View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Helsingin yliopiston digitaalinen arkisto

Analysis of exome variant data for identifying causative SNVs of infantile mitochondrial disorders

Virginia Biris Brilhante

Helsinki May 24, 2013

MSc thesis in Bioinformatics

Department of Mathematics and Statistics
Faculty of Science
and
Group Wartiovaara, Research Program for Molecular Neurology
Faculty of Medicine
UNIVERSITY OF HELSINKI

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Mathematics and Statistics

Tekijä — Författare — Author
Virginia Biris Brilhante

Työn nimi — Arbetets titel — Title

Analysis of exome variant data for identifying causative SNVs of infantile mitochondrial disorders

Oppiaine — Läroämne — Subject
Bioinformatics

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
MSc thesis in Bioinformatics	May 24, 2013	44 pages

Tiivistelmä — Referat — Abstract

This thesis presents a workflow for analysis of exome sequencing data aiming at identification of single nucleotide variants (SNVs) causing recessively inherited mitochondrial disease in children. Several variant selection criteria that are consistent with such group of genetic disorders are applied along the workflow in relation to mode of inheritance, allele frequency and the Finnish ancestry of the patients. These are combined with knowledge of nuclear-encoded mitochondrial proteins and prediction of pathogenic variants, narrowing down the total set of SNVs found in a patient to those most likely to be causative. Patient exomes are analysed individually (n=1 studies). The bioinformatic resources used for implementation include public and in-house databases of mitochondrial nuclear genes, human genetic variation and exome controls, as well as software tools for prediction of pathogenic SNVs and mitochondria-targeting proteins. Exome variant data from a cohort of 49 molecularly undiagnosed children were analysed through the workflow, leading to the identification of mitochondrial disease-causing SNVs located in nuclear genes for 10 of the patients. Therefore, a success rate of 20% was achieved. The workflow has been an important element in the use of exome sequencing as a new research tool at the Wartiovaara group of the Research Program for Molecular Neurology, Faculty of Medicine, University of Helsinki.

ACM Computing Classification System (CCS):
Bioinformatics,
Molecular Sequence Analysis

Avainsanat — Nyckelord — Keywords
exome variant data analysis, SNV, infantile-onset mitochondrial disorders

Säilytyspaikka — Förvaringsställe — Where deposited
Kumpula Campus Library

Muita tietoja — övriga uppgifter — Additional information

Supervisor: Prof. Anu Wartiovaara (Research Program for Molecular Neurology, Faculty of Medicine)

Contents

1 Introduction

The present thesis project was carried out within the Wartiovaara group of the Research Program for Molecular Neurology at Biomedicum Helsinki. The group holds as its mission "to understand the molecular background of mitochondrial disorders, and use that knowledge to develop diagnosis and therapy."

The project was motivated by the group having started to use exome sequencing, performed in cooperation with FIMM (Institute for Molecular Medicine Finland), for molecular diagnosis of patients, mostly children, with suspected mitochondrial disease. The aim was then to develop an approach for the analysis of genetic variation data resulting from exome sequencing, in order to identify the mutations and genes linked to the patients' disorders. In particular, the approach was to be applied in studies comprising a single patient exome, without associated sequence data from family members or from other patients affected by the same disorder.

An exome variant data analysis workflow was developed which is customised for the characteristics of infantile mitochondrial disorders, combining computational resources built in-house and external databases and tools. The development involved users of the workflow — several members of the Wartiovaara group — who took part in iterative rounds of proposed improvements and practical application in patient studies.

This thesis discusses all the elements and the setup of the workflow. Patient study examples illustrate how the workflow elements are put together. The results obtained in the analysis of exome variant data of a cohort of 49 paediatric patients are also presented. The workflow was effective in identifying single nucleotide variants (SNVs) in nuclear genes causing mitochondrial disease, as validated by functional studies, for 10 of the patients.

The thesis is structured as follows. Chapter 2 proceeds with some background on mitochondrial disorders and exome sequencing. The core of the thesis is Chapter 3 where the workflow itself is discussed, including application examples. Chapter 4 presents the outcome of applying the workflow to a cohort of infantile-onset patients. In Chapter 5 we briefly address current and future work that stemmed from this project. Lastly, concluding considerations are drawn in Chapter 6.

2 Mitochondrial disease and exome sequencing

Mitochondria are organelles present in almost all our cells, with mature red blood cells as the only exception. Several cellular processes that are essential for life take place in mitochondria, most prominently the production of energy which makes the organelles known as the cell's power plants. Energy is produced via cellular respiration, whereby biochemical energy, in the form of oxygen and nutrients in food molecules, is converted into ATP (adenosine triphosphate) molecules and oxygen is reduced to water.

Mitochondria have their own genome: a small and circular DNA molecule containing 16 569 base pairs, believed to have been originally acquired by endosymbiosis between our distant single-cell ancestor and a bacterium some few billion years ago. Differently from nuclear DNA (nDNA), the inheritance of mitochondrial DNA (mtDNA) is strictly maternal. The existence of mtDNA does not make the organelle genetically self-sufficient, however. Most of the hundreds of proteins involved in the energy production pathway are encoded in the nucleus, synthesised in the cytoplasm and then imported into mitochondria. Notably, energy production in mitochondria is the only process in the mammalian cell known to involve two genomes — mtDNA and nDNA — operating in fine coordination.

Mutations in mtDNA or nDNA that affect proteins involved in energy metabolism in mitochondria are the underlying cause of mitochondrial disease, although environmental factors can also play a part (Ylikallio and Suomalainen, 2012). These are usually severe and progressive disorders with an estimated minimum prevalence of one in every 5 000 births, on the basis of combined data from studies undertaken in Australia, for infantile-onset disorders, and England, for adult-onset disorders (Thorburn, 2004). Treatment remains mostly palliative with no cure available at this time (Koene and Smeitink, 2011).

Heterogeneity is the hallmark of mitochondrial disorders, from genetic to biochemical to clinical features, making the disorders complex and difficult to diagnose:

> "Oxidative phosphorylation, i.e., ATP synthesis by the oxygen-consuming respiratory chain (RC) [in mitochondria], supplies most organs and tissues with energy... Consequently, RC deficiency can theoretically give rise to any symptom, in any organ or tissue, at any age, with any mode of inheritance, due to the twofold genetic origin of RC components (nuclear DNA and mitochondrial DNA)." (Munnich and Rustin, 2001)

To date, more than 100 causative mtDNA and nDNA genes have been linked to mitochondrial disease (Tucker et al., 2010). Phenotypes usually manifest in multiple organ systems, have a wide spectrum of time of onset, from perinatal to adulthood, and vary in presentation and severity throughout an individual's life span and between individuals (Suomalainen, 2011). Children tend to be the most severely affected with the poorest prognoses. Diagnosis in children is also the most challenging. Clinical presentation is markedly variable in them and histological findings are often less specific compared to adult patients (Thorburn and Smeitink, 2001; Wolf and Smeitink, 2002). Amongst so much diversity, energy deficiency in the cells, and in the patient by consequence, is the only unifying feature of mitochondrial disorders.

Within the currently known genetic underpinnings and genotype-phenotype correlations of mitochondrial disorders, at best about half of patients studied by a particular diagnostic centre have a causative mutation found in one of the known disease genes (Kirby and Thorburn, 2008; Calvo et al., 2010). Mutant genes in mtDNA and their associated disorders have been mapped out, while many nuclear genes remain to be uncovered. It has been estimated that mutations in nuclear genes cause roughly one-third of adult-onset and three-quarters of infantile-onset mitochondrial disease (DiMauro and Schon, 2003). When current knowledge of mitochondrial disease genes is exhausted to no avail, the more exploratory approach of exome sequencing can provide some answers.

2.1 Exome sequencing in identifying disease-causing mutations

Proteins are encoded by genes in nuclear and mitochondrial DNA. Each gene has coding sections, called exons, and non-coding sections, called introns. The exome consists of all exons of all genes in a genome. The human exome is estimated to correspond to only about 1% of the total genome, amounting to approximately 30Mb. This relatively small part of the genome, however, holds most of the mutations currently known to be associated with human genetic diseases. What is more, the exome is as yet better understood than non-coding and regulatory regions of the genome. Whole-exome sequencing is seen as a middle-ground approach for identifying Mendelian disease genes: it is more comprehensive and less biased than sequencing a pre-determined gene panel while at the same time potentially more cost-effective than sequencing and studying the entire genome.

Exome sequencing arose from the development of methods that couple together targeted capture and massively parallel DNA sequencing (also referred to as 'next-generation' sequencing (NGS)). The technique of capture of targeted genomic loci, today widely used for exome capture, was proposed in (Gnirke et al., 2009). In the article, a fishing analogy illustrates the technique: exon *baits* are thrown in excess in a *pond* of total human DNA fragments for a *catch* of enriched segments of exonic DNA. The baits are long single-stranded oligonucleotide probes, each consisting of a target exome segment, long enough to hybridise in solution with protein-coding exons (they are in average 169 bp long) and flanked on both sides by primer sequences for amplification. Since many target exons are shorter than the designed probes, the captured sequence as a whole extends beyond the 30Mb long exome. Magnetic beads are used to amass the catch of exome segments, which are amplified and can then be sequenced in the chosen sequencing platform. Next, the obtained exome sequence reads are aligned to the human reference genome. Identified differences are genotyped to make up the set of genetic variants contained in the exome of the sequenced individual. Today, the most widely used software for the alignment step is the Burrows-Wheeler Alignment tool (BWA) (Li and Durbin, 2009), and for variant calling, the Genome Analysis Toolkit (GATK) (DePristo et al., 2011) developed at the Broad Institute, and SAMtools (Li et al., 2009) developed at the Sanger Institute.

The types of genetic variation that can be ascertained by current exome variant calling methods and accompanying tools are SNVs (point mutations, single nucleotide substitutions), indels (small insertions and deletions) and CNVs (copy number variants). SNVs are the simplest and most common type of variation, as well as the most prevalent in association to disease. They make up approximately 55% of the pathogenic mutations in the Human Gene Mutation Database (HGMD) (Stenson et al., 2009). So far, the methods for identifying variant in NGS data are more accurate in calling SNVs, as compared to other types of variation (McKenna et al., 2010; DePristo et al., 2011). Also, for the subsequent stage of data analysis, SNVs are the best served with a variety of computational resources, such as large-scale population databases and tools for prediction of their functional effect on proteins.

In the studies of molecularly undiagnosed patients using exome sequencing data, the Wartiovaara group concentrated first on SNVs. For this reason, SNVs are the type of genetic variation this thesis focuses on. Recessively inherited, loss-of-function mutations leading to altered, reduced or absent gene products required for normal mitochondrial function are usually implicated in infantile-onset disorders. Of

interest, therefore, are the non-synonymous SNVs (nsSNVs), which change the corresponding codon so that they either code for a different amino acid (a missense mutation) or become a stop codon (a nonsense mutation).

Discovery of Mendelian disease genes through exome sequencing has been growing at an impressive rate since it was first demonstrated a few years ago (Ng et al., 2009; 2010). As originally proposed, most studies concern well-characterised disorders with clear phenotypes affecting a small number of families or unrelated individuals (Bamshad et al., 2011). By comparing exomes grouped by disorder, the search can be narrowed down to the variants shared by all (or most) of the patients and not found in unaffected family members and controls. This approach does not apply to our cohort of children with suspected mitochondrial disease, who lack a firm clinical or, in some cases, biochemical diagnosis, and in whom the same mutation can cause different phenotypes and the same phenotype can be caused by different mutations. Indeed, only a molecular diagnosis can make certain whether patients in the cohort have particular mitochondrial disorders in common. The exome variant data analysis workflow discussed here has been applied in studies starting with a single index patient (n=1 studies) suspected of having a mitochondrial disorder with variably defined genotype-phenotype relationships.

It is clear that the challenge of identifying all genes linked to mitochondrial disease, a remarkably heterogeneous and complex group of disorders, has greatly and quickly benefited from exome sequencing (Tucker et al., 2011; Tyynismaa et al., 2012; Elo et al., 2012; Haack et al., 2012; Kornblum et al., 2013; Carroll et al., 2013; to cite some of the most recent data). Another indication of the impact of exome sequencing is the growing interest in extending its use as a diagnostic tool from research to clinical settings (Bamshad et al., 2011; Haack et al., 2012; McCormick et al., 2012).

2.1.1 Technological limitations

In spite of the many successes, there are technological limitations to exome sequencing that should be noted. To start with, causal variants located outside of coding regions (in introns, untranslated (UTRs) and regulatory regions) are missed. Moreover, a consensual map of the coding regions of the human genome is still being laid out (Pruitt et al., 2009). An exome is defined in practice by the specifications of the particular capture method (kit) employed.

Types of genetic variation that involve a genomic context broader than single exons are not well detected. These include structural rearrangements and CNVs such as repeats and larger deletions. Nonetheless, methods have been emerging for detection of CNVs from exome data (Krumm et al., 2012; Fromer et al., 2012).

Errors can occur during exome capture due to defects in probes or deletions in exons, for example. Also, hybridisation is inherently not fine enough to differentiate between exons belonging to genes with very similar sequences such as close paralogues, pseudogenes, and gene family members (Gnirke et al., 2009).

There are many parts to the process of variant calling through alignment of the captured short reads, amplified to the millions, to *a* reference sequence. It is acknowledged that the reference human genome is still not free of errors. Major quality improvements have recently been made by the GENCODE project (Harrow et al., 2012). Insufficient coverage depth of the target sequence at positions is a common problem which hinders reliable variant interpretation. A coverage depth of 20× for 80% of the sequence has been an early *de facto* standard. Newer technologies now aim at an average depth of coverage in the 60× to 180× range.

In sum, characteristics of the chemistry used in the sequencing platform, of the reference sequence and of the alignment and variant calling algorithms can all influence the extent to which all true, and only true, variants are identified. Complete and correct identification of all types of human genetic variation is a running challenge for exome sequencing and NGS technologies at large.

3 Workflow for analysis of exome SNV data

Despite the exome being a hundredth part of the genome, the number of SNVs typically found in an individual exome is still large. The patients in our cohort were found to carry in average around 30 000 SNVs, the vast majority of them being benign variants. Essentially, in a Mendelian disease gene discovery study, the task at hand is to find amongst all detected genetic variation one or two true causal variants in a single gene that associates with the patient's disease. Therefore, strategies are needed to reduce the total set of variants into a much smaller set containing the ones most likely to cause disease.

This chapter describes a workflow for analysis of exome variant data that aims at identifying candidate disease SNVs and genes in paediatric patients suffering from

suspected mitochondrial disorders. It can be thought of as a series of filters that sift potential causal mutations from benign and polymorphic variants. Multiple layers of information and computational resources pertaining to genetics of mitochondrial disorders and to common and pathogenic human genetic variation are combined to form the workflow.

More specifically, the workflow entails:

- variant selection criteria customised to the genetic characteristics of infantile mitochondrial disorders;
- public and in-house databases of exome controls, of mitochondrial nuclear genes and of human genetic variation; and
- software tools for prediction of pathogenic variants and of mitochondria-targeted proteins.

3.1 Exome variant data

The exome variant data we analyse result from first extracting genomic DNA from patient samples (usually, blood or muscle), followed by sample preparation and exome sequencing and annotation performed by the Institute for Molecular Medicine Finland (FIMM).

The FIMM pipeline applied to our patient samples comprises exome capture using Roche NimbleGen Sequence Capture 2.1M Human Exome v2.0 array (except for two of the samples for which Agilent SureSelect Human All Exon Kit was used), and exome sequencing using Illumina Sequence Analyzer-IIx platform with 2×82 bp or 2×100 bp paired-end reads (Sulonen et al., 2011). The achieved exome target coverage has been between ∼95% and 98% in recent studies where the pipeline was used (Tyynismaa et al., 2012; Elo et al., 2012; Carroll et al., 2013), with a typical coverage depth of at least 20× for ∼80% of the capture target region (Sulonen et al., 2011).

The pipeline produces exome sequence reads in the standard FASTQ format. The reads go through quality filtering and scoring procedures, including assignment of Phred scores for base calling quality. Phred is an well-established method that scores each base call in a DNA sequence by analysing characteristics of peaks (e.g., spacing, amplitude, resolution) in sequence chromatograms, such as those produced by Sanger sequencing, and scoring them in relation to reference quality scores of

correctly called bases of known sequences (Ewing et al., 1998; Ewing and Green, 1998). Phred scores are typically in the [4, 60] range, with higher values indicating higher quality. A Phred quality score, Q, maps to a base calling error probability, P, through the relation $Q = -10\ log_{10}P$, or equivalently, $P = 10^{-Q/10}$. The scores can thus be conveniently related to probabilities of base calling error and accuracy: a Phred score of 10 indicates a 1 in 10 (10%) probability that the base is erroneously called, or that the call is 90% accurate; a score of 20 indicates a 1 in 100 (1%) error probability, or 99% accuracy; and so forth.

Next in the pipeline, the QCed exome sequence is aligned to the latest human reference genome assembly, hg19/GRCh37, using the Burrows-Wheeler Alignment tool (Li and Durbin, 2009), with aligned reads given in the BAM format. After that, variant calling is performed using SAMtools (Li et al., 2009) and an in-house algorithm for refinement (Sulonen et al., 2011). The common practice of discarding variant bases with a quality score <20 is adopted, so that only high quality variant calls are filtered forward. Lastly, the called variants are annotated using the Ensembl, dbSNP and 1000 Genomes databases, rendering variant data in BED and text formats. This is the starting point of our exome variant data analysis workflow.

In the exome variant data, each SNV is described by a set of data items and their values. We now list the data items that are used along the analysis workflow for selecting the best candidate variants for association with disease. For each data item a value example is given. These data item values compose the description of a SNV that led to a successful molecular diagnosis. We will further refer to this particular variant in one of the patient study examples given in Section 3.7.

- **Sample**: patient sample ID (e.g., P6).
- **Chromosome**: the chromosome where the SNV is located (e.g., chr2).
- **Position**: the genomic nucleotide position of the SNV (henceforth, the variant position) (e.g., 224824538).
- **Reference base**: the base at the variant position in the reference genome (e.g., T).
- **VCP call base**: the variant base determined by FIMM's variant calling pipeline (VCP), summarising the bases detected in the position (e.g., G). A VCP call base value that differs from the Reference base value signals a possible variant position. Further details below.

- **1000 Genomes frequency**: the frequency of the SNV in the 1000 Genomes populations (e.g., 0).
- **SNP**: the SNP ID corresponding to the SNV, if present in the dbSNP database (e.g., no value).
- **Gene**: the gene harbouring the SNV, identified by its unique Ensembl ID and its HGNC (HUGO Gene Nomenclature Committee) official symbol (e.g., ENSG00000135900 (MRPL44)).
- **Call depth**: the number of reads at the variant position, also known as read depth (e.g., 31).
- **Reference calls**: how many reads had the reference base at the variant position (e.g., 0).
- **Variant calls**: how many reads had a non-reference base at the variant position (e.g., 31). Reference and variant calls add up to call depth.
- **A**: how many of the variant calls are adenine bases (e.g., 2).
- **T**: how many of the variant calls are thymine bases (e.g., 0).
- **C**: how many of the variant calls are cytosine bases (e.g., 0).
- **G**: how many of the variant calls are guanine bases (e.g., 29).
- **Quality ratio**: a metric for variant calling quality calculated from quality scores of reference and variant base calls (e.g., 0.062). The VCP call base value depends on the Quality ratio value, as described below.

Since the processing of samples for exome sequencing involves DNA fragmentation and amplification, each variant position can and should be covered by multiple exome segment reads. The Quality ratio attribute, Qr, is modelled to give reliable VCP call base values. A variant position where the same base is detected in all covering reads has $Qr = 0$ and receives accordingly a VCP call base value of A, C, T or G. Whereas, a position in which different bases are detected in different reads receives as VCP call base value a mixed-base code according to the calculated Qr value. The standard IUB (International Union of Biochemistry) ambiguity codes for nucleotides are used (Nomenclature Committee of the International Union of

Biochemistry, 1985). For example, R designates nucleotides G or A, Y designates nucleotides T or C, and so on.

The following algorithm describes how the Quality ratio value of each of the called SNVs is calculated.

With respect to the call bases at the variant position:

IF more than two distinct bases were called (each with a count >1),

THEN $Qr = -1$; variant position discarded

ELSE

CASE 1 the reference base and one variant base were called:

$Qr = R/(R + V)$, where R is the sum of quality scores of the reference base calls, and V is the sum of quality scores of the variant base calls

CASE 2 the reference base was not called but two distinct variant bases:

$Qr = V_{\mathrm{l}}/(V_{\mathrm{l}} + V_{\mathrm{h}})$, where, between the two called variant bases, V_{l} is the lower sum of quality scores, and V_{h} is the higher sum of quality scores

CASE 3 the reference base was not called but a single variant base:

$Qr = 0$

The example SNV falls in case 2, with no T reference base but two variant bases, A and G, being called. The rationale of the quality ratio formulae in the first two cases is that the higher the quality score sum V in relation to R in case 1, or V_{h} in relation to V_{l} in case 2, the smaller Qr gets, suggesting that the called base giving V or V_{h} is a true variant. For the example, the G calls give V_{h} and the A calls give V_{l}, as can be seen from the proportion of 29 Gs to 2 As, leading to a small Qr value of 0.062.

Furthermore, the Qr value determines the VCP call base in the following way:

IF $Qr < 0.2$ **THEN** a homozygous variant base call is asserted

CASE 1 VCP call base = the variant base

CASE 2 VCP call base = the variant base giving V_{h}

CASE 3 VCP call base = the variant base

IF $0.2 \leq Qr \leq 0.8$ **THEN** a heterozygous variant base call is asserted

CASE 1 VCP call base = mixed base code corresponding to the reference and variant bases

CASE 2 VCP call base = mixed base code corresponding to the two variant bases

IF $Qr > 0.8$ **THEN** no variant base call is asserted

CASE 1 the reference base prevails and the variant position is discarded

To conclude the example, a $Qr < 0.2$ asserts G, the variant of higher calling quality, as homozygous VCP call base.

3.2 Selection of variants based on mode of inheritance

All known inheritance patterns have been observed in infantile-onset mitochondrial disease, however two of them make the vast majority: maternal and (Mendelian) recessive inheritance (Chinnery, 2002; McCormick et al., 2012). Some known mitochondrial disorders are X-linked with recessive inheritance, usually affecting boys only. Unlike nuclear DNA which has biparental inheritance, mitochondrial DNA is inherited exclusively along the maternal line (Anderson et al., 1981). Therefore, disorders caused by non-sporadic mutations in mtDNA follow a maternal inheritance pattern.

For most of the patients in our cohort, the possibility of pathogenic mtDNA mutations was excluded by mtDNA sequencing and mutation screening, or clinical evidence, leading to a presumed recessive inheritance mode. Two additional factors reinforced such presumption. Firstly, most our patients have severe infantile-onset disorders, a presentation that is typical of recessively inherited mitochondrial disorders. Secondly, most patients are of Finnish descent. Finland being a genetic isolate (Salmela et al., 2008), the likelihood of some degree of consanguinity between the parents is increased, providing further support for disease manifesting under a recessive mode of inheritance (Kirby and Thorburn, 2008).

3.2.1 Recessive inheritance, zygosity and variant calls

In regard to zygosity, homozygous and compound heterozygous variants are consistent with recessively inherited disorders. When manifesting in a homozygous state,

the individual inherits two copies of the same pathogenic variant allele (one from each parent) (Turnpenny and Ellard, 2011). In a compound heterozygous state, the individual has in the same gene two (or more, albeit unlikely) distinct heterozygous variants which combined can cause disease.

Homozygous variants were selected first during exome variant data analysis, given the Finnish ancestry of most of our patients. Besides, any individual carries a fewer number of homozygous than heterozygous variants. The selection of variants according to zygosity is done through the VCP call base data item in the exome variant data (Section 3.1). As VCP call base value, each homozygous variant has A, C, T or G, and each heterozygous variant has a mixed-base code. Compound heterozygous variants are selected by counting heterozygous variants in each gene and observing those which occur in combination with one or more other variants in the same gene.

3.3 Selection of variants based on SNV allele frequency

Variants detected in a patient affected by a rare Mendelian disorder which are also commonly found in the general population should not be causative. Grounded on this assumption, selecting out common, hence likely benign, variants is a widely adopted practice in disease gene discovery studies. To this end, the most used resources are the public databases dbSNP of NCBI (National Center for Biotechnology Information) (Sherry et al., 2001) and that of the 1000 Genomes project (1000 Genomes Project Consortium, 2012), as well as control exomes. Suitable control exomes originate from individuals sampled from the same population as but without a familial relation to the patient(s), and who are unaffected or have an unrelated phenotype.

It has been estimated that up to 90% of non-synonymous substitutions in coding and splice site regions and small indels present in an individual's genome can also be found in public data collections of human genetic variation (Robinson et al., 2011). The estimate rises to more than 95% considering coding SNVs only (Bamshad et al., 2011). In our data derived from exome sequencing, for each called SNV there are annotations sourced from dbSNP and 1000 Genomes, namely, the SNP and 1000 Genomes frequency data items shown in Section 3.1.

A reference SNP identifier annotation (e.g., rs77655487) indicates that the variant is known (not novel) but not necessarily that it is common. The ascertainment of a

dbSNP variant does not require a global population sample. Many variants detected through narrow-scope assays with small population samples, or even a single individual, are included in the database. The 1000 Genomes database provides variant allele frequencies estimated from global population samples. The 1 092 genomes from 14 populations of African, American, East Asian and European ancestry, integrated in phase I of the project, have rapidly become the most prominent baseline data set of human genetic variation. A total of 2 500 genomes will compose the database upon completion of phase II in the near future.

Selection of variants based on allele frequency requires a threshold for distinguishing common polymorphisms from rare variants possibly linked to disease. For SNVs, this threshold has been traditionally set at 1% (Kruglyak and Nickerson, 2001), which we adopted in this work. Alleles that are globally rare, however, may be relatively more common in genetic isolates such as the Finnish population, believed to bear a founder effect (Salmela et al., 2008; Turnpenny and Ellard, 2011). In order to enhance ancestry matching in the selection of rare SNVs, we take advantage of our in-house exome variant database (Section 3.3.1) and of the SNV genotypes of the 93 Finnish individuals that are part of the 1000 Genomes EUR ancestry group. Variants found in our patients which, although rare in the 1000 Genomes populations (frequency $< 1\%$), occur repeatedly in the 93 Finnish individuals above and/or in our exome controls are not prioritised for further analysis.

3.3.1 Exome variant controls

Our in-house exome variant database contains variants found in 90 patient exomes (at time of writing), studied by the Wartiovaara and Tyynismaa groups, which are part of the Research Program for Molecular Neurology of the Faculty of Medicine, University of Helsinki. The majority of the patients are of Finnish decent and have suspected mitochondrial disease. A wide range of phenotypes are represented, including encephalopathy, cardiomyopathy, lactic acidosis, arPEO and POLG Parkinson's disease, as well as presentations suggesting well-defined mitochondrial syndromes, such as Leigh and Alpers. A few non-mitochondrial patients are also part of the database.

FIMM's exome sequencing platform and variant call methods (Section 3.1) have been applied uniformly to yield the variants in the database, rendering them useful as ancestry-matched control data. Potential candidate variants in a patient under investigation are searched for in the database. Variants that are common to several

patients with phenotypic features unrelated to those of the patient at hand are selected out as potentially disease-causing. In the few instances of patients with a very similar phenotype, the database is useful for finding variants that are private to them, hence potentially causative of the disease they seem to share.

Such views into the database are realised through queries implemented in SQL (Structured Query Language) (Date, 2009) and made available to the several investigators in the Wartiovaara and Tyynismaa groups analysing patient exome variant data.

A few other exome variant databases concerning different populations and disease groups exist and can be used as additional control resources. We have used the Exome Variant Server of the University of Washington at Seattle (Exome Variant Server, 2011), which focuses on heart, lung and blood disorders, and more recently, the Genome Variant Database for Human Diseases of the University of Miami (Genome Variant Database for Human Diseases, 2012) which started off with a focus on neuromuscular diseases.

3.4 Selection of variants based on mitochondrial nuclear genes

The full characterisation of the mitochondrial proteome is still under way. It is estimated, however, that around 1 500 nuclear genes encode proteins with a mitochondrial function (Lopez et al., 2000).

Human MitoCarta is to date the most comprehensive and robust inventory of human nuclear genes encoding mitochondrial proteins. Mitochondrial DNA genes are also included. In its original publication (Pagliarini et al., 2008), 1 013 genes were included in the inventory, on the grounds of firm biochemical, statistical or literature-based evidence of mitochondrial localisation of the encoded proteins. By applying MitoCarta, we are able to combine knowledge of the patient variants gained through exome sequencing with the best available knowledge of the mitochondrial proteome.

To do so, we use Human.MitoCarta.plus26: MitoCarta appended with 26 more genes recently shown to be linked to mitochondria in the literature, most of them also used by the MitoCarta developers in a recent study (Calvo et al., 2012; Sarah Calvo, personal communication). Human.MitoCarta.plus26 is structured as a relational table with each row characterising a gene through attributes such as its official NCBI symbol, description, genomic position, type of evidence supporting mitochondrial localisation, etc. A relational join operation (Date, 2009) is performed between Hu-

man.MitoCarta.plus26 and a table containing a patient's SNVs (Section 3.1) through the gene symbol attribute. The result is a selective table of the patient's variants located in genes with established or strongly inferred mitochondrial association.

3.4.1 Prediction of mitochondria-targeting proteins

Not all genes that encode mitochondria-associated proteins are in MitoCarta — the inventory is estimated to cover about 85% of them (Pagliarini et al., 2008).

As a secondary resource for selecting variants possibly located in mitochondrial nuclear genes, we use computational tools — MitoProt, TargetP and iPSORT — that predict whether a nuclear-encoded protein is likely to be imported into and, hence, to be functional in mitochondria.

The transport of nuclear-encoded proteins into their target organelles in the cell's cytoplasm, including mitochondria, is governed by sequence motifs called sorting signals, in that they contain biochemical information that directs a protein to the organelle it belongs to. Most signals are located at the N-terminus — one of the two ends of a protein sequence marked by an amino acid with a free amino group from where translation started — and are often cleaved off upon entry of the protein into its organelle destination (Lodish et al., 2007; Emanuelsson et al., 2007). Mitochondrial proteins have an N-terminal mitochondrial targeting peptide (mTP). The computational tools harness, primarily, known biochemical properties of mTPs, such as hydrophobicity, amino-acid composition and existence of a cleavage site, to predict whether such a signal is present in a given protein sequence. Of note, there are proteins involved in key mitochondrial functions that do not have an mTP or do not locate in mitochondria, for example, outer mitochondrial membrane proteins. Bioinformatic tools, therefore, will fail to predict such proteins to be mitochondria-targeting.

In addition to properties of the signal sequence, MitoProt employs properties of the protein sequence as a whole, such as maximum hydrophobicity and total net charge. Each property is assigned a weight statistically estimated from a large collection of proteins, and a combined mitochondrial-localisation likelihood score is calculated (Claros and Vincens, 1996). TargetP and iPSORT use N-terminal sequence information only. In TargetP, N-terminal sequence properties are mapped to numerical values and fed into a neural network that calculates an mTP score (Emanuelsson et al., 2000; 2007). In iPSORT, predictions are drawn through computational rea-

soning over rules describing the N-terminal sequence properties which, as opposed to a black box neural network model, are amenable to understanding and interpretation by the users of the tool (Bannai et al., 2002). The Swiss-Prot database was the source of protein sequences used in the development of all three prediction tools.

Of interest for in silico prediction of mitochondrial localisation of encoded proteins are the genes which harbour patient variants and which are not part of MitoCarta. The tools require only the corresponding protein sequences as input, which we obtain from the NCBI Reference Sequence (RefSeq) database. We use sequences with IDs having an NP prefix (k*n*nown *p*rotein), indicating that these are high-quality, manually curated sequences. We consider for further analysis proteins predicted to target mitochondria by at least one of the tools. More precisely, a MitoProt probability of export to mitochondria > 0.5, or a TargetP mTP score > 0.5 or an iPSORT 'having a mitochondrial targeting peptide' prediciton.

3.5 Selection of variants based on predicted damage to protein

Computational tools have been developed also to predict whether an amino acid change caused by a SNV in a gene will result in damage to the protein's structure and function, possibly leading to disease. In other words, these tools predict missense pathogenic mutations. As massively parallel sequencing technologies became more widely accessible in recent years, a greater demand for computational prediction tools ensued. The last ten years or so have seen intense development in the so called in silico prediction field and many tools exist today (for a recent review see (Rantapero, 2012)).

The central premise for most of the prediction tools is that of evolutionary conservation, or sequence homology: *protein conservation across species correlates with conserved protein function.* A high degree of similarity between protein sequences from different organisms may indicate a common evolutionary ancestor and a shared protein that persisted for having a function, likely to be disrupted should a mutation occur in its encoding gene. Note that, in regard to the possible neutral, deleterious or beneficial effects of mutations in the unfolding of evolution, it is implicitly assumed here that changes in functionally conserved proteins should be deleterious.

Protein damage prediction tools founded on the evolutionary conservation premise have better applicability in the identification of variants associated with monogenic

diseases — such as most infantile mitochondrial disorders — than with common complex diseases. This is because the evolutionary conservation patterns of variants known to be linked to complex diseases appear to be indistinguishable from the patterns of polymorphisms occurring in the general population (Kumar et al., 2011).

As part of our exome variant analysis workflow, the most used tools for pathogenic variant prediction have been SIFT (Ng and Henikoff, 2003; Kumar et al., 2009) and PolyPhen-2 (Adzhubei et al., 2010). They are commonly referred to in other disease gene discovery studies, and variants in the 1000 Genomes database are annotated with their predictions.

Both SIFT and PolyPhen-2 give probabilistic estimates of the propensity of individual amino acid changes to damage protein function, on the basis of protein conservation information obtained from aligning the protein sequence in organisms from different species. In addition to that, Polyphen-2 harnesses known biochemical properties of proteins to generate the predictions in a Bayesian fashion.

As the tools differ in their total composition of predictive features and in their inference algorithms (Sections 3.5.1 and 3.5.2), so may their predictions. Nonetheless, in terms of overall performance, SIFT and PolyPhen-2 fare similarly on prediction accuracy. Comparable performance rates can be found in their original reports (Kumar et al., 2009; Adzhubei et al., 2010), as well as in recent independent assessment studies (Li et al., 2013). New meta-tools have been proposed, e.g., Condel (González-Pérez and López-Bigas, 2011) and logit (Li et al., 2013), which appear to achieve better performance by combining prediction scores from multiple tools. Let us note, however, that widely accepted standards for assessing tools (and meta-tools) for prediction of pathogenic variants are yet to be established.

In our workflow, SIFT and PolyPhen-2 are used in a complementary manner. Variants predicted to be damaging to protein function by either of the tools are considered for further analysis. In the case of compound heterozygosity, damaging predictions for both variants are not required — suffices a damaging-predicted variant compounded with other non-synonymous or non-coding variant.

In the remainder of this section, we summarise the workings of SIFT and PolyPhen-2.

3.5.1 SIFT

SIFT (Sorting Intolerant from Tolerant) predictions draw upon protein conservation information only. In SIFT's terms, amino acid positions that appear highly conserved in a protein sequence tend to be intolerant to substitution, whereas those with a low degree of conservation tolerate most substitutions.

To obtain predictions, the user provides information on his variants of interest, such as chromosome, genomic position, reference and variant alleles at the position. From that, SIFT determines the protein query sequence and searches for its homologues in the UniProt and NCBI protein resources by applying BLAST (Basic Local Alignment Search Tool). The retrieved sequences are then aligned to the query one resulting in a multiple sequence alignment (MSA). From the amino acid frequencies in the MSA and from the amino acid substitution scores in the BLOSUM62 matrix (Henikoff and Henikoff, 1992), probabilities for all the possible amino acid substitutions at each position of the alignment are calculated. These probabilities, in turn, are used to estimate the final SIFT score. The score represents the probability of an amino acid substitution (caused by an nsSNV) being tolerated. A cutoff value of 0.05 was experimentally determined. A score $\geq$0.05 predicts a TOLERATED, or functionally neutral, substitution, whereas a score $<$0.05 predicts a DAMAGING substitution, likely to affect protein function.

SIFT calculates also a so-called conservation value, which can be thought of as a measure of sequence diversity. Apart from highly conserved protein families, too little diversity (or, too much conservation) between the aligned homologous sequences is not desirable for prediction. Closely related sequences may result, for example, from the initial set of BLAST-searched sequences belonging to the same organism, as opposed to being functionally conserved orthologous sequences. Little diversity could also result from positions in the sequences being still conserved by chance in the elapsed evolutionary time.

The conservation value calculated for each position of the MSA is in the range $[0, log_2 20 (= 4.32)]$. Zero represents minimum conservation, when all 20 amino acids occur at the position, while 4.32 represents maximum conservation, when only one amino acid occurs. SIFT sets $\sim$3 as target median conservation value for each amino acid position, aiming at optimal sequence diversity in the MSA. Positions with conservation value $>$3.25 render a low confidence prediction warning.

The prediction accuracy of SIFT was estimated (Kumar et al., 2009) at a true

positive rate of 69%, corresponding to the proportion of disease-associated nsSNVs found in a cohort of affected individuals that were correctly predicted to damage protein function. When applied to a dataset of nsSNVs found in healthy individuals, SIFT predicted 19% of them as damaging, a proportion that can be interpreted as an approximate false positive rate.

3.5.2 PolyPhen-2

Like SIFT, PolyPhen-2 also forms an MSA, searching sequences from the UniProt resource according to user input. The BLOSSUM62 matrix is another shared resource with SIFT, used for estimating intermediate conservation scores of amino acid positions (not yet the final score for protein damage prediction). These intermediate scores are distinguished between profile- and identity-based scores, depending on the MSA scope involved in the estimation. Profile-based scores reflect the substitution patterns and the relatedness of the homologous sequences in the MSA as a whole. Identity-based scores, on the other hand, reflect the identity between the query sequence and its closest homologues. Intuitively, an amino acid position within an MSA of highly diverse sequences would receive a low profile-based conservation score; an amino acid position in a query sequence that is highly identical to its closest homologues would receive a high identity-based conservation score.

Now, unlike SIFT, PolyPhen-2 combines conservation information with physicochemical features of amino acids, and with structural features of proteins, the latter being limited to proteins with known 3D structures. Some examples of features are: CpG (Cytosine–phosphate–Guanine) context of transition mutations (purine to purine, A $\leftrightarrow$ G, or pyrimidine to pyrimidine, C $\leftrightarrow$ T), deemed to correlate with mutation rate; change in the amino acid volume given the mutation; accessible surface area and B-factor, an indicator of conformational mobility, of the wild-type amino acid. To enable probabilistic inference, conservational features are represented by alignment scores, and each biochemical feature by probability distribution values.

The final PolyPhen-2 score represents the probability that an nsSNV is damaging (is pathogenic, or affects protein function) and is inferred by a naive Bayes classifier. A general interpretation of such a classifier would be that a prior probability, $P(d)$, representing an initial degree of belief in the pathogenicity of the nsSNV, is assigned and then reviewed in the light of features (observed data) that are relevant for pathogenicity — biochemical properties of amino acids, structural and conservational properties of the protein — to give $P(d|features)$. A corresponding general

formulation for the classifier, followed by the rationale of its terms, is thus:

$$P(d|features) = \frac{P(d) \times P(features|d)}{P(features)}$$

where

- $P(d|features)$ is the posterior probability of the nsSNV being damaging given the supporting features;
- $P(d)$ is the prior probability assigned to the nsSNV being damaging;
- $P(features|d)$ is the probability of the features given that the prediction d holds, in other words, the likelihood that the observed features are supported by a damaging nsSNV; and
- $P(features)$ is the marginal probability of the features, representing their overall combined likelihood.

Several different features can be exploited to inform prediction of protein-damaging variants. This is the motivation for the *naive* Bayes approach in PolyPhen-2, an approach that is particularly useful in scenarios where diverse data classes are combined to support the inference of posterior probabilities. Naive Bayes classifiers assume that the classes of supporting data (or features) are independent, in this way drastically simplifying the modelling and training of the classifier. In actuality, the features are often not independent. In the PolyPhen-2 context, amino acid volume and accessible surface area, for example, are not biologically independent protein features. Oversimplistic as it may seem, the feature independence assumption does not usually impair the classifier's performance significantly. PolyPhen-2 classifier's performance was comparable to that of other machine learning methods (Adzhubei et al., 2010; supplementary material).

PolyPhen-2 distinguishes three classes of nsSNVs, according to the inferred probabilistic score: benign (score ≤ 0.15), possibly damaging ($0.15 <$ score ≤ 0.85), and probably damaging (score > 0.85). Additionally, PolyPhen-2 reports true positive (sensitivity) and true negative (specificity) estimates for a given variant.

The prediction accuracy of PolyPhen-2 was estimated by applying it to two datasets compiled from UniProt: one dataset comprising variants associated with human Mendelian diseases, and another comprising variants associated with human genetic disease more generally. Association with disease was asserted on the basis of

UniProt annotations. Variants not annotated as linked to disease in the database were assumed to be benign. For the dataset of variants linked to Mendelian diseases, PolyPhen-2 displayed a true positive rate of 92% and a false positive rate of 20%. For the less Mendelian disease-specific dataset of variants, the true positive rate was 73%, and the false positive rate, 20% (comparable to SIFT's 69% and 19%, respectively (Section 3.5.1)).

3.6 Prioritisation of candidate variants and genes

After applying the variant selection criteria discussed in the previous sections, a best case scenario is to have at hand one to a few variants that could cause the disease in the patient. Here we consider additional requisites of these candidate variants which are used for prioritising them for further functional studies. Fulfilment of the requisites increases the likelihood of a variant being a true causal disease mutation.

Evolutionary conservation is one such requisite, for the reasons discussed in Section 3.5. We use sequence alignment software to observe the degree of conservation across species of the changed amino acids and of their flanking sequence block.

An independent form of sequence analysis, most often Sanger sequencing, is performed in order to confirm the variant and its zygosity. The most common cause of false positives in candidate variant identification is insufficient coverage depth of the genomic segment in question. We have had several instances of unconfirmed homozygosity of variants with a small coverage depth (Call depth data item in Section 3.1). When available, family samples are also sequenced for confirmation of a recessive inheritance pattern or of a *de novo* mutation. In the case of an inherited homozygous mutation: affected siblings, if any, are expected to also have the mutation; parents are expected to be heterozygous carriers; and unaffected siblings, to be heterozygous carriers or not to carry the mutation. Compound heterozygosity is confirmed by observing that the two mutations are not allelic, i.e., that each of them is inherited from a distinct parental allele. When parental samples are not available, non-allelic mutations can be confirmed by cloning the genetic region in bacterial vectors, but this can be done only if the mutations are relatively close together (Tyynismaa et al., 2012).

Approximately 400 ancestry-matched control chromosomes are screened in order to verify that the mutation does not occur in healthy individuals. A small number of heterozygous carriers was sometimes found in our studies, indicating a possible

enrichment of the concerned variants in the genetically isolated Finnish population.

For each patient case, expert biologists and clinicians in the Wartiovaara group carefully consider the candidate variants and genes in relation to the patient's phenotype. This is clearly a decisive, final step in the prioritisation of variants. Previous studies in the literature may (or may not) provide support for investigating the variant further. Once prioritised, variants are subjected to biochemical assays aiming at understanding their consequences at the molecular level and their associations with the patient's disease.

3.7 The workflow exemplified

The order in which the different variant selection criteria discussed in the previous sections are applied is not fixed. A general strategy is to give precedence to the criterion that reduces the current set of SNVs the most. In this section, a couple of examples of successful application of the workflow are given, one singling out a homozygous variant later confirmed to underlie the patient's disease, and another yielding compound heterozygous variants in a few candidate disease genes.

The first example comes from a sibship of three girls born to healthy Finnish parents without close consanguinity. The eldest daughter is healthy while the other two are affected with cardiomyopathy. The middle daughter died at six months of age from acute cardiac insufficiency which developed after a respiratory infection.

We performed exome sequencing for the second affected daughter, now a teenager. Elevated enzymes in the liver were her first sign of disease found at three months of age. At eight months, she was diagnosed with hypertrophic cardiomyopathy, which progressed initially but stabilised after 2 years of age. The patient is now 14 years old with normal psychomotor development, apart from selective mutism. The cardiomyopathy remains stable and asymptomatic.

Figure 1 shows a series of selection criteria applied to the exome variant data from the patient. The process starts off with a total of 23 958 detected SNVs. Across our cohort of 49 patients, the average number of exomic SNVs has been around 30 000, in line with other published studies where similar exome sequencing technologies were used.

Selecting variants that were either absent or present in $<1\%$ frequency in the 1000 Genomes populations resulted in 3 950 SNVs. Of these, 56 SNVs were found to locate in mitochondrial nuclear genes by use of Human.MitoCarta.plus26. Only two

of these SNVs were homozygous, one of them receiving 'damaging' predictions from both SIFT and PolyPhen-2, with scores 0.02 (probability of the resulting amino acid substitution being tolerated) and 0.989 (probability of the SNV being damaging to protein structure and function), respectively.

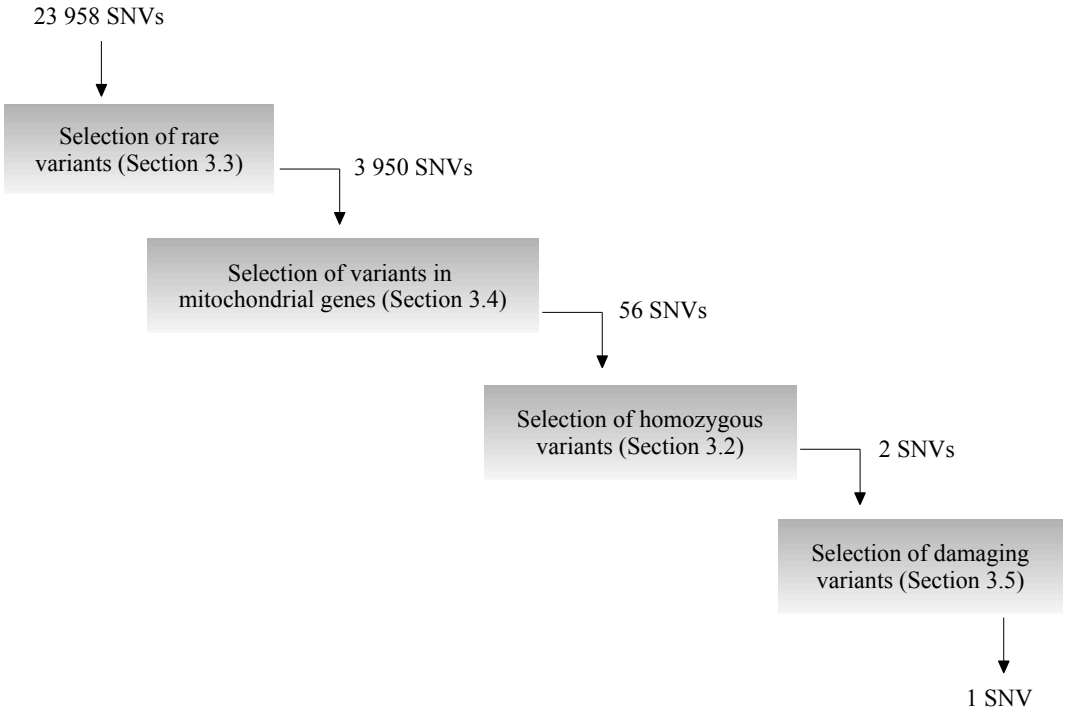

Figure 1: Example of exome variant data analysis workflow identifying a homozygous candidate variant.

The resulting SNV c.467T>G (p.Leu156Arg) — a thymine at nucleotide 467 in the coding DNA reference sequence is changed to a guanine, and consequently the amino acid Leucine-156 is changed to an Arginine in the protein sequence — is located on chromosome 2q in exon 2 (of 4) of the *MRPL44* gene. The Leu156 amino acid is highly conserved among vertebrates. Sanger sequencing confirmed the homozygous mutation in the patient as well as in her affected sibling, while the parents and the healthy sibling were found to be heterozygous carriers.

Through the 1000 Genomes Ensembl browser, one can find the SNP rs1433697995, imported from dbSNP (release 137), which corresponds to the variant but found in heterozygote carriers only. The SNP has a reported rounded frequency of 0.001 for the T/G genotype in a European-American population (3 T/G to 4 297 T/T). The

variant was not found in 436 Finnish control chromosomes, neither in our in-house 90 exome controls (Section 3.3.1).

MRPL44 (mitochondrial ribosomal protein L44) encodes a protein of the large subunit of the mitochondrial ribosomes, the structures where proteins encoded by mtDNA are synthesised. The identification of the mutant MRPL44 in the two siblings motivated functional studies, which demonstrated the role of the protein in assembly and stability of the ribosomal subunit and indicated the identified homozygous mutation as a novel genetic cause for primary mitochondrial hypertrophic cardiomyopathy with variable clinical presentation (Carroll et al., 2013).

Figure 2 gives a second example of application of the workflow, this time identifying compound heterozygous candidate variants. The studied patient suffers from cardiomyopathy and encephalopathy with onset at one year of age.

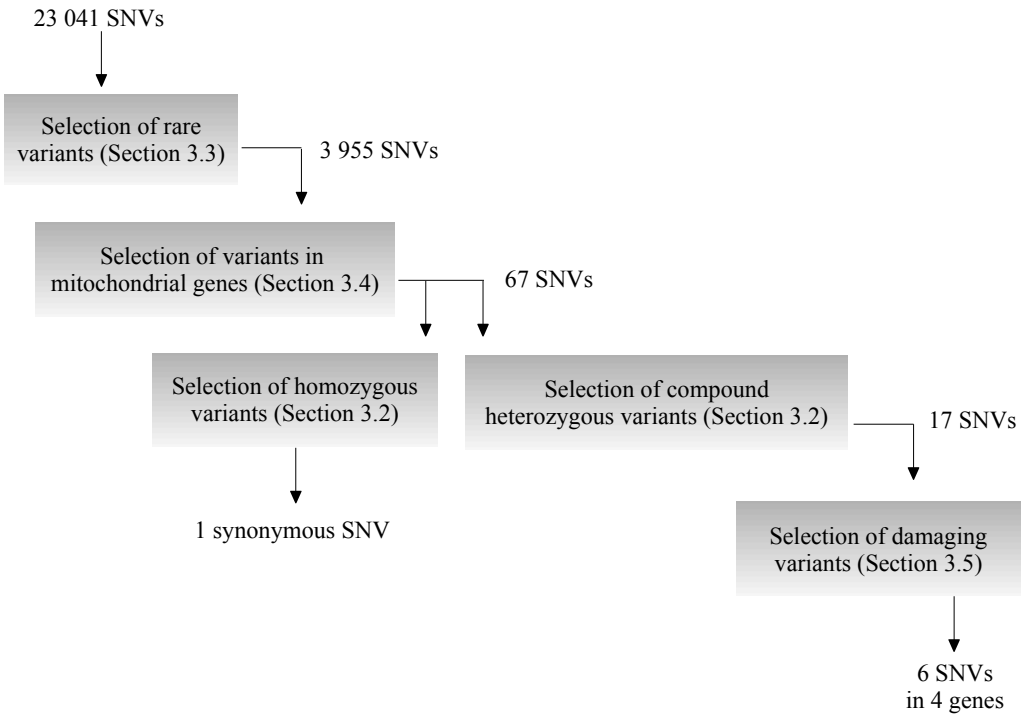

Figure 2: Example of exome variant data analysis workflow identifying compound heterozygous candidate variants.

Applying the selection criteria down to rare, homozygous variants in nuclear mitochondrial genes led to a single synonymous SNV. The selection was then extended

to compound heterozygous variants, leading to 17 SNVs. Of these, six variants distributed in four different genes were predicted to be protein damaging by SIFT and/or PolyPhen-2.

Among the genes, MNDG1 (Mitochondrial Nuclear DNA Gene 1 — fictitious name), with one missense and one nonsense mutation found in the patient, was prioritised for further studies, as the gene is known to cause severe infantile-onset mitochondrial disease. Both altered amino acids are highly conserved. The missense mutation was not found in any of our other (89) disease exomes. Interestingly, the nonsense mutation was also found, compound heterozygous with another missense mutation, in a second, unrelated patient in the cohort presenting with juvenile-onset neuropathy. The mutations in the two patients were confirmed by Sanger sequencing.

MNDG1 encodes an elongation factor which is needed, along with many other proteins, for mitochondrial translation. Its function in the elongation step affects the accuracy of protein synthesis in mitochondria. Computational structural modelling of the gene with the identified SNVs predicted them to cause protein instability progressing to degradation. The study indicates the variants as novel compound heterozygous mutations underlying less severe phenotypes with later onset (Ahola-Erkkilä et al. (unpublished)), as compared to a previously reported[1] homozygous mutation causing infantile multiorgan mitochondrial disease leading to early death.

4 Aggregate Results

Firstly, this thesis work resulted in the structured assembly of the presented data analysis workflow. The workflow was then applied to exome variant data from patients with a view to reaching genetic molecular diagnoses.

Table 1 lists a cohort of 49 paediatric patients with suspected recessively inherited mitochondrial disease. As seen on the table, the cohort encompasses a variety of phenotypes, the majority being encephalopathies, cardiomyopathies and mitochondrial syndromes, such as Leigh and Alpers. The workflow was executed on the exome variant data sets from these 49 patients jointly by this author and several other members of the Wartiovaara group.

[1]References are not given due to pending publication of the patient study.

Table 1: Cohort of 49 paediatric patients with suspected recessively inherited mitochondrial disease. Disease genes, and supporting references, identified through application of the exome variant data analysis workflow. Findings of nuclear genes with confirmed association to mitochondrial disease are highlighted in bold. MNDG1–4 (Mitochondrial Nuclear DNA Gene), NMNDG1–4 (Non-Mitochondrial Nuclear DNA Gene), and MTDG1–3 (Mitochondrial DNA Gene) are fictitious gene names, used due to pending publication of the respective patient studies. na, not available.

Patient	Age of onset (death)	Phenotype	Identified disease gene	Reference
P1	na	cardiomyopathy, mtDNA depletion		
P2	**3.5 m (10 m)**	**cardiomyopathy**	***AARS2***	**Götz et al. (2011)**
P3	**2nd day (8 m)**	**encephalomyopathy**	***FARS2***	**Elo et al. (2012), Shamseldin et al. (2012)**
P4	**3 m**	**cardiomyopathy**	***MRPL44*** (Section 3.7)	**Carroll et al. (2013)**
P5	**1 y**	**cardiomyopathy, encephalopathy**	**MNDG1** (Section 3.7)	**Ahola-Erkkilä et al. (unpublished), undisclosed supporting references**
P6	prenatal (neonatal)	infantile lactic acidosis, myopathy		
P7	birth	neuropathy, encephalopathy		
P8	na	*POLG*-like	NMNDG1	undisclosed supporting reference
P9	na	*POLG*-like		
P10	2 m (1.5 y)	*POLG*-like		
P11	na	*POLG*-like		
P12	na	*POLG*-like		
P13	na	*POLG*-like		
P14	birth (4 y)	encephalopathy	NMNDG2	Tyynismaa et al. (unpublished)
P15	1.3 y	Leigh syndrome-like		
P16	**4 m**	**cardiomyopathy, Leigh syndrome-like encephalopathy**	**MNDG2**	**undisclosed supporting reference**

Patient	Age of onset (death)	Phenotype	Identified disease gene	Reference
P17	6 y (9 y)	epileptic encephalopathy, Alpers syndrome-like		
P18	3 m	encephalopathy		
P19	<2 y (9 y)	encephalopathy		
P20	**teen-age**	**neuropathy**	**MNDG1**	**Ahola-Erkkilä et al. (unpublished), undisclosed supporting references**
P21	birth	encephalopathy		
P22	1st day	encephalopathy		
P23	8 m	encephalopathy		
P24	birth	encephalopathy	NMNDG2	Tyynismaa et al. (unpublished)
P25	birth (5 m)	Alpers syndrome-like		
P26	**1 m (2.5 y)**	**Leigh syndrome**	**MNDG3**	**Matilainen et al. (unpublished), undisclosed supporting references**
P27	1 y	cardiomyopathy, encephalopathy		
P28	<2 y	encephalopathy, renal presentation		
P29	prenatal (5 m)	failure to thrive, encephalomyopathy		
P30	1.25 y (10 y)	Leigh syndrome, hearing deficit		
P31	4 y	encephalopathy		
P32	birth (12 y)	Leigh syndrome		
P33	<1 m	encephalopathy		
P34	6 m	encephalopathy		
P35	prenatal (2nd day)	infantile lactic acidosis		
P36	**10 m (1.5 y)**	**Leigh syndrome**	**MNDG2**	**undisclosed supporting reference**
P37	**5 y**	**encephalopathy**	**MNDG4**	**undisclosed supporting reference**
P38	na	Leigh syndrome, cardiomyopathy		

Patient	Age of onset (death)	Phenotype	Identified disease gene	Reference
P39	na	encephalopathy		
P40	neonatal	encephalopathy		
P41	5-6 y (21 y)	encephalopathy	MTDG1	
P42	1st day	encephalopathy	MTDG2	
P43	1st day	encephalopathy	NMNDG3	undisclosed supporting references
P44	prenatal (4 m)	hepatoencephalopathy, hypertrophic cardiomyopathy		
P45	**1 m**	**hepatoencephalopathy**	**NMNDG4**	**undisclosed supporting reference**
P46	1 y (9 y)	Leigh syndrome	MTDG3	
P47	2 y (13 y)	Leigh syndrome-like		
P48	na	encephalomyopathy		
P49	<1 y (20 y)	Leigh syndrome, Alpers syndrome-like		

Overall, for 17 out of the 49 patients a disease gene was identified. Table 1 gives the literature references, produced by the Wartiovaara group and others, supporting each of the nuclear DNA findings. For the studies with pending publication, fictitious gene names are used and supporting references are undisclosed.

Three of the findings (patients P41, P42 and P46) were in mtDNA genes, identified by using additional data analysis resources, such as the MITOMAP and mtDB databases, which are outside of the scope of this thesis. Five other findings concern genes that do not encode mitochondrial proteins (patients P8, P14, P24, P43 and P45). However, the protein encoded by NMNDG4 (P45) participates in a reaction in the cytosol that yields a cofactor which is transported into mitochondria and is essential to the organelle's energy metabolism.[2] Therefore, a nuclear gene linked to a mitochondrial disorder was identified for 10 of the patients.

The distribution of the mutations in the 10 patients according to zygosity was as follows:

- five patients with a homozygous mutation in the identified gene (P2, P4, P36, P37 and P45);
- four patients with heterozygous mutations (P3, P5, P20 and P26); and
- one patient with a heterozygous mutation (P16).

Each of MNDG1 and MNDG2 was identified for two patients. Therefore, 8 distinct disease genes were identified. With regard to novelty of the findings, they are:

- three novel disease genes (*AARS2*, *FARS2* and *MRPL44*);
- two known disease genes with a novel phenotype association (MNDG1 (2×) and MNDG3); and
- three known genes in known phenotypes (MNDG2, MNDG4 and NMNDG4).

Patients P1 to P5 were selected first for exome sequencing and data analysis, as both their clinical and biochemical data bore the strongest evidence of a mitochondrial disease. For only one of these patients a molecular diagnosis was not reached, giving an astounding success rate of 80%. All 49 patient exomes included, the success rate was 35% (17/49) considering all findings, and 20% (10/49) considering nuclear

[2]Reference is not given due to pending publication of the patient study.

genes and confirmed mitochondrial disorders only. Not surprisingly, analysis of exome variant data resulted in a higher proportion of genetic diagnosis for clear mitochondrial pathology cases compared to a more openly defined cohort.

5 Current and Future Work

This thesis project focused on analysis of SNV data obtained from exome sequencing for molecular diagnosis of suspected mitochondrial patients. Extensions of the project have started to develop in regard to analysis of other types of genetic variation and the construction of a relational database of patient data, starting with exome sequencing variant data, for the Wartiovaara group.

5.1 Other types of genetic variation

As pointed out in Section 2.1, SNVs, indels and CNVs are ascertainable from exome sequencing data. So far, SNVs have been the most extensively and systematically examined type of variation in our patient studies.

Second to SNVs, small insertions and deletions ranging from one to a few tens of bases, or **indels**, are the next most common mutation type associated with Mendelian disease. Indels amount to about one quarter of the total of entries in the Human Gene Mutation Database (HGMD) (Stenson et al., 2009).

FIMM's exome sequencing and variant calling pipeline (Section 3.1) produces a data set of indels for each patient (in addition to a data set of SNVs). Variant calling systems such as SAMtools, which is part of the pipeline, and GATK are in general less accurate with indels in comparison to SNVs (Bamshad et al., 2011), having a tendency to overcall them (Daniel MacArthur, personal communication).

Algorithms and software tools have emerged for predicting whether an indel is damaging to gene function. There are distinct tools for in-frame and frameshifting indels. Because an amino acid is coded by three consecutive bases (a codon), a frameshift can occur when the number of inserted or deleted bases is not a multiple of three.

PROVEAN, Protein Variation Effect Analyzer (Choi et al., 2012), predicts in-frame indels, and also single and short multiple amino acid substitutions, as neutral or deleterious to the encoded protein. The prediction method uses pairwise alignments between a query protein sequence and sequences of functional homologues. The

similarity between the query sequence and a homologue is measured with an without the variant occurring in the query sequence. If the variant reduces the similarity between the two sequences, as measured by a pairwise alignment score, then the variant is predicted to be deleterious; otherwise, it is predicted to be neutral. Note that PROVEAN uses the change in the alignment score as a measure of the effect of the variant on protein function. The measuring of similarity takes into account a short sequence region encompassing the variant site and flanking amino acids on both sides. This is why the method is applicable to both indels and short amino acid substitutions. The PROVEAN tool was developed and is hosted together with SIFT, the damaging SNV predictor (Section 3.5.1), at the J. Craig Venter Institute.

In (Hu and Ng, 2012), an algorithm is described which predicts frameshifting indels to be neutral or gene-damaging. The algorithm consists of a decision tree classifier with classification rules built on the basis of a set of 20 features of the indels and of the genes where they occur. For example, number of transcripts not affected by the indel, position of the indel on the affected gene transcripts, number of overlapping amino acids between the original and the indel-modified protein, number of conserved nucleotide positions affected by the indel. The source of disease-causing frameshift indels used to train the decision tree algorithm was the Human Gene Mutation Database (HGMD).

We have used the above prediction tools for in-frame and frameshifting indels in the analysis of some of our patient exome data. Data formatting templates were built to form batch input to the tools.

SNVs and indels at **splice sites** can alter gene products and have a pathogenic role. Our exome variant data contain splice site variant annotations derived from SAMtools and Ensembl. Additionally, Eino Palin, from our group, developed a program that uses known sequence motifs at splice sites to find their genomic positions. These positions can then be matched against exome variant data to identify SNVs and indels at splice sites occurring in a patient. Furthemore, a number of algorithms and tools exist for prediction of splicing sites and alterations, among them NNSPLICE (Reese et al., 1997), GeneSplicer (Pertea et al., 2001) and Human Splicing Finder (Desmet et al., 2009).

Large-scale gains or losses of genomic segments, spanning up to several million bases, caused by duplication or deletion events are known as **CNV** (Copy Number Variation). Exome sequencing is not specifically tailored for detecting CNV, as the units of sequence capture, the exons, are short and sparsely distributed along the

genome.

Nonetheless, methods and tools have been developed for calling CNVs from exome sequencing data by leveraging depth of coverage (read depth) information. The rationale is that, given an estimated read-depth average for a set of sequenced exomes, a position of a particular exome with an above-average depth may indicate gain of sequence (duplication) around the position and, conversely, a below-average depth may indicate loss of sequence (deletion). The highly variable read-depth values that exome sequencing typically yields, however, are also influenced by a host of other factors: DNA biochemical properties, sample batch biases, experimental and bioinformatic procedures. Data analysis techniques such as principal component analysis (PCA) are applied for identification of confounding factors which, in turn, enables depth of coverage normalisation and more accurate detection of CNV signal. CoNIFER — Copy Number Inference From Exome Reads (Krumm et al., 2012) and XHMM — eXome Hidden Markov Model (Fromer et al., 2012) are two of the recently described methods. Our group has started experimenting with CoNIFER.

Once systematic analysis of all these types of variation is in place together with SNV analysis, the possibility of variants of different types acting in combination to harm gene function can also be considered — for example, an indel combined with a splice site SNV, a small-scale variant combined with CNV, and so forth.

5.2 ExoMitDB

A strength of the exome variant analysis workflow described in this thesis was its easy assimilation into the everyday work of biologists and medical doctors in the Wartiovaara group. One reason was that no computing platform shift was required. All the variant analysis steps are performed through Web-based software tools and the Microsoft Office suite (Excel and Access), familiar to the researchers. Moreover, the development wait was minimal. The workflow took shape incrementally, with new steps put into practice quickly. That said, the benefits that a relational database solution designed for patient exome analysis could bring about were apparent.

The exome variant data in its original form of a SNV table per patient was described in Section 3.1. Although variant data from all the patients are now accessible through a functioning Access database and SQL queries, the original tables had not been structured within the relational database paradigm — they are simply data sheets, one per patient, one SNV per row.

A project was initiated to build a relational database to allow storage, management and interrogation of data on patients studied by the Wartiovaara group. Exome data is the first type of patient data being tackled through development of a database prototype we are calling ExoMitDB. Jan Vollert, a student from the MSc programme in Molecular Biology/Bioinformatics at the University of Applied Sciences Gelsenkirchen, Germany, started off the design and implementation of the prototype as a summer internship project.

ExoMitDB is being developed with a view to integrate other existing and future types of patient data, such as, respectively, data on biological samples and whole-genome sequencing data. Capacity for extension is one of the advantages of a relational database solution. This requires, however, a careful design that reflects all concerned data entities and their relationships. Figure 3 shows a very high-level depiction of the conceptual data entities in ExoMitDB: *Patients* and their *Family* members, *Phenotypic* features patients present with and *Disease* groups of interest (e.g., mitochodrial disease, cardiomyopathies, etc.), genetic *Variants* identified in an individual and *Genes* where they locate, and *Variant Analysis* relating patients and their families, phenotypes and variants to draw disease associations.

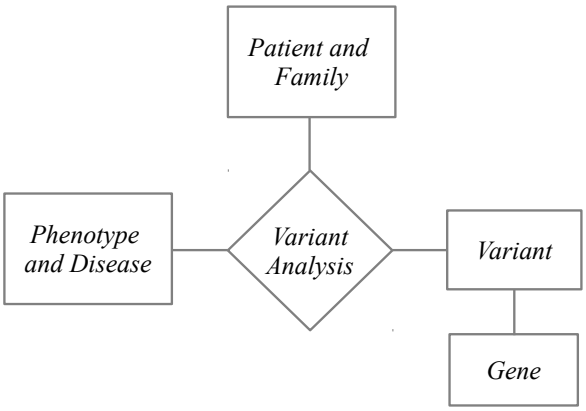

Figure 3: High-level conceptual data entities in ExoMitDB.

ExoMitDB will allow the steps composing the data analysis workflow to be further automated and better integrated. Querying flexibility will also improve, responding to diverse data analysis needs as they emerge. Analysis of patient variant data relies on external information resources, such as the human reference genome or databases of genetic variation in the general population, which are constantly evolving. Auto-

mated integration between ExoMitDB and relevant external information resources, allowing for fast data re-analysis following updates, is another desired feature.

Our choice of database management system for implementing ExoMitDB was MySQL, motivated by it being robust, open source, widely used in bioinformatics (which eases the integration with external resources just mentioned) and, importantly, by the provision of a MySQL server by FIMM. For scripting of connections to the server, procedures for loading existing data to the new relational tables, data analysis procedures, etc., we chose the Python programming language. Python is also open source and widely used in bioinformatics, has a clean syntax and supports interface with MySQL.

6 Conclusion

This thesis project set up a workflow for analysis of exome variant data for identification of mutations causing mitochondrial disease in children. The workflow has been one component in the introduction of exome sequencing as a new research tool for the Wartiovaara group, and is a current resource for actual patient studies performed by several members of the group. Rather than finished, it is in continuous development in keeping, as much as possible, with rapid advances in human genomics, sequencing technologies, and bioinformatic methods and tools. Also, the workflow has had ramifications into analysis of data on other types of genetic variation detectable through exome sequencing (Section 5.1) and into constructing a relational database of patient data on a robust platform as a bioinformatics infrastructure project for the Wartiovaara group (Section 5.2).

The workflow targets specifically SNVs in nuclear genes coding for proteins which when absent or defective can impair mitochondrial function and cause disease. The results of applying the workflow in 49 paediatric patient studies were reported (Chapter 4). SNVs in nuclear genes selected through the workflow led to a confirmed molecular diagnosis of a mitochondrial disorder for 10 of the patients. This represented a significant increase in successful diagnoses in a short period of time. Some of the patients had been previously studied at length, with traditional methods falling short of revealing the underlying cause of their disease. Moreover, with each diagnosis, enabled by exome sequencing and substantiated with solid functional studies, a better understanding of mitochondrial biology and pathogenicity comes about.

Still, most cases in our patient cohort remain unsolved. Part of them can be at-

tributed to types of genetic variation other than SNVs. Another part, to the technological limitations of exome capture, sequencing and variant calling (Section 2.1.1). Another yet is due to analytical limitations, imposed by shortcomings of the resources used to apply some of the variant selection criteria (Sections 3.3 to 3.5).

Human.MitoCarta.plus26 is not an exhaustive inventory of mitochondrial genes, and prediction tools for both mitochondria-targeting proteins and damaging variants do give false negatives.

Our main reference for selecting variants on the basis of population allele frequency has been the 1% threshold on the 1000 Genomes populations. Ancestry-specific population data provide a better ground for frequencies that distinguish between common and potentially pathogenic variants. And, clearly, patient and population should be ancestry matched, even more so for isolates, such as the Finnish population, with their private genetic make-up. We now have good prospects for this kind of resource becoming available with the Sequencing Initiative Suomi (SISu) project (Palotie et al., 2013).

Our workflow assumes a monogenic recessive pattern of inheritance for infantile mitochondrial disorders. It may be so that more complex genetic phenomena are also involved (Calvo et al., 2012; McCormick et al., 2012): interactions between multiple, low-penetrance alleles in mtDNA and/or nDNA genes, modifier genes, or epigenetics, for example.

All in all, this work joins others in showing that exome sequencing, coupled with data analysis that incorporates properties of mitochondrial disorders, is a powerful approach to keep expanding our understanding of these disorders and bringing effective therapies closer.

Acknowledgements

My opening thanks go to my first supervisor, Anu Wartiovaara, for giving me the unique opportunity to develop this thesis project, with a generous dose of belief in my capacity to pull it off. I admire Anu greatly, for her leadership and commitment to her work and that of her group.

My gratitude also to Henna Tyynismaa, who walked me through the beginnings of the project and who has since been always ready to offer advice.

Sirkka-Liisa Varvio has given me a lot of attention, way beyond the line of duty, since

the start of the MBI (Master's Degree Programme in Bioinformatics at the University of Helsinki), as studies coordinator, as lecturer, and as my second supervisor. Thank you so much, Siru.

I thank all members of the Wartiovaara group. They have been welcoming and sharing from day one and have provided me with an inspiring environment to work and learn. I wish to acknowledge in particular Chris Carroll and Eino Palin with whom I have been working the closest and learning loads from, and Erika Weckström for all the help with administrative matters, for our fun Finnish–Portuguese language tandem, and for her friendship.

Katherine Icay and I worked together in many course assignments. After that, we were fellow MBI students at the Research Programme for Molecular Neurology for a while. She is now a dear friend for life.

Doing a master's and having a child seem to correlate in my life. My first daughter, Sofia, came during the MSc in Information Technology at the University of Edinburgh, Scotland, and my second, Nina, during the MBI here in Finland. Wonderfully fruitful times! My husband, Ilias Biris, and Sofia not only allowed but supported mom to indulge in going back to school, again, to learn some more. For that I am deeply grateful.

References

1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4):248–249.

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

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet*, 12(11):745–755.

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S. (2002). Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, 18(2):298–305.

Bilgüvar, K., Oztürk, A. K., and Louvi, A. e. a. (2010). Whole-exome sequencing identifies recessive wdr62 mutations in severe brain malformations. *Nature*, 467(7312):207–210.

Calvo, S., Compton, A., Hershman, S., Lim, S. C., Lieber, D. S., Tucker, E. J., Laskowski, A., Garone, C., Liu, S., Jaffe, D. B., Christodoulou, J., Fletcher, J. M., Bruno, D. L., Goldblatt, J., Dimauro, S., Thorburn, D. R., and Mootha, V. K. (2012). Molecular diagnosis of infantile mitochondrial disease with targeted next-generation sequencing. *Sci Transl Med*, 4(118):118ra10.

Calvo, S. E., Tucker, E. J., Compton, A. G., Kirby, D. M., Crawford, G., Burtt, N. P., Rivas, M. A., Guiducci, C., Bruno, D. L., Goldberger, O. A., Redman, M. C., Wiltshire, E., Wilson, C. J., Altshuler, D., Gabriel, S. B., Daly, M. J., Thorburn, D. R., and Mootha, V. K. (2010). High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. *Nat Genet*, 42(10):851–858.

Carroll, C. J., Isohanni, P., Pöyhönen, R., Euro, L., Richter, U., Brilhante, V., Götz, A., Lahtinen, T., Paetau, A., Pihko, H., Battersby, B. J., Tyynismaa, H., and Suomalainen, A. (2013). Whole-exome sequencing identifies a mutation in the mitochondrial ribosome protein MRPL44 to underlie mitochondrial infantile cardiomyopathy. *J Med Genet*, 50(3):151–159.

Chinnery, P. F. (2002). Inheritance of mitochondrial disorders. *Mitochondrion*, 2(1–2):149–155.

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., and Chan, A. P. (2012). Predicting the functional effect of amino acid substitutions and indels. *PLoS One*, 7(10):e46688.

Claros, M. G. and Vincens, P. (1996). Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur J Biochem*, 241(3):779–786.

Date, C. J. (2009). *SQL and Relational Theory*. O'Reilly Media.

DePristo, M., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl, C., Philippakis, A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T., Kernytsky, A., Sivachenko, A., Cibulskis, K., Gabriel, S., Altshuler, D., and Daly, M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43(5):491–498.

Desmet, F. O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res*, 37(9):e67.

DiMauro, S. and Schon, E. A. (2003). Mitochondrial respiratory-chain diseases. *N Engl J Med*, 348(26):2656–2668.

Elo, J. M., Yadavalli, S. S., Euro, L., Isohanni, P., Götz, A., Carroll, C. J., Valanne, L., Alkuraya, F. S., Uusimaa, J., Paetau, A., Caruso, E. M., Pihko, H., Ibba, M., Tyynismaa, H., and Suomalainen, A. (2012). Mitochondrial phenylalanyl-tRNA synthetase mutations underlie fatal infantile Alpers encephalopathy. *Hum Mol Genet*, 21(20):4521–4529.

Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP, and related tools. *Nat Protoc*, 2(4):953–971.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol*, 300(4):1005–1016.

Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res*, 8(3):186–194.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res*, 8(3):175–185.

Exome Variant Server (2011). NHLBI GO Exome Sequencing Project (ESP), Seattle, WA. http://evs.gs.washington.edu/EVS/. Accessed May 2013.

Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., Handsaker, R. E., McCarroll, S. A., O'Donovan, M. C., Owen, M. J., Kirov, G., Sullivan, P. F., Hultman, C. M., Sklar, P., and Purcell, S. M. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet*, 91(4):597–607.

Genome Variant Database for Human Diseases (2012). University of Miami, Miller School of Medicine. https://genomics.med.miami.edu/. Accessed May 2013.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D. B., Lander, E. S., and Nusbaum, C. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol*, 27(2):182–189.

González-Pérez, A. and López-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet*, 88(4):440–449.

Götz, A., Tyynismaa, H., Euro, L., Ellonen, P., Hyötyläinen, T., Ojala, T., Hämäläinen, R., Tommiska, J., Raivio, T., Oresic, M., Karikoski, R., Tammela, O., Simola, K., Paetau, A., Tyni, T., and Suomalainen, A. (2011). Exome sequencing identifies mitochondrial alanyl-tRNA synthetase mutations in infantile mitochondrial cardiomyopathy. *Am J Hum Genet*, 88(5):1–8.

Haack, T. B., Haberberger, B., Frisch, E. M., Wieland, T., Iuso, A., Gorza, M., Strecker, V., Graf, E., Mayr, J. A., Herberg, U., Hennermann, J. B., Klopstock, T., Kuhn, K. A., Ahting, U., Sperl, W., Wilichowski, E., Hoffmann, G. F., Tesarova, M., Hansikova, H., Zeman, J., Plecko, B., Zeviani, M., Wittig, I., Strom, T. M., Schuelke, M., Freisinger, P., Meitinger, T., and Prokisch, H. (2012). Molecular diagnosis in mitochondrial complex I deficiency using exome sequencing. *J Med Genet*, 49(4):277–283.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R., and Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*, 22(9):1760–1774.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA (PNAS)*, 89(22):10915–10919.

Hu, J. and Ng, P. C. (2012). Predicting the effects of frameshifting indels. *Genome Biol*, 13(2):R9.

Kirby, D. M. and Thorburn, D. R. (2008). Approaches to finding the molecular basis of mitochondrial oxidative phosphorylation disorders. *Twin Res Hum Genet*, 11(4):395–411.

Koene, S. and Smeitink, J. (2011). *Mitochondrial medicine*. Khondrion BV, Nijmegen, The Netherlands, 1st edition.

Kornblum, C., Nicholls, T. J., Haack, T. B., Schöler, S., Peeva, V., Danhauser, K., Hallmann, K., Zsurka, G., Rorbach, J., Iuso, A., Wieland, T., Sciacco, M., Ronchi, D., Comi, G. P., Moggio, M., Quinzii, C. M., DiMauro, S., Calvo, S. E., Mootha, V. K., Klopstock, T., Strom, T. M., Meitinger, T., Minczuk, M., Kunz, W. S., and Prokisch, H. (2013). Loss-of-function mutations in MGME1 impair mtDNA replication and cause multisystemic mitochondrial disease. *Nat Genet*, 45(2):214–219.

Kruglyak, L. and Nickerson, D. A. (2001). Variation is the spice of life. *Nat Genet*, 27(3):234–236.

Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., Quinlan, A. R., Nickerson, D. A., and Eichler, E. E. (2012). Copy number variation detection and genotyping from exome sequence data. *Genome Res*, 22(8):1525–1532.

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protoc*, 4(7):1073–1081.

Kumar, S., Dudley, J. T., Filipski, A., and Liu, L. (2011). Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. *Trends Genet*, 27(9):377–386.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, M.-X., Kwan, J. S. H., Bao, S.-Y., Yang, W., Ho, S.-L., Song, Y.-Q., and Sham, P. C. (2013). Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet*, 9(1):e1003143.

Lodish, H., Berk, A., Kaiser, C. A., Krieger, M., Scott, M. P., Bretscher, A., Ploegh, H., and Matsudaira, P. (2007). *Molecular Cell Biology*. W. H. Freeman, 6th edition.

Lopez, M. F., Kristal, B. S., Chernokalskaya, E., Lazarev, A., Shestopalov, A. I., Bogdanova, A., and Robinson, M. (2000). High-throughput profiling of the mitochondrial proteome using affinity fractionation and automation. *Electrophoresis*, 21(16):3427–3440.

McCormick, E., Place, E., and Falk, M. J. (2012). Molecular genetic testing for mitochondrial disease: from one generation to the next. *Neurotherapeutics*, 10(2):251–261.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20(9):1297–1303.

Munnich, A. and Rustin, P. (2001). Clinical spectrum and diagnosis of mitochondrial disorders. *Am J Hum Genet*, 106(1):4–17.

Ng, P. C. and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*, 31(13):3812–3814.

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., Shendure, J., and Bamshad, M. J. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet*, 42(1):30–35.

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E. E., Bamshad, M., Nickerson, D. A., and Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276.

Nomenclature Committee of the International Union of Biochemistry (1985). Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. *Eur J Biochem*, 150(1):1–5.

Pagliarini, D. J., Calvo, S., Chang, B., Sheth, S. A., Vafai, S. B., Ong, S. E., Walford, G. A., Sugiana, C., Boneh, A., Chen, W. K., Hill, D. E., Vidal, M., Evans, J. G., Thorburn, D. R., Carr, S. A., and Mootha, V. K. (2008). A mitochondrial protein compendium elucidates complex I disease biology. *Cell*, 134(1):112–123.

Palotie, A., Widén, E., and S., R. (2013). From genetic discovery to future personalized health research. *N Biotechnol*, 30(3):291–295.

Pertea, M., Lin, X., and Salzberg, S. L. (2001). GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res*, 29(5):1185–1190.

Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., Diekhans, M., Maglott, D. R., Searle, S., Farrell, C. M., Loveland, J. E., Ruef, B. J., Hart, E., Suner, M. M., Landrum, M. J., Aken, B., Ayling, S., Baertsch, R., Fernandez-Banet, J., Cherry, J. L., Curwen, V., Dicuccio, M., Kellis, M., Lee, J., Lin, M. F., Schuster, M., Shkeda, A., Amid, C., Brown, G., Dukhanina, O., Frankish, A., Hart, J., Maidak, B. L., Mudge, J., Murphy, M. R., Murphy, T., Rajan, J., Rajput, B., Riddick, L. D., Snow, C., Steward, C., Webb, D., Weber, J. A., Wilming, L., Wu, W., Birney, E., Haussler, D., Hubbard, T., Ostell, J., Durbin, R., and Lipman, D. (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res*, 19(7):1316–1323.

Rantapero, T. (2012). Bioinformatic analysis of next-generation sequencing data. MSc thesis, University of Tampere.

Reese, M. G., Eeckman, F. H., Kulp, D., and Haussler, D. (1997). Improved splice site detection in Genie. *J Comp Biol*, 4(3):311–323.

Robinson, P. N., Krawitz, P., and Mundlos, S. (2011). Strategies for exome and genome sequence data analysis in disease-gene discovery projects. *Clin Genet*, 80(2):127–132.

Salmela, E., Lappalainen, T., Fransson, I., Andersen, P., Dahlman-Wright, K., Fiebig, A., Sistonen, P., Savontaus, M.-L., Schreiber, S., Kere, J., and Lahermo, P. (2008). Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. *PLoS ONE*, 3(10):e3519.

Shamseldin, H. E., Alshammari, M., Al-Sheddi, T., Salih, M. A., Alkhalidi, H., Kentab, A., Repetto, G. M., Hashem, M., and Alkuraya, F. S. (2012). Genomic analysis of mitochondrial diseases in a consanguineous population reveals novel candidate disease genes. *J Med Genet*, 49(4):234–241.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 29(1):308–311.

Stenson, P. D., Ball, E. V., Howells, K., Phillips, A. D., Mort, M., and Cooper, D. N. (2009). The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. *Hum Genomics*, 4(2):69–72.

Sulonen, A. M., Ellonen, P., Almusa, H., Lepistö, M., Eldfors, S., Hannula, S., Miettinen, T., Tyynismaa, H., Salo, P., Heckman, C., Joensuu, H., Raivio, T., Suomalainen, A., and Saarela, J. (2011). Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol*, 12(9):R94.

Suomalainen, A. (2011). Therapy for mitochondrial disorders: little proof, high research activity, some promise. *Semin Fetal Neonatal Med*, 16(4):236–240.

Thorburn, D. R. (2004). Mitochondrial disorders: prevalence, myths and advances. *J Inherit Metab Dis*, 27(3):349–362.

Thorburn, D. R. and Smeitink, J. (2001). Diagnosis of mitochondrial disorders: clinical and biochemical approach. *J Inherit Metab Dis*, 24(2):312–316.

Tucker, E. J., Compton, A. G., and Thorburn, D. R. (2010). Recent advances in the genetics of mitochondrial encephalopathies. *Curr Neurol Neurosci Rep*, 10(4):277–285.

Tucker, E. J., Hershman, S. G., Köhrer, C., Belcher-Timme, C. A., Patel, J., Goldberger, O. A., Christodoulou, J., Silberstein, J. M., McKenzie, M., Ryan, M. T., Compton, A. G., Jaffe, J. D., Carr, S. A., Calvo, S. E., RajBhandary, U. L., Thorburn, D. R., and Mootha, V. K. (2011). Mutations in MTFMT underlie a human disorder of formylation causing impaired mitochondrial translation. *Cell Metab*, 14(3):428–434.

Turnpenny, P. D. and Ellard, S. (2011). *Emery's Elements of Medical Genetics*. Elsevier, Philadelphia, PA, 14th edition.

Tyynismaa, H., Sun, R., Ahola-Erkkilä, S., Almusa, H., Pöyhönen, R., Korpela, M., Honkaniemi, J., Isohanni, P., Paetau, A., Wang, L., and Suomalainen, A. (2012). Thymidine kinase 2 mutations in autosomal recessive progressive external

ophthalmoplegia with multiple mitochondrial DNA deletions. *Hum Mol Genet*, 21(1):66–75.

Wolf, N. I. and Smeitink, J. A. (2002). Mitochondrial disorders: a proposal for consensus diagnostic criteria in infants and children. *Neurology*, 59(9):1402–1405.

Ylikallio, E. and Suomalainen, A. (2012). Mechanisms of mitochondrial diseases. *Ann Med*, 44(1):41–59.