May 2013

The use of next generation sequencing technologies to dissect the aetiologies of neurodegenerative diseases

Arianna Tucci, MD

This thesis is submitted to the University College of London for the degree of Doctor of Philosophy

Institute of Neurology
University College London

Declaration

I, Arianna Tucci, hereby declare that the work presented in this thesis was performed by myself. Exceptions to this are indicated elsewhere in the thesis.

A Martino

“Meglio aggiungere vita ai giorni che non giorni alla vita.”

Rita Levi Montalcini

Abstract

Advances in next generation sequencing technologies have brought a paradigm shift in how medical researchers investigate human disorders. Whole exome sequencing (WES) allows to comprehensively study, in a single experiment, coding variations of a human genome.

My PhD thesis focuses on the use of WES to dissect genetic aetiologies of neurodegenerative diseases. First, I present a pilot study where I prove the feasibility of this technology and I test the methods used in the following projects. I then describe the use of WES in rare Mendelian disorders: i) in patients with Kohlschütter-Tönz Syndrome, where WES has failed to identify the disease gene, because the mutation is in a non-coding region; ii) in a family with Charcot-Marie-Tooth type 6, where coupled with linkage data, WES rapidly identified the mutation causing the disease. Last, I present two projects where next generation sequencing plays an important role in elucidating the role of variants present in two loci (PARK16 and *EIF4G1*) in Parkinson's Disease.

Acknowledgements

The work presented here was made possible by the people that surround, or have surrounded myself in the past.

I am deeply grateful to my supervisor John Hardy, for his exceptional mentorship and constant support. Particular thanks to Prof. Henry Houlden for his support, Dr Vincent Plagnol for his advice, Dr Coro Paisán-Ruiz for her patience teaching me the basis of genetic research, Prof. Nick Wood for his support and advice, Dr Elia Stupka and Francesco Lescai for introducing myself to bioinformatics.

Thanks to the members of the laboratory of neurogenetics at Institute of Neurology; Anna Sailer, Boniface Mok, Reema Paudel, Nuria Seto, Lucia Schottlaender, Eleanna Kara, Niccolo' Mencacci, Una Sheerin, Alan Pittman, Mina Ryten, Daniah Trabzuni, Raquel Duran, Selina Wray, Lee Stanyer, June Smalley, Jana Vandrovcova, Dalia Kasperaviciute, Rohan de Silva and Roberto Simone.

Thanks to the members of the neurogenetics lab at NIH: Dr Andy Singleton, Dena Hernandez, Dr Mike Nalls and especially Celeste Sassi.

I am also very grateful to the people who first introduced myself to research, from the Laboratory of Genetics and Biochemistry in Milan: Stefania Corti, Monica Nizzardo and Giacomo Comi.

Thanks to my lifetime friends sorelle Rugis Stefania e Ale, Giulia. Special thanks to Viola, Silvia and Licia; to my university fellows Alberto, Sara, Mimi', Chiarina, SDM, Patoz, Luca, Leo, Tommi. Thanks to M'and the pizzica, Salvatore and Marta Wolek.

Thanks to my family: my late grandma Nonna Giovanna, Zia Margherita. Thanks to the wonderful Emilia and Luigi, to Laura e Pietro and the little cousins Elena and Alessandro.

Thanks to my parents for their encouragement, help and support. Thanks to the person that knows me better than I know myself, Fiora. Thanks to the astonishing beauty of Manarola and Tellaro.

Thanks to Martino and the brand new Riccardo, thanks for coming to this world, my little ones. And thanks to Pietro, for being always there – with a smile.

Table of contents

Abbreviations

AD	Alzheimer's Disease
ADP	adenosine diphosphate
AI	amelogenesis imperfecta
ALS	amyotrophic lateral sclerosis
ASD	autism spectrum disorder
ATP	adenosine triphosphate
ATP13A2	ATPase type 13A2
BAM	Binary of sequence Alignment/Map format
BN gel	Blue native gel
bp	base pairs
BST1	bone marrow stromal cell antigen 1
C9orf72	chromosome 9 open reading frame 72
C12orf65	chromosome 12 open reading frame 65
cDNA	coding DNA
CGH	complementary genomic hybridization
CMT	Charcot–Marie–Tooth disease
DDCt	2 delta-delta Ct
DNA	Deoxyribonucleic acid
dNTPs	deoxyribonucleoside triphosphates
DOA	dominant optic atrophy
EEG	Electroencephalography
EIF4G1	eukaryotic translation initiation factor 4 gamma, 1
FCCP	carbonyl cyanide *p*-trifluoromethoxyphenylhydrazone
GA	genome analyser
GAK	cyclin G associated kinase
GBA	glucosidase, beta, acid

GJB1	gap junction protein, beta 1
GW	genome wide
GWAS	genome wide association study
HapMap	haplotype map
HBD	homozygous by descent
HBSS	Hanks' balanced salt solution
HLA	major histocompatibility complex, class
HMSN	hereditary motor and sensory neuropathies
IBD	identical by descent
KRS	Kufor-Rakeb syndrome
KTS	Kohlschütter-Tönz syndrome
L-DOPA	L-3,4-dihydroxyphenylalanine
LD	Linkage disequilibrium
LHON	Leber hereditary optic neuropathy
LLRK2	leucine-rich repeat kinase 2
LOF	loss of function
MAF	minor allele frequency
Mb	Megabases
MFN2	mitofusin 2
MILS	maternally inherited Leigh's syndrome
MPTP	1-methyl 4-phenyl 1,2,3,6-tetrahydropyridine
MTATP6	mitochondrially encoded ATP synthase 6
mtDNA	mitochondrial DNA
NARP	neuropathy, ataxia, retinitis pigmentosa
NCBI	National Center for Biotechnology Information
NCL	neuronal ceroid lipofuscinosis
NGS	next generation sequencing
NHLBI	National Heart, Lung, and Blood Institute
NIH	National Institute of health
NMD	nonsense mediated decay
OPA1	optic atrophy 1
OXPHOS	oxidative phosphorylation enzyme

PBMC	cultivated peripheral blood mononuclear cells
PCR	polymerase chain reaction
PD	Parkinson disease
Pi	inorganic phosphate
PRKCG	protein kinase C, gamma
qRT	quantitative real-time
RefSeq	Reference Sequence Database
RF-1	release factor-1
RFLP	restriction fragment length polymorphism
RNA	Ribonucleic acid
ROGDI	rogdi homolog (Drosophila)
RPL13A	ribosomal protein L13a
RT-PCR	real-time PCR
SAM	Sequence Alignment/Map format
SDS	standard deviation score
SN	substantia nigra
SNCA	synuclein, alpha
SNP	single nucleotide polymorphisms
SNV	single nucleotide variant
TECR	trans-2,3-enoyl-CoA reductase
TGM6	TGM6 transglutaminase 6
TMRM	tetramethylrhodamine methylester
TREM2	triggering receptor expressed on myeloid cells 2
UK	United Kingdom
VCP	valosin containing protein
WDR62	WD repeat domain 62
WES	whole exome sequencing
WRN	Werner syndrome, RecQ helicase-like

Tables

Figures

1 INTRODUCTION

1.1 NEURODEGENERATIVE DISEASES: IMPORTANCE OF GENETIC ANALYSIS

Neurodegenerative diseases represent the major medical and social problem in the ageing population and have drawn a lot of attention due to their irreversibility, lack of effective treatment, and accompanied social and economical burdens. In our aging societies they afflict about 2% of the population at any time and the current demographic trends point to a likely increase their prevalence. Neurodegenerative diseases are characterized by a chronic progressive clinical course and irreversible neuronal loss in the central nervous system, leading to distinct clinical phenotypes. Although the majority of neurodegenerative diseases are sporadic, Mendelian forms have been well documented. For several inherited disorders, most of what is known about the biological mechanisms underlying the disease comes from the knowledge of the protein affected by the mutations that causes it. Interestingly, the clinical presentations and neuropathological findings of Mendelian forms of the disease are often indistinguishable from the sporadic disease, raising the possibility that common pathophysiologic mechanisms underlie both forms of disease. The study of genetics of neurodegenerative diseases is important as the identification of causative genes accelerates studies on the pathophysiologic mechanism of disease, and thus offers target pathways for therapy. It also leads to the development of animal models of the disease that can be used to develop ideas about pathogenesis and to test therapies. Lastly, but not less importantly, the identification of a mutation in a family, has a potential clinical impact on that family in terms of diagnosis and presymptomatic and prenatal testing.

For example, the identification of genetic variants causing Parkinson's Disease (PD) allows the identification of individuals at risk for disease prior to the onset of motor symptoms, since symptoms normally appear when a high percentage of neurons in the substantia nigra have already degenerated.

This thesis focuses on the use the recently developed whole exome sequencing technology (WES), to dissect the genetic aetiology of neurodegenerative diseases.

1.2 GENETIC RESEARCH OVERVIEW

The fundamental aim of genetics is to connect genotype with phenotype. However determining the DNA sequence that causes a specific trait can be particularly difficult in humans, for which experimental interventions (such as mutagenesis, crosses and selection) are unavailable and the phenotype of interest may be very subtle.

Traditionally, the search for a disease gene in a Mendelian trait starts with linkage analysis. In this approach, the aim is to find out the rough location of the gene by taking advantage of the meiotic process of recombination as manifest in families segregating for the disease. Markers (such as microsatellites) in the disease gene region show the strongest correlation with disease patterns in families; the tracking of recombination events can narrow the region harbouring a disease gene to between 100 and several thousand kilobases. This provides information regarding the location of the causative gene, after which sequencing of candidate genes is performed to pinpoint the actual mutation.

Gene discovery in Mendelian disorders advanced after the 1980s, following the discovery of restriction fragment length polymorphisms (RFLPs) first (Botstein et al., 1980) and then of the abundant highly polymorphic microsatellite (short tandemly repetitive DNA) loci (Weber and May, 1989), (Litt and Luty, 1989). Linkage analysis using RFLPs and microsatellite was established as a general method for connecting simple Mendelian diseases with the DNA sequence. In the following decades this method successfully identified over a 1000 genes causing to Mendelian diseases (Botstein and Risch, 2003). With respect to neurodegenerative diseases such as PD,

to date, 15 PD-associated loci have been identified in familial forms of the disease (also knows as PARK-, OMIM #168600) (Table 3).

This success reflects the power of linkage analysis when applied to Mendelian phenotypes, characterized by a (near) one-to-one correspondence between genotypes at a single locus and the observed phenotype. Of note, linkage analysis has been very successful in monogenic diseases presenting in families with the following features: 1) clinically well-characterized disorders, where the misdiagnosis is unlikely; 2) families with juvenile/young onset form of disease, where disease appears early in life and is unlikely to be missed; 3) families with adequate pedigree size, where sufficient number of affected individuals can be collected and 4) with sufficient locus homogeneity. This approach has led to the identification of genetic variants (mainly coding) that are usually rare but confer high risk toward disease, owing to the clear inheritance patterns they display.

The genetics of common disorders have proven much more challenging to study by linkage analysis, owing to the weak linkage signals (mainly due to locus heterogeneity) and the concomitant requirement for large sample cohorts. Indeed, complex disorders are thought to be caused by a combined effect of many different susceptibility DNA variants of low penetrance (Weeks and Lathrop, 1995). Linkage analysis was shown to have low power when a trait locus contributes only a small fraction to the disease (Risch, 2000). Moreover the mapping of human susceptibility loci for such diseases can be complicated by incomplete penetrance, phenocopies and possible epistasis (a combined effect of one or more loci). As a consequence, linkage approaches to complex diseases without Mendelian inheritance had very limited success (Altmüller et al., 2001).

A popular hypothesis about genetic variation and risk of complex diseases is the 'common disease – common variant' hypothesis, which suggests that common modest-risk alleles confer susceptibility to common disease (CDCV hypothesis). To dissect the genetic background of sporadic diseases the attention turned to association studies, in which a difference in allele frequency is sought between affected individuals and unrelated unaffected controls.

Efforts to catalogue the human genome sequence and common genetic variants were required to test whether a common genetic variant (eg. a Single Nucleotide Polymorphism, SNP) increases risk of disease by comparing allele frequencies in affected and unaffected cases.

The completion of the Human Genome Project in early 2000 laid the foundation for the development of the International Haplotype Map (HapMap) project in 2002. The HapMap project aimed at determining common patterns in DNA sequence variation in the human genome, by characterising sequence variants and their frequencies from populations with ancestry from Africa, Asia and Europe.

1.2.1 Genetic analysis using high throughput SNP genotyping

The development of the HapMap project coupled with the technological advances in ultra-high-throughput genotyping made genome-wide – SNP (GW-SNP) testing possible. The HapMap project produced a minimal set of informative SNPs to tag variation throughout the human genome (McVean et al., 2005). In parallel with the HapMap project an effort was undertaken to produce cost-effective methods to perform high throughput genotyping accurately and reproducibly. These developments provided the basis for a new era of genetic analysis such as genome wide association studies (GWAS) in complex diseases, and autozygosity mapping in consanguineous families.

Homozygosity mapping in diseases

The concept of homozygosity mapping was first proposed in 1987 by Lander and Botstein (Lander and Botstein, 1987) to map genes that cause recessive traits in consanguineous marriages. In a consanguineous marriage, both parents share some of their ancestors and may have each inherited a copy of the same ancestral allele at a locus. Their child may then be homozygous by descent (or autozygous) for that allele. If a child from consanguineous marriage is homozygous for a particular allele, this could be because of autozygosity (they are identical by descent, IBD) or it could be because a second, independent example of the same allele has entered the family at some stage (these alleles can be described as identical by state, IBS). The rarer the allele is in the population, the greater is the likelihood that homozygosity represents autozygosity. Autozygosity mapping involves the detection of the disease locus by taking advantage of the fact that adjacent region (haplotype) will be preferentially homozygous by descent in inbred children. By defining the regions of homozygosity in an affected individual from an inbred family, the disease locus can be refined, thus facilitating the eventual identification of the gene. As in traditional linkage analysis, the identification of homozygous regions is followed by candidate gene sequencing in the region to pinpoint the causal mutation, which is the rate-limiting step for the entire procedure.

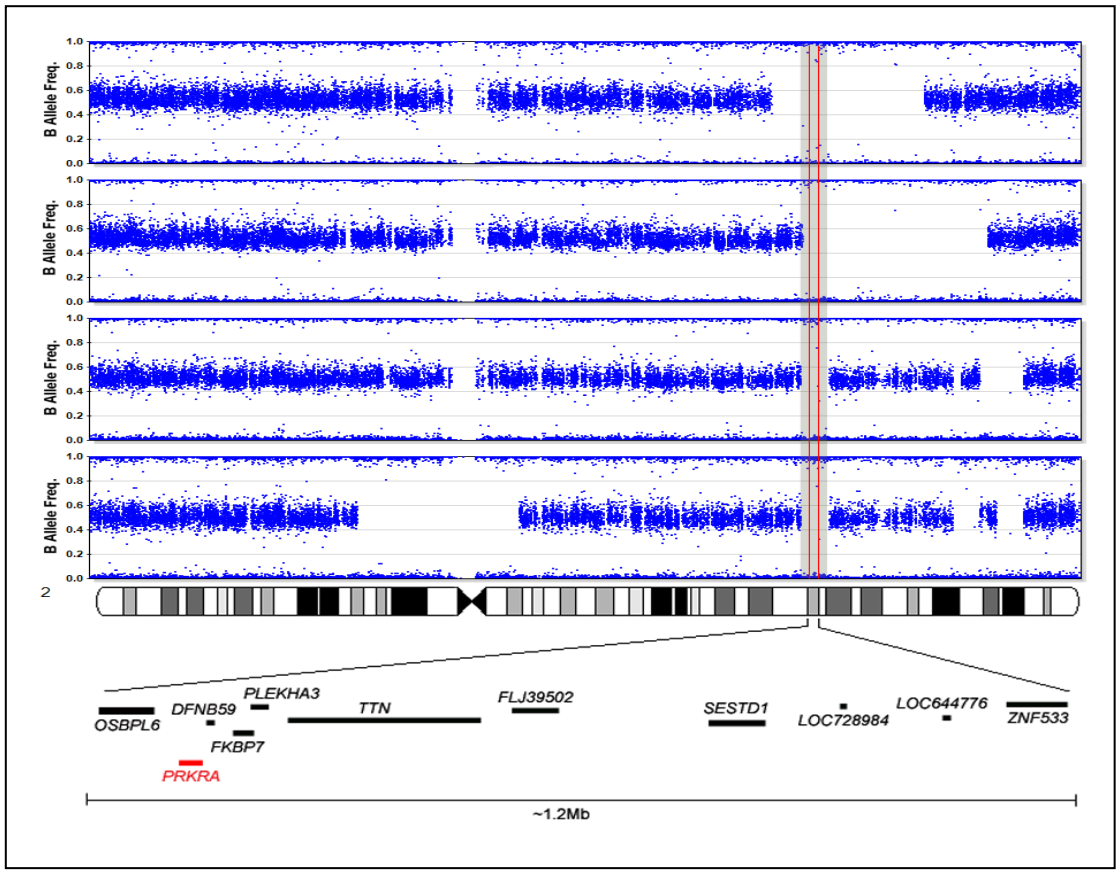
GW-SNP assays offer many advantages over microsatellite/RFLP methods to define recessive loci. In families with an apparently recessive disease, particularly one

where parental consanguinity is suspected, this approach can be used to map regions of extended homozygosity with high resolution (and essentially complete genomic coverage), and in a short amount of time. Moreover the data lends itself to immediate visualization of large homozygous tracts (Figure 1). Because all tracts of disease segregating homozygosity are identified and all heterozygous regions/nonsegregating homozygous tracts excluded, one can be confident that the region harboring the genetic lesion underlying disease has been identified - if the model is correct (i.e., the disease must be caused by a homozygous change inherited from a relatively recent common ancestor). Furthermore, this technique allows the direct visualization of structural genetic variations, such as genomic deletions or duplications (Figure 16 for an example).

This approach has been successfully applied to various neurodegenerative diseases, for example it allowed the identification of mutations in phospholipase A2, group VI (*PLA2G6*) in early onset parkinsonism dystonia (Paisan-Ruiz et al., 2009).

The main limitation of autozygosity mapping in gene identification for recessive traits is the size of the candidate fragments identified. Candidate gene sequencing by traditional Sanger sequencing can be time-consuming and expensive (Table 1).

Figure 1. Homozygous region identified using GW-SNP genotyping



Example of a disease-segregating homozygous region identified by homozygosity mapping using GW-SNP genotyping in families with dystonia-parkinsonism (genotyping performed using Illumina Infinium II HumanHap317 BeadChips). In the upper panel, B allele frequency (BAF) metrics across chromosome 2 from four affected individuals. BAF can be calculated as the proportion of the total allele signal (A + B) explained by a single allele (A). Stretches of homozygosity are denoted by a contiguous stretch of calls with BAF is equal to 1 (corresponding to A/A genotyping calls) or to 0 (corresponding to B/B) and by a lack of genotypes calls with BAF 0.5 (corresponding to A/B). Bounded in red is the primary candidate interval identical by state between all affected. The lower portion of the figure shows an ideogram of chromosome 2 and the genes in this primary critical interval.

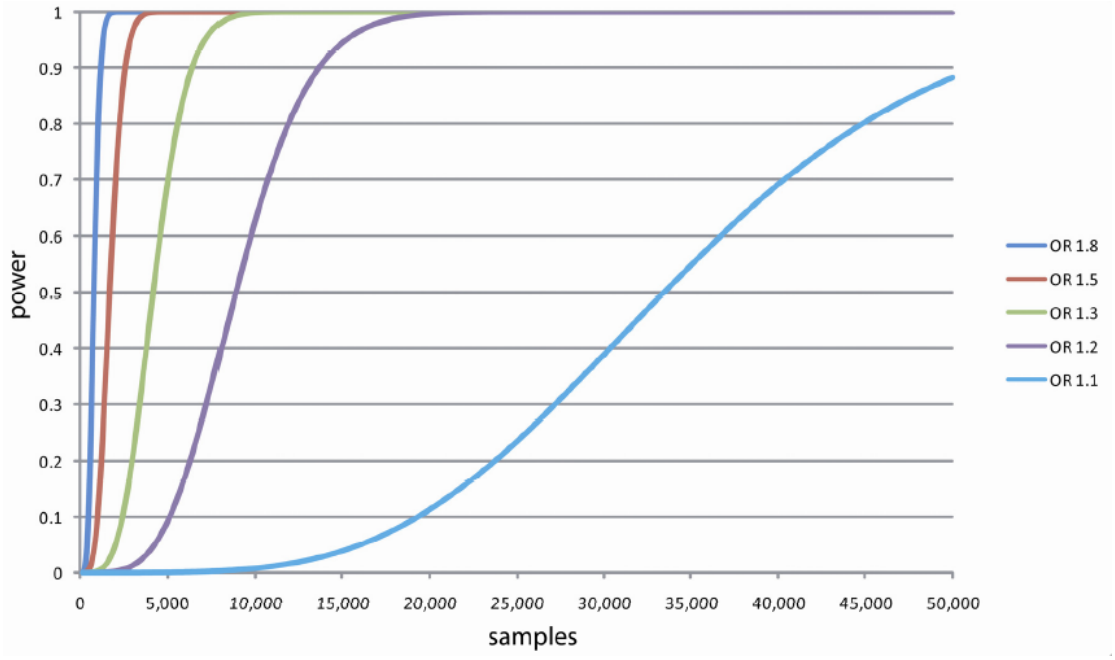
Genome-Wide association studies

This concept was first proposed by Neil Risch and Kathleen Merikangas as a mechanism to discover common disease-associated variants in complex disorders (Risch and Merikangas, 1996). GWAS involve scanning large case-control cohorts, using hundreds of thousands of SNP markers located throughout the human genome. Algorithms are applied that compare the frequencies of SNPs (alleles or genotypes) between disease and control cohorts. This analysis identifies loci with statistically significant differences in allele or genotype frequencies between cases and controls. The identification of susceptibility genes or variants depends on the existence of an association between the causal variants and surrounding SNPs nearby. Indeed, SNP markers used in GWAS are tagging-SNPs that cover common variation across the genome, targeting minor allele frequency (MAF) greater than 1-5%. The tagging SNP is a representative SNP in a region of the genome with high linkage disequilibrium (LD) (the non-random association of alleles at two or more loci). Therefore, the tagging SNP is representative of the alleles of nearby SNPs (i.e. in LD with it).

GWAS using SNPs test the hypothesis that common diseases are attributable to allelic variants present in more than 1–5% of the population. For these reasons large case-control cohort are needed: the number of required cases and controls is determined by the expected effect size of a genetic variant underlying the specific trait and by the minor allele frequency MAF of this variant (Figure 2).

GWAS have been shown to have a higher power than linkage studies to detect common variants with low effect size (Risch and Merikangas, 1996). In the last few years, several GWAS have been conducted in neurodegenerative diseases, leading to the identification of statistical association between many loci across the genome and the trait studied (see paragraph 1.4.1 for literature review).

Figure 2. Power Calculation for GWAS



This graph shows the relationship between power and sample size. For this simulation a minor allele frequency of 0.1 and a p value of 1×10^{-7}, based on the Bonferroni corrected significance threshold for testing 500,000 SNP markers, were chosen under an additive model. As for the sample size, an equally matched case-control cohort was assumed. For example, this plot shows that a GWA study with a sample size of 5,000 subjects (2,500 cases and 2,500 controls), testing 500,000 SNP markers in each individual, has an 80% power to detect variants with an odds ratio of 1.3.

Limitations of GWAS

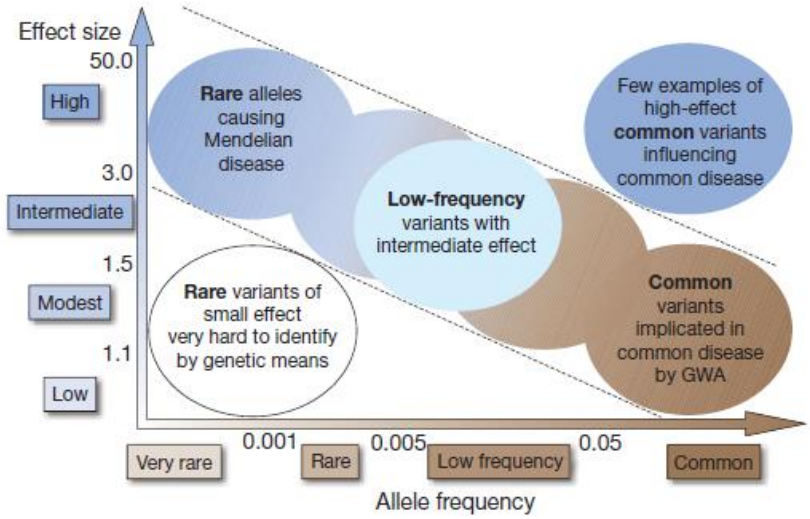
The results of GWAS have increased the understanding of molecular pathways underlying specific diseases, however they have some limitations and challenges. First, there is great difficulty finding the biological basis of the link between the locus and the complex trait. The statistical association of the GWAS is held by a tag-SNP, and does not pinpoint to the precise variant in the bin that have a causal role in the disease. A way of pinpointing the causal variant is to identify variants that are in LD with a tagging SNP and that are located in exons that truncate or alter the gene product. However, the causative variants underlying GWAS can be regulatory. For these reasons, many studies are being conducted with the aim to map genetic factors underlying differences in expression levels by assessing simultaneously global gene expression and genome wide variation (Cookson et al., 2009).

Second, SNP associations identified in one population frequently are not transferable to other populations. This is due to allele frequency differences between populations (Myles et al., 2008). For instance the MAPT H2 haplotype, reported to be almost exclusively of Caucasian origins, is rare in most other populations. This can explain the different association signal at GWAS (as discussed in paragraph 1.4.1) (Satake et al., 2009), (Simón-Sánchez et al., 2009). Different genetic markers should be used when investigating different populations because some may not be relevant to all populations.

Third, the bulk of the heritable fraction of complex traits has not been accounted for in recent GWAS. For example, despite the discovery of common low-risk susceptibility variants associated with sporadic PD, only a proportion of estimated heritability has been found (Hindorff et al., 2009). Common low risk loci identified in PD by GWAS such as *SNCA*, *MAPT* and *HLA* explain about 10% of risk each (Simón-Sánchez et al., 2009), (Hamza et al., 2010). Many factors have been suggested to explain the 'missing heritability'. Genetic risk factors identified by GWAS are common, low risk variants. Rarer variants (possibly with larger effects) poorly captured by existing arrays (which focus mainly on variants whose MAF >5%) and that do not carry sufficiently large effect-size are yet to be found. When MAF is < 0.5% detection of association becomes unlikely, unless effect sizes are very large as in monogenic conditions.

Another hypothesis on the genetic cause of complex disorders is the multiple rare variant hypothesis (Pritchard, 2001), where multiple rare high risk variants contribute to complex diseases. Unlike common variants, low-frequency alleles are not fixed in the population, and thus, there has been little selective pressure to limit the effect size of such variants (Figure 3). Rare, alleles with large effect can be hard to detect by GWAS employing common SNPs. This is well documented by the glucocerebrosidase (*GBA*) risk variants in PD. Heterozygous mutations in *GBA* are the strongest genetic risk factor for PD, yet the *GBA* locus is not detected by any GWAS, presumably because its risk variants are rare and hard to detect by using arrays representing common SNPs. So far multiple rare *GBA* variants could only be identified through comprehensive resequencing of the gene. These data show that resequencing strategies can be very powerful in identifying rarer variants.

Figure 3. Schematic genetic architecture of disease



This figure has been published elsewhere (Manolio et al., 2009). The genetic bases of disease in modern humans reflect the architecture and evolution of the human genome. The common variants implicated in diseases, such as those identified in GWAS, are likely to exert little negative selective pressure, either because they have quite small biological effect, or because the variants are associated with post-reproductive diseases. In contrast, low-frequency alleles are not fixed in the population, and thus, there has been little opportunity for selective pressure to limit the effect size of such variants.

1.3 WHOLE EXOME SEQUENCING

1.3.1 Overview: next generation sequencing technologies

Detailed study of all human genetic variants is required to elucidate the complete genetic architecture of human disease (Figure 3). This is achievable only by direct DNA sequencing. Until recently, sequencing was limited by cost and time (Table 1). In the last few years several approaches have been developed to DNA sequencing: the so called Next Generation Sequencing (NGS) technologies, making it feasible to sequence large amount of DNA in a fast and accurate way. NGS increase sequencing and reduce cost by processing in parallel hundreds of millions of DNA templates. NGS technologies provide enormous increases in speed and quantity of generated data, free of cloning biases and arduous sample preparation.

The hallmark of this technology is parallel sequencing, with determination of DNA sequence by iterative cycles of nucleotide extensions done in parallel on massive numbers of clonally amplified template molecules. Three platforms for next-generation DNA sequencing read production are most commonly used: the Roche/454 FLX (Margulies et al., 2005), the Illumina Genome Analyzer (GA) or HiSeq (Bentley et al., 2008), and Applied Biosystem SOLiD Platform (Shendure et al., 2005). They all use a reaction chamber (eg. the flow cell for Illumina GA) where the sequencing happens, but the specifics of template preparation, sequencing chemistry and the reaction chamber configuration differ among the platforms. Given the variety of NGS features, some platforms tend to have clear advantages for a particular application over the others. The Illumina technology has been widely used in human genetics and is the one we have used (chapter 2).

The development of methods for coupling targeted capture and massively parallel DNA sequencing, has made possible to determine all coding variation present in an individual genome in a single experiment, a process called 'whole exome sequencing' (WES). This technology has already become widely used in human genetics for the identification of genes that underlie Mendelian disorders.

Table 1. Comparison of sequencing features: Sanger sequencing versus NGS

Sequencing	*Sanger Sequencing (ABI 3730)*	*Next Generation Sequencing (Illumina GAIIx)*
Processing method	Individual reactions	Cloned single-molecule array (custers)
Throughput	.0001 Gb / run	50-95 Gb / run
Read Length	1000 base	100-150 base paired ends
Run time	1-2 hours	10 days
Cost	$100,000 / Gb	$500 / Gb

1.3.2 Whole exome sequencing: premises

The exome consists of all the protein coding parts of the genome. Several considerations motivated implementation of robust approaches to sequencing the whole exome. First, the exome harbours the coding variation, which is responsible for most Mendelian diseases. This is based on the observation that mutations that cause Mendelian diseases occur primarily in genes (http://www.hgmd.cf.ac.uk/ac/index.php); less than 1% of Mendelian disease mutations have been found in regulatory regions (Stenson et al., 2009). Second, mutations that cause amino acid substitutions (including changes to nonsense codons) are the most frequent type of disease causing mutation (Botstein and Risch, 2003). WES provides researchers with a powerful tool for gene discovery in an accurate and fast way.

Firstly because potentially all causal variants are identified, WES can be applied to Mendelian disorders where linkage analysis has failed (for instance in pedigrees that are too small to provide meaningful information using linkage) effectively allowing small families and even single probands to be analyzed jointly (Ng et al., 2010), or in consanguineous families with very large regions of homozygosity, where Sanger sequencing is a costly, time consuming process.

Secondly, WES has particular potential in diseases caused by germline or de-novo mutations, that were particularly resistant to previous methods for gene localization and identification. The identification of these kind of mutations could be achieved with complementary approaches: one is to sequence a cohort of cases suspected to have disease-causing de novo or non-inherited mutations and look for a gene that is commonly mutated; the other is to sequence trios (parents-affected child) to establish the variant(s) that occur only in the affected child.

Thirdly, WES can be used in complex diseases to identify variability in protein coding regions that alters risk for disease, such as *GBA* or *LRRK2* in Parkinson's disease (Tan, 2006), (Sidransky et al., 2009). WES enables researchers to identify rare high-risk alleles, with suitable statistical power and study design. This approach centers on an analytical design that involves assessment of the collective burden of multiple risk alleles at a locus.

1.3.3 The technology

The basic steps required for WES are: 1) DNA library preparation, 2) cluster generation, 3) sequencing and 4) data analysis.

1) DNA library preparation. The first step is to prepare a "library" comprising DNA fragments corresponding to exons ligated at both ends to universal oligonucleotide adapters (Figure 4a). Key performance parameters include the degree of enrichment, the uniformity with which targets are captured and the molecular complexity of the enriched library. Currently, there are three commercial exome capture platforms: NimbleGen SeqCap EZ, Agilent' s SureSelect and Illumina TruSeq. Each platform uses biotinylated oligonucleotide baits complementary to the exome targets to hybridize sequencing libraries prepared from fragmented genomic DNA. These bound libraries are enriched for targeted regions by pull-down with magnetic streptavidin beads. The capture technologies differ in their target choice, bait lengths, bait density and molecule used for capture (DNA for Nimblegen and Illumina, and RNA for Agilent). As for the target region, each platform targets particular exonic segments based on combinations of the available databases that catalogue mRNA coding sequences [for example based on Refseq (Pruitt et al., 2012), UCSC KnownGenes (Hsu et al., 2006)]. A large number of bases (29.45 Mb) are targeted by all three platforms (represented mainly by mRNA coding exons in both RefSeq and CCDS). The major difference is that Illumina Truseq captures also untranslated regions (UTRs).

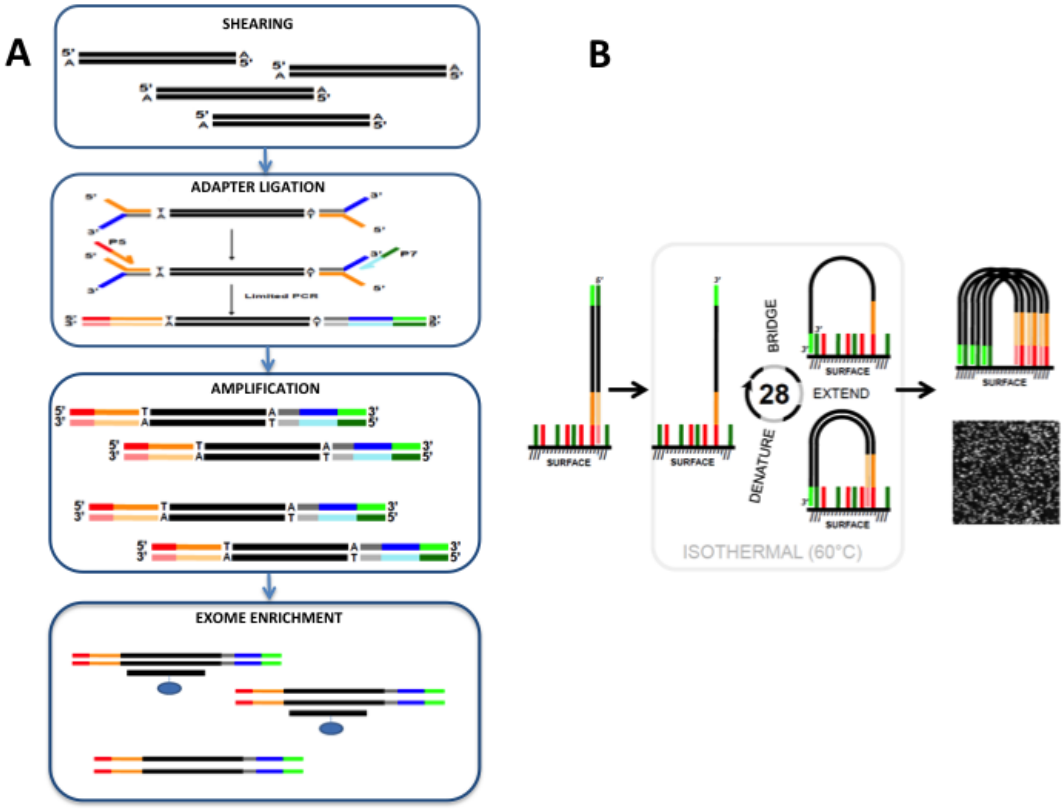
2) Cluster generation. The DNA fragments (templates) are immobilized to the surface of the flow cell. This allows billions of sequencing reactions to be performed simultaneously. The template is amplified ("bridge amplification"), because the imaging system during sequencing cannot detect single fluorescent events (Figure 4b).

3) Sequencing. The Illumina machine works on the principle of 'sequencing by synthesis' to produce sequence reads of ~75-100 bp from billions of clusters simultaneously (Figure 5). At the end of the sequencing run, the sequence of each cluster is subjected to quality filtering to eliminate low-quality reads.

4) Bioinformatic analysis. Preliminary analysis includes tools to convert intensity data into sequence reads and quality scores. Details of the process are discussed in the chapter 2.

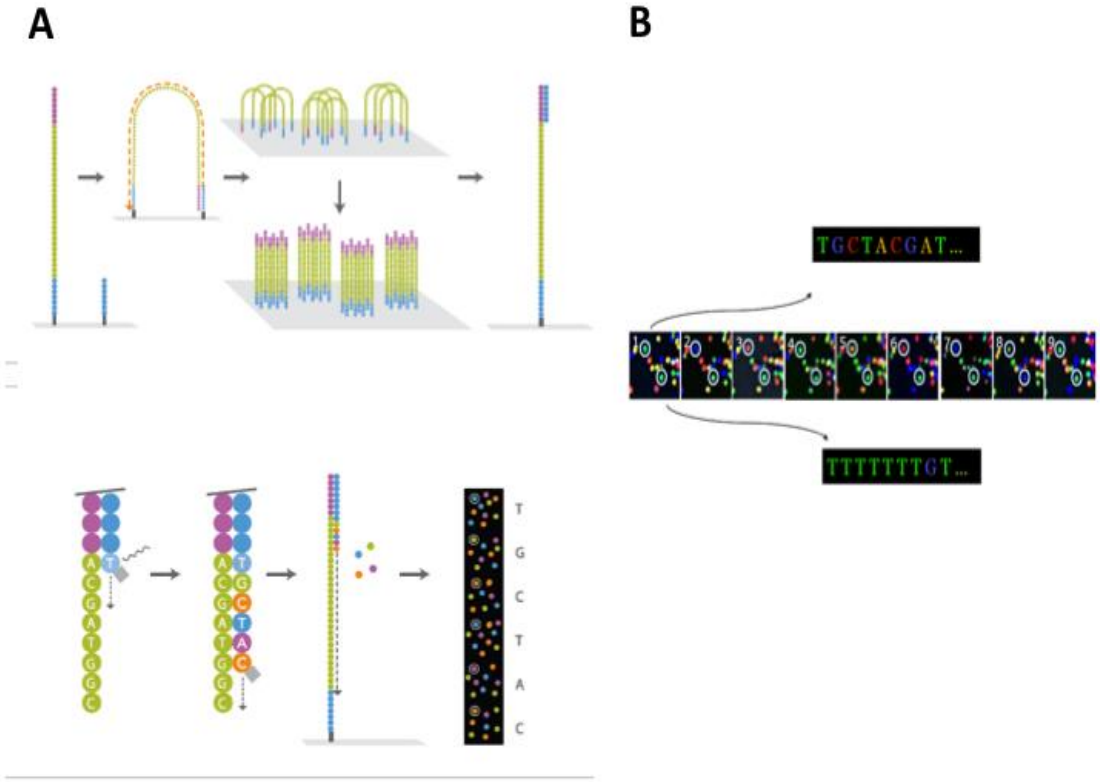
Ultimately the bioinformatic analysis produces a list of all allelic variants present in the exome sequenced.

Figure 4. WES technology



This figure shows the basic steps required for WES.
(A) Library preparation. Genomic DNA is randomly fragmented, and terminal overhangs are repaired. Then universal oligonucleotide adapters are ligated at both ends of DNA fragments. This DNA library is amplified using primers complimentary to adapters. Then, the library is enriched for fragments corresponding to the exons by hybridization capture in an aqueous-phase: the DNA fragments are hybridized to biotinylated DNA or RNA baits.
(B) Cluster generation. In a specific machine (the microfluidic cluster station) the DNA fragments are immobilized at one end on the flow cell surface. The interior surfaces of the flow cell have covalently attached oligos complementary to the universal adapters. Then they are incubated with reagents and a polymerase resulting in the amplification of the fragments in a discrete area or 'cluster' on the flow cell surfaces. This process (also called "bridge amplification") creates millions of clusters on the flow cell surface, each containing about 1000 copies of the same template. Bridge amplification produces millions clonally amplified clusters, providing free ends to which a universal sequencing primer can be hybridized to initiate the NGS sequencing.
Parts of this figure have been kindly made available by Dr. Harold Swerdlow, during the Wellcome Trust advanced Course "Next Generation Sequencing", July 2010.

Figure 5. WES Sequencing process



(A) Sequencing by synthesis. The flow cell is placed into the sequencer, in which each cluster is supplied with polymerase and four differentially labeled fluorescent nucleotides that have their 3'-OH chemically inactivated, to ensure that only a single base is incorporated per cycle. Each base incorporation cycle is followed by an imaging step to identify the incorporated nucleotide at each cluster and by a chemical step that removes the fluorescent group and deblocks the 3' end for the next base incorporation cycle. This series of steps continues for n cycles, which permits a discrete read length of n bp (maximum 150bp for the current Illumina technology) or 100 bp.
(B) Imagine and base calling. The first cycle of sequencing consists first of the incorporation of a single fluorescent nucleotide, followed by high resolution imaging of the flow cell. This image represents the data collected for the first base. Any signal above background identifies the physical location of a cluster, and the fluorescent emission identifies which of the four bases was incorporated at that position.
This cycle is repeated, one base at a time, generating a series of images each representing a single base extension at a specific cluster. Base calls are derived with an algorithm that identifies the emission color over time. At this time reports of useful Illumina reads range from 75-100 bases.
Parts of this figure have been kindly made available by Dr. Harold Swerdlow, during the Wellcome Trust advanced Course "Next Generation Sequencing", July 2010.

1.3.4 How to identify the causal variants in Mendelian disorders

Exome sequencing of a DNA sample from a single individual reveals on average 24,000 single nucleotide variants (SNV) in African American samples and ~20,000 in European American samples (Bamshad et al., 2011). ~10,000 variants are non-synonymous (lead to differences in protein sequence) and ~11,000 are synonymous. A number of variants are likely to have greater functional impact: 80-100 nonsense variants (premature stop codons), 40-50 splice site and 200 inframe indels. Of note, it is estimated that each genome is heterozygous for 50-100 variants classified by the Human Gene Mutation Database (HGMD) as causing inherited disorders (Durbin et al., 2010).

A key challenge of using WES in human genetics is how to identify disease-related alleles among the background of benign variants and sequencing errors. The 1000 Genomes Project has provided a comprehensive resource on human genetic variation by sequencing the genomes of about 2500 individuals (Via et al., 2010) and has tremendously helped defining the pattern of genetic variation in human population for rare variants (<5% MAF).

When WES is used for gene discovery in Mendelian disorders, fundamentally two assumptions are made: 1) that the mutation is rare or novel, 2) that it is protein changing (such as nonsense, splice or missense).

Different approaches are applied for identifying causal variants in exome sequencing data: discrete filtering, stratifying variants on the bases of their functional impact and the use of pedigree information.

Discrete filtering. A common step, when looking for extremely rare mutations, involves filtering for novelty, i.e. against variants present in the general population (as these variants are unlikely to cause disease). The most used reference panel is the data derived from the 1000 Genomes Project. This approach is very powerful, as only about 2% of the variants identified in an individual are novel (i.e. not previously reported). Thus sequencing only a modest number of affected individuals, then applying this discrete filtering to the data, can be very powerful for identifying new genes for Mendelian disorders. Nonetheless this approach can be problematic, as it assumes that the filter set does not contain alleles from

individuals with the phenotype being studied. This needs to be accounted for future projects, as more sequence data will accumulate in the public domain, but with this the likelihood of erroneously filtering out a disease-causing mutation increases. This error is less likely in very young-onset severe diseases, but probably quite likely in later-onset diseases (who are too young to express disease), diseases with benign phenotypes, or those with mutations that have a low penetrance (mutation carriers who do not show symptoms). Furthermore, most reference data do not include very detailed phenotypic data, so patients with, or who will get, common diseases are likely to be included but not identified.

A more reliable approach to filtering is to do it based on MAF reported in the 1000 Genomes project. For recessive disorders, filtering for variants with a maximum MAF of 1% is still a well-powered approach. A lower MAF cutoff of 0.1% can be used for dominant disorders, as the estimated prevalence of the disorder provides an upper boundary on the MAF (Bamshad et al., 2011).

A very useful panel is the National Heart, Lung, and Blood Institute (NHLBI) exome sequencing project (ESP, (Exome Variant Server, NHLBI Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) [January 2011], which includes exome data for 3,500 American individuals of European descent and 1,850 African American. Simply by looking at allele frequencies in this dataset can dramatically help distinguish between benign variants, sequencing artifacts (e.g. usually those occur at very high frequency privately to this dataset) or true highly penetrant, pathogenic variants. Both in chapter 4 and chapter 6 I used this data to define the role of variants we were interested in.

Deleterious nature of variants. Candidate variants can be further stratified on the basis of their predicted impact or deleteriousness on protein structure and function. Alleles can be stratified by their functional class by giving greater weight to stop codons, frameshifts and disruptions of canonical splice sites than to missense variants. However, this is an oversimplification that is insensitive to causal alleles that do not directly alter protein-coding sequences or canonical splice sites. Additionally, candidate alleles can be stratified by existing biological or functional information about a gene: for example, its predicted role (or roles) in a biological

pathway or its interactions with genes or proteins that are known to cause a similar phenotype.

Use of pedigree information and mode of inheritance

For Mendelian disorders, the use of pedigree information can substantially narrow the list of candidate causal alleles. When mapping data is available, the best strategy is to sequence a pair of individuals whose overlapping haplotype produces the smallest genomic region. In consanguineous pedigrees where a recessive mode of inheritance is suspected, sequencing the individual with the smallest region(s) of homozygosity, as determined by the genome-wide genotyping data, can be sufficient (Walsh et al., 2010). In the absence of mapping data, sequencing the two most distantly related affected individuals can substantially restrict the genomic search space.

WES can also be highly effective for identifying de novo coding mutations by sequencing parent-child trios, as multiple de novo events occurring within a specific gene (or within a gene family or pathway) is an extremely unlikely event (Xu et al., 2011).

The approach to reducing the size of a candidate list of variants to define the causal variants has proven successful in many diseases, as in the identification of mutations that cause Parkinson's disease [*VPS35* for example (Vilariño-Güell et al., 2011), (Zimprich et al., 2011)] .

Table 2. Mean number of coding variants per exome

Variant type	*Mean number of variants In African Americans*	*Mean number of variants In European American*
TOTAL VARIANTS	24,049	20,283
Missense	11,131	9,511
Nonsense	103	93
Splice	38	34
Synonymous	12,776	10,645
NOVEL VARIANTS (TOTAL)	(520)	(307)
Missense	303	192
Nonsense	5	2
Splice	2	2
Synonymous	209	109
KNOWN VARIANTS (TOTAL)	(23,529)	(19,976)
Missense	10,828	9,319
Nonsense	98	89
Splice	36	32
Synonymous	12,567	10,536

Parts of this table have been published elsewhere (Bamshad et al., 2011). The table lists the mean number of coding single nucleotide variants from 100 sampled African Americans and 100 European Americans.

1.3.5 How is WES changing neurogenetics

WES has already proven in few years its worth in neurogenetics for different applications: 1) genetic diagnosis and screening; 2) further defining and expanding the phenotypes associated with mutations; 3) gene identification in Mendelian disorders; 4) identification of de novo mutations. I will briefly discuss each application.

1. **Genetic diagnosis and screening**. Many neurological diseases, such as neuropathies and ataxia, are genetically heterogeneous. Current screenings are often designed to catch only the most common mutations or alterations in the most frequently mutated genes. The identification of mutations by WES is cheaper and quicker than it is by traditional sequencing approaches. Recently, two papers have shown the power of WES in this situation both in neuropathy and ataxia (Montenegro et al., 2011), (Sailer et al., 2012). Montenegro and colleagues studied a family with inherited neuropathy. Instead of screening the all the genes known to cause this disease (35 genes), the investigators used WES in two affected family members and identified a *GJB1* mutation, previously associated to X-linked neuropathy. This case also shows the power of WES in genetic diagnosis as the *GJB1* mutation would have been ruled out for screening because of a reported male-to male inheritance, incompatible with the transmission of a mutation that lies on chromosome X. Sailer and colleagues also used WES in a family a large, 5-generational British kindred with progressive cerebellar ataxia where conventional genetic testing had not revealed a causal aetiology. By sequencing one affected member they identified a novel p.Arg26Gly change in the *PRKCG* gene, known to cause spinocerebellar ataxia 14. In this case WES data enabled the identification of a novel mutation in a gene known to cause the phenotype and to provided a genetic diagnosis in a fast and cheap way.

2. **Expanding the phenotype associated with mutations**. WES has provided researchers with a powerful tool to identify mutations in genes previously associated to different disease phenotype or pathology.

For example mutations in the *VCP* gene, previously linked to Paget's disease, inclusion body myopathy and frontotemporal dementia have been shown also to

cause amyotrophic lateral sclerosis (ALS) in one of the first study involving WES in neurodegenerative diseases (Johnson et al., 2010). Of note, this group showed that *VCP* mutations substantially contribute to the cause of familial ALS, being responsible for about 2% of cases. This finding broadened the clinical and pathological phenotype of *VCP* mutations to include ALS.

Recently, another group was able to identify mutations in *ATP13A2*, a gene known to cause a form of dystonia-parkinsonism (Kufor-Rakeb syndrome, KRS), in a family with neuronal ceroid lipofuscinosis (NCL). NCL is part of a heterogeneous group of inherited progressive degenerative diseases of the brain and sometimes the retina, that are characterized by lysosomal accumulation of auto fluorescent lipopigment. The relationship between the diseases was not obvious, as the clinical features do not appear to overlap significantly. KRS typically presents with rigidity, bradykinesia, spasticity, supranuclear upgaze paresis and dementia. NCL disease varies according to the underlying gene defect and severity of the mutation, but typically includes seizures, a progressive intellectual and motor deterioration, and in children but usually not adult onset cases, visual failure.

These results indicate that broadening the phenotype associated with mutations provides information on the aetiological basis of disorders by uniting what is known about the biological underpinnings of apparently unrelated disorders into a single model. The work of Bras and colleagues also shows that KRS is indeed linked to the lysosomal pathway, a feature that was already hypothesized for a variety of parkinsonian phenotypes, but was not previously shown for KRS.

3**. Gene identification in Mendelian disorders**. Despite still being a new technology WES has already led to the identification of novel genes that cause rare, familial forms of neurodegenerative disease. Thanks to the comprehensive discovery of protein-coding variants throughout the genome, DNA samples collected from small families and isolated affected individuals, can be used to discover mutations that cause disease. Two independent studies utilized WES in a Swiss and an Austrian kindred to identify the same p.D620N (c.1858G>A) mutation in the vacuolar protein sorting 35 homolog (*VPS35*) gene as the cause of autosomal dominant Parkinson disease in these families (Zimprich et al., 2011), (Vilariño-Güell et al., 2011). This mutation was subsequently found in several additional families and is a rare cause

of Parkinson disease, with a frequency lower than 0.1%. The p.D620N mutation co-segregates with a phenotype indistinguishable from idiopathic Parkinson disease and has incomplete, age-associated penetrance.

WES has also been used to identify novel mutations in heterogeneous neurological diseases, such as trans-2,3-enoyl-CoA reductase (*TECR*) mutations in nonsyndromic mental retardation (Çalışkan et al., 2011), WD repeat domain 62 (*WDR62*) mutations in severe brain development malformations (Bilgüvar et al., 2010), TGM6 transglutaminase 6 (*TGM6)* mutations in ataxia, and a Werner syndrome, RecQ helicase-like (*WRN*) mutation in atypical Werner's syndrome (Raffan et al., 2011).

4. **De novo mutations**. In three different studies, researchers used WES to show that de novo mutations are associated with risk of autism (Sanders et al., 2012), (O'Roak et al., 2012) and (Neale et al., 2012). In the first study, Sanders et al. performed WES of 238 families from a group of families in which one child has a non-inherited autism spectrum disorder (ASD) and compared the exomes of the affected patients to that of their unaffected siblings. They found the rate of de novo SNVs expressed in brain to be higher in individuals with ASD.

O'Roak et al. performed WES of 189 trios (parents and their child with autism) and found 248 de novo autism-associated mutations, 120 of which were designated as severe (that is, the mutation affected protein function). Of interest, 39% of the most disruptive genes were part of one highly interconnected protein network, related to β-catenin and chromatin remodeling.

In the third study, Neale et al. sequenced whole exomes of 175 children with ASD and their parents, and found the rate of de novo mutations in patients with autism to be slightly higher than in unaffected controls. In agreement with the findings of O'Roak et al., proteins that were affected by structural mutations seemed to be more connected with one another in terms of function than would be expected by chance.

Most of the autism-related de novo mutations identified in the three studies were found in independent genes, indicating the genetic heterogeneity of ASDs.

These studies confirmed the potential of WES to identify 1) de novo mutations, that were particularly resistant to previous efforts aimed at gene localization and 2) mutations in genetically heterogeneous disorders.

5. **Complex diseases**. In Alzheimer's Disease (AD), two groups have recently identified a rare variant in the *TREM2* gene associated with susceptibility to the disease, with a odds ratio of ~3 (Guerreiro et al., 2013). Researchers first nominated *TREM2* as a candidate gene, following the discovery of homozygous *TREM2* mutations as a cause of Nasu-Hakola disease, a rare recessive form of dementia with leukoencephalopathy and bone cysts. They generated exome sequence data sets and identified the R47H variant, which associated with the disease in cohorts from North America and Europe. This work was confirmed by researchers at deCODE Genetics, who separately identified the R47H variant in a GWAS using the Icelandic population and replicated the association with AD in North American and European cohorts. TREM2 is an immune phagocytic receptor expressed in brain microglia. These studies suggest that reduced function of TREM2 causes reduced phagocytic clearance of amyloid proteins or cellular debris and thus impairs a protective mechanism in the brain, assuming that the risk variants impair TREM2 function.

1.3.6 Limitations and challenges

Despite the great success of WES human genetics, there are still significant limitations associated with this technology. WES can fail to identify variants for many reasons, mainly due to design and technical reasons or analytical pitfalls.

Some genes may not be in the target definition or there is a failure in the design. Of note the exome is defined by CCDS, a collaborative effort to identify a core set of protein coding regions that are consistently annotated and of high quality (http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi). WES using CCDS can result in true coding variation being missed; a more comprehensive annotation such as GENCODE (Harrow et al., 2006) provides an alternative that some companies are adopting. Nonetheless both CCDS and GENCODE exclude non-coding regions, some of which are known to contribute to disease mechanism. For example, an intronic hexanucleotide repeat in chromosome 9 open reading frame 72 (*C9orf72*) was recently identified as the cause of ALS and frontotemporal dementia (DeJesus-Hernandez et al., 2011), (Renton et al., 2011). On the other hand, a gene may be in

the target region but not adequately captured or sequenced (for example because it is too GC rich, which can cause difficulty during sample preparation) (chapter 3 as an example).

Analytical limitations can be attributed to misalignment or inadequate variant calling algorithm. For example, repetitive regions and homologous sequences may mismap to the reference genome, generating false variants. This is the case of the glucosidase, beta, acid gene (*GBA*), that has a pseudogene with ~ 96% homology in the same genomic region. The existence of this similarity complicates the determination of the source of DNA sequenced fragments during alignment. Similarly, polyglutamine-type diseases are difficult to study by WES as the underlying defect is a repetitive sequence.

As for inadequate variant calling algorithms, WES does not currently accurately calls large indels. For example, WES failed to identify the genetic defect of a simple Mendelian disorder (medullary cystic kidney disease type 1), since it is caused by a repeat unit comprising the extremely long (~1.5-5 kb), GC-rich (>80%) coding variable-number tandem repeat sequence in the *MUC1* gene encoding mucin 1 (Kirby et al., 2013).

Improvements in the technology and wider target will overcome some of these limitations.

1.4 GENETICS OF SELECTED NEURODEGENERATIVE DISEASES

In the following paragraphs I will discuss about the genetic background of some neurodegenerative diseases that are relevant to this thesis.

1.4.1 Genetics of Parkinson's Disease

Parkinson's Disease (PD, OMIM #168600) was first documented in 1817 in *An Essay on the Shaking Palsy* by James Parkinson. This publication contained the details of tremors occurring in elderly people and gave a description of the shaking palsy as a mixture of tremor, bradikynesia and gait disturbance. The term "Parkinson's disease" was coined later in the 1850s by Jean-Martin Charcot, who also added rigidity to the signs identified by Parkinson. Because those typical PD clinical features have subsequently been found to associate with other neurological disorders, the term "parkinsonism" was introduced to describe the syndrome characterized by akinesia accompanied by tremor or rigidity or postural/gait disturbance. To date many causes of parkinsonism have been identified (Dick et al., 2007). All these disorders are believed to manifest parkinsonian motor features because of a disruption of the nigrostriatal dopaminergic pathway.

PD is by far the most common form of parkinsonism, constituting about 75% of cases seen in large movement disorders centers (Colcher and Simuni, 1999) and is the second most common neurodegenerative disease after Alzheimer Disease with a prevalence of about 1% over 60 years of age (de Lau and Breteler, 2006). The main clinical features of PD are bradykinesia, resting tremor, rigor, postural instability and marked response to levo-dopa (L-DOPA). Pathologically, PD is characterized by two key events: the loss of pigmented neurons in the substantia nigra (SN) and presence of intracytoplasmatic, ubiquitin-positive inclusions that contain α synuclein in surviving neurons (also referred to as Lewy Bodies) (Braak and Del Tredici, 2008).

PD is a complex disorder whose etiology is still largely unknown. Age is the main risk factor for PD. There has been a considerable debate about the contribution of environmental versus genetic risk factors. Environmental risk factors, including

exposure to pesticides and metals, drugs (1-methyl 4-phenyl 1,2,3,6-tetrahydropyridine [MPTP]), viruses, rural living and farming have been investigated in different epidemiological studies. Those studies have shown a tenuous link between environmental factors and PD (Lees et al., 2009) and true single causes are probably restricted to a small number of poisonings. The genetic contribution to the pathogenesis has been investigated through twin studies –where discordancy between twins was interpreted as evidence against a genetic etiology of disease. Several twins studies showed low concordance rates in monozygotic and dizygotic twins (Ward et al., 1983), (Duvoisin et al., 1981). These results were controversial, since they were cross-sectional studies that did not exclude the possibility of a later disease onset in unaffected siblings. This obstacle has been partly overcome by functional imaging of the brain (positron emission tomography studies [PET]), which is sufficiently sensitive to identify decreased function in the nigrostriatal dopaminergic system. Based on PET scan data, the concordance rate becomes significantly higher for monozygotic twins compared to dizygotic twins (55% versus 18%), suggesting a substantial genetic contribution to the PD pathogenesis (Piccini et al., 1999).

Of note, discordance has been seen even in twins with *LRRK2* mutations (Xiromerisiou et al., 2012), [known to be the most common Mendelian cause of PD (Zimprich et al., 2004)], suggesting considerable variance in the penetrance of this gene.

The role of heredity in the etiology of PD is supported by the discovery of rare, Mendelian forms of disease. The mode of inheritance can be recessive or dominant, often with incomplete penetrance. In the last two decades, important advances have been made in the understanding of genetic contribution to PD owing to the identification of genes responsible for monogenic PD (Table 3). Although at the beginning the difference in clinical presentation and pathology were emphasized, it is becoming clear that genes firstly implicated in the Mendelian forms of disease also play a role in the sporadic disorder (Ahn et al., 2008), (Simón-Sánchez et al., 2011a).

Familial forms of PD related to PARK loci are characterized by the presence of parkinsonism (Table 3). Typical, late-onset Parkinson's disease is linked to

heterozygous mutations in three genes: *SNCA* (encoding α-synuclein) (Polymeropoulos et al., 1997), *LRRK2* (encoding leucine-rich repeat kinase 2) (Kitada et al., 1998) and *VPS35* (Vilariño-Güell et al., 2011). Recently, linkage and candidate gene analysis reported a mutation (the p.R1205H) in *EIF4G1*, encoding a component of the eIF4F translation initiation complex that regulates cell survival in response to stressor, in a large French family with typical late-onset PD and 7 smaller families of various origin, probably resulting from a founder effect (Chartier-Harlin et al., 2011). Screening additional patients with parkinsonism identified one less frequent putatively disease-causing mutations, p.A502V, absent from approximately 4000 controls. However, their involvement in disease pathogenesis remains inconclusive, in the absence of segregation. I will describe the first genetic analysis carried out to clarify the role of this gene in PD in chapter 6.

Mutations linked to early-onset Parkinson's disease are found in affected individuals with disease onset under the age of 45 years. They are recessive mutations in the genes encoding *PARK2* (Kitada et al., 1998), *PINK1* (Valente et al., 2004) and *DJ-1* (Bonifati et al., 2003). Some families have more complex phenotypes (atypical forms of parkinsonism) and show Lewy body pathology (eg. PARK14) (Paisan-Ruiz et al., 2009). They might therefore share with PD common pathophysiological mechanisms. Mutations in PARK9 (*ATP13A2* gene), PARK14 (*PLA2G6*) and PARK15 (*FBX07*) have been described in this kind of parkinsonism.

Table 3. Monogenic loci for Parkinson's disease

Locus	Gene symbol	Gene product	Mode of inheritance	Features	References
PARK1/PARK4	*SNCA*	α-synuclein	AD	Parkinsonism, dementia, autonomic dysfunction	(Polymeropoulos et al., 1997)
PARK2	*PARK2*	Parkin	AR	Juvenile- or young-onset parkinsonism, often with foot dystonia, slow progression, good response to L-dopa	(Kitada et al., 1998)
PARK6	*PINK1*	Pten-induced kinase 1	AR	early- onset parkinsonism, hyperreflexia, dystonia, slow progression	(Valente et al., 2004)
PARK7	*PARK7*	DJ1	AR	early- onset parkinsonism, psychiatric features, slow progression	(Bonifati et al., 2003)
PARK8	*LRRK2*	Leucine-rich repeat kinase 2	AD	mid-late onset parkinsonism, good response to L-dopa	(Paisán-Ruíz et al., 2004), (Zimprich et al., 2004)
PARK9	*ATP13A2*	ATPase type 13A2	AR	Causes Kufor-Rakeb syndrome; juvenile- onset parkinsonism, spasticity, hallucinations, dementia, supranuclear gaze paresis	(Ramirez et al., 2006)
PARK14	*PLA2G6*	Phospholipase A2, group VI	AR	L-DOPA responsive dystonia-parkinsonism	(Paisan-Ruiz et al., 2009)
PARK15	*FBX07*	F-box protein 7	AR	Early-onset parkinsonian-pyramidal syndrome	(Di Fonzo et al., 2009)
PARK17	*VPS35*	Vacuolar protein sorting 35 homolog	AD	Typical, late onset-PD	(Zimprich et al., 2011), (Vilariño-Güell et al., 2011)

AD = autosomal dominant; AR = autosomal recessive

Genetic risk in PD

In the last few years several GWAS have been conducted for PD in multiple populations (Maraganore et al., 2005), (Simón-Sánchez et al., 2009), (Edwards et al., 2010), (Hamza et al., 2010), (Spencer et al., 2011), (Saad et al., 2011), (Simón-Sánchez et al., 2011a), (Do et al., 2011), (Satake et al., 2009). These studies have identified a number of risk loci associated with PD that have been independently confirmed in follow-up studies.

The first GWAS published using GW-SNP genotyping were performed in 2009 by Simón-Sánchez (Caucasian population) and Satake (Japanese population). In the Caucasian study initial GWAS was performed in 1,820 patients with PD and 4,047 controls, and follow-up was conducted in an independent series comprising 3,452 patients with PD and 4,756 controls. Two loci, one located on chromosome 4q22 (where SNCA lies) and the other within a region encompassing the *MAPT* gene on chromosome 17q21.3, reached genome-wide significance.

The study conducted by Satake et al., consisting solely of Japanese participants identified several SNPs significantly associated with PD on 4q22, 12q12, 1q32 and 4p15. Chromosome 12q12 was tagged at loci proximal to the *LRRK2* gene, previously implicated to Mendelian PD (Table 3). The locus on chromosome 4q22 where the *SNCA* gene is located also reached genome-wide significance. By contrast, no association was observed between *MAPT* and PD, even though MAPT H1 variability contributes to PD risk in European samples H2, which is inversely associated with disease, does not occur in the Japanese population. This study also reported two novel genome-wide significant loci, one located on chromosome 1q32 (designed as PARK16), and the other on chromosome 4p15 (in the *BST1* locus). Replication of the PARK16 findings was obtained in the European study, but only a 1% difference was observed in the PARK16 minor allele frequency between cases and controls. By contrast Simón-Sánchez and colleagues did not find an association between the locus on 4p15 and PD risk. Subsequent genome-wide association studies in the Caucasian population have shown repeatedly that common variation in the loci encompassing the *SNCA* region and *MAPT* region are associated with PD (Hamza et al., 2010), (Spencer et al., 2011), (Simón-Sánchez et al., 2011b).

Additionally novel loci not previously implicated in PD have also been identified: 6p21.3, containing the *HLA* (human leukocyte antigen) region (Hamza et al., 2010), 4p16.3, containing the *GAK* locus (cyclin G associated kinase) (Latourelle et al., 2009) (Spencer et al., 2011) and 4p15.32, where *BST1* (bone marrow stromal cell antigen 1) is located (Saad et al., 2011)), confirming the initial Japanese study.

GWAS results have provided evidence that common genetic variation does play a role in the cause of Parkinson's disease, confirming the CDCV hypothesis. Moreover they highlight the commonality between monogenic Parkinson's disease and the late-onset sporadic Parkinson's disease both in terms of genetic basis and likely common pathogenic pathways. Highly penetrant mutations in *SNCA*, *MAPT* and *LRRK2* have been associated to autosomal dominant forms of parkinsonism. The strongly associated SNPs at the *SNCA* locus lie at the 3' end of *SNCA* gene, suggesting they may act on regulating expression, RNA stability or splicing. The signal observed in the *LRRK2* region, initially only found to be associated to PD in the Japanese population, might account for several *LRRK2*-associated coding polymorphisms, which have been recognized as contributors to PD susceptibility in the Asian population (Funayama et al., 2007). Further work suggests that in addition to protein-coding risk variants, there are low-risk, noncoding variants immediately 5' of *LRRK2* (Nalls et al., 2011).

The *MAPT* H1 haplotype has been reported to be associated with several neurodegenerative disorders named tauopathies (Pittman et al., 2006), and tau is known to be an aggregating protein in several neurodegenerative diseases, including Alzheimer's disease (Small and Duff, 2008). An association between tau and Parkinson's disease suggests that there may be cross talk between pathways that involve different aggregating proteins. Prior to GWAS findings, the importance of studying tau pathways in sporadic Parkinson's disease was not appreciate

Table 4. Risk loci associated to PD

Locus	Gene(s)	Map Position	Risk variants	Approx odds ratio	References
PARK1/PARK4	*SNCA*	4q21	REP1 repeat polymorphism, multiple SNPs in 3' half of gene	1.2-1.4	(Satake et al., 2009), (Simón-Sánchez et al., 2009), (Spencer et al., 2011), (Edwards et al., 2010)
PARK8	*LRRK2*	12q12	G2385R, R1628P	2.0-2.2	(Tan et al., 2010) (Mata et al., 2005)
Not assigned	*MAPT*	17q21.1	H1 haplotype	1.4	(Simón-Sánchez et al., 2009), (Spencer et al., 2011),(Edwards et al., 2010)
Not assigned	*GBA*	1q21	>300 mutations including N370S and L444P	5.4	(Sidransky et al., 2009)
PARK16	unknown	1q32	Multiple SNPs from GWASs	1.3	(Satake et al., 2009)(Nalls et al., 2011)
PARK17	*GAK, DGKQ*	4p16	Multiple SNPs from GWASs	1.5	(Pankratz et al., 2009),(Lill et al., 2012)
PARK18	HLA-DRA	6p21.3	Multiple SNPs from GWASs	1.3	(Hamza et al., 2010)
Not assigned	BST1	4p15	Multiple SNPs from GWASs	1.2	(Satake et al., 2009), (Saad et al., 2011)

1.4.2 Genetics of Kohlschütter-Tönz syndrome

Kohlschütter-Tönz syndrome (KTS, OMIM #226750), first described in 1974 by Kohlschütter (Kohlschütter et al., 1974) is a rare hereditary disorder characterized by seizures, spasticity, psychomotor delay or regression and the most distinctive feature of amelogenesis imperfecta. Onset of epilepsy is usually very early in life and seizures may be refractory to treatment; amelogenesis imperfecta affects both primary and secondary dentition: the enamel is thin and rough, and the teeth are yellow and prone to crumble. Developmental delay and spasticity occur usually in early childhood. Ataxia is also frequently reported. Cerebellum hypoplasia is frequent, and enlargement of lateral ventricles were seen in most of the affected subjects. The pregnancies are usually uneventful. Other minor symptoms include broad toes and thumbs, skull deformities, and muscle cramps. Because this syndrome is quite rare, the epidemiology remains largely unknown and the differential diagnosis is quite limited.

After the first report (Kohlschütter et al., 1974), KTS was established as a clinical entity and to date only 20 pedigrees with a total of 52 affected members have been reported in the literature, presenting with the typical combination of enamel defect tooth and neurological involvement (Christodoulou et al., 1988), (Donnai et al., 2005), (Guazzi et al., 1994), (Haberlandt et al., 2006), (Kohlschütter et al., 1974), (Musumeci et al., 1995), (Petermöller et al., 2008), (Schossig et al., 2012a), (Wygold et al., 1996), (Zlotogora et al., 1993), (Mory et al., 2012).

In the following section I will describe all the clinical features of KTS reported so far in literature. Chapter 3 describes the genetic analysis performed on KTS cases, including some of those that have previously reported.

Kohlschütter described this syndrome on 5 affected brothers in a farming family from a small valley in central Switzerland (Kohlschütter et al., 1974). The seizures onset in these 5 brothers was between 11 months and 4 years of age. Their teeth were yellowish with hypoplasia of the enamel. Significant mental deterioration was noticed as they grew older and none of them survived into their teens possibly because they were severely handicapped and prone to complications, and

adequate nursing was not available. In the following years, Kohlschütter observed another 2 Swiss families with similar clinical presentations. The seizure onset was between 11 month and 4 years of age in these 2 pedigrees. The neurological developmental milestones were normal before the seizure onset, but cognitive deterioration and spasticity gradually developed after the onset of intractable seizures. No female patients were reported in the three pedigrees Kohlschütter reported.

Later on, Christodoulou (Christodoulou et al. 1988) documented a Sicilian family comprising 4 male and 2 female affected with early onset seizures, spasticity, mental retardation or deterioration after onset of intractable seizures with amelogenesis imperfecta. The clinical presentations were similar to the kindred described by Kohlschütter. Besides the affected individuals, there were another 4 maternal female relatives who had mental retardation of unknown cause.

Zlotogora (Zlotogora et al. 1993) reported a Druze family with two affected individuals with intractable febrile seizures, spastic gait, increased deep tendon reflex, amelogenesis imperfecta with yellow teeth (amelogenesis imperfecta), and intellectual learning disability. The seizures onset was between 12 months and 3 years old. In this pedigree, the parents were unaffected and they had 4 children, only 2 younger children were affected. The parents were said to be first cousins.

A German pedigree with similar clinical presentation was reported by Petermöller (Petermöller et al. 1993). In this pedigree, there were two affected individuals. Their clinical presentations were similar to the kindred reported before and included febrile seizures and later afebrile seizures with jerks mainly on the left side, psychomotor deterioration, and amelogenesis imperfecta with yellow teeth. The parents were unaffected unrelated healthy individuals.

A variant of Kohlschütter-Tönz syndrome was reported by Guazzi in a Sicilian family (Guazzi et al. 1994). In this family, the clinical presentations were different from the pedigrees described above. The neurological deficit occurred in the second decade, and seizure onset was not in early infancy or childhood, and in some individuals the seizures were well controlled. Muscle cramps and pain were the first symptoms in one of the affected individual which started at the age of 10, with gradual psychomotor deterioration. Other symptoms in this individual were ataxia, yellow

teeth, hyperactive deep tendon reflexes, and degenerative muscular change of the calves in the lower limbs. This pedigree was quite complicated as there were 4 different phenotypes and the parents were first cousins. Some of the individuals in this pedigree had amelogenesis imperfecta only, some of them had abnormal neurological symptoms only, such as delay of psychomotor development or seizures only without amelogenesis imperfecta, and some of them had both but without signs of mental retardation. Although Guazzi concluded that the inheritance pattern of this pedigree was autosomal dominant, it could also be an autosomal recessive pattern depending on the phenotype defined.

Another variant of Kohlschütter-Tönz syndrome was described by Musumeci in another Sicilian family (Musumeci et al. 1995). In this pedigree, in addition to seizures, abnormal enamel, and mental retardation which are frequently reported, these individuals also had broad thumbs, toes and enlargement of the lateral ventricles, cerebellum malformation. Similar to the pedigree reported by Guazzi, they also had ataxia and spasticity. There was consanguinity in this pedigree.

More recently, a family with the diagnosis of Kohlschütter-Tönz Syndromes was documented by Donnai (Donnai et al. 2005). As better medical care became available, the cases reported by Donnai were able to live into their teens but with severe handicap. Besides seizures, mental retardation, and amelogenesis imperfecta, abnormal crystal sediments were found in the urine of these cases. The parents were healthy unrelated individuals.

The latest case report was documented by Haberlandt and appeared to be another form of Kohlschütter-Tönz Syndrome (Haberlandt E et al. 2006). The neurological deficits in his case were milder than the previous cases reported, and the seizures in this case showed good response to medical treatment. His seizures onset was at 8 months and developmental delay was noted at the same time. There was no consanguinity reported in this pedigree. However, the parents were from neighboring villages, and inbreeding was possible. Only one child in this pedigree was affected. MRI showed moderate ventricle enlargement and cerebellum hypoplasia.

Genetic analysis of Kohlschütter-Tönz Syndrome

In Kohlschütter's report, he noticed that geographical isolation could accidentally contribute to the development of Kohlschütter-Tönz Syndrome, although no consanguinity was found. The other phenomenon he noticed was there were no female cases. He concluded that the inheritance mode could either be autosomal recessive or X-linked. As more families were reported, it appeared that the inheritance mode was likely to be autosomal recessive as consanguinity and geographical isolation existed in most of the affected families, except the one reported by Guazzi, which was complicated by several different phenotypes.

The genetic defect had not been identified due to lack of large genetically informative families. This is a typical scenario where WES had been shown to be a powerful tool for gene identification.

Linkage analysis was carried out in our department on five families (including families from Christodoulou et al., Musumeci et al., Petermöller et al., Donnai et al. and Guazzi et al.) by using whole genome microsatellite markers as a part of a PhD thesis (Lo, C.-N., 2009). Maximum LOD score obtained was 3.05, at $\theta = 0$ at D16S423 (chromosome 16) in parametric linkage analysis for the combined families, with mode of inheritance set as autosomal recessive with full penetrance (Table 5). Haplotype analysis further suggested a disease locus between 16p13.3-13.2 present in the families reported by Christodoulou, Musumeci and Petermöller.

Table 5. LOD scores on chromosome 16.

θ / Marker	0.0	0.01	0.05	0.1	0.2	0.3	0.4
D16S521	-infini	-5.59	-2.35	-1.17	-0.31	-0.04	0.02
D16S3065	-infini	-0.57	0.56	0.82	0.75	0.47	0.18
D16S423	3.05	2.97	2.64	2.23	1.44	0.74	0.22
D16S418	-infini	-0.58	0.57	0.85	0.81	0.53	0.23
D16S3062	-infini	-3.92	-1.40	-0.54	0.00	0.09	0.06

Obtained by Merlin, table published elsewhere (Lo, C.-N., 2009).

While we were analyzing genetic data from some KTS families, loss of function mutations in *ROGDI* were identified in different pedigrees with typical features of KTS (Schossig et al., 2012a), (Mory et al., 2012): the original family reported by Kohlschütter in 1974, the Austrian family reported by Haberlandt, a novel Moroccan family and a cohort of five Israeli families, likely to be all related.

Schossig et al performed linkage analysis and autozygosity mapping in the Moroccan consanguineous family (two affected and two unaffected siblings): they identified 4 possible linked regions, comprising a total of 15.83Mb and containing 326 genes. Considering the large number of genes in the regions they did WES in one affected individual. After alignment and SNP calling (identification of about 21000 variants) they applied the auto_annovar functionality to filter variants (here described in chapter 2). After filtering, they considered only the homozygous variants in the linked region and only *ROGDI* remained, with novel frameshift deletion (c.229_230del: p.Leu77Alafs*64) predicted to disrupt aminoacid structure and cause a premature stop codon. The mutation was confirmed by Sanger sequencing and shown to segregate with disease. They also screened this gene in additional KTS families and found three novel mutations: two splice site mutations (c.531+5G.C and c.532-2A>T in intron 7) present in a compound heterozygous state

in one family and a novel homozygous nonsense mutation (c.286C>T in exon 5) in a different family.

Mory et al described 14 new KTS cases pertaining to 5 families (three closely related) from a small Druze village in northern Israeli. Homozygosity mapping on ten affected individuals, one unaffected sibling and one obligate carrier parent with 6000 SNP array identified a 2Mb candidate segment on chromosome 16p13.3. Haplotype analysis with microsatellite markers further confirmed a 500,000 base pairs segment shared by all affected individuals, with a LOD score of 6.4 under a recessive model of full penetrance. Traditional Sanger sequencing of the gene identified a nonsense mutation in *ROGDI* (c.469C>T: p.Arg157*) homozygous in all affected individuals, heterozygous in 10 unaffected relatives and absent from Druze controls.

ROGDI is the human homolog of *Drosophila ROGDI* and encodes for a protein of unknown function. *ROGDI* is widely expressed with higher levels in the adult brain, spinal cord, peripheral blood, bone marrow and heart. Genome wide expression studies show that the protein is expressed more in hippocampus than in other tissues in both human and mice (Kapushesky et al., 2012). At a cellular level, *RODGI* has been shown to localize in the nuclear envelope (Mory et al., 2012). Protein prediction methods (Rost et al., 2004) indicate that *ROGDI* is a globular protein and the secondary structure consists of 45% helix motifs, 37 percent loop structures and 17 percent loop strands. The identification of *ROGDI* mutations in KTS has provided information on the clinical effects of the loss of ROGDI function in humans. The studies described above identify loss of function *ROGDI* mutations in patients presenting with the typical core features of KTS. Future work is needed to elucidate the function of *ROGDI* in neurogenesis and amelogenesis.

1.4.3 Genetics of complex neuropathies

Inherited neuropathies are the most common genetic neurological disorders, and affect ~1 in 2,500 people. Charcot Marie Tooth (CMT) includes a group of hereditary disorders in which motor and sensory neuropathy is the sole or primary part of the disease. CMT is traditionally classified into two types based on electrophysiological and neuropathological criteria, with CMT1 defined as 'demyelinating' and CMT2 as 'axonal' (Shy et al., 2000). Symptoms of CMT include slowly progressing distal weakness, atrophy and sensory loss, which spread proximally as the disease progresses. CMT can be also associated with a variety of additional symptoms, such as spastic paraparesis, optic atrophy, cranial nerve involvement and glaucoma (Shy et al., 2000).

The search for a genetic cause in CMT started in late 1980s, and currently mutations in over 30 genes have been identified (Reilly et al., 2011). CMT1 is predominantly (70-80% of the cases) caused by the duplication of a large region on the short arm of chromosome 17, that includes the gene peripheral myelin protein (*PMP22*). The most frequent cause of CMT 2 are mutations in the mitofusin 2 gene (*MFN2*) (accounting for about 20% of CMT2 cases) (Lawson et al., 2005). Notably, CMT2 caused by mutations in the *MFN2* can be phenotypically heterogeneous, presenting as classical CMT2, CMT2 with pyramidal signs (Ajroud-Driss et al., 2009), CMT2 with optic atrophy (Züchner et al., 2006), and a severe early-onset (<10 years) or a mild late-onset (>10 years) phenotype (Chung et al., 2006).

Despite genetic and clinical heterogeneity, a classification was proposed in 1975 dividing the hereditary motor and sensory neuropathies into seven types (Dyck PJ, 1975). Types I through IV represent known genetic entities. Types V to VII represent CMT with additional features such as spastic paraparesis (type V), optic atrophy (VI) or retinitis pigmentosa (VII).

1.4.3.1 Hereditary motor and sensory neuropathy type 6 (HMSN VI)

The association of axonal neuropathy with optic atrophy is known as CMT 6 or hereditary motor and sensory neuropathy VI (OMIM #601152). This clinical entity was first report by Vizioli in 1879 (Vizioli F., 1889). Subsequently, further studies

have reported the association of axonal neuropathy and optic atrophy in different families, comprising at least 42 cases. The clinical manifestations of CMT6 consist of distal muscle weakness and wasting starting in the lower limb with reduced reflexes and glove and stocking sensory loss. There is progressive visual acuity loss due to optic atrophy, eventually leading to blindness. The age at onset is usually in childhood (first decade) for the neuropathy and in the second decade for the optic atrophy.

Vizioli (Vizioli F., 1889) described a kinship in which a father and sons were found to have amaurosis with optic atrophy in association with CMT. The father became ill with neuropathy at age 59 and the two sons at age 26 and at age 6 respectively. The father and the older son lost visual acuity and eventually became blind. Muscle weakness started in the lower limb and later affected the arm and hand.

Schneider and Abeles (Abeles M., 1937) reported two brothers aged 41 and 26 affected by primary optic atrophy with progressive peroneal atrophy. The disease manifested gradually in childhood and progressed slowly. The affected brothers had four healthy siblings and the parents were first cousins.

Milhorat (Milhorat A.T., 1943) described two brothers, who developed nystagmus at age 9 years, distal muscular atrophy at age 15, and decreased visual acuity in their 20s. The brothers were born from first-cousin parents.

Barreira (Herrera RF, 1990) reported a slowly progressing decrease of visual acuity and peroneal weakness in 12 years old boy and his 10 years old sister. The siblings were born from consanguineous parents.

Chalmers (Chalmers et al., 1997) described three siblings (two females and a male) with childhood onset of motor and sensory neuropathy and adult onset optic atrophy. Neither their parents nor four other siblings had any neurological or ophthalmological condition.

Ippel (Ippel et al., 1995) reported a father and two affected offspring (a boy and a girl) with polyneuropathy and optic atrophy. The distal muscle atrophy/weakness manifested early in childhood and the reduction of visual acuity started in late childhood/adolescence.

In another report, Chalmers described a three generations family (comprising four affected) with optic atrophy and asymptomatic neuropathy. Impaired vision

developed in early childhood and although none of the affected members had symptoms related to neuropathy, they had diminished tendon reflexes and impairment of sensation. Moreover, nerve conduction studies confirmed the presence of axonal sensory-motor neuropathy (Chalmers et al., 1996).

Voo (Voo et al., 2003) described a large family in which 58 members were affected by peripheral neuropathy and optic atrophy. Twelve members were affected by both neuropathy and optic atrophy; three other family members had either neuropathy or optic atrophy. Although the clinical syndrome was variable, most had childhood onset of progressive visual loss and abnormal gait, distal sensory impairment, and hyporeflexia. Other clinical features included hearing loss, tinnitus, and anosmia.

Züchner reported 6 unrelated families, comprising 10 affected individuals with axonal neuropathy and optic atrophy. Onset of axonal peripheral neuropathy was in childhood, ranging from 1 to 10 years. The symptoms showed severe progression, with almost all patients becoming wheelchair-bound. Onset of optic atrophy was between 5 and 50 years of age (mean age 19 years). Remarkably, 60% of the patients experienced significant recovery of their visual acuity after several years. Incomplete penetrance was observed in 1 family (Züchner et al., 2006).

Genetics of CMT6

CMT 6 is genetically heterogeneous, as both autosomal dominant and recessive inheritance has been suggested. The first reports are consistent with autosomal recessive inheritance, where the affected offspring is born to consanguineous parents (Abeles M., 1937), (Milhorat A.T., 1943), (Herrera RF, 1990) or the diseases manifest in siblings with normal parents (Chalmers et al., 1997).

Clear autosomal dominant inheritance is first described Voo in a large one multi-generation family comprising a total 97 members, with 12 members having CMT6. The disease shows great clinical variability (age at onset, severity of disability) and the penetrance is incomplete.

Züchner (Züchner et al., 2006) clarified the genetic basis of the dominant forms of CMT6 by performing genetic analysis of the mitofusin 2 (*MFN2*) gene in six families and identified a unique MFN2 mutation in each pedigree.

MFN2 encodes a mitochondrial membrane protein that participates in mitochondrial fusion and contributes to the maintenance and operation of the mitochondrial network (events that are essential for mitochondrial morphology and function).

The causes of the the recessive forms of CMT2 with optic atrophy has remained elusive due to the small number of reported recessive multigenerational families, and a genome wide linkage scan have never been attempted for this form of disease. We describe the first genetic defect in a recessive form of CMT 6 in chapter 4.

2 WHOLE EXOME SEQUENCING PILOT PROJECT

2.1 STATEMENT OF CONTRIBUTION TO THIS RESEARCH

I was involved in the experimental design of this study (choice of the sample preparation kit, sequencing machine and sequencing method). I did all the wet lab work for samples preparation (DNA library). I performed most of the data analysis, from quality control to alignment and variant calling.

2.2 BACKGROUND

WES was first commercially available in 2009 and had never been used in our Institute. We conducted a pilot project to test and validate the technology by preparing DNA libraries and sequencing from small number of samples. We generated the first sequencing data in the department, and set up an analysis pipeline which would be used for all following WES projects.

I am using the pilot data in this chapter to explain the methods that I used in the other chapters.

2.3 MATERIALS AND METHODS

2.4 DNA samples selection

We selected autopsy confirmed progressive supranuclear palsy (PSP) samples of self-reported European ancestry. Genomic DNA was extracted from brain by standard procedures. A total of four samples were selected on the basis of DNA quality (spectrophotometer) and quantity (about 10 µg).

2.4.1 Library preparation

Adapter-enriched DNA sample libraries were captured using the Illumina protocol for preparing samples for paired-end sequencing (part #1005063, June 2008) and NimbleGen SeqCap EZ Exome (v1.0, 29/10/2009) following the manufacturer's protocol.

The Illumina protocol prepares DNA for paired-end analysis on the Illumina Cluster Station and Genome Analyzer by adding adapter sequences to the ends of DNA fragments. The adapters contain sequences that correspond to the surface-bound amplification primers on the flow cell; this allow the DNA molecules to hybridize to the surface of the flow cell.

The NimbleGen SeqCap EZ Exome captures specified regions of the human genome from genomic DNA prepared with Illumina Paired-End Genomic DNA Sample Preparation Kit. The NimbleGen EZ Exome system uses variable length DNA probes (biotinylated oligonucleotide baits) complementary to the targets to hybridize sequencing libraries prepared from fragmented genomic DNA. These libraries are enriched for targeted regions by pull-down with magnetic streptavidin beads and then sequenced. The NimbleGen kit targets 174,984 Consensus Coding Sequence (CCDS) (http://www.ncbi.nlm.nih.gov/CCDS) exons of 16,008 high confidence protein coding genes (build 36.2), and 528 human miRNA genes obtained from miRBase (release 10). In sum, the coding and miRNA target covered approximately 26.2 Mb of the human genome.

2.4.1.1 Samples preparation for Paired-End sequencing

Briefly genomic DNA was fragmented using COVARIS E210 water bath sonicator, which shears DNA to fragment sizes of about 100-500 base pairs. 3 µg of genomic DNA was brought to a total volume of 100 µl with 1xTE Buffer. Sonication conditions were as in Appendix A (Duty Cycle: 10%, Intensity: 5, Cycles per Burst: 200, Time: 180 s, Temp: 4°C as suggested by the NimbleGen tech support). Then each fragmented sample was cleaned with a Qiagen QIAquick PCR Purification column (following the manufacturer's directions: MinElute Handbook PCR purification, March 2008), and finally elute with 30µl of buffer EB. Each sample was loaded on an Agilent DNA 1000 chip for confirmation of size distribution. Then, the

overhangs resulting from fragmentation were converted into blunt ends using Klenow enzyme and T4 DNA polymerase (the exonuclease activity removes the 3' overhangs, and polymerase activity fills the 5' overhangs). The sonicated DNA was incubated for 30 minutes at 20°C in a 100 µl reaction volume containing T4 DNA ligase buffer with 10mM ATP (10 µl), 10 mM dNTP mix, 10 mM dNTP mix (4 µl), T4 DNA polymerase (5µl), Klenow enzyme (1 µl), T4 PNK (5 µl), Water (45 µl), DNA (30 µl). Samples were then cleaned with QIAquick PCR Purification Kit on one QIAquick column, eluting in 32 µl of EB.

Then an 'A' base was added to the 3' end of blunt phosphorylated DNA fragments using the polymerase activity of Klenow fragment: each DNA sample was incubated for 30 minutes at 37°C with the Adenylation mastermix, containing Klenow buffer (5 µl), 1 mM dATP (10 µl), Klenow exo (3' to 5' exo minus) (3 µl), DNA sample (3 µl). Samples were cleaned using QIAquick MinElute columns, eluting in 10 µl of EB.

The adapters, that have a single 'T' overhang, were ligated to DNA fragments using a DNA ligase: DNA samples were incubated thermal cycler for 15 minutes at 20°C with the Adapter ligation mastermix, containing DNA ligase buffer, 2X (25 µl), PE adapter oligo mix (10 µl), DNA ligase (5 µl), DNA sample (10 µl). Samples were purified with QIAquick PCR Purification Kit on QIAquick columns, eluting in 30 µl of EB. The ligation products were then size-selected at approximately 300 base pairs on a gel to remove all unligated adapters and adapters that may have ligated to one another. Briefly, a 2% agarose gel was prepared with 150 ml distilled water and TAE (Certified low-range Ultra Agarose BIO-RAD, part # 161-3106) and 60 µg EtBr. All samples and one ladder (Hyperladder IV from Bioline) were loaded and run at 120 V for 120 minutes. After that, a gel slice containing the fragment of interest was then excised and DNA extracted (using the QIAquick Gel Extraction Kit on QIAquick columns), eluting in 30 ml of EB. Then the adapter ligated DNA fragments were enriched by PCR using two primers that anneal to the end of the adapters. This step was also required to amplify the amount of DNA in the library. The reaction, containing Phusion DNA polymerase (25 µl), PCR primer PE 1.0 (1 µl), PCR primer PE 2.0 (1 µl), DNA sample (10 µl) was incubated at the following conditions: 30 seconds at 98°C, 10 cycles of: 10 seconds at 98°C, 30 seconds at 65°C, 30 seconds at 72°C, 5 minutes at 72°C, then hold at 4°C. PCR products were purified across four QIAquick

columns (Qiagen) and all the elutants pooled (elution in 10µl of EB as suggested by the NimbleGen protocol). The quality of the purified and amplified DNA library was assessed on an Agilent DNA 1000 chip.

2.4.1.2 Exome capture

Adapter-enriched DNA sample libraries were captured using NimbleGen SeqCap EZ Exome following the manufacturer's protocol. DNA libraries were first amplified by ligation PCR (using primers complementary to sequencing adaptors) in 50 µl reaction volume containing PE-PRE1 Oligo (1µl), PE-PRE2 Oligo (1µl), Phusion High fidelity PCR Master mix (271µl), DNA library (1µl) at the following conditions: 30 seconds at 98°C, 11 cycles of 10 seconds at 98°C - 30 seconds at 65°C - 30 seconds at 72°C, 5 minutes at 72°C, hold at 4°C. Amplified DNA sample libraries were then hybridized to the Exome Library. They were first prepared with COT DNA (100 µl of 1mg/ml), PE-H1 and PE-H2 (1000 µM) and dried in a SpeedVac at high heat (60°C). Then SC Hybridization Buffer (7.5 µl) and Hybridization Component A (3µl) were added to the dried amplified sample library; this cocktail was placed in a 95°C heat block for 10 minutes to denature DNA. The amplified sample library/COT DNA/PE-HE Oligos/Hybridization cocktail was transferred to an aliquot (4.5 µl) SeqCap EZ Exome library and the resulting mixture was incubated at 42°C for 72 hours. The captured DNA was then washed: SC Wash Buffers and Stringent Wash Buffer were first diluted to 1x working solution, then Stringent Wash Buffer (20 µl) and SC Wash Buffer I (5 µl) were heated to 47°C in a water bath. Streptavidin Dynabead Binding and wash buffer were prepared with Trizma Hydrochloride 1M (25 µl), ETDA 0.5M (5 µl), NaCl 5M (1000 µl) and PCR grade water (1470 µl) and 100 µl of beads were aliquoted in one single 1.5 ml tube and placed in a DynaMag-2 device. All clear liquid was removed with 3 wash steps. The beads were resuspended in 100 µl Streptavidin Dynabead and Wash Buffer; the hybridization sample was finally transferred to this mix. The tube containing the beads and the hybridized sample library was placed in a thermo cycler at 47°C for 45 minutes to allow the binding of captured sample to the beads. Then the Streptavidin Dynabeads plus bound DNA were washed to remove fragment not bound to beads. First, 100 µl of SC Wash buffer I (at a temperature of 47°C) was added to the tube containing DNA bound to

beads. The tube was then placed in a DynaMag-2 device to bind the beads and the liquid was discarded. Then 200 µl of Stringent Wash Buffer (at a T of 47°C) was added to the mix and incubated at 47°C for 5 minutes. A total of two washes were performed with Stringent Wash Buffer as described earlier. Subsequent washes with 200 µl SC Wash Buffer I, 200 µl SC Wash Buffer II, 200 µl SC Wash Buffer III were performed and for each wash the tube was placed in a DynaMag-2 device and clear liquid was discarded. After the last wash, 50 µl of PCR grade water were added to each tube of bead-bound captured sample. Last, the bead bound DNA was amplified by ligation PCR in a 50 µl reaction volume containing PE-POST1 Oligo 1µl, PE-POST2 Oligo 1µl, Phusion High fidelity PCR Master mix 271µl, DNA library 1µl at following conditions: 30 seconds at 98°C, 11 cycles of 10 seconds at 98°C - 30 seconds at 60°C - 30 seconds at 72°C, 5 minutes at 72°C, hold at 4°C. Each amplified captured library was purified with QIAquick columns (Qiagen) and analyzed on an Aglient 1000 DNA chip.

2.4.1.3 Measurement of enrichment

A standardized set of qPCR SYBR Green assays were used to estimate relative fold-enrichment by measuring the relative abundance of control targets in amplified sample library and amplified captured DNA. For each amplified samples library (n=4), we used a standard set of 4 NimbleGen Sequence Capture control locus qPCR assay (table S1), and each assay was performed in triplicate. For each assay, we also included one negative control (PCR grade water) and one positive control template (control genomic DNA). Briefly, each amplified sample library and captured library was diluted to 5ng/µl in PCR grade water, for use as templates. For each assay we added: 5,9µl of PCR grade water, 0.3µl of NimbleGen Sequence Capture assay forward primer (2µM), 0.3µl of NimbleGen Sequence Capture assay reverse primer (2µM), 7,5 µl of SYBR Green Master (2X), 1µl of templates. The reactions were run on a thermocycler using the following conditions: 10 minutes at 95°C, 40 cycles of 10 seconds at 95°C and 1 minute at 60°C (Quantification), followed by Melting Curve analysis. A total of 192 assays were run: 48 assay per library (each of the 4 qPCR assay was run in triplicates for each DNA sample analyzed: 12 per amplified

sample library, 12 per captured sample library, 12 per negative control and 12 per positive control).

Following data collection, analysis was performed with Absolute Quantification Analysis Module within the LightCycler® 480 Software, where the raw Cp values were used to perform relative quantification comparing amplified sample library and captured DNA. Cp values were exported to a spreadsheet: the average Cp value from all replicate reaction was calculated. For each different sample and NSC assay combination, we calculated the delta-Cp by subtracting the average Cp value measured for the amplified captured DNA template from the average Cp value measured for the corresponding amplified sample library template. We also calculated the fold-enrichment for all NSC loci by raising the PCR Efficiency for that assay to the power of the delta-Cp measured for the corresponding control locus.

2.4.2 Sequencing

Sequencing was performed by UCL Genomics on the Illumina Genome Analyzer IIx. Sequencing was carried out as paired end 76base pairs reads, following the manufacturer's instructions and using the standard sequencing primer. Image analysis and base calling was performed by the Genome Analyzer Pipeline version v1.6. Data were handed in two separate files for each read (read_1 and read_2) for each sample.

2.4.3 Bioinformatics

A schematic representation of data analysis is represented in Figure 6.

1.QUALITY CONTROL

The sequencing data generated by the Genome Analyzer is stored in a FASTQ file**,** a text-based format for storing nucleotide sequence and its corresponding quality scores. The certainty of each base call is recorded as a 'Phred' quality score, which measures the probability that a base is called incorrectly. The quality score of a given base, Q, is defined by the equation:

$$\mathbf{Q = -10\log_{10}(e)}$$

where 'e' is the estimated probability of the base call being wrong. Thus, a higher quality score indicates a smaller probability of error. A quality score of 20 represents an error rate of 1 in 100, with a corresponding call accuracy of 99% (Table 6).

Table 6. Phred quality scores

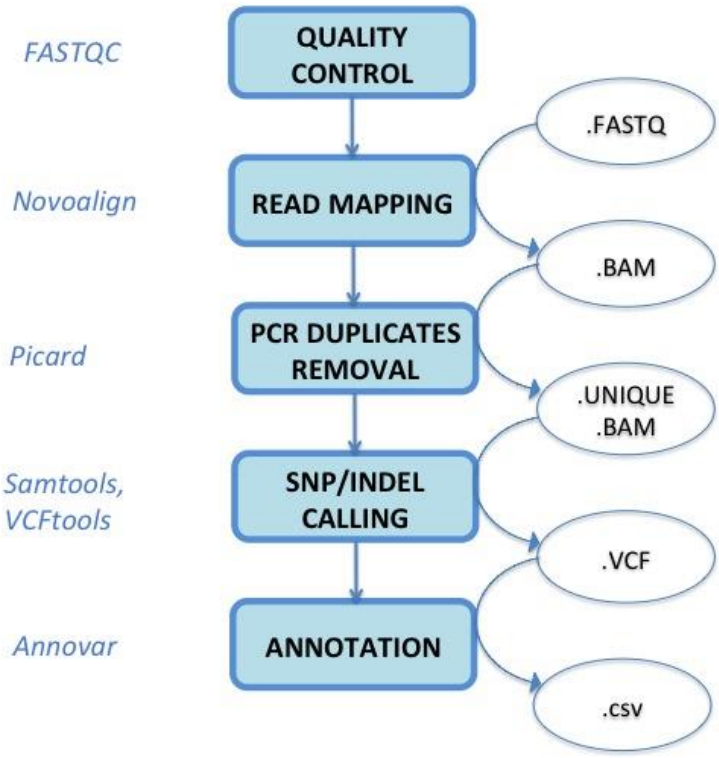
Quality Score	Probability of Incorrect Base Call	Inferred Base Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%

Phred quality scores are logarithmically linked to error probabilities

In the Illumina technology the per-base quality score is determined by background noise during imaging (Nielsen et al., 2011). Practically Phred scores over 40 are uncommon, and a cut-off Phred score of 20 is commonly used (Metzker, 2010).

We analyzed raw reads with the FastQC program (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). FastQC runs a series of tests on a fastq file to generate a comprehensive quality control report. FastQC assess data quality by evaluating: read length, per base quality score, per sequence quality scores, GC content, nucleotide content, sequence duplication and overrepresented sequences.

Figure 6. Analysis pipeline for WES



Left panel: Program used for each step. Middle: analysis steps. Right panel: data formats.

2.MAPPING

Mapping (or alignment) is the process of aligning reads to the reference genome. Illumina Genome Analyzer generates millions of 76 base pairs (comprising tens of Gb) and corresponding quality scores for each base call. Because of the large volume of reads and the huge size of the whole reference genome, alignment algorithms have been optimized for speed and memory usage. Furthermore, since the sequencing genome is usually different from the reference genome, alignment algorithms have been designed to be robust enough to sequencing errors, but do not miss true genomic polymorphism. Therefore different alignment tools are designed with different approaches to trading off speed and accuracy to optimize detection of different types of variations in donor genomes. Alignment algorithms usually follow a multistep procedure to map a sequence. First they quickly identify a small set of places in the reference sequence where the sequence is most likely to accurately align to. Then slower and more accurate alignment algorithms are run on the limited subset of possible mapping location identified in the first step. To speed up the process most alignment algorithms construct auxiliary data structures, called indices, for the reference sequence and/or read sequences.

We performed our analysis with Novoalign (www.novocraft.com), an aligner designed for single and paired-end reads from the Illumina Genome Analyzer. First, a reference assembly (downloaded from UCSC: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/) was indexed; then the paired-end reads were aligned to the indexed set of reference sequence:

```
novoindex  hg19.nix  hg19.fasta
novoalign -o SAM -F ILMFQ -f input_1 input_2 -d hg19.nix
> outputSAM
```

the options indicate:
-o = output report in the SAM format (see paragraph 3)
-F = Specifies the format of the read sequence file. In this case ILMFQ (Fastq with Illumina coding of quality values)
-f = specifies the files containing the read sequences to be aligned
-d = indicated the full pathname of indexed reference sequence

Novoalign finds global optimum alignments using an index called fast-hashing Needleman-Wunsch algorithm, and uses affine gap penalties to allow consecutive run of spaces in a single string of a given alignment. Novoalign allows gaps up to 7bp on single end reads, even longer on paired end reads. Novoalign also supports paired-end mapping: it first finds the positions of all the good hits, sorts them according to the chromosomal coordinates and then does a scan through all the potential hits to pair the two ends. The expected size distribution for these sequencing runs were on average 300 base pairs as determined in the library preparation. Importantly, Novoalign calculates a mapping quality score, which is Phred-scaled probability of the alignment for the whole read being incorrect. This probability depends on the length of alignment, on the number of mismatches and gaps and on the uniqueness of the aligned region on to the genome. It is important in the downstream analysis to distinguish the real SNPs from the mismatches between, for instance repeated homologous genomic regions.

3. PCR DUPLICATE REMOVAL: SAMTOOLS AND PICARD

Samtools (http://samtools.sourceforge.net/) is a suite of tools designed primarily to manipulate SAM files, and the binary transformed BAM files, in preparation for downstream analysis. The SAM format was created in 2009 (Li et al., 2009) to define a generic nucleotide alignment format that described the alignment of large nucleotide sequences to a reference sequence. SAM files are tab-delimitated text files that contain a header section, which carries information on the project and the genome, and an alignment section, that contains alignment information such as mapping position and information on mismatches. This information is used for downstream analysis.

Samtools was used for converting SAM and BAM files, for sorting (arranging the file according to left-most coordinates) and indexing (generating a complementary file '.bai' which aids fast access to the BAM file) the alignment files generated by Novoalign. Samtools was also used for calling SNPs and short indel variants (see paragraph on SNP calling). Below are the commands used.

To make a BAM file:

```
samtools view -bS -t ${fasta}i -o inputBAM outputSAM
```

To sort the BAM file:

```
samtools sort inputBAM output_sortedBAM
```

To build the index:

```
samtools index output_sortedBAM
```

One technical artifact of capture-sequencing procedures is the generation of duplicate DNA sequencing reads (defined as reads with the same start point and direction) that are due to PCR-induced duplication. They share the same sequence and have the same alignment position and could cause trouble during SNP calling as some allele could be overrepresented due to amplification biases. We removed PCR artifacts with Picard (http://sourceforge.net/projects/picard/), which comprises Java-based command-line utilities that manipulate SAM or BAM files. Picard removes all read pairs with identical coordinates, only retaining the pair with the highest mapping quality. Picard examines aligned records in the supplied SAM file to locate duplicate molecules and generates a SAM output file that includes all aligned reads, without the duplicate records. Picard also generates a file that contains information on the percentage of PCR duplicates found in the original aligned file.

```
java -Xmx2g -jar MarkDuplicates.jar ASSUME_SORTED=true
REMOVE_DUPLICATES=TRUE INPUT=output_sortedBAM
OUTPUT=output_sorted_uniqueBAM
METRICS_FILE=output_picard_metrics.out
```

Picard can also create a summary of alignment metrics about the alignment of reads within a SAM file. For instance, given the target intervals Picard calculates useful information on the quality of the alignment (eg. percentage of reads aligned, coverage) that can be useful to know for downstream analysis (Table 9).

```
java -Xmx2g -jar CalculateHsMetrics.jar
BAIT_INTERVALS=Nimblegen_tiled_regions.hg19.bed
TARGET_INTERVALS=CCDS_hg19
INPUT=output_sorted_uniqueBAM
OUTPUT=output.hybridMetrics
```

4.SNP CALLING

Once an alignment is generated, SNP calling can be performed by comparing the aligned SAM/BAM file to the reference genome. The end result of a SNP calling analysis is a collection of SNPs, each associated to a Phred-like quality score that takes into account base calling as well as mapping scores. A standard file has been created to hold these SNPs and related information, the Variant Call Format (VCF). BCF is the binary version of VCF: it keeps the same information in VCF, while much more efficient to process especially for many samples.

We performed SNP calling with Samtools and VCFtools (Danecek et al., 2011), a software suite that implements various utilities for processing VCF files, including validation, merging, comparing.

```
samtools mpileup -q 20 -L 400 -d 400 -ugf
indexed_hum_ref output_sorted_uniqueBAM | bcftools view
-bvcg - > output_rawVar.bcf | bcftools view
output_rawVar.bcf > output_Var.vcf
```

where the options indicate:
SNP and INDEL CALLING PARAMETERS:

q = 20. Minimum mapping quality for an alignment to be used
L = Skip the INDEL calling if the average depth is above 400
d = 400. Maximal read depth at a position
OUTPUT PARAMETERS:
u = output is uncompressed BCF
g = Compute genotype likelihoods and output them in the binary call format (BCF)
f = snp calling is based on the indexed reference file in the FASTA format
BCFTOOLS VIEW OPTIONS:
b = tells to output to BCF format (rather than VCF)
c = tells to do SNP calling
v = tells it to only output potential variant sites (i.e., exclude monomorphic ones)
g = tells to call genotypes for each sample in addition to just calling SNPs

---

Samtools collects summary information in the input BAMs, such as the number of different reads that share a mis-match from the reference, the cloning process artifacts (e.g. PCR induced mutations), the error rate associated with the sequence reads (eg. the Phred score associated to every base in the read), the error rate associated with the mapping (mapping quality) and computes the likelihood of data given each possible genotype; then it stores the likelihoods in the BCF format. Bcftools does the actual SNP calling.

Then the .bcf was converted to .vcf, and only the SNPs on the set of exons targeted by Nimblegen (i.e., CCDS) were included. Lastly, the varFilter script was used to rule out error-prone variant calls caused by factors not considered in the statistical model:

```
vcftools --vcf output_Var.vcf --bed CCDS --recode --out
output_Var_target
```

```
vcfutils.pl varFilter output_Var.vcf | awk '{if ( (\$6
>= 18) || ( \$1 ~ /^#/) ) print}' >
output_Var.vcf_filtered
```

5.INDELS CALLING

To call indels we used Dindel (Albers et al., 2011), a program specifically designed for calling small indels from next-generation sequence data by realigning reads to candidate haplotypes. Dindel considers all candidate indels in a BAM file, and tests whether each of these is a real indel, a sequencing error or mapping error. In stage I) Dindel extracts all indels from the read-alignments in the BAM file. These indels are the candidate indels around which the reads will be realigned in stage III). In this stage, Dindel also infers the library insert size distributions. These will be used in stage III) for paired-end reads. In stage II) the candidate indels obtained in stage I) are grouped into windows of $\sim$ 120 bp, into a realign-window-file. In stage III) for every window, Dindel generates candidate haplotypes from the candidate indels it detects in the BAM file, and realigns the reads to these candidate haplotypes. Realignment is the computationally most intensive step. In stage IV) indel calls and qualities are produced in the VCF4 format. This step integrates the results from all windows into a single VCF4 file.

STAGE I)

```
dindel --analysis getCIGARindels --bamFile
output_sorted_uniqueBAM --outputFile
output.dindel_output --ref fasta
```

STAGE II)

```
makeWindows.py --inputVarFile
output.dindel_output.variants.txt --windowFilePrefix
sample.realign_windows --numWindowsPerFile 1000
```

STAGE III)

```
for realign in /*.txt; do
  mbcode=\`basename \$realign\`
```

```
dindel --analysis indels --doDiploid --bamFile output_sorted_uniqueBAM --ref fasta --varFile realign --libFile output.dindel_output.libraries.txt --outputFile \${mbcode}.output

echo \${mbcode}.output.glf.txt >> output.dindelfilenames
```

STAGE IV)

```
mergeOutputDiploid.py --inputFiles output.dindelfilenames --outputFile output_dindel.vcf --ref ${fasta}
```

Then, a one line awk script line was used to extract only the indels that passed the filters and write them in a vcf file (output_dindel_filtered.vcf):

```
awk '{if ( ( \$1 ~ /^#/) || (\$7 == \"PASS\") ) print}' output_dindel.vcf > output_dindel_filtered.vcf
```

6.ANNOTATION

The ANNOVAR tool (Wang et al., 2010) is a program that utilizes information to functionally annotate genetic variants detected by SNP and indel calling, such as examining their functional consequence on genes, infer cytogenetic bands, report functional importance scores, find variants in conserved regions, or identify variants reported in the 1000 Genomes Project and dbSNP.

```
annovar/convert2annovar.pl -format vcf4 output_Var.vcf_filtered -outfile output_annovar/annovar_${code}_temp
```

```
summarize_annovar_DK_23steps.pl -genetype refgene -buildver hg19 ${output}_annovar/annovar_${code} ${annoDB}
```

A program called auto_annovar.pl is also provided to automate the procedure to identify a small subset of most likely causal variants, from a large list of variants in Mendelian disorders. This has been used for instance to identify ROGDI mutations in Kohlschütter-Tönz (Schossig et al., 2012a). This script first identifies a list of variants that are more likely to cause a Mendelian disease by identifying: splicing and exonic variants, variants located in genomic regions that are annotated as Most Conserved Elements (more likely to be functional), variants that are not located in segmental duplications regions (less likely to be affected by genotype calling issues), and by removing variants observed in the 1000 Genomes Project and dbSNP130 (CEU, YRI, JPT, CHB, respectively) as these variants are less likely to be causing Mendelian diseases.

2.5 RESULTS

2.5.1 Coupling NimbleGen SeqCap EZ to Illumina sequencing

Two commercial platforms were available for exome capture, which conducted the capture either on a solid-phase substrate (i.e. a glass microarray, NimbleGen 2.1M Human Exome Array, Roche-NimbleGen) or in solution (NimbleGen SeqCap EZ Exome). Both capture kits tiled the same oligonucleotides from approximately 180,000 exons of 18,673 protein-coding genes and 551 micro-RNAs and comprise 34.0Mb of genomic sequence. We adopted the solution-based captured kit as it required smaller amounts of input DNA (3 µg versus 10 µg required for array), had lower reagents cost and it did not require an investment in array-processing

equipment or careful training of personnel on array handling. Moreover it was shown to have to have high reproducibility and scalability because it could be conducted entirely in small laboratory tubes. In terms of performance, the two methods were shown to have similar levels of specificity (Bainbridge et al., 2010).

As for sequencing, two platforms for next-generation DNA sequencing reads production were available: the Roche/454 FLX (Margulies et al., 2005), the Illumina/Solexa Genome Analyzer (Bentley et al., 2008) (Table 7). The GA appeared better suited for WES because sequencing by synthesis used by the Illumina technology had better accuracy in base calling (more accurate for variant calling), and also because it had higher throughput at lower cost. Moreover the GA technology used paired end (PE) sequencing, where reads are generated from both ends of a captured DNA fragment. With PE sequencing, because the approximate size of the fragment is known (DNA shearing during library preparation to a fragment size of about 300 base pairs), this information is used to constrain the alignment of both reads to the human genome. PE sequencing was shown to provide more accurate read alignment and therefore better accuracy and coverage of target sequences and SNP calling (Bentley et al., 2008).

Table 7. Summary of the Genome Analyzer and Roche 454 main features

	Illumina (GA IIx)	*Roche 454*
Sequencing chemistry	Sequencing by synthesis	Pyrosequencing
Amplification approach	Bridge Amplification	Emulsion PCR
Paired end/ Separation	yes/ 200bp	yes/ 3kb
Gb/run	37-45 Gb	0.5–1 Gb
Read length	100 bp	500-700 bp
Cost per Gb	$500 per Gb	$20,000 per Gb
Pros	High accuracy in base calling	Longer reads (better mapping in repetitive regions and de novo assembly)
Cons	Low multiplexing capacity	High error rate in homopolymer repeats.

All the samples were run on the GA. The GA produced ~20 Gb data per run; each run sequenced 8 lanes on the flowcell. Each sample was run on a single lane of the flow cell, as we were aiming to obtain each targeted base sequenced 30 times (i.e., coverage 30x). We decided to aim at a coverage of 30x because SNP discovery sensitivity was shown to increase by depth of coverage by several studies (Bentley et al., 2008), (Brockman et al., 2008), (Choi et al., 2009): essentially all homozygous positions could be detected at depth of coverage of 15, whereas heterozygous positions accumulate more gradually to 33× coverage.

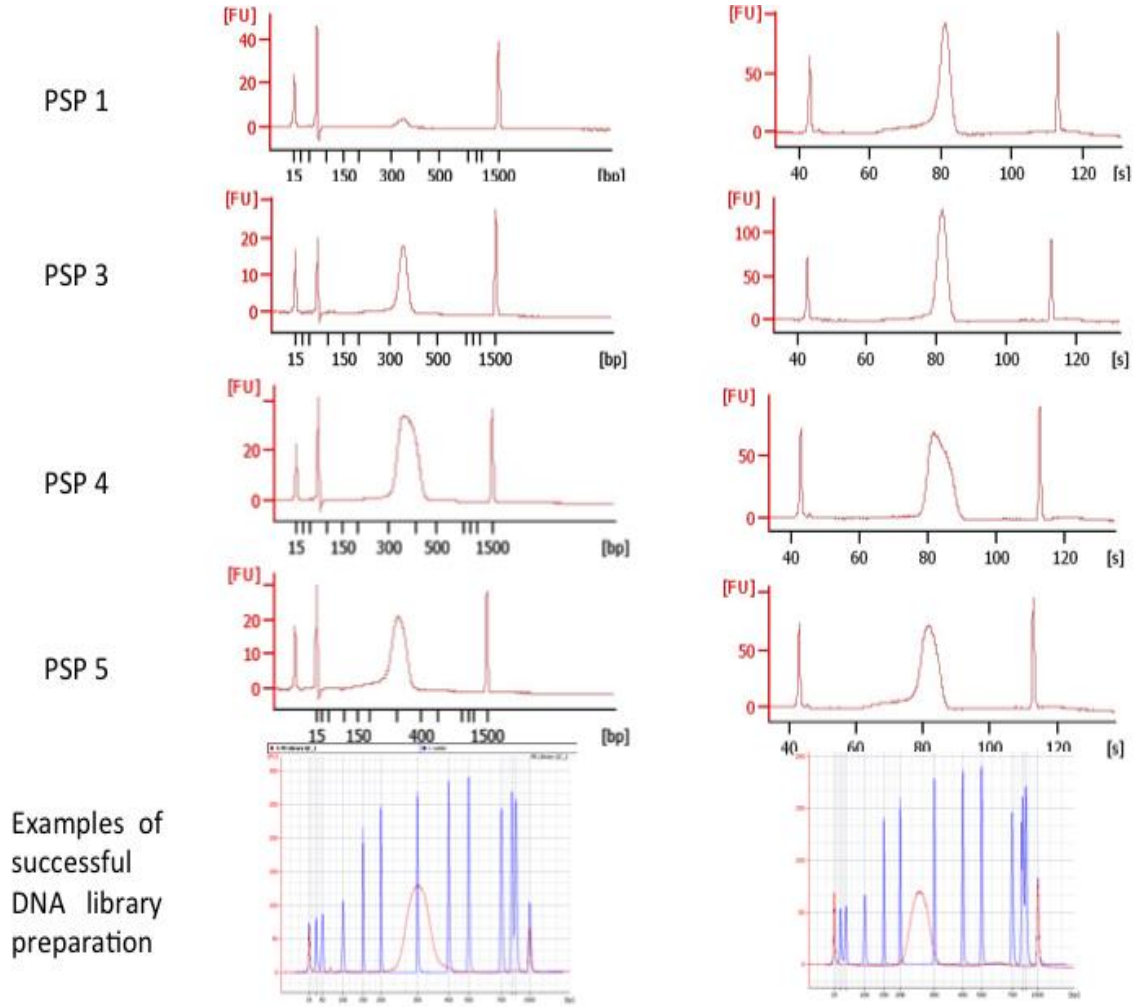
2.5.1.1 Library preparation

Library preparation was successful as showed by the quality control performed after each step of the library preparation process (Figure 7). DNA library is represented by the peak at 200-400 base pairs (which reflects the DNA fragment size) present in all samples. The peak is very low in PSP 1 in the precapture check. At the end of library preparation measure of enrichment showed good enrichment (Figure 8) across samples, indicating that the exome target (as defined by Nimblegen) was successfully captured.

2.5.1.2 Sequencing

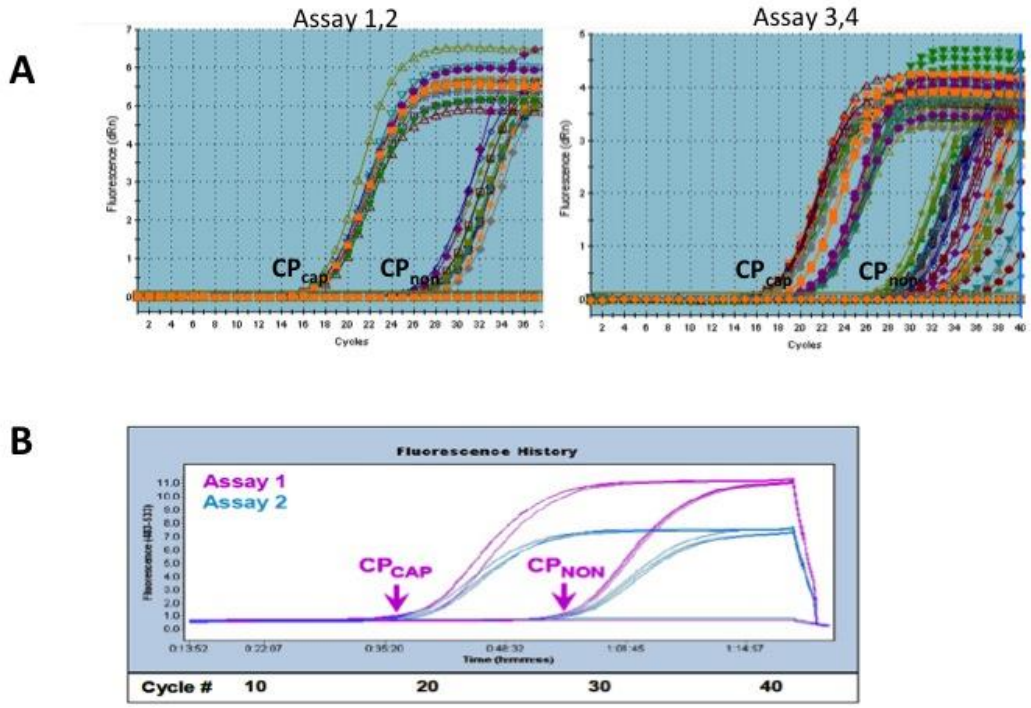
Three PSP samples (PSP 1, 3 and 4) were run on the same flow cell. PSP sample 5 was sequenced on a different run (two months later). Data were handed together with a sequencing report describing the amount and the quality of data generated by each run (Table 8). The sequencing reports showed that the lane yield (i.e. the amount of data generated for each lane, in terms of sequenced bases) almost doubled for PSP sample 5, compared PSP sample 1, 3 and 4. I believe this largely depends on the quality of the sequencing for the following reasons: i) standard amount of DNA library was sequenced in each run; ii) DNA libraries were prepared with the same protocol and showed roughly the same amount of captured library/quality or quantity (Figure 7 and Figure 8); iii) Data on a standard control sample library showed more clusters in the second run (283533 +/- 16410 versus 176749 +/- 9376) (Table 8).

Figure 7. DNA library preparation: quality control



Left panel: Bioanalyser track shows quality of the amplified paired-end library prepared using the Illumina protocol. Right panel: Bioanalyser track shows quality of the captured amplified sample library prepared using the Nimblgen exome capture protocol.

Figure 8. Measure of enrichment



A) Dissociation analysis of the qPCR assays. Four different qPCR locus assays were performed on the DNA library pre and post capture to measure the enrichment of standardized control loci. These assays act as a proxy for estimating the enrichment of the capture targets.
B) Example of sequence capture qPCR data for 2 assays. In the successful experiment the Cp value form qPCR of amplified captured DNA templates are significantly lower than Cp values form amplified sample library templates. CAP = post captured; NON = pre capture

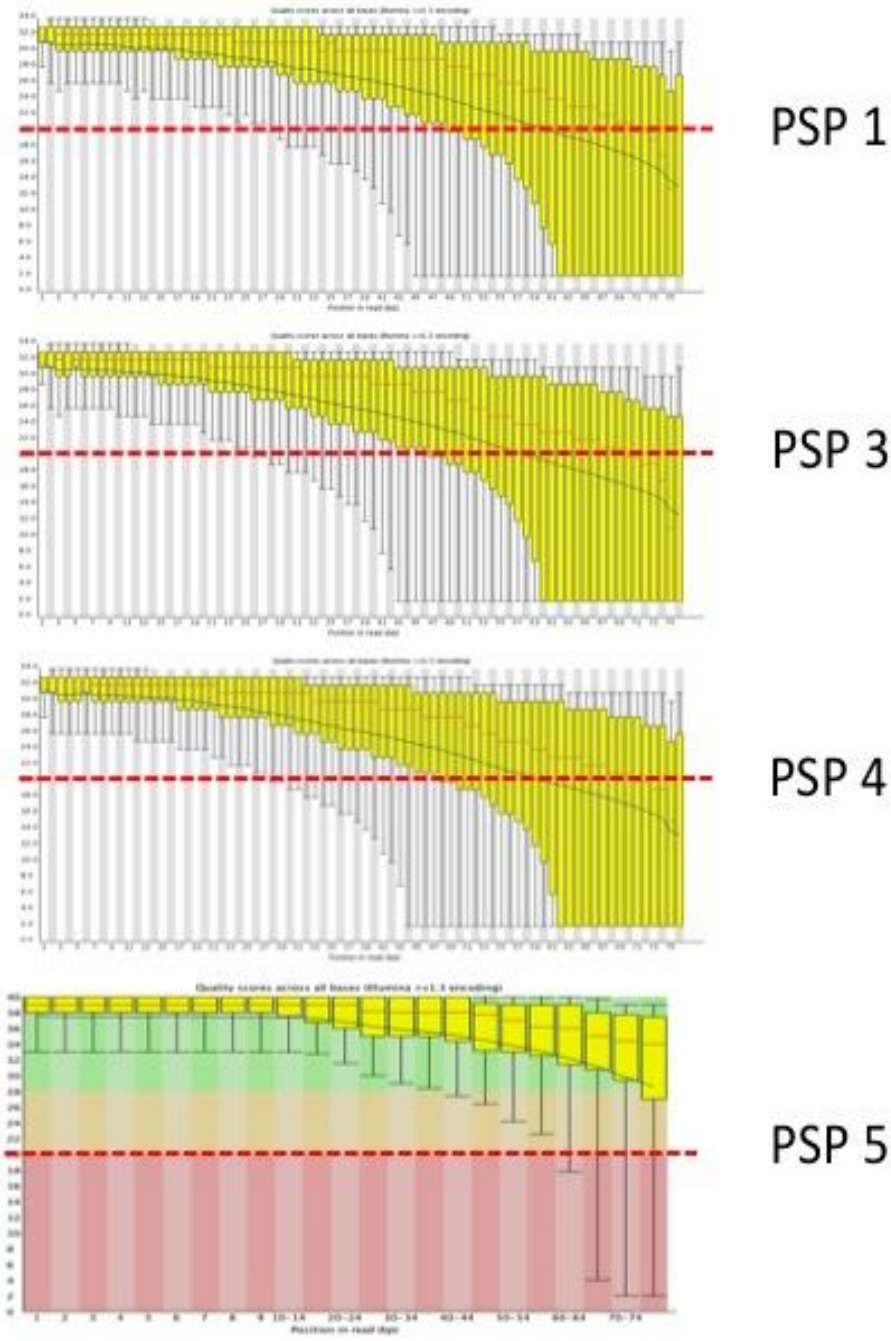
Table 8. Sequencing report

			READ 1			READ2		
Lane	**Customer Name**	**Sample ID**	**Lane Yield (kbases)**	**Clusters (raw)**	**% PF Clusters**	**Lane Yield (kbases)**	**Clusters (raw)**	**% PF Clusters**
1	Arianna Tucci	**1 PSP**	1568868	262450 +/- 4409	65.53 +/- 3.61	1568868	262450 +/- 4409	65.53 +/- 3.61
2	Arianna Tucci	**3 PSP**	1611049	258034 +/- 2973	68.47 +/- 1.87	1611049	258034 +/- 2973	68.47 +/- 1.87
3	Arianna Tucci	**4 PSP**	1641785	239310 +/- 5045	75.24 +/- 1.69	1641785	239310 +/- 5045	75.24 +/- 1.69
4	**CONTROL LANE**	**PhiX**	1398539	176749 +/- 9376	86.80 +/- 1.08	1398539	176749 +/- 9376	86.80 +/- 1.08

Lane	**Customer Name**	**Sample ID**	**Lane Yield (Kbases)**	**Clusters (raw)**	**% PF Clusters**	**Lane Yield (Kbases)**	**Clusters (raw)**	**% PF Clusters**
3	Arianna Tucci	**5 PSP**	3537000	387877 +/- 10250	76.47 +/- 1.67	3537000	387877 +/- 10250	76.47 +/- 1.67
4	**CONTROL LANE**	**PhiX**	2586000	283533 +/- 16410	89.56 +/- 0.95	2586000	283533 +/- 16410	89.56 +/- 0.95

Upper panel: sequencing report for PSP sample 1, 3 and 4. Lower panel: sequencing report for PSP sample 5. Clusters (raw): The number of clusters per tile detected by the image analysis module of the Pipeline. % PF Clusters: The percentage of clusters passing filters. Note that there are 120 tiles per lane on the GAIIx system

Figure 9. FastQC graphs



The per-base sequence quality graphs show an overview of the range of quality values across all bases at each position in the FastQ files. The y-axis shows the quality scores, the x-axis shows the position in the read. For each position a box whisker type plot is drawn. The elements of the plot are: the central red line is the median value; the yellow box represents the inter-quartile range (25-75%); the upper and lower whiskers represent the 10% and 90% points; the blue line represents the mean quality. The dotted line represents the quality score threshold of 20.

2.5.2 Data analysis

Quality check on all samples from the first run (PSP 1, 3 and 4) showed very poor quality scores of base calling towards the end of our reads (Figure 9). Considering a threshold of 20 for quality score, two thirds of the reads could not be reliably used for alignment or SNP calling. Conversely, the quality was good for PSP sample 5 (the mean quality score never falls below 20).

Further analysis confirmed the bad sequencing quality of PSP samples 1, 3 and 4, having only about 40 millions sequencing reads each (as shown by the low lane yield in Table 8). Of note PSP sample 1 was particularly poor in terms of PCR duplicates that had to be removed from the alignment, leaving only about 20% of reads for alignment and a resulting coverage on target of only 5x. Consequently, only 8848 variants could be called from this sample (Table 9). The small number of reads from PSP sample 3 and 4 resulted in a coverage of 15,2x and 19,6x respectively. A total number of variants of 1338 and 13787 were called for each sample (Table 9).

On the other hand, PSP sample 5 had higher coverage (32,7) and 17087 variants were called (Table 9,Table 10). Although current studies report 20,000 cSNVs per individual (Choi et al., 2009), the number of variants we called for PSP sample 5 was consistent with other studies at that time (Ng et al., 2009).

For PSP sample 5, about 751 variants had not been previously reported, and were classified as novel. Of all the cSNVs called, there were 7914 missense variants (316 novel), 77 premature termination codons (16 novel), 85 canonical splice site variants (22 novel) and 8760 synonymous variants (181 novel) (Table 10).

Table 9. Alignment and SNP/INDEL calling summary

	PSP 1	***PSP 3***	***PSP 4***	***PSP 5***
total number of reads	41286030	42396038	43204888	93090718
non-duplicated reads	21%	45%	52%	37,4%
reads aligned to target	24%	34%	36%	37%
target bases 10x	7,7%	55%	60%	73%
mean target coverage	5,1	15,2	19,6	32,7
Tot num of variants	8848	13338	13787	17087

total number of reads = total number of reads
non duplicated reads = reads that have not been removed by Picard after PCR
reads aligned to target = aligned, on-target bases out of the bases available
target bases 10x = target bases sequenced at a coverage of 10x or more
mean target coverage = mean coverage of targets that received at least coverage depth = 2 at one base.

Table 10. Summary of variants in PSP 5

VARIANT TYPE	NUMBER OF VARIANTS
Total variants	**17087**
Missense	7914
Nonsense	77
Splice	85
Synonymous	8760
NOVEL VARIANTS	
Total	751
Missense	316
Nonsense	16
Splice	22
Synonymous	181
KNOWN VARIANTS	
Total	15707
Missense	7227
Nonsense	53
Splice	60
Synonymous	8355

Although this was a pilot project performed mainly with the aim of validating the methods for WES, we looked at the variants present in the samples in a search for clinically relevant variants. We started by analyzing the most easily identifiable damaging (loss of function) mutations: nonsense, splice site or frameshift changes. 134 novel LOF variants (i.e.) were present in PSP sample 5 (table S6).

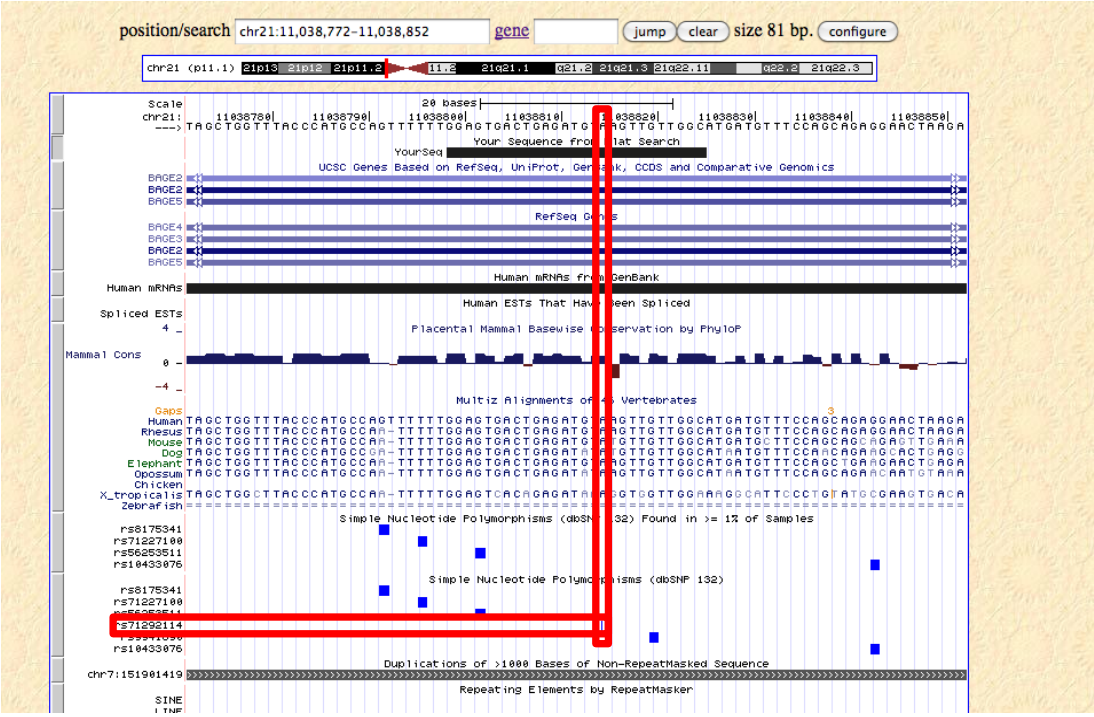
None of these were known to be linked in OMIM to PSP or forms of dementia/parkinsonism. Conversely, a number of these were in genes that had previously been associated with neurological disorders. Although the probability of two or more of the sequenced PSP samples sharing a genetic cause was extremely low, given the sporadic nature of these cases, we looked for the LOF variants present in two or more samples (Table 11), since this would have been an appropriate method with the availability of a much larger cohort.

Table 11. Variants shared by two or more PSP exomes

Gene	*chr.*	*mut*	*PSP_1*	*PSP_2*	*PSP_3*	*PSP_5*
MLL3	chr7:151945072	stop gain	v	v	v	v
NIPA2	chr15:23006219	frameshift del		v	v	v
SIGLEC12	chr19:52004794	frameshift ins	v	v	v	
PRAMEF1	chr1:12854090	stopgain			v	v
OR5K3	chr3:98110413	frameshift ins			v	v
C11orf40	chr11:4592710	frameshift ins			v	v

Surprisingly we found a LOF mutation shared by all samples in the myeloid/lymphoid or mixed-lineage leukaemia 3 gene. *MLL3* is part of the myeloid/lymphoid or mixed-lineage leukemia (MLL) family. The change occurring in all samples, a one bp insertion (uc003wla.2: c.2447_2448insA:p.Y816_I817delinsX) maps also to a duplicated region on chromosome 21 (Figure 10). To investigate this further, we compared the sequence containing the DNA change to the human genome database using blast tool (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start). This showed that the sequence containing the insertion perfectly matches to a region on chromosome 21, which contains the same change in the same position, and is also known to be a SNP (rs 71292114). This means that the read containing that sequence aligned to the wrong position in the reference genome. This is due to the high sequences homology and shows the difficulty of aligning short sequence reads. Similarly, other variants shared by two or more samples, mapped to highly homologous/duplicated sequences (variants in the *PRAMEF1, OR5K3 and SIGLEC12* genes for example). Conversely, the variant in the *NIPA2* gene, turned out to be a SNP later described in the 1000 Genomes Project.

Figure 10. The MLL3 mutation is a SNP on a different chromosome



The sequence containing the MLL3 stopain novel variants (here showns as a horizontal black bar: “Your sequence from blat search”) maps to chromosome 21 (region shown here). An in-depth analysis of the sequence shows the variant initially annotated as a novel nonsense variant is a SNP (rs7129114).

2.6 CONCLUSIONS

The promise that WES brought to genetic research in 2009 was enormous, but the hurdles to make this technique functional in the laboratory were many. The importance and usefulness of this pilot project were in understanding both technical and analytical obstacles to overcome.

On the technical point of view, a) I acquired familiarity with the cumbersome procedure of library preparation and observed the bias in results (PSP 1) that can derive from the amplification steps of the procedure. Indeed Illumina and Nimblegen later modified and optimized these steps; b) I learned how to evaluate the quality of a sequencing run based on the lane yield which directly influences the coverage. These points are particularly important as most of the WES projects in our lab are outsourced and it is crucial to understand when an exome is bad quality and where it is not worth carrying additional analysis.

This pilot project provided me with two areas of knowledge: the technological aspects for WES as well as a first feeling of human coding variation.

On the analytical point of view, the striking number of LOF variants poses major challenges when interpreting such variants:

a) The proportion of annotation error might be high due to the complexity of the genome (duplications and homology regions) and the short length of the sequencing reads.

b) Some of these variants can be true rare variants that have not been described (frequency <0.1) but that are benign. Indeed with the growing number of publicly available WES resources (ESP for instance), many of these variants are being described.

c) Clinical interpretation of LOF variants: determining which of the variants is actually responsible for causing disease remains a hurdle. In Mendelian diseases we can reduce the number of LOF by looking at segregation (i.e. filter for variants present only in affected family members). In complex diseases very large cohorts are needed.

3 KOHLSCHÜTTER-TÖNZ SYNDROME: *ROGDI* MUTATIONS AND GENETIC HETEROGENEITY

3.1 STATEMENT OF CONTRIBUTION TO THIS RESEARCH

I performed WES data analysis of patients with Kohlschütter-Tönz syndrome (KTS). I supervised the laboratory based sequencing and fragment analysis experiments. I collected all the clinical data of KTS patients seen elsewhere.

I wrote the original manuscript that has been published in the journal "Human Mutation".

3.2 BACKGROUND

Ten families with KTS were identified with the core features of epilepsy, psychomotor delay or regression and abnormal amelogenic teeth. We performed clinical and genetic investigations leading to the identification of *ROGDI* mutations in five of them, all presenting with a typical KTS phenotype. The other families, mostly presenting with additional atypical features, were negative for *ROGDI* mutations, suggesting genetic heterogeneity of atypical forms of KTS.

3.3 MATERIALS AND METHODS

3.3.1 Samples

The recruitment of subjects from the Kohlschütter-Tönz pedigrees was made by Prof. Henry Houlden, who contacted the family or researcher working on the family and obtained their permission to use DNA for genetic studies. DNA samples were

collected for a total of 22 samples, comprising 13 affected and 9 unaffected individuals.

3.3.1.1 Clinical details

A total of ten families were identified. Written informed consent was obtained before sample collection and institutional ethics approval was also obtained (Figure 11).

Family A

This family was reported in 1988 (Christodoulou et al., 1988). The parents originate from a small isolated town in Sicily. The affected individuals (6 siblings) had all a similar phenotype characterized by delayed neuromotor development, epilepsy (onset between 7 and 22 months), and amelogenesis imperfecta of the hypoplastic rough type affecting primary dentition.

Family B

This is a new case of KTS. The affected boy was the first child of healthy non-consanguineous parents. He was born after an uneventful pregnancy and his development was normal during the first half year of life. Then it slowed down and at age 7 month the boy presented with treatment resistant seizures followed by severe delay of motor and cognitive development. The primary dentition showed severe enamel defects with yellow teeth. At age 28 months the boy had microcephaly (48cm/ -2,3 SDS) and small stature (83,4 cm/-2,2SDS).

Family C

This family was reported in 1995 (Musumeci et al., 1995). Briefly, it consists of two affected siblings (male and female) born from first-degree cousins parents of Sicilian origin. They had their first seizures between the age of two and 10 months, developed psychomotor regression starting age of two and had very thin hypoplastic enamel with yellow teeth. Presence of broad thumbs and toes is reported.

Family D

This family was reported by Petermöller (Petermöller et al., 2008). Briefly, this family consists of two affected siblings, offspring from healthy unrelated parents. They both had the first epileptic seizures at the age of 8 months, yellow teeth with amelogenesis imperfecta and psychomotor regression starting at the age of two. One of the children died at the age of 21 years.

Family E

This family was reported (Schossig et al., 2012b). The affected individual was born from healthy, unrelated parents of German origin and has two healthy sisters. The boy was born at term after an uneventful pregnancy, and had normal development until he had his first seizures at age of 11 months. After that he showed delay of psychomotor development and partial regression of motor skills. The epilepsy was difficult to treat. His teeth had a brownish discoloration.

Family F

This family has not been previously reported. The child presented at 8 months with seizures and was found to have infantile spasms. MRI of the brain was normal. He went on to have profound developmental handicaps, developed spasticity in all extremities and had intractable mixed focal onset and generalized myoclonic/tonic epilepsy. He was noted to have amelogenesis imperfecta when his teeth erupted and they were yellow. Spasticity was treated with baclofen and he required bilateral femoral osteotomies for hip dislocations and spinal rod placement for scoliosis. At age 15 he was inconsistently visually interactive and has no expressive language; he was unable to sit or roll over.

Family G

This family was reported in 1994 (Guazzi et al., 1994). This pedigree was quite complicated with four different phenotypes occurring concurrently: amelogenesis imperfecta only, delayed psychomotor development, ataxia and abnormal EEG, neurological disorder or seizures and amelogenesis imperfecta. This family was atypical for KTS, as the teeth with amelogenesis imperfecta were not truly yellow

and degenerated and the later age at onset. Moreover inheritance looks dominant but affected individuals from the last generation are born from first-degree cousins parents. It is possible that more than just one disorder is responsible for the complex phenotypes in this pedigree as there were individuals suffering only from amelogenesis imperfecta.

Family H

This family was previously reported by Donnai (Donnai et al., 2005), comprising two affected siblings from healthy unrelated parents. Both siblings developed seizures at age of five weeks (male) or 11 weeks (female), delayed psychomotor development starting at the age of one year and yellow teeth with hypoplastic and hypomineralized enamel. They also had feeding problems with multiple food intolerances. The affected female was also clinically and genetically diagnosed with neurofibromatosis type 1.

Family I

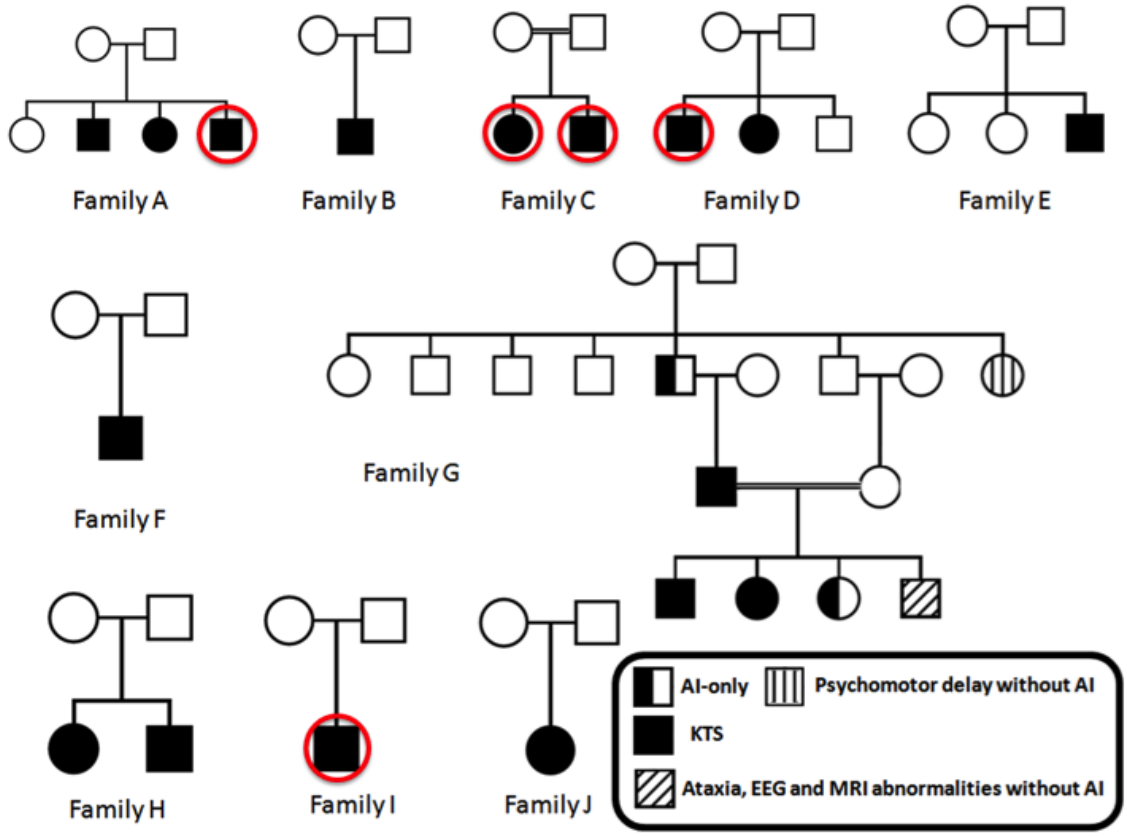
This family hasn't been previously reported. The patient is the second of two children born to healthy unrelated parents. He had a delayed motor and cognitive development. Both primary and secondary dentition showed absent enamel of several but not all teeth. A brain MRI showed mild atrophy of the cranial part of the vermis and a small pons, but otherwise was normal. He had periods of unusual quiet behavior and sleepiness, thereafter true convulsions were not recognized. EEG studies showed multiple epileptiformic activities, mainly frontal and fronto-temporal. He also had some unusual morphological signs: deeply set eyes, horizontal eyebrows, short nose, concave nasal ridge, anteverted nares, irregularly placed teeth, many of them without missing enamel, small ears, and mild hypermobility of the small joints. He had multiple warts.

Family J

This patient showed a mildly delayed global development after normal pregnancy and delivery. She had several episodes suggestive of epilepsy from the age of 18 months and a definitive diagnosis of epilepsy was made at age 6 years. She had

diverse seizures, including atypical absences and myoclonic seizures. EEG showed continuous spike waves during sleep. Epilepsy was resistant to therapy. Her teeth were first white, but discoloured some time after eruption, and the diagnosis of amelogenesis imperfecta was made by a paediatric dentist. Currently she has mild to mental retardation. Neurological examination shows no abnormalities besides mild clumsiness.

Figure 11. KTS pedigrees



Red circles indicate samples that underwent WES. AI= Amelogenesis imperfecta; KTS = Kohlschutter-Tonz syndrome.

3.3.2 Genetic investigations

3.3.2.1 Linkage analysis and homozygosity mapping

DNA array SNP analysis was run in two affected siblings from family A and one affected from family C (using Illumina Human HapMap 300), one affected from family I and one affected from family H (using Illumina HumanOmniExpress-12v1_H). Autozygosity mapping was performed using the plug-in software Homozygosity Detector within the BeadStudio suite. Regions of shared homozygosity that segregated with disease were visually inspected using the Illumina Genome Viewer within the BeadStudio suite.

As part of a project on mental retardation run by our collaborator Dr. Raoul Hennekam, array comparative genomic hybridization (aCGH) analysis with a Nimblegen 2.1M oligo array was also carried out in the affected individual from family I.

3.3.2.2 Whole exome sequencing

WES was carried out on a total of five affected samples from four different families (A, C, D and I) (Figure 11). All samples were enriched using the Illumina TruSeq exome capture system and run on an Illumina HiSeq 2000 for sequencing by AROS Applied Biotechnology, except from one affected from C, which was run using the Nimblegen SeqCap EZ enrichment kit and sequenced on one flowcell on the Illumina Genome Analyzer IIx at NIH (Bethesda, US). All sequencing reads were aligned to the hg19 build of the human reference genome using Novoalign. SNP and indel calling were performed using Samtools version 0.18 and were annotated using the software ANNOVAR (Wang et al., 2010). Candidate variants were filtered on the basis of function (as predicted by ANNOVAR), and the 1,000 genomes (www.1000genomes.org) and NHLBI exome sequencing project (http://evs.gs.washington.edu/EVS/) frequencies.

3.3.2.3 Sanger sequencing

All exons and exon-intron boundaries of the *ROGDI* gene (NM_024589.2) were sequenced by Sanger sequencing in all affected and unaffected family members for whom a sufficient amount of DNA sample was available (primer sequences in table

S2). The results were analyzed with the Sequencher 4.1.4 software. We also sequenced one additional exon based on UCSC (uc010uxu.2, exon 2) to get s comprehensive analysis of all annotated coding exons in public databases. Mutations were named based on the sequences with accession numbers NM_024589.2 and NP_078865.1.

3.3.2.4 Fragment analysis

We performed fragment analysis of *ROGDI* intron 1-2 using FAM labeled forward primer (table S2). A touchdown PCR cycling protocol was used where the initial annealing temperature was lowered from 58 °C to 52 °C in 0.3 °C decrements each time for each cycle (table S2). Fragment length analysis was performed on an ABI 3730XL genetic analyzer (Applied Biosystems, Inc., Foster City, CA, USA), after incubation at 95°C for 5 minutes of 1µl of PCR product with 10µl of formamide and 0.5µl of LIZ, and analysed using ABI GeneScan v 3.7 (Applied Biosystems, Inc., Foster City, CA, USA).

3.3.2.5 Transcript analysis

cDNA was obtained by standard procedures from cultivated peripheral blood mononuclear cells (PBMC) of the affected child and both parents from family E and family J. Relative expression of ROGDI was quantified by real time PCR with specific primers spanning exons 3–5 and the comparative DDCt method using HPRT1 (RefSeq accession number NM_000194.2) as a reference gene, as previously described (Schossig et al., 2012a). Direct cDNA sequencing was performed in family E and family J by RT-PCR amplification of the entire *ROGDI* coding region with a forward primer in exon 1, a reverse primer in exon 11 and subsequent sequence analysis using internal primers. In family E, the forward primer in exon 1 was used in combination with a wildtype-specific reverse primer at the position of mutation c.366dupA in exon 6 (the affected child of family E had inherited this mutation from his father) to preferentially amplify transcripts from the maternal allele in the affected child. Subsequent sequence analysis was performed with the wildtype-specific reverse primer.

3.4 RESULTS

The linkage analysis previously carried out on five small families including family A, C, D, H and G (Lo, C.-N., 2009, Introduction) suggested significant linkage of a disease locus on the short arm of chromosome 16 (16p13) under a recessive model. To seek further evidence to support this result, we ran microarray analysis for SNP detection on affected individuals from family A and C. The three affected individuals shared a region of homozygosity on chromosome 16 (telomeric arm from 1 to 6,098,295 bp).

Given the large number of genes in the linked region (279 genes, bases on NCBI MapView: http://www.ncbi.nlm.nih.gov/projects/mapview/), we decided to perform WES. Five KTS samples in total underwent WES. We initially sequenced one proband from family C. Shortly after that we were offered by Aros -an independent company that provides services to universities- a trial run and we decided to perform WES on two additional samples: one affected sibling from family A and one from family D. The run was successful but due to an excess number of duplicate reads the coverage was low. Therefore Aros offered a free additional run, and we ran the other affected sample from family C and from family I (Figure 11, Table 12).

Using a frequency cut off of 1% (based on both the 1,000 Genomes project and the NHLBI exome variant server), we found the affected individual from family A had a single variant in the linked region on chromosome 16. This variant is homozygous, located in the *ROGDI* gene (1bp frameshift truncating deletion c.507delC, p.Glu170Argfs*72) and was not found in any database available to us. This variant was confirmed through Sanger sequencing and was shown to segregate within the family (Figure 13). The remaining exome data did not identify additional rare variants in *ROGDI*. An in depth analysis of exome data from other affected samples showed that, owing to limited sequencing depth, some of the *ROGDI* exons were not well covered in our exome sequence data (mainly exons 1 and 2) (Figure 12). We therefore decided to pursue *ROGDI* gene and Sanger sequence the two exons not covered.

At this time *ROGDI* mutations were independently identified in Innsbruck in three KTS families (Schossig et al., 2012). Subsequently, Sanger sequencing of all coding

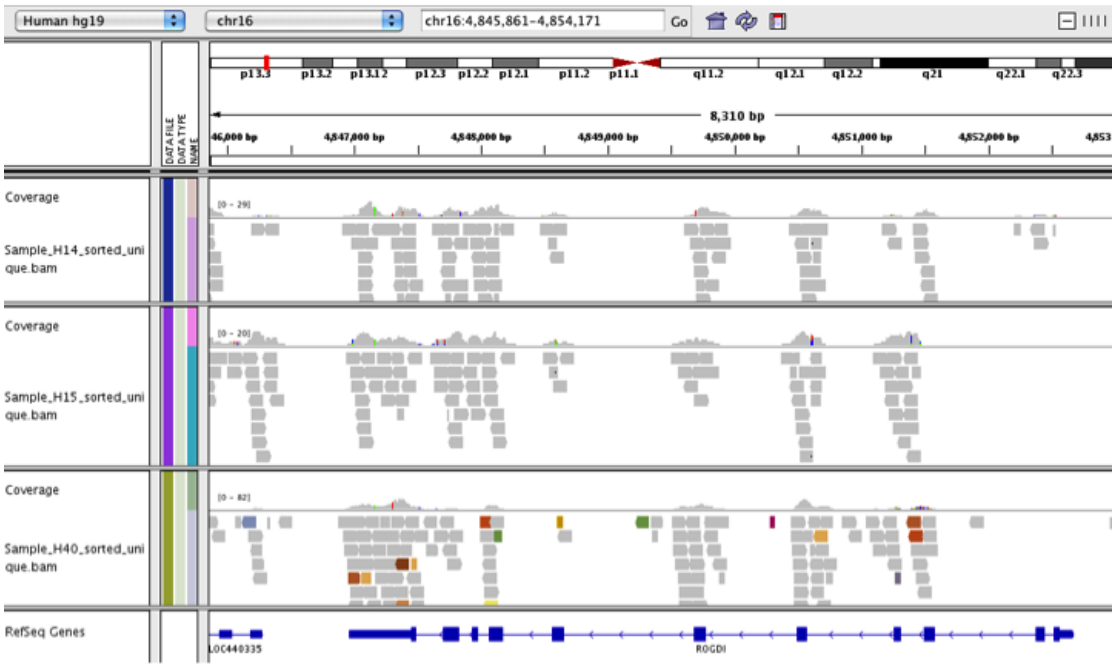
exons, exon-intron boundaries and the short introns 1, 3, 8, 9 and 10 of *ROGDI* was carried out in all samples from 7 families available in our laboratory (22 samples, of which 13 affected and 9 unaffected individuals) and in all affected members of 4 families collected in Innsbruck. Family D was investigated independently both in London and Innsbruck. Together, we identified *ROGDI* mutations in five families (families A, B, C, D and E) (Table 13). The affected individual from family B had the same mutation present in family A (c.507delC, p.Glu170Argfs*72). Affected individuals in families C and D carried different homozygous deletions in intron 1 that segregated with disease: c.46-37_46-30del and c.45+9_45+20del respectively. The mutations were heterozygous in the parents (Figure 13). The presence of these deletions was also confirmed through fragment analysis (Figure 14). The affected individual from family E was compound heterozygous for a 1 bp frameshift truncating duplication (c.366dupA, p.Ala123Serfs*19) and the same intron 1 deletion c.45+9_45+20del which was found in family D. The father was heterozygous for c.366dupA and the mother was heterozygous for c.45+9_45+20del. The intron deletion c.45+9_45+20del was assumed to alter mRNA splicing albeit this was not indicated by the splice prediction programs available through ALAMUT or Spliceview. Therefore, to further investigate the effects of c.45+9_45+20del we studied the mRNA transcripts in the cells from patients from family E. Quantitative RT-PCR showed a reduction of transcripts to about 55% and 60% in both the mother (heterozygous carrier) and the affected child (compound heterozygous) respectively when compared to controls (data not shown), while it did not show any reduction in the father. These findings indicate that the intron deletion c.45+9_45+20 is associated with markedly reduced mRNA production or stability while c.366dupA does not lead to nonsense mediated decay (NMD).

Table 12. WES summary metrics for KTS samples

	Sample I	Sample II	Sample III	Sample IV	Sample V
Family	**A**	**C**	**C**	**D**	**I**
Target capture kit	Illumina Truseq	Nimblegen SeqCap EZ	Illumina Truseq	Illumina Truseq	Illumina Truseq
Total number unique reads	7,870,652	171,506,673	35,822,484	10,002,954	28,014,742
% duplicated reads	86%	8%	39%	76%	36%
Reads aligned to target	59%	53%	24%	60%	41%
Mean target coverage	7,6	130,7	22,4	9,9	18,1
Tot number of variants	15959	16085	16231	15875	15359

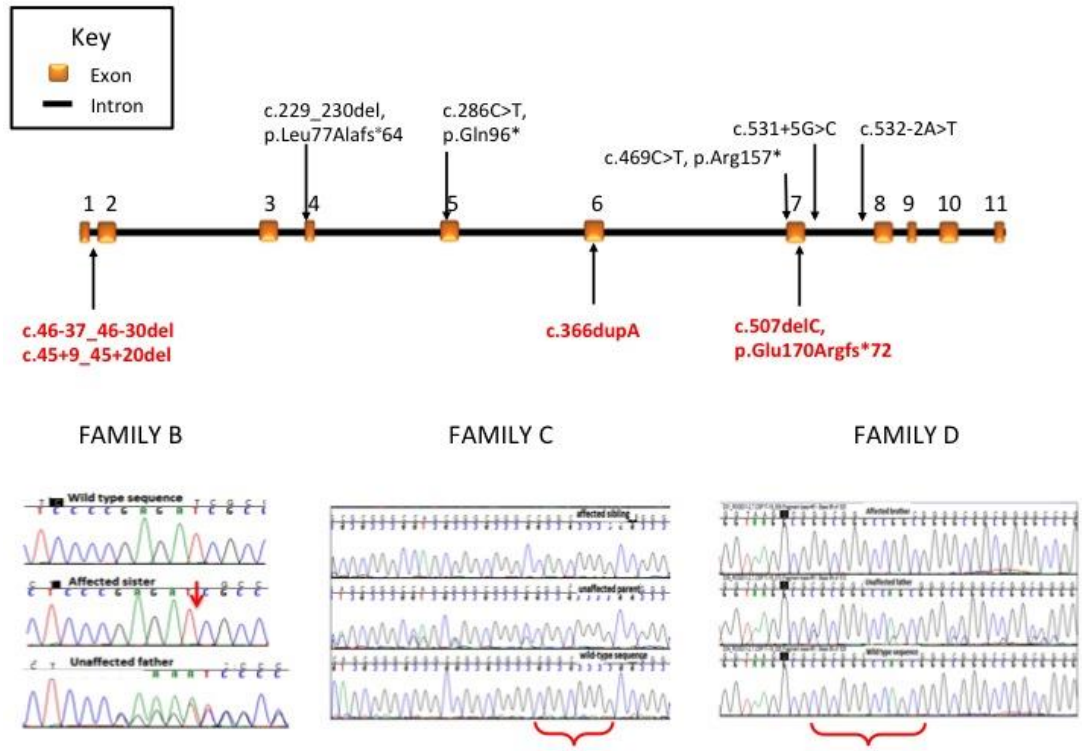
Target enrichment = method used to selectively capture the coding regions of the genome; total number of unique reads = number of reads that are not marked as duplicates after alignment; mean target coverage = mean coverage of targets that received at least coverage depth = 2 at one base

Figure 12. WES data at the *ROGDI* locus.



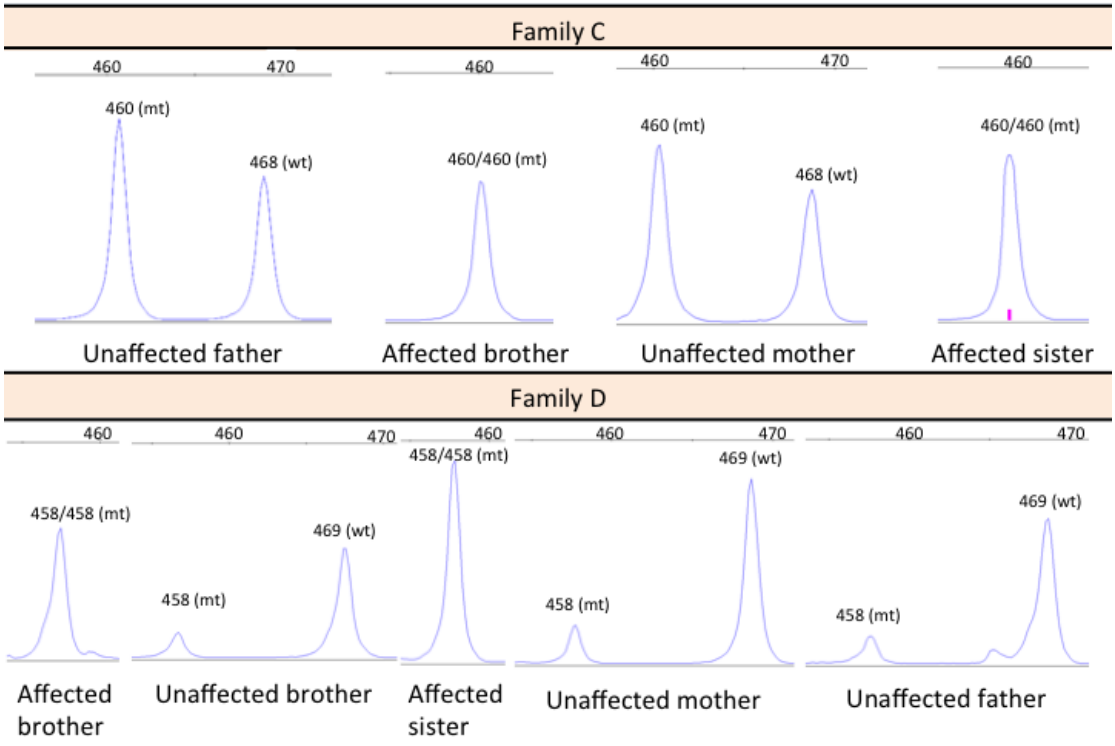
ROGDI gene structure in blue, (exons=rectangles, introns = line). Each gray line is a whole exome sequencing read. There are no reads covering exon 1 and exon 2.

Figure 13. *ROGDI* gene structure and mutations identified



Upper panel: gene structure and mutations reported in the Schossig and Mory papers. Red: novel mutations reported in the present study.
Lower panel: chromatograms showing the mutations identified in this study. Curly brackets indicate deleted nucleotides in affected members

Figure 14. Fragment analysis in families C and D



Numbers indicate fragment length.

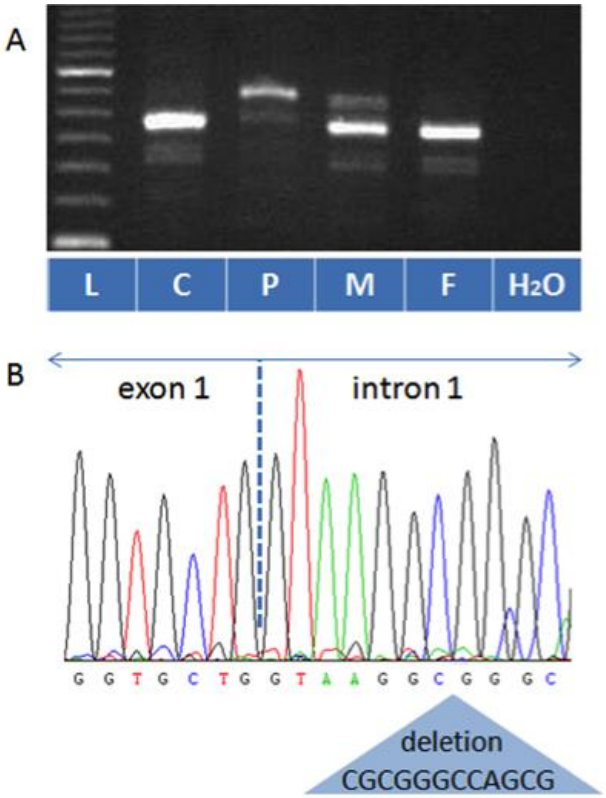
To investigate the effect of c.45+9_45+20del on pre-mRNA processing we designed an RT-PCR primer pair that specifically amplifies transcripts of the maternal allele in the affected child. We obtained and RT-PCR aberrant band approximately 60 base pairs longer than the normal amplicon. This band was not seen in the father or the control (Figure 15). Sequencing of the RT-PCR amplicon in the child showed that the aberrant transcript contains intron 1 with the 12 bp deletion (Figure 15), confirming that c.45+9_45+20 does not allow correct splicing of intron 1. This data suggest that the deletion c.45+9_45+20del prevents recognition of intron 1 by the splicing machinery and leads to frameshift and creation of a premature stop codon (p.Glu16Valfs*57) in exon 3.

We failed to identify *ROGDI* mutations in families F, G, H, I and J. To exclude the presence of large indels we ran SNP arrays on affected samples from family H and I. Over the *ROGDI* locus they were not homozygous nor showed any large indel (Figure 17).

Interestingly the proband from family I was found to have a 2.75Mb de novo duplication on chromosome 7 (hg18 coordinates chr7: 5,100,000-7,855,000), which was confirmed by aCGH array (Figure 16).

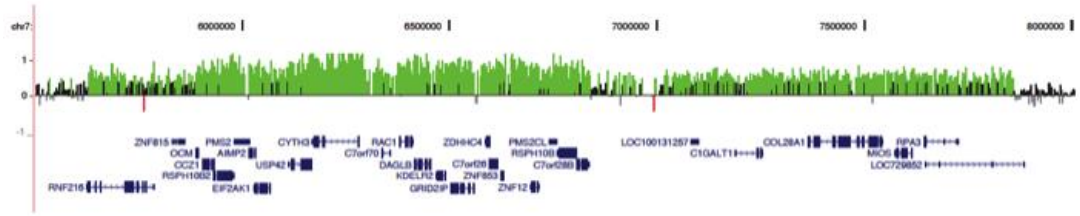
ROGDI expression levels were investigated in the affected from family J and his parents and were comparable to those in normal controls. Direct cDNA sequencing did not uncover a *ROGDI* mutation.

Figure 15. RNA analysis



(A) **Wildtype-specific rtPCR amplification in family E**. The mother (M) shows a strong 400 bp wildtype amplicon and an aberrant approx. 460 bp band. The wildtype band is also found in the father (F) and the control (C). The affected child (P) shows predominantly the aberrant 460 bp band. All samples have an additional band of approx. 340 bp which represents skipping of exon 4 in a small fraction of transcripts as shown by rtPCR sequencing (data not shown). L = Size standard. (B) **rtPCR sequence analysis in the affected child of family E.** The mutation c.45+9_20del causes inclusion of intron 1 between unaltered exons 1 and 2 in the ROGDI transcript.

Figure 16. Duplicated region on chr.7 from family I



Oligo array CGH results for the affected child from Family I shown in UCSC genome browser (chr7:5,500,000-8,000,000, NCBI Build 36, hg17). Deviations of probe log2 ratios from zero are depicted by vertical grey/black lines, with those exceeding a threshold of 1.5 standard deviations from the mean probe ratio colored green and red to represent relative gains (duplications) and losses (deletions), respectively. Genes are depicted in blue below the CGH data

Table 13. Summary of analyses and results for each KTS family
Core features = epilepsy, developmental delay and amelogenesis imperfecta

Family ID	*Phenotype*	*WES*	*Linkage and homozygosity*	*ROGDI Sanger sequencing*	*Fragment and cDNA analysis*	*ROGDI mutations*
Family A	Core Features	One affected proband	Linkage	Yes	N/A	Homozygous c.507delC
Family B	Core Features	N/A	N/A	Yes	q RT-PCR	Homozygous c.507delC
Family C	Core Features	Both affected siblings	Linkage and homozygosity mapping	Yes	Fragment segregation	Homozygous c.46-37_46-30del
Family D	Core Features	One affected proband	Linkage	Yes	cDNA and Fragment segregation	Homozygous c.45+9_45+20del
Family E	Core Features	N/A	N/A	Yes	q RT-PCR, cDNA amplification, and sequence analysis	Compound heterozygous c.366dupA/c.45+9_45+20del
Family F	Core Features + Spasticity	N/A	N/A	Yes	N/A	None
Family G	Atypical KTS	N/A	Linkage	Yes	N/A	None
Family H	Core Features+ feeding problems +neurofibromatosis 1	N/A	Linkage	Yes	N/A	None
Family I	Core Features+ dysmorphic signs+ multiple warts	Single affected member	N/A	Yes	N/A	None
Family J	Core Features	N/A	N/A	Yes	q RT-PCR, cDNA amplification, and sequence analysis	None

3.5 DISCUSSION

In this study, WES failed to identify the causal mutation in three samples from two families because the mutation is intronic. WES has been shown to detect SNVs in the introns, especially for Trusequ capture kit (this is because the bait often extends farther outside the exon target) (Clark et al., 2011). In this study WES did not sequence the intron (albeit very small, 73bp) and the flanking exons. This is probably due to the high GC content of the region.

The initial identification of a *ROGDI* mutation in family A did point to a causal role of this gene in KTS, but the concurrent absence of other *ROGDI* mutations in the remaining families initially slowed down our analysis. In this case, the absence of mutations was due to the lack of sequencing reads in some *ROGDI* exons. Only an in-depth inspection of WES data over the *ROGDI* gene allowed us to identify regions that had not been sequenced. This result highlights the importance of a detailed analysis of WES data over a candidate region.

Traditional Sanger sequencing of the regions not covered by WES, allowed us to identify the intronic deletions in families C, D and E. RNA analysis further confirmed their splicing- altering role and the consequent premature truncation of the ROGDI protein in the patients cells. Of note, splice site prediction programs (like ALAMUT and Spliceview) did not predict the intronic deletion to alter mRNA splicing.

The location of the two intronic deletions in positions distant to the exon-intron boundaries could indicate that they affect splicing through interfering with the stem-loop formation (Stallings and Moore, 1997).

The identification of homozygous mutations in the *ROGDI* gene in five families (family A, B, C, D and E), all segregating with disease, supports the fact that mutations in the *ROGD*I gene cause KTS and confirm the results of Schossig et al and Mory et al 2012. The families carrying the mutation had typical KTS, and three of them had previously been linked to chromosome 16 within which the *ROGDI* gene lies and have likely loss of function *ROGDI* mutations.

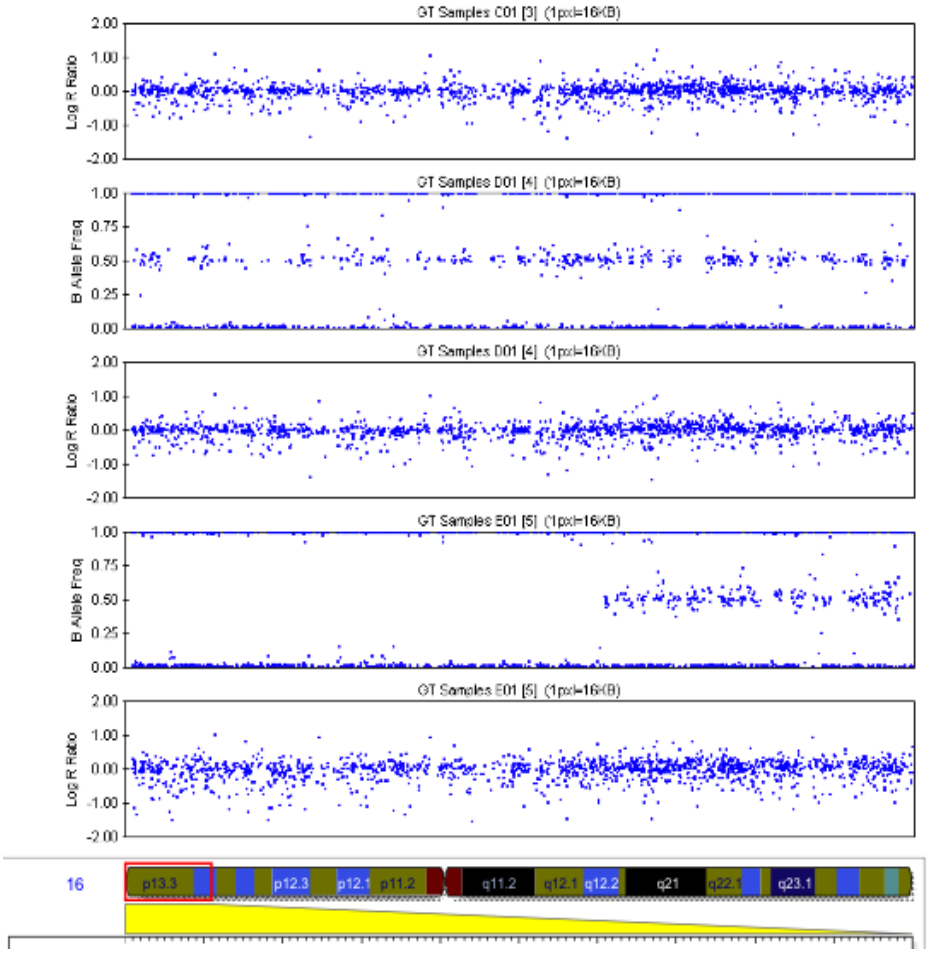
Conversely, in five KTS families (families F, G, H, I and J) no mutations were found in the *ROGDI* gene. Clinically these families had atypical KTS phenotype as they had additional features. Family F had spasticity. Family G presented late, had a likely dominant pattern of inheritance and teeth that were amelogenic but not truly yellow. Family H had all the features of KTS but were unusual, in that they also had feeding difficulties and Family I had additional dysmorphic features. These clinical differences reflect that so far *ROGDI* mutations have been identified only in typical KTS and even though the phenotype is complex there seems to be a relatively tight pattern of clinical signs associated with this gene.

ROGDI mutations were excluded by exome and Sanger sequencing, where all exons of the gene were amplified and sequenced in all samples. No heterozygous or homozygous changes were seen throughout the *ROGDI* gene making this gene unlikely to cause atypical forms for the disease, although we cannot rule out difficult to detect deep intronic mutations or compound heterozygous exonic copy number deletions. Nevertheless, this is improbable as haplotype analysis on chromosome 16p13.3 failed to show a shared haplotype on chromosome 16p13.3-13.2 in family G and H (Figure 18). Moreover, in family J additional rtPCR sequencing performed in the affected individual with primers spanning exon 1-11 and quantitative rtPCR showed no reduction of mRNA in the family members. Of note, none of the *ROGDI* negative families showed consanguinity, making the gene(s) causing atypical forms of KTS only possibly recessive. With this in mind, de novo mutations in other genes could be causing atypical phenotypes of KTS. Indeed, the proband from family I has been extensively investigated at a genetic level (WES and CHG array) and was found to have a de novo 2.75Mb duplication on chromosome 7, comprising about 60 genes (NCBI mapviewer). It is likely that a dosage effect of one of the genes in this region is causing a ROGDI like phenotype although this was not seen in any of the other families. WES in this sample showed no defects in this region.

In summary, this work confirms the role of *ROGDI* mutations in typical KTS. This work also exemplify the importance of sequencing non-coding regulatory regions, at least to within 20 bp of the exon for the analysis of Mendelian disorders.

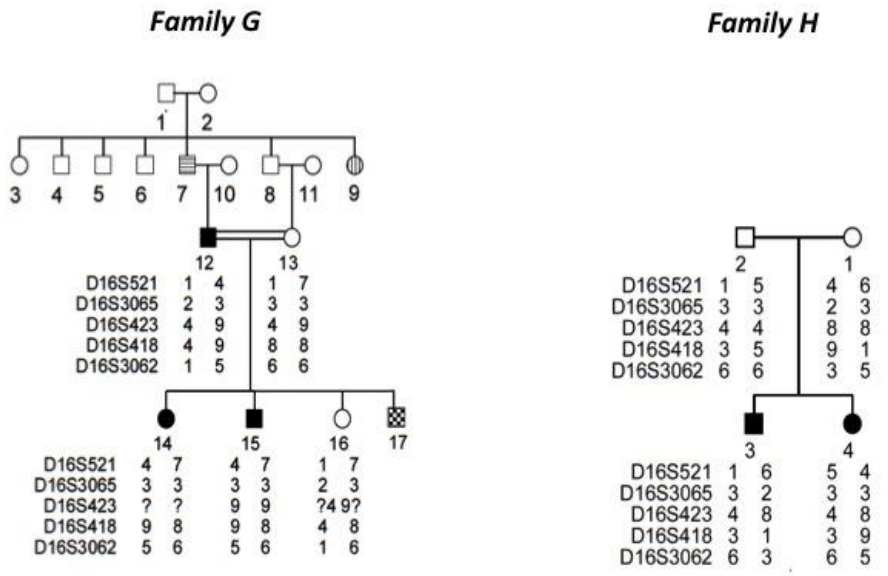
A number of families with atypical KTS phenotype were identified, who were negative for *ROGDI* mutations and failed to show homozygosity in the *ROGDI* region, suggesting genetic heterogeneity of atypical forms of disease.

Figure 17. Chromosome 16 linked region (1-10Mb)



Genome Viewer tool, showing individuals from family H (sample C1) and I (sample D1) are heterozygous, individual from family A (sample E1) block of homozygosity (1bp to 6Mb).

Figure 18. Haplotype analysis of families G and H



This is part of a PhD thesis by Lo CN (Lo, C.-N., 2009) and shoes that the affected individuals from families G and H do not share that same haplotype on the chr 16 region, suggesting genetic heterogeneity of the disease.

4 *C12orf65* MUTATIONS CAUSE AXONAL NEUROPATHY WITH OPTIC ATROPHY

4.1 STATEMENT OF CONTRIBUTION TO THIS RESEARCH

I performed autozygosity mapping and WES data analysis of a family presenting with axonal neuropathy and optic atrophy and I was involved in the laboratory based DNA sequencing and RNA analysis experiments. I made contacts with people working on mitochondrial biology, asked them to run experiments to further confirm our results and coordinated different people working on different aspects of mitochondrial function. I collected the cohort of patients to screen for the gene we found mutations in. I wrote the original manuscript that has been submitted to the Journal of Medical Genetics journal.

4.2 BACKGROUND

CMT type 6 is characterized by axonal neuropathy and optic atrophy and is genetically heterogeneous. Mutations in Mitofusin 2 have been found to cause dominant forms of CMT6, whilst until now, the genetic cause of the recessive form has remained unknown.

We describe a family with CMT6, beginning in late childhood and with a slow progression. By performing homozygosity mapping followed by WES of two affected cousins, we identify a homozygous protein-truncating mutation in the *C12orf65* gene. *C12orf65* encodes for a 166 amino acid protein, involved in mitochondrial translation. We show mitochondrial impairment in the patients lymphoblast cell lines by multiple lines of evidence: decrease of complex V activity

and a defect in stability pattern, decrease mitochondrial respiration rate as well as reduction of mitochondrial membrane potential.

This work describes for the first time the genetic defect in a recessive form of CMT6 and confirms the role of mitochondrial dysfunction in this complex axonal neuropathy.

4.3 METHODS

4.3.1 Samples

DNA samples from an Indian family with CMT6 and from 93 additional patients affected by complex neuropathy (neuropathy plus one of the following: optic atrophy, retinitis pigmentosa, psychomotor retardation, cerebellar ataxia, or pyramidal signs) were collected by Prof. Henry Houlden at the National Hospital for Neurology and Neurosurgery (NHNN), London. All patients had been previously screened for known mutations in the following genes PMP22, GJB1, MFN2, MPZ, GDAP1, BSCL2, TRPV4, NEFL, HSPB1, HSPB8, or GARS.

We also agreed to collect 90 additional samples to screen for *C12orf65* mutations from the Neurometabolic Unit (NMU) at the NHNN, a designated supra-regional centre in the UK for the investigation of neurometabolic disorders, with special interest in mitochondrial disease. Samples were chosen from their database, based on the following criteria: presence of a decreased complex V activity or combined oxidative phosphorylation enzymes deficiency in the blue native gel and/or clinical presentation with neuropathy/optic atrophy/ataxia or a Leigh-like phenotype.

Genomic DNA was purified from peripheral blood cells using standard procedures.

4.3.2 Nerve biopsy

The nerve biopsy soon after the surgical removal were immersed in a fixative for overnight containing 3% buffered glutaraldehyde and 0.2M sodium cacodylate buffer. Then specimens were cut with razor in 1-2 mm thick pieces and osmicated in secondary fixative - osmium tetroxide. After fixation the specimens were impregnated into epoxy resin from which semi-thin (~1µ) sections or ultra-thin

(~70nm) sections were cut and put on glass slides or grids, respectively. Semi-thin sections were stained either with Methylene blue - Azure A and basic fuchsin or Toluidine blue and examined by light microscopy on various magnifications. Ultra-thin (~70nm) sections were stained with Uranyl Acetate/Lead Citrate and examined by electron microscopy at various magnifications.

The paraffin blocks of the sural nerve biopsy from case 1, performed when he was investigated at the time of his deterioration, were kindly made available and examined by Prof. Sebastian Brandner. He also compared Case 1 biopsy to a *MFN2* mutation positive and a patient not carrying known mutations.

4.3.3 SNP genotyping and autozygosity mapping

Genome-wide SNP genotyping was performed in two affected family members of the family (case 1 and case 3) at NIH (Bethesda, US). Each individual was assayed on an Illumina chip (HumanHap300 BeadChip), yielding approximately 317,000 SNPs. Samples were processed, hybridized, and scanned following the instructions of the manufacturer. Clustering, normalization, and genotype calls were performed using the GenomeStudio 2010.3 Genotyping Module (Illumina). Autozygosity mapping was performed using the Homozygosity Detector plug-in software within the BeadStudio suite. Regions of shared homozygosity that segregated with disease were visually identified using the Illumina Genome Viewer tool within the BeadStudio suite.

WES was carried out on case 1 and case 3 at NIH (Bethesda, US). Nimblegen SeqCap EZ Exome (in solution capture) kit for was used for the exome capture. Shotgun sequencing libraries were generated from three µg of DNA from each individual. Sequencing was performed on a Genome Analyzer IIx, according to the manufacturer's instruction.

Raw sequencing reads were aligned to the hg18 build of the reference genome using the software Novoalign. Calling was performed using Samtools 0.18 and the resulting calls were annotated using ANNOVAR (as discussed in Chapter 2).

4.3.4 Mutation validation and screening in the additional cohorts

Primers for PCR amplification were designed by Generunner (http://www.generunner.net/). Primer sequences and the PCR thermocycling are listed in the supplementary table S3. The PCR reaction mix consisted of 30ng genomic DNA, 5nM forward primer, 5nM reverse primer and 12µl of FastStart PCR Master (www.rocheapplied-science.com). Unincorporated dNTPs, primers, salts and polymerase were removed using Multiscreen PCR Cleanup Filter Plates (Millipore, MA, USA) as per the manufacturer's instruction. After PCR amplification, coding regions and exon-intron junctions of *C12orf65* gene were sequenced by Sanger's sequencing, using the Big Dye Terminator cycle sequencing kit v3 (Applied Biosystems, Foster City, California, USA), and analyzed by a 3130 Genetic Analyser sequencing machine (Applied Biosystems). Sequences were examined in silico for mutations by Sequencher software 4.9 (Gene Codes Corporation, Ann Arbor, MI, USA).

4.3.5 Lymphoblast cells cultures

Lymphoblasts cells were obtained from case 1 and two unaffected relative carriers. Lymphoblastoid cell lines were established by Epstein–Barr virus transformation of lymphocytes isolated from peripheral blood. Cell lines were stored at the European Collection of Cell Cultures (ECACC). Informed consent was obtained from the patient and his relatives. Patient and control lymphoblasts were thawed and maintained in culture in modified RPMI-1640 medium containing 300mg/L L-Glutamine and HEPES (Invitrogen) supplemented with 10% heat inactivated FBS (Invitrogen) at 37°C, 5% CO2. Fresh medium was added every 3 days and cultures were expanded accordingly.

4.3.6 Transcript analyses

Purification of total RNA from patients' lymphoblasts was performed using the Qiagen miRNeasy Mini Kit, Hilden, Germany (catalogue #217004). Cell pellets were prepared by centrifuging 3–4 x 106 cells for 5 min at 300xg; all supernatant was removed by aspiration. Then, to disrupt cells bodies, 700µl of QIAzol Lysis Reagent

was added to the tube, mixed to the cell pellet, and incubated at room temperature for 5 minutes. 140µl chloroform were added to the tube containing the homogenate and mixed thoroughly by shaking the tube for 15s. The tube was incubated at room temperature for 5 minutes. Then the tube was centrifuged for 15 min at 12,000 x g at 4°C. After centrifugation, the sample separated into 3 phases, with the upper phase containing RNA. The upper aqueous phase (about 350µl) was transferred to a new collection tube and 525µl of 100% ethanol were added to the tube and mixed thoroughly. 700µl of the sample were pipetted into an RNeasy Mini spin column in a 2 ml collection tube and centrifuged at 10,000 rpm for 15 s at room temperature; the flow-through was discarded. Then 500µl Buffer RPE were added into the RNeasy Mini spin column; the tube was and centrifuged for 15 s at 10,000 rpm to wash the column and the flow-through was discarded. Another 500µl Buffer RPE were added to the spin column and centrifuged for 2 min 10,000 rpm to dry the column membrane. Then the spin column was transferred to a new 1.5 ml collection tube. 30µl RNase-free water were added directly onto spin column membrane and centrifuged for 1 min at 10,000 rpm to elute the RNA; this step was repeated two times. RNA concentration was then measured with spectrophotometer (all samples had a concentration of approximately 1000ng/µl) and quality was assessed evaluating 260/280 nm and 260/230 nm ratio. cDNA synthesis was performed using Qiagen Omniscript RT Kit, Hilden, Germany (catalogue #205110). Briefly, for each reaction, a fresh master mix was prepared containing: 10x Buffer RT 2µl, dNTP mix 2µl, Random primers 1µl, RNAse inhibitor 1µl, Ominscript RT 1µl and RNAse-free water 11,5µl. 1,5µl of template RNA and was added to the mix and then incubated for 60 min at 37°C.

Multiplex quantitative real-time PCR assays for the levels of C12orf65 and RPL13A were performed using SYBR Green PCR Master Mix kit (Applied Biosystems) with a Corbett Rotor-Gene real-time quantitative thermal cycler (Corbett Research/Qiagen). Thermal cycling consisted of 10 minutes at 94°C for initial denaturation and DNA polymerase activation, followed by 40 cycles of denaturation at 94°C for 15 s and annealing/extension at 60°C for 1 min. Each assay contained template negative controls and a quantitative standard curve dilution run in duplicate. All reactions were performed in triplicate in a final volume of 10 µl.

Gene-specific primers are listed in table S3. Dissociation curves were run to detect nonspecific amplification, and single amplified product was confirmed in each reaction. Standard curves showed that all duplexed assays had equal efficiencies (99.9%), satisfying criteria for the comparative Ct method of quantification. The quantities of each tested gene and internal controls were then determined from the standard curve using the Rotor-Gene 6000 series software 1.7 (Corbett Research/Quiagen). Relative quantity of each C12orf65 level normalized by the RPL13A level was then calculated. The results were expressed as percentage of control.

4.3.7 Blue native in-gel complex V assay

Blue native gel (BN gel) electrophoresis was used to examine mitochondrial respiratory chain protein levels, and complex V (ATP synthase: EC 3.6.3.14) activity.
Briefly, a mitochondrial fraction was obtained from the lymphoblastoid cell pellets (7.5 × 105 cells) using two low speed centrifugation steps (600 g × 10 min) at 4 °C, separated by a homogenisation step (30 strokes). The supernatant from each spin was combined and a higher spin (14,000 g × 10 min) was performed in order to eliminate nuclei and other subcellular membranes. Then the mitochondrial membranes were solubilised with a 750 mM amino hexanoic acid/50 mM Bis Tris buffer + 4 % n-dodecyl β-D maltoside detergent. Samples were left on ice for 30 minutes and a further high spin (14,000 g × 10 min) was used to pellet insoluble material.
An equal quantity of mitochondrial protein (30 µg) was loaded from each sample (case 1, two unaffected carriers and two controls). The BN gel was run as previously described with slight modifications (Wittig et al., 2006), using a 3-12 % Bis-Tris gel (Invitrogen) to ensure discrete separation of complex V. The complex V activity was measured in a reverse direction, where ATP is hydrolyzed into ADP and Pi (inorganic phosphate). The lead ions in the buffer combine with Pi, which results in the accumulation of Lead phosphate, a grey precipitate where the complex V band is present. Complex V assay was performed by incubating the gel overnight in stain containing 34mM Tris, 270mM glycine, 14mM magnesium chloride, 6mM Lead (II) nitrate and 8mM ATP.

4.3.8 Oxygen Consumption

Oxygen consumption was measured to investigate the mitochondrial respiration rate.

To measure mitochondrial respiration rate in intact cells, approximately 1x107 cells were suspended in respiration medium (HBSS, with 10 mM D-glucose) in a Clark-type oxygen electrode thermostatically maintained at 37ºC. The oxygen electrode was calibrated with air-saturated water, assuming 406 nmol O2 atoms/ml at 37ºC. Oxygen consumption was measured over 10 minutes with addition of oligomycin (final concentration 2µg/ml) and FCCP (0.5 µM). All data were obtained using an Oxygraph Plus system (Hansatech Instruments, UK) with chart recording software. For measurements of mitochondrial membrane potential (Δψm), cells were loaded with 25 nM tetramethylrhodamine methylester (TMRM) for 30 min at room temperature in HBSS (156 mM NaCl, 3 mM KCl, 2 mM MgSO4, 1.25 mM KH2PO4, 2 mM CaCl2, 10 mM glucose, and 10 mM HEPES, pH adjusted to 7.35), and the dye was present during the experiment. TMRM is used in the redistribution mode and therefore a reduction in TMRM fluorescence represents Δψm depolarization. Z-stack images were obtained for accurate analysis. The values for WT were set to 100% and the other genotypes were expressed relative to WT (Yao et al., 2011; J Cell Science).

Confocal images were obtained using a Zeiss 710 LSM with a 40x oil immersion objective. TMRM was excited using the 560 nm laser and fluorescence measured above 580-nm.

4.4 RESULTS

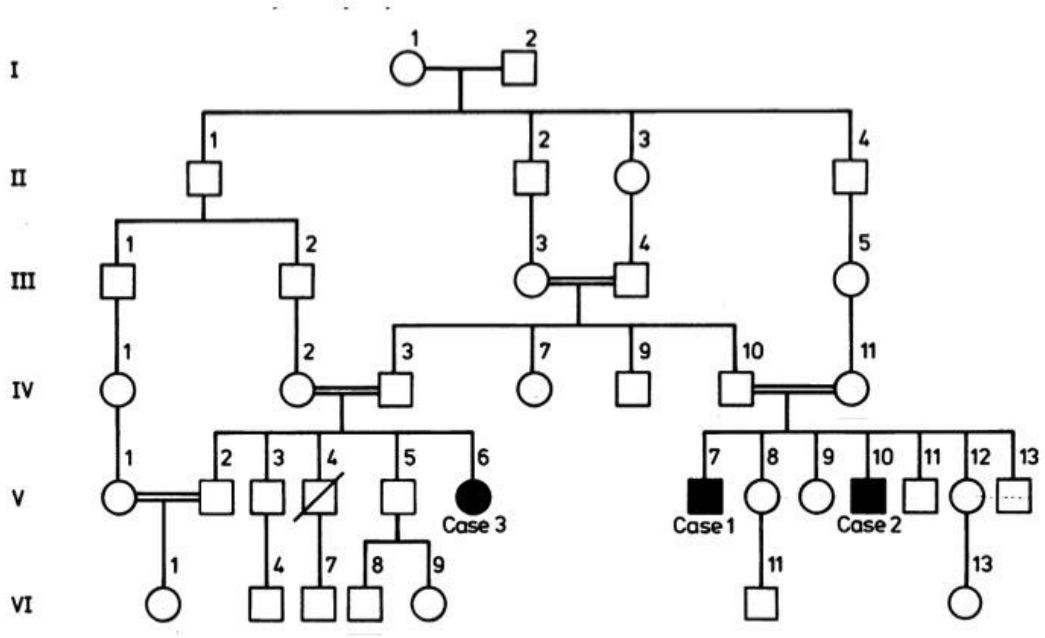
4.4.1 Clinical details

Three members of a large consanguineous Indian family (Figure 19), affected by axonal neuropathy and optic atrophy, were identified at the National Hospital for Neurology and Neurosurgery. Briefly, they all presented with a similar syndrome characterized by a very slowly progressing axonal neuropathy (onset in childhood), bilateral optic atrophy and pyramidal signs. Case 3 was first seen at the age of 35

years old and then again at the age of 51 years. She is the first cousin of cases I and 2 and presented as a child with delayed milestones, later at the age of 9 years old she had lower limb and at 11 years upper limb weakness along with static cognitive problems and visual difficulties. The examination at the age of 35 and 51 years were remarkably similar with only mild deterioration of her clinical features. She had no dysmorphic features, but severe bilateral optic atrophy and static, but significant cognitive problems. She had a brisk jaw jerk and a pout reflex. Marked distal symmetrical weakness and wasting affecting the limbs. In contrast to her cousins, tone in her upper and lower limbs was reduced with severe distal weakness in upper and lower limbs. Upper limb reflexes were normal. Knee and adductor reflexes were abnormally brisk. Ankle and plantar reflexes were absent. She had a moderately severe thoracic scoliosis. Sensation was impossible to assess reliably due to cognition but was likely to be abnormal. As a crude estimate of visual acuity, they were able to watch television and recognise faces across the room and she could do needlework.

Case 1 noted distal wasting and weakness at the age of 8, which slowly spread proximally. When he was examined (at the age of 34) he had severe muscle wasting of lower and upper limbs, bilateral optic atrophy and macular colloid bodies. Pyramidal signs were present in the upper limbs. Case 3 presented with a similar syndrome characterized by a very slowly progressing axonal neuropathy (onset in childhood), bilateral optic atrophy and pyramidal signs. This CMT6 family were originally reported as a case report in 1987 when the family members were in their middle 30s (MacDermot and Walker, 1987).

Figure 19. Pedigree of the family

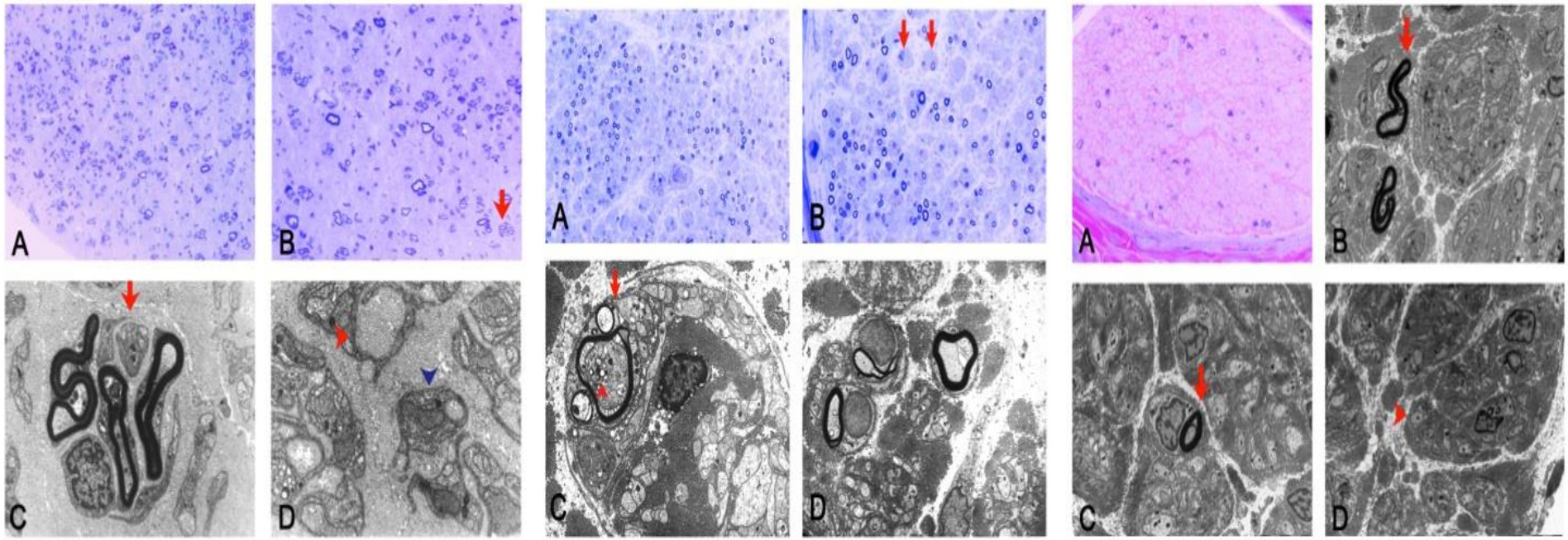


A square represents a male person, a circle represents a female. A double horizontal line indicates parental consanguinity. Black symbols indicate affected members with CMT2 and optic atrophy in whom a neurological exam was performed; blank symbols show unaffected family members. A diagonal line marks deceased individuals. This pedigree has been reported elsewhere (MacDermot and Walker, 1987).

Nerve Biopsy

The light microscopy of the sural nerve biopsy from case 1 showed a marked reduction of large myelinated fibres. The fascicles were populated by numerous small myelinated fibres majority of which were associated with regeneration clusters. Ultrastructural assessment on electron microscopy confirmed the absence of both active and chronic demyelination and showed no evidence of ongoing axonal degeneration. The remaining myelinated fibres revealed normal configuration of the myelin. Although the assessment of loss of unmyelinated fibres was confounded by frequent unmyelinated axonal sprouts associated with regeneration, the evidence of decreased numbers of unmyelinated fibres was indirectly confirmed by increased amounts of endoneural collagen with formation of collagen pockets amongst flattened Schwann cell profiles. The depicted mitochondria on the transverse sections of nerve fascicles showed no obvious ultrastructural abnormality (Figure 20).

Figure 20. Sural nerve biopsies



Sural nerve biopsies (light and electron microscopy). Case 1. Semi-thin resin section stained with toluidine blue (A and B) show markedly reduced numbers of large myelinated fibres accompanied by numerous small myelinated fibres mainly associated with regeneration clusters (arrow in B). Ultrastructural assessment (C and D) further highlights frequent regeneration clusters (arrow in C). The evidence of decreased numbers of unmyelinated fibres is confirmed by increased amounts of endoneural collagen and the formation of collagen pockets (read arrowhead in D) amongst flattened Schwann cell profiles (blue arrowhead in D). Scale bar: 60µm (A), 30µm (B), 2µm (C and D). Patient with MFN2 mutation. Semi-thin resin section stained with toluidine blue (A and B) show markedly reduced numbers of large myelinated fibres accompanied by occasional regeneration clusters and occasional fibres surrounded by concentric Schwann cell profiles forming onion bulb-like structures (arrows in B). Ultrastructural assessment (C and D) confirms that myelinated fibres are markedly reduced in numbers and shows no evidence of active or chronic demyelination; instead it reveals pseudo-onion bulbs indicative of regeneration (arrow in C). The depicted mitochondria on the transverse sections of the nerve fascicles show no obvious ultrastructural abnormality; although some intra-axonal clusters of mitochondria are occasionally observed (arrowhead in C). Scale bar: 80µm (A), 40µm (B), 5µm (C and D). Patient with no known mutation. Semi-thin resin section stained with methylene blue azure – basic fuchsin (MBA-BF) (A) shows markedly reduced numbers of large myelinated fibres with no apparent evidence of regeneration and no evidence of active macrophage-associated demyelination or chronic demyelinating/re-myelinating process. Electron microscopy (B, C and D) confirms the markedly reduced numbers of large and small myelinated fibres (arrows in B and C). The unmyelinated fibres in contrast are better preserved (arrowhead in D). Scale bar: 40µm (A), 5µm (B, C and D).

4.4.2 Genetic analyses

Autozygosity mapping was carried out on two affected cousins (case 1 and case 3) with the aim to fine map homozygous by descent (HBD) regions potentially containing the recessive disease gene. SNP analysis identified five HBD chromosomal segments of significant length (>1Mb), which were concordant in both cousins (Table 14, Figure 17) encompassing 15.8Mb and containing 226 genes (NCBI build 37.2, map viewer). Given the large number of candidate genes, we used WES to perform a comprehensive search for pathogenic mutations in both affected cases. Alignment metrics on WES data showed good coverage for all samples (Table 15).

Table 14. Homozygous regions >1Mb concordant in case 1 and case 3

Chr	Start	End	size (bp)
Chr12 Coord	rs2433345 119575350	rs12369591 125764947	6,251,757
Chr1 coord	rs2764654 77074588	rs11576605 81336822	4,262,234
Chr10 coord	rs224731 34226309	rs1480524 36513472	2,287,163
Chr20 coord	rs6087487 32015038	rs6058150 33556817	1,541,781
Chr5 Coord	rs7726515 129736249	rs2240525 131343783	1,469,711

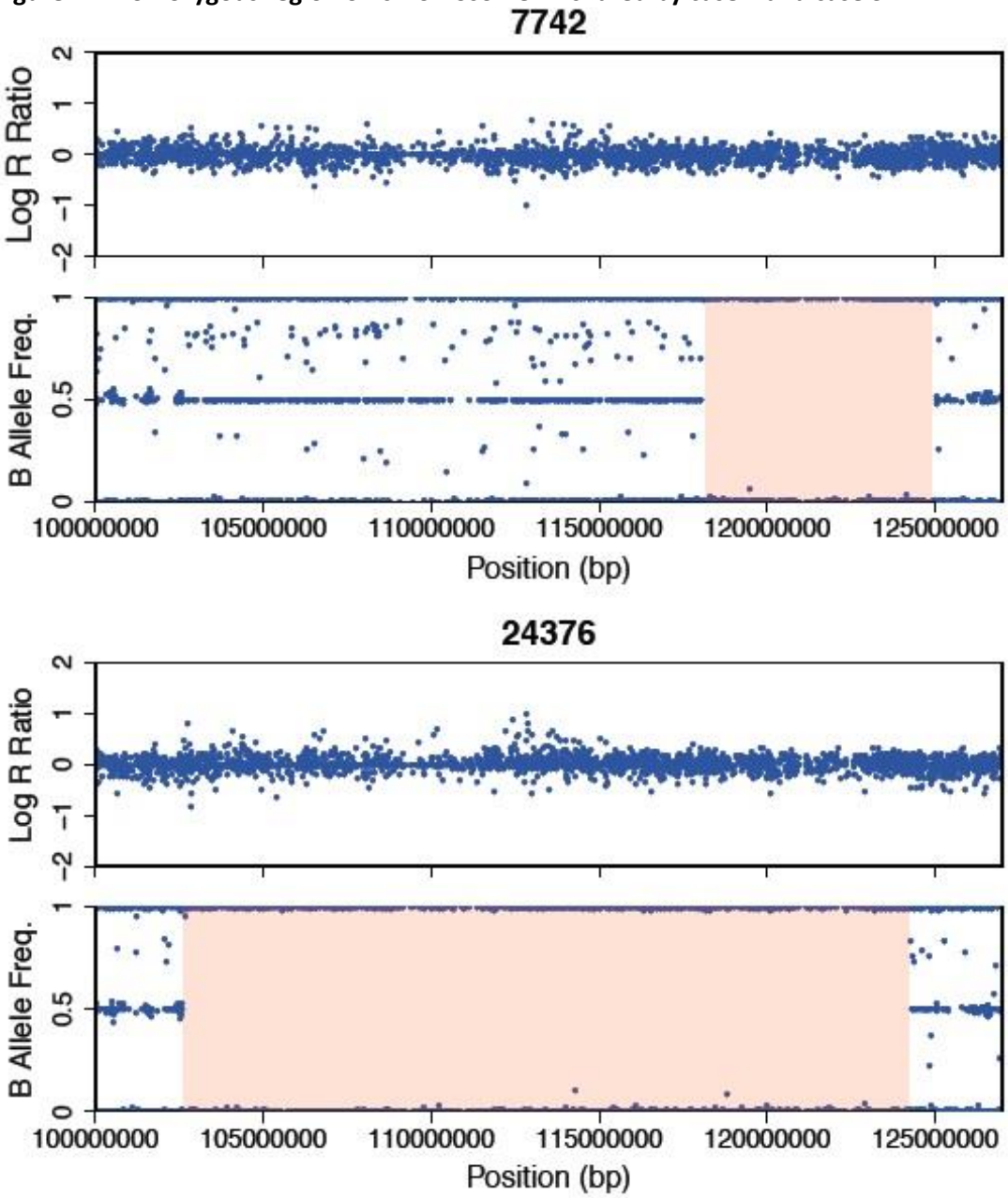
Table 15. Summary metrics of WES

	case 1	case 3
Total number of reads	75403508	72122880
Non-duplicated reads	35%	41%
Reads aligned to target	47%	44%
Mean target coverage	26,2	27,6
Tot number of variants	12965	13144

To identify potential causal variants we selected all coding variants in the homozygous regions present in both samples. Then we filtered them by discarding (i) variants documented in the Single Nucleotide Polymorphism Database and the 1000-Genomes Project and (ii) synonymous substitutions. Five variants passed these filters: two were missense, one was a non-frameshift deletion and one was nonsense mutation. The latter is a 1 bp deletion in *C12orf65* gene, resulting in a premature stop codon (NM_001143905:c.346delG: p.V116X). As this was the most protein-damaging variant, and as mutations in *C12orf65* had been described to cause a severe encephalomyopathy (Antonicka et al., 2010), we thought it was the best candidate to follow up. The presence of this variant and its segregation with disease in this family was confirmed with Sanger sequencing: the deletion was homozygous in the affected cases, and either absent or heterozygous in the unaffected relatives (Figure 17). To further investigate the presence of the mutation in patients with complex CMT, we sequenced the coding exons of *C12orf65* in an additional cohort of 93 patients. None of the patients harbored potentially pathogenic variants.

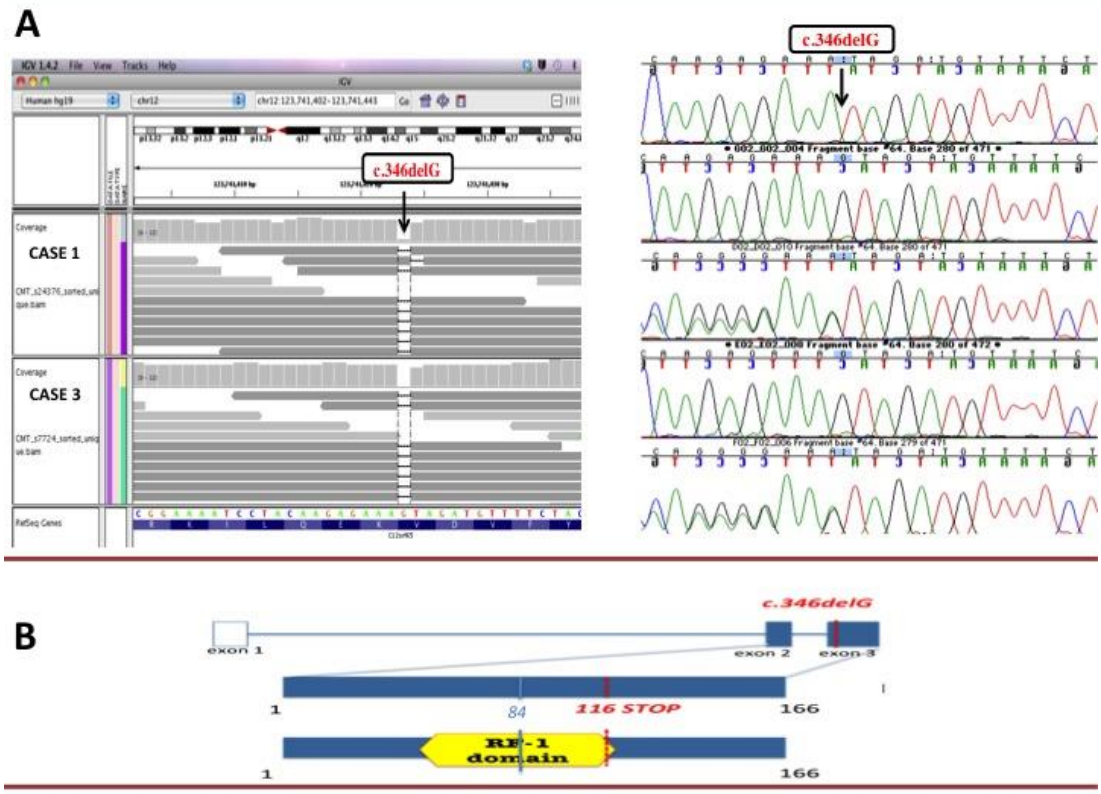
To determine whether the truncating mutation affected mRNA stability and induced mRNA nonsense mediated decay (NMD), we tested the level of expression by qPCR. This showed that level of the *C12orf65* mRNA was not reduced in the lymphoblasts from affected individuals versus controls (Figure 23) as predicited by the ’55 nucleotides rule’, an hallmark of mammalian NMD that predicates that stop codons located at least 55 nucleotides upstream of the last exon junction will be interpreted as 'premature' and trigger NMD (Nagy and Maquat, 1998).

Figure 21. Homozygous region on chromosome 12 shared by case 1 and case 3



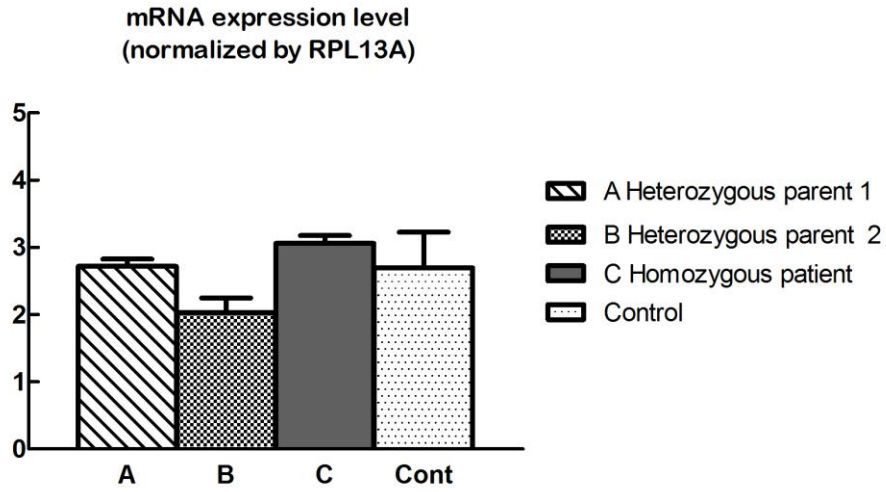
Upper panel, Log R ratio and B allele frequency metrics for Case 1. Lower panel, Log R ration and B allele frequency metrics for Case 3. Bounded in pink is the primary candidate interval identical by descent between both affected samples. Stretches of homozygosity are denoted by a contiguous stretch of genotypes where B allele frequency is either 0 or 1.

Figure 22. The *C12orf65* p.V116X mutation



A) Left panel: WES data for the p.V116X mutation in *C12orf65*. The aligned reads viewed through the IGV viewer (http://www.broadinstitute.org/igv/). Reads are depicted as arrows (grey bands). A coverage histogram per base is shown above the reads. RefSeq gene (over the *C12orf65* gene) is represented in the lower part both as aminoacid sequence (blue) and as reference sequence: green, A; orange, G; red, T; blue, C. Dashes represent the deleted base. Note the drop of coverage (black arrow) over the dashed base, representing the G base deleted in both samples. Right panel: Chromatorgrams show the p.V116X mutation in *C12orf65* and segregation in the family. From the top to the bottom panel: mutant homozygote (Case 3), wild type sequence (V:4), heterozygous p.V116X mutation (V:5), wild type sequence (VI:9) and heterozygous p.V116X mutation (VI:8).
B) Schematic diagram of the *C12orf65* gene and protein. The position of the mutation in the patient DNA and the position of the resulting premature stop codon are indicated in red; blue indicates the mutation and the position of the resulting stop codon described in patients with Leigh syndrome, optic atrophy and ophthalmoplegia.

Figure 23. The expression level of *C12orf65*



C12orf65 mRNA expression levels in the lymphoblasts were measured by quantitative real-time PCR (qRT-PCR). The bar graph represents relative C12orf65 levels normalized to RPL13A. Levels are shown as the percentage of the average of control. There is no difference between expression levels in case and controls.

4.4.3 Mitochondrial impairment

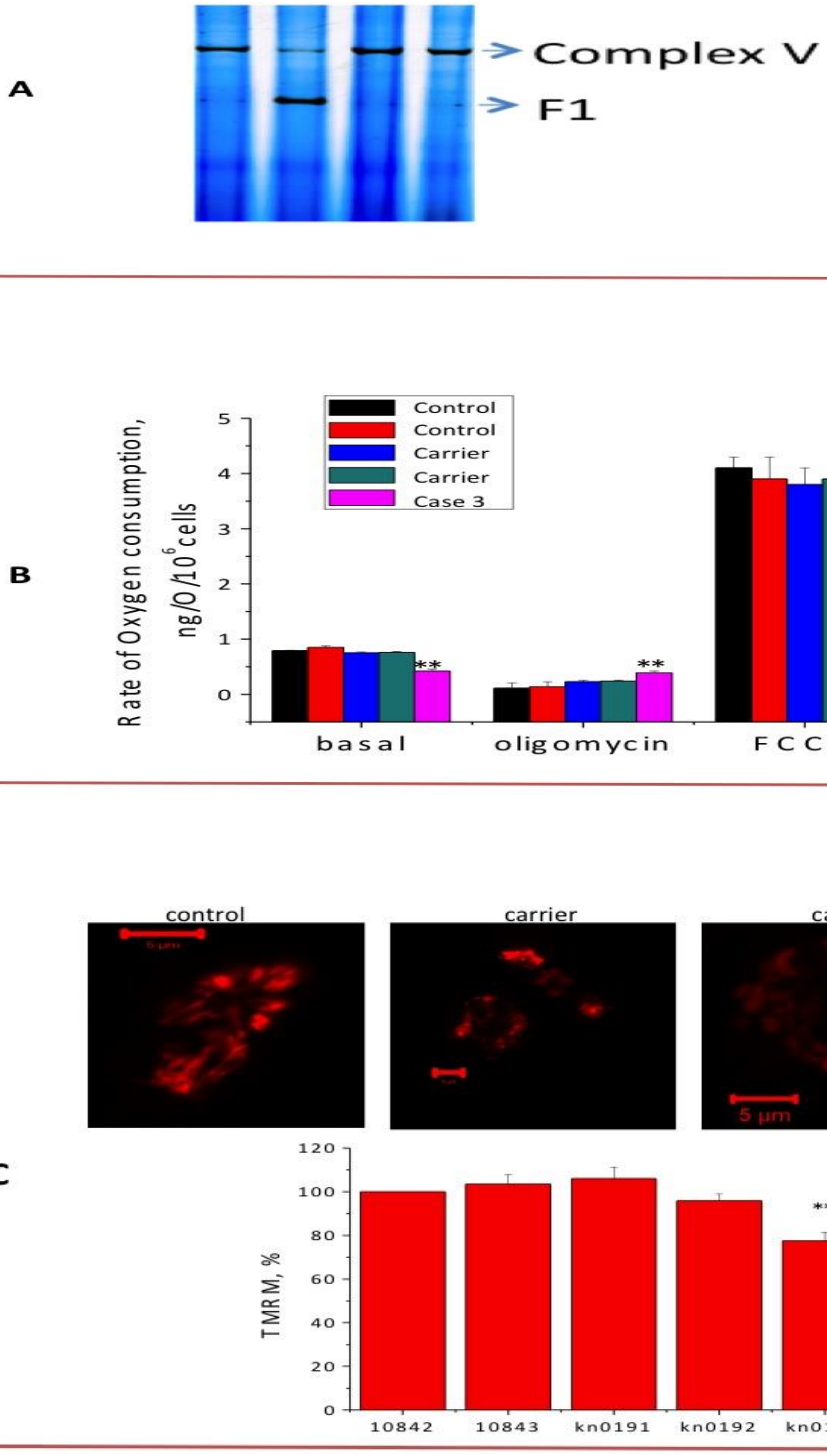
C12orf65 is a nuclear gene that encodes a mitochondrial matrix protein that appears to contribute to mitochondrial translation. Nonsense mutations in this have been found in patients with a mitochondrial disease associated with combined oxidative phosphorylation enzymes (OXPHOS) deficiency (Antonicka et al., 2010). Hence, lymphoblasts obtained from patients were analyzed by BNPAGE analysis to study mitochondrial respiratory chain protein complexes. This analysis showed a decreased activity of complex V as well as a defect in assembly or stability of complex V in the patient sample compared to controls (Figure 24).

4.4.4 Oxygen consumption

In order to investigate the effect of mutations on mitochondrial respiration we measured the rate of oxygen consumption in lymphocytes. The basal oxygen consumption in Case 3 cells was significantly reduced compare to control cells (0.42±0.02 nmol/O2/min/106 cells; n=4 experiments; compared to 0.85±0.03 nmol O2/min/106; n=5 experiments; $p<0.001$; Fig. 6A). Oligomycin (Inhibitor of complex V; 2µg/ml) inhibited the respiration coupled to oxidative phosphorylation in control lymphocytes but to a significantly lesser extent in kn0193 cells ($p<0.001$; Figure 24b). It suggest profound mitochondrial uncoupling and confirm decreased activity of the complex V in Case 3 cells. Addition of 1µM of the uncoupler FCCP accelerated respiration to maximal levels in control lymphocytes but in significant less degree in Case 3 cells, suggesting that activity respiratory chain in these cells is limited.

Mitochondrial membrane potential (Δψm) is a major indicator of mitochondrial function and health. Δψm in Case3 and control lymphoblasts was assessed by the fluorescent indicator tetramethylrhodamine methyl ester, TMRM. Case 3 cells were associated with a significant reduction in the TMRM signal (and hence in Δψm) to 77.5±3.9 of controls (n=29 cells; n=34 contr, $p<0.001$; Figure 24). There are no difference between cell groups in mitochondrial morphology.

Figure 24. Mitochondria impairment in the patient's lymphoblasts



A) BN-PAGE (in gel activity): shows decrease in complex V (F1-Fo) activity and presence of intense F1 band.
B) Oxygen consumption in Case 3 and controls. Oxygen consumption was measured in immortalised lymphocytes using a Clark oxygen electrode. Respiration was inhibited by blocking ATP production using oligomycin (2 μg/ml), and maximised by adding the uncoupler FCCP (1μM). Data is represented as mean ± SEM. In all cases, * indicates $p<0.05$ and ** indicates $p<0.01$ compared to WT values. The basal oxygen consumption in Case 3 cells was significantly reduced compared to control cells (0.42±0.02 nmol/O2/min/106 cells; n=4 experiments; compared to 0.85±0.03 nmol O2/min/106; n=5 experiments; $p<0.001$. Oligomycin (Inhibitor of complex V; 2μg/ml) inhibited the respiration coupled to oxidative phosphorylation in control lymphocytes but to a significantly lesser extent in Case 3 cells ($p<0.001$). Addition of 1μM of the uncoupler FCCP, accelerated respiration to maximal levels in control lymphocytes but to a significant less degree in the patient cells, suggesting that respiratory chain activity in these cells is limited.
C) Mitochondrial membrane potential. Mitochondrial membrane potential in control, carrier and Case3 lymphoblasts determined by TMRM fluorescence. Control was taken as 100%.

4.5 DISCUSSION

The present study describes a recessive mutation in *C12orf65* associated with CMT6. Mutations in *C12orf65* have been previously described in patients with enchephalomyopathy and combined OXPHOS deficiency (Antonicka et al., 2010). We investigated mitochondrial function in the patients' lymphoblasts and showed that it is impaired at multiple levels. Mitochondrial membrane potential and mitochondrial respiration rate were reduced; at the protein level we showed a decrease in complex V activity as well as a defect in the assembly of the complex.

When I was drafting the article, a paper was published describing a mutation in *C12orf65* in a Japanese family with complex spastic paraplegia, optic atrophy and neuropathy (Shimazaki et al., 2012).

These results highlight the importance of mitochondrial function in peripheral neuropathies and in optic atrophy. Indeed mitofusin 2, the major cause of dominant CMT2, is a mitochondrial membrane protein involved in mitochondrial fusion (Santel and Fuller, 2001) and in the regulation of mitochondrial membrane potential and the OXPHOS system (Pich et al., 2005). Similarly the two most common inherited optic neuropathies, Leber hereditary optic neuropathy (LHON, OMIM #535000) and autosomal dominant optic atrophy (DOA), are the result of mitochondrial dysfunction. LHON is caused by primary mitochondrial DNA (mtDNA) mutations affecting the respiratory chain complexes (Mackey et al., 1996), while the majority of DOA families have mutations in the *OPA1* gene, which encodes for an inner mitochondrial membrane protein important for mtDNA maintenance and oxidative phosphorylation (Alexander et al., 2000), (Zanna et al., 2008).

C12orf65 encodes for a 166 amino acid protein which contains one RF-1 domain, involved in protein translation termination and is found in peptide chain release factors. Patients carrying mutations in *C12orf65* that interrupt the RF-1 domain present with a severe infanthood encephalomyopathy (Leigh syndrome, optic atrophy and ophthalmoplegia) (Antonicka et al., 2010). The different location of the truncating mutation in our patients and the consequent sparing of the RF-1 domain

may account for the milder phenotype and a more selective deficiency of ATPase activity (Shimazaki et al., 2012) (Table 16).

Of note similarly, mutations in the mtDNA encoding MTATP6 and leading to impaired ATPase activity (Nijtmans et al., 2001), are known to cause a clinical syndrome that manifests as neuropathy, ataxia, retinitis pigmentosa (NARP) or the more severe maternally inherited Leigh's syndrome (MILS).

This study also highlights the suitability of WES in providing an accurate genetic diagnosis in heterogeneous disorders such as CMT. Given that most known CMT2 genes each account for a small proportion of CMT2 families, WES is likely to lead to the discovery of further genes for CMT

Table 16. Clinical features of patients carrying different *C12orf65* mutations

	(Antonicka et al., 2010)		(Shimazaki et al., 2012)		This study	
Patients	Patient 1	Patient 2	Patient 1	Patient 2	Case 1	Case 3
C12orf65 mutation	c.248delT p.L84X	c.210delA p.L84X	c.394C>t p.R132X	c.394C>t p.R132X	c.346delG p.V116X	c.346delG p.V116
Psychomotor milestones	Severe retardation (1 yo)	Retardation (3 yo)	Normal	Normal	Delayed (3yo)	Delayed (8yo)
Neuropathy	NA	Severe, motor + sensory (8 years)	Motor + sensory (10 yo)	Motor + sensory (10 yo)	Motor ++ Sensory NA (8yo)	Motor ++ (age 9), sensory NA
Optic atrophy	Decreased vision (3 yo)	Decreased vision (20 yo)	Reduced visual acuity (7 yo)	Reduced acuity (7 yo)	Decreased visual acuity (? yo)	Decreased visual acuity (? yo)
Pyramidal involvement	NA	NA	Bilateral leg spasticity (30 yo)	Bilateral leg spasticity	Bilateral upper limb spasticity	Brisk reflexes

5 A GENOME-WIDE ASSOCIATION STUDY FOLLOW UP: THE PARK16 LOCUS

5.1 STATEMENT OF CONTRIBUTION TO THIS RESEARCH

I was involved in the experimental design of this project. I performed the laboratory based sequencing experiments for the *NUCKS1* and *SLC41A1* genes and the statistical analysis.

I drafted parts of the original publication that has been published in European Journal of Human Genetics and reviewed the manuscript.

5.2 BACKGROUND

GWAS carried on PD patients conducted in a Japanese and a Caucasian population (Satake et al., 2009) (Simón-Sánchez et al., 2009), identified a novel risk locus for PD on chromosome 1q32, known as PARK 16 (OMIM #613164). This locus comprises 169.6kb and contains four different genes (*NUCKS1*, *RAB7L1*, *SLC41A1* and *PM20D1*). We investigated this new locus by performing sequencing analysis in a cohort of 180 pathologically proven PD cases and 480 controls. For each polymorphism identified, we performed an association test. For each novel coding variant identified in the PD cohort we further screened 350 PD (including 82 familial PD). We failed to identify any coding polymorphism associated to PD. We identified association between a novel intronic *RAB7L1* variant (c.379-12insT) and disease (P value = 0.0325). We found two novel mutations in two PD patients (one in *RAB7L1*

and one in *SLC41A1* gene, in one patient each). I will discuss these findings in the light of the most recent literature.

5.3 MATERIALS AND METHODS

5.3.1 Samples

A total of 453 PD samples were selected from brain tissues at the Queen Square Brain Bank for Neurological Disorders in the UK. All samples were clinically and pathologically diagnosed according to the PD Brain Bank criteria by an experienced neuropathologist and based on accepted morphological criteria (S. Wharton, 2008). The mean age at onset was 59 years (range 35-86) and the average of death was 78 years (range 51-94). The male-to-female ratio was 3.5: 1. 82 additional familial PD DNA samples were also employed; were considered familial those cases reporting one or more first-degree relatives with PD. Patients gave informed consent for scientific research. The control cohort (n=483) here analyzed was the "1958 British birth cohort" (http://www.b58cgene.sgul.ac.uk/) which is also used in all disease-related studies carried out by the Wellcome Trust Case Control Consortium (WTCCC).

5.3.2 PCR and sequencing analyses

We first performed PCR and sequencing analyses of all open reading frames (ORFs) of *NUCKS1* (NM_022731), *RAB7L1* (NM_003929) and *SLC41A1* (NM_173854.4) genes in 182 PD cases. Then, for each coding variant identified in the PD cohort (n=9) we also analyzed 351 neurologically normal individuals. SNPs showing association with the disease (*RAB7L1* c.379-12insT) and novel coding variants absent in controls (p.K157R in *RAB7L1* and p.A350V in *SLC41A1*) were further analyzed in larger cohort comprising a total of 453 PD cases and 483 controls. Additionally, the novel coding variants were tested in 82 familial PD cases.

All PCR analyses were performed using primers designed by ExonPrimer (http://ihg.gsf.de/ihg/ExonPrimer.html). We amplified all exons and exon-intron boundaries using the following PCR reaction mix: 20ng of genomic SNA, 5nM forward primer, 5nM reverse primer and 12µl of FastStart PCR Master

(www.rocheapplied-science.com). To amplify exon 1 of *NUCKS1* we added 1 µl of 5% dimethyl sulfoxide (DMSO; Sigma). Primer sequences and PCR thermo-cycling conditions are listed in table S4. Following PCR cleanup using MultiScreen PCR Filter Plates (Millipore), sequencing was performed with both forward and revers primers under the following protocol: 5µl of cleaned PCR products, 0.5µl of BigDye terminator v.3.1 (Applied Biosystems), 0.75µl of 5nM primer, 2µl of 5X Sequencing Buffer (Applied Biosystems) and 5,5µl of deionized molecular grade water. The resulting reactions were then purified with MultiScreen PCR Filter Plates (Millipore) and sequences resolved on ABI3730XL genetic analyzer (Applied Biosystems). Electropherograms were visualized in Sequencer software (4.9 Gene Codes Corporation).

5.3.3 Statistical analyses

For each variant identified in the PD cohort we calculated the allele frequencies in cases and controls, tested for departures from HWE and performed χ2-test and test on allelic association. All computations (test of association and permutation analyses) were performed using the Haploview 4.1 software (http://www.broad.mit.edu/haploview/).

To compare PARK16-associated allele frequencies between diverse populations, HapMap data corresponding to the PARK16 locus from Yoruba (YRI), Japan (JPT), Han Chinese (CHB) and Northern and Western European (CEUUtah residents) populations were also analyzed through haploview software (www.hapmap.org).

5.3.4 Bioinformatic Analysis

To determine the level of sequence conservation of newly identified variants we performed multiple sequence analysis and alignments with paralogues and orthologues using the National Center for Biotechnology Information domain associated homoloGene database using the MUSCLE program (Edgar, 2004).

We also used the program Alamut, a mutation interpretation software, to look for amino acid properties and for predictions of the functional and structural effects of novel coding mutations (http://www.interactive-biosoftware.com/alamut).

5.3.5 Variants definition

Any difference from the reference sequences (NM_022731, NM_003929 and NM_173854.4) was called variant. Variants with MAF >1% in controls samples (or dbSNP) were classified as polymorphism. Variants for which the following criteria applied: absent in controls or dbSNP, and stop or frameshift or missense mutations were defined as mutations.

5.4 RESULTS

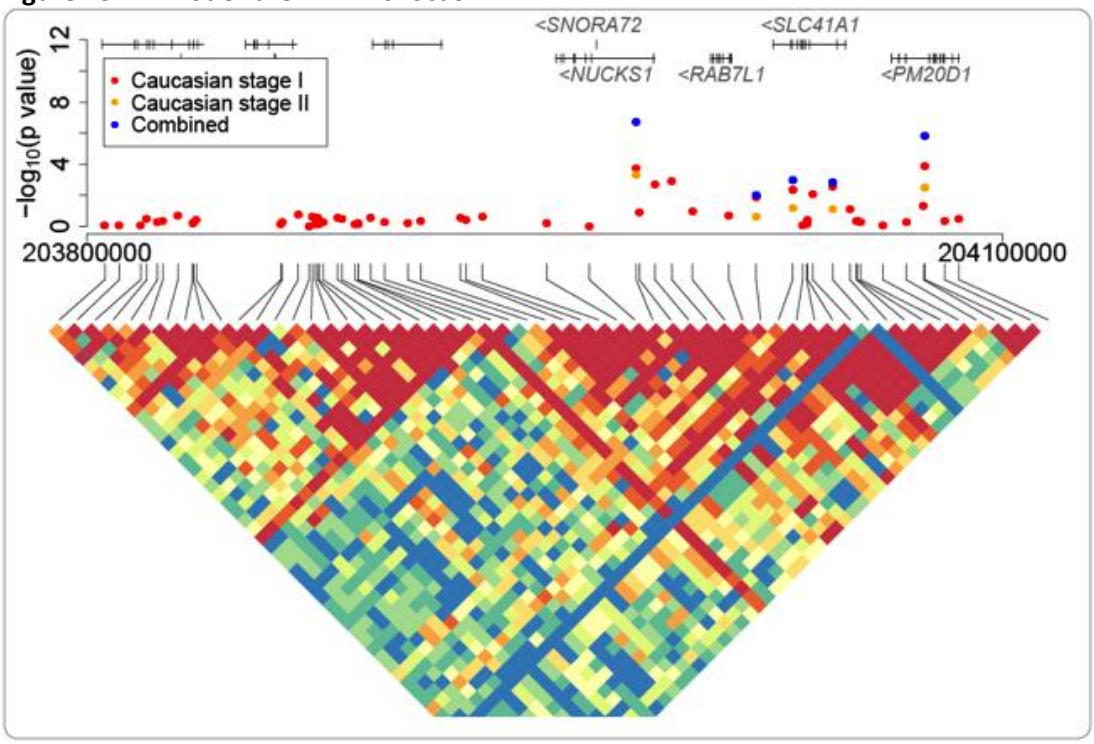
To identify coding variants underlying risk for PD in a British case-control cohort, we investigated the genomic area harboring the PARK16 locus through sequencing analysis. The PARK16 genomic area identified by both the Caucasian and the Japanese study is flanked by the SNPs rs823128 and rs11240572 and contains four genes (*NUCKS1*, *RAB7L1*, *SLC41A1* and *PM20D1*) (Table 17). However, *NUCKS1*, *RAB7L1, SLC41A1* genes were located on the same haplotype block as the most significantly associated SNPs in the Caucasian GWAS (Figure 25). In addition, the minor allele frequency of rs11240572, located in intron 10 of PM20D1, is 0.02 in the European ancestry population (Table 19), which is too low to reach significance in the Caucasian GWAS (Figure 2) and was therefore excluded. Hence only the coding regions of *NUCKS1*, *RAB7L1* and *SLC41A1* were analyzed.

Table 17. PD associated SNPs within the PARK16 locus

SNP	Chr	Position	Chromosomal Localization	Alleles (min/maj)	Caucasian P-values *	Japanese P-values**
rs16856139	1	203905087	SLC45A3 (intronic)	T/C	NA	1.02×10^{-7}
rs823128	1	203980001	*NUCKS1* (intronic)	G/A	7.29×10^{-8}	4.88×10^{-9}
rs823122	1	203991651	Genomic region	C/T	NA	5.22×10^{-9}
rs947211	1	204019288	Genomic region	A/G	NA	1.52×10^{-12}
rs823156	1	204031263	*SLC41A1* (intronic)	G/A	7.60×10^{-4}	3.60×10^{-9}
rs708730	1	204044403	*SLC41A1* (intronic)	G/A	NA	2.43×10^{-8}
rs11240572	1	204074636	PM20D (intronic)	A/C	6.11×10^{-7}	1.08×10^{-7}

Nucleotide positions refer to NCBI build 36.
RAB7L1 chromosomal localization: 204,003,738–204,011,233 bp.
* Combined p values obtained from combining stage I + stage II
** p values obtained from GWAS+replication 1+2

Figure 25. LD Plot of the PARK16 locus



LD structure and association signals across the PARK16 locus in the Caucasian population from the GWAS locus. *NUCKS1*, *RAB7L1* and *SLC41A1* genes are in the same haplotype block. This picture has been published elsewhere (Simón-Sánchez et al., 2009).

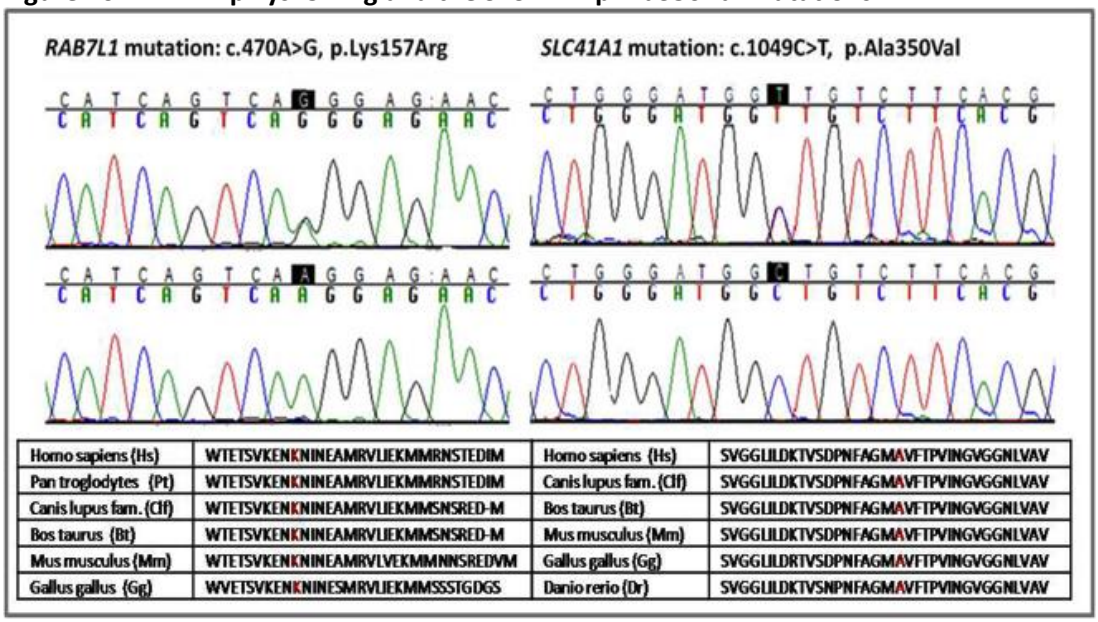
We identified seven variants in the genes *RAB7L1* (n=4) and *SLC41A1* (n=3); of these, none was a coding change. A single-marker χ2-test of association was then performed. This analysis revealed a significant association between a novel, intronic variant in the *RAB7L1* gene (c.379-12insT, intron 3) and PD (p value=0.0325), which remained significant after one million iterations of permutation testing to adjust for multiple comparisons (permuted P-value=0.0399) (Table 18). No other significant association was detected.

In the course of sequencing we also identified two novel coding mutations in *RAB7L1* (p.Lys157Arg) and *SLC41A1* (p.Ala350Val) present in one PD patient each and absent in 480 controls (Figure 26). To inspect whether these novel coding mutations were disease-causing mutations, both were screened in a larger sample size of additional pathologically proven PD cases (n=272, n(total)=454) and 82 familial cases clinically diagnosed with PD. The investigation failed to detect any other mutation carrier.

Table 18. Association tests for variants identified in the PARK16 locus

Gene	Mutation	Associated allele	Chi square	P value	Case/ctrl ratio
RAB7L1	rs708725 (5'-UTR)	A	0.609	0.4351	0.443, 0.418
RAB7L1	c.197-49InsG (intronic)	A	0.949	0.3301	0.072, 0.057
RAB7L1	rs41302139 (p.Gln104Glu)	C	0.075	0.7844	0.019, 0.017
RAB7L1	**c.379-12InsT (inronic)**	**Ins T**	**4.573**	**0.0325 (0.0399)**	**0.013, 0.004**
SLC41A1	rs708727 (p.Asn252Asn)	T	0.776	0.3783	0.438, 0.409
SLC41A1	rs41264905 (intronic)	T	1.402	0.2364	0.009, 0.003
SLC41A1	rs11240569 (p.Thr113Thr)	T	0.971	0.3243	0.301, 0.270

Figure 26. *RAB7L1* p.Lys157Arg and the *SLC41A1* p.Ala350Val mutations



Homo sapiens (Hs)	WTETSVKENKNINEAMRVLIEKMMRNSTEDIM	Homo sapiens (Hs)	SVGGLILDKTVSDPNFAGMAVFTPVINGVGGNLVAV
Pan troglodytes (Pt)	WTETSVKENKNINEAMRVLIEKMMRNSTEDIM	Canis lupus fam. (Clf)	SVGGLILDKTVSDPNFAGMAVFTPVINGVGGNLVAV
Canis lupus fam. (Clf)	WTETSVKENKNINEAMRVLIEKMMSNSRED-M	Bos taurus (Bt)	SVGGLILDKTVSDPNFAGMAVFTPVINGVGGNLVAV
Bos taurus (Bt)	WTETSVKENKNINEAMRVLIEKMMSNSRED-M	Mus musculus (Mm)	SVGGLILDKTVSDPNFAGMAVFTPVINGVGGNLVAV
Mus musculus (Mm)	WTETSVKENKNINEAMRVLVEKMMNNSREDVM	Gallus gallus (Gg)	SVGGLILDRTVSDPNFAGMAVFTPVINGVGGNLVAV
Gallus gallus (Gg)	WVETSVKENKNINESMRVLIEKMMSSSTGDGS	Danio rerio (Dr)	SVGGLILDKTVSNPNFAGMAVFTPVINGVGGNLVAV

Upper panel: Chromatograms of the sequences showing both novel mutations identified in *RAB7L1* and SCL41A1 genes. Lower panel: Tables showing conservation of both lysine and alanine amino-acids among different species. *RAB7L1* (Hs: NP_001129134.1, Pt: XP_001162428.1, Clf: XP_536104.2, Bt: NP_001092564.1, Mm: NP_659124.1, Gg: XP_417967.2); *SLC41A1* (Hs: NP_776253.3, Clf: XP_536105, Bt: XP_613469.2, Mm: NP_776290.1, Gg: XP_417968.2, Dr: XP_001922471.1).

5.5 DISCUSSION

The rationale of GWAS is to test for association of 'tag SNPs' (representing common variation of an haplotype block) in large case-control cohorts to investigate the role of common genetic variation in diseases. This association can be explained by variants (such as a common coding change, or by a SNP influencing gene expression or splicing), lying in the same haplotype block as the tag SNPs showing association in that locus (Altshuler et al., 2008). To identify changes underlying association in a novel locus (PARK16) identified by two GWAS (a Japanese and a Caucasian study), we investigated all coding regions of genes located in the same haplotype block as the most significantly associated SNPs in the GWAS in a Caucasian case-control cohort. We failed to identify any coding variants significantly associated to the disease. We identified two novel coding mutations in one PD patients each, and although they are conserved among species (Figure 26), the lack of occurrence of both mutations in a large sample of ethnicity-matched control individuals (n=483) and the lack of segregation data does not fully disclose their pathogenicity.

We found one intronic variant in *RAB7L1* gene (12 bp upstream of exon 3) associated to PD (p val = 0.03). This variant is not predicted to alter a splice site by current prediction programs. Based on our data this is a rare allele (allele frequency 0.013 in cases, 0.004 in controls), therefore unlikely to underpin by itself a strong association signal at a GWAS level. This variant has been identified by the most recent release of the 1000 Genomes Project, and catalogued in dbSNP 137 as rs34402174. The absence of frequency data for this variants confirms its rarity; only very large association studies will clarify its role in PD.

The results of this study do not explain the PARK16 association in PD in our UK cohort. Of note, the PARK16 locus was originally identified in the Asian study (with p value ranging from 10-7 to 10-12) while it did not reach a significant level in the Caucasian GWAS or its replication, but it did after combining samples from stage I and II (Simón-Sánchez et al., 2009) (Table 17). This difference can be explained by marked differences in the minor allele frequencies of the PARK16 associated SNP

between populations (Table 19). The low allele frequencies of the SNPs in the PARK16 locus in the Caucasian population have limited the statistical power to address whether variants are associated with Parkinson's disease in this locus.

Conversely, subsequent replication studies in the Asian populations have confirmed the association at the PARK16 locus even in smaller case-control cohorts (Tan et al., 2010), (Vilariño-Güell et al., 2010), (Chang et al., 2011).

Table 19. PARK16 core SNPs frequencies in diverse populations

CEU	Name	Position	ObsHET	PredHET	Hwpval	MAF	Alleles (Maj:Min)
	rs16856139	203905087	0.094	0.09	1	0.047	C:T
	rs823128	203980001	0.034	0.034	1	0.017	A:G
	rs823122	203991651	0.077	0.09	0.4364	0.047	T:C
	rs947211	204019288	0.376	0.364	0.9746	0.239	G:A
	rs823156	204031263	0.308	0.295	0.938	0.179	A:G
	rs708730	204044403	0.222	0.248	0.3983	0.145	A:G
	rs11240572	204074636	0.043	0.042	1	0.021	C:A
CHI+JPT							
	rs16856139	203905087	0.224	0.217	1	0.124	C:T
	rs823128	203980001	0.253	0.247	1	0.144	A:G
	rs823122	203991651	0.253	0.247	1	0.144	T:C
	rs947211	204019288	0.494	0.494	1	0.447	A:G
	rs823156	204031263	0.365	0.347	0.7025	0.224	A:G
	rs708730	204044403	0.371	0.35	0.6318	0.226	A:G
	rs11240572	204074636	0.324	0.294	0.3199	0.179	C:A
YRI							
	rs16856139	203905087	0.087	0.159	5.00E-04	0.087	C:T
	rs823128	203980001	0.522	0.472	0.381	0.383	A:G
	rs823122	203991651	0.478	0.499	0.7532	0.483	C:T
	rs947211	204019288	0.487	0.476	1	0.391	A:G
	rs823156	204031263	0.426	0.427	1	0.309	G:A
	rs708730	204044403	0.217	0.258	0.1759	0.152	G:A
	rs11240572	204074636	0	0	1	0	C:C

CEU = Caucasian population (Utah residents with Northern and Western European ancestry from the CEPH collection); CHB+JPT = Han Chinese in Beijing, China + Japanese in Tokyo, Japan; YRI = Yoruba in Ibadan, Nigeria. ObsHET is the marker's observed heterozygosity; PredHET is the marker's predicted heterozygosity (i.e. 2*MAF*(1-MAF)); HWpval is the Hardy-Weinberg equilibrium p value, which is the probability that its deviation from H-W equilibrium could be explained by chance. This data was available on the HapMap project (http://hapmap.ncbi.nlm.nih.gov/). The table was created with Haploview software.

Only recently, large case-controls studies have addressed the role of PARK16 in the Caucasian population. Large studies capture genetic variants that are not very common and whose genetic effects could be to be small (Ioannidis et al., 2006); moreover power is improved by combining genome-wide datasets with meta-analytic techniques. The most recent phase of GWA studies involved the combination of existing and new datasets in more extensive meta-analyses; these have provided more persuasive results that have been substantiated through replication. In 2011, two large association studies have been carried out: one, by the International Parkinson's Disease Genomics Consortium, comprising ~12000 cases and ~21000 controls, and the other by the genetic testing company 23andMe comprising ~3,426 cases and 29,624 controls (International Parkinson's Disease Genomics Consortium (IPDGC); Wellcome Trust Case Control Consortium 2 (WTCCC2), 2011), (Do et al., 2011). These groups replicated each other's signals and provided stronger evidence of association of the PARK16 locus (rs708723, p value = 8.82 x 10-15; rs823156, p value = 1.27 .x 10 -7)

A further meta-analysis of the PD GWAS data from the PDGene database (www.pdgene.org/) was also recently carried out combining different datasets from up to 16,452 PD cases and 48,810 controls. This study also confirmed the association of PARK16 to PD (rs947211, p value = 8 x 10 -10).

These results collectively provide convincing evidence of association at the PARK16 locus. However, the variants underlying associations have not been yet identified. There are two ongoing studies from the IPDGC that address this issue: i) deep sequencing of ~3000 PD cases and ~1000 controls of all the GWAS hits (including PARK16), ii) WES of ~4000 PD cases and 2000 controls. The first study will use a technology available, targeted sequencing, to sequence all coding and noncoding regions of all the loci significantly associated to PD published to date. This will ultimately identify variants associated to PD likely to underpin the association. The second study will lead to the identification of all coding variants potentially associated to PD.

In summary, this study provided limited evidence of association of PARK16 locus to PD in a UK cohort, mainly due to low statistical power to detect a significant result.

The PARK16 variants are rare in the Caucasian population and only large studies have confirmed the association of variants in the PARK16 locus. Deep-sequencing studies in large case control cohorts will ultimately identify the variants underpinning the association.

6 STUDY OF VARIABILITY OF A PARKINSON'S DISEASE GENE: *EIF4G1*

6.1 STATEMENT OF CONTRIBUTION TO THIS RESEARCH

I was involved in the conceptualization and experimental design of this project. I performed the laboratory based sequencing experiments and the statistical analysis. I wrote the original manuscript that has been published in the journal Neuroscience Letters.

6.2 BACKGROUND

Recently variants in the 4 eukaryotic translation initiation factor 4G1 gene (*EIF4G1*) were identified in familial PD, and sporadic PD cases, with a frequency of <0.2% in affected cases (Chartier-Harlin et al., 2011). We screened our familial PD cohort, to determine whether we could provide further evidence that this gene is a PD-related locus. We also investigated genetic variability around these coding positions in a cohort of Africans from the Human Diversity series (Cann et al., 2002). Moreover, to obtain an exhaustive description of the pattern of variability in the *EIF4G1* gene, we extracted genotype data from the 1000 Genomes Project and NHLBI exome sequencing project (URL: http://evs.gs.washington.edu/EVS/) [January 2011]. We failed to identify any mutation in the familial PD cohort. We describe a high number of polymorphisms in the exons where the two PD variants have been previously reported. We also identify a variant previously associated with PD in Caucasians.

6.3 METHODS

6.3.1 Samples

The study included our UK cohort of familial PD samples (n=150). All patients fulfilled criteria for clinical diagnosis of PD at the time of the study with at least 2 of 3 cardinal signs of tremor, rigidity, and bradykinesia, and a positive response to

levodopa therapy. Familial cases were defined as those reporting 1 or more first degree relatives with PD with a pedigree consistent with autosomal dominant pattern of inheritance. Patients carrying known mutations in *SNCA* and *LRRK2* pathogenic mutation were excluded.

The African cohort consisted of 114 samples from different Subsaharian African regions obtained from Centre d'Etude du Polymorphisme Humain, Human Genome Diversity Project (Table 20) (Cann et al., 2002).

The National Heart, Lung, and Blood Institute (NHLBI) exome sequencing project (ESP, (Exome Variant Server, NHLBI Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) [January 2011] includes exome data for 3,500 American individuals of European descent and 1,850 African American.

Table 20. African samples studied per population and geographic regions

Number of samples	Population	Geographic origin
35	Biaka Pygmy	Central African Republic
15	Mbuti_Pygmy	Democratic Republic of Congo
12	Bantu_N.E.	Kenya
7	San	Namibia
26	Yoruba	Nigeria
24	Mandenka	Senegal
8	Bantu_S.E._Pedi	South Africa

6.3.2 PCR and sequencing

Primers for exon 8 and exon 22 in *EIF4G1* were designed in Gene Runner (v.3.0.5, Hastings Software Inc., Hastings, NY, USA) using the GenBank reference sequence NM_182917.3 (primers listed in table S5). All exons and exon-intron boundaries were amplified using the following PCR reaction mix: 20ng of genomic DNA, 10nM forward primer, 10nM reverse primer and 12µl of FastStart PCR Master (Roche, IN,

USA). Following PCR cleanup using MultiScreen PCR Filter Plates (Millipore), bidirectional sequencing of exons and exon-intron boundaries was performed under the following protocol: 5µl of cleaned PCR product, 0.5µl of BigDye (v.3.2), 1µl of 10nM primer, 2µl 5X Sequencing Buffer (Applied Biosystems), and 5µl distilled and deionised molecular grade water. After cleanup of the sequencing products with MultiScreen PCR Filter Plates (Millipore), sequences were analyzed on a 3730xl DNA Analyzer (Applied Biosystems, CA, USA) and electropherograms were visualised in Sequencher software.

6.3.3 Variants frequency and annotation

Vcf files were downloaded from NHLBI and the 1000 Genomes Project websites: http://evs.gs.washington.edu/EVS/, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release

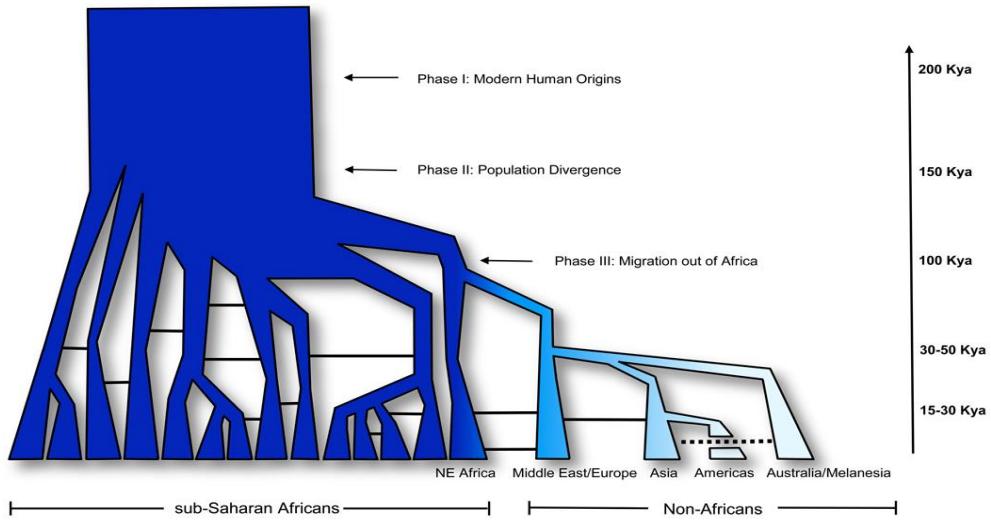
Frequencies in the 1000 Genome Project and the NHLBI were computed using VCFtools (Danecek et al., 2011), with the following command:

```
vcftools --vcf file1.vcf --chr --from-bp 184032283 --
to-bp 184053146 --freq
```

ANNOVAR was then used to annotate the function of the variants:

```
annovar/convert2annovar.pl -format vcf4
output_Var.vcf_filtered -outfile
output_annovar/annovar_${code}
annovar/summarize_annovar.pl -buildver hg19
output_annovar/annovar_${code} annovar/humandb_hg19/
```

Figure 27. The Recent African Origin' model of modern humans



The 'Recent African Origin' model of modern humans and population substructure in Africa (Campbell and Tishkoff, 2008). Humans originated in Africa ~200,000 years ago and then migrated to the rest of the world in the past ~100,000 years. Thus humans have existed continuously in Africa longer than in any other geographic region and have maintained relatively large effective population sizes, resulting in high levels of within-population genetic diversity. For these reasons African populations should be studied when investigating human genetic variation. This picture has been published elsewhere (Campbell, 2010). In this figure, decreasing intensity of color represents the concomitant loss of genetic diversity as populations migrated in an eastward direction from Africa. Solid horizontal lines indicate gene flow between ancestral human populations and the dashed horizontal line indicates recent gene-flow between Asian and Australian/Melanesian populations.

6.4 RESULTS

We failed to identify any previously reported variant in the PD familial samples. We identified one coding change (P486S) in one PD individual. This variant was initially identified by Farrer et al (personal communication) in 5 PD cases and 2 neurologically normal controls (in a cohort of 3225 cases and 2273 controls). However, this variants was subsequently reported in dbSNP 35 (rs112545306), in African-Americans (http://snp.gs.washington.edu/EVS), with a frequency of 0.15%.

We identified six non-synonymous changes in the African cohort from the Human Genome Diversity Project. Of these, one is a novel change (P382L), the others are variants recently reported in dbSNP and found mainly in African populations (http://snp.gs.washington.edu/EVS). To predict the impact on protein function of these non-synonymous variants, we performed an in silico analysis using the software PolyPhen and SIFT (Kumar et al., 2009) and all were predicted to be benign (Table 21).

Analysis of the NHLBI samples allowed us to detect the A502V variant in the European-American individuals, with frequency of 0.02%. We identified a total of 95 nonsynonymous SNP over 32 exons in total. Of note, about a third of all coding variants (36) are found in exon 8 and exon 22 (table S7), making these exons extremely polymorphic.

Table 21. *EIF4G1* coding variants identified in the African cohort

Nucleotide change	Protein change	SNP accession number	Frequency	Population	Effect (SIFT)
c.870G>A	M290I	rs144947145	0.01	San, Bantu SE	tolerated
c.913C>T	R305C	rs116508885	0.01	Youruba	tolerated
c.932A>G	Y311C	rs16858632	0.03	Bantu SW, Bantu NE, Yoruba, Mandenka	tolerated
c.1145C>T	P382L	NA	0.004	San	tolerated
c.1429G>A	E477K	rs145228718	0.004	Mandenka	tolerated
c.3918G>A	R1216H	rs34086109	0.004	Biaka Pygmy	Tolerated

6.5 DISCUSSION

We took three approaches to assess whether mutations in *EIF4G1* cause PD: first, we screened a cohort of PD familial cases; then, we investigated genetic variation in the African samples from Human Genome Diversity Panel (HGDP); and last, we described variability in the whole gene occurring in the 1000 Genomes Project and the NHLBI.

We screened our UK PD familial cohort in which we have previously identified *LRRK2*, *VPS35* and *SNCA* mutations (Sheerin et al., 2011), (Khan et al., 2005). If we had found a mutation in another family we would have provided further evidence that *EIF4G1* is a gene involved in PD. We failed to find any mutation, suggesting that mutations in the EIF4G1 gene are very rare in PD. Further sequencing studies will clarify if the original mutation occurred privately in the original French family, making it difficult to assess its role in PD.

We then assessed the exons containing the mutations originally described in PD in the African samples from HGDP, as they have the greatest diversity (Campbell and

Tishkoff, 2008) and offer a rapid route to the identification of benign polymorphisms (Guerreiro et al., 2010). We found six nonsynsonymous changes (five in exon 8 and one on exon 22) in a small number of samples, suggesting that they are benign.

The in-depth analysis of variability occurring in the *EIF4G1* gene in the 1000 Genomes Project and NHLBI revealed the presence of a variant originally described as pathogenic (A502V) in two samples of European ancestry, making it likely to be a rare benign polymorphism. The high number of coding polymorphisms in the exons where the two PD variants have been previously reported suggests that the protein can tolerate some extent of variability particularly at this point of the protein.

The study here presented was the first negative report of *EIF4G1* mutations in PD. Subsequent studies in larger case-control cohort have confirmed our findings in Caucasians and Asian populations (Lesage et al., 2012), (unpublished results, in press in Neurobiology of Aging journal). Of note, recently the original variant found in the French family (R1205H) was found in Caucasian controls samples and not in PD (Schulte et al., 2012). Overall these results do not support the pathogenicity of *EIF4G1* mutations in PD.

7 CONCLUSIONS AND FUTURE DIRECTIONS

My PhD research focused on the use of WES to dissect the genetic aetiologies of neurodegenerative diseases. WES was first available in 2009 and I was the first in our department to use it. Therefore I had the opportunity to fully appreciate WES strengths, pitfalls and challenges (Chapter 2) and to start applying this technology where it was most promising: rare Mendelian disorders (Chapters 3 and 4). The in-depth understanding of the technology also provided me with the appreciation of its potential utility in sporadic diseases. The work on the PARK16 locus (Chapter 5) and the study of variability of the *EIF4G1* gene (Chapter 6), that had ben carried out before WES was available in our lab, clearly show that WES and related technologies will help us elucidating the genetic architecture of sporadic diseases and will provide us with a better knowledge of coding variations as more samples will be sequenced.

I will briefly discuss the major findings and limitations of the projects presented in this thesis.

WHOLE EXOME SEQUENCING PILOT PROJECT (Chapter 2)

This pilot project was conducted to test and validate the technology by preparing DNA libraries and sequencing from small number of DNA samples. We generated the first sequencing data in the department, and set up an analysis pipeline which would be used for all following WES projects. The importance and usefulness of this project were in understanding both technical and analytical obstacles to overcome.

KOHLSCHÜTTER-TÖNZ SYNDROME: ROGDI MUTATIONS AND GENETIC HETEROGENEITY PROJECT (Chapter 3)

WES had been shown to be very suitable for gene identification in recessive, clinically well-defined conditions, especially when coupled with linkage data. We

had collected 10 families with the clinical presentation of Kohlschütter-Tönz syndrome, a very rare recessive Mendelian disorder. WES was performed on a number of patients that had homogeneous clinical features and shared a disease candidate region based on linkage analysis. However, WES failed to identify a mutation in our patients, as it was intronic. This study shows one of WES limitation, that is the absence of data in intronic regions.

Overall we identify mutations in the *ROGDI* gene by a combination of WES and traditional Sanger sequencing in five patients that had typical Kohlschütter-Tönz syndrome. We failed to identify *ROGDI* mutations in five patients, who had additional atypical features. We anticipate that those patient carry mutations in other genes and that an accurate description of endophenotypes could help defining the genetic defect.

C12orf65 MUTATIONS CAUSE AXONAL NEUROPATHY WITH OPTIC ATROPHY (Chapter 4)

Charcot-Marie Tooth disease (CMT) forms a clinically and genetically heterogeneous group of inherited peripheral nerve disorders.

This study shows for the first time the genetic defect individuals affected by recessive CMT type 6. Homozygosity mapping coupled with WES quickly identified a nonsense mutation in *C12orf65* gene as a strong candidate for this disease. *C12orf65* encodes for a mitochondrial protein and *C12orf65* mutations had been previously described in patients with Leigh-like syndrome and a mitochondrial translation defect. In his study we report mitochondrial dysfunction in the patient's lymphoblasts at different levels (decreased complex V stability, decreased mitochondrial respiration rate and reduction of mitochondrial membrane potential).

In this project WES proved its utility in different ways: it provided a genetic diagnosis in a heterogeneous disorders such as CMT; it expanded the phenotype associated to *C12orf65* mutations. Moreover, the identification of this novel gene brought new biological insight to the pathogenesis of CMT type 6 and the role of mitochondria impairment in this disease.

A GENOME-WIDE ASSOCIATION STUDY FOLLOW-UP: THE PARK16 LOCUS (Chapter 5)

In this project we investigated a novel locus for PD (PARK16), previously identified in two GWAS, in an attempt to pinpoint the variant(s) underlying the GWAS signal. We used traditional Sanger sequencing to sequence the genes in the locus in a cohort of UK PD patients and controls. We identified one rare intronic variant (rs34402174; MAF 0.013 in cases, 0.004 in controls) weakly associated to PD (p value = 0.03). We also identified two mutations in one PD patient each and absent in controls. The rarity of the variants here identified makes them unlikely to underpin the GWAS signal.

A complete catalogue of all variations in the locus is required in the search for causal variants; this can be achieved nowadays by using targeted resequencing, an application of next generation sequencing technology. This technology allows to capture all variation of a genomic region in a cost-effective (especially compared to Sanger sequencing) and efficient manner in multiple DNA samples. Targeted resequencing projects in PD are currently ongoing (in large case controls cohort) and they will likely reveal the full allelic spectrum of causal variants underpinning GWAS signals.

STUDY OF VARIABILITY OF A PARKINSON'S DISEASE GENE: EIF4G1 (Chapter 6)

We performed a study of variants in the *EIF4G1* gene previously reported in PD in our UK PD cases, in a cohort of African samples and by searching all public databases. Here, the rarity of the variants and the absence of segregation data made it very difficult to assess their pathogenicity. The study of variability of *EIF4G1* in a WES database (Exome Variant Server) was pivotal, as it revealed the presence of one of the variant previously associated to PD in the normal Caucasian population. This was the first study published questioning the role of *EIF4G1* variants in PD; subsequent studies in other populations have confirmed the presence of variants previously associated to PD in controls.

Here WES has allowed a full appreciation of benign variability across populations and we envisage that as more samples will be sequence, we will further define normal coding variability in this gene.

Future directions

Most of the successes of WES have been in studies on rare Mendelian disorders, caused by variants of high penetrance segregating in families, such as the *C12orf65* mutation in complex CMT (chapter 4). In the immediate future, the power of WES should enable the identification of genes underlying a large fraction of Mendelian disorders that are currently unsolved.

The main challenge of WES remains to interpret the data in a way that is clinically useful to the individual, as the number of genetic variants predicted to disrupt protein-coding genes in each individual is much more than what was previously suspected (Tennessen et al., 2012), (MacArthur et al., 2012). Determining which of many potentially damaging variants in an individual play a role in more complex diseases is an open issue. There is a significant shortfall in our current ability to systematically interpret genetic variants detected in human exomes outside the context of Mendelian diseases, and this represents the biggest obstacle for the translation of WES to a clinical setting. As more and more complete exomes/ genomes will be sequenced for a wide variety of human populations, our understanding of normal pattern of coding variation will expand. Comprehensive catalogues of the location, frequency and properties of the full spectrum of human variation will provide an important resource for the investigation of clinically - relevant variants.

8 REFERENCES

Abeles M., S., D. E. (1937). Charcot-Marie-Tooth disease with primary optic atrophy: report of two cases occurring in brothers. J Nerv Ment Dis *85*, 541–547.

Ahn, T.-B., Kim, S.Y., Kim, J.Y., Park, S.-S., Lee, D.S., Min, H.J., Kim, Y.K., Kim, S.E., Kim, J.-M., Kim, H.-J., et al. (2008). alpha-Synuclein gene duplication is present in sporadic Parkinson disease. Neurology *70*, 43–49.

Ajroud-Driss, S., Fecto, F., Ajroud, K., Yang, Y., Donkervoort, S., Siddique, N., and Siddique, T. (2009). A novel de novo MFN2 mutation causing CMT2A with upper motor neuron signs. Neurogenetics *10*, 359–361.

Albers, C.A., Lunter, G., MacArthur, D.G., McVean, G., Ouwehand, W.H., and Durbin, R. (2011). Dindel: accurate indel calls from short-read data. Genome Res. *21*, 961–973.

Altmüller, J., Palmer, L.J., Fischer, G., Scherb, H., and Wjst, M. (2001). Genomewide scans of complex human diseases: true linkage is hard to find. Am. J. Hum. Genet. *69*, 936–950.

Antonicka, H., Ostergaard, E., Sasarman, F., Weraarpachai, W., Wibrand, F., Pedersen, A.M.B., Rodenburg, R.J., van der Knaap, M.S., Smeitink, J.A.M., Chrzanowska-Lightowlers, Z.M., et al. (2010). Mutations in C12orf65 in patients with encephalomyopathy and a mitochondrial translation defect. Am. J. Hum. Genet. *87*, 115–122.

Bainbridge, M.N., Wang, M., Burgess, D.L., Kovar, C., Rodesch, M.J., D'Ascenzo, M., Kitzman, J., Wu, Y.-Q., Newsham, I., Richmond, T.A., et al. (2010). Whole exome capture in solution with 3 Gbp of data. Genome Biol. *11*, R62.

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*, 745–755.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53–59.

Bilgüvar, K., Oztürk, A.K., Louvi, A., Kwan, K.Y., Choi, M., Tatli, B., Yalnizoğlu, D., Tüysüz, B., Cağlayan, A.O., Gökben, S., et al. (2010). Whole-exome sequencing

identifies recessive WDR62 mutations in severe brain malformations. Nature *467*, 207–210.

Bonifati, V., Rizzu, P., van Baren, M.J., Schaap, O., Breedveld, G.J., Krieger, E., Dekker, M.C.J., Squitieri, F., Ibanez, P., Joosse, M., et al. (2003). Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism. Science *299*, 256–259.

Botstein, D., and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat. Genet. *33 Suppl*, 228–237.

Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. *32*, 314–331.

Braak, H., and Del Tredici, K. (2008). Invited Article: Nervous system pathology in sporadic Parkinson disease. Neurology *70*, 1916–1925.

Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W.L., Russ, C., Lander, E.S., Nusbaum, C., and Jaffe, D.B. (2008). Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res. *18*, 763–770.

Çalışkan, M., Chong, J.X., Uricchio, L., Anderson, R., Chen, P., Sougnez, C., Garimella, K., Gabriel, S.B., dePristo, M.A., Shakir, K., et al. (2011). Exome sequencing reveals a novel mutation for autosomal recessive non-syndromic mental retardation in the TECR gene on chromosome 19p13. Hum. Mol. Genet. *20*, 1285–1289.

Campbell, M.C. (2010). The Evolution of Human Genetic and Phenotypic Variation in Africa. Curr. Biol. *20*, R166–R173.

Campbell, M.C., and Tishkoff, S.A. (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu. Rev. Genomics Hum. Genet. *9*, 403–433.

Cann, H.M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A human genome diversity cell line panel. Science *296*, 261–262.

Chalmers, R.M., Bird, A.C., and Harding, A.E. (1996). Autosomal dominant optic atrophy with asymptomatic peripheral neuropathy. J. Neurol. Neurosurg. Psychiatry *60*, 195–196.

Chalmers, R.M., Riordan-Eva, P., and Wood, N.W. (1997). Autosomal recessive inheritance of hereditary motor and sensory neuropathy with optic atrophy. J. Neurol. Neurosurg. Psychiatry *62*, 385–387.

Chartier-Harlin, M.-C., Dachsel, J.C., Vilariño-Güell, C., Lincoln, S.J., Leprêtre, F., Hulihan, M.M., Kachergus, J., Milnerwood, A.J., Tapia, L., Song, M.-S., et al. (2011).

Translation Initiator EIF4G1 Mutations in Familial Parkinson Disease. Am. J. Hum. Genet. *89*, 398–406.

Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Ozen, S., Sanjad, S., et al. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc. Natl. Acad. Sci. U. S. A. *106*, 19096–19101.

Christodoulou, J., Hall, R.K., Menahem, S., Hopkins, I.J., and Rogers, J.G. (1988). A syndrome of epilepsy, dementia, and amelogenesis imperfecta: genetic and clinical features. J. Med. Genet. *25*, 827–830.

Chung, K.W., Kim, S.B., Park, K.D., Choi, K.G., Lee, J.H., Eun, H.W., Suh, J.S., Hwang, J.H., Kim, W.K., Seo, B.C., et al. (2006). Early onset severe and late-onset mild Charcot-Marie-Tooth disease with mitofusin 2 (MFN2) mutations. Brain J. Neurol. *129*, 2103–2118.

Clark, M.J., Chen, R., Lam, H.Y.K., Karczewski, K.J., Chen, R., Euskirchen, G., Butte, A.J., and Snyder, M. (2011). Performance comparison of exome DNA sequencing technologies. Nat. Biotechnol. *29*, 908–914.

Colcher, A., and Simuni, T. (1999). Clinical manifestations of Parkinson's disease. Med. Clin. North Am. *83*, 327–347.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. Nat. Rev. Genet. *10*, 184–194.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinforma. Oxf. Engl. *27*, 2156–2158.

DeJesus-Hernandez, M., Mackenzie, I.R., Boeve, B.F., Boxer, A.L., Baker, M., Rutherford, N.J., Nicholson, A.M., Finch, N.A., Flynn, H., Adamson, J., et al. (2011). Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron *72*, 245–256.

Dick, F.D., De Palma, G., Ahmadi, A., Scott, N.W., Prescott, G.J., Bennett, J., Semple, S., Dick, S., Counsell, C., Mozzoni, P., et al. (2007). Environmental risk factors for Parkinson's disease and parkinsonism: the Geoparkinson study. Occup. Environ. Med. *64*, 666–672.

Do, C.B., Tung, J.Y., Dorfman, E., Kiefer, A.K., Drabant, E.M., Francke, U., Mountain, J.L., Goldman, S.M., Tanner, C.M., Langston, J.W., et al. (2011). Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. PLoS Genet. *7*, e1002141.

Donnai, D., Tomlin, P.I., and Winter, R.M. (2005). Kohlschutter syndrome in siblings. Clin. Dysmorphol. *14*, 123–126.

Durbin, R.M., Altshuler, D.L., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Collins, F.S., De La Vega, F.M., Donnelly, P., et al. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

Duvoisin, R.C., Eldridge, R., Williams, A., Nutt, J., and Calne, D. (1981). Twin study of Parkinson disease. Neurology *31*, 77–80.

Dyck PJ (1975). Inherited neuronal degeneration and atrophy affecting peripheral motor, sensory and autosomic neurons. In Peripheral Neuropathy, (Philadelphia), pp. 825–867.

Edwards, T.L., Scott, W.K., Almonte, C., Burt, A., Powell, E.H., Beecham, G.W., Wang, L., Züchner, S., Konidari, I., Wang, G., et al. (2010). Genome-wide association study confirms SNPs in SNCA and the MAPT region as common risk factors for Parkinson disease. Ann. Hum. Genet. *74*, 97–109.

Di Fonzo, A., Dekker, M.C.J., Montagna, P., Baruzzi, A., Yonova, E.H., Correia Guedes, L., Szczerbinska, A., Zhao, T., Dubbel-Hulsman, L.O.M., Wouters, C.H., et al. (2009). FBXO7 mutations cause autosomal recessive, early-onset parkinsonian-pyramidal syndrome. Neurology *72*, 240–245.

Funayama, M., Li, Y., Tomiyama, H., Yoshino, H., Imamichi, Y., Yamamoto, M., Murata, M., Toda, T., Mizuno, Y., and Hattori, N. (2007). Leucine-rich repeat kinase 2 G2385R variant is a risk factor for Parkinson disease in Asian population. Neuroreport *18*, 273–275.

Guazzi, G., Palmeri, S., Malandrini, A., Ciacci, G., Perri, R.D., Mancini, G., Messina, C., and Salvadori, C. (1994). Ataxia, mental deterioration, epilepsy in a family with dominant enamel hypoplasia: A variant of Kohlschütter-Tönz syndrome? Am. J. Med. Genet. *50*, 79–83.

Guerreiro, R., Wojtas, A., Bras, J., Carrasquillo, M., Rogaeva, E., Majounie, E., Cruchaga, C., Sassi, C., Kauwe, J.S.K., Younkin, S., et al. (2013). TREM2 variants in Alzheimer's disease. N. Engl. J. Med. *368*, 117–127.

Guerreiro, R.J., Washecka, N., Hardy, J., and Singleton, A. (2010). A thorough assessment of benign genetic variability in GRN and MAPT. Hum. Mutat. *31*, E1126–1140.

Haberlandt, E., Svejda, C., Felber, S., Baumgartner, S., Günther, B., Utermann, G., and Kotzot, D. (2006). Yellow teeth, seizures, and mental retardation: a less severe case of Kohlschütter-Tönz syndrome. Am. J. Med. Genet. A. *140*, 281–283.

Hamza, T.H., Zabetian, C.P., Tenesa, A., Laederach, A., Montimurro, J., Yearout, D., Kay, D.M., Doheny, K.F., Paschall, J., Pugh, E., et al. (2010). Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. Nat. Genet. *42*, 781–785.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.-K., Chrast, J., Lagarde, J., Gilbert, J.G.R., Storey, R., Swarbreck, D., et al. (2006). GENCODE: producing a reference annotation for ENCODE. Genome Biol. *7 Suppl 1*, S4.1–9.

Herrera RF, B.A., Junior EC, Junior WM (1990). Hereditary sensorimotor neuropathy associated with optic atrophy (type VI HSMN).

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U. S. A. *106*, 9362–9367.

Hsu, F., Kent, W.J., Clawson, H., Kuhn, R.M., Diekhans, M., and Haussler, D. (2006). The UCSC Known Genes. Bioinforma. Oxf. Engl. *22*, 1036–1046.

Ippel, E.F., Wittebol-Post, D., Jennekens, F.G., and Bijlsma, J.B. (1995). Genetic heterogeneity of hereditary motor and sensory neuropathy type VI. J. Child Neurol. *10*, 459–463.

Johnson, J.O., Mandrioli, J., Benatar, M., Abramzon, Y., Van Deerlin, V.M., Trojanowski, J.Q., Gibbs, J.R., Brunetti, M., Gronka, S., Wuu, J., et al. (2010). Exome sequencing reveals VCP mutations as a cause of familial ALS. Neuron *68*, 857–864.

Kapushesky, M., Adamusiak, T., Burdett, T., Culhane, A., Farne, A., Filippov, A., Holloway, E., Klebanov, A., Kryvych, N., Kurbatova, N., et al. (2012). Gene Expression Atlas update--a value-added database of microarray and sequencing-based functional genomics experiments. Nucleic Acids Res. *40*, D1077–1081.

Khan, N.L., Jain, S., Lynch, J.M., Pavese, N., Abou-Sleiman, P., Holton, J.L., Healy, D.G., Gilks, W.P., Sweeney, M.G., Ganguly, M., et al. (2005). Mutations in the gene LRRK2 encoding dardarin (PARK8) cause familial Parkinson's disease: clinical, pathological, olfactory and functional imaging and genetic data. Brain J. Neurol. *128*, 2786–2796.

Kirby, A., Gnirke, A., Jaffe, D.B., Barešová, V., Pochet, N., Blumenstiel, B., Ye, C., Aird, D., Stevens, C., Robinson, J.T., et al. (2013). Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. Nat. Genet. *45*, 299–303.

Kitada, T., Asakawa, S., Hattori, N., Matsumine, H., Yamamura, Y., Minoshima, S., Yokochi, M., Mizuno, Y., and Shimizu, N. (1998). Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. Nature *392*, 605–608.

Klassen, T., Davis, C., Goldman, A., Burgess, D., Chen, T., Wheeler, D., McPherson, J., Bourquin, T., Lewis, L., Villasana, D., et al. (2011). Exome sequencing of ion channel genes reveals complex profiles confounding personal risk assessment in epilepsy. Cell *145*, 1036–1048.

Kohlschütter, A., Chappuis, D., Meier, C., Tönz, O., Vassella, F., and Herschkowitz, N. (1974). Familial epilepsy and yellow teeth--a disease of the CNS associated with enamel hypoplasia. Helv. Paediatr. Acta *29*, 283–294.

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

Lander, E.S., and Botstein, D. (1987). Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science *236*, 1567–1570.

Latourelle, J.C., Pankratz, N., Dumitriu, A., Wilk, J.B., Goldwurm, S., Pezzoli, G., Mariani, C.B., DeStefano, A.L., Halter, C., Gusella, J.F., et al. (2009). Genomewide association study for onset age in Parkinson disease. BMC Med. Genet. *10*, 98.

De Lau, L.M.L., and Breteler, M.M.B. (2006). Epidemiology of Parkinson's disease. Lancet Neurol. *5*, 525–535.

Lawson, V.H., Graham, B.V., and Flanigan, K.M. (2005). Clinical and electrophysiologic features of CMT2A with mutations in the mitofusin 2 gene. Neurology *65*, 197–204.

Lees, A.J., Hardy, J., and Revesz, T. (2009). Parkinson's disease. Lancet *373*, 2055–2066.

Lesage, S., Condroyer, C., Klebe, S., Lohmann, E., Durif, F., Damier, P., Tison, F., Anheim, M., Honoré, A., Viallet, F., et al. (2012). EIF4G1 in familial Parkinson's disease: pathogenic mutations or rare benign variants? Neurobiol. Aging *33*, 2233.e1–5.

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. Bioinforma. Oxf. Engl. *25*, 2078–2079.

Lill, C.M., Roehr, J.T., McQueen, M.B., Kavvoura, F.K., Bagade, S., Schjeide, B.-M.M., Schjeide, L.M., Meissner, E., Zauft, U., Allen, N.C., et al. (2012). Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: The PDGene database. PLoS Genet. *8*, e1002548.

Litt, M., and Luty, J.A. (1989). A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am. J. Hum. Genet. *44*, 397–401.

Lo, C.-N. (2009). Genetics in Epilepsy [PhD thesis].

MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science *335*, 823–828.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Maraganore, D.M., de Andrade, M., Lesnick, T.G., Strain, K.J., Farrer, M.J., Rocca, W.A., Pant, P.V.K., Frazer, K.A., Cox, D.R., and Ballinger, D.G. (2005). High-resolution whole-genome association study of Parkinson disease. Am. J. Hum. Genet. *77*, 685–693.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376–380.

Mata, I.F., Kachergus, J.M., Taylor, J.P., Lincoln, S., Aasly, J., Lynch, T., Hulihan, M.M., Cobb, S.A., Wu, R.-M., Lu, C.-S., et al. (2005). Lrrk2 pathogenic substitutions in Parkinson's disease. Neurogenetics *6*, 171–177.

McVean, G., Spencer, C.C.A., and Chaix, R. (2005). Perspectives on human genetic variation from the HapMap Project. PLoS Genet. *1*, e54.

Metzker, M.L. (2010). Sequencing technologies - the next generation. Nat. Rev. Genet. *11*, 31–46.

Milhorat A.T. (1943). PROGRESSIVE MUSCULAR ATROPHY OF PERONEAL TYPE ASSOCIATED WITH ATROPHY OF THE OPTIC NERVES; REPORT ON A FAMILY. Arch Neurol Psychiatry *50*, 279–287.

Montenegro, G., Powell, E., Huang, J., Speziani, F., Edwards, Y.J.K., Beecham, G., Hulme, W., Siskind, C., Vance, J., Shy, M., et al. (2011). Exome sequencing allows for rapid gene identification in a Charcot-Marie-Tooth family. Ann. Neurol. *69*, 464–470.

Mory, A., Dagan, E., Illi, B., Duquesnoy, P., Mordechai, S., Shahor, I., Romani, S., Hawash-Moustafa, N., Mandel, H., Valente, E.M., et al. (2012). A Nonsense Mutation in the Human Homolog of Drosophila rogdi Causes Kohlschutter-Tonz Syndrome. Am. J. Hum. Genet. *90*, 708–714.

Musumeci, S.A., Elia, M., Ferri, R., Romano, C., Scuderi, C., and Del Gracco, S. (1995). A further family with epilepsy, dementia and yellow teeth: the Kohlschütter syndrome. Brain Dev. *17*, 133–138; discussion 142–143.

Nalls, M.A., Plagnol, V., Hernandez, D.G., Sharma, M., Sheerin, U.-M., Saad, M., Simón-Sánchez, J., Schulte, C., Lesage, S., Sveinbjörnsdóttir, S., et al. (2011). Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet *377*, 641–649.

Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature *485*, 242–245.

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., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature *461*, 272–276.

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., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet. *42*, 30–35.

Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. *12*, 443–451.

O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246–250.

Paisan-Ruiz, C., Bhatia, K.P., Li, A., Hernandez, D., Davis, M., Wood, N.W., Hardy, J., Houlden, H., Singleton, A., and Schneider, S.A. (2009). Characterization of PLA2G6 as a locus for dystonia-parkinsonism. Ann. Neurol. *65*, 19–23.

Paisán-Ruíz, C., Jain, S., Evans, E.W., Gilks, W.P., Simón, J., van der Brug, M., López de Munain, A., Aparicio, S., Gil, A.M., Khan, N., et al. (2004). Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease. Neuron *44*, 595–600.

Pankratz, N., Wilk, J.B., Latourelle, J.C., DeStefano, A.L., Halter, C., Pugh, E.W., Doheny, K.F., Gusella, J.F., Nichols, W.C., Foroud, T., et al. (2009). Genomewide association study for susceptibility genes contributing to familial Parkinson disease. Hum. Genet. *124*, 593–605.

Petermöller, M., Kunze, J., and Groß-Selbeck, G. (2008). Kohlschütter Syndrome: Syndrome of Epilepsy - Dementia - Amelogenesis Imperfecta. Neuropediatrics *24*, 337–338.

Piccini, P., Burn, D.J., Ceravolo, R., Maraganore, D., and Brooks, D.J. (1999). The role of inheritance in sporadic Parkinson's disease: evidence from a longitudinal study of dopaminergic function in twins. Ann. Neurol. *45*, 577–582.

Pittman, A.M., Fung, H.-C., and de Silva, R. (2006). Untangling the tau gene association with neurodegenerative disorders. Hum. Mol. Genet. *15 Spec No 2*, R188–195.

Polymeropoulos, M.H., Lavedan, C., Leroy, E., Ide, S.E., Dehejia, A., Dutra, A., Pike, B., Root, H., Rubenstein, J., Boyer, R., et al. (1997). Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. Science *276*, 2045–2047.

Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. *69*, 124–137.

Pruitt, K.D., Tatusova, T., Brown, G.R., and Maglott, D.R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. *40*, D130–135.

Raffan, E., Hurst, L.A., Turki, S.A., Carpenter, G., Scott, C., Daly, A., Coffey, A., Bhaskar, S., Howard, E., Khan, N., et al. (2011). Early Diagnosis of Werner's Syndrome Using Exome-Wide Sequencing in a Single, Atypical Patient. Front. Endocrinol. *2*, 8.

Ramirez, A., Heimbach, A., Gründemann, J., Stiller, B., Hampshire, D., Cid, L.P., Goebel, I., Mubaidin, A.F., Wriekat, A.-L., Roeper, J., et al. (2006). Hereditary parkinsonism with dementia is caused by mutations in ATP13A2, encoding a lysosomal type 5 P-type ATPase. Nat. Genet. *38*, 1184–1191.

Reilly, M.M., Murphy, S.M., and Laurá, M. (2011). Charcot-Marie-Tooth disease. J. Peripher. Nerv. Syst. JPNS *16*, 1–14.

Renton, A.E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J.R., Schymick, J.C., Laaksovirta, H., van Swieten, J.C., Myllykangas, L., et al. (2011). A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron *72*, 257–268.

Risch, N.J. (2000). Searching for genetic determinants in the new millennium. Nature *405*, 847–856.

Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516–1517.

Rost, B., Yachdav, G., and Liu, J. (2004). The PredictProtein server. Nucleic Acids Res. *32*, W321–326.

Saad, M., Lesage, S., Saint-Pierre, A., Corvol, J.-C., Zelenika, D., Lambert, J.-C., Vidailhet, M., Mellick, G.D., Lohmann, E., Durif, F., et al. (2011). Genome-wide association study confirms BST1 and suggests a locus on 12q24 as the risk loci for Parkinson's disease in the European population. Hum. Mol. Genet. *20*, 615–627.

Sailer, A., Scholz, S.W., Gibbs, J.R., Tucci, A., Johnson, J.O., Wood, N.W., Plagnol, V., Hummerich, H., Ding, J., Hernandez, D., et al. (2012). Exome sequencing in an SCA14 family demonstrates its utility in diagnosing heterogeneous diseases. Neurology *79*, 127–131.

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237–241.

Satake, W., Nakabayashi, Y., Mizuta, I., Hirota, Y., Ito, C., Kubo, M., Kawaguchi, T., Tsunoda, T., Watanabe, M., Takeda, A., et al. (2009). Genome-wide association

study identifies common variants at four loci as genetic risk factors for Parkinson's disease. Nat. Genet. *41*, 1303–1307.

Schossig, A., Wolf, N.I., Fischer, C., Fischer, M., Stocker, G., Pabinger, S., Dander, A., Steiner, B., Tönz, O., Kotzot, D., et al. (2012a). Mutations in ROGDI Cause Kohlschütter-Tönz Syndrome. Am. J. Hum. Genet.

Schossig, A., Wolf, N.I., Kapferer, I., Kohlschütter, A., and Zschocke, J. (2012b). Epileptic encephalopathy and amelogenesis imperfecta: Kohlschütter-Tönz syndrome. Eur. J. Med. Genet. *55*, 319–322.

Schulte, E.C., Mollenhauer, B., Zimprich, A., Bereznai, B., Lichtner, P., Haubenberger, D., Pirker, W., Brücke, T., Molnar, M.J., Peters, A., et al. (2012). Variants in eukaryotic translation initiation factor 4G1 in sporadic Parkinson's disease. Neurogenetics.

Sheerin, U.-M., Charlesworth, G., Bras, J., Guerreiro, R., Bhatia, K., Foltynie, T., Limousin, P., Silveira-Moriyama, L., Lees, A., and Wood, N. (2011). Screening for VPS35 mutations in Parkinson's disease. Neurobiol. Aging.

Shimazaki, H., Takiyama, Y., Ishiura, H., Sakai, C., Matsushima, Y., Hatakeyama, H., Honda, J., Sakoe, K., Naoi, T., Namekawa, M., et al. (2012). A homozygous mutation of C12orf65 causes spastic paraplegia with optic atrophy and neuropathy (SPG55). J. Med. Genet. *49*, 777–784.

Shy, M.E., Lupski, J.R., Chance, P.F., Klein, C.J., and Dyck, P.J. (2000). Hereditary Motor and Sensory Neuropathies: An Overview of Clinical, Genetic, Electrophysiologic, and Pathologic Features. In Peripheral Neuropathy, (Elsevier), pp. 1623–1658.

Sidransky, E., Nalls, M.A., Aasly, J.O., Aharon-Peretz, J., Annesi, G., Barbosa, E.R., Bar-Shira, A., Berg, D., Bras, J., Brice, A., et al. (2009). Multicenter Analysis of Glucocerebrosidase Mutations in Parkinson's Disease. N. Engl. J. Med. *361*, 1651–1661.

Simón-Sánchez, J., Schulte, C., Bras, J.M., Sharma, M., Gibbs, J.R., Berg, D., Paisan-Ruiz, C., Lichtner, P., Scholz, S.W., Hernandez, D.G., et al. (2009). Genome-wide association study reveals genetic risk underlying Parkinson's disease. Nat. Genet. *41*, 1308–1312.

Simón-Sánchez, J., van Hilten, J.J., van de Warrenburg, B., Post, B., Berendse, H.W., Arepalli, S., Hernandez, D.G., de Bie, R.M.A., Velseboer, D., Scheffer, H., et al. (2011a). Genome-wide association study confirms extant PD risk loci among the Dutch. Eur. J. Hum. Genet. EJHG *19*, 655–661.

Simón-Sánchez, J., van Hilten, J.J., van de Warrenburg, B., Post, B., Berendse, H.W., Arepalli, S., Hernandez, D.G., de Bie, R.M.A., Velseboer, D., Scheffer, H., et al. (2011b). Genome-wide association study confirms extant PD risk loci among the Dutch. Eur. J. Hum. Genet. EJHG *19*, 655–661.

Small, S.A., and Duff, K. (2008). Linking Abeta and tau in late-onset Alzheimer's disease: a dual pathway hypothesis. Neuron *60*, 534–542.

Spencer, C.C.A., Plagnol, V., Strange, A., Gardner, M., Paisan-Ruiz, C., Band, G., Barker, R.A., Bellenguez, C., Bhatia, K., Blackburn, H., et al. (2011). Dissection of the genetics of Parkinson's disease identifies an additional association 5' of SNCA and multiple associated haplotypes at 17q21. Hum. Mol. Genet. *20*, 345–353.

Stallings, S.C., and Moore, P.B. (1997). The structure of an essential splicing element: stem loop IIa from yeast U2 snRNA. Struct. Lond. Engl. 1993 *5*, 1173–1185.

Stenson, P.D., Mort, M., Ball, E.V., Howells, K., Phillips, A.D., Thomas, N.S., and Cooper, D.N. (2009). The Human Gene Mutation Database: 2008 update. Genome Med. *1*, 13.

Tan, E.-K. (2006). Identification of a common genetic risk variant (LRRK2 Gly2385Arg) in Parkinson's disease. Ann. Acad. Med. Singapore *35*, 840–842.

Tan, E.-K., Peng, R., Teo, Y.-Y., Tan, L.C., Angeles, D., Ho, P., Chen, M.-L., Lin, C.-H., Mao, X.-Y., Chang, X.-L., et al. (2010). Multiple LRRK2 variants modulate risk of Parkinson disease: a Chinese multicenter study. Hum. Mutat. *31*, 561–568.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science *337*, 64–69.

Valente, E.M., Abou-Sleiman, P.M., Caputo, V., Muqit, M.M.K., Harvey, K., Gispert, S., Ali, Z., Del Turco, D., Bentivoglio, A.R., Healy, D.G., et al. (2004). Hereditary early-onset Parkinson's disease caused by mutations in PINK1. Science *304*, 1158–1160.

Via, M., Gignoux, C., and Burchard, E.G. (2010). The 1000 Genomes Project: new opportunities for research and social challenges. Genome Med. *2*, 3.

Vilariño-Güell, C., Wider, C., Ross, O.A., Dachsel, J.C., Kachergus, J.M., Lincoln, S.J., Soto-Ortolaza, A.I., Cobb, S.A., Wilhoite, G.J., Bacon, J.A., et al. (2011). VPS35 mutations in Parkinson disease. Am. J. Hum. Genet. *89*, 162–167.

Vizioli F. (1889). Dell' atrofia muscolare progressiva nevrotica.

Voo, I., Allf, B.E., Udar, N., Silva-Garcia, R., Vance, J., and Small, K.W. (2003). Hereditary motor and sensory neuropathy type VI with optic atrophy. Am. J. Ophthalmol. *136*, 670–677.

Walsh, T., Shahin, H., Elkan-Miller, T., Lee, M.K., Thornton, A.M., Roeb, W., Abu Rayyan, A., Loulus, S., Avraham, K.B., King, M.-C., et al. (2010). Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. Am. J. Hum. Genet. *87*, 90–94.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164–e164.

Ward, C.D., Duvoisin, R.C., Ince, S.E., Nutt, J.D., Eldridge, R., and Calne, D.B. (1983). Parkinson's disease in 65 pairs of twins and in a set of quadruplets. Neurology *33*, 815–824.

Weber, J.L., and May, P.E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am. J. Hum. Genet. *44*, 388–396.

Weeks, D.E., and Lathrop, G.M. (1995). Polygenic disease: methods for mapping complex disease traits. Trends Genet. TIG *11*, 513–519.

Wygold, T., Kurlemann, G., and Schuierer, G. (1996). [Kohlschütter syndrome--an example of a rare progressive neuroectodermal disease. Case report and review of the literature]. Klin. Pädiatrie *208*, 271–275.

Xiromerisiou, G., Houlden, H., Sailer, A., Silveira-Moriyama, L., Hardy, J., and Lees, A.J. (2012). Identical twins with Leucine rich repeat kinase type 2 mutations discordant for Parkinson's disease. Mov. Disord. Off. J. Mov. Disord. Soc. *27*, 1323.

Xu, B., Roos, J.L., Dexheimer, P., Boone, B., Plummer, B., Levy, S., Gogos, J.A., and Karayiorgou, M. (2011). Exome sequencing supports a de novo mutational paradigm for schizophrenia. Nat. Genet. *43*, 864–868.

Zimprich, A., Biskup, S., Leitner, P., Lichtner, P., Farrer, M., Lincoln, S., Kachergus, J., Hulihan, M., Uitti, R.J., Calne, D.B., et al. (2004). Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. Neuron *44*, 601–607.

Zimprich, A., Benet-Pagès, A., Struhal, W., Graf, E., Eck, S.H., Offman, M.N., Haubenberger, D., Spielberger, S., Schulte, E.C., Lichtner, P., et al. (2011). A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset Parkinson disease. Am. J. Hum. Genet. *89*, 168–175.

Zlotogora, J., Fuks, A., Borochowitz, Z., and Tal, Y. (1993). Kohlschütter-Tönz syndrome: epilepsy, dementia, and amelogenesis imperfecta. Am. J. Med. Genet. *46*, 453–454.

Züchner, S., De Jonghe, P., Jordanova, A., Claeys, K.G., Guergueltcheva, V., Cherninkova, S., Hamilton, S.R., Van Stavern, G., Krajewski, K.M., Stajich, J., et al. (2006). Axonal neuropathy with optic atrophy is caused by mutations in mitofusin 2. Ann. Neurol. *59*, 276–281.

Table S1. Primers sequences used for NimbleGen Sequence Capture control locus assays.

qPCR assay	*Primer sequences (5' -3')*	*Tm (°C)*	*Product length*	*qPCR efficiency*
NSC-0237	F: CGCATTCCTCATCCCAGTATG R: AAAGGACTTGGTGCAGAGTTCAG	81.15	80bp	1.84
NSC-0247	F: CCCACCGCCTTCGACAT R: CCTGCTTACTGTGGGCTCTTG	81.03	74bp	1.80
NSC-0268	F: CTCGCTTAACCAGACTCATCTACTGT R: ACTTGGCTCAGCTGTATGAAGGT	78.99	75bp	1.78
NSC-0272	F: CAGCCCCAGCTCAGGTACAG R: ATGATGCGAGTGCTGATGATG	82.23	71bp	1.93

Table S2. PCR and sequencing primers for *ROGDI*

Exon(s)	**Product length (bp)**	**Forward primer**	**Reverse primer**
Exon 1-2	326	GAGGACCGAGGACAGAAAGA	GAGCCAGGGAAATGAGGAG
Exon 3-4	589	CAGCTGAACCCAGGTGAG	TTCCAAAGGAGAAAACTGAG
Exon 5	352	GAGACCCTCTGCCCTATAC	GCTGTTAGGAAACACCCTC
Exon 6	305	GAGAGTGGAATGACGTGG	GTGCTTACAGGTTGCACAG
Exon 7	308	ACAGTAGCCACTGGGCTG	CCACCTATAATCTGGGAGTG
Exon 8-9	467	CATTGATGAGAAGAGCATC	AGTGAGAGGGTC CCTGAG
Exon 8-10	735	GATGAGAAGAGCATCTCTGG	GACAATATGTCAGGACAGGG
Exon 10-11	586	CATGTTGTAAGTAGGCAGG	GAGATAAATAGCAGCCTGG
Exon 11	241	GCCCTGTCCTGACATATTG	GGAGATAAATAGCAGCCTGG
FAM labeled primers:			
Exon 1-2	483	CAGTCTGGTTCCAGGGCTC	5'-FAM-GAGCCAGGGAAATGAGGAG
cDNA analysis (RT PCR)			
Exon 1-11		CCATGGCCACCGTGATG	TTCCTGGAGACAAGCTCCTG
Exon 1-6 (wild-type specific)		CCATGGCCACCGTGATG	GGTAAGCAGGTAAATGGCTTC

Table S3. PCR, qPCR and sequencing primers for *C12orf65*

Exon(s)	Product length (bp)	Forward primer	Reverse primer
C12orf65 DNA			
Exon 2	584	GGTAACATGGCAGACAGTG	CACACTTGGGAGAAATCTC
Exon 3	490	ATCACTTGGACCCAGGAG	ACTCGCCACCATTATTCC
C12orf65 cDNA[i]			
Exon 2-3	336	CTACCCTGCACTGCTTTC	TCCCACAGTTCTTTAAGTAGC
C12orf65 cDNA[ii]			
Exon 2	164	GCACTGCTTTCCTTGGATGA	CTGTTCTGATCAACTGATCTTG
Exon 3	135	GCTGAAGCACATCCCCTCA	TGAACAGGACTGTTTTCACC

[i)]primers used to confirm the mutation at the cDNA level; [ii)]primers used for qPCR.

Table S4. PCR primers for genes in the PARK16 locus

Exon(s)	Product length (bp)	Forward primer	Reverse primer
NUCKS1			
Exon 2	285	TTATGAAAGAGAAATGTGTG	TTGCTTTGGTTGACACAGATC
Exon 3	262	TCCAACAATTTAGAGATCAG	TTATGTCAGTAAATATAGTGGTG
Exon 4	262	AAACGTCTTTGTAATCCC	CTTCTCAAGTCAGAAATCAG
Exon 5	322	GCCAAAGTAACCAAGGTC	CAGTCAGTTATGTTTTCAGTC
Exon 6	597	AATGAGGAATTTCTGTATTG	CCTCTAGTCATTCTTAGCAC
SLC41A1			
Exon 1	597	AAAGATGAAGCAGTCAATC	GATTGTTGTATTTTGTAGCTG
Exon 2	458	CATCTGAAGAGCAGGGAG	GAATTTGGTTGAGCTCTTCTAC
Exon 3	274	GTGGAGAAGGGCTCGTAG	TCAGACTACCTGTTGAACCTC
Exon 4	544	TGAGAGGAAGAGAGGCGACTG	GCACCACCAAGAACTGAGAAC
Exon 5-6	597	ACACAGAAGAAACTCTTGCAC	CCAGATGCTTATGACCTCAG
Exon 7	346	TGAACACATTTCTAGGGTCTG	ATCTGTTGGGACATTTTCAG
Exon 8	392	CTTGTTCTGGAGTCACTGTGC	GCTGGAACCCTGACTGATAC
Exon 9	311	AGGAGAACCTGTGGATCTG	AGTGAGAAGATCATTTTGTTCC
Exon 10	310	CCTGCTCCTGCCACTTTGTAC	AGACCACAGGATGCAGAAGTAG
Exon 11	563	GACAAAACTCCCCTGCTC	GAATGGGTCTGATCATGTG

Table S5. PCR and sequencing primers for *EIF4G1* gene

Exon(s)	Product length (bp)	Forward primer	Reverse primer
Exon 8a	605	TGGAGTGACTTGAACTGGGTAC	GCTCATGAATTTCCACTGTGTG
Exon 8b	662	TCTCGCCGAACCCATACTG	CAGGGACCCAGAAACATGTC
Exon 22	403	TGCTAAGAACAAGGCCCAACAG	CTAGTCCCAAGGCAGCCAATG

Table S6. 60 Touchdown 50 PCR cycling program

	1 cycle	8 cycles			20 cycles			16 cycles			1 cycle
Phase	den	den	ann	ext	den	ann	ext	den	ann	Ext	ext
Temp. °C	94	94	60	72	94	60	72	94	50	72	72
Temp. gra	0	0	0	0	0	-0.5	0	0	0	0	0
Duration	5min	30sec	30sec	30sec	30sec	30sec	30sec	30sec	20sec	30sec	5min

den = denaturation; ann = annealing; ext = extension; temp = temperature; gra = gradient.

Table S7. 57 touchdown 52 PCR cycling program

	1 cycle	15 cycles			16 cycles			14 cycles			1 cycle
Phase	den	den	ann	ext	den	ann	ext	den	ann	ext	ext
Temp. °C	94	94	57	72	94	57	72	94	52	72	72
Temp. gra	0	0	0	0	0	-0.5	0	0	0	0	0
Duration	1min	30sec	30sec	30sec	30sec	30sec	30sec	30sec	30sec	30sec	5min

den = denaturation; ann = annealing; ext = extension; temp = temperature; gra = gradient.

Table S6. PSP 5 novel and damaging variants (i.e. nonsense, splicing, frameshift)

Gene	AAChange	Exonic	Depth	Q	Seg	omimID	esp540	Chr;pos
ISG15	NA	splice	12	14	NA	147571	NA	1:94884
H6PD	H6PD:uc001apt.2:wholegene,	frame	31	21	NA	138090;604931	NA	1;93049
PRAM	uc001auj.1:c.T314A:p.L105X	stopai	28	10	0.96	NA	0.17	1;12854
CASP9	uc001awm.1:c.1234_1235ins	frame	20	81	NA	602234	NA	1;15820
NBPF1	NA	spliceg	10	14	0.99	NA	NA	1;14461
KDM4	NA	spliceg	52	78	NA	609764	NA	1;44125
NBPF1	uc010oyd.1:c.G790T:p.E264X	stopai	19	22	0.98	NA	0.02	1;14482
GPATC	uc001fpl.2:c.1068_1069insTG:	frame	51	21	NA	NA	NA	1;15656
FCRL5	uc001fqv.1:c.1731_1732insC:	frame	26	14	0.91	605877	NA	1;15750
CD34	uc001hgv.1:c.71_72insT:p.V2	frame	7	43	NA	142230	NA	1;20806
ANGEL	uc001hkb.2:c.659_660insT:p.L	frame	35	25	NA	NA	NA	1;21318
OR2T2	uc001ieo.1:c.C52T:p.Q18X	stopai	11	51	0.99	NA	0.00	1;24872
OR2T3	uc001ies.1:c.604_610del:p.20	frame	65	21	0.98	NA	NA	1;24880
SLC3A	uc002rty.2:c.1139delT:p.L380	stopai	52	19	NA	220100;606407;104614	NA	2;44528
CD207	uc002shg.2:c.73_74insG:p.E2	frame	2	33	NA	604862	NA	2;71062
CRELD	uc003buh.2:c.959delA:p.Q320	frame	41	21	NA	600309;607170;606217	NA	3;99851
ZNF16	uc011azx.1:c.1472_1473del:p.	frame	8	20	NA	NA	NA	3;44540
OR5K3	uc011bgw.1:c.904_905insA:p.	frame	15	16	NA	NA	NA	3;98110
ATG3	uc003dzc.2:c.911_912insT:p.I	frame	41	20	NA	609606	NA	3;11225
HSPBA	uc003efu.1:c.763_764insTG:p.	frame	25	20	NA	144700;608263	NA	3;12247
SLC41	uc003eik.2:c.1388_1391del:p.	frame	26	18	NA	610803	NA	3;12572
ALDH1	uc010hsf.1:c.1610delG:p.R53	frame	15	22	NA	600249	NA	3;12585
ALG1L	uc011bld.1:c.469delC:p.L157f	frame	28	21	0.96	NA	NA	3;12981
CEP70	uc011bmm.1:c.1338_1522del:	frame	29	21	NA	NA	NA	3;13821
ZNF59	uc011but.1:c.12_13insC:p.S4f	frame	3	54	NA	NA	NA	4;86102
CXCL6	uc003hhf.2:c.239_240insT:p.V	frame	16	18	NA	138965	NA	4;74702
NHED	uc011cev.1:c.C232T:p.R78X	stopai	17	48	0.96	611527	0.07	4;10383
ALPK1	uc010imo.2:c.103_104insT:p.	frame	36	74	NA	607347	NA	4;11334
ZNF47	uc003ksv.2:c.953delT:p.L318X	stopai	42	18	NA	NA	NA	5;12148
ACSL6	uc010jdn.1:c.615_616insCA:p.	frame	3	77	NA	604443	NA	5;13132
TIGD6	uc003lri.2:c.1031delA:p.Q344f	frame	57	21	NA	NA	NA	5;14937
CYFIP2	uc003lwq.2:c.280_281insC:p.	frame	3	35	NA	606323	NA	5;15672
NOP1	uc003med.2:c.576_577insAC:	frame	2	74	NA	NA	NA	5;17581
DSP	uc003mxp.1:c.1_2insA:p.M1fs	frame	6	19	NA	125647;605676;607655;	NA	6;75421
TPMT	uc010jpm.1:c.594_595insTT:p	frame	20	25	NA	187680	NA	6;18134
ZNF18	NA	spliceg	9	93	NA	NA	NA	6;28239
HLA-A	uc011dmd.1:c.673delG:p.V22	frame	31	50	0.91	142800	NA	6;29912
CDSN	uc003nsm.1:c.1589_1617del:	frame	44	21	NA	602593	NA	6;31083
PSORS	uc003nso.3:c.280delC:p.P94fs	frame	14	66	NA	NA	NA	6;31105
HLA-	uc003obj.2:c.C97T:p.R33X	stopai	56	11	0.89	604776;142857	NA	6;32497
HLA-	uc003obt.1:c.745_746insTT:p.	stoplo	6	12	NA	146880	NA	6;32610
HLA-	uc003obt.1:c.684delC:p.S228f	frame	21	25	NA	146880	NA	6;32610
MANE	uc003pon.2:c.567_568insATG	frame	11	10	NA	612327	NA	6;96034
PBOV1	uc003qhv.2:c.347_348insC:p.	frame	46	17	NA	605669	NA	6;13853
AOAH	uc010kxf.2:c.1913_1914insA:	frame	19	21	NA	102593	NA	7;36552
AOAH	uc010kxf.2:c.1973_1974insA:	frame	18	15	NA	102593	NA	7;36552
PKD1L	NA	spliceg	35	12	NA	609721	NA	7;47869
WNT1	uc003vjv.2:c.1_2insCCCA:p.M	frame	45	21	NA	NA	NA	7;12096
MLL3	uc003wla.2:c.2447_2448insA:	stopai	98	21	0.97	606833	NA	7;15194
EIF3E	uc010mcj.1:c.489_492del:p.1	frame	22	21	NA	602210	NA	8;10924
PYCRL	uc003yyy.2:c.743delC:p.T248f	frame	35	10	NA	NA	NA	8;14468
CBWD	uc004agl.3:c.189_190insA:p.K	frame	4	60	0.99	611080	NA	9;70900
OR1B1	uc011lyz.1:c.38_39insT:p.V13f	frame	20	21	NA	NA	NA	9;12539
GPSM	uc004chc.2:c.1242delC:p.P41	frame	3	22	NA	609491	NA	9;13923
ITIH5	uc001ijp.2:c.2154delG:p.M71	frame	46	21	NA	609783	NA	10;7605
PTCHD	uc001itu.2:c.918_919insG:p.L	frame	23	46	0.9	611791	NA	10;2770
CCAR1	NA	spliceg	29	21	NA	612569	NA	10;7054
C10orf	uc009xsn.2:c.123_124insAA:p	frame	10	53	0.9	NA	NA	10;8184
ATRNL	uc001lcf.2:c.1400_1401insTT:	frame	16	21	NA	612869	NA	10;1169
CYP2E	NA	spliceg	19	21	NA	124040	NA	10;1353
C11orf	uc010qyg.1:c.597_598insCT:p.	frame	58	21	NA	NA	NA	11;4592
OR52R	uc010qym.1:c.C637T:p.R213X	stopai	46	22	NA	NA	0.00	11;4825
OR51I	uc010qzf.1:c.715_716insCA:p.	frame	48	21	NA	NA	NA	11;5475
CYB5R	uc001mfm.2:c.830_831insAA	frame	35	21	NA	608342	NA	11;7686
SPI1	uc009ylp.1:c.579_580insGT:p.	frame	9	24	NA	165170	NA	11;4738
OR4C3	uc010rhv.1:c.G522A:p.W174X	stopai	75	22	NA	NA	NA	11;4834
FAM1	uc010rko.1:c.304delT:p.Y102f	frame	32	13	NA	NA	NA	11;5889
SNX32	uc009yqt.2:c.206delG:p.G69fs	frame	19	52	NA	NA	NA	11;6561
ALDH3	uc001ona.2:c.790_791insC:p.	frame	16	16	0.93	600466	NA	11;6778
GDPD	uc001oyf.2:c.1191_1192insT:	frame	32	21	NA	NA	NA	11;7695
MMP1	uc001phi.2:c.572delT:p.I191fs	frame	52	21	NA	120353	NA	11;1026
CASP5	uc010rvb.1:c.19_20insA:p.K7f	frame	25	19	NA	602665	NA	11;1048
HYOU	uc010ryu.1:c.28_29insG:p.G1	frame	26	19	NA	601746	NA	11;1189
TULP3	uc001qlj.2:c.C70T:p.R24X	stopai	38	19	NA	604730	0.00	12;3018
CLECL	uc001qwj.2:c.149_150insACT	frame	30	21	NA	607467	NA	12;9885

GUCY2	uc001rcd.2:c.T2943G:p.Y981X	stopai	35	19	NA	601330	NA	12;1476
C12orf	uc001rcj.3:c.550_551del:p.18	frame	28	21	NA	NA	NA	12;1497
OR10A	uc001rrl.1:c.200_201insT:p.L6	frame	62	21	NA	NA	NA	12;4859
OR9K2	uc010spe.1:c.38delT:p.L13fs	frame	48	18	NA	NA	NA	12;5552
OR6C1	uc010spi.1:c.24_25insA:p.T8fs	frame	32	21	NA	NA	NA	12;5571
RXFP2	NA	spliceg	26	21	NA	606655	NA	13;3236
OR11	uc010tkp.1:c.T699G:p.Y233X	stopai	41	62	1	NA	0.73	14;1937
OR11	uc010tlb.1:c.687_688insA:p.K	frame	59	21	NA	NA	NA	14;2066
FANC	NA	spliceg	10	10	NA	609644;227650	NA	14;4562
FANC	NA	spliceg	18	16	NA	609644;227650	NA	14;4565
NIPA2	uc001yva.2:c.1025_1381del:p	frame	17	21	NA	608146	NA	15;2300
TPSD1	uc002clb.1:c.127delG:p.E43fs	frame	11	99	0.97	609272	NA	16;1306
PHKG2	NA	spliceg	4	11	NA	172471	NA	16;3076
MNT	NA	spliceg	35	14	NA	603039	NA	17;2297
P2RX5	uc002fwh.1:c.329delC:p.T110f	frame	25	15	NA	602836	NA	17;3594
VMO1	NA	spliceg	25	71	NA	NA	NA	17;4689
MAP2	uc002gys.2:c.C304T:p.Q102X	stopai	68	20	NA	602315	NA	17;2120
POLDI	NA	spliceg	2	32	NA	611519	NA	17;2668
SARM	NA	spliceg	3	71	NA	607732	NA	17;2669
SLC46	uc002hbg.1:c.1142delT:p.I381	frame	2	22	NA	611672;229050;607732	NA	17;2672
MMP2	uc002hjy.1:c.436_437insG:p.P	frame	7	12	NA	608417	NA	17;3410
DHRS1	NA	spliceg	50	21	NA	NA	NA	17;3495
DSC3	uc002kwi.3:c.1540delC:p.P51	frame	23	11	NA	600271	NA	18;2858
KIAA1	uc002lik.1:c.3008_3009insT:p.	frame	29	81	NA	NA	NA	18;5994
SERPI	uc010dqb.2:c.354_355insA:p.	frame	42	21	0.93	600518;600517	NA	18;6132
ZNF56	uc010xkx.1:c.420_424del:p.14	frame	12	21	0.91	NA	NA	19;9801
HKR1	uc002ofz.2:c.512_513del:p.17	frame	2	71	NA	165250	NA	19;3785
CYP2F	uc002opu.1:c.15_16insC:p.S5f	frame	49	21	NA	124070	NA	19;4162
BCKD	uc010xvz.1:c.972_973insC:p.P	frame	26	21	NA	248600;608348	NA	19;4192
CEACA	uc010ejp.1:c.1169delA:p.E390	frame	2	27	NA	NA	NA	19;4501
HAS1	NA	spliceg	48	19	NA	601463	NA	19;5222
ZNF48	uc010ydl.1:c.9_10del:p.3_4de	frame	48	21	NA	NA	NA	19;5280
ZNF76	uc010eqp.2:c.139delC:p.L47fs	frame	10	94	0.91	NA	NA	19;5395
AURKC	NA	spliceg	6	74	NA	243060;603495	0.76	19;5774
ZNF27	uc002qrs.1:c.214_215insG:p.A	frame	7	10	NA	605467	NA	19;5871
DEFB1	uc002wcx.2:c.317_318del:p.1	frame	31	15	NA	NA	NA	20;1263
TTLL9	NA	spliceg	2	23	NA	NA	NA	20;3052
CPNE1	uc002xdc.2:c.187_188insT:p.F	frame	32	20	NA	604205	NA	20;3421
BIRC7	uc010gkc.1:c.672_673insG:p.	frame	13	16	NA	605737	NA	20;6187
DNAJC	uc002yrv.2:c.940_944del:p.31	frame	40	21	NA	NA	NA	21;3486
CLTCL	NA	spliceg	2	34	NA	601273	NA	22;1918
PI4KA	uc002zuv.3:c.1624_1627del:p	frame	7	55	0.99	NA	NA	22;2182
APOL4	uc003aox.2:c.330_331insAG:p	frame	3	12	NA	607254	NA	22;3658
C22orf	uc003aqe.2:c.G498A:p.W166X	stopai	34	20	NA	NA	NA	22;3738
SSTR3	uc003ara.2:c.1257_1258insG:	frame	27	21	NA	182453	NA	22;3760
RRP7A	NA	spliceg	17	20	0.97	607979	NA	22;4291
PHKA1	uc010nll.2:c.89_90insA:p.Q30	frame	8	27	NA	300559;311870	NA	X;71813
TCEAL	uc004eiq.2:c.521_522insC:p.A	frame	8	13	0.92	NA	NA	X;10139
FAM5	uc010nug.2:c.50_51insG:p.R1	frame	3	30	NA	300707;300708	NA	X;15286
Nbla1	NA	spliceg	2	81	NA	605477	NA	13;1119
BCR	uc002zwy.1:c.577_578insC:p.	frame	35	91	NA	608232;151410	NA	22;2363
AX746	uc003lhy.1:c.523delG:p.G175f	frame	12	70	NA	NA	NA	5;14024
PCDH	uc003liw.1:c.1141_1142insA:	frame	13	12	NA	606335	NA	5;14056
MSTP	uc010ock.1:c.117_121del:p.3	frame	60	12	0.99	NA	NA	1;17087
AL359	uc003vwg.1:c.G35A:p.W12X	stopai	2	20	NA	NA	NA	7;14117
KIAA1	uc002ysf.1:c.1320_1321insA:	frame	33	15	0.96	182465	NA	21;3494
pp144	uc003yyi.2:c.442_443insAGG	frame	10	49	NA	NA	NA	8;14464
AK126	uc002jku.2:c.170delG:p.G57fs	frame	14	79	NA	605949	NA	17;7243
FRG1B	uc010ztk.1:c.260_261insA:p.K	frame	28	11	0.96	NA	NA	20;2963

Table S7: *EIF4G1* coding variants present in NHLBI Exome Sequencing Project.

EA= European American population; AA=African American population

Nucleotide change	Protein change	SNP accession number	Frequency (EA)	Frequency (AA)	Exon
c.C71T	P24L		0.000142	0	2
c.C167G	A56G		0.000142	0	3
c.C211T	P71S	rs113810947	0.000285	0	3
c.282C>G	I94M		0	0.000268	3
c.451C>G	V151L		0.000143	0	5
c.481A>G	A161T	rs13319149	0.996722	0.999465	5
C.602G>A	R201H	rs34838305	0.000427	0	6
C.608C>T	A203V		0.000142	0	6
C.704G>A	R235Q	rs144543953	0.000142	0.000268	8
C.731G>A	R244Q	rs147855566	0.000142	0	8
C.779C>T	S260L		0.000142	0	8
c.821C>T	P274L	rs139626338	0	0.000268	8
c.870G>A	M290I	rs144947145	0	0.001338	8
c.913C>T	R305C	rs116508885	0	0.00321	8
c.914G>A	R205H	rs151151194	0.000142	0.000268	8
c.926A>G	E309G		0.000142	0	8
c.932A>G	Y311C	rs16858632	0.000427	0.059658	8
c.1001A>C	P334Q		0.000285	0	8
c.1013C>T	S338F	rs139021806	0	0.000268	8
c.1036C>A	Q346K		0	0.000268	8
c.1054G>T	A352S		0.000142	0	8
c.1063A>G	T355A		0	0.000268	8
c.1064C>T	T355I		0	0.000268	8
c.1142C>G	A381G	rs142095694	0.000142	0	8
c.1256G>T	S419I	rs138207269	0	0.000535	8
c.1294A>G	M432V	rs2178403	0.759544	0.943553	8
c.1298C>T	A433V	rs145998921	0.000142	0	8
c.1309A>G	I437V	rs144222028	0	0.000268	8
c.1316C>A	S439Y	rs148709174	0.000142	0	8
c.1331C>T	T444M	rs143014570	0	0.000268	8
c.1352C>A	P451Q	rs147419996	0.000142	0	8
c.1429G>A	E477K	rs145228718	0	0.000268	8
c.1456C>T	P486S	rs112545306	0.00057	0.000803	8
c.1505C>T	**A502V**	**rs111290936**	**0.000285**	**0**	**8**
c.1610C>T	A537V		0	0.000268	10
c.1648G>C	A550P	rs111924994	0.001994	0.000535	10
c.1679G>A	G560D	rs149685875	0.000142	0	10
c.1696C>T	R566C	rs145521479	0.000142	0.000268	10
c.1700C>A	P567H	rs140212150	0	0.000268	10
c.1754A>C	E585A		0	0.000268	10
c.1801T>C	W601R	rs145247318	0	0.000268	11
c.1831C>T	R611C		0	0.000268	11
c.1980A>G	I660M		0.000142	0	12

c.1982A>G	N661S	rs145780534	0.000142	0	12
c.2096G>C	G699A		0.000142	0	13
c.2114C>G	S705C	rs141054452	0.000142	0	13
c.2152G>C	A718P	rs111396765	0.001567	0.000268	13
c.2225C>T	T742M	rs147678593	0	0.004013	13
c.2276A>G	Q759R		0.000142	0	14
c.2278G>C	D760H	rs142947014	0.000142	0	14
c.2386A>G	K796E		0	0.000268	14
c.2419A>G	I807V	rs62287499	0.000427	0	14
c.2488A>T	T830S	rs111500185	0.000285	0	15
c.2612A>C	E871A		0.000142	0	15
c.2671A>G	I891V		0	0.000268	16
c.2882A>G	N961S	rs191888688	0	0.000268	17
c.3187C>T	R1063C		0.000142	0	19
c.3221C>T	T1074I	rs146433145	0.000142	0	19
c.3343C>T	R1115C	rs150054202	0	0.000268	21
c.3428A>G	Q1143R	rs145414660	0	0.000268	21
c.3482G>A	R1161H	rs139135683	0.000427	0	22
c.3511C>T	R1171C	rs141684202	0	0.000268	22
c.3529C>T	R1177C		0	0.000268	22
c.3580C>T	R1194W		0.000142	0	22
c.3584G>A	S1195N		0.000142	0	22
c.3592C>T	R1198W	rs113388242	0	0.000268	22
c.3617G>A	R1206H	rs112176450	0.000285	0	22
c.3649C>T	R1207C		0.000143	0	22
c.3650G>A	R1217H	rs34086109	0	0.010433	22
c.3652G>A	G1217R	rs138270117	0	0.000268	22
c.3686C>T	P1299L		0	0.000268	23
c.3688C>G	P1230A	rs35629949	0.005842	0.001606	23
c.3701T>C	L1234P	rs2230570	0.021937	0.080257	23
c.3743A>G	K1248R		0.000142	0	23
c.3773A>G	N1258S	rs73053766	0	0.001873	23
c.3935C>T	S1312F		0	0.000268	24
c.3937A>G	T1313A	rs144570332	0.000142	0.000803	24
c.3988A>G	M1330V	rs112809828	0.000285	0	25
c.4067T>C	M1356T	rs144059151	0.000855	0	25
c.4068G>C	M1356I	rs145975905	0.000285	0	25
c.4081A>G	R1361G	rs139793721	0	0.000268	25
c.4106C>T	P1369L	rs142064428	0	0.000803	26
c.4184C>T	T1395M	rs112441721	0	0.000268	27
c.4201G>A	G1401R	rs149821418	0.000142	0.000803	27
c.4229A>C	E1410A	rs141776790	0	0.000268	27
c.4259A>G	E1420G		0	0.000268	27
c.4292C>T	S1431L		0.000142	0.000268	28
c.4300C>T	P1434S	rs147696097	0.000142	0	28
c.4379G>A	R1460Q		0.000142	0	28

c.4399G>A	A1467T	rs148270724	0.000142	0	29
c.4433C>T	T1478M	rs141379472	0.000142	0	29
c.4454C>T	T1485M		0.000142	0	29
c.4486A>T	T1496S		0	0.000268	29
c.4712C>T	A1571V	rs144462594	0.000142	0	31
c.4772G>A	R1591H		0.000142	0	31