T_m (°C)	GC (%)
SURF1 Fwd	tgtaaaacgacggccagtATGCAGATGCTTCCTGCGTC	20	63.8	55
SURF1 Rev	caggaaacagctatgaccAGCTACCTCTCGCGTTGTGA	20	61.1	55

Table 3 PCR primer sequences for Sanger sequencing. The primer sequence is shown in capital letters, while the universal M13 tales are shown in small letters. The length of the amplified fragment is 455 base pairs and occupies the chromosome region chr9:133356046-133356500. The melting temperatures were calculated by In-Silico PCR assuming 50 mM salt and 50 nM annealing oligo concentration.

PCR amplification of the 455bp DNA segment was carried out on a Veriti 96-Well Thermal Cycler (Applied Biosystems) according to standard PCR protocols using AmpliTaq Gold DNA Polymerase with Gold Buffer (Applied Biosystems). PCR conditions and cycle parameters are shown in Table 4. To check the specificity of the designed primers, the amplified DNA fragments from the three subjects were run on a 1% agarose gel in 1X TAE buffer and visualized with GelRed Nucleic Acid Gel Stain (Biotium) on an ImageQuant Las 4000 instrument (GE Healthcare Life Sciences). The DNA was then purified on Biomek FXP Laboratory Automation Workstation (Beckman Coulter) using the Agencourt AMPure XP

system (Beckman Coulter). Paramagnetic beads (see 2.5; Footnote 6) were added in 1.8 bead:DNA ratio by volume in order to remove primer dimers, excess primers and dNTPs, salts and other contaminants. The DNA fragments were washed with 70% ethanol and eluted in water.

Component	Volume (µL)
H_2O	18.4
Primers (10 µM)	1.0
10 mM dNTP	0.5
25 mM $MgCl_2$	1.5
AmpliTaq Gold (5U/µL)	0.1
10X Gold Buffer	2.5
DNA (60 ng/µL)	1.0
Total	25.0

Stage	Temperature	Time
Stage 1	96°C	5 min
Stage 2	94°C	30 sec
	60°C	45 sec
	72°C	45 sec
Stage 3	72°C	10 min

Table 4 PCR conditions and cycle parameters used for PCR amplification. Stage 1, 1 cycle of initial denaturation; stage 2, 30 cycles of denaturation, annealing and elongation; stage 3, 1 cycle of final extension. The first stage is also necessary in order to activate the enzyme.

2.3.2 Cycle sequencing and purification

The cycle sequencing was carried out using BigDye Terminator v3.1 Cycle Sequencing Kit (Life Technologies) supplied with universal M13 primers. Veriti 96-Well Thermal Cycler (Applied Biosystems) was used for the cycle sequencing (Table 5). The samples were then purified on the Biomek FXP Laboratory Automation Workstation (Beckman Coulter) using Agencourt CleanSEQ (Beckman Coulter). CleanSEQ is a solid-phase reversible immobilization (SPRI) paramagnetic bead-based system similar to AMPure but is especially design to remove dye-terminators in addition to the other contaminants. This significantly improves the final sequencing data quality [70]. 85% ethanol was used for washing before the DNA was eluted in 0.05 mM EDTA.

Component	Volume (µL)
H_2O	4.75
5X Sequencing Buffer	2.0
Ready Reaction Mix	0.25
M13 Primer	1.0
Purified PCR product	2.0
Total	10.0

Stage	Temperature	Time
Stage 1	96°C	10 sec
	50°C	5 sec
	60°C	4 min
Stage 2	10°C	∞

Table 5 Conditions and cycle parameters used for cycle sequencing. 5X Sequencing Buffer, Ready Reaction Mix and M13 Primer supplied with the BigDye Terminator v3.1 Cycle Sequencing Kit. The BigDye enzyme is contained in the reaction mix. Stage 1: 25 cycles of denaturation, annealing and elongation.

2.3.3 Capillary electrophoresis and sequence analysis

Cycle sequencing produces PCR fragments of different sizes. To separate the fragments, the samples were subjected to capillary electrophoresis (see 1.3.3) on the 3730xl DNA Analyzer (Applied Biosystems). This instrument separates the extension fragments and detects the fluorescent ddNTP (dye-terminator) attached to the end of each fragment during cycle sequencing. The digital raw data produced were then analyzed using Sequencing Analysis 5.1 (http://sequencing-analysis.software), which processes the raw data and displays it as an electropherogram.

2.4 Sequencing of transcripts

The *SURF1* splice site mutation we identified with whole exome sequencing was predicted by the mutation analysis software Alamut Visual 2.6.1 (Interactive Biosoftware) to cause a skip of exon 2. To test this prediction, we used the RNA extracted from the blood samples (see 2.1) to synthesize cDNA. As described below, the cDNA was PCR amplified, checked on an agarose gel, and Sanger sequenced.

2.4.1 cDNA synthesis and PCR amplification

Total RNA from patient and parents was reversely transcribed to single-stranded cDNA with the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems) according to producer's specifications. Approximately 800 ng total RNA was used in each 20 µL reaction. Reverse transcription was performed in a Veriti 96-Well Thermal Cycler (Applied Biosystems) using the following cycling conditions: 10 min at 25°C, 120 min at 37°C, and 5 min at 85°C. For amplification of the cDNA produced from *SURF1* transcripts, primers were designed to specifically target exon 1 and 5 of the transcripts in order to detect a potential skipping of exon 2 (Table 6). The primer design was carried out manually due to the small size and high GC content of the exons. The *SURF1* cDNA was amplified with the Platinum *Taq* PCR_x DNA Polymerase Kit using $MgSO_4$ in the mix instead of $MgCl_2$ (Table 7). PCR mix components and cycling parameters used are summarized in Table 6. The annealing temperature (57°) was optimized experimentally.

Name	Sequence (5'- 3')	Primer size (bp)	T_m (°C)	GC (%)	Exon
cDNA Fwd	TGCGTTGCAGCTGGGGCT	18	68.9	66.7	1
cDNA Rev	CCTCCCGGACAGGGTCCAC	19	67.5	73.7	5

Table 6 Primers used in PCR amplification of cDNA. The forward primer was targeted to exon 1 of the *SURF1* transcript, while the reverse primer was targeted to exon 5. *SURF1* contains 9 exons in total. The full amplicon size – including exon 2 – is 422bp; without exon 2 the length of the amplicon is 370bp. Melting temperatures were calculated by In-Silico PCR assuming 50 mM salt and 50 nM annealing oligo concentration.

Component	Volume (µL)
H_2O	14.9
10X PCR buffer	2.0
10 mM dNTP	0.4
50 mM $MgSO_4$	0.8
Primers (10 µM)	0.8
Platinum *Taq* Polymerase	0.1
cDNA (40 ng/µL)	1.0
Total	20.0

Stage	Temperature	Time
Stage 1	95°C	2 min
Stage 2	95°C	45 sec
	57°C	30 sec
	68°C	1 min
Stage 3	4°C	∞

Table 7 PCR components and cycling conditions used in PCR amplification of cDNA.

2.4.2 Gel electrophoresis and Sanger sequencing of cDNA

The amplified cDNA fragments from the three subjects were run for 90 min on a 1.5% agarose gel in 1X TAE buffer and visualized with GelRed Nucleic Acid Gel Stain (Biotium) on the ImageQuant Las 4000 instrument (GE Healthcare Life Sciences). The gel electrophoresis resulted in one band for the patient sample and two bands for the parent samples. Of the two bands in the parent samples, the biggest was extracted and purified using the Wizard SV Gel and PCR Clean-Up System (Promega) according to manufacturer's instructions. This was not necessary to do for the patient sample as only one band was produced. The three samples were then Sanger sequenced following the same procedure as described above (2.3.1-2.3.3), consisting of AMPure, cycle sequencing, CleanSEQ, and capillary electrophoresis. The sequencing results were finally analyzed with Sequencing Analysis 5.1.

2.5 Sequencing the mitochondrial genome

To reliably sequence the mitochondrial genomes of patients 2 and 3, mtDNA enriched samples were prepared according to the mtDNA extraction protocol designed by Quispe-Tintaya *et al.* [52]. This method consists of two steps: mtDNA extraction using a standard kit for isolation of bacterial plasmids, and additional purification using AMPure beads. The principles underlying these two techniques are described here.

Conventional kits for isolation of bacterial plasmids are based on the alkaline extraction method described by Birnboim and Doly [71]. After harvesting and resuspension, bacterial cells are lysed in NaOH and sodium dodecyl sulfate (SDS). SDS solubilizes the phospholipid and protein components of the cell membrane, leading to lysis and release of the cell contents, while NaOH denatures proteins and DNA. The lysate is then neutralized under high salt conditions, causing denatured proteins, SDS and high molecular weight chromosomal DNA to co-precipitate and form insoluble complexes, while the plasmids, because they are closed circular molecules, renature and stay in solution [72]. In addition to neutralizing the lysate, the high salt concentration adjusts the DNA to adsorption onto the silica membrane that is now often used in plasmid extraction kits [72, 73]. Because mtDNA is also covalently closed circular molecules, bacterial plasmid isolation kits can be used to purify mtDNA. It should be noted, however, that nuclear DNA is prone to shearing during the lysis procedure because it consists of large linear molecules. In the protocol, mixing must thus be carried out carefully since nuclear DNA fragments are chemically indistinguishable from mtDNA under the conditions used. The two species will consequently not be separated on the silica membrane and will both elute under the same low-salt conditions [72].

The use of solid-phase reversible immobilization (SPRI) for purification of double-stranded PCR products was first described by Hawkins *et al.* [74]. They discovered that carboxyl-coated paramagnetic[6] beads could reversibly bind DNA under conditions of high polyethylene glycol (PEG) and salt. The size of the DNA fragments that bind to the beads depends on the PEG concentration: the lower the bead:DNA ratio, the larger the fragments that will bind. For kits that are commercially available today, the recommended ratio is usually 1.8 so that the beads will selectively bind DNA fragments of 100 base pairs and larger, while excess primers, nucleotides, salts, and enzymes are washed away [75]. But in the method for mtDNA

[6] Paramagnetic beads are magnetic only in a magnetic field. This prevents them from clumping together and falling out of solution.

enrichment that Quispe-Tintaya *et al.* designed, beads were added in a 0.4x proportion by volume in order to remove excess primers as well as the bulk of shared nuclear DNA. In this way, they were able to achieve more than 2000-fold enrichment for mtDNA when compared with total cellular DNA as demonstrated by real-time PCR [52].

2.5.1 Establishment and culture of skin fibroblasts

Fibroblasts were established from a skin biopsy of patients 2and 3 under local anesthesia. The skin fibroblasts were cultured at 37°C and 5% CO_2 in 75-cm^2 tissue culture flasks containing 10 mL of Dulbecco's Modified Eagle Medium (DMEM) (Life Technologies) supplemented with 1 M HEPES buffer solution, Penicillin-Streptomycin, and 10% fetal bovine serum (FBS). After the cells had grown to confluence, the medium was removed and the culture surfaces of the 75-cm^2 flasks were washed twice with 10 mL of Dulbecco's Modified Phosphate-Buffered Saline (DPBS) (Life Technologies). To allow cells to detach from the culture flask surface, 2 mL of 0.05% trypsin/EDTA were added to each flask. After 3-5 min of incubation at 37°C, the cell solution was placed in a 15 mL conical tube containing 10 mL DMEM, and 50 µL of the solution was used to count the cells with NucleoCounter NC-100 (Chemometec) according to manufacturer's instructions. The cell solutions were then centrifuged for 10 min at 400 x *g*. The pelleted cells were washed once with 1.5 mL DPBS and transferred to 2 mL micro centrifuge tubes. Approximately 10 million cells were used for each mtDNA extraction. In addition, 3 million cells were used to extract genomic DNA from each sample for control purposes.

2.5.2 mtDNA and gDNA extraction

The QIAprep Spin Miniprep Kit (Qiagen) was applied as the first step for mtDNA extraction and enrichment from the cell pellets. The procedure was performed according to manufacturer's recommendations, and DNA was eluted in 80 µL of elution buffer that was pre-heated to 70°C. Optional LyseBlue reagent was used to avoid inefficient cell lysis and incomplete precipitation of SDS, nuclear DNA, and cell debris. Mixing during the lysis procedure was carried out by slow and gentle inversion of the tube to reduce shearing of nuclear DNA. The extracted DNA was quantified and assessed for purity on the NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific). In parallel, GenElute Mammalian Genomic DNA Miniprep Kit (Sigma) was used to extract genomic DNA (gDNA) from the

cell pellets of both samples. DNA was eluted in 200 µL of elution solution, and purity and quantity was measured with NanoDrop.

2.5.3 Further purification and enrichment with AMPure

Fractions of the mtDNA-enriched samples were further purified manually with Agencourt AMPure XP system (Beckman Coulter) using a DynaMag-96 Side Skirted magnet (Life Technologies). The magnetic beads were added in a 0.45x proportion by volume, collected on the magnetic stand, and washed twice with freshly prepared 70% ethanol. After air-drying the beads, the mtDNA was resuspended in 40 µL of 10 mM Tris-Cl, pH 8.5. For comparison purposes, the gDNA samples were likewise purified with the AMPure XP system and resuspended in 40 µL of Tris-Cl buffer.

2.5.4 Relative quantification of mtDNA enrichment

To measure the relative level of mtDNA enrichment obtained with this method we performed real-time qPCR using SYBR Green JumpStart *Taq* ReadyMix (Sigma) for detection of amplification. Two primer pairs were used, one specific for the mitochondrial gene $tRNA^{Leu(UUR)}$, and the other specific for the nuclear gene β-2-microglobulin (*β2M*) [76] (Table 8). The mitochondrial $tRNA^{Leu(UUR)}$ gene region is rarely deleted and contains only a few single nucleotide polymorphisms, while *β2M* is a single copy nuclear gene and was used for normalization [37, 76]. The specificity of the primers was checked with In-Silico PCR (http://genome.ucsc.edu/cgi-bin/hgPcr). The real-time qPCR reaction (Table 9) was performed in triplicate for each reaction and fluorescent signal intensity of PCR products was recorded and analyzed on QuantStudio 12K Flex (Applied Biosystems) using QuantStudio 12K Flex Software v1.2.2. Relative levels of mtDNA enrichment were calculated with the formula $RQ=2^{(-\Delta\Delta Ct)}$ using the gDNA sample as reference and *β2M* as endogenous control.

Name	Sequence (5'- 3')	Amplicon size (bp)	T_m (°C)	GC (%)
tRNA F3212	CACCCAAGAACAGGGTTTGT	107	59.9	64.2
tRNA R3319	TGGCCATGGGTATGTTGTTA		59.7	45.0
ß2M F594	TGCTGTCTCCATGTTTGATGTATCT	86	62.1	65.9
ß2M R679	TCTCTGCTCCCCACCTCTAAGT		61.7	65.6

Table 8 qPCR primer sequences. Melting temperatures were calculated by In-Silico PCR assuming 50 mM salt and 50 nM annealing oligo concentration.

Component	Volume (µL)
H_2O	2.0
Fwd primer (5 µM)	1.0
Rev primer (5 µM)	1.0
2X SYBR Green ReadyMix	5.0
DNA (1.0 ng/µL)	1.0
Total	10.0

Stage	Temperature	Time
Stage 1	50°C	2 min
	95°C	10 min
Stage 2	95°C	15 sec
	60°C	1 min
Stage 3	95°C	15 sec
	60°C	1 min
	95°C	15 sec

Table 9 qPCR reaction mixture and cycle parameters. The internal reference dye ROX was added to the SYBR Green mix for reaction normalization. Stage 3 of the qPCR reaction was for melting curve analysis. Stage 2: 40 cycles.

2.5.5 PCR amplification of mtDNA

Fractions of the mtDNA enriched samples were subjected to 10 cycles of PCR amplification to check whether PCR improves the sequencing results as compared to using non-amplified mtDNA samples. By limiting the number of PCR cycles, the potential production of artifacts is also reduced. The mtDNA was amplified using a set of 9 overlapping primer pairs covering the whole mitochondrial genome (Table 11; Figure 8). These primers were validated and tested by Ramos *et al.* [51] to selectively amplify the mtDNA and avoid co-amplification of nDNA sequences of mitochondrial origin (NUMTs). Each of the 9 mtDNA fragments was amplified separately on a Veriti 96-Well Thermal Cycler (Applied Biosystems) using AmpliTaq Gold DNA Polymerase with Gold Buffer (Applied Biosystems). PCR reaction mix components for each primer pair and cycle parameters are shown in Table 10. The amplified PCR fragments were pooled together, purified with the AMPure XP system adding beads in a 1.8x proportion by volume, and eluted in 25 µL Tris-Cl buffer. The mtDNA samples were then quantified using Qubit 2.0 Fluorometer (Life Technologies).

Component	Volume (µL)
H_2O	16.4
Primers (10 µM)	1.0
10 mM dNTP	0.5
25 mM $MgCl_2$	1.5
AmpliTaq Gold (5U/µL)	0.1
10X Gold Buffer	2.5
mtDNA (~7 ng/µL)	3.0
Total	25.0

Stage	Temperature	Time
Stage 1	96°C	5 min
Stage 2	94°C	1 min
	56°C	40 sec
	72°C	2.5 min
Stage 3	72°C	5 min
	4°C	∞

Table 10 PCR reaction mixture and cycle parameters used for amplification of the whole mtDNA.

Fragment	Amplicon size	Name	Sequence (5'-3')	Primer length	Tm	GC (%)
1	1822	1489F	tagccatgcactactcaccaga	22	60.8	50
		151R	ggatgaggcaggaatcaaagac	22	62.2	50
2	1758	1648F	ctgtatccgacatctggttcct	22	60.4	50
		1677R	gtttagctcagagcggtcaagt	22	60.1	50
3	2543	1404F	acttaagggtcgaaggtggatt	22	60.2	45.5
		3947R	tcgatgttgaagcctgagacta	22	60.0	45.5
4	3005	3734F	aagtcaccctagccatcattcta	23	59.2	43.5
		6739R	gatatcatagctcagaccatacc	23	54.6	43.5
5	2709	6511F	ctgctggcatcactatactacta	23	54.2	43.5
		9220R	gattggtgggtcattatgtgttg	23	61.3	53.5
6	1738	8910F	cttaccacaaggcacacctaca	22	60.1	50
		10648R	ggcacaatattggctaagaggg	22	62.0	50
7	1866	10360F	gtctggcctatgagtgactaca	22	57.0	50
		12226R	cagttcttgtgagctttctcgg	22	61.5	50
8	1853	11977F	ctccctctacatatttaccacaac	24	56.1	41.7
		13830R	aagtcctaggaaagtgacagcga	23	61.6	47.8
9	1872	13477F	gcaggaatacctttcctcacag	22	60.1	50
		15349R	gtgcaagaataggaggtggagt	22	59.6	50

Table 11 Validated primers to amplify the complete mtDNA in nine overlapping fragments. Melting temperatures were calculated by In-Silico PCR assuming 50 mM salt and 50 nM annealing oligo concentration (from Ramos *et al.* [51]).

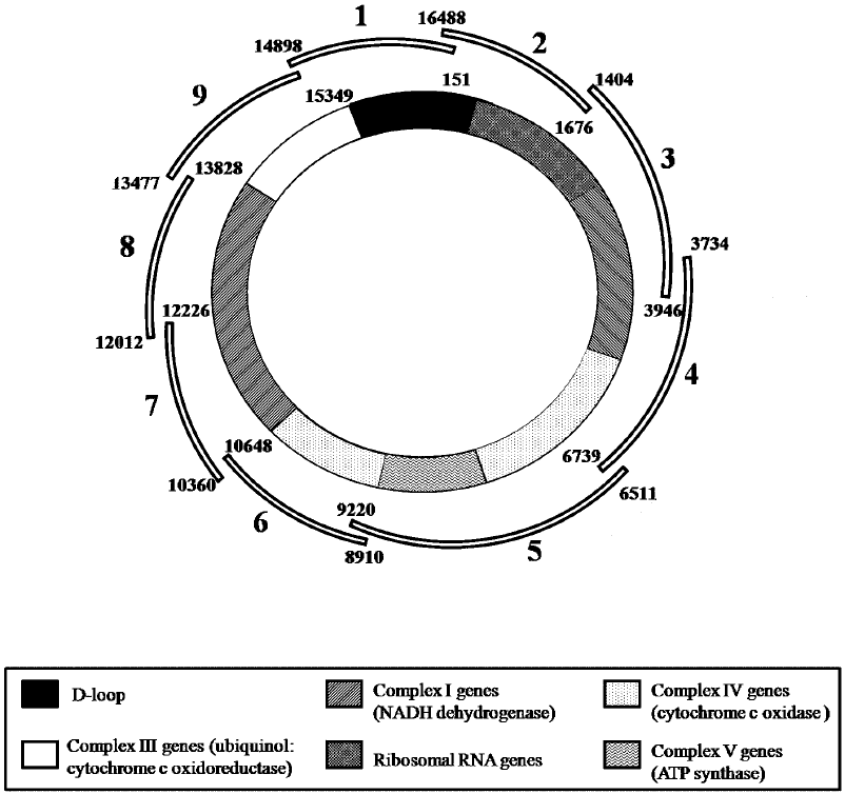


Figure 8 Schematic representation of mtDNA. The nine overlapping fragments defined to PCR amplify the complete mtDNA genome are represented as well as the different gene regions in human mtDNA (from Ramos *et al.* [51]).

2.5.6 Massively parallel sequencing of mtDNA

In total, five mtDNA enriched samples from patients 2 and 3 were sequenced by massively parallel sequencing (Table 12). Sample preparation method and sequencing platform were chosen in consultation with Gregor Duncan Gilfillan of the Norwegian Sequencing Centre (NSC). Individual libraries were constructed for each of the five samples by randomly shearing the mtDNA with a Covaris E220 ultrasonicator (Covaris) producing double-stranded fragments with an average size of 500bp. The sheared mtDNA was concentrated using 1.8 volumes of AMPure XP beads (Beckman Coulter) according to manufacturer's instructions and eluted in 33 µL 10 mM Tris. Preparation of the libraries was carried out using the MicroPlex Library Preparation kit (Diagenode) according to manufacturer's recommendations. Maximum input volume of 10 µL (1-3 ng) was used for all samples (Table 12). Indices were assigned to allow sequencing of all samples in one lane. The Microplex libraries were amplified by a total of 14 PCR cycles (4 cycles in step 4 and 10 cycles in step 5 of Section 3 in the Microplex procedure, version 3). Sequencing was performed by NSC on an Illumina MiSeq using paired end 300bp V3 reagents according to manufacturer's instructions. Image analysis and base calling were performed using Illumina's RTA software version 1.18.54. Reads were filtered to remove those with low base call quality using Illumina's default chastity criteria.

Sample	Patient	mtDNA enrichment method	Input volume (µL)	mtDNA input (ng)
1	2	MP + PCR	10	1.29
2		MP + AMP	10	0.96
3	3	MP + PCR	10	3.12
4		MP + AMP	10	1.36
5		MP + AMP + PCR	10	2.42

Table 12 mtDNA enriched samples and input amount for massively parallel sequencing. MP, miniprep; AMP, purification using 0.45x AMPure beads; PCR, 10 cycles of PCR amplification using nine overlapping primer pairs.

Reads that passed Illumina's quality filter were aligned to the reference human genome (GRCh37/hg19) using Burrows-Wheeler Aligner tool (http://bwa.sourceforge.net/). PCR duplicates were marked with Picard tools (http://broadinstitute.github.io/picard/). Variant calling and filtering was performed using the Genome Analysis Toolkit (GATK) (https://www.broadinstitute.org/gatk/), and variants were annotated with SnpEff (http://snpeff.sourceforge.net/).

2.6 Detection of mtDNA deletion in patient 3 muscle tissue

A frozen muscle sample from patient 3 was sent to the Nijmegen Center for Mitochondrial Disorders (NCMD) for mitochondrial DNA analysis. From DNA isolated from the muscle sample the complete mtDNA was screened for mtDNA rearrangements using Long Template PCR. This revealed the almost homoplasmic deletion m.6211_8681del in the patient's muscle tissue.

2.6.1 Detection and verification of mtDNA deletion in skin tissue

To determine whether the same mtDNA deletion (m.6211_8681del) was present in the patient's skin tissue, we performed PCR on genomic DNA isolated from skin fibroblasts using a set of designed primers flanking the deletion breakpoints (Table 13). PCR reaction mixture and cycling conditions were the same as in 2.5.5 with the exception that 50-60 ng of genomic DNA was used as template. In addition to genomic DNA from fibroblasts, we also performed PCR on genomic DNA isolated from the patient's blood. DNA from a commercial fibroblast cell line was used as control.

Name	Sequence (5'- 3')	T_m (°C)	GC (%)
m6073For	TCGTCACAGCCCATGCATTT	64.9	50.0
m9065Rev	GCTAGGGTGGCGCTTCCAAT	65.8	60.0

Table 13 PCR primers used for detection of mtDNA deletion. The primers were designed using Primer3 and checked for sequence specificity using In-Silico PCR. Universal M13 tales (not shown) were also added for later Sanger sequencing. The amplicon size is 2993bp with the deleted region included and 579bp without it. With the M13 tales included, the amplicon size is 3029bp and 559bp, respectively.

PCR amplification resulted in two distinct bands for the patient samples and one band for the control sample. To verify that these bands corresponded to the full-length and deleted fragments, we Sanger sequenced the amplified PCR fragments. DNA was first extracted from

the agarose gel and purified using QIAquick Gel Extraction Kit (QIAGEN) and then Sanger sequenced following exactly the same procedure as in 2.3-2.3.3. The raw sequence data generated was finally analyzed using Sequencing Analysis 5.1.

2.6.2 Quantification of heteroplasmy levels of mtDNA deletion

The mtDNA deletion m.6211-8681del was verified by Sanger sequencing in both muscle and blood. The heteroplasmy levels of the deletion in the two tissues were quantified with qPCR using two primer pairs (Table 14). The first pair was designed to flank the deleted region. The short extension time used in this qPCR experiment (1 min) means that amplification will take place only if the deletion is present. This is because the amplicon size would be 2561bp with the deleted fragment included, which is too much to be amplified during the short extension time. On the other hand, the amplicon size is only 91bp without the deleted fragment and amplification therefore readily takes place. This primer pair was thus used to detect only those mtDNA molecules that contain the deletion. The second primer pair was targeted to the $tRNA^{Leu(UUR)}$ gene region, which is not deleted in the patient and which is not a common site for single nucleotide polymorphisms. In addition, this region does not harbor any of the point mutations detected by massively parallel sequencing. The tRNA primer pair was therefore designed to detect all mtDNA molecules regardless of whether the deletion was present or not. When designing the primers, amplicon size and melting temperatures were adjusted as much as possible and sequence specificity was checked with both BLAT and In-Silico PCR. The PCR conditions and amplification program were the same as in 2.5.4 except that 2 ng of gDNA was used.

Name	Sequence (5'- 3')	Amplicon size (bp)	T_m (°C)	GC (%)
mDEL F6170	CCCCGATATGGCGTTTCC	91	63.5	61.1
mDEL R8730	TCAGGTTCGTCCTTTAGTGTTGTG		62.6	45.8
tRNA F3244	GCCCGGTAATCGCATAAAACTTA	98	63.2	43.5
tRNA R3341	GGTACAATGAGGAGTAGGAGGTTGG		63.3	52.0

Table 14 Primers used for deletion quantification with qPCR.

We also performed a primer efficiency test for more accurate relative quantification measurements. The test consisted of five points and was run in triplicates using 16 ng as starting quantity and 1:2 as serial factor. Genomic DNA extracted from patient 3 fibroblasts was used as template.

3 Results

3.1 Whole exome sequencing and filtering of genetic variants

Whole exome sequencing (WES) of DNA from the three families was performed as described in 2.2-2.2.2. Analysis and filtering of called variants (see 2.2.3) resulted in 25 potentially pathogenic variants in patient 1 (Table 15), 19 variants in patient 2 (Supp. Table S1), and 21 variants in patient 3 (Supp. Table S2). Literature searches using OMIM and PubMed were carried out to identify causative links between the detected variants and the clinical manifestations of the patients. Based on these searches, the homozygous *SURF1* variant c.106+1G>C was concluded to be the best candidate in patient 1. None of the putative variants found in patients 2 and 3 were considered to be causative of the clinical phenotypes.

Table 15 Putative pathogenic variants in patient 1 detected by whole exome sequencing and retained after filtering. The *SURF1* variant is marked in red.

Gene name	Chr:pos	Ref	Alt	Amino acid change	Effect	Depth
TENM1	X:123519883	C	T	R1900H	Non-synonymous	86
NOS2	17:26093598	A	C	H689Q	Non-synonymous	109
TCHH	1:152080993	C	G	R1567T	Non-synonymous	93
GTF3C5	9:135930371	G	C	S278T	Non-synonymous	86
PDCD11	10:105202091	G	T	R1610L	Non-synonymous	42
CEBPD	8:48650242	C	A	*134L	Stop lost	10
UBTF	17:42289077	T	G	N278T	Non-synonymous	56
SCN4A	17:62018451	C	T	A1731T	Non-synonymous	101
FCN1	9:137804452	G	A	R148W	Non-synonymous	46
FRAS1	4:79188076	T	G	L259R	Non-synonymous	91
ARIH1	15:72767034	TGAG	T	E19-	In-frame codon deletion	50
SURF1	**9:136223123**	**G**	**C**	**c.106+1**	**Splice site substitution**	**11**
SMTN	22:31486854	C	A	A301D	Non-synonymous	103
PROM2	2:95941792	G	A	G137R	Non-synonymous	97
WRAP53	17:7604812	T	C	I223T	Non-synonymous	104
LAMC3	9:133920949	C	T	P474L	Non-synonymous	89
PCDHA12	5:140256415	C	T	P453L	Non-synonymous	38
TRIM45	1:117659283	C	T	G379E	Non-synonymous	32
BMX	X:15548115	T	G	S302A	Non-synonymous	47
SLC27A3	1:153752324	C	G	T141S	Non-synonymous	99
PCYT1B	X:24597545	C	G	G181A	Non-synonymous	37
RAB25	1:156035733	G	C	K25N	Non-synonymous	58
SRRM2	16:2820684	C	T	P220S	Non-synonymous	70
CCDC120	X:48926120	C	T	R638C	Non-synonymous	63
GABRE	X:151123944	C	T	A345T	Non-synonymous	42

3.2 Verification of c.106+1G>C variant in *SURF1*

SURF1 is one of at least six nuclear genes controlling the assembly of the 13 subunits constituting complex IV (cytochrome c oxidase (COX)) of the respiratory chain. Mutations in *SURF1* are primarily associated with Leigh and Leigh-like syndrome [77]. The WES results showed a low depth of coverage (11x) of the *SURF1* variant position compared to the other variants detected (Table 15). However, Sanger sequencing confirmed that the patient was homozygous for the variant while the parents were heterozygous (Figure 9).

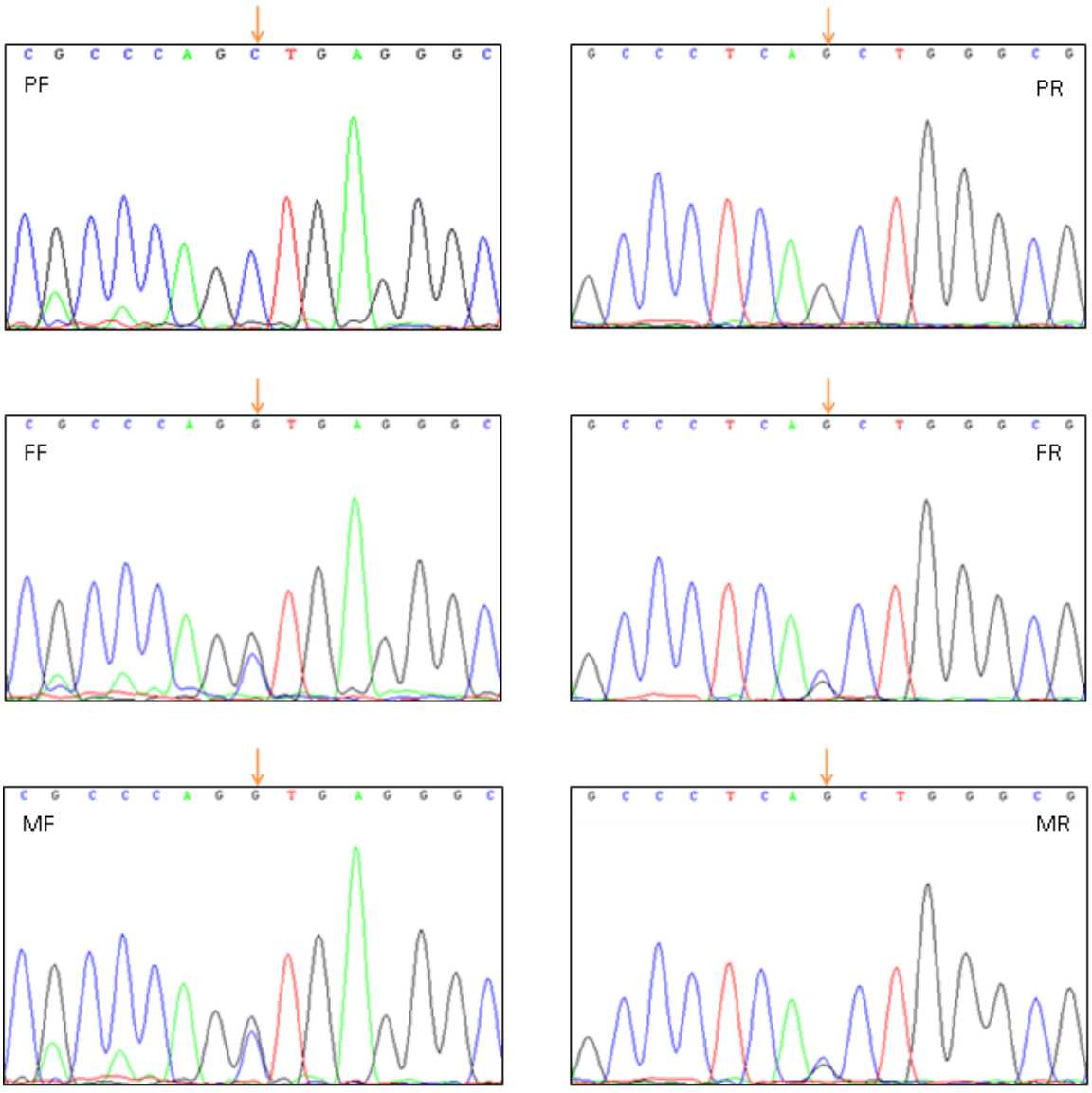


Figure 9 Confirmation by Sanger sequencing of the c.106+1G>C variant in *SURF1*. The variant was confirmed in both forward and reverse directions. PF, patient forward; FF, father forward; MF, mother forward; PR, patient reverse; FR, father reverse; MR, mother reverse.

3.2.1 Skipping of exon 2 in *SURF1*

The c.106+1G>C variant is located in the donor splice site of intron 2 and was predicted by Alamut Visual 2.6.1 (Interactive Biosoftware) to cause a skip of exon 2. Since exon 2 is 52bp, skipping results in a shift of the reading frame, and nonsense-mediated decay (NMD) of the transcript is likely to result [78]. This variant has previously been reported in a Pakistani patient with Leigh syndrome [57], but the authors did not document the skipping of exon 2. Using primers targeted to exon 1 and 5 of the transcript (Supp. Figure S1), PCR amplification of cDNA gave only one fragment of approximately 400bp in the patient, while this same fragment and one larger fragment were detected in both parents (Figure 10). The expected size of the amplicon was 422bp if exon 2 was retained and 370bp if it was skipped. Our PCR result was therefore in agreement with the prediction that the c.106+1G>C splice site mutation results in skipping of exon 2. Furthermore, the result supported the finding that the patient was homozygous for the mutation while the parents were heterozygous.

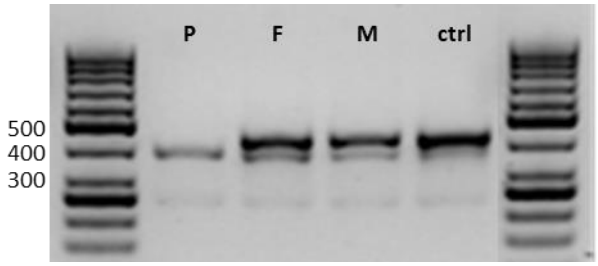


Figure 10 Exon 2 skipping in *SURF1*. The PCR results support the prediction that the splice site variant leads to a skip of exon 2. P, patient; F, father; M, mother.

To assess if the genetic variant causes exon 2 skipping, the fragments obtained with PCR were Sanger sequenced (see 2.4.1-2.4.2). Sanger sequencing confirmed that exon 2 was absent in the fragment amplified from patient DNA only, as exon 3 was continuous with exon 1 in the patient, while it was continuous with exon 2 in both parents (Figure 11).

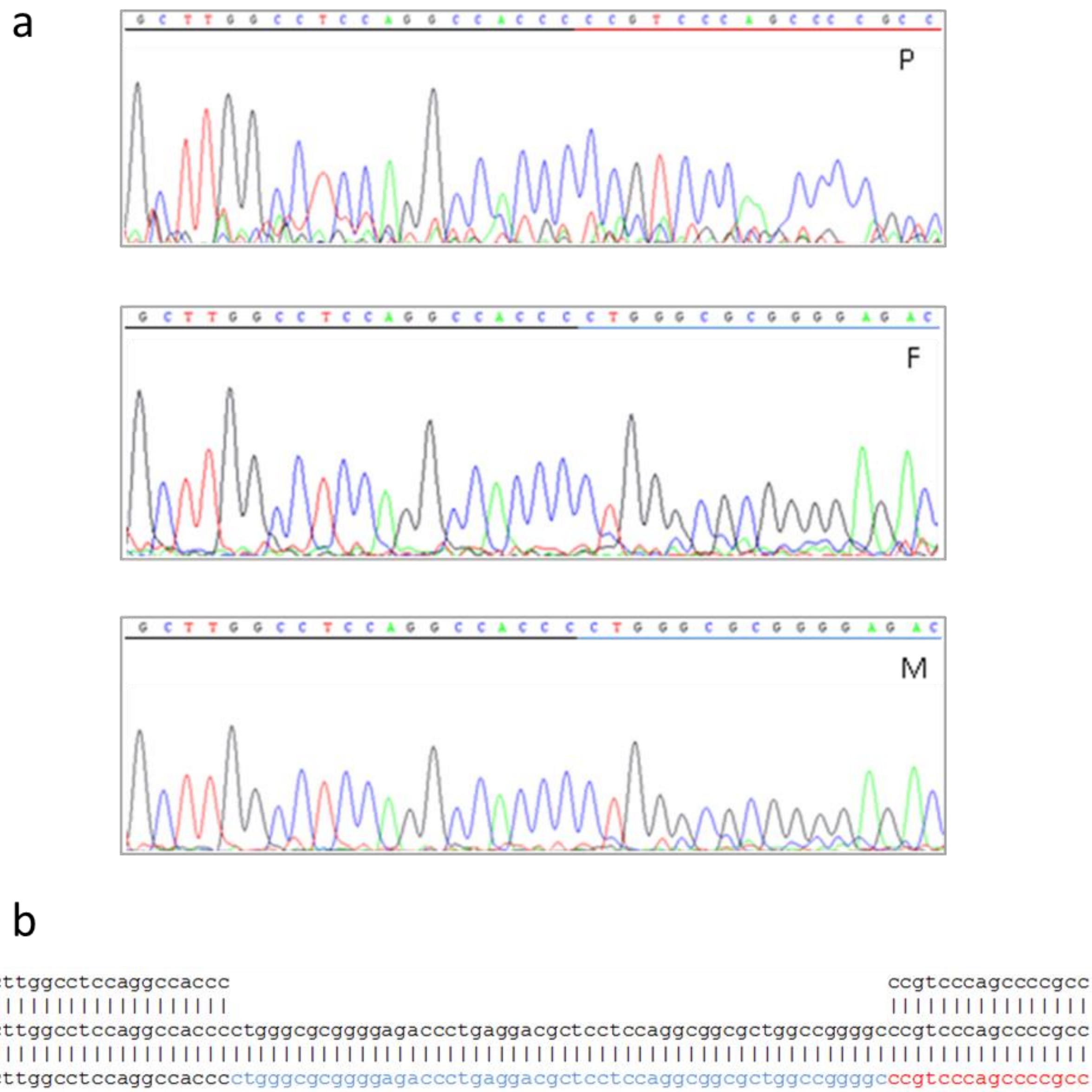


Figure 11 Verification of exon 2 skipping in patient 1. a) The electropherograms show that exon 3 (underlined in black) is continuous with exon 1 (underlined in red) in the patient, while it is continuous with exon 2 (underlined in blue) in both parents. The sequencing data shown is for the reverse strand. **b)** Alignment of cDNA sequence from patient (top) and parents (middle) to the human reference sequence (GRCh38 assembly) shows that exon 2 is absent in patient DNA but retained in parental DNA. In the reference sequence, exons 3, 2, and 1 are shown in black, blue, and red, respectively. The sequence data shown is for the reverse strand. P, patient; F, father; M, mother.

3.3 Sequencing the mitochondrial genome

Since we could not detect any potentially pathogenic variants with whole exome sequencing in patients 2 and 3, we sought to sequence the mitochondrial genome to search for pathogenic mtDNA point mutations or large deletions that could explain the clinical manifestations of these two patients. In order to avoid the generation of artifacts associated with PCR

amplification, we enriched the samples for mtDNA using the method developed by Quispe-Tintaya *et al.* [52]. In addition, we carefully chose a DNA library preparation protocol and high-throughput sequencing platform that would result in deep and even coverage of the entire mtDNA (see 2.5-2.5.6). Deep coverage allows detection and quantification of low-level heteroplasmic point mutations that may not be detected by other sequencing methods like Sanger sequencing [79].

3.3.1 Relative quantification of mtDNA enrichment

Enrichment of the samples for mtDNA was carried out in two stages using QIAGEN Miniprep Kit and AMPure purification beads. Relative mtDNA enrichment levels were measured by qPCR using two primer sets (see 2.5.4: Table 8; Figure 12a). The tRNA primers were specific to an mtDNA region that is rarely deleted and contains only a few rare single nucleotide polymorphisms (mtSNP), while the β2M primers were specific to a single copy nuclear gene. Relative enrichment (RQ) was calculated by the formula $RQ = 2^{(-\Delta\Delta Ct)}$, in which the total cellular DNA sample was used as reference and the β2M target as endogenous control. In other words, the difference between the Ct values of tRNA and β2M in the mtDNA enriched samples was normalized to the difference between the Ct values of tRNA and β2M in the total cellular DNA sample. The qPCR measurement indicated a 280 and 730 fold enrichment after AMPure purification of the mtDNA samples for patients 2 and 3, respectively (Figure 12b). Although the relative enrichment was apparently higher in patient 3, the actual number of mtDNA copies in the enriched samples themselves was 9 times greater in patient 2 than in patient 3, as indicated by the difference in Ct values (ΔCt) between tRNA and β2M in the "miniprep and AMPure" samples (Figure 12a; Table 16). The reason for this is that the ΔCt value in the total cellular DNA sample, to which the enriched samples were normalized, was much higher in patient 2 than it was in patient 3 (Figure 12a). The relative enrichment value is therefore not a measurement of the actual mtDNA copy number in the sample.

a

Patient 2	Mean Ct values tRNA	β2M
Total cellular DNA	19.38	28.44
QIAGEN Miniprep Kit	14.34	28.40
QIAGEN Miniprep Kit and AMPure	13.32	30.49

Patient 3	Mean Ct values tRNA	β2M
Total cellular DNA	21.56	26.51
QIAGEN Miniprep Kit	16.65	27.80
QIAGEN Miniprep Kit and AMPure	16.65	30.66

b

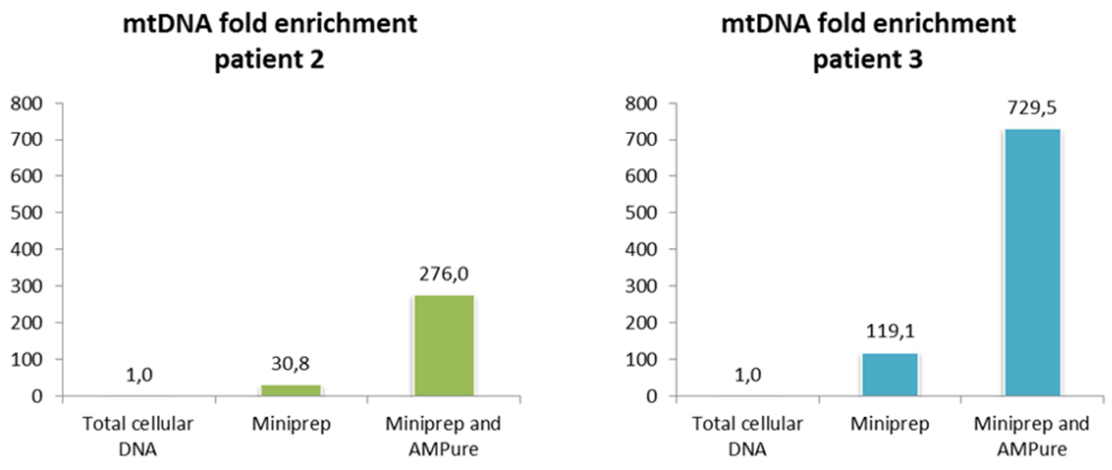


Figure 12 Relative quantification of mtDNA enrichment. a) Mean Ct values obtained after qPCR quantification of mitochondrial and nuclear DNA. tRNA is the mtDNA target, while β2M is the nuclear DNA target. Note that the ΔCt value for the AMPure sample is much higher in patient 2 than it is in patient 3, but the same is the ΔCt value for the total cellular DNA sample. That is why the mtDNA fold enrichment was greater for patient 3. **b)** Relative enrichment of mtDNA over nuclear DNA using miniprep kit alone and miniprep kit plus AMPure. In both patients, AMPure significantly increased the mtDNA enrichment levels.

3.3.2 Massively parallel deep sequencing of mtDNA

In total, five mtDNA enriched samples were massively parallel sequenced; two samples from patient 2 and three samples from patient 3 (see 2.5.6: Table 12). The samples were prepared using the MicroPlex Library Preparation Kit (Diagenode), which allows for as little as 50 picogram input of fragmented dsDNA. Paired-end sequencing was then performed on an Illumina MiSeq platform. Paired-end sequencing enables both ends of the DNA fragments to be sequenced. Since the distance between each paired read is known, alignment algorithms can use this information to map reads over regions containing repetitive sequences and to detect structural rearrangements. Because paired-end reads are more likely to align correctly to a reference, the data quality as well as the overall coverage level greatly improve [80]. Improved depth of coverage is one of the crucial factors in the detection of low-frequency heteroplasmic mtDNA variants [42].

As Figure 13 shows, a depth of coverage ranging from 10,000 to 130,000 per base position was achieved for all the samples, although the coverage was not uniform in all cases. Only with the two PCR free samples did we achieve uniform coverage across the entire mitochondrial genome. The depth of coverage in these two samples was approximately 60,000 for patient 2 and 40,000 for patient 3 (Figure 13 A and B). This difference in coverage was probably due to the fact that the patient 2 sample had an mtDNA:nDNA copy number ratio that was 9 times greater than the patient 3 sample had (Table 16). Therefore, neither input amount nor relative mtDNA enrichment was predictive of the coverage level to be achieved by massively parallel sequencing of mtDNA. Both these values were greater in the sample from patient 3, but that was not reflected in the coverage results (Table 16).

On the other hand, the three samples that were enriched for mtDNA by PCR resulted in uneven coverage of the mtDNA genome (Figure 13 C-E). Moreover, the coverage pattern for each sample clearly mirrored the positions of the nine overlapping PCR fragments produced during amplification. The exception was fragments 3, 4 and 5, which had reduced coverage in all three samples (Figure 13 C-E). These three fragments were also the largest fragments and were probably amplified less efficiently do to the relatively short extension time that was used (2.5 min) (see 2.5.5: Table 11). The first 1000bp of the mtDNA genome had poor sequence coverage regardless of the enrichment method (Figure 13 A-E). The reason for this is unclear, but this region is part of the D-loop, which is a highly polymorphic non-coding mtDNA region. The more polymorphic a DNA region is, the less likely it is that it will align correctly to the reference genome, leading to apparently reduced coverage of that region [81]. Similar low coverage of this region has been observed before [82].

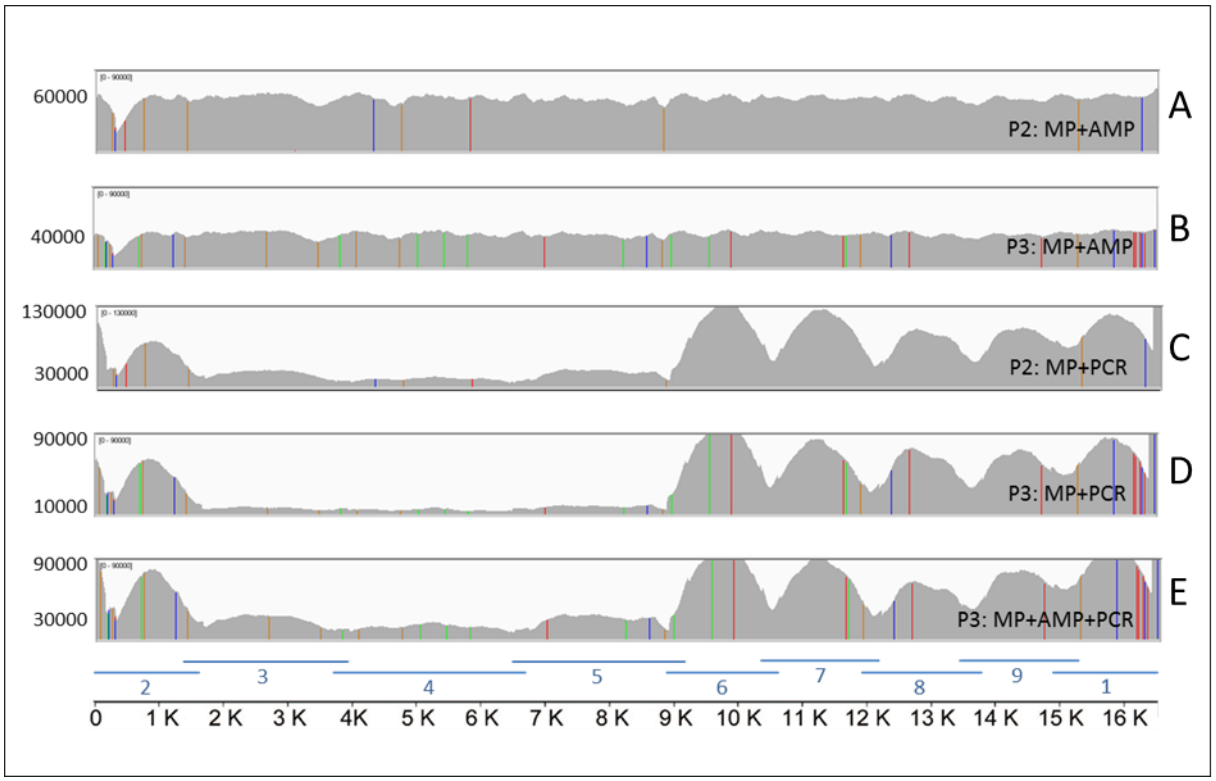


Figure 13 Depth of coverage achieved with and without the use of PCR for mtDNA enrichment. Samples presented in C, D, and E were enriched by 10 cycles of PCR using nine overlapping primer pairs. In all of these samples, deep sequencing resulted in an uneven coverage across the entire mitochondrial genome, ranging from 10,000 to as high as 130,000 (indicated in the Y-axis). In contrast, the coverage was uniform throughout the mtDNA genome when PCR was not used for target enrichment. The horizontal lines numbered from 1 to 9 represent the nine overlapping fragments amplified by PCR (See Figure 8). As can be seen, the coverage pattern in each of the three PCR samples correlates very closely with the distribution of the fragments. The colored vertical lines indicate the positions of the single nucleotide variants detected. P2, Patient 2; P3, Patient 3; MP, Miniprep; AMP, 0.45xAMPure; PCR, 10 cycles of PCR amplification.

Miniprep and AMPure		
	Patient 2	Patient 3
mtDNA:nDNA	147 464	16 498
Relative enrichment	276	729
Input (ng)	0.96	1.36
Coverage	~60 000	~40 000

Table 16 Relationship between depth of coverage and mtDNA:nDNA copy number ratio in "miniprep and AMPure" samples. The mtDNA:nDNA copy number ratio was calculated using the formula $2^{\Delta Ct}$, where ΔCt is the difference in Ct values between mtDNA target (tRNA) and the nDNA target (β2M). Note that relative enrichment levels and DNA input amount do not correlate with the achieved coverage.

3.3.3 Summary of sequencing results

PCR amplification during target enrichment for massively parallel sequencing complicates genotyping because PCR introduces sequence errors and duplicates template molecules [83]. Duplicates are defined as reads with the same start position and orientation [84]. Since they are identical copies of template molecules, they can strongly inflate and distort the read coverage data, making it necessary to remove them before carrying out coverage calculations in the alignment [85]. The percentage of duplicate reads was consistently higher in the PCR enriched samples than in the samples not amplified by PCR. Removing the duplicates therefore resulted in a significantly lower number of unique reads in the PCR amplified samples (Table 17). However, when considering the percentage of unique bases that aligned to the mtDNA genome, the number was significantly higher in two of the three PCR enriched samples. One advantage of PCR enrichment over PCR free enrichment seems therefore to be that the former is more effective in that it can achieve high target coverage with fewer reads. This advantage is also reflected in the fold enrichment, which is the fold by which the target region has been amplified above the reference genome (Table 17).

Sample	Total reads	Unique reads	% unique reads	Unique bases aligned to mtDNA	% unique bases aligned to mtDNA	Fold enrichment
P2: MP+PCR	11495296	4564993	39,7	856874111	87,7	166012
P2: MP+AMP	11106524	8873390	79,9	955347809	50,8	96280
P3: MP+PCR	9508602	5897473	62,0	581856061	49,5	93640
P3: MP+AMP	9076904	8390339	92,4	612600347	33,8	64028
P3: MP+AMP+PCR	9000376	3846280	42,7	731523477	92,0	174204

Table 17 Summary of sequencing results. Unique reads are reads not marked as PCR duplicates. MP, Miniprep; AMP, 0.45xAMPure; PCR, 10 cycles of PCR amplification

3.3.4 Quality assessment of sequencing data

Quality assessment of the raw sequence data from the mtDNA enriched samples was carried out with FastQC quality control tool (http://.bioinformatics.babraham.ac.uk/projects/fastqc), which runs a set of quality checks producing a report that allows for quick assessment of the overall quality of the run. One of these checks calculates the average Phred quality score per sequence read, and gives an overview of the quality score distribution over all sequences in the sample. The Phred score (Q) is logarithmically linked to the base calling error probabilities (P): Q = – 10 x log P [86]. For example, if Q is 30, this is equivalent to the probability of an incorrect base call being 1 in 1000. The most commonly used cutoff, however, is a Phred quality score of Q = 20 [87], meaning that the probability of incorrect base calling is 1 in 100. In each of the five mtDNA enriched samples, the majority of reads (>95%) had an average quality score above 20 (Figure 14; Supp. Figure S2; Supp. Figure S3). Interestingly, there were no marked differences in average quality per read between the PCR enriched samples and the PCR free samples.

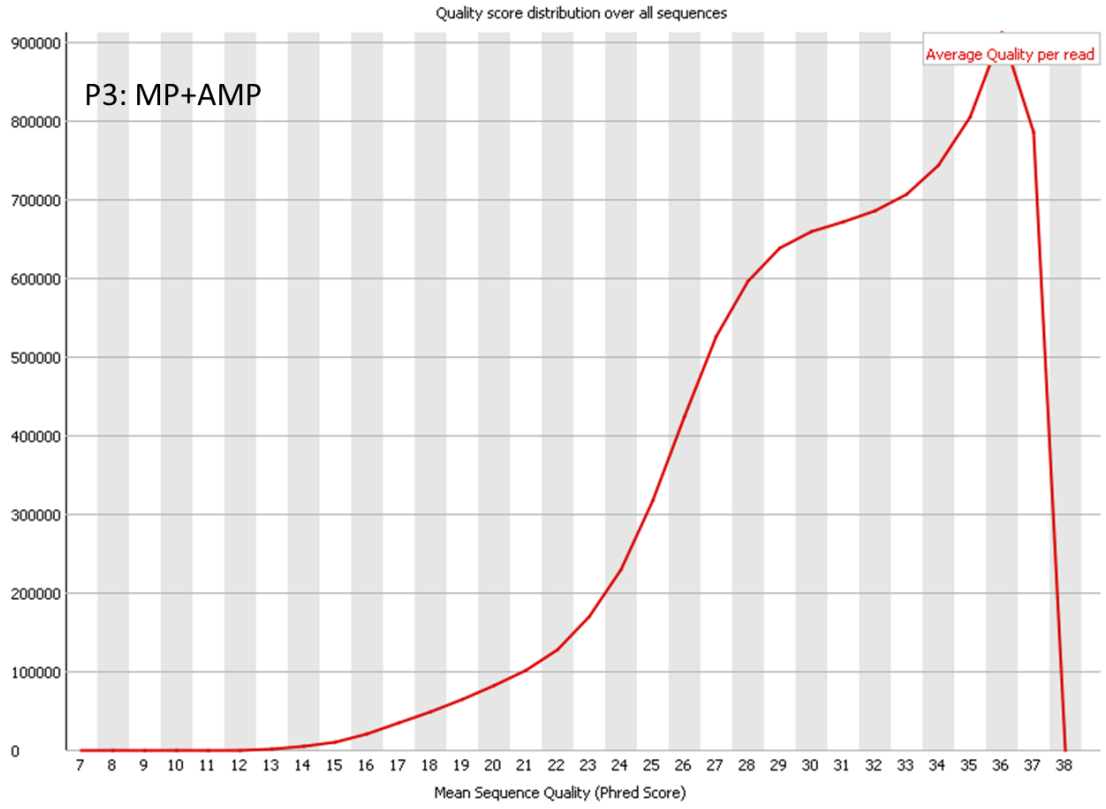


Figure 14 Distribution of average quality score per sequence. Ideally, only one peak should be seen to the far right, indicating that all sequences have the same average Phred score above 30. But even though our results were short of the ideal, the majority of sequences (>95%) had an average score above 20, which is the generally accepted cutoff. Only the results from one sample are shown here, the remaining samples are shown in Supplementary Figures S2 and S3. P3, patient 3; MP, miniprep; AMP, AMPure.

3.4 Detection of mtDNA deletion in patient 3

Mitochondrial DNA analysis of muscle tissue from patient 3 was carried out at the Nijmegen Center for Mitochondrial disorders (NCMD), which identified the almost homoplasmic deletion m.6211_8681del (Figure 15). To assess if the same deletion was present in cells from the patient's blood and skin as well, the genomic region containing this deletion was PCR amplified using a set of primers flanking the deletion breakpoints (see 2.6.1: Table 13). PCR amplification resulted in two fragments in DNA isolated from the patient's blood and skin tissue samples, while PCR in the control sample only produced one fragment (Figure 16). The theoretical size of PCR fragments from mtDNA molecules without the deletion is 3029bp, corresponding very well with the large fragment produced by PCR in both patient and control samples. The additional small PCR fragment produced in the patient samples indicated that the deletion was also present in the patient's blood and skin mtDNA.

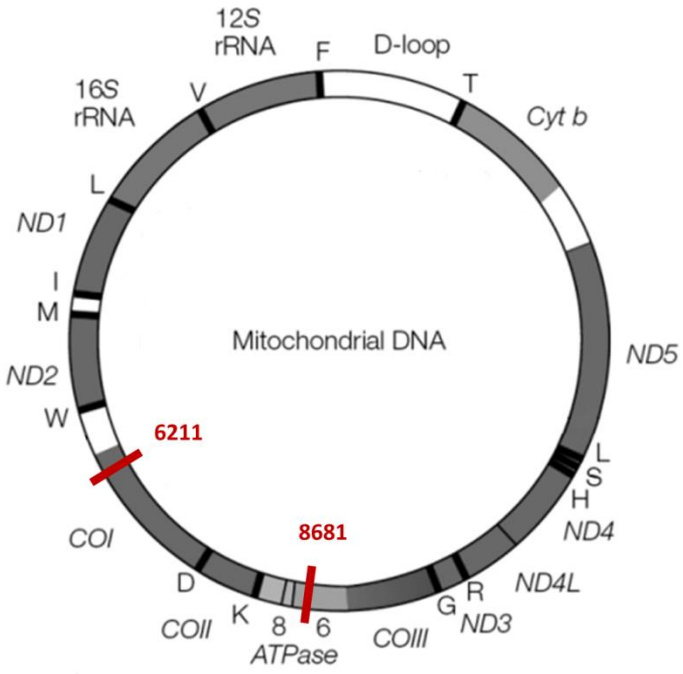


Figure 15 The mtDNA deletion detected in patient 3. The deletion m.6211_8681del is 2469bp and spans the genes *COI*, *TRND*, *COII*, *TRNK*, *ATPase8* and part of *ATPase6* (adapted from Zeviani *et al.* [56]).

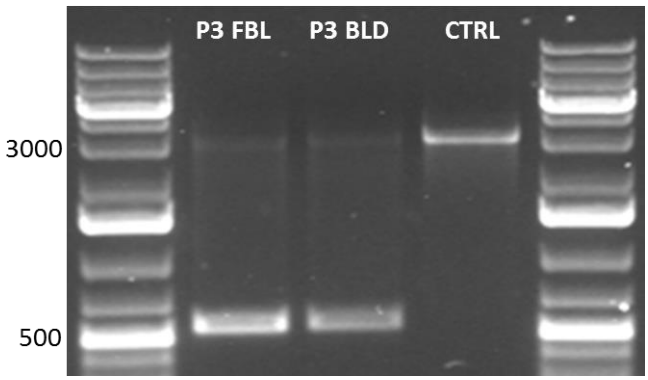


Figure 16 Detection of mtDNA deletion in blood and skin tissues from patient 3. The two PCR bands seen in the patient samples indicate that the mtDNA deletion was present in the patient's blood and skin in addition to muscle. The control sample only resulted in one PCR band corresponding to the full-length amplicon. The reason why the biggest bands in the patient samples is weaker than the corresponding band in the control sample, is probably that the DNA polymerase used in this PCR run is more efficient for smaller amplicons, which therefore out-compete the bigger ones. We also see that the small band in P3 FBL has stronger intensity than in PR BLD. This indicates that the heteroplasmy level of the deletion is greater in the patient's skin tissue than in blood. P3, patient 3; FBL, fibroblast; BLD, blood.

The small PCR fragments from the two patient samples were excised from the gel, purified, and Sanger sequenced to determine the sequence of the mtDNA region containing the deletion in the patient's fibroblast and blood DNA (see 2.6.1). As Figure 17 shows, the deletion was detected in both tissues and was verified in both forward and reverse directions, clearly indicating that the mtDNA region 6211-8681 was deleted in the patient's fibroblast and blood tissues.

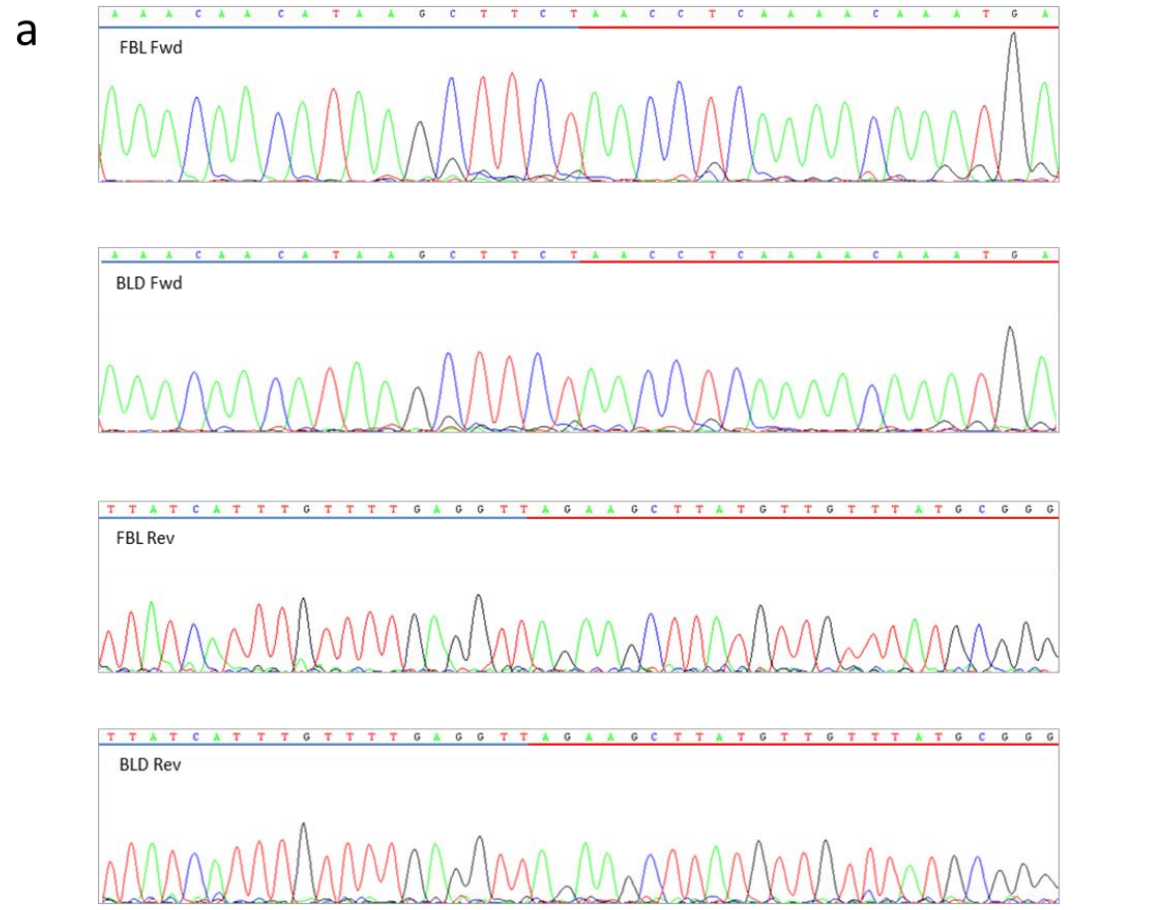


Figure 17 Verification of mtDNA deletion in patient 3 by Sanger sequencing. a) Bases underlined in blue indicate the mtDNA sequence upstream of the deleted region, while the bases underlined in read indicate the mtDNA sequence downstream of the deletion. All sequences are shown in the 5'-3' direction. **b)** Alignment of patient mtDNA sequence (blood and skin mtDNA) to the human reference sequence (GRCh38) in forward (top) and reverse (bottom) directions. The m.6211_8681del deletion is marked in red. Only the ends of the deletion are shown. FBL, fibroblasts; BLD, blood.

3.4.1 Relative quantification of heteroplasmy levels

Relative heteroplasmy levels of the verified mtDNA deletion were measured by qPCR using two different primer pairs; one flanking the deletion breakpoints (mDEL) and the other targeting a conserved mtDNA region outside the deletion (tRNA) (see 2.6.2: Table 14). Mean Ct values for the qPCR experiments are shown in Table 18. Melt curve analysis showed a single peak for all samples, indicating that both primer pairs produced a single, specific product (Figure 18). Relative heteroplasmy levels were calculated using the formula:

$$2^{([Ct(tRNA)\cdot E(tRNA)] - [Ct(mDEL)\cdot E(mDEL)])}$$

in which "Ct" is the mean Ct value for each primer pair and "E" is the PCR primer efficiency as determined by running a primer efficiency test, which was included to increase the accuracy of the calculation (see 2.6.2; Figure 19). Relative heteroplasmy levels of the mtDNA deletion in the patient's skin and blood were calculated using tRNA as reference and a PCR efficiency of 73.879% for tRNA and 73.056% for mDEL (Figure 19). Plotting these values into the formula, the relative heteroplasmy level of the mtDNA deletion was determined to be 4.6% in the patient's skin tissue and 2.9% in the patient's blood (Table 18).

Sample	Mean Ct values	
	tRNA	mDEL
P3 FBL	18,22	24,51
P3 Blood	19,16	26,39
Control	17,82	Undetermined

Sample	Relative quantification	
	tRNA	mDEL
P3 FBL	1	0.0459
P3 Blood	1	0.0286

Table 18 Mean Ct values and relative quantification of mtDNA deletion heteroplasmy levels in skin and blood. The tRNA primer pair was targeted to a conserved region outside the mtDNA deletion to detect all mtDNA molecules. The mDEL primers were design to flank the deletion breakpoints in order to detect only those mtDNA molecules containing the deletion. The Ct value for mDEL in the control sample could not be determined for any of the wells in the triplicate, indicating that the deletion was not present at all in the control mtDNA. P3, patient 3; FBL, fibroblasts.

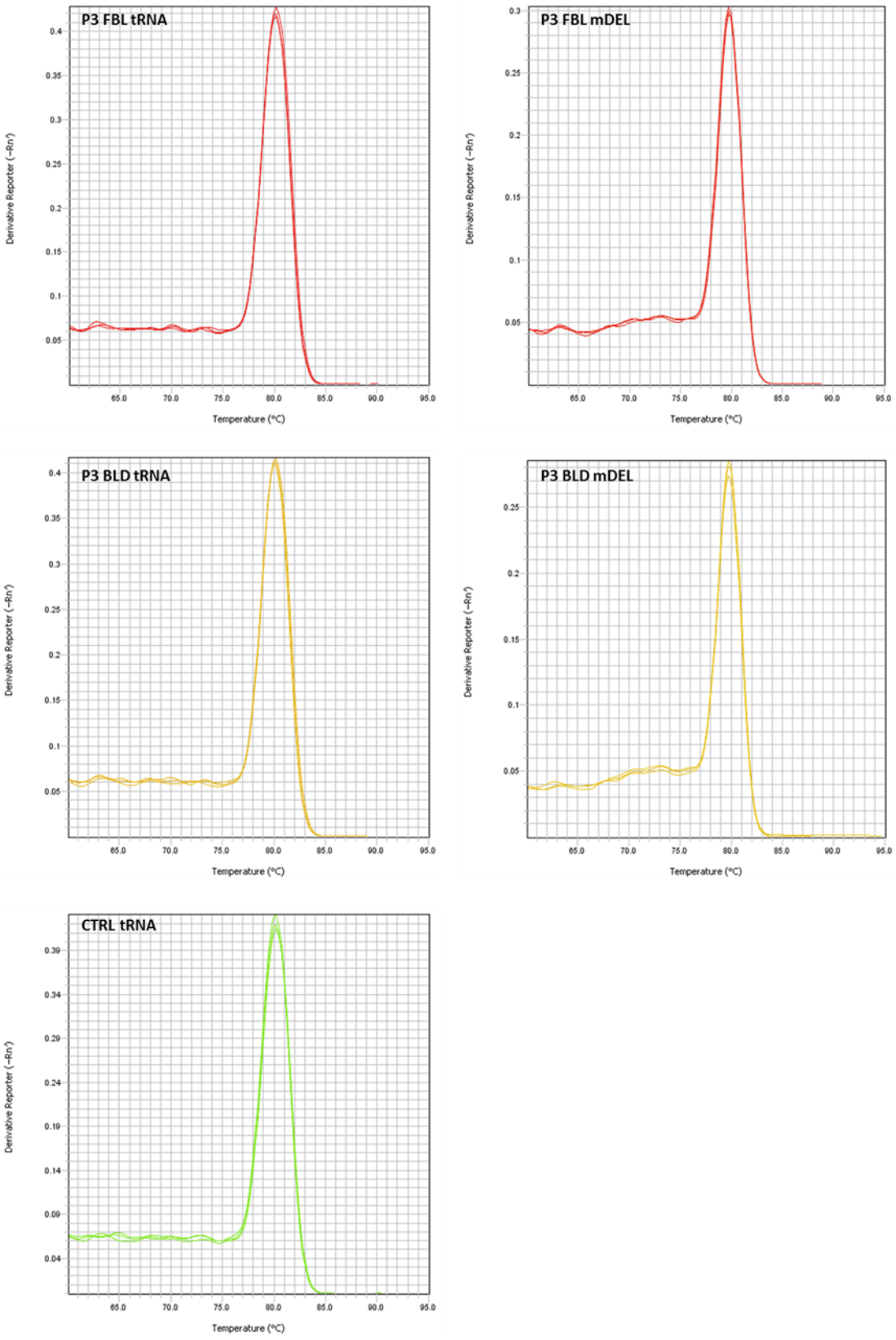
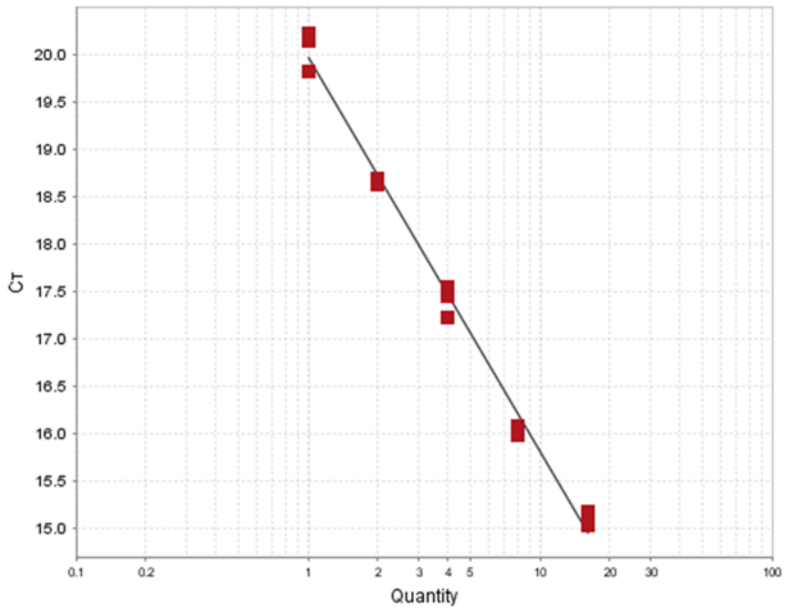
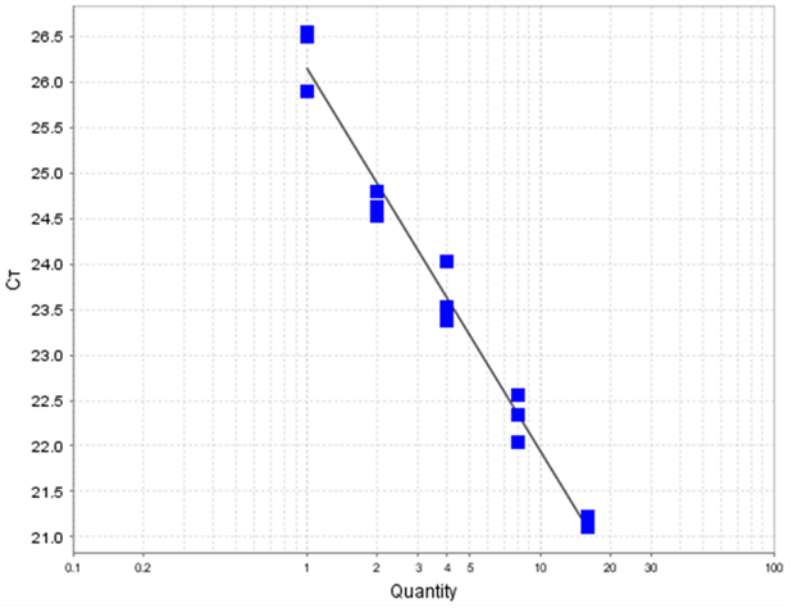


Figure 18 Melt curve analysis. The melting curves show that the qPCR run produced only one product in all samples, indicating that the primers were specific for the targeted region. P3, patient 3; BLD, blood tissue; FBL, skin fibroblasts.



Target: tRNA Slope: -4.162 R^2: 0.992 Effectivity (%): 73.879



Target: mDEL Slope: -4.198 R^2: 0.980 Effectivity (%): 73.056

Figure 19 Primer efficiency test. The efficiency of the tRNA and mDEL primers was evaluated by performing a 5-fold dilution series experiment. The slope of the standard curve indicates PCR efficiency, which is calculated by the formula $E = 10^{(-1/slope)}$. A slope value of -3.322 would give E = 2, which means exact doubling for each cycle and therefore 100% efficiency. The R^2 value (correlation coefficient) is a measure of the closeness of fit between the standard curve and the individual Ct values of the standard reaction. A value of 1.00 indicates a perfect fit. An R^2 value >0.99 is desirable.

3.5 mtDNA point mutations detected by MPS

Massively parallel sequencing of mtDNA from fibroblasts detected 11 single nucleotide variants (SNVs) in patient 2 (Table 19) and 43 single nucleotide variants in patient 3 (Supp. Table S3). Since the m.6211_8681del deletion detected in patient 3 most likely is the disease-causing variant, we did not analyze any further the mtDNA SNVs identified in this patient. Of the 11 genetic variants found in patient 2, only three affected protein coding regions, while the rest affected non-coding tRNAs, rRNAs, or the D-loop (Table 19). The depth of coverage achieved for these 11 positions using the non-PCR sample ranged from 26,000 to 60,000. With one exception, all the single nucleotide variants detected in this patient were present in a nearly homoplasmic state. The one exception was the D-loop mutation m.310T>C, which had a heteroplasmy level of 87%. Furthermore, approximately 65% of the mtDNA molecules had a T>TC insertion in this very same position (m.310T>TC).

Table 19 Mitochondrial DNA single nucleotide variants detected in patient 2 by massively parallel sequencing of fibroblast derived mtDNA. With the exception of the m.310T>C variant, all point mutations were present in an almost homoplasmic state as seen by the read count and percentage values. The m.310 position had two heteroplasmic variants: the m.310T>C variant, which was present in 12% of mtDNA molecules, and the m.310T>TC insertion variant, which was present in ~66% of mtDNA molecules.

					Nucleotide count					
Position	**Gene name**	**Nucleotide change**	**Amino acid change**	**Total count**	**A (%)**	**C (%)**	**G (%)**	**T (%)**	**Del**	**Ins**
263	MT-DLOOP	A>G	non-coding	43283	271 (0.63)	171 (0.40)	42649 (98.54)	192 (0.44)	1	1
310	MT-DLOOP	T>C	non-coding	26580	313 (1.18)	23059 (86.75)	36 (0.14)	3172 (11.93)	28	17416 (T>TC)
456	MT-DLOOP	C>T	non-coding	34915	166 (0.48)	362 (1.04)	295 (0.84)	34092 (97.64)	25	1
750	MT-RNR1	A>G	non-coding	58715	334 (0.57)	100 (0.17)	58073 (99.91)	208 (0.35)	8	2
1438	MT-RNR1	A>G	non-coding	55839	247 (0.44)	37 (0.07)	55426 (99.26)	129 (0.23)	4	0
4336	MT-TQ	T>C	non-coding	58019	176 (0.30)	57580 (99.24)	48 (0.08)	215 (0.37)	2	2
4769	MT-ND2	A>G	syn:M-M	52410	271 (0.52)	60 (0.11)	51816 (98.87)	263 (0.50)	4	1
5839	MT-TY	C>T	non-coding	59214	139 (0.23)	106 (0.18)	132 (0.22)	58837 (99.36)	2	0
8860	MT-ATP6	A>G	non-syn:T-A	48111	236 (0.49)	183 (0.38)	47495 (98.97)	197 (0.41)	4	0
15326	MT-CYB	A>G	non-syn:T-A	57310	266 (0.46)	135 (0.24)	56722 (98.97)	187 (0.33)	0	1
16304	MT-DLOOP	T>C	non-coding	60184	277 (0.46)	59589 (99.01)	52 (0.09)	266 (0.44)	8	2

4 Discussion

The great majority of respiratory chain subunits, translation machinery components, and almost all of the factors involved in mtDNA replication, transcription, translation, and repair are encoded by the nuclear genome [37]. In addition, nuclear DNA mutations account for more than 90% of all mitochondrial disorders in children [88]. For these reasons, analysis of the nuclear genome was the logical first step in the search for disease-causing variants in the three patients in this study.

4.1 Genetic variants identified by whole exome sequencing

Whole exome sequencing of DNA isolated from blood from the three patients and filtering of the called variants (see 2.1-2.2) resulted in the identification of 25 potentially pathogenic variants in patient 1 (Table 15), 19 variants in patient 2 (Supp. Table S1), and 21 variants in patient 3 (Supp. Table S2). To assess if any of these variants were likely to be causative of the clinical phenotypes of the patients, literature searches using OMIM and PubMed databases were carried out to identify potentially causal links between the genetic variants and the clinical phenotypes. Genes and genetic variants previously reported in these databases and implied in mitochondrial disease were specifically looked for. In addition, the mutation analysis software Alamut Visual 2.6.1 (Interactive Biosoftware) was used to assess the pathogenicity of the detected variants based on evolutionary conservation and molecular consequences of the mutations. Based on these assessments, the homozygous *SURF1* variant c.106+1G>C was concluded to be the best candidate for further analysis in patient 1. None of the putative variants identified in patients 2 and 3 were considered to be good candidates for further analysis.

4.2 *SURF1* variant causes Leigh syndrome

The c.106+1G>C variant, predicted by Alamut Visual 2.6.1 to cause skipping of exon 2, was verified in the patient by Sanger sequencing (see 2.3-2.3.3). The same homozygous variant has previously been reported in a Pakistani individual with Leigh syndrome (LS); however, the authors did not show the consequences of the mutation [57]. In the current work, skipping of exon 2 was demonstrated by PCR and Sanger sequencing (see 2.4-2.4.2).

The nuclear gene most frequently mutated in Leigh syndrome is *SURF1* [77]. To date, more than 70 mutations in *SURF1* have been reported in patients with LS [57]. The gene encodes the 300 amino acid mitochondrial protein SURF1, which is located in the mitochondrial inner membrane [89]. SURF1 contains a central part and two transmembrane domains, one in the N-terminus and the other in the C-terminus, that are thought to be essential for the protein function (Figure 20) [90]. Although SURF1 is known to be necessary for the assembly and maintenance of cytochrome c oxidase (COX; complex IV of the respiratory chain), its exact physiological function is still not fully understood [91]. Truncating mutations in the central part and both transmembrane domains result in an unstable protein that does not rescue COX activity in SURF1-deficient cell lines [89].

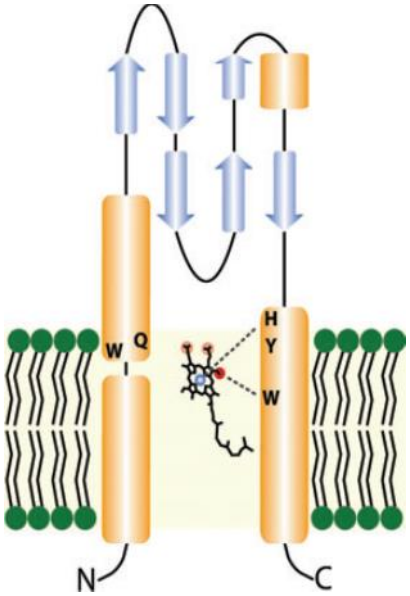


Figure 20 Schematic model of SURF1. The central part and the two transmembrane domains in the N- and C-termini of SURF1 are shown. SURF1 is believed to bind heme a (shown in middle) and incorporate it into COX subunit I as part of its role in COX assembly (from Hannappel et al. [91]).

The majority of the 74 different LS-associated *SURF1* mutations reported so far are splice site, nonsense, and frameshift mutations that may lead to nonsense-mediated decay (NMD) of the transcript (Supp. Table S4). Exon 2 of *SURF1* is 52bp, and skipping of this exon results in frameshift early in the transcript, which may render the remaining protein non-functional even if the transcript is not targeted for NMD. This, and the fact that *SURF1* c.106+1G>C has been reported in a patient with similar clinical manifestations as our patient, lead us to conclude that this mutation is likely the cause of disease in patient 1.

4.3 Mitochondrial DNA analysis by massively parallel sequencing

Because none of the genetic variants identified by whole exome sequencing in patients 2 and 3 were considered likely to be causative of the clinical phenotypes, the mitochondrial genomes of the two patients were analyzed by massively parallel sequencing. Fibroblasts from the patients were cultured and the mtDNA isolation protocol designed by Quispe-Tintaya *et al.* was used to prepare mtDNA enriched samples (see 2.5-2.5.3). A fraction of each mtDNA sample was then subjected to direct massively parallel sequencing without prior amplification by PCR, while another fraction of each sample was first amplified by 10 cycles of PCR using 9 overlapping primer pairs covering the whole mtDNA genome (see 2.5.5-2.5.6). Massively parallel sequencing resulted in deep coverage (10,000-130,000) of the entire mtDNA genome in all samples; however, the coverage was uniform only in the samples that were not PCR amplified prior to sequencing (see 3.3.2). In addition, massively parallel sequencing identified 11 single nucleotide variants in patient 2 and 43 variants in patient 3 (see 3.5).

4.3.1 Low-level heteroplasmy detection

Although the mitochondrial genome remains one of the most widely studied DNA segments in humans [43], the several unique features of mitochondrial genetics still pose many challenges for the complete and accurate characterization of mtDNA genomes by conventional sequencing technologies [92]. Three challenges are of particular significance: the presence of thousands of copies of mtDNA in each human cell; the need for mtDNA enrichment prior to sequencing; and the presence of mtDNA sequences in the human nuclear genome.

The first of these challenges is the presence of multiple mtDNA copies in each nucleated human cell, leaving open the possibility that all copies are not identical. Originally, heteroplasmy was believed to be quite rare in healthy individuals [93, 94]. However, recent work has shown that very low-level (>0.2%) heteroplasmic variance is universally present in healthy people [42]. Although heteroplasmic mtDNA mutations usually need to be present in at least 70-90% of mtDNA molecules in the relevant tissue to cause observable phenotypes [43], even mutations present at low levels in a female may result in an affected offspring because of the germ-line bottleneck in mtDNA segregation [95]. Therefore, sensitive

detection and accurate quantification of heteroplasmic pathogenic changes is of great importance in genetic counselling.

The small size of the human mitochondrial genome and the resulting high coverage for each nucleotide position generated by massively parallel sequencing allow for the detection of low-level mtDNA heteroplasmy. The mean coverage achieved in our study (>40,000) was much higher than what is typically seen in similar mtDNA sequencing studies [42, 43, 96-99] and should in principle allow us to detect very low heteroplasmy levels. However, the relatively high sequence error rate associated with massively parallel sequencing technologies can lead to false positives, meaning that sequencing errors may falsely be considered to be heteroplasmies [43]. There is therefore a need to establish a critical threshold by which background sequencing noise can be weeded out.

In the present work, massively parallel sequencing detected four alleles in each position in the mtDNA. Moreover, one allele was typically overrepresented, while the three other alleles were present in only ~0.5% of mtDNA molecules (Table 19). Since it is unlikely that every position should harbor three low-frequency alleles and one high-frequency allele, the low-frequency alleles most probably represent background noise generated by the sequencing process. This means that the critical threshold value for background noise in our experiments was most likely below 1%, meaning that all alleles with a frequency less than 1% can be discarded. This cut-off value is also in agreement with earlier studies [92, 97]. However, the exact detection threshold needs to be determined experimentally before any firm conclusion can be arrived at. One general method that could be applied to estimate the level of background noise experimentally is to use external control libraries with known DNA sequence [92, 99]. Comparing the known sequence with the called sequence could then be used to determine the critical detection threshold.

4.3.2 Advantages and disadvantages of PCR amplification

The second challenge associated with massively parallel sequencing of the mitochondrial genome is the need for mtDNA enrichment. Despite the presence of multiple mtDNA copies in each cell, mtDNA only comprises a small portion of the total cellular DNA, making it necessary to enrich the samples for mtDNA before sequencing. The most commonly used method for enrichment is PCR amplification [52]. Although this method is relatively cheap and efficient, it can produce sequencing errors in several ways. Firstly, since high number of

amplification cycles is usually needed for sufficient enrichment, PCR can lead to sequencing artifacts that may result in false positives [100]. Secondly, PCR may alter the proportion of different alleles as not all mtDNA molecules (or segments within a molecule) are amplified equally [100]. Lastly, there is the presence of so-called nuclear DNA sequences of mitochondrial origin (NUMTs; see below), which are a potential source of contamination when PCR is used to analyze mtDNA. Although this problem to a large extent can be solved by designing and validating a set of primers that prevents the amplification of nuclear mtDNA sequences and specifically amplifies mitochondrial DNA (see 2.5.5), finding the right set of primers can be a very labor-intensive task [51]. Moreover, the process of producing and validating a consensus compilation of human NUMTs is still ongoing, meaning that new NUMTs could be added to the list in the future and old NUMTs could be removed [101]. This makes it even more difficult, if not impossible, to completely prevent the amplification of NUMTs when PCR is used for mtDNA enrichment.

For these reasons, we enriched our samples for mtDNA using a reliable protocol involving only a conventional plasmid preparation method and purification with paramagnetic beads, or followed by 10 cycles of PCR amplification. As shown in Figure 13, when PCR was included this resulted in uneven coverage of the mitochondrial DNA with a pattern corresponding to the 9 pairs of overlapping primers used to cover the entire mtDNA genome. In contrast, the PCR free method resulted in uniform coverage throughout, which is important in order to detect and determine the size of large-scale mtDNA deletions. The coverage achieved by direct sequencing thus indicates that the PCR free method is the preferred approach when analyzing mitochondrial DNA by massively parallel sequencing.

4.3.3 The presence of nuclear DNA sequences of mitochondrial origin (NUMTs)

The third challenge associated with mtDNA analysis by massively parallel sequencing is the presence of mitochondrial DNA sequences in the nuclear genome (NUMTs), which are the result of transfer of DNA segments from mitochondria to the nucleus. Although the full details of how mitochondrial DNA actually enters the nucleus are not understood, several mechanisms have been proposed (Figure 21). To date, approximately 200 NUMTs in the human nuclear genome are known [101]. These are evenly distributed within and among chromosomes, and up to 90% (14,654bp) of the human mtDNA can be found at a single

nuclear locus [102]. However, only about a third of all NUMTs present in the nuclear DNA are due to insertions of mitochondrial sequences, the rest being originated as duplications of pre-existing NUMTs [103]. Moreover, a subset of NUMTs is highly rearranged, including sequences derived from different regions of the mitochondrial genome that have undergone inversions, deletions, and duplications [51]. The sequence similarity between NUMTs and their corresponding mitochondrial sequences is therefore highly variable, and not a single known NUMT is identical to its mitochondrial counterpart [101]. In addition, only a few human NUMTs have so far been validated and verified by both *in silico* and *in vitro* approaches, meaning that the existence of only some NUMTs is confirmed. The existence of the rest is more or less probable as determined by various bioinformatics approaches [101].

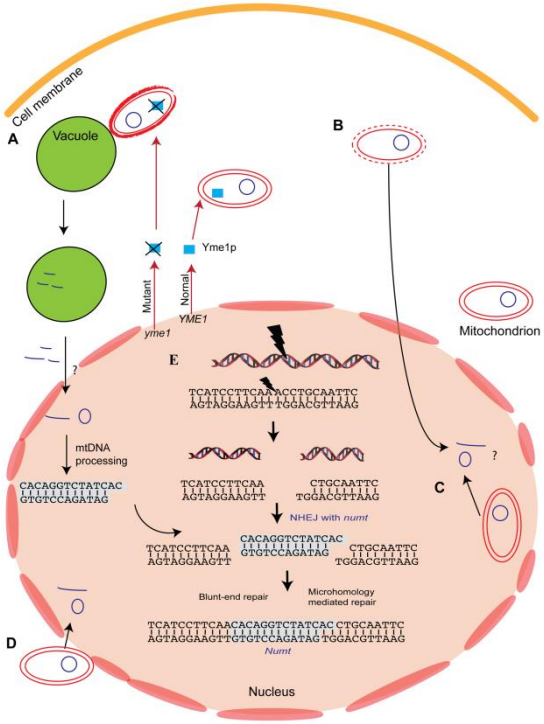


Figure 21 Proposed mechanisms of NUMT entrance into the nucleus. For NUMTs to persist in the nuclear genome, mtDNA must first physically reach the nucleus and then integrate into the nuclear chromosome, with intragenomic processes of amplification, mutation, or deletion following. Work so far has focused on the escape of DNA from the mitochondria and on the integration of mtDNA within the nucleus but not on its physical entrance into the nucleus. Although the full details of this process are not understood, a few different pathways have been proposed (not all details in the figure are described here): A) Abnormal mitochondria are taken up for degradation by the vacuole (in plant and fungal cells). B) Lysis of mitochondrial compartment. C) Encapsulation of mitochondrial DNA inside the nucleus. D) Direct physical association between the mitochondria and the nucleus and membrane fusion. E) After mtDNA has entered the nucleus, it can integrate into nuclear chromosomes during the repair of double-strand breaks in a mechanism known as non-homologous end-joining (NHEJ) (adapted from [102]).

These facts strongly suggest that NUMTs constitute a real source of contamination when PCR is used to study mtDNA. Even if primers are carefully designed to prevent co-amplification of nuclear and mitochondrial DNA, we cannot be sure whether some contamination has occurred or not since the full list of validated NUMTs has yet to be defined. Therefore, by avoiding PCR amplification altogether as we did in this study, we also avoided one important source of contamination associated with the presence of NUMTs. When using massively parallel sequencing to analyze the mtDNA genome, however, NUMTs constitute still another potential source of error as there is the possibility of mtDNA reads being aligned to the nuclear genome instead of the mitochondrial DNA target. However, our results demonstrate that this possibility is negligible. Figure 22 shows the alignment data of three of the NUMTs predicted to be real based on various *in silico* methods [101]. In all three cases, only a few reads, at most 6 at a given position, aligned to the nuclear DNA, while the rest of the reads (>40,000 per nucleotide position) aligned to the mtDNA genome (compare with Figure 13). These results indicate that the presence of mtDNA sequences in the nuclear DNA does not constitute a problem even when the whole human genome is used as reference during alignment.

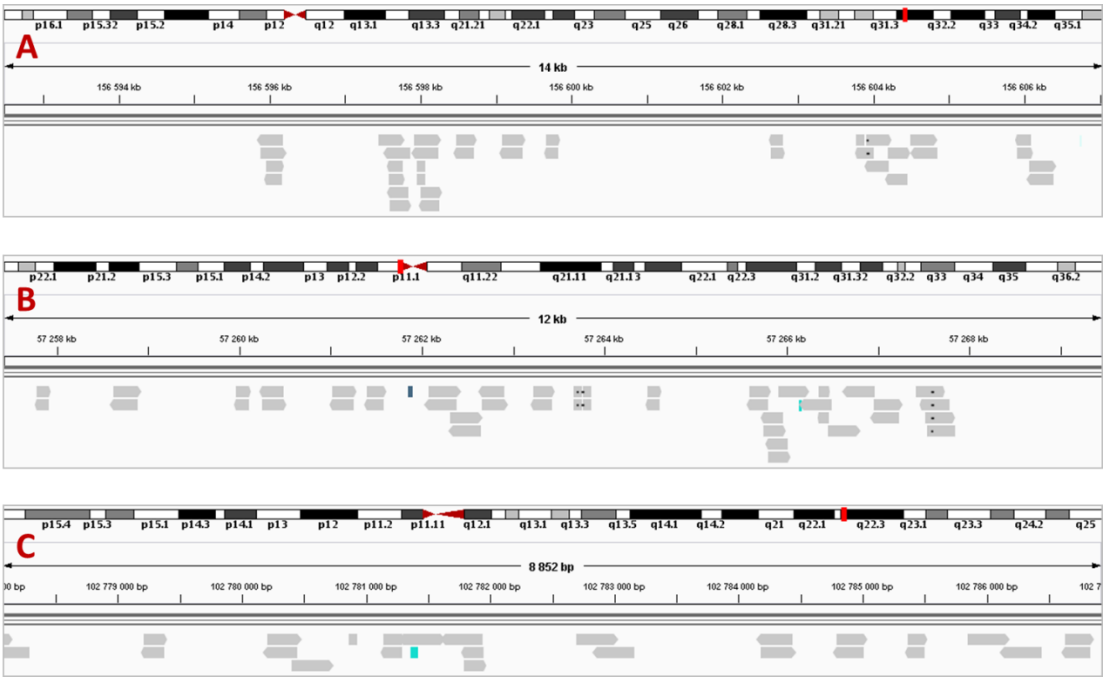


Figure 22 Alignment data for three human NUMTs. The figure shows alignment data for three human NUMTs predicted to be real, including the long 14,654bp NUMT. Somewhat surprisingly, only a few reads aligned to the three nuclear DNA loci. In comparison, more than 40,000 reads on average per base position aligned to the mitochondrial sequences homologues to these NUMTs. The data shown are from the PCR free sample from patient 3. A: m.672-15325/chr4:156592474-156607061; B: m3819-15924/chr7:57257414-57269467; C: m.724-9666/chr11:102778067-102786933. Nucleotide positions refer to the hg18/GRCh38 genome assembly.

The reason why so few reads aligned to the nuclear DNA loci may in part be that all NUMTs have an incomplete sequence similarity with their mitochondrial homologues. This makes it more likely that a read will align to the mtDNA genome than to the nuclear DNA since the alignment tool only allows for a limited amount of error. Moreover, the use of paired-end sequencing further increases the likelihood that a read will align to the mitochondrial genome. For a paired-end read to align to the nuclear genome, both ends of the read must fall within the nuclear sequence, making it less likely that the alignment will occur. This is especially true for NUMTs shorter than 500bp, as the DNA was sheared with a sonicator to produce fragments with an average size of 500bp during library preparation. In any case, our results show that the potential problems associated with PCR and NUMTs can be overcome while still achieving an ultra-deep and uniform coverage of the entire mtDNA genome, making it possible to accurately detect and quantify low-level heteroplasmic mutations.

4.4 Assessment of mtDNA variants identified in patient 2

Current molecular diagnostic methods for the detection and quantification of mitochondrial DNA mutations are based on the screening for common point mutations followed by quantification of the mutation load. If the screening is negative, the entire mtDNA genome is sequenced to find causative rare variants or novel mutations [99, 104]. In the current work, however, these steps were performed simultaneously, resulting in the identification of 11 mtDNA single nucleotide variants (SNVs) in patient 2, all of which were nearly homoplasmic (Table 19). Predicting the pathogenicity of mtDNA SNVs is a challenging task due to the complicated nature of mitochondrial genetics combined with the genetic heterogeneity of mitochondrial diseases. Thus, very useful pathogenicity criteria have been proposed for the identification of disease-causing mtDNA mutations [105, 106].

One of these criteria is whether a variant has previously been reported to cause a similar or identical phenotype in other individuals. Of the 11 mtDNA variants detected in patient 2, only two are reported in the comprehensive human mitochondrial genome database MITOMAP (www.mitomap.org), which so far includes almost 600 different mtDNA point mutations implicated in disease. The first of the two mutations, the D-loop insertion m.310T>TC, has been proposed to be involved in malignant melanoma etiology and pathogenesis [107]; while the second variant, the $tRNA^{Gln}$ point mutation m.4336T>C, has been implicated in several disorders, primarily Alzheimer's and Parkinson's disease [108]. None of these disorders,

however, are directly relevant for the clinical phenotype of our patient. The remaining 9 single nucleotide variants, including the 3 protein-coding variants m.4769A>G, m.8860A>G, and m.15326A>G, are all reported in MITOMAP as mtDNA polymorphisms, and are therefore unlikely to be the cause of the patient's phenotype. This conclusion is also consistent with the observation that homoplasmic mtDNA point mutations tend to be less pathogenic than heteroplasmic mutations due to purifying selection acting on deleterious heteroplasmies to keep them at low frequencies [109].

Considering this patient's clinical picture, the most prominent feature of which is the signs and symptoms of partial PEO, the genetic cause is most probably a large-scale mtDNA deletion. Although a firm diagnosis is yet to be established, PEO is almost invariably associated with mtDNA deletions confined to skeletal muscle [66]. Even if the signs of PEO are only the first manifestation of a developing KSS, it is expected that the genetic defect responsible for the patient's phenotype is a large-scale mtDNA deletion, which is more easily detected in muscle tissue than in fibroblasts, explaining why we did not detect any deletion by massively parallel sequencing of fibroblast-derived DNA. In any case, a molecular diagnosis for patient 2 is likely to be established by a genetic analysis of skeletal muscle tissue.

4.5 Detection of mtDNA deletion in patient 3

DNA isolated from muscle tissue from patient 3 was analyzed at the Nijmegen Center for Mitochondrial Disorders (NCMD), revealing the nearly homoplasmic deletion m.6211_8681del. This mtDNA deletion is most probably the cause of the Kearns-Sayre syndrome like phenotype in the patient (see 1.7). Approximately 90% of KSS patients have a large-scale (1-10 kb) mtDNA deletion present in most tissues, although it is often detectable in muscle only [66]. The high level of mutant mtDNA molecules present in patient 3 may explain the ataxia, ptosis, and especially ophthalmoplegia. Extraocular muscles have a high energy requirement, and mitochondria make up approximately 60% of the cell volume in eye muscles [67]. To date, 133 different deletions with known breakpoints have been reported in the MITOMAP database. The ~2.5 kb deletion m.6211_8681del identified in our patient was not among the reported deletions. Therefore, we here present another case of KSS caused by a novel deletion present in an almost homoplasmic state in the patient's muscle tissue.

The mtDNA deletion m.6211_8681del was also identified in blood leukocytes and skin fibroblasts from the patient, and it was estimated to be present in 2.9% and 4.6% of mtDNA genomes in these tissues, respectively (see 3.4-3.4.1). As shown in Figure 13, sequencing of the PCR-free DNA sample from patient 3 fibroblasts resulted in uniform coverage across the entire mtDNA genome, with no visible reduction of coverage between the deletion junctions (6211-8681 kb). The low-level deletion heteroplasmy of 4.6% was therefore too low to be detected by the deep-sequencing method used in this study.

4.6 Are fibroblasts suitable for mtDNA analysis by massively parallel sequencing?

It is well known that mitochondrial disorders exhibit clinical and biochemical tissue-specificity; however, the underlying genetic defects are generally less specific, as the same mtDNA defect can be detected in a variety of tissues [110]. While pathogenic mtDNA point mutations are often detected in blood leukocytes and, in some cases, also in urinary epithelial cells, hair follicles, buccal mucosa, and skin fibroblasts, large-scale deletions are more likely found in muscle [110].

The three main syndromes observed in patients with mtDNA deletions are KSS, chronic progressive external ophthalmoplegia (CPEO), and Pearson's syndrome (PS), the latter being a severe systemic illness of childhood associated with sideroblastic anemia (see 1.6.1; Footnote 3). Whereas in CPEO, the mtDNA deletion is found exclusively in muscle, in both KSS and PS the deletion is usually present in many tissues and can be detected in both blood leucocytes and skin fibroblasts [66, 111]. However, the occurrence of heteroplasmy in disorders caused by mtDNA deletions can result in varying tissue distribution of mitochondrial DNA molecules containing the deletion. Our results showed that the percentage of deleted mtDNA in patient 3 was much lower in fibroblasts and leucocytes than in muscle cells (see 3.4-3.4.1). A similar tissue distribution of mtDNA deletions have also been observed in other KSS patients [112]. The reason for this pattern is unclear, but it has been attributed to the frequent cell division of leucocytes and skin cells, allowing selection against the survival of cells with deletions in mitochondrial DNAs [112]. It appears, however, that this selection is not always very efficient since mtDNA deletions in PS are usually more abundant in blood than in other tissues, making the molecular diagnosis of PS reliably made by genetic analysis of leucocyte DNA [66].

If possible, the most easily accessed symptomatic tissue should be tested. However, if the clinical or family history is suggestive of an mtDNA point mutation, various tissues can be analyzed to find the mutation. This is generally also the case when an mtDNA deletion is suspected, unless the clinical picture is indicative of either PS or CPEO, in which case blood leucocytes and skeletal muscle should be tested respectively. In any case, one should always bear in mind that a normal result does not eliminate the possibility of a heteroplasmic mitochondrial DNA mutation being present, either in the tested tissue or in another tissue.

4.6.1 Advantages and disadvantages of using skin fibroblasts for mtDNA analysis

In the present work, cultured human fibroblasts were used to analyze mitochondrial DNA by PCR free mtDNA enrichment and ultra-deep massively parallel sequencing. An obvious advantage of skin fibroblasts is the ease with which they are cultured and the resulting high number of cells that can be obtained. Due to the small size of the human mitochondrial genome, having many cells available is an important factor for the complete and accurate analysis of mtDNA, especially when PCR amplification is excluded from the protocol. However, care must be exercised when culturing fibroblast for study, because cells with electron transport chain deficiencies are fragile. They may die in culture, leaving only the healthy cells to grow [113]. This may in part explain why approximately half of all children with deficits in their muscles have normal enzyme activity in cultured skin fibroblasts [114]. In addition, mtDNA deletions tend to be present at low levels of heteroplasmy in fibroblasts, making their detection very challenging. But despite these limitations, cultured skin fibroblasts remain the choice of preference in some situations because of the ease with which they are obtained and handled [115]. Furthermore, adding pyruvate and uridine to the culture medium has previously been indicated to allow cells with respiratory chain deficiencies to survive, thus preventing false-negative results [113].

4.6.2 Analyzing mtDNA from blood leukocytes and skeletal muscle

When investigating patients with a suspected mitochondrial disorder, peripheral blood is in a sense the ideal tissue to analyze because it is obtained in a non-invasive manner. As was pointed out, blood leucocytes are the best choice when dealing with cases of Pearson's syndrome, but they can also be useful in detecting mtDNA point mutations and low-level

heteroplasmic deletions in general. Moreover, the relatively high concentration of white blood cells in normal human blood further suggests that they are suitable for mtDNA studies. In the mtDNA enrichment protocol applied in our work, approximately 10 million fibroblast cells were used per extraction (see 2.5.1). The normal leukocyte count in adults is 4,000-11,000 cells per microliter of blood [116], meaning that a 10 mL blood sample gives sufficient amount of mtDNA for direct sequencing.

Despite the invasive nature of muscle biopsies, skeletal muscle remains the tissue of choice for genetic analyses of mitochondrial DNA. Because of the great energy requirement of muscle cells, they are highly vulnerable to respiratory chain deficiencies. Consequently, many of the signs and symptoms most often seen in patients with mitochondrial disorders are directly related to muscle [110]. In addition, the relative abundance of mitochondrial DNA copies in muscle cells and their retention of mtDNA mutations over time, mean that defects in the mitochondrial genome are generally easier to detect in muscle than in other tissues. For these reasons, it is desirable to apply the two-step method of Quispe-Tintaya *et al.* [52] to human muscle tissue to prepare samples sufficiently enriched for mtDNA without the use of PCR amplification. This method, however, involves the application of a conventional plasmid extraction protocol, which is based on the lysis of resuspended cells in solution, and is consequently not directly applicable to intact tissues (see 2.5). It is therefore necessary to dissociate the muscle tissue into single cells before using the plasmid isolation protocol for mtDNA extraction. One way to dissociate muscle tissue is by enzymatic digestion [117, 118].

4.7 Conclusion

The main aim of this thesis was to find the genetic causes of the clinical phenotypes of three patients with suspected mitochondrial disease. Whole exome sequencing (WES) resulted in the detection of the homozygous *SURF1* mutation c.106+1G>C, which was verified by Sanger sequencing and concluded to be the cause of disease in patient 1. However, no mutations were found to be clinically relevant in the two other patients. Therefore, as a sub aim of this thesis, a direct deep-sequencing method was used to analyze the whole mitochondrial genome of the two patients aiming to identify disease-causing variants. Using skin fibroblasts as proof of concept, mtDNA enriched samples were prepared both with and without the involvement of PCR amplification. Massively parallel sequencing on the Illumina MiSeq platform resulted in a coverage of 10,000-130,000 in all samples, but uniform

coverage of the entire mitochondrial genome was only achieved when PCR amplification was excluded from the protocol. Because of this, and the fact that PCR is a potential source of several sequencing errors, we concluded that the PCR free protocol is the method of choice when mtDNA is analyzed by massively parallel sequencing. We also suggest that the same protocol could be used on blood leucocytes, both because blood is obtained in a non-invasive manner and because a blood sample of normal size contains enough leukocytes for direct sequencing.

Analysis of a muscle sample from patient 3 was performed at the Nijmegen Center for Mitochondrial Disorders (NCMD), revealing the nearly homoplasmic mtDNA deletion m.6211_8681del. This deletion, which was concluded to be the cause of the Kearns-Sayre syndrome like phenotype in the patient, was also a novel finding not previously reported in the comprehensive MITOMAP database. In addition, analysis of mtDNA from the patient's blood leukocytes and skin fibroblasts showed that this deletion was present in 2.9% and 4.6% of mtDNA genomes in these cells, respectively. However, the low-level deletion heteroplasmy of 4.6% was too low to be detected by the deep-sequencing method used in the present study. In general, massively parallel sequencing was found to be highly sensitive for the detection of low-level (>1%) heteroplasmic single nucleotide variants, but less sensitive for large-scale (~2.5kb) heteroplasmic deletions.

In summary, the genetic defect underlying the clinical phenotype was identified in two of the three patients included in this study: the nuclear DNA mutation *SURF1* c.106+1G>C in patient 1, and the mitochondrial DNA deletion m.6211_8681del in patient 3. The genetic defect responsible for the clinical picture of patient 2 remains unknown.

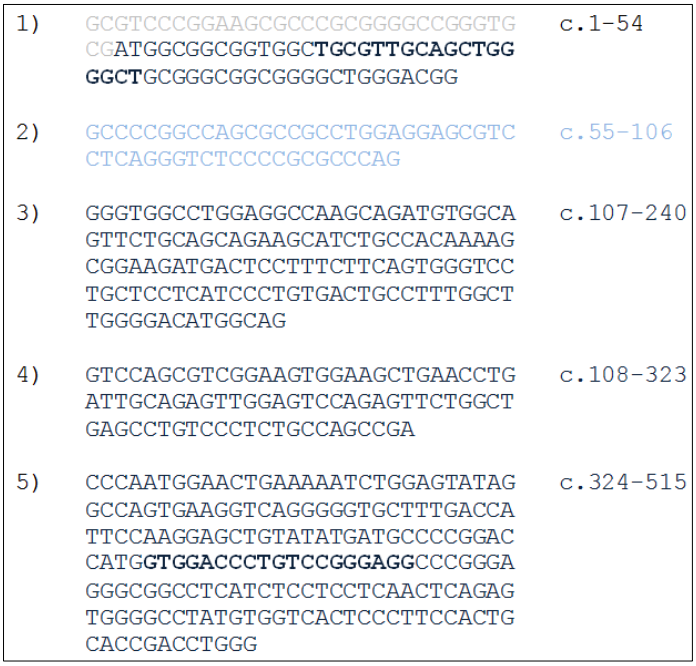
Supplementary data

Supp. Table S1 Putative pathogenic variants in patient 2 detected by whole exome sequencing and retained after filtering. None of the variants found were considered to be disease-causing.

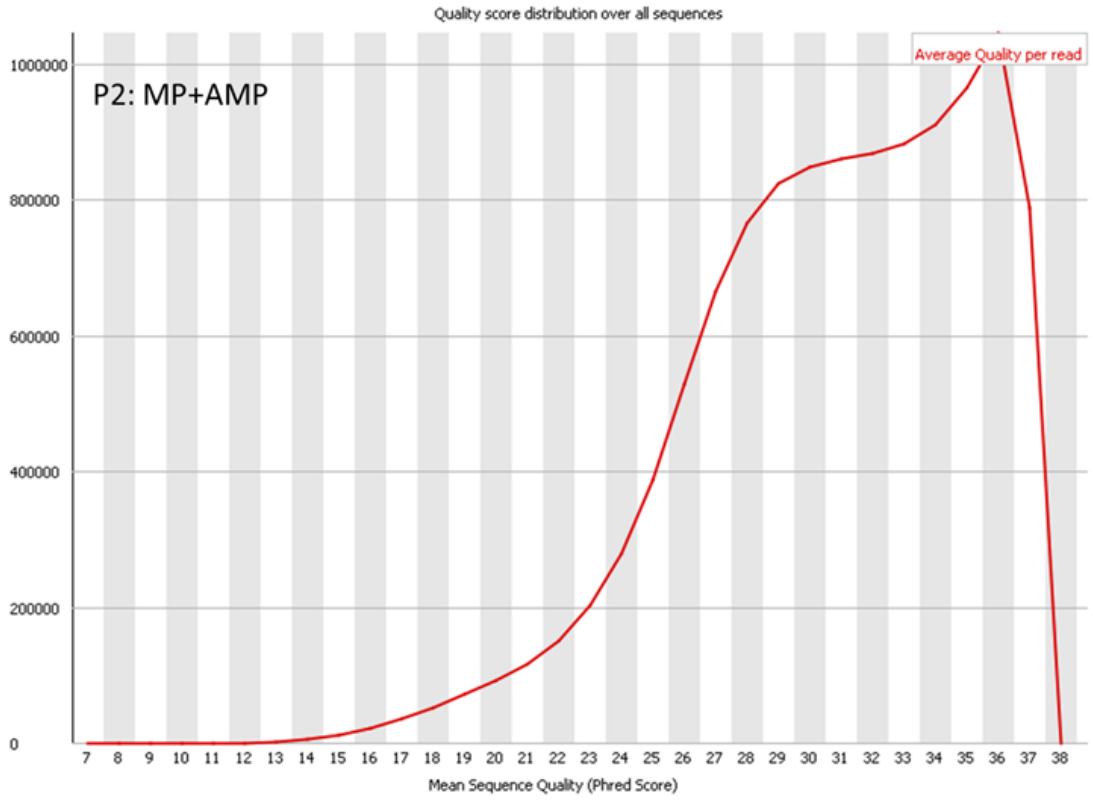
Gene name	Chr:pos	Ref	Alt	Amino acid change	Effect	Depth
MST1L	1:17083807	T	C	T664A	Non-synonymous	43
MST1L	1:17086870	T	C	S152G	Non-synonymous	28
PDE4DIP	1:144852477	G	A	P2235L	Non-synonymous	1
NBPF10	1:145296478	G	T	D134Y	Non-synonymous	7
HRNR	1:152187562	A	C	H2181Q	Non-synonymous	79
FCGR2C	1:161569419	C	G		Splice site substitution	78
OR6B2	2:240969483	G	A	R122C	Non-synonymous	58
POLR2J3	7:102180473	C	T	R184H	Non-synonymous	19
YWHAZ	8:101965496	C	A		Splice site substitution	2
CNTNAP3	9:39118196	C	A	G620V	Non-synonymous	16
ORM2	9:117092701	C	A	P173T	Non-synonymous	60
SFTPA2	10:81319267	C	T		Splice site substitution	17
FAM22D	10:89124914	T	A	L40Q	Non-synonymous	5
OR8B2	11:124252750	G	A	L164F	Non-synonymous	87
KIR3DL1	19:55329851	A	G	Y29C	Non-synonymous	44
PLA2G6	22:38519143	C	A	G378V	Non-synonymous	2
A4GALT	22:43091638	C	A		Splice site substitution	2
TEX11	X:69828956	A	G	W312R	Non-synonymous	115
NXF3	X:102334690	C	T		Splice site substitution	18

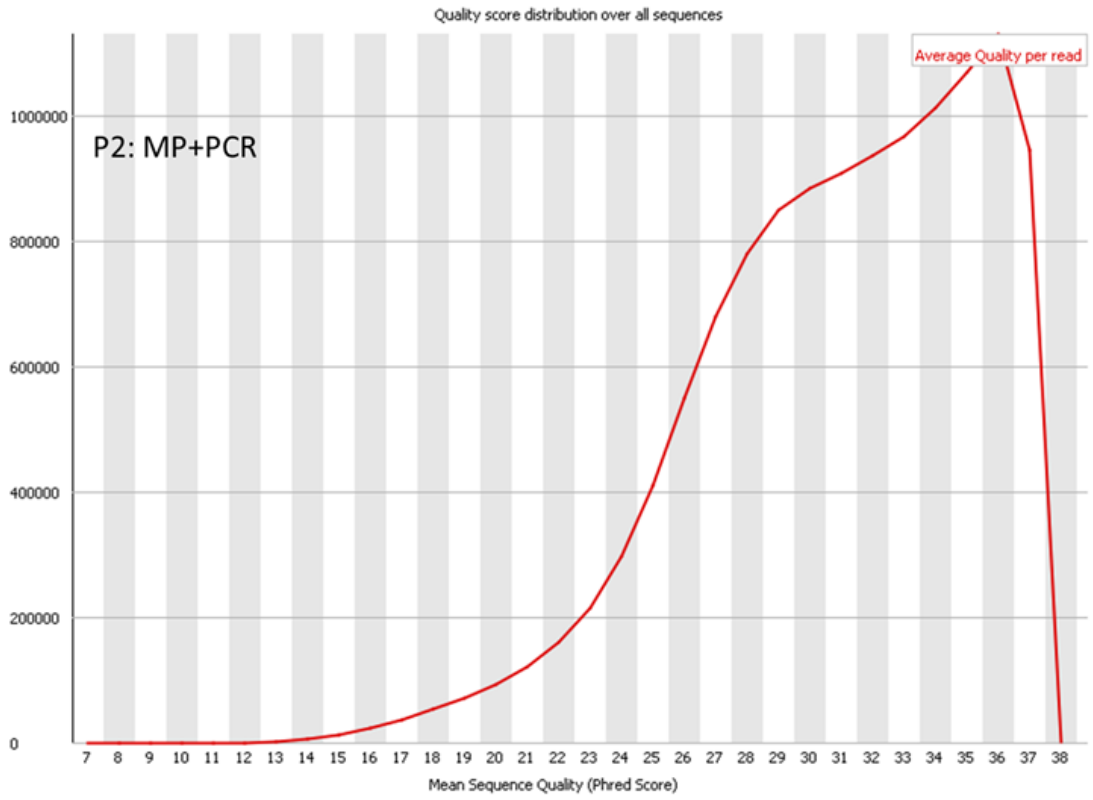
Supp. Table S2 Putative pathogenic variants in patient 3 detected by whole exome sequencing and retained after filtering. None of the variants found were considered to be disease-causing.

Gene name	Chr:pos	Ref	Alt	Amino acid change	Effect	Depth
RTKN	2:74657626	C	T	M118I	Non-synonymous	127
PCMTD1	8:52811580	G	A	P31L	Non-synonymous	88
GJB2	13:20763113	A	G	I203T	Non-synonymous	128
MST1	3:49725242	C	A	Q47H	Non-synonymous	82
SFTPA2	10:81319267	C	T		Splice site substitution	40
OR2T10	1:248756470	G	C	Y200*	Stop gained	127
CYP4A22	1:47608983	G	T	V185F	Non-synonymous	130
MST1P9	1:17083807	T	C	T664A	Non-synonymous	71
C7orf28B	7:6859453	C	G	E214D	Non-synonymous	261
HEG1	3:124774665	G	GCAT	L23LW	Codon change plus deletion	28
PSORS1C1	6:31106499	G	A	R37H	Non-synonymous	217
HLA-DRB1	6:32551996	G	T	A87D	Non-synonymous	32
FCGR2C	1:161565381	A	G		Splice site substitution	146
FCGR2C	1:161569419	C	G		Splice site substitution	193
C14orf57	14:72457365	T	G	E93A	Non-synonymous	127
ZNF589	3:48289146	G	A	W25*	Stop gained	114
CACNA2D3	3:54676222	A	G	D279G	Non-synonymous	127
C10orf93	10:134663933	A	G	W1923R	Non-synonymous	127
IP6K2	3:48726170	T	G	N273H	Non-synonymous	127
PTPRS	19:5223204	G	A	R845C	Non-synonymous	127
PPIP5K1	15:43973324	T	C	Q292R	Non-synonymous	159

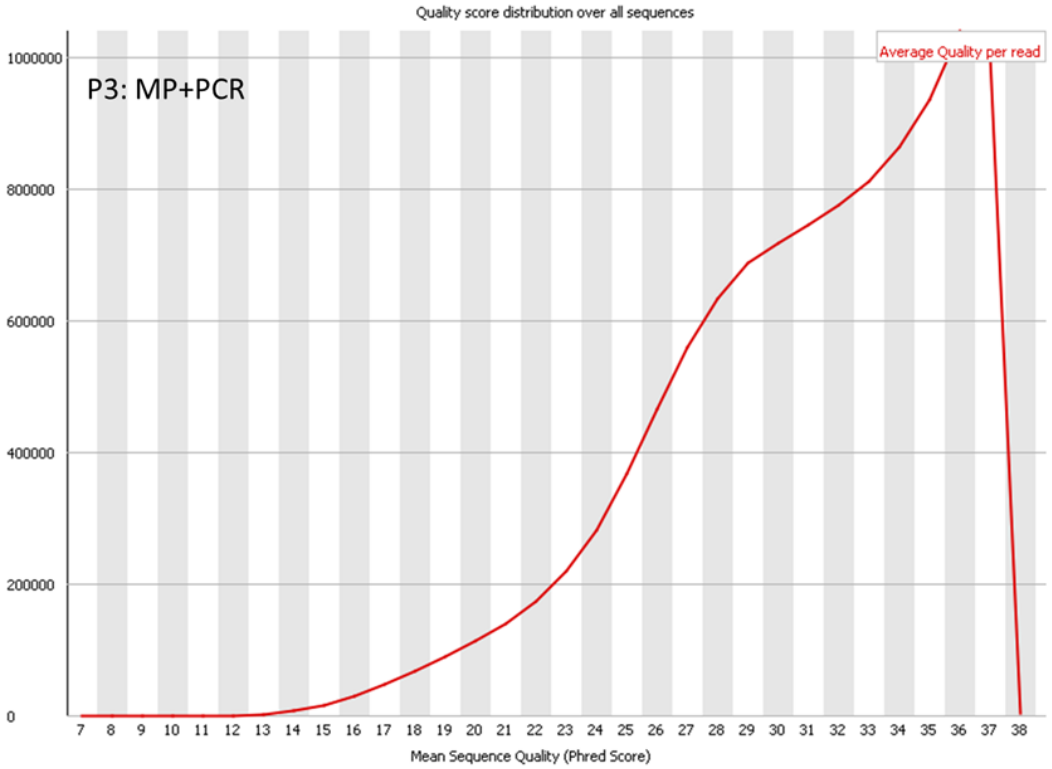


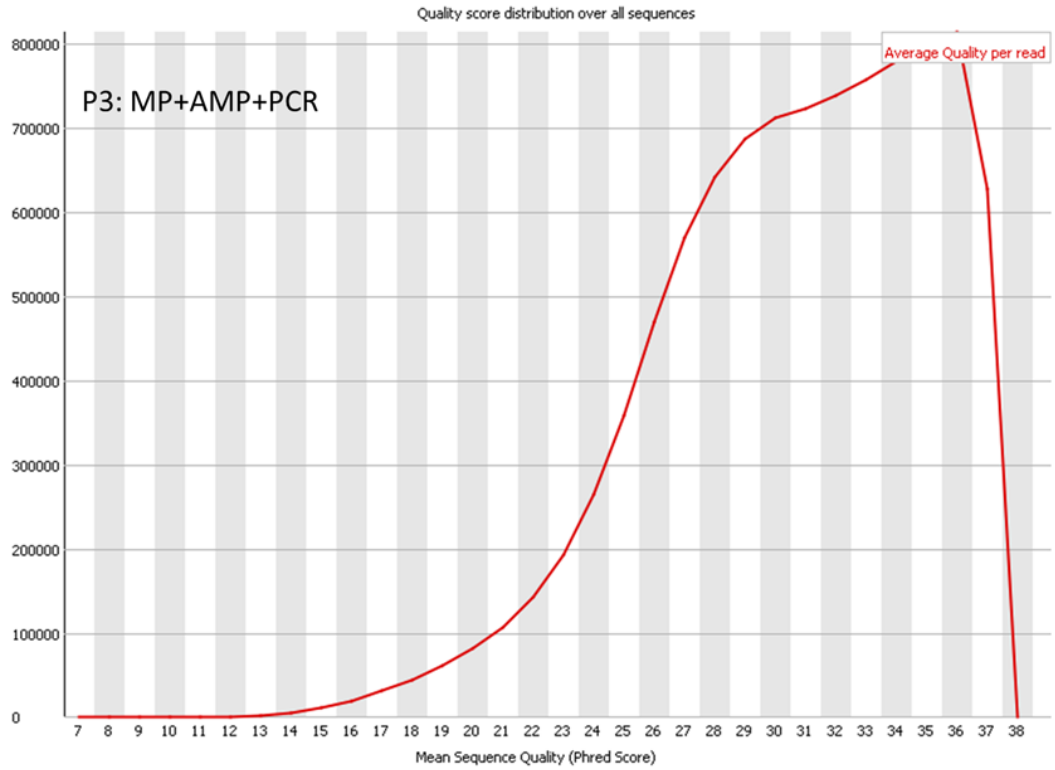
Supp. Figure S1 Exon 1-5 of *SURF1*. Nucleotides in bold type indicate the target sequences of the forward and reverse cDNA primers in exon 1 and 5, respectively. They gray sequence in exon 1 is the 5'-UTR. The amplicon size is 422bp with exon 2 (light blue) and 370bp without exon 2. The sequence depicted is the forward strand shown in 5'-3' direction.





Supp. Figure S2 Distribution of average quality score per sequence for samples from patient 2. Note that there is no significant difference in average quality score between PCR and non-PCR samples. P2, patient 2; MP, miniprep; AMP, AMPure; PCR, 10 cycles of PCR amplification.





Supp. Figure S3 Distribution of average quality score per sequence for samples from patient 3. The non-PCR sample results from patient 3 are shown in Figure 15. P3, patient 3; MP, miniprep; AMP, AMPure; PCR, 10 cycles of PCR amplification.

Supp. Table S3 Mitochondrial DNA point mutations in patient 3 detected by massively parallel sequencing.

Gene name	Position	Nucleotide change	Amino acid change	Effect
MT-DLOOP	73	A>G		non-coding
MT-DLOOP	185	G>A		non-coding
MT-DLOOP	189	A>G		non-coding
MT-DLOOP	194	C>T		non-coding
MT-DLOOP	195	T>C		non-coding
MT-DLOOP	204	T>C		non-coding
MT-DLOOP	207	G>A		non-coding
MT-DLOOP	263	A>G		non-coding
MT-DLOOP	302	A>AC; A>ACC		non-coding
MT-DLOOP	310	T>TC		non-coding
MT-RNR1	709	G>A		non-coding
MT-RNR1	750	A>G		non-coding
MT-RNR1	1243	T>C		non-coding
MT-RNR1	1438	A>G		non-coding
MT-RNR2	2706	A>G		non-coding
MT-ND1	3505	A>G	Thr>Ala	non-synonymous
MT-ND1	3834	G>A	Leu>Leu	synonymous
MT-ND1	4093	A>G	Thr>Ala	non-synonymous
MT-ND2	4769	A>G	Met>Met	synonymous
MT-ND2	5046	G>A	Val>Ile	non-synonymous
MT-ND2	5460	G>A	Ala>Thr	non-synonymous
MT-TC	5824	G>A		non-coding
MT-CO1	7028	C>T	Ala>Ala	synonymous
MT-CO2	8251	G>A	Gly>Gly	synonymous
MT-ATP6	8614	T>C	Leu>Leu	synonymous
MT-ATP6	8860	A>G	Thr>Ala	non-synonymous
MT-ATP6	8994	G>A	Leu>Leu	synonymous
MT-CO3	9591	G>A	Val>Ile	non-synonymous
MT-CO3	9920	C>T	Ala>Ala	synonymous
MT-ND4	11674	C>T	Thr>Thr	synonymous
MT-ND4	11719	G>A	Gly>Gly	synonymous
MT-ND4	11947	A>G	Thr>Thr	synonymous
MT-ND5	12414	T>C	Pro>Pro	synonymous
MT-ND5	12705	C>T	Ile>Ile	synonymous
MT-CYB	14766	C>T	Thr>Ile	non-synonymous
MT-CYB	15326	A>G	Thr>Ala	non-synonymous
MT-CYB	15884	G>C	Ala>Pro	non-synonymous
MT-DLOOP	16192	C>T		non-coding
MT-DLOOP	16223	C>T		non-coding
MT-DLOOP	16292	C>T		non-coding
MT-DLOOP	16325	T>C		non-coding
MT-DLOOP	16360	C>T		non-coding
MT-DLOOP	16519	T>C		non-coding

Supp. Table S4 Summary of the 74 different *SURF1* mutations that have been reported to date. 57 of the 74 mutations are frameshift, splice site, and nonsense mutations that may result in nonsense-mediated decay of the transcript. The homozygous splice site mutation c.106+1G>C (marked in red) detected in our patient has previously been reported once in a Pakistani individual with Leigh syndrome. References to all the mutations listed below can be found in Lee *et al.* [66]. Homo, homozygous; Het, heterozygous; NR, not recorder in reference.

Mutations (n=74)	Consequence	Number of alleles (n=248)	Ethnicity
Frameshift (n=30)			
c.37_38ins17	p.V13AfsX65	2 (1 Homo)	1 European
c.39delG	p.A14RfsX57	1 (1 Het)	1 Slovenian
c.169delG	p.E57KfsX15	1 (1 Het)	1 Hispanic
c.183_186delTCTT	p.L62SfsX9	1 (1 Het)	1 Caucasian/ native American
c.312_321del10insAT	p.L105X	45 (10 Homo, 25 Het)	11 European
			6 Slovenian
			3 Denmark
			10 Caucasian
			3 NR
			1 Sweden
			1 Caucasian/ Native American
c.367_368delAG	p.R123GfsX4	2 (1 Homo)	1 Japanese
c.472_473delAG	p.S158WfsX21	1 (1 Het)	1 Caucasian/ Asian
c.531_534del AAAT	p.V177VfsX183	2 (2 Het)	2 European
c.536delG	p.R179KfsX9	2 (1 Homo)	1 European
c.552delG	p.K185RfsX3	5 (2 Homo, 1 Het)	3 European
c.555_556delGA	p.K186SfsX4	2 (2 Het)	2 Caucasian
c.566delG	p.P568LfsX19	1 (1 Het)	1 European
c.572_573insCCCT	p.T192PfsX8	1 (1 Het)	1 NR
c.574_575insCTGC	p.R192PfsX8	9 (1 Homo, 7 Het)	3 European
			1 Czech
			4 Caucasian
c.587_588insCAGG	p.Q196HfsX4	2 (2 Het)	2 NR
c.622delA	p.R218GfsX1	1 (1 Het)	1 Chinese
c.653_654delCT	p.P218VfsX30	2 (2 Het)	2 Chinese
c.675_692del18	p.N225NfsX71	1 (1 Het)	1 NR
c.756_759delCA	p.T253SfsX37	3 (1 Homo, 1 Het)	1 European
			1 Czech
c.772_773delCC	p.P258HfsX33	3 (1 Homo, 1 Het)	2 European
c.790_791delAG	p.R263SfsX27	3 (1 Homo, 1 Het)	1 Japanese
			1 European
c.800_801insT	p.R268QfsX23	1 (1Het)	1 Slovenian
c.802_827del26insGG	p.R268GfsX26	1 (1 Het)	1 Japanese
c.807_810del4ins9	p.E270QfsX9	1 (1 Het)	1 Asian
c.812del18	p.H271PfsX1	1 (1 Het)	1 Czech
c.814_815delCT	p.L272AfsX19	1 (1 Het)	1 European

c.820_824dupTACAT	p.V276TfsX2	1 (1 Het)	1 NR
c.821_838del18	p.Y274X	4 (1 Homo, 2 Het)	3 Slovenian
c.845_846delCT	p.S282CfsX9	59 (21 Homo, 17 Het)	28 Slovenian
			6 European
			1 Denmark
			2 Caucasian
			1 Caucasian/ Asian
c.868_869insT	p.K291X	4 (2 Homo)	2 European
Splice (n=18)			
c.55-1G>A	Splice	1 (1 Het)	1 Caucasian
c.106+1G>C	**Splice**	**2 (1 Homo)**	**1 Pakistani**
c.107-2A>G	Splice	1 (1 Het)	1 Caucasian
c.240+1G>T	Splice	6 (1 Homo, 4 Het)	4 European,
			1 Caucasian/ Asian
c.240+1G>C	Splice	1 (1 Het)	1 Chinese
c.323+2T>C	Splice	1 (1 Het)	1 European
c.324-11T>G	Splice	2 (1 Homo)	1 Asian
c.515+2T>G	Splice	1 (1 Het)	1 European
c.516-2_516-1delAG	Splice	2 (1 Homo)	1 European
c.516-2A>G	Splice	1 (1 Het)	1 Caucasian/ Asian
c.588+1G>A	Splice	1 (1 Het)	1 European
c.588+1delG	Splice	1 (1 Het)	1 European
c.603-1G>C	Splice	1 (1 Het)	1 European
c.751+1G>A	Splice	1 (1 Het)	1 Sweden
c.751+6T>C	Splice	2 (2 Het)	1 European
			1 Caucasian
c.752-3C>G	Splice	1 (1 Het)	1 European
c.752-1G>C	Splice	1 (1 Het)	1 Slovenian
c.833+1G>A	Splice	2 (1 Homo)	NR
Nonsense (n=9)			
c.74G>A	p.W25X	1 (1 Het)	1 European
c.244C>T	p.Q82X	4 (2 Homo)	1 Japanese
			1 Turkish
c.574C>T	p.R192X	1 (1 Het)	1 Czech
c.640C>T	p.Q214X	1 (1 Het)	1 Japanese
c.688C>T	p.R230X	11 (3 Homo, 5 Het)	2 Sweden
			2 Czech
			1 Denmark
			1 French
			1 Germany
			1 NR
c.751C>T	p.Q251X	3 (1 Homo, 1 Het)	2 European
c.808G>T	p.E270X	1 (1 Het)	1 European

c.834G>A	p.W278X	4 (2 Homo)	2 NR
c.867G>A	p.W289X	2 (1 Homo)	1 European
Duplication (n=1)			
c.808_822dup 15	p.269_274dup5	1 (1 Het)	1 Sweden
Missense (n=16)			
c.167C>G	p.A56G	1 (1 Het)	1 Caucasian
c.239A>G	p.Q80R	1 (1 Het)	NR
c.269T>C	p.L90P	2 (2 Het)	2 Caucasian
c.370G>A	p.G124R	3 (1 Homo, 1 Het)	2 European
c.371G>A	p.G124E	1 (1 Het)	1 European
c.530T>G	p.V177G	3 (1 Homo, 1 Het)	1 Hispanic
			1 Turkish
c.539G>A	p.G180E	1 (1 Het)	1 European
c.574C>G	p.R192G	1 (1 Het)	1 Chinese
c.604G>C	p.D202H	8 (8 Het)	8 Chinese
c.608T>C	p.L203P	1 (1 Het)	1 NR
c.614G>A	p.G205E	**2 (1 Het)**	2 Caucasian
c.704T>C	p.M235T	1 (1 Het)	1 Czech
c.737T>C	p.I246T	2 (2 Het)	1 European
			1 French
c.769G>A	p.G257R	1 (1 Het)	1 Caucasian
c.820T>G	p.Y274D	1 (1 Het)	1 Japanese
c.821A>G	p.Y274C	1 (1 Het)	1 Czech

References

1. Garrod, A.E., *The incidence of alkaptonuria a study in chemical individuality.* Lancet, 1902. **2**: p. 1616-1620.
2. Nussbaum, R.L., et al., *Introduction*, in *Thompson & Thompson genetics in medicine*. 2007, Saunders/Elsevier: Philadelphia. p. xi, 585 s.
3. Guttmacher, A.E. and F.S. Collins, *Genomic medicine--a primer.* N Engl J Med, 2002. **347**(19): p. 1512-20.
4. Nussbaum, R.L., et al., *Genetics of Common Disorders with Complex Inheritance*, in *Thompson & Thompson genetics in medicine*. 2007, Saunders/Elsevier: Philadelphia. p. xi, 585 p.
5. Ezkurdia, I., et al., *The shrinking human protein coding complement: are there now fewer than 20,000 genes?* 2014: eprint arXiv:1312.7111
6. Ng, S.B., et al., *Targeted capture and massively parallel sequencing of 12 human exomes.* Nature, 2009. **461**(7261): p. 272-6.
7. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
8. Majewski, J., et al., *What can exome sequencing do for you?* Journal of Medical Genetics, 2011. **48**(9): p. 580-589.
9. Frazer, K.A., et al., *Human genetic variation and its contribution to complex traits.* Nat Rev Genet, 2009. **10**(4): p. 241-51.
10. Lettre, G., *Rare and low-frequency variants in human common diseases and other complex traits.* J Med Genet, 2014.
11. Cirulli, E.T. and D.B. Goldstein, *Uncovering the roles of rare variants in common disease through whole-genome sequencing.* Nat Rev Genet, 2010. **11**(6): p. 415-25.
12. Nussbaum, R.L., et al., *Genetic Variation in Individuals and Populations: Mutation and Polymorphism*, in *Thompson & Thompson genetics in medicine*. 2007, Saunders/Elsevier: Philadelphia. p. xi, 585 p.
13. Driscoll, D.A. and S. Gross, *Clinical practice. Prenatal screening for aneuploidy.* N Engl J Med, 2009. **360**(24): p. 2556-62.
14. Patterson, D., *Molecular genetic analysis of Down syndrome.* Hum Genet, 2009. **126**(1): p. 195-214.
15. Moore, C.M. and R.G. Best, *Chromosomal Genetic Disease: Structural Aberrations*, in *eLS*. 2001, John Wiley & Sons, Ltd.
16. Cassidy, S.B., E. Dykens, and C.A. Williams, *Prader-Willi and Angelman syndromes: sister imprinted disorders.* Am J Med Genet, 2000. **97**(2): p. 136-46.
17. Zhou, X., E.S. Iversen, Jr., and G. Parmigiani, *Classification of Missense Mutations of Disease Genes.* J Am Stat Assoc, 2005. **100**: p. 51-60.
18. Flanagan, S.E., A.M. Patch, and S. Ellard, *Using SIFT and PolyPhen to Predict Loss-of-Function and Gain-of-Function Mutations.* Genetic Testing and Molecular Biomarkers, 2010. **14**(4): p. 533-537.
19. Miller, M.P. and S. Kumar, *Understanding human disease mutations through the use of interspecific genetic variation.* Human Molecular Genetics, 2001. **10**(21): p. 2319-2328.
20. Blair, D.R., et al., *A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk.* Cell, 2013. **155**(1): p. 70-80.
21. Franco, B. and A. Ballabio, *X-inactivation and human disease: X-linked dominant male-lethal disorders.* Curr Opin Genet Dev, 2006. **16**(3): p. 254-9.

22. McKusick-Nathans Institute of Genetic Medicine, J.H.U.B., MD), *Online Mendelian Inheritance in Man, OMIM*. 2014.
23. Yang, Y., et al., *Clinical whole-exome sequencing for the diagnosis of mendelian disorders.* N Engl J Med, 2013. **369**(16): p. 1502-11.
24. Ng, S.B., et al., *Exome sequencing identifies the cause of a mendelian disorder.* Nat Genet, 2010. **42**(1): p. 30-5.
25. Goh, G. and M. Choi, *Application of whole exome sequencing to identify disease-causing variants in inherited human diseases.* Genomics Inform, 2012. **10**(4): p. 214-9.
26. Bamshad, M.J., et al., *Exome sequencing as a tool for Mendelian disease gene discovery.* Nature Reviews Genetics, 2011. **12**(11): p. 745-755.
27. Stitziel, N.O., A. Kiezun, and S. Sunyaev, *Computational and statistical approaches to analyzing variants identified by exome sequencing.* Genome Biol, 2011. **12**(9): p. 227.
28. Koboldt, D.C., et al., *Challenges of sequencing human genomes.* Brief Bioinform, 2010. **11**(5): p. 484-98.
29. Gilissen, C., et al., *Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome.* Am J Hum Genet, 2010. **87**(3): p. 418-23.
30. Buske, O.J., et al., *Identification of deleterious synonymous variants in human genomes.* Bioinformatics, 2013. **29**(15): p. 1843-50.
31. Wang, J.L., et al., *TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing.* Brain, 2010. **133**: p. 3510-3518.
32. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
33. Karger, B.L. and A. Guttman, *DNA sequencing by CE.* Electrophoresis, 2009. **30 Suppl 1**: p. S196-202.
34. Dovichi, N.J. and J.Z. Zhang, *How capillary electrophoresis sequenced the human genome.* Angewandte Chemie-International Edition, 2000. **39**(24): p. 4463-+.
35. Taylor, R.W. and D.M. Turnbull, *Mitochondrial DNA mutations in human disease.* Nat Rev Genet, 2005. **6**(5): p. 389-402.
36. Robin, E.D. and R. Wong, *Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells.* J Cell Physiol, 1988. **136**(3): p. 507-13.
37. Venegas, V., et al., *Real-time quantitative PCR analysis of mitochondrial DNA content.* Curr Protoc Hum Genet, 2011. **Chapter 19**: p. Unit 19 7.
38. Zeviani, M. and E. Lamentea, *Genetic Disorders of the Mitochondrial OXPHOS System.* Science & Medicine, 2007. **10**(3).
39. Griswold, A., *Genome packaging in prokaryotes: the circular chromosome of E. coli.* Nature Education, 2008. **1(1):57**.
40. Birky, C.W., Jr., *The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models.* Annu Rev Genet, 2001. **35**: p. 125-48.
41. Birky, C.W., *Relaxed and Stringent Genomes - Why Cytoplasmic Genes Dont Obey Mendels Laws.* Journal of Heredity, 1994. **85**(5): p. 355-365.
42. Payne, B.A., et al., *Universal heteroplasmy of human mitochondrial DNA.* Hum Mol Genet, 2013. **22**(2): p. 384-90.
43. Li, M., et al., *Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes.* Am J Hum Genet, 2010. **87**(2): p. 237-49.
44. Longo, D., et al., *Harrison's Principles of Internal Medicine*. 18 ed. 2011: The McGraw-Hill Companies.
45. Schwartz, M. and J. Vissing, *Paternal inheritance of mitochondrial DNA.* N Engl J Med, 2002. **347**(8): p. 576-80.

46. Filosto, M., et al., *Lack of paternal inheritance of muscle mitochondrial DNA in sporadic mitochondrial myopathies.* Ann Neurol, 2003. **54**(4): p. 524-6.
47. Sato, M. and K. Sato, *Maternal inheritance of mitochondrial DNA by diverse mechanisms to eliminate paternal mitochondrial DNA.* Biochim Biophys Acta, 2013. **1833**(8): p. 1979-84.
48. Sutovsky, P., et al., *Ubiquitin tag for sperm mitochondria.* Nature, 1999. **402**(6760): p. 371-2.
49. Nishimura, Y., et al., *Active digestion of sperm mitochondrial DNA in single living sperm revealed by optical tweezers.* Proc Natl Acad Sci U S A, 2006. **103**(5): p. 1382-7.
50. Sutovsky, P., et al., *Ubiquitinated sperm mitochondria, selective proteolysis, and the regulation of mitochondrial inheritance in mammalian embryos.* Biol Reprod, 2000. **63**(2): p. 582-90.
51. Ramos, A., et al., *Human mitochondrial DNA complete amplification and sequencing: a new validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification.* Electrophoresis, 2009. **30**(9): p. 1587-93.
52. Quispe-Tintaya, W., et al., *Fast mitochondrial DNA isolation from mammalian cells for next-generation sequencing.* Biotechniques, 2013. **55**(3): p. 133-+.
53. Chinnery, P.F., *Mitochondrial Disorders Overview*, in *GeneReviews(R)*, R.A. Pagon, et al., Editors. 1993: Seattle (WA).
54. Wong, L.J., *Molecular genetics of mitochondrial disorders.* Dev Disabil Res Rev, 2010. **16**(2): p. 154-62.
55. Ghezzi, D. and M. Zeviani, *Mitochondrial Disorders: Nuclear Gene Mutations.*, in *Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.* 2011.
56. Zeviani, M. and E. Lamantea, *Genetic Disorders of the Mitochondrial OXPHOS System.* Science & Medicine, 2007. **10**(3).
57. Lee, I.C., et al., *SURF1-associated Leigh syndrome: a case series and novel mutations.* Hum Mutat, 2012. **33**(8): p. 1192-200.
58. Andreu, A.L. and S. DiMauro, *Current classification of mitochondrial disorders.* J Neurol, 2003. **250**(12): p. 1403-6.
59. Man, P.Y.W., D.M. Turnbull, and P.F. Chinnery, *Leber hereditary optic neuropathy.* Journal of Medical Genetics, 2002. **39**(3): p. 162-169.
60. Zhu, Z., et al., *SURF1, encoding a factor involved in the biogenesis of cytochrome c oxidase, is mutated in Leigh syndrome.* Nat Genet, 1998. **20**(4): p. 337-43.
61. Han, J., et al., *Ophthalmological manifestations in patients with Leigh syndrome.* Br J Ophthalmol, 2014.
62. Rossi, A., et al., *Leigh Syndrome with COX deficiency and SURF1 gene mutations: MR imaging findings.* AJNR Am J Neuroradiol, 2003. **24**(6): p. 1188-91.
63. Shoubridge, E.A., *Cytochrome c oxidase deficiency.* American Journal of Medical Genetics, 2001. **106**(1): p. 46-52.
64. Debray, F.G., et al., *LRPPRC mutations cause a phenotypically distinct form of Leigh syndrome with cytochrome c oxidase deficiency.* J Med Genet, 2011. **48**(3): p. 183-9.
65. Mehndiratta, M.M., et al., *Neurological mitochondrial cytopathies.* Neurol India, 2002. **50**(2): p. 162-7.
66. DiMauro, S. and M. Hirano, *Mitochondrial DNA Deletion Syndromes*, in *GeneReviews(R)*, R.A. Pagon, et al., Editors. 1993: Seattle (WA).
67. Khambatta, S., et al., *Kearns-Sayre syndrome: a case series of 35 adults and children.* Int J Gen Med, 2014. **7**: p. 325-32.
68. Laloi-Michelin, M., et al., *Kearns Sayre syndrome: an unusual form of mitochondrial diabetes.* Diabetes Metab, 2006. **32**(2): p. 182-6.

69. Fajardo, K.V.F., et al., *Detecting false-positive signals in exome sequencing.* Human Mutation, 2012. **33**(4): p. 609-613.
70. Tusneem, N., et al., *Using the Agencourt AMPure and Agencourt CleanSEQ methods to manage a variety of input sample quality.* Beckman Coulter, 2006.
71. Birnboim, H.C. and J. Doly, *A rapid alkaline extraction procedure for screening recombinant plasmid DNA.* Nucleic Acids Res, 1979. **7**(6): p. 1513-23.
72. Qiagen, *QIAprep Miniprep Handbook*. 2 ed. 2012.
73. Bio-Rad, *Quantum Prep Plasmid Miniprep Kit Instruction Manual*. Catalog #732-6100.
74. Hawkins, T.L., et al., *DNA purification and isolation using a solid-phase.* Nucleic Acids Res, 1994. **22**(21): p. 4543-4.
75. Beckman Coulter, I., *Agencourt AMPure XP Information For Use Guide - PCR Purification*. 2013.
76. Bai, R.K. and L.J. Wong, *Simultaneous detection and quantification of mitochondrial DNA deletion(s), depletion, and over-replication in patients with mitochondrial disease.* J Mol Diagn, 2005. **7**(5): p. 613-22.
77. Finsterer, J., *Leigh and Leigh-like syndrome in children and adults.* Pediatr Neurol, 2008. **39**(4): p. 223-35.
78. Pecina, P., et al., *Functional alteration of cytochrome c oxidase by SURF1 mutations in Leigh syndrome.* Biochim Biophys Acta, 2003. **1639**(1): p. 53-63.
79. Rohlin, A., et al., *Parallel Sequencing Used in Detection of Mosaic Mutations: Comparison With Four Diagnostic DNA Screening Techniques.* Human Mutation, 2009. **30**(6): p. 1012-1020.
80. Illumina, *Data Sheet: Genomic Sequencing*. 2010.
81. Li, H., J. Ruan, and R. Durbin, *Mapping short DNA sequencing reads and calling variants using mapping quality scores.* Genome Res, 2008. **18**(11): p. 1851-8.
82. Zhang, W., H. Cui, and L.J. Wong, *Comprehensive one-step molecular analyses of mitochondrial genome by massively parallel sequencing.* Clin Chem, 2012. **58**(9): p. 1322-31.
83. Casbon, J.A., et al., *A method for counting PCR template molecules with application to next-generation sequencing.* Nucleic Acids Res, 2011. **39**(12): p. e81.
84. Bainbridge, M.N., et al., *Whole exome capture in solution with 3 Gbp of data.* Genome Biol, 2010. **11**(6): p. R62.
85. Schweyen, H., A. Rozenberg, and F. Leese, *Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters.* Biol Bull, 2014. **227**(2): p. 146-60.
86. Richterich, P., *Estimation of errors in "raw" DNA sequences: a validation study.* Genome Res, 1998. **8**(3): p. 251-9.
87. Nielsen, R., et al., *Genotype and SNP calling from next-generation sequencing data.* Nat Rev Genet, 2011. **12**(6): p. 443-51.
88. Lamont, P.J., et al., *Clinical and laboratory findings in referrals for mitochondrial DNA analysis.* Arch Dis Child, 1998. **79**(1): p. 22-7.
89. Yao, J. and E.A. Shoubridge, *Expression and functional analysis of SURF1 in Leigh syndrome patients with cytochrome c oxidase deficiency.* Hum Mol Genet, 1999. **8**(13): p. 2541-9.
90. Salviati, L., et al., *Novel SURF1 mutation in a child with subacute encephalopathy and without the radiological features of Leigh Syndrome.* Am J Med Genet A, 2004. **128A**(2): p. 195-8.

91. Hannappel, A., F.A. Bundschuh, and B. Ludwig, *Characterization of heme-binding properties of Paracoccus denitrificans Surf1 proteins.* FEBS J, 2011. **278**(10): p. 1769-78.
92. He, Y., et al., *Heteroplasmic mitochondrial DNA mutations in normal and tumour cells.* Nature, 2010. **464**(7288): p. 610-4.
93. Monnat, R.J., Jr. and L.A. Loeb, *Nucleotide sequence preservation of human mitochondrial DNA.* Proc Natl Acad Sci U S A, 1985. **82**(9): p. 2895-9.
94. Monnat, R.J., Jr. and D.T. Reay, *Nucleotide sequence identity of mitochondrial DNA from different human tissues.* Gene, 1986. **43**(3): p. 205-11.
95. Cree, L.M., et al., *A reduction of mitochondrial DNA molecules during embryogenesis explains the rapid segregation of genotypes.* Nat Genet, 2008. **40**(2): p. 249-54.
96. Calvo, S.E., et al., *Molecular diagnosis of infantile mitochondrial disease with targeted next-generation sequencing.* Sci Transl Med, 2012. **4**(118): p. 118ra10.
97. Goto, H., et al., *Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study.* Genome Biol, 2011. **12**(6): p. R59.
98. Tang, S. and T.S. Huang, *Characterization of mitochondrial DNA heteroplasmy using a parallel sequencing system.* Biotechniques, 2010. **48**(4): p. 287-292.
99. Zhang, W., H. Cui, and L.J.C. Wong, *Comprehensive One-Step Molecular Analyses of Mitochondrial Genome by Massively Parallel Sequencing.* Clinical Chemistry, 2012. **58**(9): p. 1322-1331.
100. Acinas, S.G., et al., *PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample.* Appl Environ Microbiol, 2005. **71**(12): p. 8966-9.
101. Lascaro, D., et al., *The RHNumtS compilation: Features and bioinformatics approaches to locate and quantify Human NumtS.* Bmc Genomics, 2008. **9**.
102. Hazkani-Covo, E., R.M. Zeller, and W. Martin, *Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes.* PLoS Genet, 2010. **6**(2): p. e1000834.
103. Hazkani-Covo, E., R. Sorek, and D. Graur, *Evolutionary dynamics of large numts in the human genome: rarity of independent insertions and abundance of post-insertion duplications.* J Mol Evol, 2003. **56**(2): p. 169-74.
104. Tang, S., et al., *Transition to next generation analysis of the whole mitochondrial genome: a summary of molecular defects.* Hum Mutat, 2013. **34**(6): p. 882-93.
105. Montoya, J., et al., *20 years of human mtDNA pathologic point mutations: carefully reading the pathogenicity criteria.* Biochim Biophys Acta, 2009. **1787**(5): p. 476-83.
106. DiMauro, S. and E.A. Schon, *Mitochondrial DNA mutations in human disease.* Am J Med Genet, 2001. **106**(1): p. 18-26.
107. Ebner, S., et al., *Mitochondrial haplogroups, control region polymorphisms and malignant melanoma: a study in middle European Caucasians.* PLoS One, 2011. **6**(12): p. e27192.
108. Brown, M.D., et al., *Mitochondrial DNA sequence analysis of four Alzheimer's and Parkinson's disease patients.* Am J Med Genet, 1996. **61**(3): p. 283-9.
109. Ye, K., et al., *Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals.* Proc Natl Acad Sci U S A, 2014. **111**(29): p. 10654-9.
110. Koenig, M.K., *Presentation and diagnosis of mitochondrial disorders in children.* Pediatr Neurol, 2008. **38**(5): p. 305-13.
111. Blakely, E.L., et al., *Mitochondrial DNA deletion in "identical" twin brothers.* J Med Genet, 2004. **41**(2): p. e19.

112. Moraes, C.T., et al., *Mitochondrial DNA deletions in progressive external ophthalmoplegia and Kearns-Sayre syndrome.* N Engl J Med, 1989. **320**(20): p. 1293-9.
113. Chretien, D. and P. Rustin, *Mitochondrial oxidative phosphorylation: pitfalls and tips in measuring and interpreting enzyme activities.* J Inherit Metab Dis, 2003. **26**(2-3): p. 189-98.
114. Thorburn, D.R., C.W. Chow, and D.M. Kirby, *Respiratory chain enzyme analysis in muscle and liver.* Mitochondrion, 2004. **4**(5-6): p. 363-75.
115. Dimauro, S., S. Tay, and M. Mancuso, *Mitochondrial encephalomyopathies: diagnostic approach.* Ann N Y Acad Sci, 2004. **1011**: p. 217-31.
116. Hollowell, J.G., et al., *Hematological and iron-related analytes--reference data for persons aged 1 year and over: United States, 1988-94.* Vital Health Stat 11, 2005(247): p. 1-156.
117. Lu, S., et al., *Isolation and culture of smooth muscle cells from human acute type A aortic dissection.* J Cardiothorac Surg, 2013. **8**: p. 83.
118. Yi, L. and F. Rossi, *Purification of progenitors from skeletal muscle.* J Vis Exp, 2011(49).