UNIVERSITÀ
DEGLI STUDI
DI PADOVA

UNIVERSITA’ DEGLI STUDI DI PADOVA

Dipartimento di Biologia

SCUOLA DI DOTTORATO DI RICERCA IN BIOSCIENZE E BIOTECNOLOGIE
INDIRIZZO: GENETICA E BIOLOGIA MOLECOLARE DELLO SVILUPPO
CICLO XXVI

TESI DI DOTTORATO

IDENTIFICATION AND CHARACTERISATION OF NOVEL GENES IN MOTOR NEURON DISORDERS

Direttore della Scuola: Ch.mo Prof. Giuseppe Zanotti
Coordinatore d’indirizzo: Ch.mo Prof. Rodolfo Costa
Supervisore: Ch.ma Prof. Maria Luisa Mostacciuolo
Co-supervisore: Dr. Giovanni Vazza

Dottoranda: Elisa Gregianin

31 Gennaio 2014

TABLE OF CONTENTS

SUMMARY

Introduction. Hereditary Motor Sensory Neuropathy (HMSN), also known as Charcot-Marie-Tooth disease (MIM118300), is an heterogeneous group of Mendelian diseases which affects the peripheral motor and sensory nervous system (PNS). With a prevalence of about 1/2500 individuals, it is considered the most common inherited neuromuscular disorder. The clinical hallmarks include slow and progressive muscular weakness of distal limbs, peroneal atrophy and bilateral *pes cavus*, however they are often associated with other signs such as spastic paraplegia, ataxia, mental retardation and incontinence. Distal HMN disorders are a subgroup of HMSN characterized by a predominant motor involvement and minor or no sensory loss. To date, more than 50 genes have been detected for HMSN and 17 genes for dHMN, but many disease-genes have to be identified yet.

Aim of the study. This study aims to identify and, whenever possible, functionally characterise novel genes causing different forms of peripheral neuropathy. In order to achieve this objective, four families affected by complex HMSN or distal HMN and showing no mutations in known disease-genes were examined.

Materials and methods. The identification of novel genes was performed by using a strategy which integrates the traditional positional cloning approach with the next-generation whole-exome sequencing. In all the four families candidate linkage regions were identified by genome-wide linkage analysis, specifically for the recessive forms they were identified also by homozygosity mapping and identity-by-descent analysis using high-density SNPs arrays and STR markers. In one family a candidate-genes screening by Sanger sequencing was preferred for the small size of the critical region. In the remaining three families the whole-exome sequencing (WES) approach was adopted in order to inspect all coding variants within the previously identified linkage regions. Considering the advantage of having variants within the whole exome, known disease-related genes were also analysed and less stringent genome wide searches were even conducted. Moreover, coverage analysis was performed and poorly-covered exons were sequenced. The best candidate variants were selected by filtering and prioritization analyses which allowed to evaluate the putative pathogenicity of variants. For the variants which represented potential mutations, further confirmations were obtained by *in silico* analyses, control population screening, unrelated patients characterization and functional studies.

Results. In Family 1, affected by a recessive HMSN and spastic paraplegia, only one small candidate region encompassing 5 genes was identified by linkage studies on chr13q12.1-

q12.12. Direct sequencing highlighted in the patients a novel homozygous missense mutation in the *SACS* gene (c.11104A>G). *SACS* was already associated with a different neurodegenerative disorder termed ARSACS. *In silico* predictions showed that the p.Thr3702Ala change falls in a surface-exposed and evolutionarily conserved region of XPCB protein domain, suggesting a key role in the interaction with E3 ubiquitin ligase UBE3A. Recent studies in mutated fibroblasts indicated an alteration in the dimensions of mitochondria and such impairment was assessed for the p.Thr3702Ala mutation identified. A higher number of significantly more little and spherical mitochondria was observed in proband's fibroblasts compared with controls. These findings were indicative of a higher fragmentation of mitochondria in presence of the p.Thr3702Ala sacsin mutation, in agreement with the hypothesis of impairment in fusion/fission mitochondrial dynamics.

Family 2, affected by a distal form of motor neuropathy (dHMN), is characterized by the presence of three consanguineous marriages. Despite this, the homozygosity mapping failed to detected a unique candidate autozygous region shared by all affected subjects. In order to investigate the hypothesis of two genetic causes, the family was splitted into two nuclei on the basis of phenotypic differences between patients. In the first nucleus, linkage analysis highlighted a single candidate autozygous region on chr8p23.1-p22 (rs2738148; rs6997599) and the analysis of the WES variants identified a novel missense substitution in the *SGK223* gene (c.1529T>C). *In silico* predictions strongly supported the effect of this variant in a splicing alteration. In the second nucleus, 8 candidate regions (>1 Mb) were identified by homozygosity mapping but only one of them (chr9p21.1-p13.2) was pinpointed by the IBD analysis and was likely to be identical-by-descent (IBD). The analysis of the WES variants selected a candidate mutation (c.412G>C) in *SIGMAR1*. This gene was already associated with the motor neuron disorder ALS16. By screening *SIGMAR1* in unrelated patients, a second variant (c.448G>A) was identified in a different family displaying a recessive form dHMN.

In the consanguineous Family 3, affected by a recessive complex form of neuropathy with spastic paraplegia and mental retardation, linkage analysis detected 4 candidate regions on chr2p13.3-p12, chr3q27-q28 (2 regions) and chr21q11.2-q21.1. The analysis of WES variants enabled to identify in chr2p13.3-p12 an interesting variant (c.950G>A) in the *FBXO41* gene. This substitution has been never reported in any variant database, was predicted to map in a evolutionary conserved region and to play a disruptive functional effect in a α-helix structure.

In Family 4, affected by an autosomal dominant form of HMSN and spastic paraplegia with a marked clinically heterogeneity, different genetic models were investigated including monogenic, digenic and the co-occurrence of two distinct diseases. A high number of candidate regions was thus identified by the linkage analysis, with the best candidate region on chr9q22.33-9q33.2. Due to a low quality of WES data, the analysis of sequence variants highlighted many false positive calls and no variants displayed perfect co-segregation with the disease. Also the CNV analysis and the direct sequencing of the poorly-covered exons in the chr9q22.33-9q33.2 region did not identified any other significant mutation.

Discussion. The high clinical and genetic heterogeneity which characterizes distal neuropathies represents a clue that the presence of a single genetic cause and a straightforward genotype-phenotype correlation cannot explain all Mendelian diseases. Since to date many of these cases still remain “orphan” of a molecular explanation, the research of novel disease-genes attempts to shed light on the genetics and pathogenic mechanisms which are not-well known yet. The combinatorial strategy used in the present study was powerful for identifying candidate disease-genes in the families with a recessive transmission (Families 2 and 3). Indeed, the information on the candidate regions enabled to reduce the great deal of variants in order to perform a deeper and more effective analysis. On the other hand, this approach revealed more challenging in the picture of dominant transmission and clinical heterogeneity of Family 4, proving the already-reported difficulty to unravel dominant traits. However this part of the study put in evidence technical and analytical weaknesses which this analysis attempted to limit and that the currently-performed WES will enable to overcome.

This work represents the first evidence for a *SACS* mutation associated with a non-ataxic clinical picture, and for two *SIGMAR1* mutations associated with a juvenile distal neuropathy. Considering also that the *SIGMAR1* gene maps to the “orphan” dHMN Jerash type *locus* (dHMN-J), this finding supports the causality of this gene in dHMN disorder. The demonstration of a growing phenotypic heterogeneity up to an overlap with other neurological disorders can be indicative of a more extended phenotypical spectrum associated with these genes. This work represents the first step to demonstrate the pathogenicity of a candidate variant and dissect the mechanisms underlying distal neuropathies. Indeed, further genetic screenings in a large cohort of patients and functional studies will elucidate the actual involvement of the novel candidate genes in these disorders.

RIASSUNTO

Introduzione. La neuropatia ereditaria sensitivo-motoria (HMSN), anche denominata malattia di Charcot-Marie-Tooth (MIM118300), costituisce un gruppo eterogeneo di disordini mendeliani che colpiscono il sistema nervoso periferico motorio e sensitivo. Tale malattia è considerata il disordine neuromuscolare ereditario più comune, con una prevalenza stimata di 1 caso su 2500 nella popolazione. I segni clinici distintivi includono una progressiva ipostenia e ipotrofia dei muscoli distali degli arti, un'atrofia peroneale e piede cavo bilaterale. Questo fenotipo clinico è di frequente associato ad altri sintomi, quali la paraparesi spastica, il ritardo mentale, atassia e disturbi urinari. Le neuropatie ereditarie motorie distali (dHMN) costituiscono un sottogruppo delle neuropatie ereditarie sensitive-motorie, caratterizzato da un prevalente coinvolgimento motorio e un minore o assente interessamento sensitivo. Ad oggi, più di 50 geni sono stati associati alle neuropatie ereditarie sensitivo-motorie e 17 alle forme motorie distali, tuttavia molti geni rimangono ancora sconosciuti.

Obiettivo dello studio. Il presente studio si prefigge di identificare e possibilmente caratterizzare dal punto di vista funzionale, nuovi geni causativi di varianti ereditarie di neuropatia periferica. A tale scopo sono state studiate quattro famiglie in cui erano presenti soggetti affetti da HMSN complesse o dHMN e nelle quali erano state escluse le mutazioni nei geni malattia già noti.

Materiali e Metodi. Al fine di identificare nuovi geni causativi, in questo lavoro è stata adottata una strategia che va ad integrare l'approccio tradizionale di clonaggio posizionale con il sequenziamento next-generation dell'esoma. Nelle quattro famiglie oggetto di studio, mediante SNPs ad alta densità e marcatori microsatelliti, è stata svolta una analisi di linkage genome-wide e si sono identificate regioni candidate; inoltre per le forme recessive si è proceduto con il mappaggio per omozigosità e l'analisi di identità per discendenza (IBD). In una di queste famiglie, data la ridotta estensione della regione candidata, lo screening mutazionale dei geni candidati è avvenuto per mezzo del sequenziamento diretto. Nelle rimanenti tre famiglie, invece, è stato adottato l'approccio di sequenziamento next-generation dell'esoma, al fine di analizzare tutte le varianti presenti negli esoni codificanti delle regioni in linkage. Il vantaggio di avere un'informazione estesa all'intero esoma ha permesso un'analisi genome-wide meno stringente e la ricerca delle varianti nei geni malattia già noti. L'analisi dell'efficienza di copertura delle regioni candidate ha permesso invece di sequenziare gli esoni poco coperti di queste regioni. Un'approfondita e dettagliata analisi di filtraggio e prioritizzazione ha reso possibile valutare la possibile patogenicità delle varianti e di identificare le migliori candidate. A partire da queste, ulteriori validazioni sono state poi

ottenute mediante studi *in silico*, screening di popolazione di controllo, caratterizzazione di pazienti non imparentati e studi funzionali.

Risultati. Nella prima famiglia, in cui segrega una forma recessiva di HMSN e paraparesi spastica, si è identificata una sola regione di linkage sul cromosoma 13q12.1-q12.12 e contenente 5 geni. Per mezzo del sequenziamento diretto è stata individuata una nuova mutazione missenso in omozigosi nel gene *SACS* (c.11104A>G). Tale gene era già stato associato in precedenza a una diversa malattia neurodegenerativa denominata ARSACS. Predizioni *in silico* hanno evidenziato che il cambiamento aminoacidico (p.Thr3702Ala) si localizza esposto sul dominio proteico XPCB e in una regione evolutivamente conservata nella sequenza aminoacidica. Alla luce di questi dati, si ipotizza per l'aminoacido mutato un ruolo chiave nell'interazione con l'ubiquitin-ligasi E3 UBE3A. Visti i recenti studi condotti su fibroblasti mutati che hanno riportato un'alterazione delle dimensioni dei mitocondri, si è ritenuto interessante verificare se tale alterazione avvenisse anche nei fibroblasti con la mutazione identificata. In tali fibroblasti di paziente, confrontati con fibroblasti di controllo, si è osservato un maggior numero di mitocondri caratterizzati da dimensioni statisticamente inferiori e da una maggiore sfericità. Tali dati potrebbero essere indicativi di una loro maggiore frammentazione in presenza della mutazione missenso, in accordo con il dato di alterazione riportato in precedenza.

La seconda famiglia investigata, affetta da una forma distale di neuropatia motoria, è caratterizzata dalla presenza di tre matrimoni fra consanguinei. Tuttavia il mappaggio per omozigosità ha escluso un'unica regione di autozigosi condivisa da tutti i soggetti affetti. Al fine di investigare l'ipotesi della presenza di due cause genetiche diverse, la famiglia è stata divisa in due nuclei sulla base delle differenze fenotipiche dei pazienti. Per il primo nucleo l'analisi di linkage ha identificato una singola regione di autozigosi sul cromosoma 8p23.1-p22 (rs2738148; rs6997599) e la successiva analisi di varianti ha individuato una nuova sostituzione missenso nel gene *SGK223* (c.1529T>C) come la migliore candidata. Il suo effetto nell'alterazione dello splicing è stato fortemente supportato da studi di predizione *in silico*. Nel secondo nucleo invece, l'approccio di mappaggio per omozigosità ha identificato 8 regioni candidate (>1 Mb) ma solo una di queste (chr9p21.1-p13.2) è stata trovata in autozigosi dall'analisi IBD. L'indagine delle varianti ottenute dal sequenziamento dell'esoma e presenti in questa regione ha selezionato una mutazione fortemente candidata (c.412G>C) in *SIGMAR1*. Tale gene è già stato associato in precedenza ad una diversa patologia neurologica quale la sclerosi laterale amiotrofica giovanile denominata ALS16. Mediante lo screening mutazionale di *SIGMAR1* in altri pazienti con dHMN e non imparentati tra loro, è stata identificata una seconda variante (c.448G>A) in un'altra famiglia che presentava una analoga trasmissione autosomica recessiva.

La terza famiglia oggetto di studio, anch'essa con consanguineità, presenta invece una forma complessa di neuropatia con paraparesi spastica e ritardo mentale. Mediante analisi di linkare sono state identificate quattro regioni candidate e nello specifico sul cromosoma 2p13.3-p12, 3q27-q28 (2 regioni) e 21q11.2-q21.1. L'analisi delle varianti ha evidenziato nella regione candidata sul cromosoma 2p13.3-p12 un'interessante sostituzione (c.950G>A) nel gene *FBXO41*, che ad oggi non risulta essere riportata in alcun database di varianti. Studi *in silico* hanno indicato un effetto funzionale del cambiamento aminoacidico nella struttura ad α-elica dove si localizza e una regione altamente conservata a livello evolutivo.

La quarta famiglia, affetta da una forma autosomica dominante di neuropatia ereditaria sensitivo-motoria, è caratterizzata da una marcata eterogeneità clinica in quanto in più soggetti è presente anche paraparesi spastica ereditaria. Per spiegare tale eterogeneità, sono stati adottati diversi modelli genetici che prevedono l'esistenza una singola causa genetica, due cause che concorrono allo stesso fenotipo (modello digenico), e la co-presenza di due malattie distinte. Dall'analisi di linkage è emerso un alto numero di regioni, tra cui la miglior candidata mappante sul cromosoma 9q22.33-9q33.2. Tuttavia la bassa qualità dei dati di sequenziamento ha portato ad un'alta percentuale di falsi positivi, e nessuna variante nelle regioni candidate ha dimostrato una perfetta co-segregazione con la malattia. Risultati negativi sono stati ottenuti anche dall'analisi di Copy Number Variations (CNVs) e dal sequenziamento diretto degli esoni a basso coverage della regione 9q22.33-9q33.2.

Discussione. L'alta eterogeneità clinica e genetica che caratterizza le neuropatie distali è indicativa del fatto che un'unica causa genetica e un'univoca correlazione tra fenotipo e genotipo non possono essere associate a tutti i disordini mendeliani. Alla luce dei molti casi che ad oggi rimangono irrisolti dal punto di vista genetico, la ricerca di nuovi geni malattia ha il fine di far luce sulle basi genetiche e i meccanismi patogenetici che per molti aspetti sono ancora sconosciuti. La strategia usata nel presente studio è risultata efficace ai fini dell'identificazione di geni malattia candidati nelle famiglie con forme a trasmissione recessiva (seconda e terza famiglia). La conoscenza delle regioni candidate ha infatti permesso di ridurre le varianti ottenute col sequenziamento dell'esoma ad un gruppo limitato e analizzabile mediante un'approccio più dettagliato e critico e che ha reso possibile l'efficacia dell'analisi. Lo stesso approccio si è rivelato meno efficace in un quadro di trasmissione dominante ed eterogeneità clinica (quarta famiglia), mettendo in luce le difficoltà di risolvere i tratti dominanti. Tuttavia questa parte dello studio ha permesso di comprendere le criticità tecniche ed analitiche presenti al fine di limitarle, ma che sicuramente verranno superate dai sequenziamenti più recenti.

Tale lavoro rappresenta la prima evidenza di una mutazione nel gene *SACS* associata a un quadro clinico non atassico, e di due mutazioni nel gene *SIGMAR1* associate a una forma giovanile di neuropatia distale. Inoltre *SIGMAR1* mappa in un locus già descritto per una forma di neuropatia distale (tipo Jerash) e tale dato supporterebbe la causatività di *SIGMAR1* per tale malattia. La dimostrazione di una crescente eterogeneità fenotipica che si estende fino alla sovrapposizione con altre malattie neurologiche può essere indicativa di uno spettro fenotipico ben più ampio di quello associato fino ad oggi a questi geni. Tale studio costituisce il primo step fondamentale per trovare nuovi geni causativi e studiare nel dettaglio i meccanismi alla base delle neuropatie distali. A tale scopo screening genetici su coorti di pazienti e studi funzionali potranno far luce sull’effettivo coinvolgimento di questi nuovi geni candidati nelle malattie.

1. INTRODUCTION

1.1

1.1.1 Approaches for causal genes identification in Mendelian disorders: state of the art

Identifying genes responsible for inherited diseases is fundamental to elucidate their pathogenetic mechanisms and better understand the biological pathways and cellular processes involved. Uncovering the molecular basis of a genetic disorder opens the possibility to molecular diagnosis for patients and carriers, and to prenatal testing; moreover it represents the starting point for developing a future therapy. Although many advances in human and medical genetics have occurred, the identification of disease genes within the whole genome still represents a big challenge, and more than half of about 7500 Mendelian disorders classified based on clinical features have not been genetically characterized yet. Indeed, up to 25 November 2013, a total of 22,111 entries were reported in OMIM (Online Mendelian Inheritance in Man) describing 14,432 genes, 3,977 phenotypes with identified molecular basis and 3,593 with still unknown genetic cause (http://omim.org/statistics/entry).

Amongst the traditional approaches of gene discovery used for Mendelian disorders, Sanger sequencing and linkage analysis with positional cloning found the broadest applicability in the previous three decades. Thus far, Sanger sequencing method has been effective for screening a limited number of genes which have already been associated with a specific disease. This approach revealed suitable for disorders that are characterized by a clear and straightforward genotype-phenotype correlation and for which a deep medical and biological prior knowledge is available.

On the other hand, when the prior genetic knowledge is limited or the phenotype is not closely related to a underlying unique genotype, genetic mapping has turned out the approach of choice. This represents an unbiased approach and allows to analyse the whole genome. This method has been particularly powerful in identifying genetic mutations with a large effect size and rare prevalence in the population.

Over the last years, with the introduction of massively parallel or Next Generation Sequencing (NGS) technologies, several limits of the traditional approaches have been overcome and novel causal genes have been discovered. The whole genome sequencing (WGS) and the target sequencing, such as the whole exome sequencing (WES), have been increasingly used to shed light on Mendelian disorders.

1.1.2 Linkage analysis

Before the advent of the NGS, the genome-wide linkage analysis followed by positional cloning was considered the main tool to unravel the genetics of Mendelian diseases in families. Noteworthy examples are cystic fibrosis and Huntington disease, for which *CFTR* and *HTT* genes were identified (Group 1993;Tsui *et al.,* 1985).

The statistical method of linkage analysis is based on the theory that the absence of recombination events between genetic markers and the disease gene during meiosis makes to infer they are physically close together on a chromosome. This analysis starts from genotyping data of polymorphic genetic markers such as short tandem repeats (STRs) and single nucleotide polymorphisms (SNPs), and aims to highlight genomic *loci* co-segregating with the disease allele, which represent the candidate regions where the putative disease-associated gene map. In order to pinpoint the causal mutations in the causal gene, genes within candidate regions are subsequently prioritised based on expression, function and homology data available in bioinformatic databases (candidate positional cloning method). The best candidate genes are thus screened with Sanger sequencing to look for variants which are potential mutations in the disease.

Even if many cases have been solved by this traditional approach, some disorders present a challenge for them. For instance, the linkage approach may display some weaknesses in cases of incomplete penetrance, misdiagnosis and occurrence of *de novo* variants. Moreover low-prevalence disorders with a small number of available unrelated cases, or a pedigree with few individuals, constrain the effectiveness of this analysis (Ku *et al.,* 2011). The following step of candidate positional cloning could be arduous as well, even if the human genome reference sequence is available today (Consortium 2004). The causal gene can indeed be missed due to incomplete information about genes or difficulties in sequencing large linkage regions.

1.1.3 Next Generation Sequencing (NGS) approach and “second generation” sequencers

In 2003 the draft sequence of the human genome was completed by the Human Genomic Project by using almost exclusively the “first generation” sequencing or Sanger’s method (Consortium 2004;Lander *et al.,* 2001). This goal fuelled the advances in sequencing technology, and in few years the first individual genome (James Watson) was successfully sequenced by the Illumina NGS platform, termed “second generation” sequencer (Wheeler *et al.,* 2008).

Although the capillary-based Sanger sequencing remains the "gold standard" due to its 99.999% per-base accuracy, the revolutionary NGS technique reveals extremely advantageous in terms of costs, time and number of samples that can be analysed (Shendure and Ji 2008). High-throughput NGS method is characterised by highly parallelized reactions which use smaller reagent volumes in little areas, thus enabling the sequencing of thousands to millions molecules at once. NGS generates shorter sequences termed "reads", with a lower per-base accuracy compared with the Sanger sequencing. However this limitation is compensated by the high number of overlapping reads, quantified as "depth of coverage". Furthermore, the most time-consuming step of cloning genomic libraries in bacteria, is here overcome by an *in vitro* reaction.

Despite of traditional linkage approach, NGS directly detects both common and rare variants, and enables to identify causal mutations even starting from a small number of patients. In addition the discrete information of genotyping moves to a continuous one with a single-nucleotide resolution in NGS. The gene identification in Mendelian disease is thus transforming and shifting from the identification to the interpretation of data.

In general, NGS relies on DNA synthesis or ligation chemistry strategies to read through clones of DNA templates simultaneously and in a parallel fashion. By exploiting DNA synthesis cycles alternating with imaging-based acquisitions, sequencers collect data as short reads, each of them corresponds to a clone that is spatially clustered on substrate. For signal detection, "second generation" sequencers take advantage of Sanger's synthesis chemistry, based on fluorescently labelled nucleotides and optical recording. NGS technology varies amongst the main platforms provided by different sequencing companies. Roche 454 pyrosequencing employs emulsion PCR for clonal amplification and cyclic synthesis; Illumina (Solexa) uses bridge PCR and reversible terminator sequencing, Applied Biosystem/Life Technologies SoLiD requires emulsion PCR on beads and ligation chemistry for sequencing (Hui 2012).

1.1.4 Whole-genome and exome sequencing

Exome comprises near to 180,000 protein-coding exons in 23,500 total genes and constitutes approximately 1-2% of the whole genome. Coding sequence has proven to be the prevalent source for the study of disease genes, as exome harbors 85% of the causal mutations of Mendelian diseases that have been solved up to now (Choi *et al.*, 2009).

As the research of causing mutations requires sifting through a great deal of sequence data, the whole-exome sequencing (WES) allows to reduce them to a relatively limited subset compared with genome sequencing. Indeed, each genome contains 3.2 billion nucleotide and 3-4 million sequence variants, whereas each exome about 30 million nucleotides and 25,000 high-quality single nucleotide variants (Marian 2012). In addition, coding regions are less polymorphic because subjected to a major selective pressure, and less repetitive, thus reducing misalignment of reads and errors of interpretation. Furthermore the current ability to interpret the functional effects of non-coding sequence variations is highly limited. For all these reasons, to date an increasing number of genetic studies for Mendelian disorders uses the WES approach. Indeed, in the last four years the WES has been confirmed the successful approach for identifying genes where traditional methods failed.

1.1.5 Exome-sequencing technology

The WES experimental pipeline is characterized by three main steps, with 2 and 3 steps that vary depending on capture kit and platform employed:

1. genomic library preparation
2. target enrichment
3. NGS of the eluted target fragments.

1. The genomic DNA is randomly sheared by sonication, nebulization or enzymatic digestion to get desired fragments of about 200-250 bp. Fragments ends are repaired by T4 DNA ligase and ligated to universal adaptors. Taking advantage of standard primers that anneal with the adaptors, fragments are amplified by low-cycle PCR (Shendure and Ji 2008).

2. The coding-sequences of the genomic library are captured by hybridization of biotinylated RNA or DNA probes (Bainbridge *et al.,* 2010;Gnirke *et al.,* 2009). This represents the main method, but other methods such as selective amplification by PCR or enzymatic methods can be used. The hybridization can be performed in a solid or liquid phase, depending on whether the probes are fixed on a solid support or in solution (with biotinylated probes and streptavidin-coated magnetic beads) respectively. Different capture kits can be preferred because they vary in efficiency and specificity (Parla *et al.,* 2011). Through the target enrichment, random fragments corresponding to coding exons/adjacent intronic regions are thus selected, whereas the unbound DNA fraction is removed.

3. The third step is the sequencing of the exome library by the NGS technology. Amongst the most commonly used (in this project as well) platforms there is the Illumina

(Solexa) based on bridge PCR. The technique uses universal primers attached to a solid substrate through a flexible linker, and these primers are recognized by universal adaptors of the exome library. Once single DNA fragments are immobilised on the substrate, they also anneal to another adjacent primer, thus creating a bridge structure (Figure 1.1). A single fragment is amplified in about 1000 clonal amplicons, which are linearized and sequenced to obtain one sequence read. Each added nucleotide is fluorescently labeled and has a 3'-OH reversible terminator that blocks the DNA synthesis, to enable fluorescence detection. The terminator is then removed and the next nucleotide can be incorporated into a new cycle. Reads generated by the sequencer have approximately the same length (50-100 bp) and start from one extremity of the library fragment (single-end read) or from both extremities (pair-end reads).

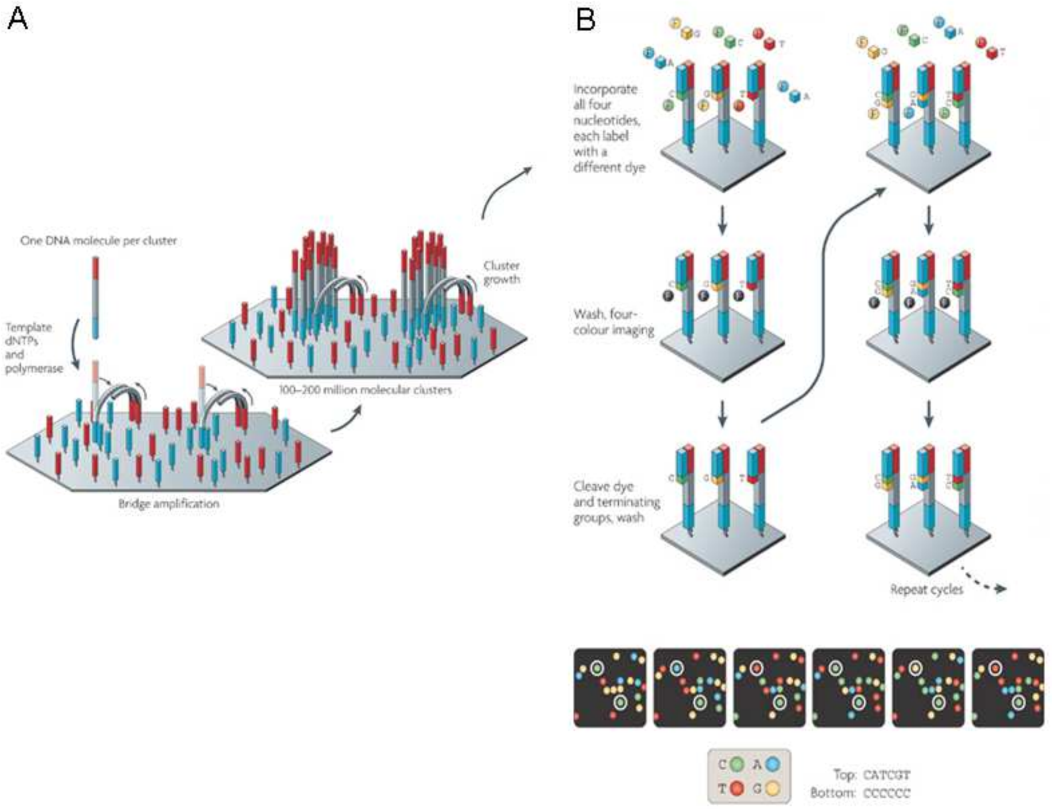

A
One DNA molecule per cluster
Template dNTPs and polymerase
Cluster growth
100–200 million molecular clusters
Bridge amplification
B
Incorporate all four nucleotides, each label with a different dye
Wash, four-colour imaging
Cleave dye and terminating groups, wash
Repeat cycles
C
A
T
G
Top: CATCGT
Bottom: CCCCCC

Figure 1.1: Illumina NGS technology based on solid-phase bridge PCR. (A) Library fragments bind complementary primers on the array surface and are amplified in clusters. (B) Clusters are sequenced by PCR with four-colour cyclic reversible termination method (Metzker 2010, with permission).

The following phase is characterised by a computational pipeline, in order to analyse sequencing data and manage all detected variants. First the reads are aligned and mapped against the human genomic reference sequence, a consensus sequence created by the Human Genome Project. All sequence differences between reads and reference are annotated as variants, by means of “variants calling” process. Variants are filtered for quality criteria and annotated by using several information from databases. The last and the most challenging step is the biological interpretation and management of the variation data, with the aim of identifying the disease-causing mutations.

1.1.6 Successful applications of the Whole-Exome Sequencing

The first publication of the WES as a means for identifying disease genes dates at 2009 (Ng *et al.,* 2009). Up to now, WES has led to the identification of over 100 new genes in Mendelian diseases, and there is a increasing number of studies which employes it. Although publication bias makes it arduous to estimate the actual success rate of the WES, the major disease gene seems identified in at least 60% of projects for Mendelian diseases, mostly for recessive traits (Gilissen *et al.,* 2012).

The study of familial cases has proven that the WES is successful for extremely rare Mendelian disorders. For a monogenic inherited disorder, multiple family members can be sequenced to search variants shared by affected and absent in healthy individuals. Affected members that are preferred for a resolving WES study are the most distantly related, in order to reduce the amount of shared “benign” variations. Even non-affected individuals can be useful to exclude private benign variations. To cite an example, through the WES of four family members affected by dominant spinocerebellar ataxia, a mutation in the *TGM6* gene was identified (Wang *et al.,* 2010).

Even the exome sequencing of unrelated individuals proved to be successful for rare Mendelian disorders with a little number of available cases. In 2009, WES study of four unrelated individuals affected by Freeman-Sheldon syndrome demonstrated its potential by finding *MYH3*, the unique gene with good candidate variants in all subjects (Ng *et al.,* 2009). In 2010, a study identified for the first time the genetic cause of a rare recessive disorder, Miller’s syndrome; the sequencing of four affected individuals belonging to three independent families led to discover the *DHODH* causative gene (Ng *et al.,* 2010). Other studies identified the most likely candidate disease gene by the sequencing of a single individual, but in other unrelated patients the same gene was excluded. Two examples are *WDR35* for Sensenbrenner syndrome (Gilissen *et al.,* 2010) and *HSD17B4* for Perrault syndrome (Pierce *et al.,* 2010). Exome sequencing of a considerable number of index cases can be even effective for phenotypically heterogeneous disorders. A considerable number of samples can improve the analysis since the pathogenic mutation could be missed or not shared by all patients. WES of ten sporadic cases allowed to identify *MLL2* as the gene underlying the Kabuki syndrome, an extremely rare and heterogeneous dominant Mendelian disease. At the beginning, this study failed to identify a common genetic cause; only by considering the sample heterogeneity and performing a less stringent filtering, *MLL2* was successfully identified in seven individuals. Sanger sequencing of *MLL2* in the remaining subjects did identify mutations in two out of three unsolved cases (Ng *et al.,* 2010).

WES approach even enables to overcome difficulties in studying sporadic disorders with low reproductive fitness, which are potentially caused by *de novo* mutations. For instance, *de novo* variants in *SETBP1,* the causal gene of Schinzel-Gedion syndrome, would not otherwise be identified without the WES, in this case by analysing four unrelated individuals (Hoischen *et al.,* 2010). Another good strategy for identifying *de novo* mutations is the WES of the case-parents trios, and the following exclusion of all inherited variants. A recent study involving 175 autism trios put in evidence risk alleles in *KATNAL2* and *CHD8* genes (Neale *et al.,* 2012).

Furthermore, the WES sequencing allows to successfully solve other issues. Noteworthy sequencing studies have unmasked misdiagnosis cases or enlarged phenotypic spectrum, by revealing mutations in known disease genes which have not been associated with the phenotype diagnosed before. For instance, Choi and colleagues identified mutations in *SLC26A3,* a gene coding for an epithelial exchanger, in a case with misdiagnosis of Bartter renal syndrome (Choi *et al.,* 2009).

More complex clinical pictures, characterized by the combination of more than one Mendelian phenotype and explained by the co-occurrence of mutations in multiple genes, can be unravelled by WES approach. For example in a study of two siblings, mutations in *DHODH* and *DNAH5* were identified as the explanation of combined phenotype with Miller syndrome and primary ciliary dyskinesia (Ng *et al.,* 2010).

Other applications of WES approach are the study of common diseases and somatic mutations in tumours, and the support to diagnosis of disorders with high genetic heterogeneity.

1.1.7 High variability of the human genome

With the advent of the NGS, studies of large cohorts became more feasible and the human genetic variability more characterised. Several projects collecting variants have highlighted higher genome variability than previously reported, with nearly half of the genes that are polymorphic in a genome (Levy *et al.,* 2007). This plethora of variants discloses the complexity of the genetic diversity of humans that has been furthered by the recent rapid expansion of populations. Considering such a background of thousands of variants (as mentioned in 1.3, 3-4 millions in a genome and approximately 25,000 in an exome), pinpointing a single mutation becomes extremely difficult. For instance, in every exome approximately 13,000 nonsynonymous variants (nsSNVs) could exert biological effects, there are more than 100 disease-causing or predisposing variants and 1000 novel variants, of which understanding the meaning still remains a challenging topic (Table 1.2). Multiple bioinformatics tools for predicting the functional effect tried to improve the analysis, but often they do not offer concordant results and still thousands of nsSNVs remain potentially “damaging” with this approach (Tennessen *et al.*, 2012).

However, the deep variant characterization which has been performing by recent projects will be useful for filtering analyses, interpreting WES data and better identifying the pathogenic effect of low-frequency variants. Indeed, current projects are demonstrating that rare variants tend to be recent events that display population-specificity (Abecasis *et al.*, 2012). This finding implies that for investigating variant frequency, a control cohort belonging to the same geographically context is preferred in order to have the same genetic background. Highly polymorphic genes (i.e. olfactory and taste receptor gene families) and a pathway specificity for low-frequency variants load have been even detected by these studies. This information will be definitely useful for improving the variant filtering in a specific disease and to decipher the biological and clinical significance of a variant (Moore *et al.*, 2011).

Total variants	4×10^6
SNVs	3.5×10^6
CNVs	10^4-10^5
nsSNVs	10,000-13,000
nsSNV potentially damaging	100s-1,000s
Loss of Function variants	120
Associated with inherited diseases	50-100
Stop codon	25-35
De novo	30

Table 1.1: DNA sequence variants in the human genome. SNVs: Single Nucleotide Variants. CNVs: Copy Number Variants. nsSNV: non synonymous single nucleotide variants (modified by Marian 2012, with permission).

1.1.8 Combinatorial approach of linkage analysis and WES

One of the main problems of WES is the manipulation of a large amount of data. This can be partially overcome in family studies, and specifically in the last years the integrated approach of WES with linkage analysis or homozygosity mapping has facilitated the identification of many disease genes. The traditional mapping approaches have the potential to narrow down first the number of variants identified by WES, second the number of individuals to be sequenced. In addition, the knowledge of candidate regions can be advantageous to apply a target sequencing approach specific for these *loci*. For instance, the combinatorial strategy was useful for identifying the disease gene *IFRD1* in autosomal-dominant sensory/motor neuropathy with ataxia, *PIGV* as underlying recessive Hyperphosphatasia-Mental retardation Syndrome and the causal *VCP* gene for dominant amyotrophic lateral sclerosis (Brkanac *et al.*, 2009;Johnson *et al.*, 2010;Krawitz *et al.*, 2010). In cases of consanguineous families with a recessive mode of inheritance, a homozygosity mapping approach identifies *loci* where searching causal variants amongst WES data. For example, in a study where the *FADD* gene was associated with autoimmune lymphoproliferative syndrome, this strategy allowed to shortlist 23,146 variations to only 81 ones (Bolze *et al.*, 2010).

Considering these advantages, over the last years the interest in linkage analyses is undergoing a renaissance due to NGS, which effectively scan nucleotide-by-nucleotide the even extended linkage regions.

1.2

1.2.1 Vulnerability of the neuron

Ubiquitously expressed proteins often play crucial roles in specific cell types or tissues. This is particularly frequent for proteins that are necessary to maintain the proper functionality of one of the most sensitive tissues, the nervous system. Peculiar features of neurons predispose to this characteristic sensitivity. First, cell size and axonal length (1 meter axon is over 10,000 times its cell body dimension and over 100 times its volume) require an efficient intracellular transport system to connect synapse and dendrites to cell body (Coleman 2013). For this reason cytoskeleton and motor proteins are critical for a proper retrograde and anterograde trafficking of molecules, membrane and organelles. Molecules involved in this process use just one dimension, while most cells can use all three dimensions. Second, branching morphology of the neuron makes cell membrane particularly exposed and susceptible to mechanical damage and other exogenous stresses. Thus, an efficient stress-response repertoire with chaperone and other proteins is vital for the neuron. Third, neuron is reliant on high energy demands and high metabolic rate, which are supplied by a suitable mitochondrial activity. For this reason, mitochondria need to maintain functional dynamics and protein quality controls. Fourth, the neuronal peculiarity of having synapse requires specialised processes in that districts, like calcium buffering and signalling pathways triggered by specific ion channels and receptors.

An impairment of one of these physiological cellular processes and the presence of stress factors including oxidative stress and misfolded proteins can lead to neuronal damages. In addition, recent evidences demonstrate that also neighbourhood cells like microglia and astrocytes may contribute to this condition (Brambilla *et al.,* 2013;Di Malta *et al.,* 2012). As neurons are not promptly regenerated, over time they can undergo a slow process of neurodegeneration, which starts long before visible outcomes. An example is the “dying back” process, a progressive distal to proximal length-dependent axonal degeneration that precedes cell death. Only when many cells die or undergo a functional decline and a part of the nervous system becomes damaged, clinical manifestations come out. Specifically, Shaw and colleagues tried to explained different neurodegenerative disorders with a selective neuronal vulnerability and the consequent neurodegeneration of definite regions, however to date the bases of this neuronal specificity remain elusive (Shaw 2005). In addition, a stressors-threshold model have recently attempted to explain this selective neuronal vulnerability in different genetic disorders. Specific genetic causes appear to enhance the sensitivity to stressors of specific subpopulations of neurons, characterized by specific cellular proteins, energy and organelle homeostasis processes. Moreover, the accumulation of specific combinations of intrinsic and environment-induced stressors can play an important role in disease aetiology and progression of neurodegeneration (Saxena and Caroni 2011).

1.2.2 Complexities of the inherited neurodegenerative disorders

Human neurodegenerative disorders are among the most heterogeneous diseases currently known in terms of clinical and genetic features, and for this reason among the most difficult disorders to dissect. Variable clinical manifestations, uncertainty of the diagnosis, ambiguity in the phenotype and incomplete penetrance represent examples of clinical heterogeneity that can hamper the study of these disorders. In some of these cases the initial diagnosis need to be re-evaluated after the identification of the genetic cause or the clinical spectrum become wider than previously appreciated. Even in cases with unambiguous diagnosis and clear mode of inheritance, there could be a complicated connection between phenotypes and their corresponding genetic changes.

Genetic heterogeneity characterizes several Mendelian disorders. Hereditary Motor Sensory Neuropathies (HMSNs) and Hereditary Spastic Paraplegias (HSPs) are exemplifying disorders with high genetic heterogeneity: more than 50 genes for HMSNs and 30 for HSPs are reported as causal to date. Moreover, specific mutations in the same gene can underlie diverse phenotypes. A noteworthy example is the *L1CAM* gene responsible for agenesis of the corpus callosum, CRASH syndrome, hydrocephalus and MASA syndrome or SPG1 (Jouet *et al.,* 1994;Menkes *et al.,* 1964). Different mutations can even underlie different modes of segregation, such as in *MPZ* and *PMP22* genes which is causal in both autosomal dominant and recessive diseases (Baets *et al.,* 2011).

A detailed description of clinical and genetic features of the diseases studied in this project is reported in the following chapters.

1.2.3 Hereditary Motor Sensory Neuropathies and their heterogeneity

Hereditary Motor Sensory Neuropathy (HMSN), also known as Charcot-Marie-Tooth disease (MIM118300), is a remarkable phenotypically and genetically heterogeneous group of diseases which affect the peripheral motor and sensory nervous system (PNS). HMSN is considered the most common inherited neuromuscolar disorder, given the prevalence of about 1 in 2500 individuals (Szigeti and Lupski 2009). The common clinical hallmarks in HMSN variable phenotypic picture include slow and progressive muscular weakness of distal limbs and peroneal atrophy, commonly associated with sensory loss, reduced tendon reflexes and skeletal deformities (*pes cavus*, hammer toes and scoliosis). The age of onset, disease course and severity are all variable features, but in general symptoms are displayed in the first/second decades of life, with a slowly progressive course. Severity is mainly displayed in recessive forms. Moreover, variable expression and oligosymptomatic patients were described even in the same families. Unfortunately at present, rehabilitation therapy and surgical procedures are the only available treatments for HMSN (Pareyson and Marchesi 2009).

Peripheral neuropathies have been traditionally classified according with neurophysiology and clinical data. A median nerve motor conduction velocity (MCV) below 38 m/s is associated with demyelinating forms (dominant CMT1 and recessive CMT4) and prominent neuropathologic features are myelin abnormalities. A MCV above 38 m/s instead indicates axonal forms (CMT2) with evidence of chronic axonal degeneration and regeneration. Several exceptions to this division and mixed features have been increasingly reported. A MCV between 25 and 45 m/s characterises dominant-intermediate (DI-CMT) and X-linked forms (CMTX). Other CMT-related neuropathies with a normal MCV are distal hereditary motor neuropathies (dHMN) and hereditary sensory and autonomic neuropathies (HSAN), depending on the affected motor or sensory nerves respectively. More recently, a growing number of complex CMT forms involving other tissues such as the central nervous system, muscle, bone and skin have emerged. For instance, the *FBLN5* gene identified for CMT was previously implicated in cutis laxa and macular degeneration (Auer-Grumbach *et al.*, 2011) and the *TRPV4* gene was already associated with 'neuro-skeletal' phenotypes (Chen *et al.*, 2010). Moreover, cases with phenotypes complicated by pyramidal involvement have been described in HMSN type V and optic atrophy in HMSN type VI (Barisic *et al.*, 2008;Pareyson *et al.*, 2006).

At the present time, CMT is still classified into subtypes based on the pattern of inheritance and the electrophysiological features (Vallat *et al.*, 2013). However over the last years, a classification based on genetics was performed and highlighted a growing genetic heterogeneity associated with this disease. The number of causal genes has increased rapidly since the first discovery of the most frequent cause (accounts for 40-50% of all CMT cases) in 1991, the duplication of the *PMP22* gene associated with CMT1A (Lupski *et al.*, 1991). Up to now more than 50 disease related genes have been identified through classical linkage studies followed by Sanger sequencing and recently by NGS approach. These genes are inherited through all possible Mendelian transmission patterns. The more common is autosomal-dominant way of inheritance, followed by X-linked transmission; autosomal-recessive is generally rare, except in contexts with a high rate of consanguineous marriages. The analysis of HMSN genes and their encoded proteins sheds light on the different molecular mechanisms underlying the disease and which convey in a common degenerative process of the peripheral neuron. Defects in myelin maintainment, axonal transport and cytoskeletal apparatus, membrane trafficking, lysosomal degradation, phosphoinositide metabolism, mitochondrial fission and fusion, protein folding and gene transcription were reported (Hui 2012). Besides this genetic heterogeneity, genetic studies even highlighted an overlap of genetic causes between different mode of transmission (both dominant and recessive mutations in *NEFL*, *EGR2*, *MFN2* and *GDAP1*), between different CMT forms (i.e. *MPZ*, *NEFL* for both CMT1 and 2; *GDAP1* for CMT2 and 4), between CMT2 and dHMN (the *HSPB1*, *HSPB8*, *BSCL2* and *GARS* genes), and between CMT and other motor neuron diseases such as axonal CMT2 and spastic paraplegia (HMSN type V caused by *BSCL2*, *GJB1* and *MFN2* mutations) (Baets and Timmerman 2011;Reilly and Shy 2009).

1.2.4 Distal Hereditary Motor Neuropathies: a subgroup of HMSN

Distal Hereditary Motor Neuropathies (dHMN) are characterized by a predominant motor involvement and minor or no sensory loss, with a very rare prevalence (no data available) (Siskind *et al.*, 2013). Moreover, dHMN accounts for about 10% of all cases of spinal muscular atrophy (SMA) and specifically distal forms (DSMA). The main clinical features are usually slowly progressive length-dependent weakness and wasting with hyporreflexia. The onset generally occurs in the first two decades of life. Neurophysiology studies show increased motor amplitude and duration potentials suggesting a chronic distal denervation, and this enables to differentiate dHMN from CMT2 and distal myopathy.

The first classification of dHMN was based upon clinical phenotype and mode of inheritance (Harding 1993). Harding proposed a division in seven categories (types I-VII), with types I, II, V and VII autosomal dominant forms and types III, IV and VI autosomal recessive forms. Type I and II are typical dHMN, III and IV are chronic forms of dHMN differentiated by the diaphragmatic palsy in type IV, type V is characterised by upper-limb onset, type VI is instead characterised by distal weakness and respiratory failure, and type VII is mainly defined by vocal-cord paralysis (Table 1.2). Considering all dHMN and/or DSMA annotated in OMIM database, up to now 17 disease related genes and 4 genetic *loci* with unidentified genes have been identified (Table 1.2). The analysis of encoded proteins sheds light on the molecular mechanisms underlying the disease. Defects in protein folding, neuroprotective signalling, DNA/RNA processing and metabolism, axonal transport and cation-channel dysfunction were described as pathogenetic mechanisms. These diverse functions suggest a multifactorial process or numerous possible ways causing the primary damage in the cell body of the ventral horn cell (differently from the axonal degeneration of HMSN). Since the primary pathologic process occurs in the cell body, neuropathies are even referred to as “neuronopathies”.

Molecular genetic studies have highlighted a high phenotypic heterogeneity of dHMN. Phenotypic variability is both intrafamilial as well as interfamilial, even for the same mutation. This suggests additional modifying factors, environmental or genetic, influencing the phenotype. An overlap of genetic causes between different modes of transmission and different dHMN types has been indeed reported. Moreover, some of these mutated genes are common between dHMN and CMT2 (*HSPB1*, *HSPB8*, *BSCL2*, *GARS*, *TRPV4*), also for the same mutations. For instance, p.S135F *HSPB1* mutation results in both dHMN2 and CMT2, p.K141N *HSPB8* mutation causes both dHMN-II and CMT2L in unrelated families and *GARS* mutations cause both dHMN-V and CMT2 (Antonellis *et al.,* 2003;Houlden *et al.,* 2008;Irobi *et al.,* 2004). Clinical variability of dHMN also extends up to upper motor neuron involvement, specifically in dHMN complicated by pyramidal signs. This form has been associated with mutations in *BSCL2*, *SETX*, *HSPB1* and three as-yet unidentified genes in *loci* 9p21.1-p12 (HMN-Jerash), 7q34-q36 and 4q34.3-q35.2 (Rossor *et al.,* 2012). It is interesting to note that *BSCL2* p.S90L mutation causes the more severe SPG17 phenotype, with upper motor neuron signs and lower limbs spasticity in addition to the muscle atrophy of upper limbs (Dierick *et al.,* 2008). Moreover, a digenic inheritance of the neighboring *BSCL2* p.N88S mutation with a second chr16p *locus* was reported in a dHMN5 family with variable presentation and additional pyramidal features (Brusse *et al.,* 2009). In addition, dHMN genetic causes can be found in other motor syndromes including juvenile amyotrophic lateral sclerosis (ALS), Kennedy disease, myopathy and spastic paraplegia (Table 1.2).

Despite all these advances in genes discovery, the causal gene has not been identified in more than 80% of dHMN patients yet (Rossor *et al.,* 2012).

Locus	Disease	Inheritance	Overlapping Diseases	Gene	Protein	Clinical phenotype
7q34-q36	dHMN I	AD		-		Juvenile onset, lower limb predominant distal weakness and wasting
12q24.23	dHMN IIA	AD	CMT2L	*HSPB8*	heat-shock 22-kD protein-8	Adult onset, lower limb predominant distal weakness and wasting
7q11.23	dHMN IIB	AD/AR	CMT2F	*HSPB1*	heat-shock 27-kD protein-1	
5q11.2	dHMN IIC	AD		*HSPB3*	heat-shock 27-kD protein-3	
11q13	dHMN III DSMA3	AR		-		Adult onset, distal weakness and wasting
11q13	dHMN IV DSMA3	AR		-		Juvenile onset, severe muscle weakness and wasting and paralysis of diaphragm
7p14.3	dHMN VA DSMA5A	AD	CMT2D	*GARS*	glycyl tRNA synthetase	Upper limb predominant distal muscle weakness and wasting
11q12.3	dHMN VA DSMA5	AD	Silver syndrome, SPG17	*BSCL2*	seipin	
2p11.2	dHMN VB DSMA5B	AD	SPG31	*REEP1*	receptor expression enhancing protein 1	onset in the first or second decade, distal muscle weakness and atrophy primarily of hand muscles
11q13.3	dHMN VI DSMA1 (SMARD1)	AR		*IGHMBP2*	immunoglobulin mu binding protein 2	Infantile onset, severe form with respiratory distress
2q12.3	dHMN VIIA	AD	Harper-Young Myopathy	*SLC5A7*	solute carrier family 5 member 7	Adult onset, with vocal cord paralysis
2p13.1	dHMN VIIB	AD	ALS1, Perry syndrome, dSBMA	*DCTN1*	dynactin1	Adult onset, breathing difficulties due to vocal cord paralysis, facial weakness and hand muscle atrophy
9q34.13	dHMN with pyramidal features	AD	ALS4	*SETX*	senataxin	Juvenile onset, distal weakness and wasting with pyramidal tract signs
4q34.3-q35.2	dHMN with pyramidal features	AD	CMT5	-		Adult onset, lower limb distal weakness and wasting with pyramidal tract signs
Xq12	DSMA X1	AR	Kennedy disease	*AR*	androgen receptor	Variable onset, spinal and bulbar muscular atrophy, facial fasciculations
Xp11.23	DSMA X2	X-linked		*UBE1*	ubiquitin activating enzyme 1	Infantile onset, severe hypotonia, areflexia, and multiple congenital contractures
Xq21.1	DSMA X3	X-linked	Menkes disease, Occipital horn syndrome	*ATP7A*	transmembrane copper-transporting P-type ATPase	Juvenile onset, distal weakness and wasting
2q35	DSMA5	AR		*DNAJB2 (HSJ1)*	heat-shock 40-kD protein DNAJ-like 1	Young adult onset, slowly progressive distal muscle weakness and atrophy
1p35	DSMA6	AR	Recessive intermediate CMTC	*PLEKHG5*	pleckstrin homology domain-containing protein, family G member 5	Childhood onset muscle weakness and severely decreased respiratory function
9p21.1-p12	dHMN-J DSMA2	AR		-		Juvenile onset, distal weakness and wasting with pyramidal tract signs
12q24.11	Congenital distal SMA	AD	CMT2C, Brachyolmia type 3, Scapuloperoneal spinal muscular atrophy	*TRPV4*	transient receptor potential cation channel subfamily V member 4	Nonprogressive lower limb weakness and paralysis with contractures
14q32.31	SMA lower extremity-predominant	AD	CMT2O, Kugelberg-Welander syndrome	*DYNC1H1*	cytoplasmic dynein 1 heavy chain 1	Juvenile delayed motor development and lower limb weakness

Table 1.2: Classification for dHMN, with causal genes and characteristic signs, based on OMIM annotation (http://www.omim.org/). AD = autosomal dominant, AR = autosomal recessive.

1.2.5 Co-occurring clinical features of axonal HMSN and HSP

In axonal HMSN the most frequent eight genes account for only 25% of cases (Reilly and Shy 2009). In addition, several complicated forms are observed for axonal CMT. By the traditional classification which divides HMSN from HSP, complex HMSN with pyramidal signs and HSP complicated by peripheral neuropathy are separately reported even if they have overlapping phenotypes. This dichotomous classification is probably due to the observed variable phenotypic spectrum which can ranges from HMSN to HSP even in the same family. Indeed, patients with both syndromes, others with a prevalence of peripheral neuropathy or pyramidal signs have been described. Genetic studies have increasingly confirmed that variable phenotypes are due to the same underlying genetic basis. Interesting examples are *BSCL2* mutations associated with Silver syndrome and HSP phenotype; *KIF5A* determining peripheral neuropathy and pure HSP; *MFN2* responsible for CMT2 and spastic paraplegia; the *atlastin-1* (*SPG3A*), *REEP1* (*SPG33*), *NIPA1* (*SPG6*) and *spastin* (*SPG4*) genes commonly associated with HSP and also extended to peripheral neuropathy (Auer-Grumbach *et al.,* 2005;Goizet *et al.,* 2009;Hewamadduma *et al.,* 2009;Ivanova *et al.,* 2007;Liu *et al.,* 2008;Schulte *et al.,* 2003;Zuchner and Vance 2006).

HMSN type V was the first complicated form of HMSN described in 1968 (Dyck and Lambert 1968). It is a rare autosomal dominant form of axonal CMT (CMT2), characterised by normal or slightly reduced nerve conduction velocity and lower limb atrophy, with additional pyramidal signs typical of spastic paraplegia. Pyramidal tract involvement is mainly displayed by spastic gait and variable tendon hyperreflexia manifestations. The onset usually occurs in the second decade of life or later and the course is slowly progressive. From a genetic point of view, no data are available so far about the genetic mapping of this form. Previous studies excluded already known genes and *loci*, but the rarity of the disorder and the limited number of studies hind further genetic investigations (Mostacciuolo *et al.,* 2000;Vucic *et al.,* 2003). Defining HMSN type V as a complicated form of CMT2 or HSP, or a stand-alone clinical entity is still matter of debate.

1.2.6 Overlapping molecular mechanisms between HMSN and HSP

HMSN and HSP primarily affect specific target tissues, that are peripheral nerves (PNS) and corticospinal tracts/dorsal column (CNS) respectively. However both HMSN and HSP are considered progressive neurodegenerative disorders, with a length-dependent affection and a possible late onset. Besides the above-mentioned common genes, many other genes implicated solely in HMSN or HSP are instead involved in similar cellular processes. Examples of genes belonging to the same gene family, encoding proteins involved in the same biochemical pathway or directly interactors have been reported (Figure 1.2). Common cellular processes are membrane traffic, mitochondrial function, myelination, axonal transport and cytoskeletal organization (Timmerman *et al.,* 2013).

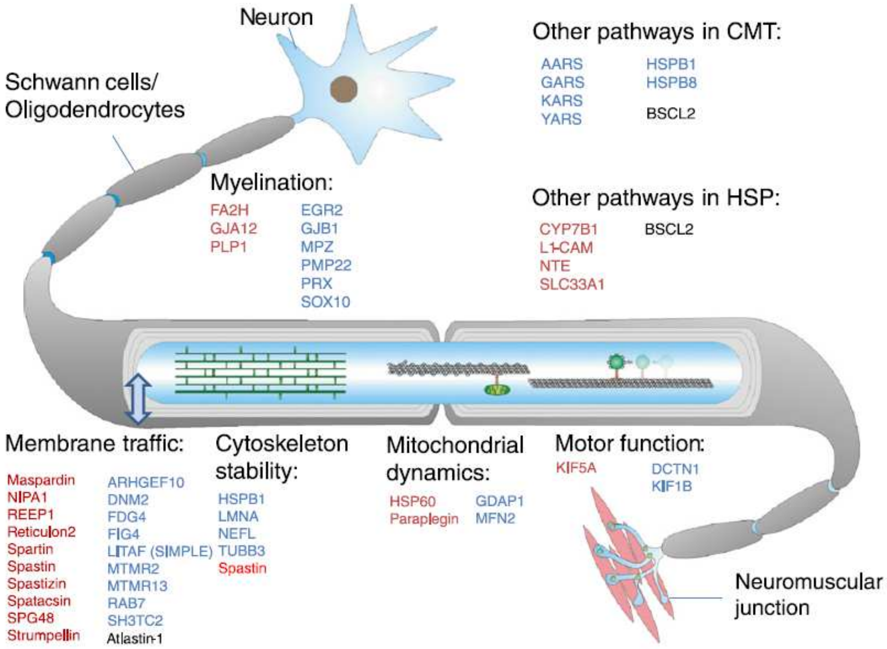


Figure 1.2: Similar and different cellular processes involved in the pathogenesis of HMSN and HSP. Genes associated with HMSN are indicated in blue and with HSP in red. Common genes are in black (Timmerman *et al.*, 2013, with permission).

Membrane traffic

The complexity of intracellular membrane structures is fundamental to define different compartments in neuron, its polarity, branching and long axons. Furthermore an efficient transport from cell body to synapses and neurites is necessary for their maintainment. Vesicles trafficking and membrane shaping processes can be impaired in HMSN and HSP. In HMSN mutated *RAB7A* and *DNM2* genes code for RAB GTPases that play a key role in the vesicle formation and transport. Mutated SH3TC2 and SIMPLE proteins interact with other proteins to recycle endosomes and for degradation *via* lysosomal pathway respectively. Mutated *MTMR2*, *MTMR13* and *FIG4* genes encode Phosphatidyl-Inositol-phosphatases located in the endosomes of Schwann cells. Furthermore, mutated frabin (*FGD4*) and ARHGEF10 are GTP/GDP exchange factors of Rho GTPases implicated in the myelination and cytoskeleton maintainment.

In the same way, HSP genes code for transmembrane proteins involved in membrane trafficking. Reticulon2, spastin, atlastin1 and REEP1 are all involved in the endoplasmic reticulum (ER) shaping; strumpellin acts to shape endosome, spartin for lipid droplets and lysosome shaping.

Mitochondrial dynamics

Dynamics of fusion/fission of mitochondria determine their morphology and size, but also regulate their distribution and function. Mitochondrial distribution and transport along the long axons play a primary role in the energy transport up to distal parts. Neurons are particularly vulnerable to mitochondrial dysfunctions because they require a high metabolic rate even in distal terminations and ATP is produced mainly by the mitochondrion. In addition, mitochondria are fundamental for the neuron survival because they synthesise key metabolites, buffer calcium and protect against oxidative stresses. In HMSN, mutated MFN2 and GDAP1 proteins impair mitochondria fusion/fission leading to defects in the transport and energy production. In HSP, paraplegin AAA ATPase participates in protein quality control of misfolded proteins, similarly to the mitochondrial chaperone HSP60.

Myelination

Myelin sheaths around the axons allow their insulation and the proper action potential conduction. Schwann cells in the PNS and oligodendrocytes in the CNS are responsible for a correct myelin production. For instance, causal genes for HMSN code for PMP22 and P0 myelin proteins of Schwann cells. Moreover, connexin32 of gap junctions and periaxin for the Schwann cell-axon contacts can be also mutated. Mutations in EGR2 and SOX10 transcription factors regulating the expression of these four proteins have been also identified in HMSN. Similarly in HSP, mutated PLP1 and FA2H proteins play a role in the CNS myelination, and even the intercellular communication can be involved in the pathogenesis of HSP disorders with mutations in the connexin 47 gap junction protein.

Cytoskeleton stability and motor proteins

The neuronal intracellular transport is essential to supply the axon with newly synthesized molecules and for anterograde transport of neurotrophic factors and damaged organelles. The neuronal transport requires functional motor proteins and cytoskeletal network. For instance, NEFL protein that plays a fundamental role in the intermediate neurofilaments assembly and transport, and tubulinβIII for microtubules stabilization are mutated in HMSN. Other mutations in dynactin-1, interactor of dynein for the retrograde axonal transport, have been associated with HMSN. Likewise in HSP, spastin protein catalyses the microtubule severing and KIF5A kinesin has essential roles in the anterograde axonal transport.

2. AIM OF THE RESEARCH

This study aims to identify and, whenever possible, functionally characterise genes causing highly heterogeneous forms of hereditary peripheral neuropathy. With the purpose of shedding light on the genetics of these diseases, four families affected by complex forms of hereditary peripheral neuropathy and showing no mutations in already known disease-genes have been examined. Three of these families display an autosomal recessive mode of inheritance and one family presents an autosomal dominant transmission.

The identification of novel genes is performed by using a comprehensive strategy of traditional approaches coupled with the next-generation approach of whole-exome sequencing. Linkage analysis, homozygosity mapping and identity-by-descent analysis enable to detect candidate regions in which the disease-associated gene potentially segregates. The whole-exome sequencing allows to inspect the variants in the coding exons of the candidate regions. Considering the advantage of having variants within the whole exome, the known disease-related genes are also analysed and less stringent genome-wide searches are even conducted. The best candidate variants are selected by filtering and prioritization analyses which allow to evaluate the possible pathogenicity of the variants. For the potential mutations, further confirmations are obtained by means of *in silico* analyses, the study of allelic frequencies in the specific population, the genetic screening of unrelated patients and functional studies.

3. RESULTS

3.1 General approach

With the aim to identify candidate genes associated with the disease, the study of 4 independent families started from a genome-wide search performed with SNPs and confirmed by STR markers. For all families the most frequently mutated genes causing neuropathy and spastic paraplegia have been screened by Sanger sequencing in previous studies, without finding the genetic cause amongst the genes identified before 2009.

In the current project, the candidate linkage regions found by traditional genetic mapping strategies were better refined and less stringent analyses were performed to overcome technical and computational limits of the technique. For the families with an uncertain segregation, different models were taken into account. Whether a low number of genes mapped to the linkage regions, the candidate positional cloning by Sanger sequencing was the approach of choice. The study of Family 1 is an example of application of this traditional method.

In three other families (Family 2, 3 and 4) large linkage regions with many genes were identified and the whole-exome sequencing (WES) approach was used. The first sequencing dated 2011, a more recent sequencing dates December 2013. The exome analysis was performed by the Beijing Genomics Institute (BGI) which used the *Agilent SureSelect Human All Exon v4* or *v4+UTRs* kits for the targeted enrichment and the Illumina (Solexa) Hiseq2000 platform for the sequencing. End-paired reads of about 90 bp were obtained and aligned against the human reference sequence (*hg18*) (NCBI build 36.3 assembly), in order to detect the mismatches and call the variants. The analysis of nearly 100,000 high-confidence variants identified by WES focused on the known disease-genes and on genes located in the linkage regions. In parallel, the depth of coverage was analysed and the poorly-covered exons were sequenced to identify undetected variants. Variants were critically evaluated, filtered and prioritized using information annotated in databases and in other in-house exomes. A manageable pool of good candidate variants with a putative pathogenicity was obtained and validated by Sanger sequencing. The approach to identify a recessive trait was exemplified by the studies of Families 2 and 3, whereas the study of a dominant trait was exemplified by Family 4.

3.2 Identification and characterisation of a novel mutation in the *SACS* gene

3.2.1 Clinical picture of Family 1

This consanguineous three-generation family comes from a little village of Northern Italy and presents three brothers (II-2, II-3 and II-4) affected by a severe form of peripheral neuropathy and spastic paraplegia. All family members were seen as adults by Dr. Pegoraro of the Department of Neurosciences, University of Padova, except the affected brother (II-2) who died before the study for lung cancer, but was reported to have the same gait disturbance. The age of onset was about 20 years. The patients presented mild hand muscle atrophy and distal hypotrophy of the lower limbs, with bilateral Babinski sign and foot deformities including bilateral *pes cavus* and hammer toes. Progressive lower limb spasticity with foot drop, incontinence and hyperreflexia were also displayed. Electrophysiological studies and nerve biopsy revealed the presence of a severe peripheral sensorimotor neuropathy with mixed axonal and demyelinating findings. The muscle biopsy showed a normal mitochondrial histochemistry and fibre type grouping, suggesting a re-innervation process. Brain magnetic resonance imaging (MRI) showed non-significant evidence of cerebellar and spinal cord atrophy. Moreover, ocular fundus appearance was normal without retinal nerve hypermyelination.

The analysis of the pedigree suggested an autosomal recessive way of inheritance of the disease. Affected sib's healthy parents were second cousins and the recurrence of the clinical phenotype was observed only in the second generation, without finding any manifestation in other generations.

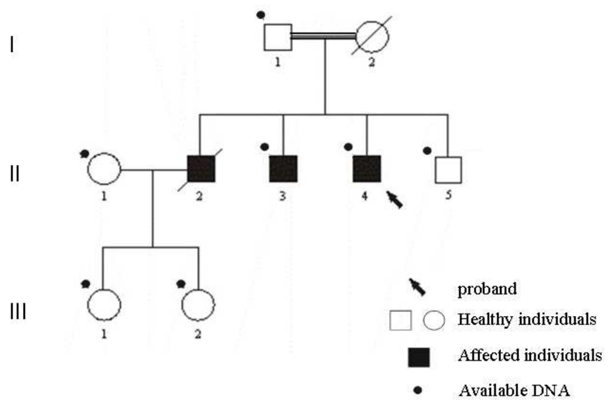


Figure 3.1: Pedigree of family 1.

3.2.2 Preliminary results

The first studies on this family excluded mutations on the disease-genes most frequently associated with similar clinical pictures including *PMP22*, *SPAST*, *ATL1*, *SPG7* and *KIAA1840.* Considering the presence of a severe peripheral neuropathy, the *MFN2* gene, responsible for the most common form (20-30%) of axonal CMT type 2 (CMT2A), was also analysed (Ortega-Roldan *et al.*, 2013). Interestingly, a heterozygous missense substitution (c.749G>A, p.Arg250Gln) was identified in the proband (II-4) and in two asymptomatic subjects (I-2 and III-2) but not in the affected brother (II-3). Despite the specific mutation was already identified in literature (McCorquodale *et al.*, 2011;Zuchner *et al.*, 2004), c.749G>A substitution did not co-segregate with the disease and it was unlikely to be the primary cause of the clinical phenotype in this family. In addition, this variant has been recently reported in EVS database at the heterozygous state in 4 out of 6503 human genotypes and in dbSNP database as polymorphism (rs140234726). Accordingly, its pathogenic role in CMT2A needs further confirmation.

The presence of a novel disease gene was investigated by a whole-genome scan. The analysis was first performed by using 368 STR markers and subsequently by genotyping over 200,000 SNPs (Illumina Human CNV370-Quad platform). Copy number variant analysis did not identify any candidate chromosomal deletion/duplication (>10 kb) and all genes and *loci* already associated with this phenotype were excluded by linkage analysis. However a maximum linkage signal on chromosome 13q12.1-q12.12 was highlighted. The LOD score value (2.58) was not significant but the only one confirmed in the analyses with different SNP sets. It was close to the theoretical maximum value that this relatively small family could generate in the case of linkage.

3.2.3 Candidate gene screening

The haplotype reconstruction on chr13q12.1-q12.12 confirmed and better refined a large homozygous region of 1.6 Mb (312 consecutive homozygous SNPs) which co-segregated with the disease. The critical interval (rs12868337; rs7997447) contained only 5 genes: *SGCG* already associated with muscular dystrophy, *SACS* associated with the spastic ataxia of Charlevoix-Saguenay, non-coding RNA *LINC00327*, *TNFRSF19* involved in the embryonic development and *MIPEP* implicated in the protein maturation. The *SACS* gene was considered the best candidate as it is already associated with a neurodegenerative disease. Direct sequencing of the entire gene identified a missense substitution (c.11,104A>G), which leads to the p.Thr3702Ala amino acid change. This substitution co-segregated with the disease in the family: both available affected patients (II-3, II-4) were homozygous, whereas their parents (I-1, I-2), the unaffected brother (II-5) and the two unaffected nephews (III-1, III-2) were heterozygous. Furthermore, haplotype reconstruction allowed to infer the status of the third unavailable affected brother (deceased subject II-2) who could be homozygous for the mutation (Figure 3.2). The c.11,104A>G variant was not present in the NHLBI Exome Variant Database amongst 6503 exomes, nor in the 1000 Genomes data set, nor in dbSNP (build 137) neither in LOVD, HGMD, Uniprot and ClinVar databases which annotate disease mutations. In addition, by tri-primer ARMS-PCR (Allele Refractory Mutation System PCR) it was excluded from 700 control chromosomes from Northern Italy.

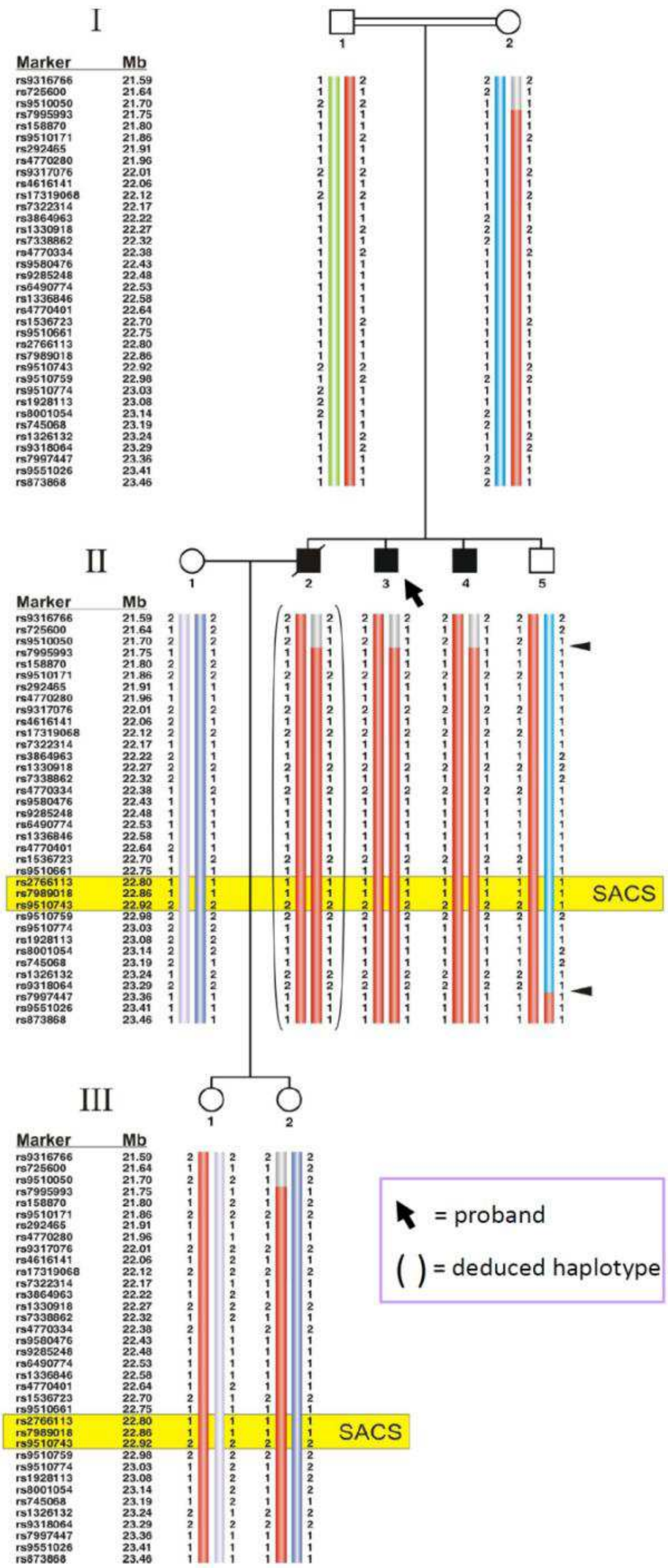


Figure 3.2: Haplotype on chr 13q12.1-q12.12. Chromosomes in red bear the 1.6 Mb region (limits pointed by black arrows) co-segregating with the disease, where the *SACS* gene carries the c.11,104A>G mutation.

3.2.4 *In silico* analyses

SACS (NM_014363.5) was considered the best candidate gene of the chr13q12.1-q12.12 linkage region as it has been already associated with a neurodegenerative disease (Engert *et al.,* 2000). To date the enormous size of sacsin encoded protein (NP_055178) (4,579 amino acids and with the largest exon amongst vertebrates of 12.8 Kb) has considerably hindered biochemical studies. For this reason, also if more than 100 *SACS* mutations have been identified in patients, very few studies have been published on sacsin function (Duquette *et al.*, 2013).

Considering the limits in the functional studies, the first investigation was performed *in silico*, in order to evaluate whether the c.11,104A>G variant could have a pathogenic significance. Sequence conservation was assessed by significant scores obtained by phyloP, GERP and phastCons tools, and its pathogenicity was predicted by PolyPhen-2, SIFT and MutationTaster prediction softwares.

Secondly, an assessment of whether the amino acid substitution could affect the stability of the protein structure was performed. The secondary structure of the oligopeptide where p.Thr3702Ala falls was predicted by Jpred3, Psipred and SOPMA softwares and highlighted a longer α-helix for the mutated sequence compared with the reference sequence. This result suggested that the change from a polar amino acid (threonine) to an apolar one (alanine) potentially destabilizes the sheet structure in favor of an α-helix structure (Figure 3.3).

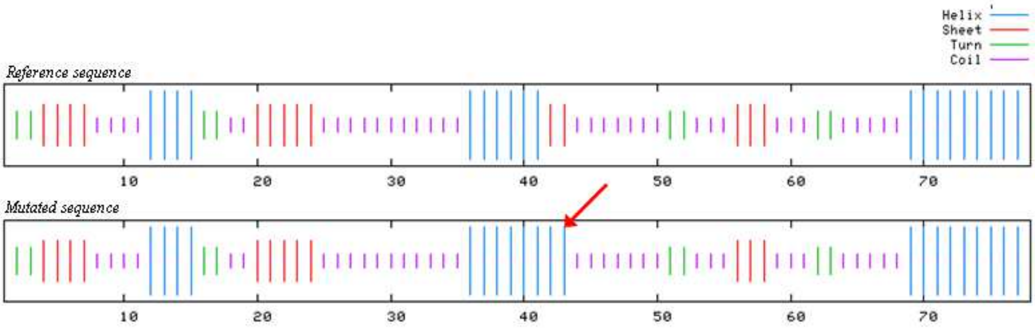
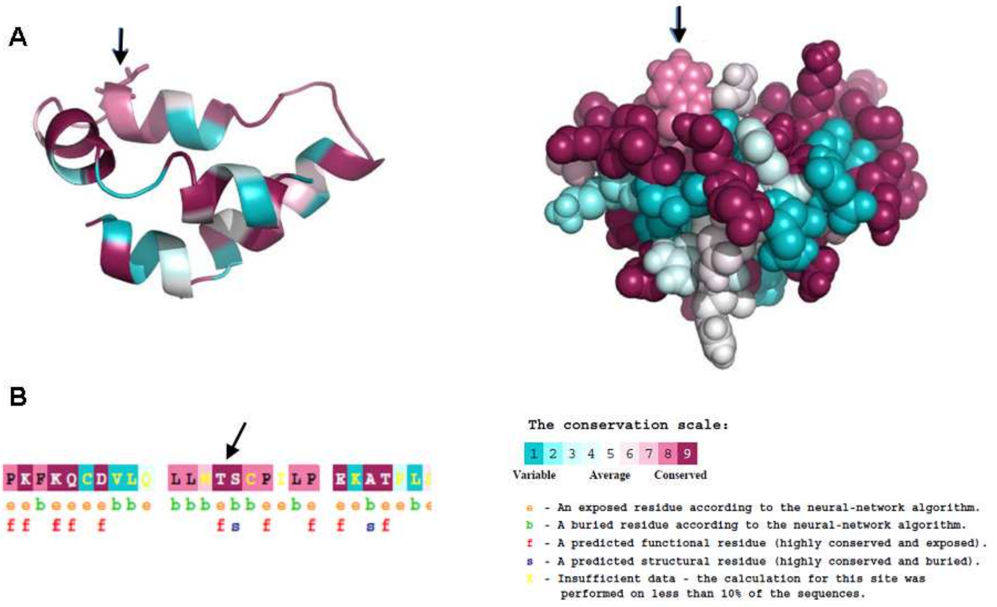


Figure 3.3: The domain of sacsin protein where p.Thr3702Ala falls. It displays a longer α-helix at 42 and 43 amino acid positions, in correspondence with the mutation (Prediction by SOPMA software).

The domain where p.Thr3702Ala falls is a 75 amino acid sequence (3660–3735) presenting 35% sequence identity with the NMR structure of Xeroderma pigmentosum C binding (XPCB) domain of the hHR23A protein (amino acids 231–285; PDB code 1TP4) (Kamionka and Feigon 2004). Binding experiments in mouse demonstrated that hHR23A requires the

XPCB domain to interact with UBE3A protein and to deliver ubiquitinated proteins to the proteasome (Greer *et al.*, 2010). The interaction with the E3 ubiquitin-ligase UBE3A was speculated also for the XPCB domain of sacsin, probably to be itself ubiquitinated and degraded *via* proteasome, or to degrade other misfolded proteins which sacsin chaperonin attempts to correctly fold. An inspection of the key amino acids that are potentially involved in the interaction with UBE3A was carried out by three-dimensional modeling. The homology modeling of the sacsin XPCB domain was performed by HOMER server and took advantage from the sequence similarity with hHR23A (Figure 3.4). The model showed the threonine residue at 3702 amino acid position exposed on the protein surface, thus revealing a possible protein-protein interaction site. Moreover, threonine and its surrounding residues are conserved amongst vertebrates (pink colour), suggesting a key role for the interaction with UBE3A. A confirmation was obtained by the prediction of ConSeq server which starts from the linear amino acid sequence and automatically collects the homologous sequences of the database. It predicted threonine position to be "functional" because highly conserved and exposed in the protein. In conclusion, *in silico* prediction of p.Thr3702Ala is in agreement with a potential functional effect of the mutation.


A
B
P K F K Q C D V L Q L L N T S C P I L P E K A T P L
e e b e e e e b b e b b b e b b e e b e e e b e e b e
f f f f f f s f f f s f
The conservation scale:
1 2 3 4 5 6 7 8 9
Variable
Average
Conserved
e - An exposed residue according to the neural-network algorithm.
b - A buried residue according to the neural-network algorithm.
f - A predicted functional residue (highly conserved and exposed).
s - A predicted structural residue (highly conserved and buried).
X - Insufficient data - the calculation for this site was performed on less than 10% of the sequences.

.Figure 3.4: (A) Structural model and surface representation of the tertiary structure of sacsin XPCB domain, visualized by PyMol (cartoon and surface representations). The threonine at position 3702 (indicated by the arrow) shows localization within an exposed region which is conserved during evolution across sacsin orthologues (pink colour) as shown by the ConSurf program. Prediction obtained in collaboration with Dr. Tosatto's laboratory. (B) Specific amino acid sequence where the change falls, with threonine position (indicated by the arrow) predicted to be functional by the algorithm of ConSeq server, because highly conserved and exposed.

3.2.5 Studies on fibroblasts of the proband

Early studies in literature revealed *SACS* transcript in fibroblasts, skeletal muscle and brain motor system at high levels, whereas the sacsin protein presented a cytoplasmic distribution with a partial mitochondrial localization (near to 30%) in different cell lines (Parfitt *et al.,* 2009). In 2012 Girard and colleagues demonstrated that sacsin is required for the mitochondrial functionality, morphology and localization in neurons. They explained these findings through an imbalance between mitochondrial fusion and fission, probably due to the disruption of Drp1 function. In addition to sacsin knockdown studies, an alteration of mitochondrial dimensions was even indicated in the fibroblasts of patients homozygous for a frame-shift *SACS* mutation (c.8844delT) (Girard *et al.,* 2012). Starting from this background of knowledge, the current study aimed to assess such mitochondrial impairment in the fibroblasts of the proband. The experiments were performed in the laboratory of Dr. Chapple (Barts and The London, School of Medicine and Dentistry, London), a collaborator in Girard's project. Before studying mitochondria, *SACS* mRNA and sacsin protein were quantified by quantitative PCR and Western blot assays respectively, in order to check whether the mutation could affect the expression. The comparison were performed between fibroblasts derived from the proband and healthy controls. No evident differences were found, suggesting a normal mRNA and protein expression in the patients homozygous for the c.11,104A>G variant (Figure 3.5). Moreover, full-lenght protein of 520 kDaltons was detected by Western blot, thus excluding a mutational effect in splicing processes and a truncated protein.

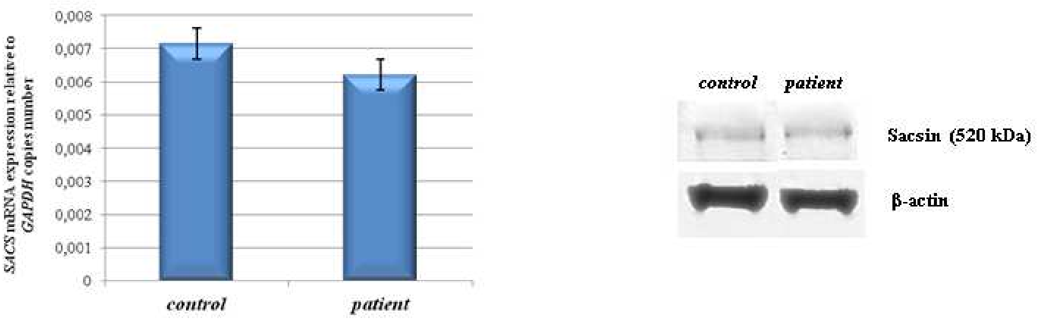


Figure 3.5: *SACS* expression assays. (A) real-time PCR analysis of *SACS* mRNA expression in fibroblasts; values were normalized to *GAPDH* gene expression. The data represent mean ± SEM (n = 3). (B) Sacsin protein expression detected by Western blot method.

After the confirmation of unaltered sacsin expression levels in patients, the mutational effect on mitochondria was investigated by evaluating mitochondria volumes and surface areas.

Immunostaining experiments with the antibody against TOM20 which labels mitochondria were carried out for 100 mutated and 100 control fibroblasts. Fluorescent z-stacks acquired by confocal microscope were volume rendered by Imaris software, in order to obtain 3D mitochondria as objects (Figure 3.6). First, the analysis highlighted an increased number of mitochondria in mutated fibroblasts. Indeed, 3849 mitochondria in proband's fibroblasts and 3482 in control fibroblasts were counted. On the other hand, the analysis of mitochondria volumes showed a mean value of 20.8 μm^3 for controls and 17.4 μm^3 for mutated fibroblasts, displaying a statistically significant decrease in presence of the p.Thr3702Ala mutation (p-value= 0.031). Similarly, the surface area was significant decreased in mutated fibroblasts (p-value= 0.036; mean value of 53 μm^2 for controls and 46.7 μm^2 for mutated fibroblasts). The opposite trend was showed by sphericity data, which reflect how well an object fits a sphere (maximum score of 1 indicates a sphere). The mean value of 0.813 for mutated fibroblasts is significantly higher than 0.8 of controls (p-value= 0.003), indicating an increased sphericity for the more fragmented mitochondria of patients, as expected. Obviously these three-dimensional morphometry measures require further supporting evidences, however these first findings are indicative of a higher fragmentation of mitochondria in presence of p.Thr3702Ala sacsin mutation, whereas in control fibroblasts the mitochondrial network is characterized by a typical interconnected tubular structure. These data are in agreement with the hypothesis of impairment in fusion/fission mitochondrial dynamics.

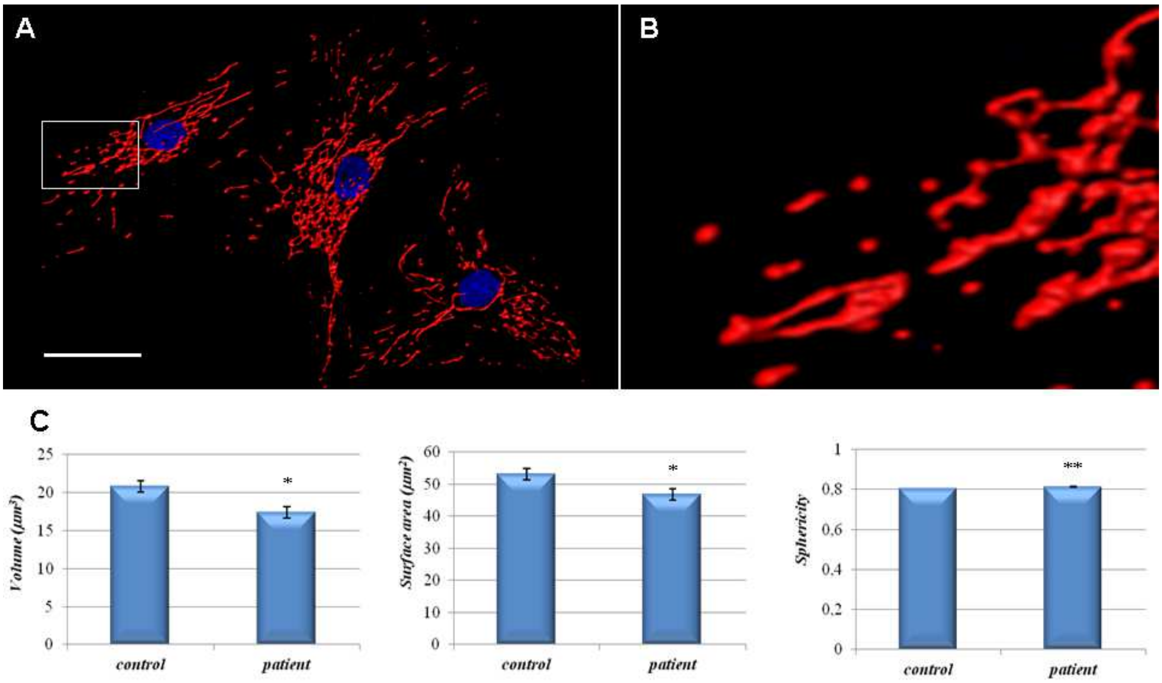


Figure 3.6: (A) An example of the three dimensional rendering of mitochondria in fibroblasts, labelled by TOM20 antibody (red). Nuclei are labelled by DAPI (blue). Scale bar, 20 μm. Box indicates the region enlarged at right (B).

(C) Histograms of volumes, surface area and sphericity of mitochondria. Non parametric Mann-Whitney test of medians was used for the analyses, since a non-gaussian distribution of data was confirmed by D'Agostino test (p-value<0.05). Error bars represent standard error; one asterisk indicates a p-value<0.05, two asterisks indicate a p-value<0.01.

Concluding, the findings obtained by the study of Family 1 seem consistent with the pathogenicity of the novel variant c.11,104A>G identified in the *SACS* gene. They suggest a role of the mutated sacsin protein in the pathogenic mechanisms underlying the clinical picture of peripheral neuropathy and spastic paraplegia.

3.3 Identification of *SIGMAR1* and *SGK223* as candidate genes associated with distal Hereditary Motor Neuropathy

3.3.1 Clinical picture of Family 2

This consanguineous family comes from a small and isolated village of Southern Italy, with a reported high rate of inbreeding. A detailed anamnesis of the whole family enabled to reconstruct a complex, seven-generation pedigree. This family presents four affected subjects, and amongst them the patient VII-1 refused the genetic investigation. Clinical data were collected by Dr. Cavallaro at the Hospital of Verona, and Dr. Tessarolo at S. Camillo-Forlanini Hospital of Rome, who diagnosed a distal Hereditary Motor Neuropathy (dHMN). Even if the clinical hallmarks of lower limb weakness and bilateral *pes cavus* seemed shared by all patients and suggested a dHMN clinical picture, a more detailed diagnosis highlighted differences between the affected subjects V-3, VI-6 (gray symbols in the pedigree of Figure 3.7) and patients VII-1, VII-2 (black symbols in the pedigree). Subjects V-3 and VI-6 displayed a late-onset peripheral neuropathy (in third and second decade respectively), with a clinical picture characterized by lower limb weakness and atrophy, denervation of muscles, reduced osteotendinous reflexes and mild disfagia. Electrophysiological studies and nerve biopsy confirmed the diagnosis of dHMN. On the other hand, two other patients (VII-1 and VII-2) showed an earlier onset (in the first decade of life) and more severe clinical features, with even proximal signs, atrophy of the upper limbs, absent osteotendinous reflexes and a reduced motor conduction velocity with proximal conduction blocks (not altered in patients V-3 and VI-6).

The analysis of the pedigree suggested an autosomal recessive mode of inheritance for the disease. Three consanguineous marriages (between IV-1 and V-3; IV-2 and IV-3; VI-1 and VI-2) with a clear kinship between spouses were showed, except for the union between individuals VI-3 and VI-4. Even if cases of incomplete penetrance have been already reported in literature for dHMN, an autosomal dominant transmission seemed less likely in this consanguineous family.

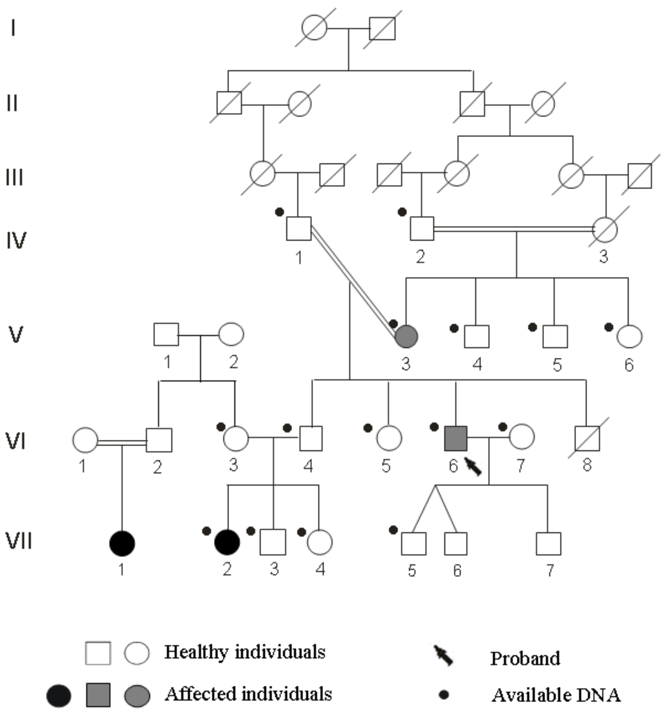


Figure 3.7: Pedigree of Family 2. Differently coloured symbols indicate distinguishable clinical pictures.

3.3.2 Preliminary results

Over the previous years, the most frequently mutated genes for dHMN forms and for phenotypically overlapping forms of SMA and HMSN (i.e. *HSPB1*, *HSPB8*, *GARS*, *BSCL2*, *IGHMBP2*, *SETX*, *DCTN1; SMN1* for SMA, *PMP22* and *MPZ* for HMSN) were excluded by direct sequencing of genomic DNA. First linkage studies by using STR markers excluded also the already known *loci* for recessive forms of dHMN (11q13 for dHMN III, 11q13.3 for dHMN VI and 9p21.1-p12 for HMN-J). The presence of a novel disease *locus* was hypothesized and it was investigated by a whole-genome scan. Homozygosity mapping identified a candidate 1.1 Mb region on chromosome 11p12 (N11S409754; N11S23943), the only candidate region shared by all three patients available (V-3, VI-6 and VII-2). In this region, only the 5'-UTR of the *LRRC4C* gene, encoding NGL-1 protein, maps. Direct sequencing did not identify any variant co-segregating with the disease and unmasked a false-positive autozygous region (identical-by-status) in VII-2. A high-density genome-wide linkage analysis was thus performed with over 160,000 SNP markers (Affymetrix Mendel Nsp 250K chip) distributed evenly across the genome. The multipoint linkage analysis was carried out by assuming an autosomal recessive model with 95% penetrance and disease allele frequency of 0.0001. Since the analysis excluded a common genetic cause amongst all three patients, the model with two distinct genetic causes was considered.

3.3.3 Analysis of two nuclei

The lack of shared candidate regions and the presence of two distinguishable clinical pictures led to the exclusion of a common genetic cause among all patients. Moreover, the family originates from a little village and the high inbreeding increases the probability that more disease alleles can segregate in the same family. For these reasons the whole family was splitted into two nuclei (Figure 3.8) by taking into account the patients with a more similar phenotype. The nucleus with V-3 and VI-6 patients was analysed separately from the nucleus with the two more severely affected subjects VII-1 and VII-2.

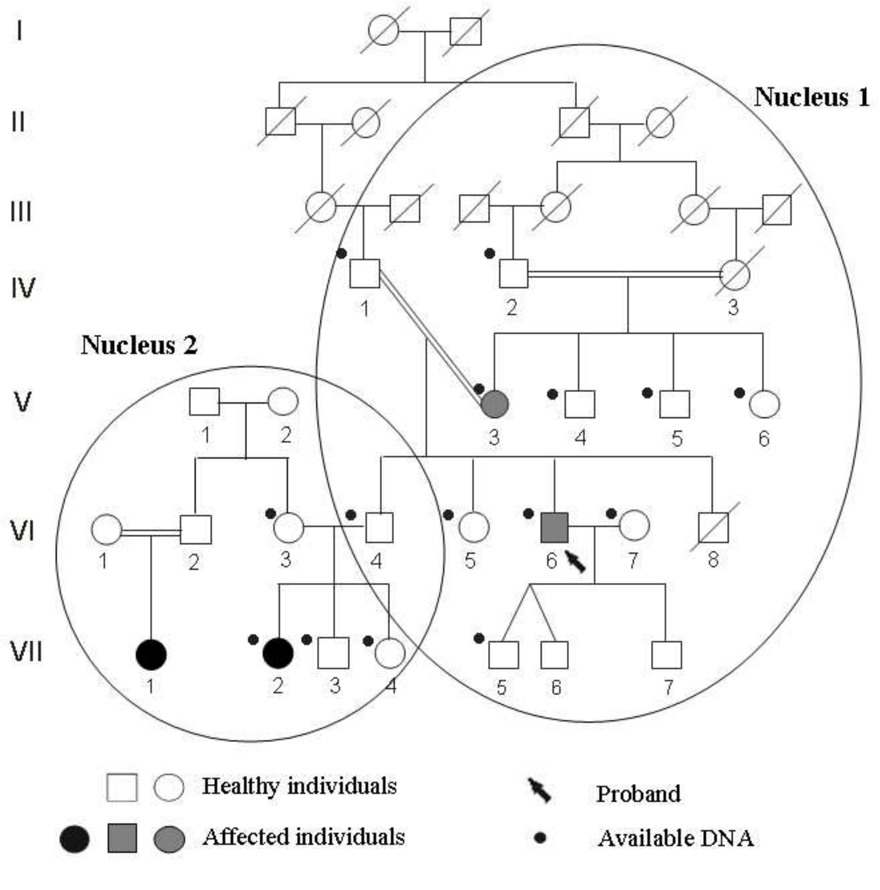


Figure 3.8: Pedigree of Family 2 was splitted into 2 nuclei on the basis of the phenotypes.

3.3.4 Study of nucleus 1

3.3.4.1 Homozygosity mapping

For identifying the disease-gene in this little consanguineous nucleus, the traditional homozygosity mapping approach was used to identify candidate regions co-segregating with the disease. Indeed, affected individuals are expected to share the mutation and the surrounding regions in an autozygous state because transmitted as identical by descent (IBD) by both consanguineous parents. The analysis of 160,000 SNP genotypes with HomozygosityMapper software highlighted 14 homozygous regions shared by the two affected subjects (V-3 and VI-6) with a homozygosity score higher than 0.8 (probability of homozygosity compared with control haplotypes of the CEPH collection of the HapMap project) (Figure 3.9). For each of these peaks, the haplotypes were evaluated in order to exclude regions shared by healthy subjects (V-4, V-5, V-6, VI-4 and VI-5). Through this analysis, two candidate regions were obtained: the interval on chromosome 11p12 previously identified and already excluded by direct sequencing and a novel candidate region on chromosome 8p23.1-p22 (Figure 3.9 and Table 3.1).

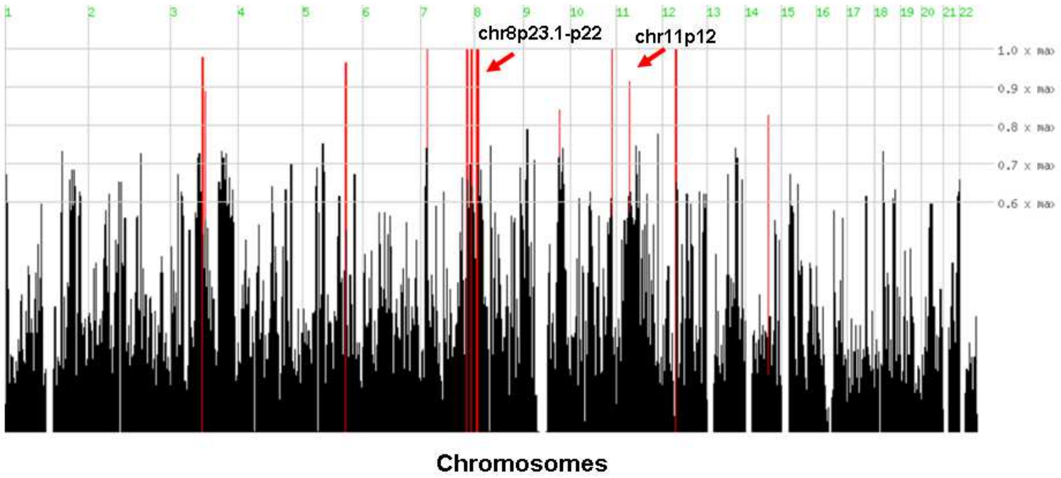


Figure 3.9: Graphical representation of the homozygosity regions obtained for the two patients of nucleus 1. Red bars indicate the most promising genomic regions above the threshold value (0.8). Candidate regions selected after the haplotype evaluations and segregation are indicated by the two arrows.

Chr	from (bp)	to (bp)	from SNP	to SNP	Mb	Total RefSeqs
8	6754919	12718090	rs2738148	rs6997599	5.96	120
11	39300178	40369557	rs10837138	rs11035793	1.06	1
				total	7.02	121

Table 3.1: Candidate regions selected after haplotype evaluations. For each region the chromosomal positions (hg19 release), flanking SNP markers, total validated Refseqs and OMIM genes mapping to the region were annotated.

Haplotype reconstruction of chr8p23.1-p22 region pointed out a telomeric recombination event occurred at SNP_A-2207444 in previous generations and transmitted to both patients (II-2 and III-4 in Figure 3.10) by their healthy fathers (I-1 and II-1 respectively in the haplotype). This crossing-over defines the telomeric limit at 17.16 cM corresponding to 6,734,409 genomic position. On the other hand, the recombination event occurred in patient II-2 at SNP_A-1911827 defines the centromeric limit at 26.05 cM (chr9: 12,964,191 genomic position). It is interesting to note that the patient III-4 displayed downstream a more extended homozygosity than his affected mother (II-2).

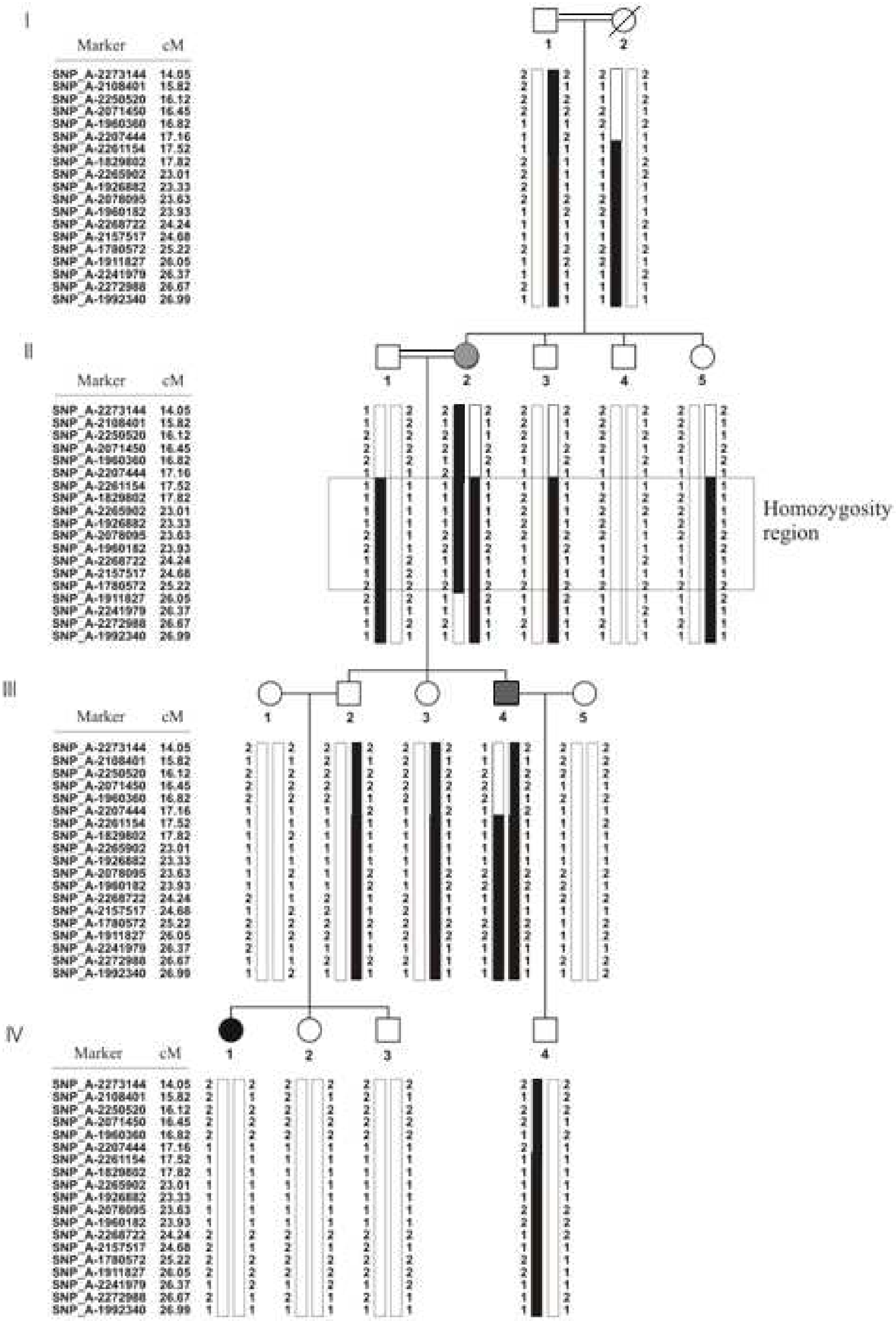


Figure 3.10: Haplotype reconstruction of chr8p23.1-p22 region in a subpedigree with the genotyped family members. Haplotypes associated with the disease are indicated in black. The autozygous region shared by the 2 patients of nucleus 1 (here II-2 and III-4 indicated by gray symbols) is indicated by the box.

A less stringent homozygosity mapping was performed for the 2 patients separately in order to overcome any limit due to genotyping errors and for identifying also very narrow regions. It must be considered that a far common ancestor can imply many possible recombination events in previous meioses and the presence of small shared regions. Moreover, given the complexity of this family, the alternative hypothesis of transmission with one patient in homozygous state and the other one compound heterozygous in the same gene was investigated by this analysis. It is interesting to note that two narrow regions were identified on chr7q33-q36.2 and chr12q12 and were shared by both patients (Table 10 in Appendix E). For a complete inspection, all shared/not shared candidate regions were considered during the subsequent analyses of the variants identified by the exome sequencing.

3.3.4.2 Whole-exome sequencing

Considering the large number of genes mapping to the candidate homozygosity regions (more than 120 genes for shared regions and more than 1000 genes for not shared regions), the identification of disease-causing mutations in the nucleus 1 was carried out by the whole-exome sequencing approach (WES). Exome capture and sequencing were performed by the BGI for the two affected individuals V-3 and VI-6 (more technical details in Materials and Methods).

3.3.4.3 Statistical evaluation of the performance for patients V-3 and VI-6

First the performance of the WES technology applied to these samples was evaluated. To address this, the critical steps of exome capture (here with *Agilent SureSelect Human All Exon v4* kit*)*, sequencing (of *Illumina-Solexa* NGS platform) and read alignment (by *SOAPaligner/SOAP2* program) were considered. The specificity of the technique is mainly due to the target (50 Mb) enrichment and the read alignment steps. The percentage of 90-bp reads that mapped to target regions was about 72% for both subjects, whereas the uniquely mapped reads were 81.5% for V-3 and 84.7% for VI-6 (more data in Appendix F). These relatively low percentages of good-quality data were expected for an exome sequencing performed in 2011 since the technique still required improvements.

The read depth was one of the most relevant parameters that were taken into consideration (number of reads mapping to a single genomic position). Since a high inter-exon and intra-exon variability was observed, the mean values of 77X and 91X read depth for V-3 and VI-6 individuals respectively should be regarded as indicative only. A more interesting parameter was the coverage of the target regions, that is the percentage of target regions covered by a certain read depth. 97% target regions was covered with at least 1X read depth in both samples, conversely a more desirable read depths above the threshold of 10X and 20X, which contributed to a higher confidence of the variant calls, displayed lower coverage values. With a threshold of 10X read depth, only 87% (V-3) and 88.5% (VI-6) of target regions were covered, whereas 79% and 81% respectively with a minimum of 20X read depth. The subject VI-6 obtained higher values of sequencing depth and cumulative read depth as showed in the graphs of Figures 3.11 and 3.12.

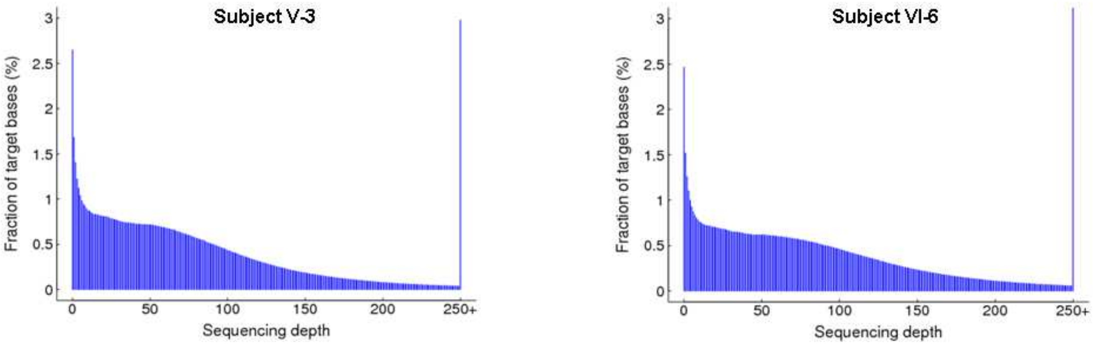


Figure 3.11: Histograms of distributions for sequencing depth in target regions.

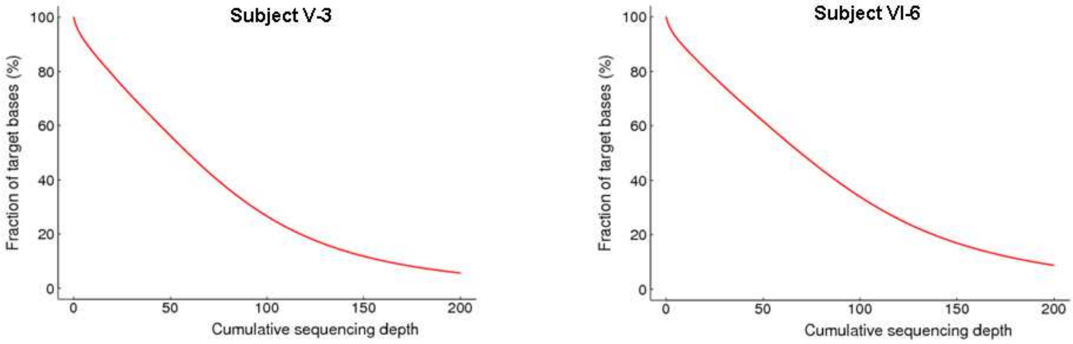


Figure 3.12: Cumulative distributions for sequencing depth in target regions.

3.3.4.4 Single Nucleotide Variants (SNVs) and indels annotation

The final set of high-confidence variants obtained by the exome sequencing and called by SOAPsnp program was annotated in the following tables (Table 3.2 and 3.3). A high confidence was established on the basis of the minimal distance from another variant to avoid error-prone nucleotide stretches (5 bp), a minimum 4X of read depth and a Phred-like quality score higher than 20.

Categories of SNVs	Patient V-3	Patient VI-6
Number of genomic positions for calling SNVs(1)	134752579	135216361
Total number of high-confidence SNVs	84213	90776
Homozygous	30959	32804
Heterozygous	53254	57972
Intron	50682	55690
Intergenic	1347	1584
5’-UTR	3169	3374
3’-UTR	4043	4493
Synonymous	10148	10359
Splice site(2)	2472	2539
Missense	12122	12510
Stop codon abolishing	54	59
Nonsense	176	168

Table 3.2: High-quality SNVs annotated in the *snp.filter.gff* file. (1) Capture target regions and its 200bp flanking regions; (2) Intronic SNPs within 10bp of exon/intron boundary.

Categories of indels	Patient V-3	Patient VI-6
Total number	6780	7168
Total insertions	3215	3434
Total deletions	3565	3734
Homozygous	2761	2845
Heterozygous	4019	4323
Intron	5102	5416
Intergenic	61	66
Promoter (1000bp upstream)	22	32
5’-UTR	278	305
3’-UTR	415	445
Non-frameshift	175	181
Splice site	404	395
Insertions in coding sequence	242	250
Deletions in coding sequence	256	259
Frameshift	323	328

Table 3.3: Indels annotated in the *indel.gff* file.

The higher number of variants in the subject VI-6 compared with V-3 reflects the higher coverage seen beforehand. It is interesting to note that the number of total high-confidence SNVs is higher compared with other data in literature (Abecasis *et al.*, 2012). More than 80,000 SNVs are reported here, whereas an exome of about 30 million nucleotides commonly contains 25,000 high-quality SNVs. The reason is the inclusion of all the SNVs in the 200bp-flanking regions in this analysis. Without intron and intergenic variants, the total number of coding/splice site variants detected here becomes similar to other studies. For the current study, the analysis started from the entire pool of nearly 100,000 total SNVs and indels annotated in *.snp.filter and *.indel.gff files, focusing on the most probably pathogenic indels and non-synonymous SNVs.

3.3.4.5 Variants in disease genes and loci and CNV analysis

All known disease-genes and loci implicated in dHMN, as well as all genes annotated in OMIM database and associated with similar phenotypes were analysed for WES variants. Variants in coding exons and close to splice sites were firstly considered. Their potential pathogenicity was evaluated by taking advantage of the available information on variants already annotated in databases and on known or predicted functional domains. No homozygous variant calls shared by both patients were identified. A further investigation for compound heterozygous or hemizygous traits was performed and highlighted 4 variants in the *PLEKHG4* and *ABCD1* genes (Table 3.4).

Chr	Candidate Variant	Variant call V-3	Variant call VI-6	Predicted Effect	Allele frequency	Haplotype	Functional Role	Confirmation by Sanger sequencing
16	*PLEKHG4* c.1658C>G	C(ref)13/G7 (Q=32)	C(ref)10/G3 (Q=0)	Missense tolerated	Novel variant	Combination of variants excluded from unaffected subjects	*PLEKHG4* causes SCA31 (AD) Rho GTPases pathway for cytoskeleton dynamics	False Positive call
16	*PLEKHG4* c.1986G>T	G(ref)5/T0 (Q=35)	G(ref)6/T3 (Q=32)	missense damaging	Novel variant			False Positive call
16	*PLEKHG4* c.3190C>T	C(ref)32/T12 (Q=99)	C(ref)22/T24 (Q=99)	Missense damaging				-
X	*ABCD1* c.1823G>A	G(ref)8/A4 (Q=63)	G(ref)3/A5 (Q=32)	missense damaging	Rs, no in EVS not validated, unknown frequency	Not informative (few SNPs)	Cause of Adrenomyeloneuropathy, adult (X-linked,Recessive)	False Positive call

Table 3.4: Filtered SNVs found in genes annotated in OMIM and related to neurodegenerative disorders. Variant calls are expressed with number of reads for each allele (ref=reference allele) and Q=Phred-like quality score of variant call. Homozygous and compound heterozygous calls were considered. AD= autosomal dominant, AR= autosomal recessive, SCA= spinocerebellar ataxia, DSMA= distal spinal muscular atrophy.

3 out of 4 variants were not confirmed by Sanger sequencing, highlighting the presence of false-positive calls even after the filtering for quality. Heterozygous genotypes seemed to be more error-prone than homozygous ones as the background noise can be misinterpreted as an effective heterozygous allele, mainly for calls with a low read depth. The remaining fourth

variant c.3190C>T in *PLEKHG4*, heterozygous in both patients, also segregated in 2 healthy subjects. Therefore, all variants mapped to disease and OMIM genes, and loci implicated in dHMN were excluded or could not explain alone the disease.

The whole SNV dataset from the exome sequencing was also analyzed by the BGI using Copy Number Inference From Exome Reads (CoNIFER) program. A total of 11 copy-number variants (CNVs) were identified (>1.5 kb), but none of them was shared by the 2 affected subjects (Table 17 in Appendix G).

3.3.4.6 Variants in the candidate region chr8p23.1-p22

The exclusion of the known disease-related *loci* suggested the presence of a novel disease gene. For this reason the analysis of WES variants was performed for the candidate homozygous regions detected before. The candidate region on chromosome **8p23.1-p22**, which was shared by the two patients and contained 120 genes, was characterized by a high number of pseudogenes, paralogs and genes belonging to gene families. For instance, the *DEF* genes encode defensin proteins belonging to the family of microbicidal and cytotoxic peptides for host defense. However, they seemed unlikely to be the primary cause of the neurological clinical phenotype in this family. Unfortunately other genes with a more interesting function such as *USP17L* (ubiquitin specific peptidase) and *ZNF205* (zinc finger protein) showed many paralogs and pseudogenes. As a consequence, several reads mapping to multiple regions led to a high number of false positive calls in these genes. Parameters of read aligners, even if restricted and with maximum 3 tolerated mismatches, cannot prevent this problem. In order to address this issue, a multiple sequence alignment of the *USP17L* and *ZNF205* genes, with their corresponding pseudogenes and paralogs were performed, enabling to identify and exclude all variants mapping to them.

The analysis of the remaining variants in the chr8p23.1-p22 region started from a manageable pool of 221 SNVs and 4 indels for patient V-3, and from 232 SNVs and 5 indels for individual VI-6. Amongst these variants, a total of 78 and 81 for V-3 and VI-6 respectively were prioritized because in coding exons and close to splice sites. After the filtering process only one variant in the *SGK223* gene (NM_001080826, chr8: 8234390 in hg19 release) was identified as putative disease-causing. Since it was present only in one of the two patients, a further inspection of *.snp file containing all raw calls was carried out. Surprisingly, the analysis highlighted the same variant also in the second patient (Table 3.5).

Chr	Candidate Variant	Variant call V-3	Variant call VI-6	Predicted Effect	Allele frequency	Haplotype	Functional Role	Confirmation by Sanger sequencing
8	*SGK223* c.1529T>C	A(ref)0/G5 (Q=6)	A(ref)0/G10 (Q=26)	Missense tolerated	Novel variant	Excluded from unaffected subjects	Rho GTPases pathway	yes

Table 3.5: Filtered SNV found in genes of the 8p23.1-p22 homozygous candidate region. Variant calls are expressed with the number of reads for each allele (ref=reference allele) and Q=Phred-like quality score of the variant call.

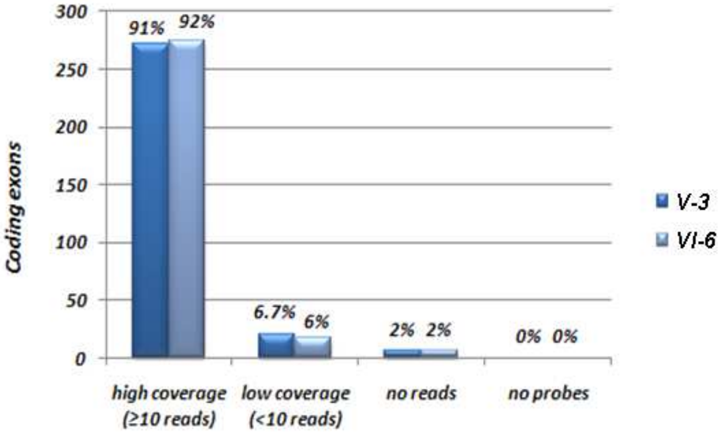
It is interesting to note that this was due to the low quality score (Q = 6) of the call in the patient V-3, which led to its exclusion from *.snp.filter file. The c.1529T>C substitution has never been identified before in variant databases and Sanger sequencing confirmed the homozygous status in both patients.

Homozygosity regions separately detected for the two patients were even investigated by the same analysis, and the two found candidate variants were excluded (Table 11 in Appendix E). Specifically, these genes displayed interesting functional roles and the variants were annotated in the variant databases with low allele frequencies. It seemed reasonable to take into account also rare polymorphisms given that the variant databases could collect healthy subjects that should display a late-onset phenotype in the future. However, these two variants could explain the trait of only one patients and were not homozygous or compound heterozygous in the second one.

Therefore, this analysis highlighted the *SGK223* c.1529T>C homozygous substitution shared by both patients as the only good candidate that alone can explain the disease in nucleus 1. This finding was further enforced by the exclusion of all high-quality variants within the whole exome. Indeed, for this purpose all variants of the final *.xls file were even analysed.

3.3.4.7 Coverage analysis for the chr8p23.1-p22 candidate region

In order to enforce the hypothesis of causality of the *SGK223* c.1529T>C substitution, the exclusion of all other possible variants in chr8p23.1-p22 candidate region was carried out. Indeed, false negative calls are heavily influenced by a poor sequence coverage. The read depth within the region was analyzed through the visualization of the read alignment by NCBI Genome Browser and MapView tool. On the basis of many observations, a read depth under 10X was arbitrarily evaluated insufficient to accurately call homozygous or two compound heterozygous mutations, or to pinpoint a false positive call (Figure 3.13 and Table 3.6).



	V-3	VI-6
Total coding exons	298	298
High coverage depth (≥10 reads)	272	275
Low coverage depth (<10 reads)	20	18
No reads	6	6
No probes	0	0

Figure 3.13: Depth of coverage per exon, on chr8p23.1-p22 region. Y-axis indicates the number of coding exons; the correspondent percentage is indicated above each histogram bar. Poorly-covered coding exons were annotated for each patient. Table 3.6: Exons for each category of coverage depth on chr8p23.1-p22 region.

This analysis allowed to highlight problematic regions that were not captured by probes or not well-sequenced. The first finding was a comparable read depth for both subjects with slightly higher values in VI-6, turning out in agreement with the above-mentioned statistical WES data. However, the coverage over 10X read depth was improved for this specific region (91% and 92% for V-3 and VI-6 respectively) compared with WES data (mean value 88%). The capture kit appeared adequately designed for all coding exons, with a capture efficiency that showed little variability between samples. Interestingly, only a small portion of the coding exons characterized by a GC content >60% or <40% was more difficult to capture.

On the basis of the gene function, the poorly covered exons were prioritized and selected for direct sequencing in order to identify potentially not-detected variations. Table 3.7 reports the analyzed exons together with the new variants identified.

Gene and Exon	Gene function	Expression (Gene Cards)	Variants	Comments
*SGK223*_exon2	Tyrosine-protein kinase, cytoskeleton regulation	High levels in nervous system	rs150979349 c.1050_1051insAGCGGC unknown frequency	even in related and independent healthy individuals
*SGK223*_exon5	Tyrosine-protein kinase, cytoskeleton regulation	High levels in nervous system	-	-
*RP1L1*_exon4	microtubule polymerization	Also in brain	-	-
*ERI1*_exon1	mRNA Exonuclease1, histone mRNA decay after replication	Also in nervous system	rs2288672 (MAF>1%) rs28510464 (MAF>1%)	Only polymorphisms
*TNKS*_exon1	ADP-Ribose Polymerase, vesicle trafficking	Also in brain	-	-
*NDRG1*_exon16	cell trafficking, notably of the Schwann cell; causes CMT4D	Also in nervous system	-	-

Table 3.7: Poorly-covered exons selected for Sanger sequencing with the relative information and variants eventually detected.

Besides the two high-frequency polymorphisms in *ERI1* gene, the homozygous rs150979349 6-nucleotides insertion was identified in the same *SGK223* gene where the candidate variant was detected by the WES. The hypothesis of a pathogenic effect exerted by the co-occurrence of these two homozygous variants exclusive of the 2 patients was also supported by the presence of the same insertion with a second synonymous substitution in another dHMN patient (data of Genomes Management Application database, with data from patients with neuromuscular disorders and unidentified genetic cause). This insertion was annotated without allele frequency data and by several submissions, suggesting a certain frequency in the population. On the contrary, its mapping to different positions in a tandem 6-nucleotide repetition corroborated the hypothesis of a very rare frequency. The involvement of this second variant in the disease was subsequently excluded by a genotyping study of healthy controls, where a high frequency was found. Concluding, also this analysis excluded other candidate variants, thus enforcing the hypothesis of causality of the *SGK223* c.1529T>C missense substitution in the disease.

3.3.5 Study of the nucleus 2

In this second nucleus the DNA samples of the left branch of the pedigree were not available for the current study. However, the information about the consanguineous marriage between VI-1 and VI-2 subjects corroborates the hypothesis of a recessive mode of inheritance of the disease. Considering also that VI-2 and VI-3 are brothers and that they should carry the same disease allele, this data gave the idea to verify the consanguinity between VI-3 and VI-4 spouses that was only reported but not confirmed.

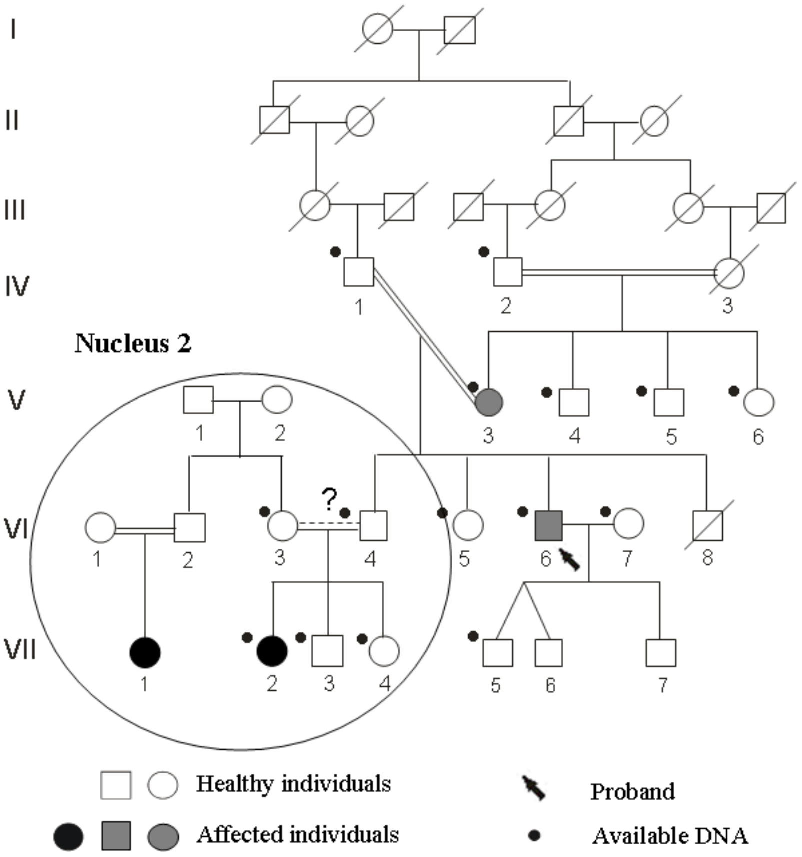


Figure 3.14: Pedigree of the nucleus 2 of Family 2.

3.3.5.1 Homozygosity mapping

As for the nucleus 1, the recessive trait was studied by using the genome-wide homozygosity mapping approach. The analysis of 160,000 SNP genotypes identified a high number of homozygous regions in patient VII-2 and not shared by the healthy brothers VII-3 and VII-4 (Figure 3.15). For each of these regions, haplotypes were compared with the healthy sibs and evaluated for possible genotype-calling errors. The 11 candidate regions identified are listed in Table 3.8.

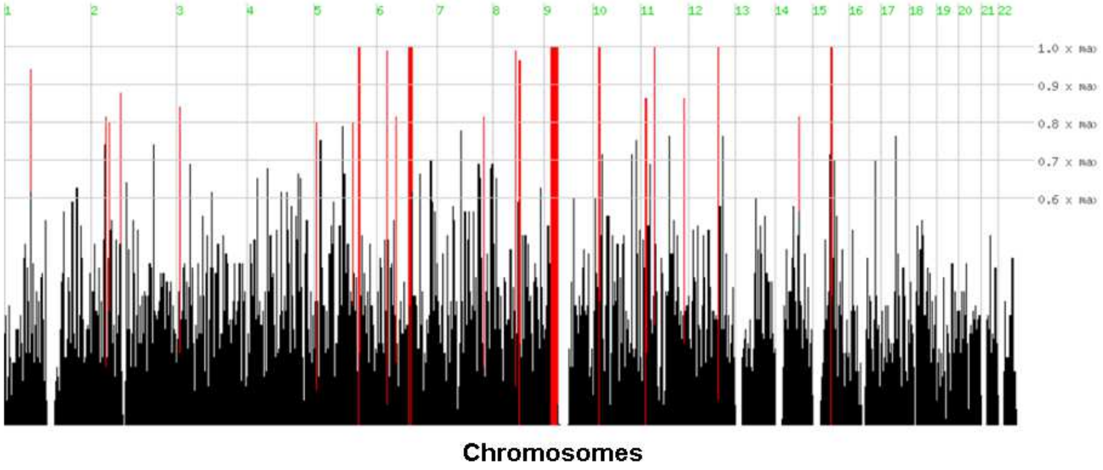


Figure 3.15: Graphical representation of homozygosity regions obtained for patient VII-2 of Nucleus 2. Red peaks indicate the most promising genomic regions.

Chr	from (bp)	to (bp)	from SNP	to SNP	Mb	Total RefSeqs	OMIM genes	OMIM genes and *loci* for recessive neurological disorders
1	75263485	76401812	rs1969111	rs1770887	1.14	6	1	-
2	82286435	83378207	rs17642475	rs1430271	1.09	1	1	-
3	8046509	8601764	rs3902639	rs7639343	0.55	2	-	-
6	53292013	53876547	rs4715392	rs1409101	0.58	3	1	-
6	90651785	92241291	rs210052	rs10498974	1.59	2	-	-
6	97234028	99159601	rs1854269	rs7776218	1.93	4	1	*NDUFAF4* for Mitochondrial complex I deficiency
8	63999561	65448437	rs6994076	rs12056691	1.45	4	-	-
8	76140789	77390676	rs1351225	rs6985886	1.25	2	-	-
9	30504349	38362713	rs17775810	rs1022770	7.86	95	19	- *APTX* for Ataxia; - *B4GALT1* for disorders of glycosylation; - *SIGMAR1* for Juvenile amyotrophic lateral sclerosis 16; - *SPG46* for Hereditary Spastic Paraplegia; - *GNE* for Inclusion Body Myopathy-2 and Nonaka Myopathy; - *EXOSC3* for Pontocerebellar hypoplasia 1B - DHMN-J locus for Jerash type dHMN
11	13886922	15220073	rs1827516	rs1540151	1.33	9	2	-
11	123880722	124345791	rs4936879	rs10431091	0.46	2	-	-
				total	19.03	130		

Table 3.8: Candidate regions identified after haplotype evaluations. For each region the chromosomal positions (hg19 release), flanking SNP markers, total validated Refseqs and OMIM genes mapping to the region were annotated.

3.3.5.2 Identical-By-Descent (IBD) analysis

The kinship between VI-3 and VI-4 spouses was assessed by Identity-By-Descent (IBD) analysis. This approach estimates the probability of sharing alleles identical by descent (IBD) between pairs of subjects. Starting from about 64,000 genotypes of SNPs that were in approximate linkage equilibrium with each other, only one region displayed an IBD1 probability close to the maximum of 1 within the whole genome (Figure 3.16). This chromosomal segment maps on chr9p21.1-p13.2 and it is likely to be inherited by VI-3 and VI-4 from a distant common ancestor.

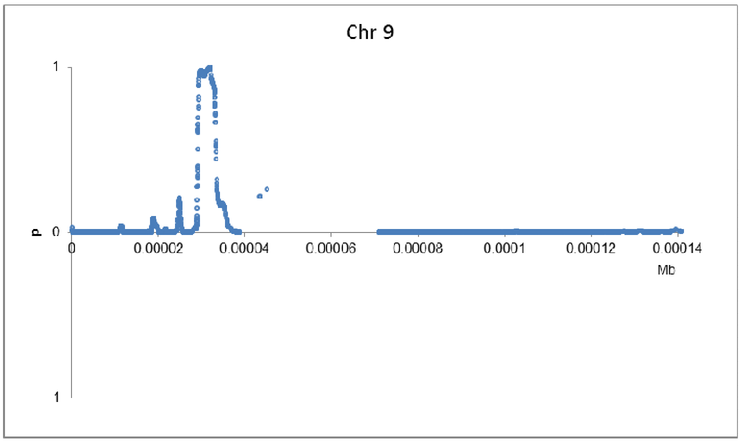


Figure 3.16: The IBD1 analysis between patient's parents (VI-3 and VI-4) identified an unique peak on chr9p21.1-p13.2. This peak reaches the maximum probability (value of 1) of this analysis.

Interestingly, the shared IBD1 region was transmitted by the two analysed individuals exclusively to their affected daughter (IV-1 in the haplotype of Figure 3.17) and corresponded to one of the homozygous regions identified in the previous analysis. The haplotype reconstruction shows the recombination event in unaffected patient's sister, which defines the telomeric limit of the homozygous candidate region (in the haplotype SNP_A-1933680 at 61.33 cM). Centromeric limit is instead defined by a recombination event occurring in patient's father (in the haplotype SNP_A-1812761 at 70.13 cM).

It is noteworthy that this autozygous region overlaps with a locus previously associated to a rare form of dHMN termed Jerash type (Table 3.8) (Christodoulou *et al.*, 2000).

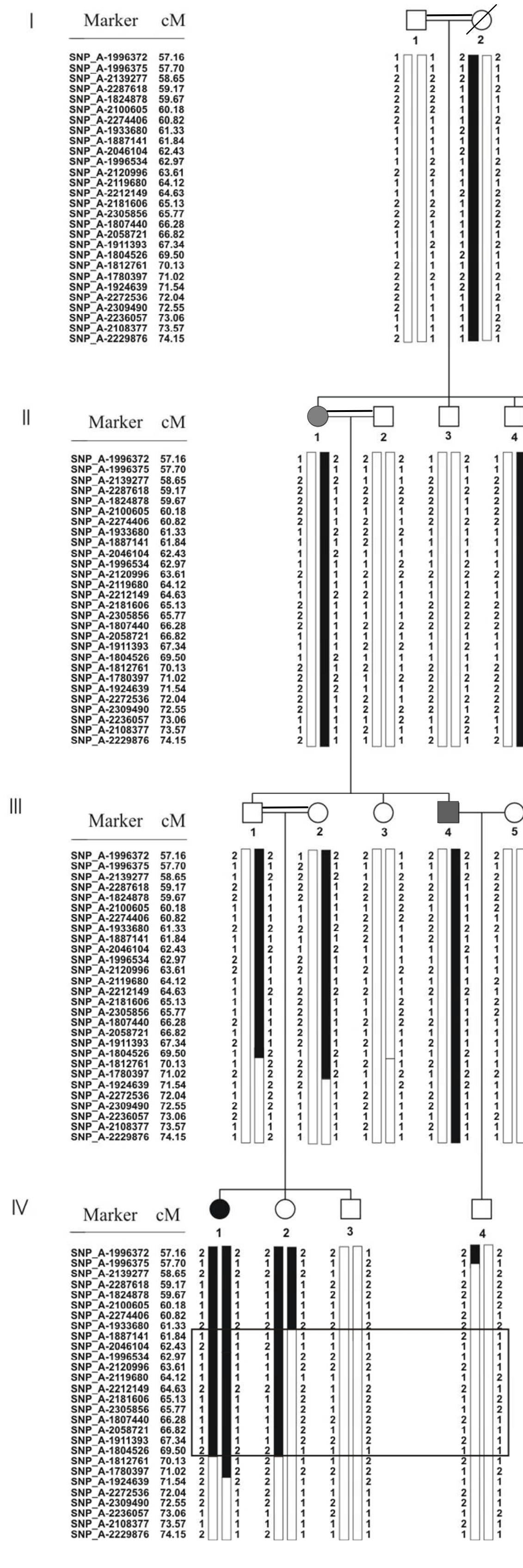


Figure 3.17: Haplotype segregation on chromosome 9p21.1-p13.2. The disease haplotype is in black, the candidate homozygous region shared exclusively by the patient is indicated by the box.

3.3.5.3 Whole-exome sequencing

The identification of potential disease-causing mutations in the 95 genes mapping to the chr9p21.1-p13.2 candidate region was carried out by the exome-sequencing approach. When this analysis was performed, WES data were available for the affected individuals V-3 and VI-6 of nucleus 1. Considering the haplotype segregation reported in Figure 3.17, these individuals carried in heterozygosis the IBD segment. Accordingly, heterozygous variants shared by both patients were searched on WES data. On December 2013, the high-coverage WES (at least 100X read depth) of patient VI-2 was performed in order to get further confirmations and exclusions. UTR regions were also included in the analysis by using the *Agilent Select Human All Exon v4+UTRs* kit for the target enrichment.

3.3.5.4 Variants in the candidate IBD region on chr9p21.1-p13.2

WES variants of the subjects V-3 and VI-6 were inspected in the IBD region paying particular attention to the genes underlying neurodegenerative and movement disorders with recessive transmission. Amongst them, the *APTX* gene is associated with a form of ataxia; *B4GALT1* with disorders of glycosylation; *SIGMAR1* underlies Juvenile amyotrophic lateral sclerosis 16; *SPG46* is associated with Hereditary Spastic Paraplegia, *GNE* with Inclusion Body Myopathy-2 and Nonaka Myopathy, and *EXOSC3* resulting in Pontocerebellar hypoplasia type 1B (Al-Saif *et al.*, 2011;Date *et al.*, 2001;Eisenberg *et al.*, 2001;Hansske *et al.*, 2002;Martin *et al.*, 2013;Wan *et al.*, 2012). Filtering and prioritization steps allowed to identify one highly candidate variant in the *SIGMAR1* gene (34,637,027 genomic position in hg19 release) (Table 3.9).

Chr	Candidate Variant	Variant call V-3	Variant call VI-6	Predicted Effect	Allele frequency	Haplotype	Functional Role	Confirmation by Sanger sequencing
9	*SIGMAR1* c.412G>C	C(ref)11/G9 (Q=99)	C(ref)12/G18 (Q=99)	Missense damaging	Novel variant	Homozygosity exclusive of patient VII-2	ER intracellular receptor, nervous system development	yes

Table 3.9: The best candidate SNV found in the two exome-sequenced patients, heterozygous for the IBD candidate region of patient VII-2. Variant calls are expressed with the number of reads for each allele (ref=reference allele in the genome) and Q=Phred-like quality score of the variant call. ER=endoplasmic reticulum. *SIGMAR1* showed the complementary alleles in the genome since the transcript is oriented on minus strand.

All undetectable variations (SNVs, indels and CNVs) of the poorly-covered exons and UTR regions were excluded by means of the high-coverage exome sequencing (mean read-depth of 100X) that have been recently carried out for patient VII-2 (data not shown). This analysis excluded a common genetic cause between the 2 nuclei because no homozygous mutations shared by all affected subjects were identified. WES analysis pointed to the *SIGMAR1* c.412G>C substitution that was confirmed by Sanger sequencing in homozygous state exclusively in patient VII-2.

3.3.6 Confirmations for *SIGMAR1* and *SGK223* variants

The analyses of Family 2 pinpointed two novel variants, the c.1529T>C in *SGK223* and the c.412G>C in the *SIGMAR1* gene, as the best candidate mutations for nucleus 1 and 2 respectively.

The segregation of both variants (obtained by genotyping all the available family members) confirmed previous haplotype analyses as reported in Figure 3.18.

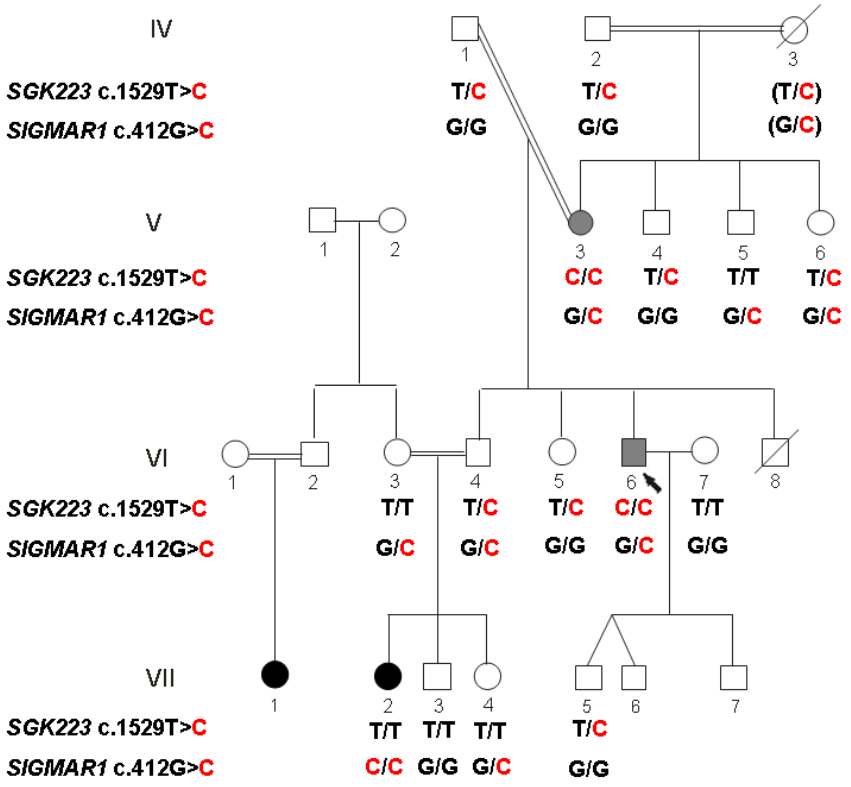


Figure 3.18: Genotypes of the two candidate variants identified for the *SGK223* and *SIGMAR1*genes in Family 2. Genotypes in parentheses of individual IV-3 are inferred.

3.3.7 Characterization of the c.1529T>C variant in the *SGK223* gene

In order to characterize the functional effect of the c.1529T>C variant, further information from literature, databases and prediction tools were collected. The *SGK223* gene (NM_001080826) codes for PRAGMIN (NP_001074295, homolog of rat pragma of Rnd2), a still not well characterized Tyrosine-protein kinase, probably involved in the regulation of Rho GTPase pathway for the cytoskeleton dynamics (Tanaka *et al.*, 2006). Unfortunately, the paralogous gene *PEAK1* is not deeply characterized as well. Full-length protein were first investigated to study potential functional domains of the N-terminus half where the substitution falls. InterproScan server showed that several tools predict at C-terminus a catalytic kinase domain confirming data of literature, but did not provide insights into the function of the N-terminus half (Figure 3.19).

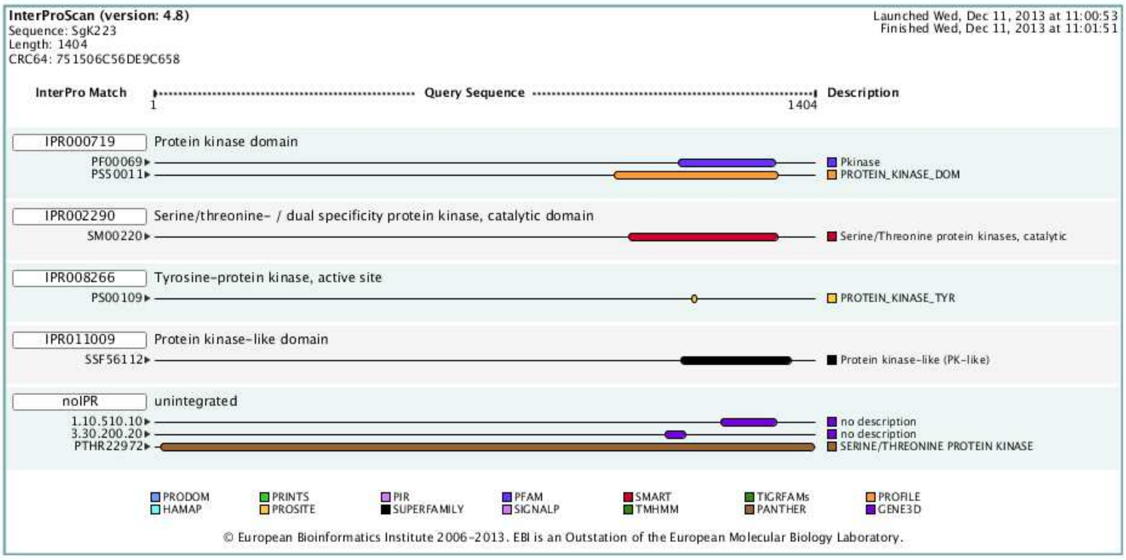
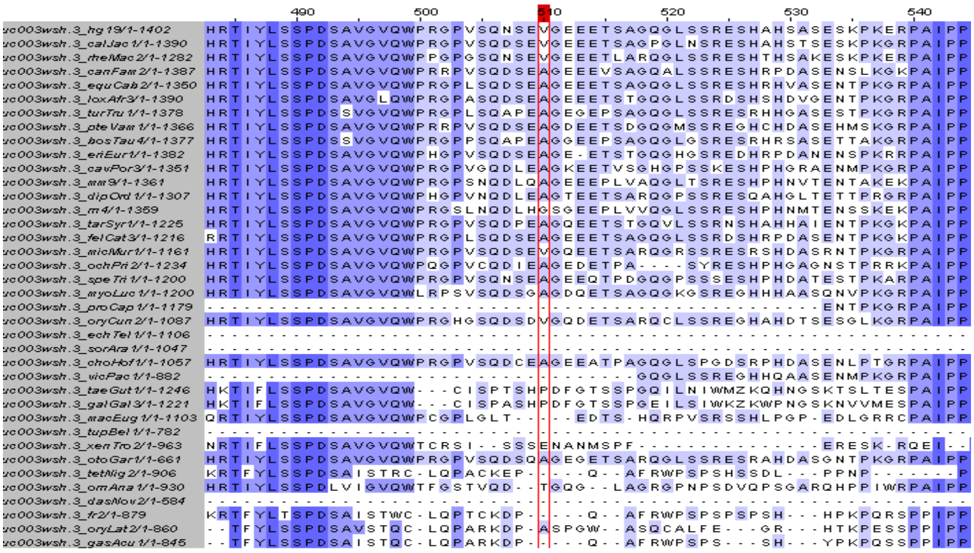

InterProScan (version: 4.8)
Sequence: SgK223
Length: 1404
CRC64: 751506C56DE9C658
Launched Wed, Dec 11, 2013 at 11:00:53
Finished Wed, Dec 11, 2013 at 11:01:51
InterPro Match
Query Sequence
1
1404
Description
IPR000719
Protein kinase domain
PF00069
PS50011
Pkinase
PROTEIN_KINASE_DOM
IPR002290
Serine/threonine- / dual specificity protein kinase, catalytic domain
SM00220
Serine/Threonine protein kinases, catalytic
IPR008266
Tyrosine-protein kinase, active site
PS00109
PROTEIN_KINASE_TYR
IPR011009
Protein kinase-like domain
SSF56112
Protein kinase-like (PK-like)
noIPR
unintegrated
1.10.510.10
3.30.200.20
PTHR22972
no description
no description
SERINE/THREONINE PROTEIN KINASE
PRODOM
HAMAP
PRINTS
PROSITE
PIR
SUPERFAMILY
PFAM
SIGNALP
SMART
TMHMM
TIGRFAMs
PANTHER
PROFILE
GENE3D
© European Bioinformatics Institute 2006-2013. EBI is an Outstation of the European Molecular Biology Laboratory.

Figure 3.19: Result of PRAGMIN protein for domain prediction/identification by InterProScan.

The missense substitution determines the p.Val510Ala amino acid change from valine to alanine. Even if this variant has never been reported in variant databases, it was not predicted to be deleterious by SIFT, Polyphen2, MutationTaster and LRT tools, suggesting a neutral effect at protein level. This result was probably influenced by the biochemical similarity between valine and alanine, and by the presence of alanine in the wild-type protein of other organisms (Figure 3.20).



```
uc003wsh.3_hg19/1-1402      HRTIYLSSPDSAVGVQWPRGPVSQNSEVGEEETSAGQGLSSRESHAHSASESKPKERPAIPP
uc003wsh.3_calJac1/1-1390   HRTIYLSSPDSAVGVQWPRGPVSQDSEVGEEETSAGPGLNSRESHAHSTSESKPKGRPAIPP
uc003wsh.3_rheMac2/1-1282   HRTIYLSSPDSAVGVQWPPGPGSQNSEVGEEETLARQGLSSRESHTHSAKESKPKERPAIPP
uc003wsh.3_canFam2/1-1387   HRTIYLSSPDSAVGVQWPRRPVSQDSEAGEEEVSAGQALSSRESHRPDASENSLKGKPAIPP
uc003wsh.3_equCab2/1-1350   HRTIYLSSPDSAVGVQWPRGPLSQDSEAGEEETSAGQGLSSRESHRHVASENTPKGRPAIPP
uc003wsh.3_loxAfr3/1-1390   HRTIYLSSPDSAVGLQWPRGPASQDSEAGEEETSTGQGLSSRDSHSHDVGENTPKGRPAIPP
uc003wsh.3_turTru1/1-1378   HRTIYLSSPDSSVGVQWPRGPLSQAPEAGEGEPSAGQGLSSRESRHHGASESTPKGRPAIPP
uc003wsh.3_pteVam1/1-1366   HRTIYLSSPDSAVGVQWPRRPVSQDSEAGDEETSDGQGMSSREGHCHDASEHMSKGRPAIPP
uc003wsh.3_bosTau4/1-1377   HRTIYLSSPDSSVGVQWPRGPPSQAPEAGGEEPSAGQGLGSRESRHRSASETTAKGRPAIPP
uc003wsh.3_eriEur1/1-1382   HRTIYLSSPDSAVGVQWPHGPVSQDSEAGE-ETSTGQGHGSREDHRPDANENSPKRRPAIPP
uc003wsh.3_cavPor3/1-1351   HRTIYLSSPDSAVGVQWPRGPVGQDLEAGKEETVSGHGPSSKESHPHGRAENMPKGRPAIPP
uc003wsh.3_mm9/1-1361       HRTIYLSSPDSAVGVQWPRGPSNQDLQAGEEEPLVAQGLTSRESHPHNVTENTAKEKPAIPP
uc003wsh.3_dipOrd1/1-1307   HRTIYLSSPDSAVGVQWPHGPVNQDLEAGTEETSARQGPSSRESQAHGLTETTPRGRPAIPP
uc003wsh.3_rn4/1-1359       HRTIYLSSPDSAVGVQWPRGSLNQDLHGSGEEPLVVQGLSSRESHPHNMTENSSKEKPAIPP
uc003wsh.3_tarSyr1/1-1225   HRTIYLSSPDSAVGVQWPRGPVSQDPEAGQEETSTGQVLSSRNSHAHHAIENTPKGKPAIPP
uc003wsh.3_felCat3/1-1216   RRTIYLSSPDSAVGVQWPRGPLSQDSEAGEEETSAGQGLSSRDSHRPDASENTPKGKPAIPP
uc003wsh.3_micMur1/1-1161   HRTIYLSSPDSAVGVQWPRGPVSQDSEVGQEETSARQGRSSRESRSHDASRNTPKGRPAIPP
uc003wsh.3_ochPri2/1-1234   HRTIYLSSPDSAVGVQWPQGPVCQDIEAGEDETPA----SYRESHPHGAGNSTPRRKPAIPP
uc003wsh.3_speTri1/1-1200   HRTIYLSSPDSAVGVQWPRGPVSQNSEAGEEQTPDGQGPSSSESHPHDATESTPKARPAIPP
uc003wsh.3_myoLuc1/1-1200   HRTIYLSSPDSAVGVQWLRPSVSQDSGAGDQETSAGQGKGSREGHHHAASQNVPKGRPAIPP
uc003wsh.3_proCap1/1-1179   --------------------------------------------------ENTPKGRPAIPP
uc003wsh.3_oryCun2/1-1087   HRTIYLSSPDSAVGVQWPRGHGSQDSDVGQDETSARQCLSSREGHAHDTSESGLKGRPAIPP
uc003wsh.3_echTel1/1-1106   --------------------------------------------------------------
uc003wsh.3_sorAra1/1-1047   --------------------------------------------------------------
uc003wsh.3_choHof1/1-1057   HRTIYLSSPDSAVGVQWPRGPVSQDCEAGEEATPAGQGLSPGDSRPHDASENLPTGRPAIPP
uc003wsh.3_vicPac1/1-882    -----------------------------------GQGLSSREGHHQAASENMPKGRPAIPP
uc003wsh.3_taeGut1/1-1246   HKTIFLSSPDSAVGVQW---CISPTSHPDFGTSSPGQILNIWMZKQHNGSKTSLTESPAIPP
uc003wsh.3_galGal3/1-1221   HKTIFLSSPDSAVGVQW---CISPASHPDFGTSSPGEILSIWKZKWPNGSKNVVMESPAIPP
uc003wsh.3_macEug1/1-1103   QRTIYLSSPDSAVGVQWPCGPLGLT-----EDTS-HQRPVSRSSHLPGP-EDLGRRCPAIPP
uc003wsh.3_tupBel1/1-782    --------------------------------------------------------------
uc003wsh.3_xenTro2/1-963    NRTIFLSSPDSAVGVQWTCRSI--SSSENANMSPF---------------ERESK-RQEI--
uc003wsh.3_otoGar1/1-661    HRTIYLSSPDSAVGVQWPRGPVSQDSQAGEGETSARQGLSSRESRAHDASGNTPKGRPAIPP
uc003wsh.3_tetNig2/1-906    KRTFYLSSPDSAISTRC-LQPACKEP-----Q--AFRWPSPSHSSDL---PPNP-------P
uc003wsh.3_ornAna1/1-930    HRTIYLSSPDLVIGVQWTFGSTVQD--TGQG--LAGRGPNPSDVQPSGARQHPPIWRPAIPP
uc003wsh.3_dasNov2/1-584    --------------------------------------------------------------
uc003wsh.3_fr2/1-879        KRTFYLTSPDSAISTWC-LQPTCKDP-----Q--AFRWPSPSPSPSH---HPKPQRSPPIPP
uc003wsh.3_oryLat2/1-860    --TFYLSSPDSAVSTQC-LQPARKDP-ASPGW--ASQCALFE---GR---HTKPESSPPIPP
uc003wsh.3_gasAcu1/1-845    --TFYLSSPDSAISTQC-LQPARKDP-----Q--AFRWPSPS---SH---YPKPQSSPPIPP
```

Figure 3.20: Multiple amino acid sequence alignment of PRAGMIN where the substitution falls. Higher amino acid identity corresponds to more intense colour. Amino acid at position 510 is marked by a box. Alignment was provided by *UCSC MultiZ46Way GRCh37/hg19.*

On the other hand, valine 510 is conserved in primates, indicating a possible primate-specific evolution of this allele. Phosida, NetPhos and NetAcet prediction tools did not showed evident alterations in the most common post-translational modifications such as phosphorylation and acetylation (data not shown). Conversely, the c.1529T>C nucleotide substitution could affect the mRNA splicing. Indeed, data obtained by different prediction tools were in agreement with an alteration of the correct splice site selection and usage. (Table 3.10). To increase the reliability of this *in silico* analysis, multiple tools were used. They rely on the sequence information, the intrinsic strength by which the splice sites are recognized by the spliceosome, as well as the antagonistic dynamics of proteins that bind Exonic Splicing Enhancers (ESEs) and Exonic Splicing Silencers (ESSs). ESE-finder and FAS-ESS have already successfully predicted the disruption of ESEs/ESSs in a variety of genes, including *SMN1*, *SMN2* and *POMGNT1* involved in neuromuscular diseases (Cartegni and Krainer 2002;Oliveira *et al.*, 2008).

Prediction tool	Result	Reference motif	Mutated motif	Wild Type	Mutant	Variation
Human Splicing Finder	Acceptor Splice site	ctccgaggtaggt	ctccgaggcagGT	75.65	83.33	10,2% (significant difference >10%)
	Donor Splice site	gaggtaggt	GAGgcaggt	89.9	63.06	-30% **Site broken** (significant difference >10%)
ASSEDA server	Splice site	gaggt	GAggc	4.9 (ΔR)	6.5 (ΔR)	**Increased** (Fold change -131, %binding 0.7)
		gaggtAggt	gaggcaggt	8.2 (ΔR)	1.2 (ΔR)	**Decreased** (Fold change +3, %binding 303.3)
MaxEntScan	Splice site	taggtgaag	CAGgtgaag	3.29	6.66	+102% (significant difference >20%)
	Splice site	gaggtaggt	GAGgcaggt	9.65	1.9	-81% (significant difference >20%)
ESE Finder	Exonic Splicing Enhancer	ggtaggt	ggcaggt	-	82.59	**New Site, specific for ASF splicing factor** (significant difference >10%)
- Computational approach for Silencer motifs (Sironi *et al.*, 2004) **- FAS-ESS server**	Exonic Splicing Silencer	ggtaggtg	ggcaggtg	69.88	-	**Site broken**

Table 3.10: Predictions of splice site and ESE/ESS elements for the c.1529T>C in *SGK223* and relative scores. Mutated position is underlined in the nucleotide sequence.

Interestingly, these analyses predicted in the reference sequence the presence of a natural splice site that has never been annotated among the human *SGK223* transcripts of NCBI Genome Browser database. Conversely, the splicing event seemed to occur in non-human transcripts displaying the same allelic variant in the reference sequence. A further analysis of human expressed sequenced tags (ESTs) identified 4 human ESTs encompassing this region: two not-spliced and expressed at embryonal stages (CN341734 and CN341745) and two spliced and expressed in adult brain (HY116488 and HY306767). These findings might indicate the presence of a cryptic splice site with a tissue-specific and developmental activation. According to the late onset of the disease, the c.1529T>C substitution could modify the strength of this cryptic splice site or its regulation in the adult.

3.3.8 Characterization of the c.412G>C variant in the *SIGMAR1* gene

Ascertained from literature that *SIGMAR1* underlies a motor neuron disorder (ALS16) and is a good candidate gene, the pathogenic effect of the specific missense variant was investigated. The c.412G>C substitution has never been reported in variant databases neither in LOVD, HGMD and ClinVar mutation databases. It was predicted to be deleterious by Polyphen2, MutationTaster and LRT tools. Moreover, in the current study it was excluded from 200 control chromosomes belonging to same geographic area of this family. The c.412G>C substitution falls within exon 3 which is not present in all alternative *SIGMAR1* transcripts (isoforms 2 and 10 lack exon 3 as reported in Figure 3.21). However, the pathogenic effect of the c.412G>C substitution implies the expression of full-length isoforms in the nervous system. The majority of isoforms that are currently annotated in UCSC Genome Browser presents the exon 3 (Figure 3.22). Moreover, the analysis of spliced ESTs confirmed the presence of isoforms with exon 3 in the nervous system (for instance ESTs DA075272, CD519094, BI596428 and AL537799).

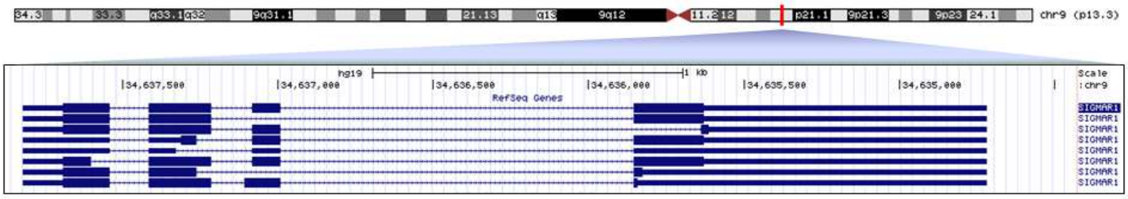


Figure 3.21: All transcript variants for *SIGMAR1* currently annotated in RefSeq Genes track of UCSC Genome Browser. From top to bottom, transcript variant 1, 2, 6, 7, 11 (non-coding), 8, 10 and 9, all under review and not still validated. From the genomic positions, transcripts are oriented on minus strand and are here reversed.

The exon 3 of *SIGMAR1* encodes the ligand-binding domain of the protein where specific anionic residues were showed to be critical for its function (Seth *et al.*, 2001). The residue at position 138 (glutamic acid) is negatively-charged and is here substituted by an uncharged one (glutamine). It is interesting to note that the whole gene is highly conserved and only 6 polymorphisms are reported in the coding region (dbSNP137 and dbSNP138 annotations, MAF>1%). Also p.Glu138Gln change falls in a highly conserved sequence (Figure 3.23), as confirmed by phyloP, GERP and phastCons tools.

In order to confirm the *SIGMAR1* involvement in dHMN, a genetic screening was performed in other 12 unrelated index cases displaying a clinical picture similar to patient VII-2 and without an identified genetic cause yet. Surprisingly, a second variant (c.448G>A) was identified in one patients belonging to a small family with two affected brothers. The family came from South of Italy and was clinically ascertained by Dr. Bengala and Petrucci at S. Camillo-Forlanini Hospital of Rome. Both patients presented a phenotypic onset and progression similar to that of the subject VII-2 of nucleus 2, and the absence of the disease in the parents suggested a recessive inheritance. Sanger sequencing enabled to confirm the homozygous segregation of the c.448G>A variant in the two affected subjects II-1 and II-2 (Figure 3.22).

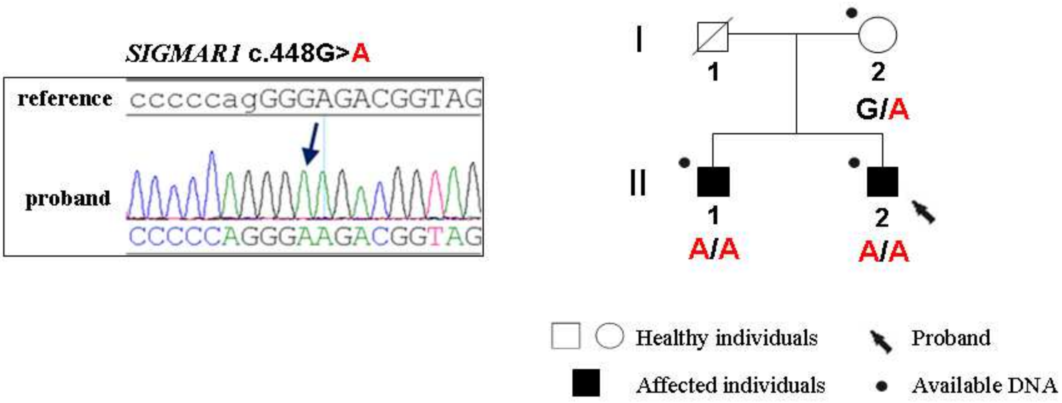


Figure 3.22: Electropherogram with the homozygous *SIGMAR1* c.448G>A variant indicated by the arrow, and the genotyping in the available family members.

Also the c.448G>A variant has never been reported in variant databases, neither in LOVD, HGMD and ClinVar mutation databases. It was predicted to be deleterious by different tools and was excluded from 200 control chromosomes from South of Italy. It maps to exon 4 near the splice-site junction with exon 3 and leads to the substitution of a glutamic acid residue with a hydrophobic phenylalanine at position 150 (p.Glu150Phe). The change occurred in the same conserved amino acid motif of the first mutation, for which the anionic residues have been previously described as playing a critical role (Figure 3.23). The two mutations fell in the same C-terminus chaperone domain that was demonstrated exposed in the inner ER membrane, and a similar pathogenic role was thus attributed to two variants (Ortega-Roldan *et al.*, 2013).

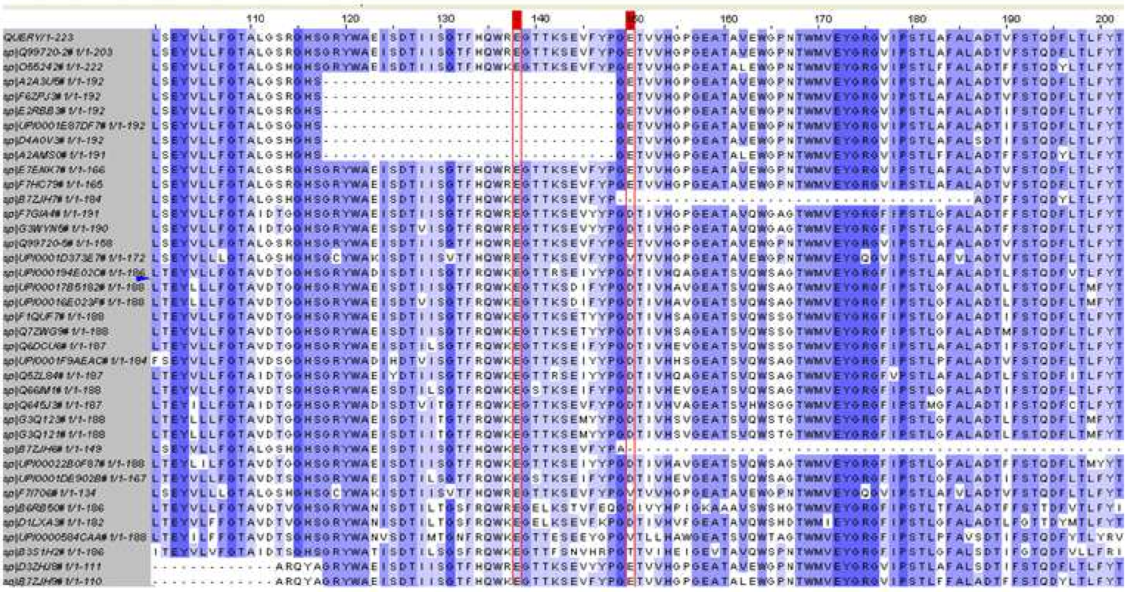


Figure 3.23: Multiple amino acid sequence alignment of SIGMAR1where p.Glu138Gln and p.Glu150Phe substitutions fall. Higher amino acid identity corresponds to more intense colour. Amino acid at position 138 and 150 are marked by a box. Alignment was provided by *UCSC MultiZ46Way GRCh37/hg19.*

In conclusion, these findings strongly support the causality of the *SIGMAR1* c.412G>C and c.448G>A mutations in the dHMN disease and prompt to consider *SIGMAR1* the gene of the dHMN Jerash type *locus*.

3.4 Identification of the candidate gene *FBXO41* associated with a complicated form of peripheral neuropathy and spastic paraplegia

3.4.1 Clinical picture of Family 3

This consanguineous family comes from a little village in Veneto region and presents three brothers (IV-2, IV-6, IV-7) affected by a complex form of motor neuropathy with spastic paraplegia and mental retardation. Neurological exam was performed in all subjects of the fourth generation (both healthy and affected individuals) by Dr. Micaglio of Neurology Division, Hospital of Montebelluna (Treviso). The three patients displayed a similar clinical picture, characterized by weakness of the lower limbs, bilateral *pes cavus*, spastic gait, hyperreflexia and mild hypertonicity. The age of onset was the third decade of life. Electrophysiological studies and nerve biopsy revealed a reduced motor conduction velocity and the presence of a distal motor neuropathy. Brain magnetic resonance imaging, computed axial tomography and electroencephalography showed non-significant CNS involvement. In addition, all affected members showed a neuropsychological impairment characterized by mild mental retardation, visual agnosia, short- and long-term memory deficiency, signs of perseveration and confabulation.

The analysis of the pedigree suggested an autosomal recessive way of inheritance of the disease. Indeed, as reported in Figure 3.24, the healthy parents (III-1 and III-2) of the affected sibs were first cousins and the recurrence of the clinical phenotype was observed only in one generation.

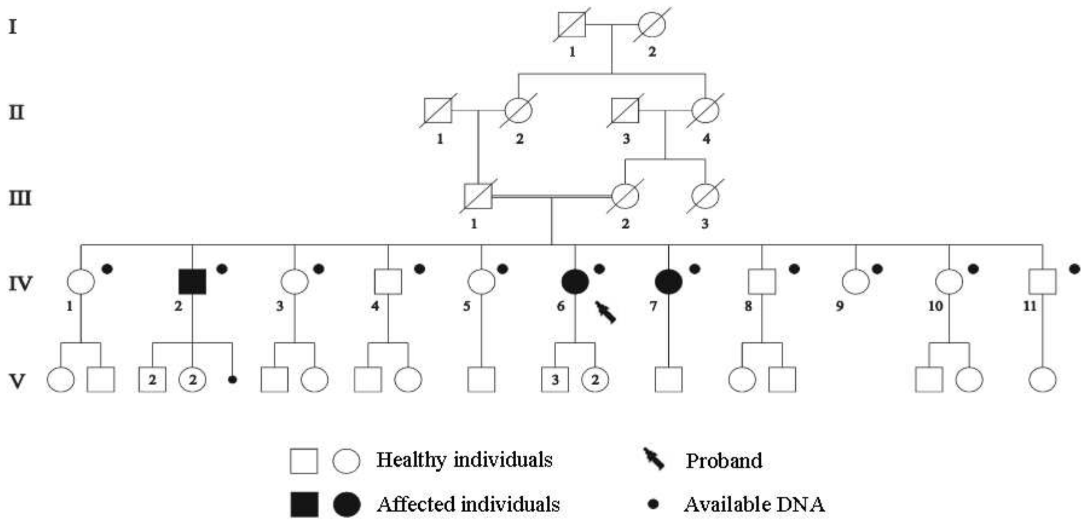


Figure 3.24: Pedigree of Family 3.

3.4.2 Preliminary results

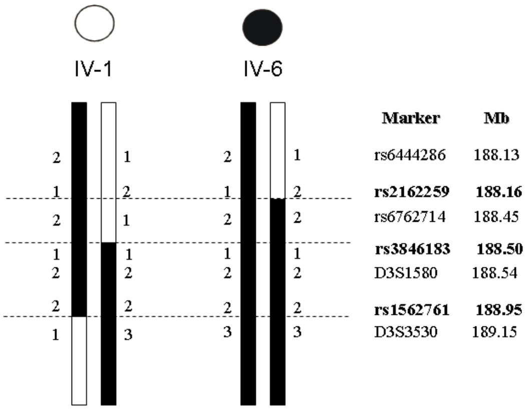
First studies of whole-genome scan were performed with 382 STR markers and identified a candidate region on chromosome 3q27-q28 (D3S1580 and D3S3669 flanking markers), defined *SPG14 locus* (Vazza *et al.*, 2000). A further investigation with other 50 STRs allowed to detect a second chromosomal region on chr21q11.1-q21.1. Screening of candidate genes in this region, identified the c.*377T>C substitution in the 3'UTR of the *HSPA13* gene (NM_006948.4). This sequence variant was absent in 300 healthy individuals and a luciferase assay showed a decrease of the protein translation in presence of the substitution, probably due to a mechanisms of post-transcriptional regulation by microRNAs (unpublished data). A further genome-wide search by using over 200,000 SNP markers (Illumina Human CNV370-Quad platform) excluded all known HMSN and HSP genes and *loci* associated with this phenotype, confirmed the two candidate regions already detected, and highlighted a new linkage signal on chr2p13.3-p12. The homozygous status was assessed for three candidate regions.

3.4.3 Better definition of the candidate regions

The high number of genotyped individuals (3 affected and 8 unaffected individuals) and the clear recessive mode of inheritance enabled to use well-defined parameters for a robust linkage analysis (autosomal recessive model, full penetrance and disease allele frequency of 0.001). Indeed, maximum linkage peaks displayed significant LOD score values of 3.86 (3 is the minimum value to accept the hypothesis of linkage), indicating 3 reliable candidate regions. For each linkage peak identified, haplotypes with SNPs were combined with STR markers previously genotyped, thus unifying the advantages and compensating the limits of both (i.e. STR markers are more informative for their high variability but more interspersed than SNPs, that are more dense in the genome).

For instance, this analysis allowed to divide the centromeric terminus of linkage region on chr3q27-q28 into 2 chromosomal segments according with the haplotype (Figure 3.25). Flanking makers were identified by taking advantage of recombination events in unaffected and affected brothers IV-1 and IV-6 respectively. All candidate linkage regions with refined limits are reported in Table 3.11.

Furthermore, considering the recessive trait and the consanguinity, a more-dense homozygosity mapping (with 350,000 SNPs) was performed to overcome computational limits of multipoint linkage analysis and to confirm the linkage regions, since it required a lower number of SNPs (about 25,000 SNPs). Also in this family a less-stringent analysis was carried out by considering the homozygosity of 2 patients at least, thus limiting genotyping errors and incomplete detections (regions reported in Table 12 of Appendix E).



Linkage region	Previous centromeric limit (SNP)	Previous centromeric limit (micro
3q27-q28	rs6444286	D3S1580

Figure 3.25: Key recombination events occurred in unaffected and affected sisters define upper and lower linkage regions on the centromeric terminus of chr3q27-q28 region. Markers reported in bold type are novel flanking markers, in the table previous flanking markers were annotated.

Linkage region	from (bp)	to (bp)	from SNP	to SNP	Mb	Total RefSeqs	OMIM genes for neurological disorders
2p13.3-p12	72251751	75681462	rs1878503	rs10496198	3.4	56	- *ALMS1* for Alstrom syndrome (AR) - *DCTN1* for HMNVIIB (AD)
3q27-q28 Centromeric region	188161706	188500658	rs2162259	rs3846183	0.34	1	-
3q27-q28 Telomeric region	188952524	192088525	rs1562761	rs9859577	3.14	16	-
21q11.2-q21.1	14658830	19044397	rs2258300	rs2824435	4.4	21	-
				total	11.3	94	

Table 3.11: Limits of linkage regions obtained by the combinations of haplotypes with SNPs and STRs. AR= autosomal recessive, AD= autosomal dominant.

3.4.4 Whole-exome sequencing

Considering the 94 genes mapping to the linkage regions, even in Family 3 the identification of disease-causing mutations was carried out by the exome-sequencing approach. Exome capture, sequencing and computational analysis were performed by the BGI for the proband IV-2 in 2011 and recently (December 2013) for the patient IV-7 (more technical details in Materials and Methods).

3.4.5 Statistical evaluation of the performance

The performance of WES for IV-2 appeared similar to that of Family 2. On the other hand, recent data obtained on December 2013 from patient IV-7 have shown a higher performance due to the optimization of the technique and the higher coverage achieved. The specificity, which was calculated through the percentage of reads mapped to the target regions (70% in 2011 and 78% in 2013) and uniquely mapped (81% in 2011 and 87% in 2013), has been increased over two years as expected. It is interesting to note that the mean read depth of 149X for IV-7 strongly overcame the value of IV-2 (101X), even if this value is only indicative because of the high variability between genes and exons. The coverage for target regions (more than 1X read-depth) was high for both samples (97.8% for IV-2 and 99.8% for IV-7), but the 10X read-depth required for a high-confident variant calling increased from 89% in 2011 to 99.5% in 2013 The better quality of the more recent techniques can be also appreciated through the following histograms (Figure 3.26).

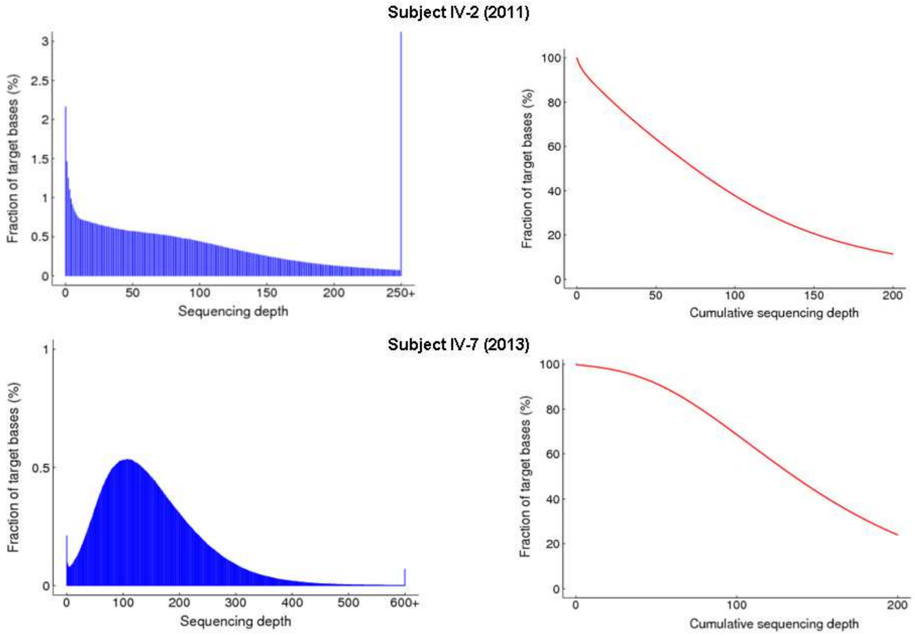


Figure 3.26: Histograms of distribution and cumulative distribution for sequencing depth in the target regions. Subject IV-2 was analysed in 2011, IV-7 has been recently sequenced (December 2013).

3.4.6 Single Nucleotide Variants (SNVs) and indels annotation

The final set of high-quality variants obtained by the exome sequencing was annotated in the following tables (Table 3.12 and 3.13).

Categories of SNVs	**Patient IV-2 (2011)**	**Patient IV-7 (2013)**
Number of genomic positions for calling SNPs(1)	135216361	134975362
Total number of high-quality SNVs	91935	108898
Homozygous	33653	41279
Heterozygous	58282	67619
Intron	56440	64477
Intergenic	1553	3450
5’-UTR	3519	3922
3’-UTR	4555	7257
Synonymous	10440	5973
Splice site(2)	2549	2729
Missense	12620	11119
Stop codon abolishing	69	60
Nonsense	190	120

Table 3.12: High-quality SNVs annotated in the *snp.filter.gff* files. (1) Capture target regions and its 200bp flanking regions; (2) Intronic SNPs within 10bp of exon/intron boundary.

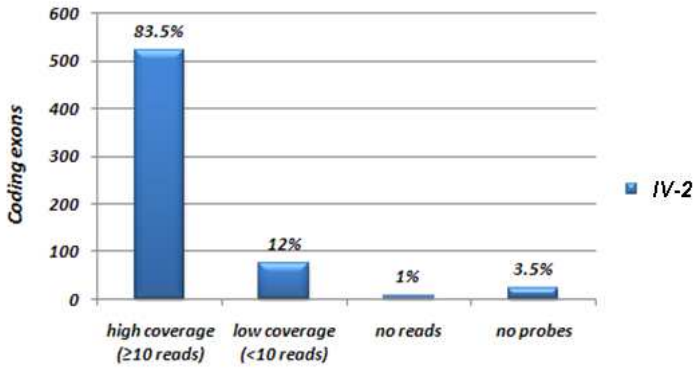
Categories of indels	**Patient IV-2 (2011)**	**Patient IV-7 (2013)**
Total number	7067	11855
Total insertions	3335	5471
Total deletions	3732	6384
Homozygous	2926	4831
Heterozygous	4141	7024
Intron	5343	8537
Intergenic	63	325
Promoter (1000bp upstream the gene)	29	120
5’-UTR	319	354
3’-UTR	393	761
Non-frameshift	193	258
Splice site	386	499
Insertions in coding sequence	277	267
Deletions in coding sequence	257	282
Frameshift	341	291

Table 3.13: Indels annotated in the *indel.gff* files.

The analysis of the current study started from 91,935 high-quality SNVs and 7067 total indels for the subject IV-2. From a comparison between variants obtained in 2011 (sample IV-2) and 2013 (sample IV-7), a higher accuracy and the inclusion of UTR regions were evident for the most recent data. A decrease of false positive variants in non-synonymous SNVs and a higher sensitivity in indels detection were observed.

3.4.7 Coverage analysis

Before having the confirmation with the recent WES, the coverage was analysed for the patient IV-2 who displayed a low accuracy of the technique. The read depth within the 4 linkage regions on 2p13.3-p12, 3q27-q28 (centromeric and telomeric regions) and 21q11.2-q21.1 was visualized through the read alignment. After many observations, a read depth under 10X was arbitrarily evaluated as insufficient to accurately call homozygous mutations or to assess false positive calls (Figure 3.27 and Table 3.14).



	IV-2
Total coding exons	625
High coverage depth (≥10 reads)	522
Low coverage depth (<10 reads)	75
No reads	6
No probes	22

Figure 3.27: Depth of coverage per exon, on 3 linkage regions. Y-axis indicates the number of coding exons; the correspondent percentage of total is indicated above each histogram bar. Poorly-covered coding exons were annotated for IV-2 patient. Table 3.14: Exons for each category of coverage depth on 3 linkage region.

This analysis allowed to highlight problematic exons that were not captured by probes (1%) or not well-sequenced (12%) within the 4 candidate regions. Moreover, the capture kit was not adequately designed for 22 exons. For these reasons the 3.5% coding exons required a coverage by direct sequencing. The performance for these regions was lower than for the whole exome. Over 10X read depth, the 89% coverage for the whole exome decreased to 83.5% for the 4 candidate regions. All variants identified by Sanger sequencing in the selected poorly-covered exons were detected in the patient IV-7 recently sequenced but none of them was considered a good candidate (poorly-covered exons in Table 8, Appendix D; variants in Table 13, Appendix E). The missense variant and the deletion mapping in *CLDN16* have been reported with several ID in the same G homopolymeric stretch, suggesting different submissions for a unique variant. The high-frequency of the missense variant (MAF>5%) could be thus extended to the deletion. Moreover the *CLDN16* gene was reported to be the primary cause of Renal hypomagnesemia 3, characterized by the specific involvement of renal epithelial cells and not related with nervous system (Simon *et al.*, 1999). Also the *ALMS1* gene, where the second candidate deletion (c.63delGAGGAGGAG) was detected, have been associated with Alström syndrome that is mainly characterized by renal involvement (Hearn *et al.,* 2002). This in-frame deletion of 9 nucleotides has never been reported before but maps in a GGA simple repeat, where many other in-frame insertions and deletions have been annotated. A relative frequency and a neutral effect were thus suggested for the indels of this region.

3.4.8 Variants in disease genes and loci and CNV analysis

Whole SNV dataset deriving from the WES was analyzed by CoNIFER program and a total of 21 and 8 CNVs (>1.5 kb) were identified for patients IV-2 and IV-7 respectively. No CNVs shared by the 2 affected subjects were observed (Table 14 in Appendix F).

The analysis of the variants in the known disease-associated genes and *loci* implicated in distal neuropathies and HSP, as well as all genes annotated in OMIM database associated with similar phenotypes enabled to exclude them. Indeed, very few variants in these genes showed a homozygous call different from the reference allele and with a lower than 2% frequency. Two variants in the *SHROOM4* (c.3385A>C) and *FAM58A* (c.49A>G) genes explained a dominant X-linked trait and they were not consistent with the autosomal mode of inheritance in this family. In addition, their disease haplotype segregated in more than one unaffected subject. Therefore, all OMIM genes or implicated in neurodegenerative disorders seemed unlikely to be the primary cause of the clinical phenotype in this family.

Chr	**Candidate Variant**	**Variant call IV-2 (2011)**	**Variant call IV-7 (2013)**	**Predicted Effect**	**Allele frequency**	**Haplotype**	**Functional Role**	**Confirmation by Sanger sequencing**
X	*SHROOM4* c.3385A>C	T(ref)0/G2 (Q=35)	T(ref)3/G9 (Q=73)	missense tolerated	rs199502054, not validated, unknown frequency	X-linked transmission, in more than one unaffected individual	Cause of Stocco dos Santos mental retardation syndrome (X-linked)	-
X	*FAM58A* c.49A>G	T(ref)0/G0 (Q=1)	T(ref)0/G15 (Q=37)	missense tolerated	Novel variant	X-linked transmission, in more than one unaffected individual	Cause of STAR syndrome (Dominant X-linked)	-

Table 3.15: Filtered SNVs found in genes annotated in OMIM and related to neurodegenerative disorders. Variant calls are expressed with the number of reads for each allele (ref=reference allele) and Q=Phred-like quality score of variant call. Homozygous status was considered.

3.4.9 Variants in the candidate linkage regions

The exclusion of the known disease-related loci suggested the presence of a novel disease gene and a similar analysis was performed for the WES variants mapping to the candidate linkage regions **chr2p13.3-p12**, **chr3q27-q28** (centromeric and telomeric regions) and **chr21q11.2-q21.1**. A total of 214 variants mapped to these regions and 45 were prioritized because in coding exons and close to splice sites. Finally 7 SNVs with a lower than 2% frequency were filtered. Amongst them, the analysis highlighted 3 best candidate variants (Table 3.16).

Chr	Candidate Variant	Variant call IV-2 (2011)	Variant call IV-7 (2013)	Predicted Effect	Allele frequency	Haplotype	Functional Role	Confirmation by Sanger sequencing
2	*FBXO41* c.950G>A	C(ref)1/T18 (Q=26)	C(ref)1/T135 (Q=64)	Missense damaging	Novel variant	Excluded from unaffected subjects	F-box protein 41, ubiquitin ligase	yes
2	*NAT8B* c.355G>T	C(ref)1/A118 (Q=72)	C(ref)1/A254 (Q=85)	Missense ?	rs62619834, Rare, MAF= 0.356%	Excluded from unaffected subjects	N-acetyltransferase (pseudogene)	-
3	*LPP* c.1142C>T	T83/A2(Cref) (Q=68)	C(ref)0/T253 (Q=99)	Missense tolerated	rs139075681, in EVS: MAF=0.02% (1 / 4545)	Excluded from unaffected subjects	Cause of Lipoma and acute myeloid Leukemia	yes

Table 3.16: Filtered SNVs found in genes of the homozygosity candidate regions. Variant calls are expressed with number of the reads for each allele (ref=reference allele) and Q=Phred-like quality score of variant call. Alleles in the genome and in the transcript are complementary for transcripts oriented on minus strand.

The variant c.355G>T in *NAT8B* seemed unlikely to be the genetic cause of the disease. Even if not-well characterized yet, *NAT8B* has been reported to be the pseudogene of *NAT8*. Unlike *NAT8* that encodes a N-acetyltransferase, *NAT8B* is described not encoding a functional protein in humans due to premature stop codons (Veiga-da-Cunha *et al.,* 2010). The variant c.1142C>T in the *LPP* gene maps to the small centromeric region on chr3q27. This missense variant was predicted to be “tolerated” although the gene is causal of disorders, probably because this specific variant is annotated in variant databases. Even though any point mutations have not been reported in the *LPP* gene to date, rearrangements have been associated with lipoma and acute myeloid leukemia, characterized by pathogenetic mechanisms and involvement of tissues apparently not related with neurodegenerative disorders and nervous system (COSMIC and OMIM databases). In addition, LPP protein seems mainly expressed in internal organs such as kidney, liver and lung, suggesting a non-neuronal primary function and specificity.

On the contrary, the “damaging” effect predicted for the novel missense variant c.950G>A in the *FBXO41* gene was in agreement with a putative pathogenic role. The involvement in the ubiquitin-proteasome pathway was reported for FBXO41 protein, which represents a pathogenic mechanism already associated with neurodegenerative disorders. In addition FBXO41 is ubiquitously expressed but mainly in the nervous system, suggesting a specific key role in it (data of GeneCard database). For these reasons, the c.950G>A substitution in *FBXO41* gene was considered the best putative mutation in this family.

3.4.10 Confirmation and characterization of the c.950G>A variant in the *FBXO41* gene

Previous analyses of this study pinpointed one novel variants, c.950G>A in the *FBXO41* gene, as the candidate disease-causing mutation in Family 3. Segregation analysis was performed by genotyping all family members for which the genomic DNA was available and as expected the c.950G>A variant was homozygous exclusively in the affected subjects IV-2, IV-6 and IV-7 (Figure 3.28).

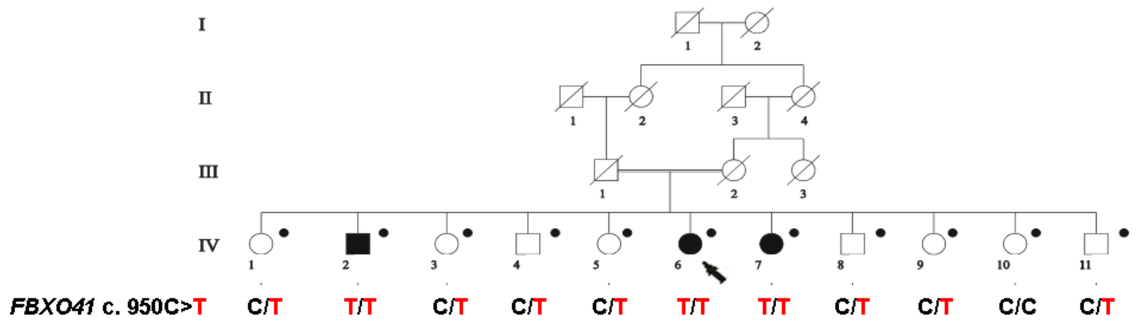


Figure 3.28: Genotypes of the candidate variant identified for the *FBXO41*gene in Family 3.

This substitution has been never reported in any variant database and was predicted to be "damaging" and "conserved" by the majority of prediction tools (Table 3.17). Moreover, the c.950G>A variant was excluded from 400 chromosomes of control subjects from Northern Italy.

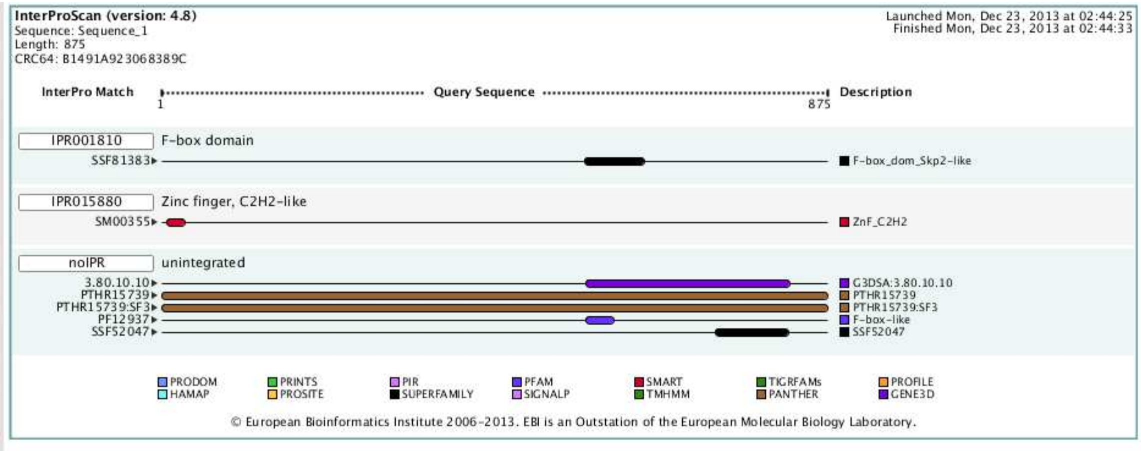
	SIFT (deleterious<0.05)	**Polyphen2 (probably damaging: 0.9-1)**	**Mutation Taster (0-1)**	**GERP++ score (>2)**	**PhyloP score**	**SiPhy score**
FBXO41 c.950G>A	Tolerated (0.15)	Probably damaging (0.99)	Damaging (0.99)	Conserved (5.17)	Conserved (2.4)	Conserved (17.2)

Table 3.17: Prediction tools with relative scores (in parenthesis)

The *FBXO41* gene (NM_001080410) encodes for the 875 amino acids protein FBXO41 (F-box only protein 41, NP_001073879), one of 46 members of the F-box protein family.

FBXO41 is not-well functionally characterized, and the only defined role is due to the F-box domain. Indeed, all proteins presenting F-box are involved in the E3 ubiquitin ligase complex SCF (SKP1-CUL1-F-box protein) for the recognition and ubiquitination of other target proteins.

In order to characterize the functional effect of the c.950G>A mutation, the full-length protein was first investigated to study potential functional domains in the N-terminus half where the substitution falls. InterproScan server showed that several tools predicted the well-characterized F-box domain in C-terminus half, confirming data in literature, but it did not provide insights into the function of the N-terminus (Figure 3.29).



InterProScan (version: 4.8)
Sequence: Sequence_1
Length: 875
CRC64: B1491A92306838 9C

Launched Mon, Dec 23, 2013 at 02:44:25
Finished Mon, Dec 23, 2013 at 02:44:33

© European Bioinformatics Institute 2006-2013. EBI is an Outstation of the European Molecular Biology Laboratory.

Figure 3.29: Result of FBXO41 protein for domain prediction/identification by InterProScan.

On the contrary, literature data reported at the N-terminus half an apolipophorin III-like domain, suggesting a role in cholesterol transport (Jin *et al.*, 2004). Lipophorin ApoL-III was already described in insects as a multifunctional lipoprotein involved in lipid binding and transport, displaying an amphipathic structure of 5 α-helices (Weers and Ryan 2006). Specifically, hydrophylic residues were described to be crucial for the lipid-induced opening of the structure and the protein-lipid interactions (Weers *et al.*, 2005). The secondary structure of FBXO41 was predicted by Jpred3 and Psipred softwares and revealed the p.Arg317Gln amino acid change localizing to a α-helix structure. The α-helix was also confirmed by ConSeq server which predicted an alternation between exposed and buried amino acids, indicative of a α-helix. Arginine localizes to an exposed position (indicated by "e" letter) like all charged amino acids in a α-helix (Figure 3.30). The p.Arg317Gln determines a change from the strongest positive-charged amino acid (arginine) to an uncharged residue (glutamine), leading to strongly change chemical properties. Glutamine could destabilize the structure of α-helix or disrupt a crucial function already described for hydrophylic residues by Weers and colleagues.

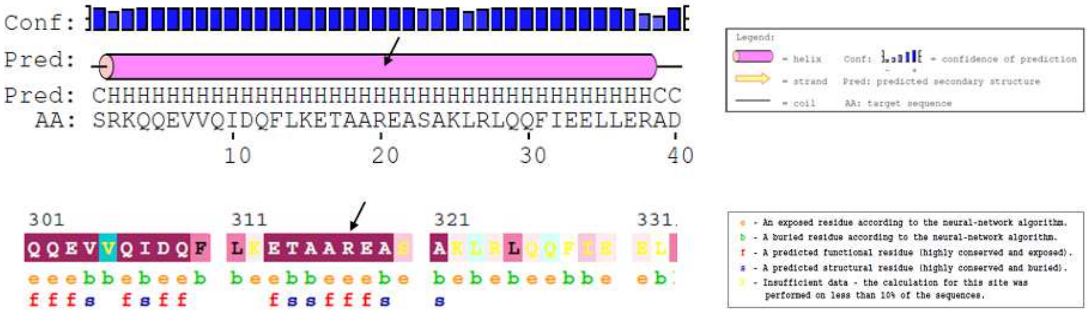

Conf:
Pred:
Pred: CHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCC
AA: SRKQQEVVQIDQFLKETAAREASAKLRLQQFIEELLERAD
10
20
30
40
Legend:
= helix
= strand
= coil
Conf: = confidence of prediction
Pred: predicted secondary structure
AA: target sequence
301
311
321
331
e - An exposed residue according to the neural-network algorithm.
b - A buried residue according to the neural-network algorithm.
f - A predicted functional residue (highly conserved and exposed).
s - A predicted structural residue (highly conserved and buried).
- Insufficient data - the calculation for this site was performed on less than 10% of the sequences.

Figure 3.30: (Top) α-helix structure predicted by Psipred software for the specific amino acid sequence where p.Arg317Gln falls (indicated by the arrow). (Bottom) Alternating exposed and buried amino acids predicted by the algorithm of ConSeq server, indicative of a α-helix structure with arginine in exposed position.

This position and the surrounding amino acid region appeared extremely conserved, in agreement with a functional role at protein level (Figure 3.31).

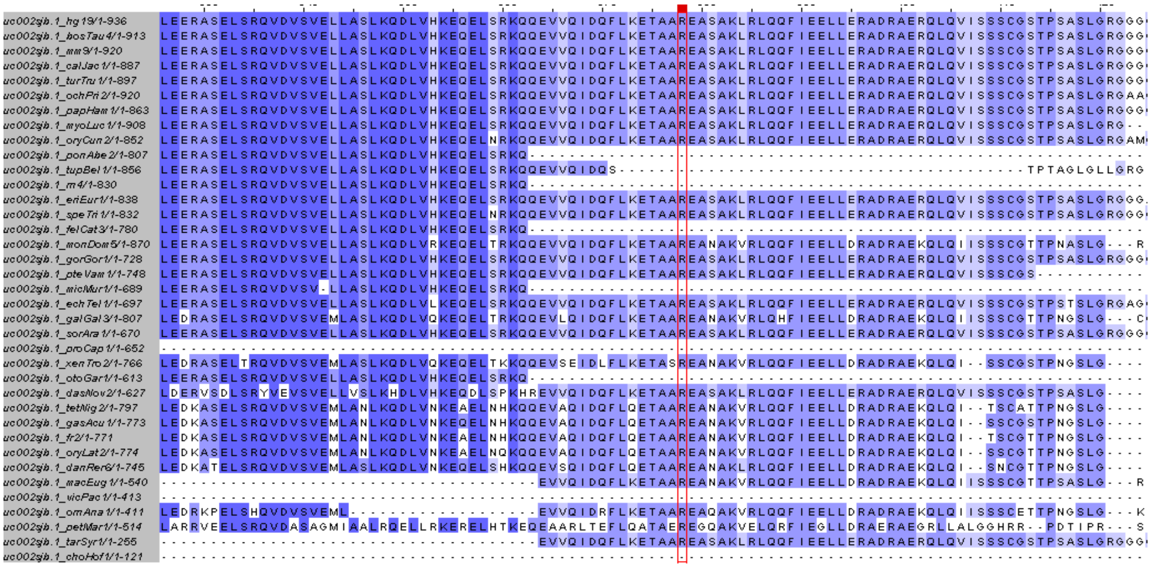


Figure 3.31: Multiple amino acid sequence alignment of FBXO41 where the substitution falls. Higher amino acid identity corresponds to more intense colour. Amino acid at position 317 is marked by a box. Alignment was provided by *UCSC MultiZ46Way GRCh37/hg19.*

Unfortunately, the homology modelling based on the sequence similarity was hampered by the lack of templates already crystallized and displaying a more than 30% of sequence identity for N-terminus half. Global and local alignments of full-length protein performed with BLAST tool found sequence conservation exclusively for the F-box domain, whereas the N-term portion displayed no significant similarity with other proteins.

Despite the paucity of information on the function of *FBXO41*, our findings strongly support the role of the c.950G>A mutation in the development of the disease in this family. For this reason, a mutation screening of the *FBXO41* gene in patients exhibiting a similar clinical picture is currently under way.

3.5 Insight into the genetic causes of HMSN and HSP with a dominant transmission

3.5.1 Clinical picture of Family 4

This four-generation family presents patients affected by a typical form of HMSN type V with heterogeneous clinical signs. Clinical features of axonal peripheral neuropathy (HMSN type II) such as weakness, atrophy of lower limb muscles and bilateral *pes cavus* were associated with pyramidal signs with spasticity of lower limbs and hyperreflexia. Two clinical pictures typical of HMSN and HSP co-occurred in many cases, in patients II-8, III-11, III-13, IV-3 and IV-5 the involvement of peripheral nervous system was predominant or exclusive, whereas in the patients III-2 and III-3 the central nervous system was mainly affected. Family members were seen by Dr. Angelini of the Department of Neurosciences, University of Padova. Motor conduction velocity was slowed ranging from 38 to 44 m/s, and cerebral and spinal MRI revealed no abnormalities. The age of onset was about the third or fourth decade of life, with a variable and progressive course.

The analysis of the pedigree suggested an autosomal dominant mode of inheritance of the disease, since the recurrence of the clinical phenotype was observed in all generations, in both male and female and there was one male-to-male transmission. To explain the phenotypic heterogeneity, an unique genetic cause or two distinct genetic causes for HMSN and HSP were proposed.

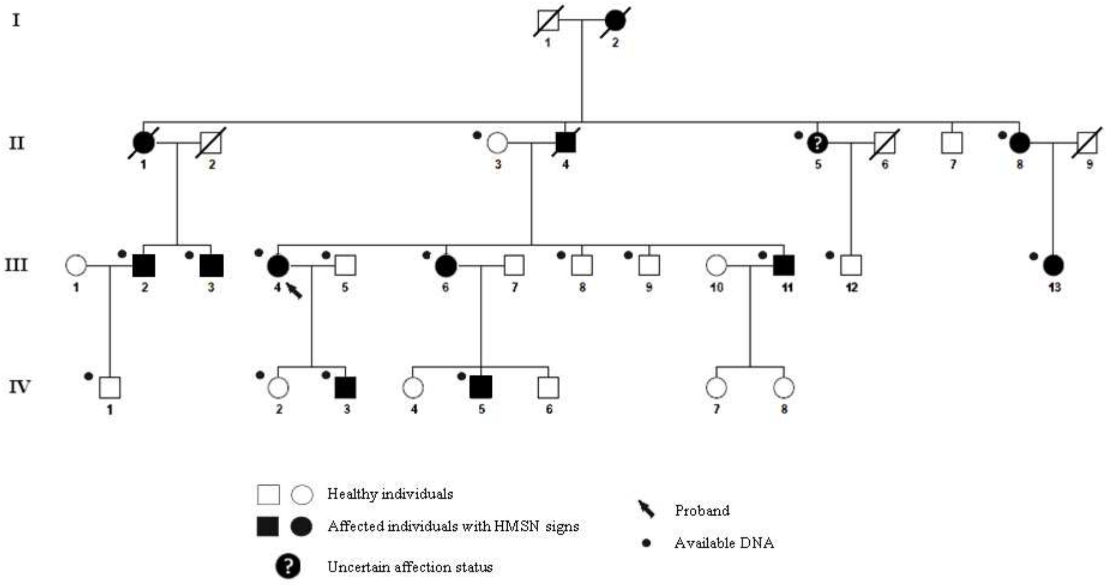


Figure 3.33: Pedigree of Family 4.

3.5.2 Preliminary results

First studies on this family were focused on the disease genes most frequently associated with a similar clinical picture (i.e. *MPZ, PMP22, MFN2, SETX, GJB1* and *SPG4*) and they were excluded by Sanger sequencing. A linkage analysis performed with 368 STRs enabled to highlight two linkage regions on chr1q33.3 and on chr9q33.1-34.11 (LOD score value higher than 2). A second high-density genome-wide linkage analysis with over 200,000 SNP markers (Illumina Human CNV370-Quad platform) allowed to ruled out all genes and *loci* already associated with this phenotype. A linkage signal on chr9q31.33-q31.2 (LOD score value of 2.75) was detected; a lower peak was even obtained on chr5q33.1-q35.2 (LOD score value of 0.75) which co-segregated in all affected subjects and was shared by also 4 unaffected individuals (II-7, III-8, IV-1 and IV-2). Haplotype reconstruction confirmed and better refined the region on chr9q31.1-q31.2 of 2.98 Mb (rs2049347; rs7875152) co-segregating with the disease. Moreover, CNV analysis did not identify candidate deletions/duplications spanning more than 10 kb.

3.5.3 Re-evaluation of the linkage data

In addition to the linkage regions identified by previous studies, a less stringent linkage analysis with only affected subjects (“affected only” approach) was performed to overcome cases of incomplete penetrance and later onset. It broadened the region on chr9q31.1-q31.2 and highlighted a novel candidate region on chr5q33.1-q35.2, explaining together a digenic model. In addition, considering the clinical heterogeneity, a model accounting for two distinct genetic causes for HMSN and HSP was used. A genome-wide linkage analysis was thus carried out by stratifying patients displaying predominant HMSN or HSP clinical features. For the HMSN phenotype (individuals III-4, III-11, III-13, IV-3) linkage signals on chr1q21.3-q31.1, chr4q28.3-q35.1, chr8p21.2-q13.3 and chr9q21.13-q34.3 were identified; conversely, for the HSP phenotype (individuals III-2, III-3, III-4 and III-6) linkage peaks on chr1p36.32-p36.31, chr1p32.1-p31.3, chr2q36.3-q37.3, chr4q22.1-q26, chr5p13.2, chr9q31.1-q33.2 and chr21q22.11-q22.12 were obtained.

3.5.4 Whole Exome Sequencing

Patients III-2, III-4 and III-6 were chosen for the whole-exome sequencing in order to identify disease-causing mutations among the shared variants in all candidate regions.

The performance of WES technology appeared similar to that of previous families (Table 12, Appendix E). It is interesting to note that the mean read depth of 65X is lower than in other families (77X and 91X in Family 2 and more than 100X in Family 3). Also the coverage of target regions was lower and a read depth more than 20X (required for confident calls of heterozygous variants) was achieved for only 76% of the target. These low values could negatively affect the performance of variant callings in particular for heterozygous variants, which are those expected in patients for a dominant trait.

The final set of high-quality variants obtained by the exome sequencing was similar to other families (about 85,000 high-quality SNVs). The higher number of intergenic variants (more than 12,000 variants in this family versus the 1500 in other families) and the lower number of coding SNVs (about 7200 missense variants *vs* more than 12000 in other families) confirmed the low quality of these WES data.

3.5.5 Variants in disease-genes and *loci* and CNV analysis

As the HMSN type V was not genetically characterized and the clinical classification not well-established, all the known disease-genes and *loci* implicated in HMSN and HSP, as well as all the genes annotated in OMIM database were analysed for the WES variants. Variants in at least one of the 3 patients, in coding exons and close to splice-sites, with a MAF lower than 1% were considered. After the filtering and prioritization, 5 candidate variants were detected in the known disease-genes, but none of them was confirmed by Sanger sequencing even if with a high quality score (Table 3.18). All these false positive calls mapped in homopolymeric stretches and 3 of them were identified in other in-house unrelated exomes confirming error-prone positions. Therefore, no candidate mutations were identified in the known HMSN and HSP genes and in other disease-related OMIM genes. The analysis of the whole dataset of WES variants identified 9 total CNVs (>1.5 kb), but none of them was shared by at least 2 affected subjects (Table 15, Appendix F).

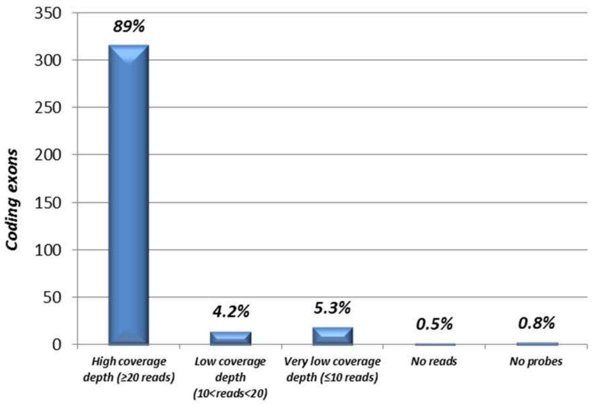
Chr	Candidate Variant	Variant call III-2	Variant call III-4	Variant call III-6	Predicted Effect	Allele frequency	In-house exomes/in stretch	Functional Role (of mutations)	Confirmatio n by Sanger sequencing
1	*MAPKAPK2* c.643A>C	A(ref)43/C17 (Q=0)	A(ref)44/C14 (Q=89)	A(ref)39/C17 (Q=78)	Missense damaging	Novel variant	Yes / yes	HSPB1 target, cause of HMSN	False Positive call
6	*SYNE1* c.20242C>T	C(ref)28/T17 (Q=44)	C(ref)26/G6 (Q=21)	C(ref)36/T10 (Q=0)	Missense damaging	Novel variant	No / yes	cause SCA, HMSN and HSP	False Positive call
7	*hnRNPA2B1* c.1048T>C	T(ref)46/C35 (Q=99)	T(ref)38/C39 (Q=99)	T(ref)48/C34 (Q=92)	Missense damaging	rs117917826, MAF<1%	Yes / yes	cause ALS	False Positive call
9	*FNBP1 c.941T>G*	A(ref)201/C45 (Q=0)	A(ref)185/G35 (Q=0)	A(ref)198/G41 (Q=0)	Missense damaging	Novel variant	Yes / yes	regulation of the actin cytoskeleton	False Positive call
12	*KCNMB4* c.604G>A	G(ref)22/A13 (Q=27)	G(ref)26/A16 (Q=31)	G(ref)18/A9 (Q=10)	Missense damaging	Novel variant	No / yes	cause Dyskinesia	False Positive call

Table 3.18: Filtered SNVs found in genes annotated in OMIM and related to neurodegenerative disorders. Variant calls are expressed with number of reads for each allele (ref=reference allele) and Q=Phred-like quality score of variant call.

3.5.6 Coverage analysis for the best linkage region on chr9q22.33-9q31.2

The accuracy in heterozygous variants calling is necessary to unravel an autosomal dominant trait. The variants at the heterozygous state are more susceptible to technical artefacts and more difficult to identify than homozygous sites, particularly in regions of low coverage.

The coverage in the best candidate region on chr9q22.33-9q31.2 was analyzed using the visualization of read alignments. After the observation of many regions, a read-depth under 20X was arbitrarily evaluated as insufficient to accurately call heterozygous mutations or to assess a false positive call. The analysis was performed for the three patient separately and gave comparable results, reported with mean values in Figure 3.34 and Table 3.19.



	Mean value of 3 patients
Total coding exons	355
High coverage depth (≥20 reads)	316
Low coverage depth (10<reads<20)	15
Very low coverage depth (≤10 reads)	19
No reads	2
No probes	3

Figure 3.34: Depth of coverage per exon, on chr9q22.33-9q31 linkage region. Y-axis indicates the number of coding exons; the correspondent percentage of total is indicated above each histogram bar. Means values of poorly-covered coding exons were calculated considering 3 patients. Table 3.19: Exons for each category of coverage depth for linkage region.

The 21% targeted coding exons displayed a low read depth due to problems of probe design, capture, preferential amplification, low complexity sequence or poor alignments. On the basis of the gene function and pathways in which the encoded protein is involved, the poorly-covered exons were prioritized and selected for Sanger sequencing, in order to identify potential mutations still not detected by the WES (Table 9, Appendix D). From this analysis no significant variants were identified.

3.5.7 Variants in the linkage regions

The exclusion of the known disease-related loci suggested the presence of a novel disease gene. For this reason a similar analysis of the WES variants was performed for the linkage regions. Considering the above-mentioned low quality of the WES data, variants called in at least 2 out of 3 patients were scored. No WES variants were found in the best linkage region on chr9q31.1-q31.2, thus the analysis was extended to the other regions identified by the “affected only” linkage analysis. The analysis highlighted 7 candidate variants (Table 3.20).

Chr	**Candidate Variant**	**Variant call III-2**	**Variant call III-4**	**Variant call III-6**	**Predicted Effect**	**Allele frequency**	**In-house exomes/in stretch**	**Confirmation by Sanger sequencing**
9	*RNF20* c.2854A>C	A(ref)41/C18 (Q=0)	A(ref)33/C26 (Q=26)	A(ref)33/C12 (Q=59)	Missense damaging	Novel variant	No / Yes	False Positive call
9	*SMC2* c.1583C>T	C(ref)37/T20 (Q=99)	C(ref)33/T12 (Q=37)	C(ref)26/T15 (Q=98)	Missense damaging	rs201083116, No data	Yes / Yes	False Positive call
9	*SVEP1* c.4781C>A	G(ref)158/T35 (Q=0)	G(ref)158/T35 (Q=0)	G(ref)158/T35 (Q=0)	Missense tolerated	Novel variant	Yes / No	False Positive call
9	*HSDL2* c.130G>C	G(ref)21/C11 (Q=77)	G(ref)38/C10 (Q=29)	G(ref)49/C14 (Q=78)	Missense damaging	rs184202621, MAF(A)<0.2%	Yes / Yes	False Positive call
9	*MEGF9* c.380C>T	G(ref)1/A2 (Q=19)	G(ref)3/A3 (Q=31)	G(ref)1/T0 (Q=2)	Missense tolerated	rs149898889, MAF(A)= 0.96%	No / No	yes
5	*CLINT1_intron10* g.157218708 T>C	T(ref)124/C90 (Q=99)	T(ref)126/C117 (Q=99)	T(ref)124/C90 (Q=99)	Splice Site activation	rs14385761, MAF=0.09%	No / No	yes
5	*RANBP17* c.2548G>A	G(ref)74/A81 (Q=99)	G(ref)55/A56 (Q=99)	G(ref)82/A56 (Q=99)	Missense tolerated	rs14366219, MAF=0.1%	No / No	yes

Table 3.20: Filtered SNVs found in genes of the 9q22.33-9q31.2 and chr5q33.1-q35.2 candidate regions. Variant calls are expressed with number of reads for each allele (ref=reference allele) and Q=Phred-like quality score of variant call.

The 3 variants confirmed by Sanger sequencing were not identified in other in-house exomes or in homopolymeric stretches. Even if their functional role could be interesting for the pathogenesis of the disease, they were already annotated in variant databases with a low frequency and did not displayed a correct segregation without assuming non-penetrant subjects. Also the *MEGF9* c.380C>T variant mapping to the best linkage region on chr9q31.1-q31.2 segregated in 3 unaffected subjects (IV-4, IV-6 and IV-8). These data were not able to fully explain the disorder except for the digenic model which relies on the co-occurrence of the two candidate regions chr9q31.1-q31.2 and chr5q33.1-q35.2 exclusively in affected subjects. In addition, further analyses were performed with the stratification of patients according to their prevalent HMSN or HSP phenotype. Also in this case 3 candidate variants were identified and listed in Table 3.21.

Candidate Variant	Predicted Effect	Allele frequency	In-house exomes/in stretch	Phenotype considered for analysis	Linkage region	Segregation of the haplotype	
						Affected	Unaffected
SLC26A7 c.808T>C	Synonymous tolerated	rs373170016 Unknown frequency	No / No	HMSN	chr8p21.2-q13.3	All the affected by HMSN	III-8
c21orf67 c.579C>A	Stop gaining tolerated	Novel variant	No / No	HMSN	chr21q22.11-q22.12	All the affected by HMSN except for IV-5	III-8
FARP2 c.2530A>T	Missense tolerated	rs150786024, MAF= T: 0.15%	No / No	HSP	chr2q36.3-q37.6	All affected by HSP and IV-3 (HMSN)	III-8, IV-1, IV-2

Table 3.21: Candidate SNVs confirmed by Sanger sequencing and correct segregation in patients with HMSN or HSP phenotype.

Despite the *FARP2* variant with a partial co-segregation with the disease, the variant displaying the best segregation was the synonymous substitution (c.808T>C) in the *SLC26A7* gene. Several tools displayed notable difficulties to reliably predict the effect of a synonymous variant, however the *SLC26A7* gene characterized by a renal expression seemed unlikely to be the primary cause of a neurological clinical phenotype. On the other hand, the c.579C>A substitution in *C21orf67* was interesting because of the stop codon introduction. The interpretation of a possible effect was hindered by the lack of information on this gene that is annotated as long non-coding RNA. Therefore, also these variants or a combination of them were not able to fully explain the affection status without assuming non-penetrant individuals.

In conclusion, the analysis of Family 3 has identified a pool of variants within genes displaying functions of interest. All variants are unable to explain the disease without assuming non-penetrant subjects or the involvement of a second genetic cause (such as the digenic model). Except for one candidate variant mapping to a long non-coding RNA, other variants are low-frequency polymorphisms. In order to unravel the genetic cause in this family, further evidences are required. For this reason other 2 patients have been sequenced with high-coverage WES and the analysis of their data is currently under way.

4. DISCUSSION AND CONCLUSIONS

The present study was focused on searching novel causal genes in different forms of hereditary peripheral neuropathies or HMSN (Hereditary Motor Sensory Neuropathy). The growing genetic heterogeneity of these disorders is shown by more than 50 genes identified up to now. Although several efforts have been recently made by genetic studies, many cases still remain “orphan”. For instance, in more than 80% distal HMN patients which undergo molecular diagnosis, the causal mutation is not identified (Rossor *et al.*, 2012). In the four families of the present work, a previous exclusion of the known disease-genes associated with the peripheral neuropathies failed to detect causative mutations, indicating a further genetic heterogeneity.

This already complex background is further complicated by the phenotypic heterogeneity of the disease. Indeed, clinical pictures associated with these neuropathies display variable symptoms, age of onset, disease course, severity and expression even in the same family (Brusse *et al.*, 2009). The four families object of this study are all affected by peripheral neuropathy, but different additional signs and modes of inheritance are described. Considering these dissimilarities, the study of each family was conducted separately by taking into account the specific peculiarities and developing, whenever required, more sophisticated strategies of study. Indeed, in a context of clinical and genetic heterogeneity, unravelling the genetic bases can be more challenging respect to the majority of Mendelian disorders with a straightforward genotype-phenotype correlation. However the recessive transmission and the consanguinity loops which characterize three of these families represented a considerable advantage for the study. Indeed, consanguineous pedigrees, even with a low number of individuals, are invaluable to mapping recessively acting disease genes since rare alleles are expected to be overrepresented and with homozygous state in the patients. For these families the approach of homozygosity mapping enabled to successfully identify homozygous segments in patients, but at the same time it highlighted striking unexpected findings that further support the growing genetic heterogeneity which is increasingly being characterised the distal neuropathies and other neurological disorders.

Family 2 represents a very good example of this fact. It displays three consanguineous marriages, suggesting a recessive transmission. The presence of such a high inbreeding rate in the same family apparently seemed unlikely, but became reasonable considering the origin from a closed community located in a little village. Unexpectedly, homozygosity mapping failed to identify an autozygous region shared by all patients suggesting a more complex inheritance model. The presence of two slightly different clinical pictures between patients prompted the existence of two different genetic causes that were effectively identified by studying separately two nuclei of the family.

In the first nucleus, the homozygosity mapping strategy efficiently detected a single autozygous regions of 6 Mb on chr8p23.1-p22 (rs2738148; rs6997599) shared exclusively by the two patients of the I and II generation. On the other hand, the availability of only one patients in the second nucleus (III generation) decreased the robustness of this analysis, identifying 11 homozygous regions with many identity-by-status regions due to high-frequency homozygous SNPs. The combination of this approach with the Identity-By-Descent (IBD) analysis enabled to definitely overcome the pitfall of IBS regions. In this study IBD analysis has turned out the approach of choice to confirm the distant kinship among the patient's parents, relying on rare IBD chromosomal segments shared in single copy. It revealed very powerful as enabled to define an ancestral haplotype (IBD1) which surprisingly matched with one of the 11 homozygous regions inherited in double copy exclusively by the affected child. Therefore, this more specific approach allowed to pinpoint one candidate region on chr9p21.1-p13.2 (rs17775810; rs1022770) amongst the IBS and IBD regions detected by the single-patient homozygosity mapping. Given that more than 120 genes and 95 genes (respectively for nucleus 1 and nucleus 2) mapped in the candidate regions, the Whole Exome Sequencing (WES) approach was applied. The information of candidate regions was absolutely necessary to reduce the number of the sequenced family members and the relative costs, but especially to dramatic reduce the high number of WES variants. Indeed, nearly 100,000 high-confidence variants were markedly narrowed down to less than 250 between SNVs and indels. Starting from this manageable pool of variants, a deeper investigation further supported by manual reviews was more feasible for both low and high-confidence variants. This strategy have been powerful in identifying the two different genetic causes expected by this complex model of segregation. In the first nucleus, the candidate variant was identified in the ***SGK223*** gene (**c.1529T>C**) of the chr8p23.1-p22 candidate region, and it was absent in the variant databases. The hypothesis of its causality in the disease is enforced by the interesting functional role of the PRAGMIN encoded protein. Even if not-well characterized yet, it has been described as regulator of the RhoA protein for the reorganization of cytoskeleton filaments (Tanaka *et al.*, 2006). Interestingly, dysfunctions in cytoskeleton maintenance have already been considered one pathogenic mechanism of neuropathies. Moreover, PRAGMIN seems to plays the same role of FRABIN protein, another Rho

GDP/GTP exchange factor that has been associated with CMT4H (Delague *et al.,* 2007). *In silico* predictions of this study strongly support the effect of this variant in a splicing alteration. Even if to date the *SGK223* transcript is not well-characterized, these findings are indicative of the presence of a natural cryptic splice site that is activated by a developmental regulation (probably after differentiation) with a tissue-specificity. According to the late onset of the phenotype, the effects of this substitution on the strength of the splice site could be evident only after the development, when this splice site becomes activated. Further characterisations of the *SGK223* transcript in normal and mutated conditions will help to verify this hypothesis. Therefore, *SGK223* is strongly candidate to be a novel causal gene for the dHMN and further evidences could confirm this finding.

For the patient of nucleus 2, the analysis of the WES variants mapping to the IBD region effectively detected the candidate mutation (**c.412G>C**) in ***SIGMAR1***, a gene that was already associated with the Juvenile amyotrophic lateral sclerosis 16 (ALS16) (Al-Saif *et al.,* 2011). As expected, Sigma receptor-1 protein plays more than one function in pathways involved in the pathogenic mechanisms of neuropathies. It has been described as an endoplasmic reticulum chaperone and a regulator of Ca^{2+} signaling, with a main expression in motor neurons (Crottes *et al.,* 2013;Pabba 2013;Prause *et al.,* 2013). By the *SIGMAR1* screening, a second variant (**c.448G>A**) was identified in another unrelated case displaying a recessive form of dHMN. Both mutations were novel and excluded from 200 control chromosomes. At a protein level, the two amino acid substitutions (p.Glu138Gln and p.Glu150Phe) seem to modify the chemical properties of the chaperone domain that is exposed to the endoplasmic reticulum (ER) lumen. Even if the involvement in different pathways makes it difficult to link the mutations to a specific role, a mutational effect on the folding/degradation of proteins in the ER is speculated. In line with this hypothesis, ongoing experiments in NSC34 cells will attempt to verify an effect in the unfolded protein response. All these notable findings confirm the causality of the *SIGMAR1* gene in dHMN disorder. This is also supported by the mapping to the “orphan” dHMN Jerash type *locus* (dHMN-J) (Christodoulou *et al.,* 2000). Indeed, a similar form of autosomal recessive dHMN was described by Christodoulou and colleagues in their families.

In literature only one homozygous mutation (p.Glu102Gln) has been reported in *SIGMAR1* (Al-Saif *et al.,* 2011). This is interesting because in spite of the paucity of mutations worldwide, two mutations (p.Glu138Gln and p.Glu150Phe) have been identified in the same Italian region and do not share a common ancestral allele. A possible explanation could be a mutational frequency in *SIGMAR1* higher than expected. The known association with ALS16 could have prevented the screening of *SIGMAR1* in many dHMN cases which still remain unsolved. For this reason we believe that our finding will have strong consequences for the diagnosis of dHMN.

Interestingly, the problem we face is the traditional categorization of diseases. Whether dHMN and ALS16 are considered two distinct disorders, we can state that this study allows to extend the phenotypic spectrum associated with *SIGMAR1* mutations. In regard to this, similar genetic overlaps between dHMN and ALS diseases have been observed for *DCTN1* and *SETX* genes (Pierce *et al.,* 2010). On the other hand, the dichotomous classification has been questioned by other evidences. First, a difficulty in differentiating the diagnosis of ALS from hereditary peripheral neuropathies emerged in several cohort studies (Davenport *et al.,* 1996;Riva *et al.,* 2011). Second, SigmaR1 knock-out mouse model showed more slight motor abnormalities than observed for classical ALS phenotype, more similar to neuropathy (Mavlyutov *et al.,* 2013). Third, also the neuropathy could hamper a differential diagnosis, as it is a common sign in many disorders and not a specific hallmark. Considering these limitations, a classification based on molecular basis could be preferable but even necessary to unravel and clarify a definitive genotype/phenotype correlation. However our data enable to link the *SIGMAR1* gene to the dHMN Jerash type *locus*, thus explaining it for the first time.

Also the study of **Family 1** put in evidence the heterogeneity of the neurodegenerative disorders and the consequent problem of the traditional categorization of these diseases. The analysis of this family, affected by a recessive form of distal neuropathy and spastic paraplegia, is an example of application of the traditional approach. Only one candidate region was highlighted by the genome-wide linkage analysis on chr13q12.1-q12.12, and only 5 genes map to this interval. These findings allowed to use the traditional Sanger sequencing of the coding exons to identify the variants in this region. A homozygous missense mutation (**c.11104A>G**, p.Thr3702Ala), co-segregating with the disease in the patients was identified in the ***SACS*** gene. The mutation was absent in all variant databases and was not detected in 350 healthy controls. Interestingly, *SACS* is responsible for the autosomal recessive spastic ataxia of Charlevoix-Saguenay (ARSACS) and more than 100 different missense, nonsense and frameshift mutations have been reported since 2000 up to now. (Engert *et al.,* 2000;Thiffault *et al.,* 2013). The identified substitution p.Thr3702Ala is the first amino acid change affecting the XPCB domain of sacsin protein, which was reported to interact with the E3 ubiquitin ligase UBE3A (Greer *et al.,* 2010). In the current study, *in silico* predictions showed that p.Thr3702Ala change falls in a surface-exposed and evolutionarily conserved region, suggesting a key role in UBE3A binding and the consequent ubiquitination of proteins. Indeed, even if sacsin has not been well characterized because of its large size, a role in the protein quality control has already been reported for sacsin (Kozlov *et al.,* 2011;Parfitt *et al.,* 2009). It is interesting to note that the same function was reported for LRSAM1 and TRIM2 E3 ubiquitin ligases responsible for HMSN (HMSN2P and HMSN2R respectively) and for the E3 interactor spartin which causes HSP (Bucci *et al.,* 2012;Finsterer *et al.,* 2012).

Recently, Girard and colleagues described an imbalance between mitochondrial fusion and fission in fibroblasts of patients homozygous for the c.8844delT *SACS* mutation (Girard *et al.,* 2012). In the present research such impairment was assessed for the mutation identified, and a higher number of significantly more little and spherical mitochondria was observed in the proband's fibroblasts. These findings were indicative of a higher fragmentation of the mitochondria in presence of the p.Thr3702Ala sacsin mutation, in agreement with the hypothesis of impairment in the fusion/fission mitochondrial dynamics. Previous studies for mutations in *MFN2* complained several limitations in studying mutational effects in patient's fibroblasts (Amiott *et al.,* 2008). This aspect could explain the reason why the differences here observed between mutated and control fibroblasts are significant but not so striking. Nevertheless a deeper investigation is required in order to confirm and lend further weight to the hypothesis supported by these data. A better model for studying neuron mechanisms could be obtained by reprogramming these fibroblasts into neurons (Vierbuchen *et al.,* 2010). Considering that the current knowledge on sacsin is related to the protein quality control, the p.Thr3702Ala mutation could impair the mitochondrial proteostasis specifically through UBE3A binding, thus affecting the mitochondrial functionality and consequently causing the morphological changes here observed. In line with this speculation, a study revealed smaller mitochondria and "dense spheroids" associated with UBE3A deficiency (Su *et al.,* 2011). Mitochondrial impairment has been also reported in HMSN patients with *MFN2* and *GDAP1* mutations, whereas the paraplegin and HSP60 which are mutated in HSP, participate to the mitochondrial protein-quality control as hypothesized for sacsin (Timmerman *et al.,* 2013). Therefore, further studies on the mitochondrial functionality (by checking the membrane potential or oxidative damages) might explain whether this finding is solely a secondary effect of the mutation.

The results obtained in this family represent the first evidence for a *SACS* mutation associated with a non-ataxic clinical picture (Gregianin *et al.,* 2013). Indeed, ARSACS is mainly characterized by early-onset spastic ataxia (i.e. uncoordination of voluntary movements) and pyramidal tract signs. Even if a clinical variability have been increasingly described, the cerebellar ataxia and typical features in the magnetic resonance imaging (MRI) are fundamental hallmarks for the diagnosis of ARSACS (Masciullo *et al.,* 2012;Pyle *et al.,* 2012;Shimazaki *et al.,* 2007;Synofzik *et al.,* 2013;Takiyama 2006). Interestingly, the patients in this family did not show overt evidence of cerebellar abnormalities in MRI. This novel and atypical non-ataxic phenotype may be effectively explained by the first substitution affecting the XPCB domain. Indeed, ARSACS phenotype has been mainly defined on the basis of cases in Quebec region (Canada), where a founder effect for truncating mutations has increased the prevalence of a severe and relatively homogenous disease phenotype. On the other hand, a growing phenotypic variability has been recently demonstrated in other populations, mainly associated with missense mutations which cause only a partial loss of function (Thiffault *et al.,* 2013). This could explain the later onset in this family compared with the early childhood onset of ARSACS. In agreement with this suggestion, other cases displaying a later onset have been reported for *SACS* missense mutations (Baets *et al.,* 2010).

In conclusion, our findings expand the genetic spectrum of *SACS* mutations and further broaden the

ARSACS phenotype described up to now. In line with a growing number of reports, the increasing clinical variability supports the hypothesis that *SACS* mutations are more common than previously known and that they might be underdetected (Synofzik *et al.,* 2013). Accordingly, the *SACS* screening has to be considered in still-unsolved cases with unusual clinical presentations such as late onset, prominent peripheral neuropathy and spastic paraplegia. A not-biased approach of whole-exome sequencing could be useful to find these cases. Considering the limitations of the traditional classification, ARSACS disease could be better indicated by the term “sacsinopathy” to specify all diseases caused by *SACS* mutations (Takiyama 2006). A disease classification based on molecular evidences is becoming increasingly necessary to unravel and clarify a definitive genotype/phenotype correlation. Moreover, as for other not-well characterized proteins, mutation discoveries like this one represents a stimulus for shedding light on the sacsin role which is still poorly characterized.

The approach of homozygosity mapping/linkage analysis combined with WES enabled to powerfully shed light on the genetic bases of the third consanguineous family. The **Family 3**, affected by a recessive complex form of neuropathy with spastic paraplegia and mental retardation, was investigated by robust homozygosity mapping/linkage analysis because of the clear recessive monogenic model and the high number of genotyped individuals (3 affected and 8 unaffected family members). Despite the high informativeness of the family, 4 candidate regions on chr2p13.3-p12, chr3q27-q28 (centromeric and telomeric regions) and chr21q11.2-q21.1 were surprisingly identified. Amongst the 94 genes mapping to these regions, the analysis of WES variants enabled to identify in chr2p13.3-p12 region the best candidate variant in the ***FBXO41*** gene. The missense variant **c.950G>A** has never been reported in variant databases and was excluded from 400 control chromosomes. FBXO41 protein has been described with a role in the E3 ubiquitin ligase complex SCF and probably in cholesterol transport, with a main expression in the nervous system (Jin *et al.,* 2004). Several data in literature are in agreement with a putative pathogenic role of this variant. First, the ubiquitin proteasome pathway is a mechanism already associated with the pathogenesis of HMSN, with mutations in the *TRIM2* and *LRSAM1* genes encoding E3 ubiquitin ligases (Guernsey *et al.,* 2010;Ylikallio *et al.,* 2013). Second, the paralogs *FBXO48, FBXO7* and *FBXO38* have been already associated with clinical features found in Family 3, such as neuropsychological signs (Parkinson Disease), spastic paraplegia and neuropathy (Shojaee *et al.,* 2008;Sumner *et al.,* 2013). Third, spartin protein (SPG20) that has already been associated with a very similar complex picture, presents the two same functional roles of FBXO41 (Eastman *et al.,* 2009). Concluding, even if the underlying mechanism remains unclear and functional studies on FBXO41 will be required to gain further information, at present the *FBXO41* gene appears the best candidate disease-causing gene in Family 3. In order to shed light on many of these aspects, a mutation screening of *FBXO41* gene is currently under way in unrelated patients. At this time, *FBXO41* is strongly candidate to be a novel causal gene for this disease and further evidences could confirm the findings of the current study.

As reported in literature, the majority of the novel genes identified by the WES approach underlies recessive diseases, whereas dominant disorders are more difficulty to unravel to date (Gilissen *et al.,* 2012). Indeed, data obtained in the current study for **Family 4**, which is affected by an autosomal dominant form of HMSN and/or spastic paraplegia, have been insufficient to pinpoint a candidate genetic cause. Alternative genetic hypotheses such as the digenic model and the co-occurrence of two independent causes were also proposed in order to explain the intrafamilial clinical heterogenity, with patients exhibiting either a prevalence of peripheral neuropathy or pyramidal signs or both of them. Linkage analysis enabled to detect a high number of candidate regions, with the best candidate one on chr9q22.33-9q33.2 (rs2297602; rs1041356) that was defined by the analysis with only surely affected individuals and the integration with microsatellite genotypes. WES variants were inspected in all these regions but no variants displaying a correct segregation without assuming at least one non-penetrant subject were selected. Also the CNV analysis and the direct sequencing of the poorly-covered exons in the chr9q22.33-9q33.2 region did not identify any other significant variant. However, through this study the weaknesses of the WES technique came to light. Firstly, technical artefacts created by the sequencing and read alignment steps were highlighted. The presence of allelic variants in homopolymer nucleotide stretches and in other in-house unrelated exomes proved to be good selection criteria to filter them out. Also the presence of multiple-mapping reads due to repetitive regions in paralogs and pseudogenes were indicative of spurious variant calls. For instance, in all exomes a huge number of variants in olfactory and taste receptor gene families was found. Second, the automatic thresholds that are commonly used by the algorithms of aligners and variant callers to select high-quality variants, may instead produce unfavourable results. Only a manual review of the allele-specific read depths allowed to unmask missed variants and miscalled heterozygous calls. As expected, the heterozygous calls were more error-prone than homozygous ones due to misinterpreted background noise, and strong evidences were observed in this family. Indeed, a high number of variants called in poorly-covered regions failed to be confirmed by Sanger sequencing. Third, the coverage had a systematic impact on the sensitivity and specificity of the technique. The visualisation of the depth of coverage in the critical intervals revealed a similar trend among different samples, but an inter-exonic and intra-exonic uneven distribution. The sequences characterized by extremely low (<40%) or high (>60%) GC contents appeared the most affected in capture and sequencing yields. Therefore, Sanger sequencing of a high number of poorly-covered exons was necessary to detect missed variants and exclude homozygous deletions. Only recent improvements in the accuracy have been able to bridge this gap. Indeed, thanks to technology advances, a higher quality of WES data was observed in the high-coverage exomes (100X of depth) sequenced at December 2013. Samples sequenced in 2011 (with a mean 75X read-depth) displayed 80% target sequences with more than 20X depth of coverage, whereas in recent WES the value increased to 98%. Unfortunately optimal values of sensitivity and specificity and their relationship with the mean on-target read depth are not well-defined to date. Guidelines and standardization criteria have only recently been taken into account for diagnostic purposes. Unfortunately a better understanding is also hampered by the multiplicity of capture technologies, sequencing platforms and aligners. Moreover longer reads, local realignments and more efficient algorithms should improve the currently limited efficiency of investigating structural variations, chromosome rearrangements and short indels.

In conclusion, this project supplies an original contribution for the study of peripheral neuropathies, for which many cases are still “orphan” today. In detail, the results provide the first evidence that *SACS* mutation can be associated with non-ataxic phenotypes, and that *SIGMAR1* mutations cause a form of distal motor neuropathy. Considering also that the *SIGMAR1* gene maps to the “orphan” dHMN Jerash type *locus* (dHMN-J), this finding supports the causality of this gene in dHMN disorder. Moreover, the *SGK223* and *FBXO41* genes, identified by the combinatorial strategy of linkage and WES approach in families with recessive transmission, are strongly candidate to be novel causal genes for the inherited peripheral neuropathies. This study represents the first step to demonstrate the pathogenicity of a candidate variant and dissect the mechanisms underlying the pathology. Indeed, further genetic screenings in a large cohort of patients and functional studies are currently undergone and will elucidate the actual involvement of the novel candidate genes in these disorders. The demonstration of a growing phenotypic heterogeneity up to an overlap with other neurological disorders can be indicative of a more extended phenotypic spectrum associated with these genes, which could be masked by a neat categorization based on traditional diagnostic criteria and clinical features. Even if a complicated task, the classification of these diseases based on genetics criteria is becoming increasingly useful to unravel and clarify the heterogeneity of these disorders.

5. MATERIALS AND METHODS

5.1 Traditional approach of genome-wide scan

Linkage analysis and haplotype reconstruction

This analysis starts from high-density SNP genotyping obtained by different platforms. For the families 1, 2 and 4 genotyping was carried out with *Illumina Infinium HD HumanCNV370-QuadV3 BeadChip* (average density of 1 SNP per 8 Kb), for the family 3 with the *Affymetrix Mendel Nsp 250K chip* (average density of 1 SNP per 12 Kb).

Linkage analysis was performed for each family by GENEHUNTER-PLUS program (Kruglyak *et al.,* 1996) through the graphical user interface easyLINKAGE PLUS (Lindner and Hoffmann 2005).

The program calculates a parametric multipoint LOD score, which indicates the eventual presence of linkage and considers the probability of co-segregation between disease and allele markers. This analysis takes into account parameters that define the disease model, such as the mode of inheritance, disease allele frequency, penetrance and possibility of phenocopy.

For each region of interest, since the genome-wide analysis uses a limited number of SNPs, haplotype reconstruction allows to consider a higher SNP number and to obtain a more detailed segregation. Moreover EasyLINKAGE provides input files for the HaploPainter program (Thiele and Nurnberg 2005). HaploPainter is used to draw pedigrees and visualise coloured haplotypes of SNP markers with their genetic positions and eventual recombination events. In order to confirm haplotypes of SNP markers, they are manually integrated by microsatellite genotypes collected in previous analyses.

Homozygosity mapping analysis

For consanguineous families with a recessive mode of inheritance and few affected individuals, autozygosity regions was confirmed by the analysis carried out by HomozygosityMapper software (Seelow *et al.,* 2009). Moreover, for a dominant trait the homozygosity mapping can highlight possible hemizygous deleted regions. Starting from SNP genotyping data of more than one individual, HomozygosityMapper screens patients (cases) for blocks of homozygous genotypes in contiguous markers. Homozygosity scores are reported and candidate genomic regions are visualized and compared with genotypes of unaffected individuals, in order to exclude shared homozygous regions. Frequencies of homozygous genotypes for SNP markers are obtained from the CEPH (Centre d'Etude du Polymorphisme Humain) collection of HapMap project. To circumvent the problem of genotyping errors, single heterozygous genotypes with seven or more homozygous genotypes on each side are ignored. A homozygosity score higher than 80% of the maximum value is considered suggestive of the most interesting genomic regions. A manual inspection of genotypes allowed to evaluate genotype-calling errors and better define recombination events of candidate regions.

IBD sharing analysis

The program performs a non-parametric statistical analysis and estimates the probabilities of sharing alleles identical by descent (IBD) between pairs of relatives. Multipoint estimates of IBD sharing at any genomic location are based on the implementation of Lander-Green algorithm (Kruglyak *et al.,*1996). The obtained scores represent the probabilities that two relatives share 0, 1 or 2 IBD alleles (maximum value =1) and a sharing of the same ancestral alleles can be suggestive of a kinship. To prevent bias in the analysis due to linkage disequilibrium (LD), SNPs in strong LD with other SNPs were previously removed and a pruned subset of about 64,000 SNPs in approximate linkage equilibrium was used.

5.2 Whole-exome sequencing approach

Exome sequencing

Exome capture and following sequencing were performed by Beijing Genomics Institute (BGI) in affected individuals. The first sequencing of subjects V-3 and VI-6 of Family 2, IV-2 of Family 3 and III-2, III-4 and III-6 of Family 4 dated 2011; the second sequencing of VII-2 of Family 2, IV-7 of Family 3 dated December 2013.

In this project, *Agilent SureSelect Human All Exon v4 kit* and *Agilent Select Human All Exon v4+UTRs kit* were used for the targeted enrichment of coding regions and all exons respectively.

The first kit contains probes designed with the aim of capturing 51 Mb of the genome by hybridization and to obtain 4 Gb of sequences, the second kit covers 71 Mb of genome with 6 Gb of sequences. Probes are designed for capturing exon sequences, their flanking regions and a pool of non-coding RNAs based on ENCODE, Ensemble (GRCh36 assembly) and Consensus Coding Sequence (CCDS) gene annotation. Two kits are composed by 120-mer biotinylated single strand RNA probes, synthesized by a transcription from a DNA library and with biotin-conjugated uridine-5'-triphosphate incorporation.

Genomic DNA of patients (5 μg) was randomly fragmented by sonication and ligated to standard adaptors. Fragments of 250 bp were selected and amplified by polymerase chain reaction (PCR) with standard primers. This genomic library was hybridized in liquid-phase with the biotinylated RNA probes, which in turn were captured by streptavidin-coated magnetic beads to purify targets from nonspecific fragments. Captured fragments were quantified and sequenced by Illumina Hiseq2000 platform. Illumina (Solexa) NGS platform relies on bridge PCR amplification, where single strand templates bind a solid support and form a “bridge” structure by means of their standard adaptors which recognize complementary primers. Each single template is amplified to obtain a cluster that will be sequenced with reversible terminators and cycles of single-base extension. After each incorporation of a single nucleotide, fluorescence imaging allows the base calling. In this project sequences were generated as 90 bp reads and from both extremities of each library fragment (end-paired reads).

Output files from the exome sequencing

The creation and managing of output files was entirely performed by the BGI. The Genomic Institute developed a pipeline of analysis and provided files for each sequenced individual. In the last steps data from different family members were pooled together.

The raw data obtained by the sequencer were processed by Illumina base-calling Software 1.7, which converts fluorescence signals into base calls. All reads were annotated in FASTQ format file (*. fq) with their identifier, sequence and single-nucleotide quality scores.

1	@FC81EBGABXX:7:1101:1359:2102#GCCAATAT/1 (or 2)
2	ATCCTGAAGACTTCTAGAGAGCTATCCACTTCCCCATGTAATCCCATAGTTCGCCAGAAAGAATCTGACTTACAAATGAC ATCTGCAGCC
3	+
4	gggggggggegggggdgfgeeggggggegggggggggee[gcggggegegcge^egggbgbdccUedccSbddfa`efaeegggfc^^dec

Table 5.1: Example of *.fq file. Row 1: read identifier; row 2: read sequence; row 3: eventual annotations; row 4: string of single-nucleotide quality scores, expressed in ASCII character and corresponding to a probability of error for single base calls.

Reads were aligned against the existing UCSC reference human genome by SOAPaligner/SOAP2 (Short Oligonucleotide Analysis Package) program created by the BGI (http://soap.genomics.org.cn/). This software calculated the best alignment with maximum 3 mismatches per read and the parameters -a -b -D -o -u -p -2 -m -x -s 40 -l 35 -v 3. The used reference genome was based on March 2006 human reference sequence (*hg18*) (NCBI build 36.3 assembly) from University of California, Santa Cruz (UCSC). The *.soap output file contains for each read information about the mapping position and nucleotide mismatches.

1	2	3	4	5	6	7	8	9	10	11	12	13
73	ATAAATCTGAATGGCATATTGTACA TGTAGAAATTTGTGGTGTTCACTTT TTATCAGAACTAAATCCGGATTCTG GGGCTGGAGGACAGA	gggggggggggggggggggggeegfff effeegggggeggegggggggggggeg gefgfdfebeeggfdcgggfggfgegdc W_d[dcd\	1	a	90	+	chr12	108946177	0		90M	90
96	GGGACAATGAACTATTCATTAAAA AAAATACGCTTAGATACCTACCTCA CACCATATACAAAAATCATTTCCAG ATGGATTAAAGAGCTC	SUQSSWWWSVYYYYY]_ccZY YXXXYc_[cc_____ccccc^^^^^ __________cXXcccccccccccccc ccccccc_[[[]]YWXUB	1	b	90	-	chr15	86229884	1	A->89C21	90M	89 A

Table 5.2: Example of *.soap file: Column 1: read identifier; 2: read sequence; 3: string of single-nucleotide quality scores; 4: number of read mapping in the genome; 5: forward “a” or reverse “b” sequence of pair-end reads; 6: nucleotide lenght; 7-9: strand and chromosomal position; 10: number of mismatches; 11: nucleotide and position of mismatch (reference allele >alternative sequenced allele, quality score); 12: matches number; 13: consecutive matches number.

The consensus sequence was obtained by SOAPsnp (BGI), a program that calls consensus genotypes and measures their call accuracy. The following parameters were set: -i -d -o -r 0.0005 -e 0.001 -u -L 150 -T -s -2. SOAPsnp takes into account different scores indicating data quality, alignment and recurring experimental errors, and it converts them into a single Phred-like quality score. The *.cns file indicates for each position the consensus diploid genotype and the more likely nucleotides, based on Bayes’ theorem.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
chr1	19850	T	T	99	T	36	67	67	C	0	0	0	67	1	255	0
chr1	19851	G	G	99	G	39	29	35	A	0	0	0	35	1	255	0

Table 5.3: Example of *.cns file. Column 1 and 2: chromosomal position; 3: reference allele; 4: consensus genotype of the sample; 5: Phred-like quality score of the genotype; 6-9: more representative allele, quality score, uniquely mapping reads number and total reads; 10-13: second more representative allele, quality score, uniquely mapping reads number and total reads; 14: total reads or read-depth; 15: p-value; 16: copy number of adjacent regions; 17: annotation in dbSNP database, “1” if annotated, “0” if absent.

All genomic positions where the consensus genotype differed from the reference sequence were annotated in the *.snp file as candidate variants.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
chr1	25621	C	T	1	C	0	0	129	T	0	0	14	151	1	2,81457	1	33138
chr1	58759	G	A	3	A	38	2	6	G	0	0	0	6	1	2,16667	1	167
chr1	58926	T	Y	0	C	39	1	46	T	0	0	217	263	0	2,61217	0	167

Table 5.4: Example of *.snp file. Column 1 and 2: chromosomal position; 3: reference allele; 4: consensus genotype of the sample; 5: Phred-like quality score of the genotype; 6-9: more representative allele, quality score, uniquely mapping reads number and total reads; 10-13: second more representative allele, quality score, uniquely mapping reads number and total reads; 14: total reads or read-depth; 15: p-value; 16: copy number of adjacent regions; 17: annotation in dbSNP database, “1” if annotated, “0” if absent; 18: distance from the nearest variant (bp). The distance from the nearest variant allows to detect regions prone to high error rates.

In order to discard false positive calls, filtering criteria were applied. For each variant call, 5 bp minimal distance from another variant, ≤2 estimated copy number of a nearby region, ≥20 Phred-like quality score and 4X read-depth were considered. Resulting high-confidence variants were reported in *.snp.filter file, which shows the same structure of *.snp file.
Further information was added in *.snp.filter.gff file, such as rs identifier for already reported variants or snp number for new variants, gene name and other annotations.

1	2	3	4	5	6	7	8	9
chr1	SOAPsnp	SNP	743132	743132	46	+	1	ID=rs3115861; status=dbSNP; ref=C; alleles=G/G; support=17/17; name=FAM87B; geneID=ENSG00000177757; mutType=Hom; transcriptID=ENST00000326734; mRNAtoChr='+'; exonNum=1; mRNA_pos=380; codonNum=127; codonChange='ACT=>AGT'; residueChange='T=>S'; function=missense;
chr1	SOAPsnp	SNP	798791	798791	42	+	.	ID=rs11240780; status=dbSNP; ref=C; alleles=T/T; support=247/247; name=AL669831.13-2; geneID=ENSG00000209354; mutType=Hom; transcriptID=ENST00000386619; mRNAtoChr='-'; exonNum=1; function=5-UTR;

Table 5.5: Example of *.snp.filter.gff. file. Column 1, 4, 5 and 7: chromosomal position and strand; 2: variant-caller program; 3, 6: variant type and Phred-like quality score; 8: codon position; 9: SNP identifier, status, reference allele, genotype, mapped reads, homozygosity or heterozygosity, transcript, exon and mRNA position, codon and residue change and function.

Short insertions and deletions (*indels*) were identified by Genome Analysis ToolKit software (GATK, Broad Institute) and collected in *.indel.gff file.

1	2	3	4	5	6	7	8	9
chr1	GATK	Indel	1148397	1148397	5	+	.	ID=rs59317408; status=-8; indelType=+2; Base=AC; mutType=Het; name=SDF4; geneID=51150; transcriptID=CCDS12.1; mRNAtoChr='-'; exonNum=3,4; function=intron;
chr1	GATK	Indel	1637753	1637753	87	+	1	ID=rs35961525; status=-8; indelType=+6; Base=TTTCTT; mutType=Het; name=CDK11B; geneID=*; transcriptID=NM_033489; mRNAtoChr='-'; exonNum=6; mRNA_pos=278; codonNum=93; function=cds-Indel;

Table 5.6: Example of *.indel.gff file. Column 1, 4, 5 and 7: chromosomal position and strand; 2: variant-caller program; 3, 6: variant type and read-depth; 8: codon position; 9: more details such as variant identifier, status, type, genotype, homozygosity or heterozygosity, transcript, exon and mRNA position, codon and residue change and function.

The high-confidence indels and single nucleotide variants were filtered out whether annotated with a MAF (minor allele frequency) ≥0.5% in the following variant databases: dbSNP130 and 132 (different versions depending on more or less recent sequencing), 1000 genome project (pilot 1, 2, 3), eight HapMap exomes, YanHuang project and BGI database. Rare and novel variants were annotated in a *.xls file.
Upon request, a re-alignment was performed by BGI in 2013 and new *.xls file was created. Variants were updated to the assembly GRCh37 (hg19) and information from KEGG database (Kyoto Encyclopedia of Genes and Genomes) and Gene Ontology were added.
Furthermore a *.xls file containing CNV (copy number variants) called by Copy Number Inference From Exome Reads (CoNIFER) program were analyzed (Krumm *et al.*, 2012).

Coverage analysis

Output files obtained by exome sequencing, for which a visualization standard tool is not provided, were displayed in UCSC Genome Browser after modification (http: //genome. ucsc. edu/). *.soap, *.snp, *.snp.filter, *.indels files were converted to *.bed files and displayed as custom tracks of a work session. Moreover, a file with sequences of Agilent probes was added. By a manual investigation, these data were analysed in a context of already existing information of the database.

The MapView tool was also used for the navigation through the alignment since *.soap file was converted to a compressed MVF binary file (Bao *et al.*, 2009). MapView allows the alignment visualisation of only uniquely mapped reads and their sequences could add further information to Genome Browser visualisation.

For each linkage region, exon coverage was analyzed and a table with the following annotations was created (Table 5.7). Read-depth categories were arbitrarily established in order to unequivocally evaluate variant calls. In a family with a suspected recessive trait, a homozygous or two compound heterozygous mutations were expected. A read-depth of 10X was evaluated sufficient to detect them or to pinpoint a false positive call. On the other hand, in a dominant disease a heterozygous mutation was expected and a higher read depth was required (20X).

GENE of LINKAGE REGION	**Exon**	**UTR/ coding**	**no probe (only for coding Exons)**	**with probe**		
				no reads	**low coverage (<10 reads)**	**high coverage (≥10 reads)**

GENE of LINKAGE REGION	**Exon**	**UTR/ coding**	**no probe (only for coding Exons)**	**with probe**			
				no reads	**Very low coverage ≤10 reads**	**low coverage (10<reads<20)**	**high coverage (≥20 reads)**

Table 5.7: Tables headers used in coverage analyses for a recessive and dominant disease respectively.

Through these tables, technical problems due to probes design, capture or sequencing steps were highlighted, in order to better understand technical limits and to cover low-coverage target regions by direct sequencing.

Analysis of the WES variants

Single nucleotide variants (SNVs) and indels were analysed first for linkage regions and genes listed in OMIM database (OMIM and “review” genes). A detailed analysis was performed starting from *.snp.filter and *.indels files. When required, a deeper investigation of *.snp files was carried out. In order to rule out all other variants in the whole-exome, final *.xls files containing all rare and novel variants were also investigated.

The nearly 100,000 total variants in *.snp.filter and *.indel.gff files were reduced to manageable pools by filtering out all variants which clearly represented technical errors based on the presence of homopolymeric nucleotide sequences flanking the variant, multiple-mapping reads, the occurrence in other non-correlated in-house exomes, the evaluation of the number of reads representing each allele in a variant call. Limits of variant detection was overcome by considering calls in at least one individual and evaluating whether the read depth in the not-sharing patient could be sufficient to exclude its presence.

The prevalence and the late-onset of the disease can lead to the annotation of causative mutations in variant databases, thus a threshold of minor allele frequency (MAF) of 1% or 2% was used to filter out polymorphisms from dbSNP137, dbSNP138, NHLBI Exome

Variant Server and 1000 Genome Browser. In addition, the HapMap and Human Genome Diversity Project allele tracks in NCBI Genome Bowser were visualised. The false positive rate which could characterize a single database was thus minimized by the integration of allele frequency data annotated in different databases. Moreover, information on the validation status and number of submissions in dbSNP database was indicative of a rare or false positive variant.

Variants were analyzed also by functional predictors, in order to prioritize them on the basis of a putative pathogenic effect. The type of amino acid substitution, the localization in known or predicted functional domains, the high sequence conservation were evaluated. The potential pathogenic effect was predicted using PolyPhen2 (http://genetics.bwh.harvard.edu/pph2) MutationTaster (http://www.mutationtaster.org) and SIFT (http://sift.jcvi.org) programs and likelihood ratio test LRT (Chun and Fay 2009). The evolutionary conservation of sequence was calculated by by phyloP (Pollard *et al.,* 2010), GERP (Davydov *et al.,* 2010), phastCons (Siepel *et al.,* 2005) and SiPhy tools (Garber *et al.,* 2009). Further information about the functional role of the encoded protein was searched in databases such as Gene Card and STRING servers in order to find a link with the pathogenic pathways already known for the disorder (following chapters). Finally, the correct segregation analysed by means of haplotype reconstructions led to filter out other variants.

5.3 Bioinformatic tools

UCSC Genome browser

UCSC Human Genome Browser (http://genome. ucsc. edu/) is an online bioinformatics tool created by the University of California, Santa Cruz (UCSC). Interesting annotated tracks include reference genomic sequence (RefSeq), OMIM genes and updated DNA variations. Useful information for understanding technical yield of high throughput sequencing alignment are displayed by "Hi Seq Depth" and "Mapability" tracks. Genomic regions affected by problems of alignment can be detected by these tracks, which indicate troublesome regions due to repetitive elements and sequence similarities throughout the genome. "GC Percent" track can indicate critical regions for the library enrichment step.

Putatively pathogenic variants are annotated by LOVD, HGMD, UniProt, ClinVar mutation databases or published in literature ("publications" track), while frequent polymorphisms were annotated in dbSNP, 1000 genome Project, HapMap, HGDP allele tracks.

Moreover the user can upload and visualize his own files or retrieve the data of already existing tracks by "Table Browser" application.

Variant databases

dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/) is a free public archive developed by the National Center for Biotechnology Information (NCBI) and the National Human Genome Research Institute (NHGRI). This database includes different classes of polymorphisms, such as SNPs, indels and short tandem repeats. Information on allele frequency and validation status allows to understand whether a submitted variant represents a real polymorphism rather than a rare or false positive variant. Exome sequencing output data (*snp.filter.gff) refer to the build 130 of dbSNP, created in 2009. By more recent releases (dbSNP132, 135, 137 and 138) more details have been added, and SNPs are divided into "Common" (in at least 1% of the population), "Multiple" (mapped in more than one genomic position) and "Flagged" (probably clinically associated).

The false positive rate of dbSNP was minimized by the integration of allele frequency data annotated in other databases, such as **Exome Variant Server** (NHLBI GO Exome

Sequencing Project (ESP), Seattle, WA; http://evs.gs.washington.edu/EVS/). EVS has been developed by National Heart, Lung, and Blood Institute (NHLBI) and currently collects exome variant calls of 6503 samples (13,006 chromosomes) from different projects. Samples have been selected in order to represent controls from African-American and European-American population.

1000 genome browser (http://browser.1000genomes.org/) represents a deep catalog of human genetic variation from 1092 individuals belonging to different populations (phase 1 of the 1000 genome project).

GEnomes Management Application (GEM.app, https://genomics.med.miami.edu/) is a software tool to annotate and visualise datasets of variants from exome sequencing. GEM.app currently contains 1,600 whole exomes belonging to patients affected by 50 different phenotypes.

In silico study of proteins

For not functionally characterized proteins, conserved protein domains were assessed by InterProScan, which allows to scan protein sequence for matches against the InterPro collection of protein signature databases (http://www.ebi.ac.uk/Tools/pfa/iprscan). The search for known and predicted direct and indirect interactions between proteins was performed by STRING database. It currently covers 5214234 proteins from 1133 organisms (Franceschini *et al.,* 2013).

Jpred3 (Cole *et al.,* 2008), Psipred (Jones 1999) and SOPMA (Geourjon and Deleage 1995) prediction tools were used to investigate the effect of a mutation in the secondary structure of the protein. Alterations in the most common post-translational modifications, such as phosphorylation and acetylation were predicted by Phosida, NetPhos and NetAcet servers (Blom *et al.,* 1999;Gnad *et al.,* 2011;Kiemer *et al.,* 2005).

Three-dimensional structure of the full-length protein or a single domain (target) was predicted by homology modelling starting from the amino acid sequence alignment with a crystallized protein (template with a pdb code available). PDB database contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies obtained by means of NMR spectroscopy or X-ray crystallography. The amino acid sequence alignments were obtained by NCBI-BLAST (Altschul *et al.,* 1997) or ClustalW tool (Larkin *et al.,* 2007) whether target and template were already known. For oligomeric proteins the modelling of a single domain is considered more reliable. SWISS-MODEL gave the output file of the best target-template alignment, and HOMER server performed the homology modeling which was visualised by PyMol program (Kiefer *et al.,* 2009) (URL:http://protein.bio.unipd.it/homer/) (URL:http://www.pymol.org). The surface conservation of the structure was evaluated by Consurf server (Ashkenazy *et al.,* 2010). The evolutionary conservation starting from linear sequence was assessed by ConSeq server (Berezin *et al.,* 2004).

A potential regulation of splicing site was predicted by ASSEDA server (Mucaki *et al.,* 2013), ESEfinder (Smith *et al.,* 2006), Human Splicing Finder (Desmet *et al.,* 2009), MaxEntScan (Yeo and Burge 2004), Computational approach for Silencer motifs (Sironi *et al.,* 2004) and FAS-ESS server (Wang *et al.,* 2004).

5.4 Molecular studies

Salting-out genomic DNA extraction

The genomic DNA extraction was performed by using a modified protocol based on the salting-out method described by Miller et al., which relies on the physical-chemical rationale that macromolecules are less soluble at high salt concentrations (Miller *et al.*, 1988).

10 ml of peripheral blood were collected in tubes containing disodium-EDTA and stored at -20°C. After thawing, 40 ml of isotonic N-N solution (NaCl 0.9%, Nonidet 0.1%) were added and leukocytes were pelleted by centrifugation (30 min at 4000 rpm, 4°C). Pellet was rinsed twice in 30 ml of N-N solution and resuspended in 4 ml of hypotonic TEN buffer (Tris-HCl 10mM pH 8, EDTA 2mM, NaCl 400 mM). After the addition of 300 µl of 20% SDS solution, samples were incubated at 80°C for 3 hours under vigorous mixing, in order to lyse nuclei and solubilise proteins. The supernatant was collected after adding 1 ml of saturated NaCl (higher than 6 M) and centrifugation (10 min at 4000 rpm), followed by the addition of an equal volume of chloroform and another centrifugation (10 min at 4000 rpm). The supernatant containing DNA was thus separated from proteins. An equal volume of isopropanol was added and after centrifugation (10 min at 4000 rpm) the DNA was precipitated. 2 washes of pellet were performed with 1-2 ml of ethanol 70%. The pellet with DNA was dried from all ethanol and resuspended in 300-500 µl TE buffer (10 mM Tris-HCl pH 8, 1 mM EDTA).

DNA quantification

DNA concentration was quantified by measuring the optical density (OD) at 260 nm with NanoDrop ND-1000 UV-Vis spectrophotometer (CELBIO). Starting from 1.5 µl of DNA sample, the 260nm/280nm absorbance ratio assesses eventual contaminations by proteins (Absorbance of proteins at 280 nm). Values between 1.8 and 2 are indicative of a sufficiently pure DNA sample. In addition, the 260nm/230nm absorbance ratio assesses eventual contaminations by carbohydrates and phenols, which absorb near 230 nm. Values higher 2 are indicative of a sufficiently pure sample.

Genome amplification with GenomiPhi DNA amplification kit

Whether the genomic DNA was not sufficient or partially degraded for the following analyses, it was amplified with GenomiPhi DNA amplification kit (GE Healthcare). Few nanograms of genomic template were added to the sample buffer (50 mM Tris-HCL pH 8.2, 0.5 M EDTA), containing random hexamer primers. After denaturation of 3 min at 95°C, reaction buffer (dNTPs and buffer) and enzyme mix (Phi29 DNA polymerase and other hexamers) were added. The polymerization reaction was carried out at 30° C for 90 min. The final step of enzyme inactivation was performed at 65°C for 10 min.

Amplified DNA was then purified through ethanol precipitation.

DNA amplification by Polymerase Chain Reaction (PCR) and Sanger sequencing

All exons and adjacent intron regions of candidate disease genes, not well covered target regions and regions where interesting variants map were amplified by PCR reaction. Primers were designed by Primer3 software (http://primer3. wi. mit. edu/) based on the sequence data in UCSC Human Genome Browser and taking into account unspecific annealings predicted by Oligo Analyzer v 1.0.3. tool. Large exons were covered by multiple amplicons.

PCR reactions were carried out in a 15 µl volume, which contains 50 ng of DNA template and the remaining volume of master mix. The master mix includes primer (4 pmoles each), dNTPs (200 µM), PCR buffer (1X), $MgCl_2$ (1.5 mM) and of DNA polymerase (0.8 U). DNA polymerase was either Taq Gold® DNA Polymerase (Applied Biosystems) or Accuprime^{TM} GC-Rich DNA Polymerase (Life Technologies) or Fast Start® Taq DNA Polymerase (Roche) or GoTaq® Polymerase (Promega) or AmpliTaq® 360 DNA Polymerase (Life Technologies) depending on the different amplicon. PCR reaction was performed in a Peltier PTC-200 thermal cycler (MJ Research) using standard or touch-down protocols. Whether a high GC content was present in the template sequence, DMSO or a specific GC buffer was added according to the manufacturer's protocol.

The PCR yield was checked by gel-electrophoresis in 2% w/v agarose gel stained by GelRed (Biotium). DNA bands were visualized under ultraviolet trans-illumination and compared with a known molecular-weight marker. Amplified exons were purified in order to remove unincorporated dNTPs and primers, which might interfere with the following sequencing process. A 5 µl of PCR product was purified by Antarctic Phosphatase (1 U) and Exonuclease I (5 U). Samples were incubated at 37°C for 15 min and 80°C for 15 min for the inactivation of the enzymes. The purified PCR products with 3.2 pmoles of primer were sent to BMR-Genomics (Padua, Italy) for the Sanger sequencing. Sequencing was carried out by ABI3730XL DNA Sequencer using Big Dye dideoxy-terminator biochemistry (Applied Biosystem). Electropherograms obtained in *.abi format were analyzed by SeqManII software (DNASTAR).

Candidate variants identified with the exome sequencing were validated by Sanger sequencing. Unsufficiently covered coding exons which map in the linkage regions were sequenced as well. The most promising disease genes (*SACS* and *SIGMAR1*) were also investigated by direct sequencing in other unrelated patients. Amplicon sizes, primer sequences and condition of amplification are reported in Appendix A, B, C and D.

Test of variants by restriction assay and ARMS PCR

The amplification products can be also digested by restriction endonucleases which specifically recognize reference or alternative allele, thus producing restriction fragments of different lengths. The genotyping of a specific genomic position was alternatively inspected by tri-primer ARMS (Allele Refractory Mutation System) PCR reaction. In a standard PCR master mix, outer and inner primers specific for the amplicon with the variant were favoured by a higher concentration (3 pmoles of each primer) while the outer primer competitor of the inner one was added at a lower concentration (0.3 pmoles).

The presence of the best candidate variants mapping in *SACS*, *FBXO41*, *LPP* and *SIGMAR1* genes was examined in a pool of unrelated healthy subjects in order to study variant frequency in the same population and to exclude a polymorphism. Other candidate variants identified with the exome sequencing were validated by enzyme restriction. All tested variants with primer sequences and conditions are reported in Appendix C.

RNA isolation and Reverse Transcription

Total RNA was extracted by fibroblasts pellet using RNeasy Mini Kit (Qiagen), according to the manufacturer's protocol. The content of RNA of each sample was quantified using Nanodrop 2000 spectrophotometer (Thermo Scientific). RNA was reverse-transcribed in a 20 µL reaction volume consisting of $MgCl_2$ (5 mM), PCR BufferII (1X), RNase inhibitor (1U/µL), Moloney Murine Leukemia Virus Reverse Transcriptase (50 U, Promega), random primers (2.5 µM) and dNTPs (1 mM each). Reaction was performed in the thermal cycler at 37°C for 1 hour and 85°C for 5 sec. 20 µL of RNase-free water was added to each tube at the end of the reaction.

Quantitative real-time PCR (qPCR)

Quantitative real-time polymerase chain reaction was performed in a 10 µl reaction containing 5 µl of KAPA SYBR® FAST qPCR Master Mix 2X (KAPABiosystems) with fluorescent dye SYBR® Green I, 2 µl of the reverse transcription reaction and 600 nM primers, by using Mx3000P™ real-time PCR system (Stratagene). The amplification program consisted of denaturation at 95°C for 3 min, 40 cycles of denaturation at 95°C for 3 sec, annealing 60°C for 30 sec and extension at 72°C for 1 sec. Finally 1 min at 72 °C. Real-time PCR assays were performed in triplicate for each condition. The cycle threshold (Ct) value was measured for *SACS* gene and the Ct value of *GAPDH* gene was used as internal reference for normalization.

Primers specific for *SACS* cDNA that were used are:
Forward (exon9): TTTAAAGGAAGCTGCCCAAA
Reverse (exon10): CCAAACCATCTTAAGCCATGA

Specific primers for *GAPDH* cDNA that were used are:
Forward (exon2): GAAGGTGAAGGTCGGAGTC
Reverse (exon4): GAAGATGGTGATGGGATTTC

5.5 Functional studies on fibroblasts

Fibroblasts culture

Skin fibroblasts were obtained from the skin biopsy of X patient of the family 1.

Patient and control fibroblasts were maintained into flasks (CORNING) and culture medium containing DMEM Glutamax (GIBCO) supplemented with Fetal Bovine Serum 10% (v/v) (Life Technologies), 100U/mL of penicillin and 100µg/mL of streptomycin (Life Technologies). Cell growth were in controlled temperature (37°C) and atmosphere (5% CO2 v/v). When cells reached confluence were detached with trypsin and seeded at a density of 5,000 cells/cm^2 for propagation.

Protein extraction and Western blotting

Cell monolayer was detached by trypsine, pelleted and resuspended in lysis buffer with 1x complete protease inhibitor cocktail (Roche). Supernatant was collected after centrifugation and proteins were quantified using a Bradford colorimetric assay with bovine serum albumin (Sigma). 40µg of proteins were denatured at 95°C for 5 min and loaded in each well of a 4-12% gradient NuPAGE Bis-Tris gel (Life Technologies). Electrophoresis was performed for 5 hours at 200 V with morpholinepropanesulfonic acid (MOPS)-sodium dodecyl sulfate (SDS) running buffer (Life Technologies). Proteins were blotted overnight at 4°C onto a nitrocellulose membrane (Whatman). Membrane was incubated with rabbit anti-sacsin (1:500) and mouse β-actin (1:10,000, Santa Cruz) primary antibodies diluted in BSA 5% Tween20-PBS. Fluorescent immunodetection was carried out using secondary antibodies conjugated to fluorescent dyes (1:10,000, Dako).

Immunostaining of the mitochondria and Imaging

Cells were grown at 70% confluence on glass coverslips and fixed in 4% paraformaldehyde in phosphate buffered saline (PBS) for 15 min. Permeabilisation was performed in 0,2% Triton X-100 in PBS for 5 min. Cells were washed and incubated in blocking solution (10%goat serum, 1% BSA, 0,02% Triton X-100 in PBS) for 45 min. Rabbit polyclonal anti-Tom20 antibody (1:500, Santa Cruz) was used in blocking solution for 2 hours, and AlexaFuor (568 nm) goat anti-rabbit antibody (1:1,000, Life Technologies) in blocking solution for 1 hour. Three washes with PBS were followed by the nuclear staining with 1 µg/ml DAPI (4,6′-diamidino-2-phenylindole) for 2 min.

Cover slips were mounted and images were acquired with a Zeiss LSM510 inverted microscope equipped with a 63X oil-immersion objective (1.4 NA) and zoom 1.5. For each Z-stack about 16 cross-sectional images were taken in increments along the Z-axis to encompass the entire fibroblast volume.

Image processing and 3D reconstruction

Image processing, 3D reconstructions and surface rendering were performed with Surpass mode of Imaris software (Bitplane AG, St. Paul, MN). Each mitochondria were considered one object for the analysis and three-dimensional morphometry measures (area, ellipticity, sphericity and volume) were conducted. To reduce background noise, background fluorescence was subtracted by manual threshold adjustment and invalid objects were manually deleted. Experiment was repeated three times, with a total of 100 patient fibroblasts and 100 control fibroblasts collected.

Statistical analysis of data

Statistical analyses were performed using GraphPad Prism software (GraphPad Prism version 5.0 for Windows, San Diego, CA). In normally distributed data, parametric Student’s t-test was used. In not normally distributed data (confirmed by D’Agostino test), non parametric Mann Whitney test was used. Differences were considered significant for p-values < 0.05.

APPENDIX A

Mutational screening of unrelated index cases

Amplicon	Forward and Reverse Primer sequences 5’-3’	Amplicon lenght (bp)	PCR conditions	DNA Polymerase
EX10_1	F: CCTTCCAGTACTGTGTTATTTGTGA R: TTCTGCAATGGCATCTTCTC	562	STD 60°C 30"30"45" x35	Taq Gold
EX10_2	F: CAAAAACTTGGAGGGTTTGTC R: AAACTGAGGGTGGGAAATAGG	700	TD 70-60°C 30"30"45" x10+25	Taq Gold
EX10_3	F: TCCAAATGTGCTTGAGTGGT R: GCAAAATATGCTGGAATTGATAGTAG	677	TD 70-60°C 30"30"45" x10+25	Taq Gold
EX10_4	F: AAACCTGGAAAAAGCATTAGG R: GGCATGGAACTTTTAGCCATT	695	STD 60°C 30"30"45" x35	Taq Gold
EX10_5	F: AAGCCCCAACACACCAGTT R: CCCAGGATTGGATTTGTCTT	570	TD 70-60°C 30"30"45" x10+25	Taq Gold
EX10_6	F: AAAAAGGGGAGAAGTTGACAAAG R: TGATCTGAAGAATGCAAGATGAC	653	TD 70-60°C 30"30"45" x10+25	Taq Gold
EX10_7	F: ACCAACCCCAGTTTAGCACA R: CCAGGTCCCGTAAGACACTC	687	TD 72-62°C 30"30"45" x10+25	Taq Gold
EX10_8	F: TATCAATGGGTGCTTTGCTG R: TCCCGAGAACTCATCAACTTTT	616	STD 60°C 30"30"45" x35	Taq Gold
EX10_9	F: TGTTGAACTTCCTTCTTCGGTAA R: AACTGTTGCCTTTCCAGTCC	623	STD 60°C 30"30"45" x35	Taq Gold
EX10_10	F: GCTAGAACGTGCAGTGTCAGTAG R: CCCACGGTTTCAAAAAGTTC	681	STD 60°C 30"30"45" x35	Taq Gold
EX10_11	F: GAAGCCTTGATGCAAAATGA R: GGATGCTCTTAATTCTGCTGGT	634	STD 60°C 30"30"45" x35	Taq Gold
EX10_12	F: GACATACCCAGGGAAGTAGCA R: CTGATGCTGGAACAGACGAA	700	STD 60°C 30"30"45" x35	Taq Gold
EX10_13	F: TTTGGATGCAGATTTTAGGACA R: CATTATCATCACGCCACAGG	698	TD 70-60°C 30"30"45" x10+25	Taq Gold
EX10_14	F: TCAGCTCACAAGAACCAAGA R: TGAAACCAATTTCTAAAAGGAGATG	691	STD 60°C 30"30"45" x35	Taq Gold
EX10_15	F: TGATGGCTCTGACTTGCACT R: AAGTCTTTGCGGGATGGA	558	STD 60°C 30"30"45" x35	Taq Gold
EX10_16	F: GAAGCTGCCTTGTCGTCTG R: AAGATGCTTGTGGGGCTCT	741	STD 60°C 30"30"45" x35	Taq Gold
EX10_17	F: TTCCAAATGCCCAGAGTGA R: GAAATGCTTTGCTTGCTTTAGT	686	STD 60°C 30"30"45" x35	Taq Gold
EX10_18	F: GAAACACCTCTTACCAAAAATTGA R: TACCTCTTGATATTGAGGATGAAA	586	STD 60°C 30"30"45" x35	Taq Gold
EX10_19	F: TGAGGGCTAATACAGAAAACTGG R: TGAAATGTGCCAAGTTCTAAAGG	682	TD 71-61°C 30"30"45" x10+25	Taq Gold
EX10_20	F: CGAGGGGTTGCTTTTGTG R: GAACACAACGCTCCAAACTG	617	TD 70-60°C 30"30"45" x10+25	Taq Gold
EX10_21	F: AAATGTTAGTTGATCTCAGCCAGT R: TCATTGGGTCCATAAGCAGA	699	STD 60°C 30"30"45" x35	Taq Gold
EX10_22	F: AGGATGCAATGATATTTACAGGA R: GGTTCTCTGGATTTTTGTCAGG	662	STD 60°C 30"30"45" x35	Taq Gold
EX10_23	F: CAAACATCAGTCCCCCAAAA R: TTCCCCTCACAGCATAGTCA	614	STD 60°C 30"30"45" x35	Taq Gold
EX10_24	F: GGCTAAGACAAGCCAGAGCA R: TCATCCCAATCATTCAAATCC	639	TD 67-57°C 30"30"45" x10+25	Taq Gold

Table 1: Primers and amplification conditions for the *SACS* gene. For the proband of Family 1, full-lenght cDNA from fibroblasts were screened. The mutational screening of unrelated cases displaying a similar clinical phenotype was performed for the most extended and frequently mutated exon10.

Amplicon	Forward and Reverse Primer sequences 5’-3’	Amplicon lenght (bp)	PCR conditions	DNA Polymerase
EX1	F: GAGGAAATGGTTCAACCGAAG R: GTCCTAGGTCCGGGGATG	422	STD 58°C 30"30"30" x35	360 Taq GC buffer
EX2	F: CGGCAGTACGCTGGTGAG R: ATGGCCCAGCCAAACATC	423	TD 68-58°C 30"30"30" x10+25	360 Taq GC buffer
EX3	F: CATGCCCTCCTTTCTGATGT R: CCCAACACACTCCTTTTCCA	267	STD 58°C 30"30"30" x35	FastStart Taq
EX4	F: TCCCCACCCCTAGTTAGTCC R: GCTCCAGCAAGTGGATATGTG	394	STD 58°C 30"30"30" x35	FastStart Taq

Table 2: Primers and amplification conditions for the *SIGMAR1* (NM_005866) gene. The mutational screening of 12 unrelated cases displaying a similar clinical phenotype to the Family 2 was performed.

APPENDIX B

Validation of candidate variants identified by the exome sequencing

Family 2

Chr	Candidate Variant (gene, position, effect)	Forward and Reverse Primer sequences 5’-3’	Amplicon lenght	PCR conditions	DNA Polymerase
Chr8	*SGK223* c.1529T>C missense	F: TGTCAGCCACCATCACAGTC R: CTCTCCTTGGGCTTGCTCTC	214	STD 59°C 30”30”40” x35	FastStart Taq
Chr8	*MTMR7* c.1749T>G missense	F: ATGATTTTTCCCCTGAAATTGTT R: CAGAATCTTCATCGCCTTGTG	220	STD 60°C 30”30”30” x35	FastStart Taq
Chr11	*ABTB2*_intron5 g.34189543A>G (hg19) splice site	F: GTGCCCGATCTGTGGTTT R: AGCTCTGTGGGGCTTCCT	206	STD 60°C 30”30”30” x35	FastStart Taq
Chr16	*PLEKHG4* c.1986G>T missense	F: TTCCGAAAGATGTGGGCTCT R: GGGAGTCAGGACCCAAGACA	442	STD 60°C 30”30”30” x35	FastStart Taq
Chr16	*PLEKHG4* c.1658C>G missense	F: GTCTCTGCAGGAGCAGGTCA R: GGCAGCATGACCAAGAAAAG	290	STD 58°C 30”30”30” x35	FastStart Taq
ChrX	*ABCD1* c.1823G>A missense	F: TGTTGGGGCTTGaACTCCACCGTA R: CACGGCGtTGGTGCATTCATCCAG	489	STD 67°C 30”30”40” x35	FastStart Taq +digestion with HpnI

Table 3: Primers and amplification conditions for validation of candidate variants by Sanger sequencing or enzymatic digestion (when indicated) in Family 2.

Family 3

Chr	Candidate Variant (gene, position, effect)	Forward and Reverse Primer sequences 5’-3’	Amplicon lenght (bp)	PCR conditions	DNA Polymerase
Chr2	*FBXO41* c. 950G>A missense	F: TGAGGCTCTGGGGTAAGAGA R: GTAGAGGAGGGAAGGCAGGT	291	TD 68–58°C 30”30”30”x10+26	Taq Gold +digestion with BstUI
Chr3	*LPP* (NM_001167671.1) c.1142C>T missense	F: TGCAATCGCTGCTTGTTTAG R: TTCCAATTACTAAGGAAGGAAATCA	279	TD 65–55°C, 30”30”30”x10+26	FastStart Taq

Table 4: Primers and amplification conditions for validation of candidate variants by Sanger sequencing or enzymatic digestion (when indicated) in Family 3.

Family 4

Chr	Candidate Variant (gene, position, effect)	Forward and Reverse Primer sequences 5’-3’	Amplicon lenght (bp)	PCR conditions	DNA Polymerase
Chr1	*MAPKAPK2* c.643A>C missense	F: GAGTAGAGGAAGGGAGGAACC R: CCCCAGATAGCCAAGAGGAG	296	STD 57°C 30"30"30" x35	FastStart Taq
Chr2	*FARP2* c.2530A>T missense	F: GCCCTACACAAGGAAGAGCA R: CAATGGGGACCAAGGAAAA	255	STD 60°C 30"30"30" x35	FastStart Taq
Chr4	*PALLD* c.1394G>A missense	F: TGTATCTGTTTTGTCTGAGGTTTGG R: GGGCACTTCTCCTTACTTTTCTG	264	STD 60°C 30"30"30" x35	FastStart Taq
Chr5	*CLINT1*_intron10 g.157151286T>C (hg19) Splice site activating	F: CAGCCAGCGGTAGAACTTG R: GGGGAGAGTATGAGGTAAATGAA	367	STD 59°C 30"30"30" x35	FastStart Taq
Chr5	*RANBP17* c.2548G>A missense	F: AGTGTGTCCCATAACTACTTTGAAT R: CTAACTCCCTGGGCTTCATTA	374	STD 57°C 30"30"30" x35	FastStart Taq
Chr6	*SYNE1* c.24422G>C missense	F: GGGATGAAATGTTTACTTAATGGACT R: CTTTTCTACTGGTGTATGGGTCAG	248	STD 59°C 30"30"30" x35	FastStart Taq
Chr7	*hnRNPA2B1* c.1048A>G missense	F: TGGATGGATATAAAATAGAATCAACTG R: CGATAACCTGAAGCTGTTCTG	396	STD 59°C 40”40”40”x35 and Digestion BseRI	FastStart Taq
Chr8	*SLC26A7* c.808T>C Stop codon abolishing	F: TCTAAGGGCAAGGCTGCT R: CTGGAGTGGAGAGGGACTG	328	STD 59°C 30"30"30" x35	FastStart Taq
Chr9	*SMC2* c.1583C>T missense	F: TTTTGCTGATTCTACTTTCATTTCA R: GAGTCGTTCTCCAGCCACTAA	236	STD 57°C 30"30"30" x35	FastStart Taq
Chr9	*SVEP1* c.4781C>A missense	F: TGAAAGAAGGAAATTCTCACCA R: CTCCACCCCCATCTCTAAAA	260	STD 59°C 30"30"30" x35	Taq Gold
Chr9	*MEGF9* c.380C>T missense	F: GACTGCTGGACCCTCTTCCA R: GGTACGGTGGTCGCTACAGG	174	STD 59°C 30"30"30" x35	Taq Gold + digestion with MnlI
Chr9	*HSDL2* c.130G>C missense	F: GGTCGTATTTGTTCTGTTGTAGGA R: AAACTCTCCAGCAAATTCTCTATGT	260	STD 59°C 30"30"30" x35	Taq Gold
Chr9	*CTNNAL1*_intron18 g. 111705122G>T (hg19) Splice site activating	F: GCTCTCTTAGTCCAACTTCTTTCA R: TTAAAAATGTCAGCTTCATCACA	295	STD 58°C 30"30"30" x35	FastStart Taq
Chr9	*RNF20* c.2854A>C missense	F: TGGGTTTTCTTCTTGGTTGTG R: GCTCCTCTTCTTCTCTTGACTTAGA	292	STD 58°C 30"30"30" x35	FastStart Taq
Chr9	*FNBP1* c.1646A>C missense	F: GCCGCTCTAACTCTAACCCA R: GACCACCAAAAGCAACCATC	390	STD 60°C 30"30"30" x35	FastStart Taq
Chr11	*SHANK2* c.1664C>T missense	F: CCCTCAGTCCCCTTCGTT R: CCACATTCTCTCCACCTTCG	248	STD 60°C 30"30"30" x35	FastStart Taq
Chr12	*KCNMB4* c.604G>A missense	F: GCCCTTTCTTTCTTATTCTCCA R: GTCTTCCAGTTGTGCCTGTTT	365	STD 60°C 30"30"30" x35	FastStart Taq
Chr21	*TMEM50B*_ex3 g.34839414A>C missense	F: TGAAGAAAGCCAATGTGGAA R: TGATCTGTAGGACAAAAGGAATG	199	STD 59°C 30"30"30" x35	Taq Gold
Chr21	c21orf67 c.579C>A nonsense	F: CTCACTCGCTTCCTCCTGT R: AGCCATTTACCCCGTTTG	328	STD 58°C 30"30"30" x35	FastStart Taq

Table 5: Primers and amplification conditions for validation of candidate variants by Sanger sequencing or enzymatic digestion (when indicated) in Family 4.

APPENDIX C

Frequency study of candidate variants in the healthy population

Gene and variant	Primers for PCR or ARMS PCR	Reaction
SACS c.11104A>G missense	F outer: TTCTGCATCATATATTCCAAGAACGAAT R outer: GTTATTGATTACCTTATCAAGAGGAGGATC R inner(mut): CAATGTGATGTACTCCAGCTGTTATTGG	Tri-primer ARMS TD 76-66ºC 30”30”30”x10+25
FBXO41 c. 950G>A missense	F: TGAGGCTCTGGGGTAAGAGA R: GTAGAGGAGGGAAGGCAGGT	TD 68–58°C 30”30”30”x10+27 and Digestion with BstUI (New England Biolabs)
LPP (NM_001167671.1) c.1142C>T missense	F outer: CTTTTTCCTCTCCAGTGCAATCGCTGCT R outer: CGAAGTCAGCTCTAGGCCCCACTCACCA R inner(mut): GGAATGAAGGGGCAACTGCCA	Tri-primer ARMS TD 70–60°C 30”30”30”x10+24
SIGMAR1 (NM_005866) c.412G>C	F: ATTGTCACTCAGGGCGCTAC R: CCCAACACACTCCTTTTCCA	STD 61°C 30"30"30"x35 and digestion with MnlI (New England Biolabs)
SIGMAR1 (NM_005866) c.448G>A	F: TCCCCACCCCTAGTTAGTCC R: GCTCCAGCAAGTGGATATGTG	STD 58°C30"30"30"x35 and digestion with Alw26I (Promega)

Table 6: *SACS*, *FBXO41* and *LPP* variants were investigated in 200 healthy Italian subjects by these assays. *SIGMAR1* variants were investigated in 100 healthy subjects belonging to the same Southern region of Family 2.

APPENDIX D

Direct sequencing of poor-covered coding exons

Family 2

Chromosome	Amplicon	Forward and Reverse Primer sequences	Amplicon lenght (bp)	PCR conditions	DNA Polymerase
Chr8	*NDRG1*_ex16	F: GAGAGGGCACCCACGTAATAG R: AGCGTCACTTCTCTGGATGG	365	STD 61°C 30”30”40” x 35	FastStart Taq
Chr8	*SGK223*_ex2	F: GGCATTACCTGTGCATACCTG R: CCCAAGAAACTGTCCCTCAC	372	STD 60°C 30”30”40” x 35	FastStart Taq
Chr8	*SGK223*_ex5	F: CCTTCTCCGCAAACTTCATC R: GATCGTGTCTGCTTCCCAGT	383	STD 60°C 30”30”40” x 35	FastStart Taq
Chr8	*RP1L1*_ex4	F: TCCTGATTGGGGACCAGTGT R: TTCAGGGGCATCAAGGAGAA	406	STD 62°C 30”30”40” x 35	FastStart Taq
Chr8	*ERI1*_ex1	F: GCCTCCCTCTTGTTCGTCCT R: GCGGGACACCTGAGAAGAAA	457	STD 61°C 30”30”40” x 34	FastStart Taq
Chr8	*TNKS*_ex1	F: GTTCCCTTGGCTGTTCTCTG R: GAGGACGACGGTGAATTGTT	565	STD 59°C 30”30”40” x 34	FastStart Taq

Table 7: Primers and amplification conditions to sequence poorly-covered coding exons in Family 2.

Family 3

Chromosome	Amplicon	Forward and Reverse Primer sequences 5’-3’	Amplicon lenght (bp)	PCR conditions	DNA Polymerase
Chr2	*C2orf65*_ex10	F: CCAGTCATTTGCCAGGTTTT R: TTCAACCCCAACTGTCTTCA	180	STD 57°C 30”-30”45” x 34	Taq Gold
Chr2	*AUP1*_ex1	F: AGCAGGCGGACAGTAGGAC R: AGCAGCACGAGCAGTAGGA	485	STD 61°C 30”30”45” x 34	FastStart Taq
Chr2	*WDR54*_ex1	F: CATCCCCAGCTGAACTGAA R: CGAGCCAAATAAGTGTGATGA	385	STD 57°C 30”30”45” x 34	Taq Gold
Chr2	*PRADC1*_ex1	F: GACCCGGAGAGTATGCTG R: ACCACCACCCACTTATTCC	533	STD 61°C 30”30”45” x 35	Taq Gold
Chr2	*BOLA3*_ex1	F: GGGACCCTATCCATGAAACAG R: GGGAGCCAGTCCTCAAGC	449	TD 67-57°C 45”45”45” x10+26	360 Taq GC buffer
Chr2	*FBXO41*_ex1A	F: CCCTGAGCCTTCCTGACC R: CTCGGCCAACTCCTCACA	516	STD 61°C 30”30”45” x 34	FastStart Taq
Chr2	*FBXO41*_ex1B	F: CTGCTGCACCACCACCAT R: GATCTTCTGCTCCACCTCCTC	378	STD 60°C 30”30”45” x 35	Taq Gold
Chr2	*FBXO41*_ex1C	F: TGATGTGGCCTACGAAGAGG R: CACAGTGACCCGCACAGG	377	STD 61°C 30”30”45” x 35	Taq Gold
Chr2	*EGR4*_ex1A	F: AGGTGGGAAGCGCATCTA R: TTCGGAAAACTCGCTAAGGT	578	TD 68-58°C 30”30”45” 10 + 30	360 Taq GC buffer
Chr2	*EGR4*_ex1B	F: TGTCCATGTTTGGGCATTT R: TCCTGGTTCTCAGGTATGGTG	569	TD 66-56°C 30”30”45” x10+30	Accuprime GC-rich
Chr2	*EGR4*_ex2A	F: TTGTAGGTAGGGGCTGTGGA R: CAGAACGCCTCTGGGAAA	411	STD 58°C 30”30”45” x 34	Taq Gold
Chr2	*EGR4*_ex2B	F: AACCTCATGTCGGGCATCTT R: CCCACTAGGAGGGGTCAGG	570	STD 61°C 30”30”45” x 36	Taq Gold
Chr2	*EGR4*_ex2C	F: GGGGCCTATGACGCTTTC R: GCTGTGCCGTTTCTTCTCAT	582	STD 59°C 30”30”45” x 35	Taq Gold
Chr2	*EGR4*_ex2D	F: CTGCCTCCGCAACTTCAG R: GCCTGTCTCTGGGGGTTATAG	556	TD 67-57°C 30”30”45” x10+26	Taq Gold
Chr2	*DCTN1*_ex8	F: CCTTCCCTTGCCATATTCTC R: CCTCACCCTTTCTGATCCAA	408	STD 61°C 30”30”45” x35	Taq Gold
Chr2	*ALMS1*_ex1A	F: CAACGTCGCCTGTAGCAAA R: CGACAGCGGAGGCAAAAT	482	TD 68-58°C 30”30”30” x10+25	Taq Gold
Chr2	*ALMS1*_ex1B	F: GGACTCCGACTCTCACTACGG R: GAGTCTGGGCCGCCTACTA	302	STD 60°C 30”30”45” x 34 cicli	FastStart Taq
Chr3	*LEPREL1*_ex1A	F: CGAGGGAAGGTGGGAGAG R: CCAACAAGGAGCGGAAAA	430	STD 56°C 30”30”45” x 35	Accuprime GC-rich
Chr3	*LEPREL1*_ex1B	F: CGCCTACTACAGCGGAGACTA R: TCTGGATTCACAAACTGAGACAC	463	TD 67-57°C 45”45”45” x10+26	360 Taq GC buffer
Chr3	*CLDN16*_ex1	F: ACCACCACTAGCCCACAGTT R: CACCATCCAACAGTCAGTCC	422	STD 57°C 30”30”45” x 34	Taq Gold
Chr21	*USP25*_ex1	F: CCTCTCTCCCTTCCCCAAA R: CGGCAGAAGGAAGTGGATT	653	STD 55°C 30”30”45” x 35	FastStart Taq
Chr21	*SAMSN1*_ex1	F:ACATCCCCATCTGTTCCTGA R: TGGAATCCCTTTAAATCCAAAC	300	STD 56°C 30”30”30” x 34	Taq Gold

Chr21	*BTG3*_ex1	F: GGTAGGGCGAGGGTGTGT R: CTCCCCCGATACCCACAG	454	STD 59°C 30”30”45” x 35	FastStart Taq
Chr21	*NRIP1*_ex1	F: GCTGGTCGGAGGGAAGAG R: GGCAGCAGAGGCAGGATT	273	TD 64-54°C 30”30”45” x10+28	FastStart Taq
Chr21	*NRIP1*_ex2	F:CAACAACTATGTCCAAAGAAAGCA R: ACAAATGGCAAGGTAGTTTATCC	397	STD 61°C 30”30”30” x 34	Taq Gold
Chr21	*NRIP1*_ex3	F: AATGAACCAGCATTCATAAACAA R:AAGGAAGGATTGTAGCTCTTTCA	467	STD 60°C 30”30”30” x 34	Taq Gold

Table 8: Primers and amplification conditions to sequence poorly-covered coding exons in Family 3.

Family 4

Chromosome	**Amplicon**	**Forward and Reverse Primer sequences**	**Amplicon lenght (bp)**	**PCR conditions**	**DNA Polymerase**
Chr9	*ALG2*_ex1	F: GCAGAAGACCCCCATCAG R: GAACCGCAGACAGGGAAG	495	STD 59°C 40”40”40”x35	FastStart Taq
Chr9	*ABCA1*_ex2	F: GCAGTCCTCATTGGTGTATGG R: ATCCCCAACTCAAAACCACA	247	STD 60°C 40”40”40”x35	FastStart Taq
Chr9	*COL15A1*_ex16	F: CCTTTTCTAGCAAGCGTGTGT R: AGGCAAGGTCAGGTTTAGGG	220	STD 60°C 40”40”40”x35	FastStart Taq
Chr9	*NIPSNAP3A*_ex1	F: GCTCAGCACAGCAGAGAAAGA R: CATTGCTCGCACGTACCC	343	STD 60°C 40”40”40”x35	FastStart Taq
Chr9	*NIPSNAP3B*_ex1	F: CCGAAGAGAAAGACGCCAAC R: TCGCACGCTTCAGAAATACC	325	STD 60°C 40”40”40”x35	FastStart Taq
Chr9	*SEC61B*_ex2	F: CGGGTGTGGGTGTCTAGG R:CAATCTTCACATGCTAGGGTCTC	270	STD 60°C 40”40”40”x35	FastStart Taq
Chr9	*GRIN3A*_ex1	F: CTCCTGGGACCGCTTCAC R: GTTGTCCACGGCAAATAGGA	532	TD 65-55°C 30”30”40”x10+25	FastStart Taq
Chr9	*KLF4*_ ex2	F: GGGTTTTGGCTTCGTTTCTT R: CTCGTTCAGTGGCTCTTGGT	343	STD 60°C 40”40”40”x35	FastStart Taq
Chr9	*KLF4*_ ex3a	F: CGCCAGCACGTCAGTATGT R: ACGACGAAGAGGAGGCTGA	350	STD 63°C 30”30”30”x35	FastStart Taq
Chr9	*KLF4*_ ex3b	F: CGGAGAGAGACCGAGGAGT R: CCTTTGCTGACGCTGATG	573	TD 66-56°C 30”30”40”x10+25	FastStart Taq
Chr9	*KLF4* _ex3c	F: GCAAGTTCGTGCTGAAGG R: GAGCATCATCCCGTGTGT	500	STD 59°C 40”40”40”x35	FastStart Taq
Chr9	*SLC44A1*_ex1	F: GCCGCCTCTTGAGTACCAG R: TGGGGACAGCGAGAGGTAT	292	STD 56°C 40”40”40”x35	Accuprime GC-rich
Chr9	*TGFBR1*_ex1	F: CCTCCGAGCAGTTACAAAGG R: GCCATGTTTGAGAAAGAGCA	317	STD 56°C 40”40”40”x35	Accuprime GC-rich
Chr9	*TMEFF1*_ex1a	F: ACAAAGGGAAGGCGAGGA R: GAGAGAAGGCGAAGAGCAGA	464	TD 69-59°C 30”30”40”x10+23	Taq Gold
Chr9	*TMEFF1*_ex1b	F: TCCAGGGGCACCAGTCAT R: AACGGAGGGGTGGGAAGA	257	TD 69-59°C 30”30”40”x10+23	Taq Gold

Table 9: Primers and amplification conditions to sequence poorly-covered coding exons in Family 4.

APPENDIX E

Family 2 nucleus 1

Homozygosity mapping for each patient

Patient V-3							
chr	**from (bp)**	**to (bp)**	**from SNP**	**to SNP**	**Mb**	**Total RefSeqs**	**Comments about relevant genes/loci**
1	194072229	201517349	rs533405	rs11577209	7.4	40	-
1	203281175	211082285	rs17534202	rs1777250	7.8	98	*SYT14* for Spinocerebellar ataxia 11 (AR)
1	220009989	224897974	rs12143315	rs6669720	4.9	39	-
2	155651200	156917835	rs2591154	rs7578557	1.3	1	-
8	6754919	12718090	rs2738148	rs6997599	5.96	120	-
9	5487980	8342540	rs10975134	rs1846695	2.9	14	-
11	33904286	68855954	rs4756076	rs4930265	35	>100	- *BSCL2* for dHMNV and SPG17 (AR/AD) - *IGHMBP2* for dHMNVI (AR) - *B3GNT1* for Muscular dystrophy-dystroglycanopathy
15	27706279	35370551	rs6497269	rs6495749	7.7	70	-
18	55201461	61901187	rs754789	rs4328542	6.7	44	-
21	42694667	48061211	rs2838011	rs2256070	5.4	>100	-
Patient VI-6							
chr	**from (bp)**	**to (bp)**	**from SNP**	**to SNP**	**Mb**	**Total RefSeqs**	**Comments about relevant genes/loci**
1	168986374	170383224	rs2146201	rs10919392	1.4	18	-
3	143505568	144824171	rs1868179	rs7432650	1.3	2	-
6	118543515	119841070	rs9398485	rs10485006	1.3	1	-
6	128829399	130019354	rs6938035	rs17057640	1.2	3	*LAMA2* for muscular distrophy
7	135650363	152608306	rs10261276	rs6954929	17	>100	- DHMN1 locus - *CLCN1* for Myotonia congenita/levior (AR/AD) - myotonia-c locus
8	6754919	22111871	rs2738148	rs4872456	15.4	>100	- *RP1L1* for occult macular dystrophy (AD) - Vps37A (SPG53 -AR) - *ASAH1* for SMA with progressive myoclonic epilepsy - distal Myopathy 3 locus
8	139484985	146292734	rs16909195	chrom end	6.8	>100	*SLC52A2* for Brown-Vialetto-Van Laere syndrome 2 (AR)
11	15829043	17007220	rs11023727	rs2353369	1.2	3	-
11	27266037	28629115	rs10835148	rs11030385	1.4	11	-
12	41301332	42140826	rs11178982	rs1497166	0.8	2	- *CNTN1* for congenital Myopathy - CMT2G locus - *SPG26* locus (AR)
12	122471031	131490145	rs7137708	rs12313737	9	81	-

Table 10: Candidate homozygous regions selected after haplotype evaluations for each patient. Regions spanning more than 1 Mb were reported here (even variants were checked also for smaller regions). For each region chromosomal positions (hg19 release), flanking SNP markers, total validated Refseqs and relevant genes mapping to the region were annotated. In gray the regions totally or partially shared by the second patient. AR= autosomal recessive; AD= autosomal dominant.

Variants identified in the candidate regions (Table 10)

Chr	Candidate Variant	Variant call V-3	Variant call VI-6	Predicted Effect	Allele frequency	Haplotype	Functional Role	Confirmation by Sanger sequencing
8	*MTMR7* c.1749T>G	A(ref)87/C82 (Q=99)	A(ref)0/182C (Q=89)	Missense tolerated	rs145244130, MAF= 0.3% EVS: 1 homozygous subject	Excluded from unaffected subjects	Lipidic pathway, myotubularin-related MTMR2 cause of CMT4B1 (AR)	Homozygous only in VI-6
11	*ABTB2*_intron5 g.34189543A>G (hg19)	A(ref)0/19G (Q=52)	A(ref)19/G13 (Q=99)	Splice Site damaging	rs370573932, unknown frequency	Excluded from unaffected subjects	Cytoskeleton regulation, ion transporter	Homozygous only in V-3

Table 11: Filtered SNVs found in homozygosity regions for single patients. Variant calls are expressed with number of reads for each allele (ref=reference allele) and Q=Phred-like quality score of variant call. In gray the variant calls which could explain the mode of transmission for the disease trait are highlighted.

Family 3

Homozygosity mapping for at least 2 out of 3 patients

chr	from (bp)	to (bp)	from SNP	to SNP	Mb	Total RefSeqs	Shared by	Comments about relevant genes/loci
2	50754813	53356999	rs759507	rs1424972	2.6	1	IV-6; IV-7	-
2	72251751	75681462	rs1878503	rs10496198	3.4	56	All patients	- *ALMS1* for Alstrom syndrome (AR) - *DCTN1* for HMNVIIB (AD)
3	188161706	188500658	rs2162259	rs3846183	0.34	1	All patients	-
3	188952524	192088525	rs1562761	rs9859577	3.14	16	All patients	-
6	22248576	23401282	rs9356810	rs6932335	1.2	2	IV-6; IV-7	-
21	14658830	19044397	rs2258300	rs2824435	4.4	21	All patients	-

Table 12: Candidate homozygosity regions selected after haplotype evaluations. For each region chromosomal positions (hg19 release), flanking SNP markers, total validated Refseqs and relevant genes mapping to the region were annotated. AR= autosomal recessive, AD= autosomal dominant.

Variants identified in the poorly-covered exons in the patient IV-2 (WES in 2011)

Gene and Exon	Gene function	Expression (Gene Cards)	Variants	Comments
*CLDN16*_exon1	Claudin in tight junctions	Mainly in kidney	Indel rs56086318 Missense rs76555381	cause of Renal hypomagnesemia 3
*ALMS1*_exon1A	Protein for ciliogenesis in collecting duct cells	Also in brain	c.63delGAGGAGGAG, No rs	cause of Alström syndrome (cone-rod retinal dystrophy, cardiomyopathy and type 2 diabetes mellitus)

Table 13: Poorly-covered exons selected for Sanger sequencing with relative informations, where variants were detected.

APPENDIX F

Statistics of exome-sequencing data

Family 2

Patients	V-3	VI-6
Target region (bp)[(1)]	49874469	50031585
Raw reads	79933210	96182744
Raw data yield (Mb)	7194	8656
Reads mapped to genome	70830461	82046555
Reads mapped to target region[(2)]	50591639	60247551
Data mapped to target region (Mb)	3826.33	4545.89
Mean depth of target region(X)	76.72	90.86
Coverage of target region (%)[(3)]	97.35	97.53
Average read length (bp)	89.85	89.86
Rate of nucleotide mismatch (%)	0.33	0.33
Fraction of target covered >=4X (%)	93.05	93.65
Fraction of target covered >=10X (%)	87.14	88.45
Fraction of target covered >=20X (%)	78.78	81.19
Capture specificity (%)	72.35	74.72
Reads mapped to flanking region[(4)]	6824962	8573186
Mean depth of flanking region(X)	18.34	22.44
Coverage of flanking region (%)	91.38	92.43
Fraction of flanking covered >=4X (%)	71.44	75.82
Fraction of flanking covered >=10X (%)	47.65	53.60
Fraction of flanking covered >=20X (%)	28.79	34.27
Fraction of unique mapped bases on or near target (%)	81.50	84.67
Duplication rate (%)[(5)]	6.26	10.06
Mean depth of chrX(X)	96.07	58.51
Mean depth of chrY(X)	-	144.85
GC rate (%)	43.82	43.81
Gender test result	F	M

Table 14: Statistical analysis of output data of patients V-3 and VI-6. Data were provided by BGI. (1) regions actually covered by the designed probes, (2) reads within or overlapping with target region, (3) the percentage of uniquely mapped reads aligning to target region, (4) regions +/-200 bp on both sides of each target region, (5) PCR duplicates would have the same start and end for both mates, which rarely occur by chance. Duplication rate is the fraction of duplicated reads in raw data.

Family 3

Patients	IV-2 (2011)	IV-7 (2013)
Target region (bp)[1]	50031585	51339787
Raw reads	118816412	151049882
Raw data yield (Mb)	10693	13594
Reads mapped to genome	95752656	124889294
Reads mapped to target region[2]	67005013	97603841
Data mapped to target region (Mb)	5042.99	7663.11
Mean depth of target region(X)	100.80	149.26
Coverage of target region (%)[3]	97.84	99.79
Average read length (bp)	89.80	89.90
Rate of nucleotide mismatch (%)	0.29	0.20
Fraction of target covered >=4X (%)	94.03	99.53
Fraction of target covered >=10X (%)	88.91	99.04
Fraction of target covered >=20X (%)	81.87	97.96
Capture specificity (%)	71.02	78.70
Reads mapped to flanking region[4]	9947003	11046587
Mean depth of flanking region(X)	25.55	29.62
Coverage of flanking region (%)	93.48	98.26
Fraction of flanking covered >=4X (%)	77.93	88.98
Fraction of flanking covered >=10X (%)	56.55	68.94
Fraction of flanking covered >=20X (%)	37.34	48.49
Fraction of unique mapped bases on or near target (%)	80.87	87.24
Duplication rate (%)[5]	15.82	14.81
Mean depth of chrX(X)	65.28	165.50
Mean depth of chrY(X)	151.49	-
GC rate (%)	43.48	48.69
Gender test result	M	F

Table 15: Statistical analysis of output data of patient IV-2 and IV-7. Data were provided by BGI. (1) regions actually covered by the designed probes, (2) reads within or overlapping with target region, (3) the percentage of uniquely mapped reads aligning to target region, (4) regions +/-200 bp on both sides of each target region, (5) PCR duplicates would have the same start and end for both mates, which rarely occur by chance. Duplication rate is the fraction of duplicated reads in raw data.

Family 4

Patients	III-2	III-4	III-6
Target region (bp)(1)	50238347	50080510	50080510
Raw reads	74388348	76554006	71048998
Raw data yield (Mb)	6695	6890	6394
Reads mapped to genome	64020072	65485534	61714360
Reads mapped to target region(2)	44510426	44978621	43793351
Data mapped to target region (Mb)	3299.62	3339.55	3268.15
Mean depth of target region(X)	65.68	66.68	65.26
Coverage of target region (%)(3)	96.92%	97.36%	96.95%
Average read length (bp)	89.8	89.8	89.8
Rate of nucleotide mismatch (%)	0.38	0.39	0.4
Fraction of target covered >=4X (%)	92.07	92.88	92.18
Fraction of target covered >=10X (%)	85.43	86.7	85.66
Fraction of target covered >=20X (%)	75.85	77.62	76.25
Capture specificity (%)	70.58	69.71	71.94
Reads mapped to flanking region(4)	8568356	9045966	7258570
Mean depth of flanking region(X)	20.06	20.62	18.14
Coverage of flanking region (%)	92.26	93.3	91.16
Fraction of flanking covered >=4X (%)	75.63	78.2	71.83
Fraction of flanking covered >=10X (%)	53.25	55.77	48.79
Fraction of flanking covered >=20X (%)	32.89	34.18	29.47
Fraction of unique mapped bases on or near target (%)	83.35	82.75	83.19
Duplication rate (%)(5)	8.14	8.56	6.98
Mean depth of chrX(X)	43.06	82.11	80.84
Mean depth of chrY(X)	94.56	-	-
GC rate (%)	43.44	43.8	43.89
Gender test result	M	F	F

Table 16: Statistical analysis of output data of patient III-2, III-4 and III-6. Data were provided by BGI. (1) regions actually covered by the designed probes, (2) reads within or overlapping with target region, (3) the percentage of uniquely mapped reads aligning to target region, (4) regions +/-200 bp on both sides of each target region, (5) PCR duplicates would have the same start and end for both mates, which rarely occur by chance. Duplication rate is the fraction of duplicated reads in raw data.

APPENDIX G

Copy Number Variants (CNVs) analysis

Family 2

Sample ID	Chromosome	Start	Stop	State	bp
V-3	chr1	192129479	192321379	duplication	191900
V-3	chr6	31949857	31952238	duplication	2381
V-3	chr6	31991645	31994158	duplication	2513
V-3	chr6	31999315	32002118	duplication	2803
VI-6	chr6	138566629	138582839	duplication	16210
V-3	chr7	5939904	5944932	duplication	5028
VI-6	chr10	37451926	37478499	duplication	26573
VI-6	chr10	46961921	47161432	duplication	199511
VI-6	chr10	48247655	48264451	duplication	16796
V-3	chr11	60971540	60978753	duplication	7213
V-3	chr11	60989841	60999048	duplication	9207

Table 17: CNVs obtained from exome-sequencing data of patients V-3 and VI-6, obtained by Copy Number Inference From Exome Reads (CoNIFER) program. Data were provided by BGI.

Family 3

Sample ID	Chromosome	Start	Stop	State	bp
IV-2	chr1	12921055	13001464	deletion	80409
IV-2	chr1	1640152	1644787	duplication	4635
IV-2	chr1	1650711	1663987	duplication	13276
IV-7	chr2	73215372	73249722	duplication	34350
IV-7	chr2	96943497	96945302	duplication	1805
IV-7	chr2	219288977	219292801	duplication	3824
IV-2	chr2	233244190	233245765	deletion	1575
IV-2	chr3	48606963	48609515	deletion	2552
IV-7	chr4	3230424	3237187	duplication	6763
IV-7	chr4	5800296	5811351	duplication	11055
IV-2	chr5	140710198	140726105	deletion	15907
IV-7	chr5	149441046	149449650	duplication	8604
IV-2	chr5	176310998	176316593	deletion	5595
IV-7	chr6	33631436	33633029	duplication	1593
IV-7	chr6	34496387	34497650	duplication	1263
IV-2	chr7	141764133	141786132	deletion	21999
IV-2	chr7	143880564	144070392	duplication	189828
IV-2	chr8	7753944	7808529	duplication	54585
IV-2	chr8	11965649	11969283	duplication	3634
IV-2	chr9	117072756	117092301	duplication	19545
IV-2	chr10	37486558	37505335	duplication	18777

IV-2	chr10	47911019	47920112	duplication	9093
IV-2	chr12	129569025	129822390	deletion	253365
IV-2	chr15	75575199	75581841	deletion	6642
IV-2	chr16	15035613	15043997	duplication	8384
IV-2	chr17	78972133	79059549	deletion	87416
IV-2	chr19	7030585	7049467	duplication	18882
IV-2	chr19	52130647	52217377	deletion	86730
IV-2	chr20	21314159	21319763	duplication	5604

Table 18: CNVs obtained from exome-sequencing data of patients IV-2 and IV-7, obtained by Copy Number Inference From Exome Reads (CoNIFER) program. Data were provided by BGI.

Family 4

Sample ID	Chromosome	Start	Stop	State	bp
III-2	chr3	180320621	180322450	duplication	1829
III-2	chr3	180334050	180337798	duplication	3748
III-4	chr4	69340364	69426440	duplication	86076
III-2	chr6	32317502	32410748	duplication	93246
III-4	chr14	19553373	19571443	duplication	18070
III-4	chr14	20014465	20249432	duplication	234967
III-6	chr16	70180024	70190819	duplication	10795
III-6	chr17	18387175	18395342	duplication	8167
III-2	chr19	14142675	14157177	duplication	14502

Table 19: CNVs obtained from exome-sequencing data of patients III-2, III-4 and III-6, obtained by Copy Number Inference From Exome Reads (CoNIFER) program. Data were provided by BGI.

REFERENCES

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* (2012) 491, 56-65.

Al-Saif, A., Al-Mohanna, F. & Bohlega, S. A mutation in sigma-1 receptor causes juvenile amyotrophic lateral sclerosis. *Ann Neurol* (2011) 70, 913-919.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* (1997) 25, 3389-3402.

Amiott, E. A., Lott, P., Soto, J., Kang, P. B., McCaffery, J. M. *et al.* Mitochondrial fusion and function in Charcot-Marie-Tooth type 2A patient fibroblasts with mitofusin 2 mutations. *Exp Neurol* (2008) 211, 115-127.

Antonellis, A., Ellsworth, R. E., Sambuughin, N., Puls, I., Abel, A. *et al.* Glycyl tRNA synthetase mutations in Charcot-Marie-Tooth disease type 2D and distal spinal muscular atrophy type V. *Am J Hum Genet* (2003) 72, 1293-1299.

Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* (2010) 38, W529-533.

Auer-Grumbach, M., Schlotter-Weigel, B., Lochmuller, H., Strobl-Wildemann, G., Auer-Grumbach, P. *et al.* Phenotypes of the N88S Berardinelli-Seip congenital lipodystrophy 2 mutation. *Ann Neurol* (2005) 57, 415-424.

Auer-Grumbach, M., Weger, M., Fink-Puches, R., Papic, L., Frohlich, E. *et al.* Fibulin-5 mutations link inherited neuropathies, age-related macular degeneration and hyperelastic skin. *Brain* (2011) 134, 1839-1852.

Baets, J., Deconinck, T., Smets, K., Goossens, D., Van den Bergh, P. *et al.* Mutations in SACS cause atypical and late-onset forms of ARSACS. *Neurology* (2010) 75, 1181-1188.

Baets, J., Deconinck, T., De Vriendt, E., Zimon, M., Yperzeele, L. *et al.* Genetic spectrum of hereditary neuropathies with onset in the first year of life. *Brain* (2011) 134, 2664-2676.

Baets, J. & Timmerman, V. Inherited peripheral neuropathies: a myriad of genes and complex phenotypes. *Brain* (2011) 134, 1587-1590.

Bainbridge, M. N., Wang, M., Burgess, D. L., Kovar, C., Rodesch, M. J. *et al.* Whole exome capture in solution with 3 Gbp of data. *Genome Biol* (2010) 11, R62.

Bao, H., Guo, H., Wang, J., Zhou, R., Lu, X. *et al.* MapView: visualization of short reads alignment on a desktop computer. *Bioinformatics* (2009) 25, 1554-1555.

Barisic, N., Claeys, K. G., Sirotkovic-Skerlev, M., Lofgren, A., Nelis, E. *et al.* Charcot-Marie-Tooth disease: a clinico-genetic confrontation. *Ann Hum Genet* (2008) 72, 416-441.

Berezin, C., Glaser, F., Rosenberg, J., Paz, I., Pupko, T. *et al.* ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics* (2004) 20, 1322-1324.

Blom, N., Gammeltoft, S. & Brunak, S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol* (1999) 294, 1351-1362.

Bolze, A., Byun, M., McDonald, D., Morgan, N. V., Abhyankar, A. *et al.* Whole-exome-sequencing-based discovery of human FADD deficiency. *Am J Hum Genet* (2010) 87, 873-881.

Brambilla, L., Martorana, F. & Rossi, D. Astrocyte signaling and neurodegeneration: new insights into CNS disorders. *Prion* (2013) 7, 28-36.

Brkanac, Z., Spencer, D., Shendure, J., Robertson, P. D., Matsushita, M. *et al.* IFRD1 is a candidate gene for SMNA on chromosome 7q22-q23. *Am J Hum Genet* (2009) 84, 692-697.

Brusse, E., Majoor-Krakauer, D., de Graaf, B. M., Visser, G. H., Swagemakers, S. *et al.* A novel 16p locus associated with BSCL2 hereditary motor neuronopathy: a genetic modifier? *Neurogenetics* (2009) 10, 289-297.

Bucci, C., Bakke, O. & Progida, C. Charcot-Marie-Tooth disease and intracellular traffic. *Prog Neurobiol* (2012) 99, 191-225.

Cartegni, L. & Krainer, A. R. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nat Genet* (2002) 30, 377-384.

Chen, D. H., Sul, Y., Weiss, M., Hillel, A., Lipe, H. *et al.* CMT2C with vocal cord paresis associated with short stature and mutations in the TRPV4 gene. *Neurology* (2010) 75, 1968-1975.

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

Christodoulou, K., Zamba, E., Tsingis, M., Mubaidin, A., Horani, K. *et al.* A novel form of distal hereditary motor neuronopathy maps to chromosome 9p21.1-p12. *Ann Neurol* (2000) 48, 877-884.

Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res* (2009) 19, 1553-1561.

Cole, C., Barber, J. D. & Barton, G. J. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res* (2008) 36, W197-201.

Coleman, M. P. The challenges of axon survival: introduction to the special issue on axonal degeneration. *Exp Neurol* (2013) 246, 1-5.

Consortium, I. H. G. S. Finishing the euchromatic sequence of the human genome. *Nature* (2004) 431, 931-945.

Crottes, D., Guizouarn, H., Martin, P., Borgese, F. & Soriani, O. The sigma-1 receptor: a regulator of cancer cell electrical plasticity? *Front Physiol* (2013) 4, 175.

Date, H., Onodera, O., Tanaka, H., Iwabuchi, K., Uekawa, K. *et al.* Early-onset ataxia with ocular motor apraxia and hypoalbuminemia is caused by mutations in a new HIT superfamily gene. *Nat Genet* (2001) 29, 184-188.

Davenport, R. J., Swingler, R. J., Chancellor, A. M. & Warlow, C. P. Avoiding false positive diagnoses of motor neuron disease: lessons from the Scottish Motor Neuron Disease Register. *J Neurol Neurosurg Psychiatry* (1996) 60, 147-151.

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* (2010) 6, e1001025.

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

Di Malta, C., Fryer, J. D., Settembre, C. & Ballabio, A. Astrocyte dysfunction triggers neurodegeneration in a lysosomal storage disorder. *Proc Natl Acad Sci U S A* (2012) 109, E2334-2342.

Dierick, I., Baets, J., Irobi, J., Jacobs, A., De Vriendt, E. *et al.* Relative contribution of mutations in genes for autosomal dominant distal hereditary motor neuropathies: a genotype-phenotype correlation study. *Brain* (2008) 131, 1217-1227.

Duquette, A., Brais, B., Bouchard, J. P. & Mathieu, J. Clinical presentation and early evolution of spastic ataxia of Charlevoix-Saguenay. *Mov Disord* (2013).

Dyck, P. J. & Lambert, E. H. Lower motor and primary sensory neuron diseases with peroneal muscular atrophy. II. Neurologic, genetic, and electrophysiologic findings in various neuronal degenerations. *Arch Neurol* (1968) 18, 619-625.

Eastman, S. W., Yassaee, M. & Bieniasz, P. D. A role for ubiquitin ligases and Spartin/SPG20 in lipid droplet turnover. *J Cell Biol* (2009) 184, 881-894.

Eisenberg, I., Avidan, N., Potikha, T., Hochner, H., Chen, M. *et al.* The UDP-N-acetylglucosamine 2-epimerase/N-acetylmannosamine kinase gene is mutated in recessive hereditary inclusion body myopathy. *Nat Genet* (2001) 29, 83-87.

Engert, J. C., Berube, P., Mercier, J., Dore, C., Lepage, P. *et al.* ARSACS, a spastic ataxia common in northeastern Quebec, is caused by mutations in a new gene encoding an 11.5-kb ORF. *Nat Genet* (2000) 24, 120-125.

Finsterer, J., Loscher, W., Quasthoff, S., Wanschitz, J., Auer-Grumbach, M. *et al.* Hereditary spastic paraplegias with autosomal dominant, recessive, X-linked, or maternal trait of inheritance. *J Neurol Sci* (2012) 318, 1-18.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* (2013) 41, D808-815.

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* (2009) 25, i54-62.

Geourjon, C. & Deleage, G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput Appl Biosci* (1995) 11, 681-684.

Gilissen, C., Arts, H. H., Hoischen, A., Spruijt, L., Mans, D. A. *et al.* Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. *Am J Hum Genet* (2010) 87, 418-423.

Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Disease gene identification strategies for exome sequencing. *Eur J Hum Genet* (2012) 20, 490-497.

Girard, M., Lariviere, R., Parfitt, D. A., Deane, E. C., Gaudet, R. *et al.* Mitochondrial dysfunction and Purkinje cell loss in autosomal recessive spastic ataxia of Charlevoix-Saguenay (ARSACS). *Proc Natl Acad Sci U S A* (2012) 109, 1661-1666.

Gnad, F., Gunawardena, J. & Mann, M. PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res* (2011) 39, D253-260.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* (2009) 27, 182-189.

Goizet, C., Boukhris, A., Mundwiller, E., Tallaksen, C., Forlani, S. *et al.* Complicated forms of autosomal dominant hereditary spastic paraplegia are frequent in SPG10. *Hum Mutat* (2009) 30, E376-385.

Greer, P. L., Hanayama, R., Bloodgood, B. L., Mardinly, A. R., Lipton, D. M. *et al.* The Angelman Syndrome protein Ube3A regulates synapse development by ubiquitinating arc. *Cell* (2010) 140, 704-716.

Gregianin, E., Vazza, G., Scaramel, E., Boaretto, F., Vettori, A. *et al.* A novel SACS mutation results in non-ataxic spastic paraplegia and peripheral neuropathy. *Eur J Neurol* (2013) 20, 1486-1491.

Group, T. H. D. R. C. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell* (1993) 72, 971-983.

Guernsey, D. L., Jiang, H., Bedard, K., Evans, S. C., Ferguson, M. *et al.* Mutation in the gene encoding ubiquitin ligase LRSAM1 in patients with Charcot-Marie-Tooth disease. *PLoS Genet* (2010) 6.

Hansske, B., Thiel, C., Lubke, T., Hasilik, M., Honing, S. *et al.* Deficiency of UDP-galactose:N-acetylglucosamine beta-1,4-galactosyltransferase I causes the congenital disorder of glycosylation type IId. *J Clin Invest* (2002) 109, 725-733.

Harding, A. *Inherited neuronal atrophy and degeneration predominantly of lower motor neurons*. 1051-64 (1993).

Hearn, T., Renforth, G. L., Spalluto, C., Hanley, N. A., Piper, K. *et al.* Mutation of ALMS1, a large gene with a tandem repeat encoding 47 amino acids, causes Alstrom syndrome. *Nat Genet* (2002) 31, 79-83.

Hewamadduma, C., McDermott, C., Kirby, J., Grierson, A., Panayi, M. *et al.* New pedigrees and novel mutation expand the phenotype of REEP1-associated hereditary spastic paraplegia (HSP). *Neurogenetics* (2009) 10, 105-110.

Hoischen, A., van Bon, B. W., Gilissen, C., Arts, P., van Lier, B. *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat Genet* (2010) 42, 483-485.

Houlden, H., Laura, M., Wavrant-De Vrieze, F., Blake, J., Wood, N. *et al.* Mutations in the HSP27 (HSPB1) gene cause dominant, recessive, and sporadic distal HMN/CMT type 2. *Neurology* (2008) 71, 1660-1668.

Hui, P. Next Generation Sequencing: Chemistry, Technology and Applications. *Top Curr Chem* (2012).

Irobi, J., Van Impe, K., Seeman, P., Jordanova, A., Dierick, I. *et al.* Hot-spot residue in small heat-shock protein 22 causes distal motor neuropathy. *Nat Genet* (2004) 36, 597-601.

Ivanova, N., Claeys, K. G., Deconinck, T., Litvinenko, I., Jordanova, A. *et al.* Hereditary spastic paraplegia 3A associated with axonal neuropathy. *Arch Neurol* (2007) 64, 706-713.

Jin, J., Cardozo, T., Lovering, R. C., Elledge, S. J., Pagano, M. *et al.* Systematic analysis and nomenclature of mammalian F-box proteins. *Genes Dev* (2004) 18, 2573-2580.

Johnson, J. O., Mandrioli, J., Benatar, M., Abramzon, Y., Van Deerlin, V. M. *et al.* Exome sequencing reveals VCP mutations as a cause of familial ALS. *Neuron* (2010) 68, 857-864.

Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* (1999) 292, 195-202.

Jouet, M., Rosenthal, A., Armstrong, G., MacFarlane, J., Stevenson, R. *et al.* X-linked spastic paraplegia (SPG1), MASA syndrome and X-linked hydrocephalus result from mutations in the L1 gene. *Nat Genet* (1994) 7, 402-407.

Kamionka, M. & Feigon, J. Structure of the XPC binding domain of hHR23A reveals hydrophobic patches for protein interaction. *Protein Sci* (2004) 13, 2370-2377.

Kiefer, F., Arnold, K., Kunzli, M., Bordoli, L. & Schwede, T. The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res* (2009) 37, D387-392.

Kiemer, L., Bendtsen, J. D. & Blom, N. NetAcet: prediction of N-terminal acetylation sites. *Bioinformatics* (2005) 21, 1269-1270.

Kozlov, G., Denisov, A. Y., Girard, M., Dicaire, M. J., Hamlin, J. *et al.* Structural basis of defects in the sacsin HEPN domain responsible for autosomal recessive spastic ataxia of Charlevoix-Saguenay (ARSACS). *J Biol Chem* (2011) 286, 20407-20412.

Krawitz, P. M., Schweiger, M. R., Rodelsperger, C., Marcelis, C., Kolsch, U. *et al.* Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nat Genet* (2010) 42, 827-829.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. & Lander, E. S. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* (1996) 58, 1347-1363.

Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M. *et al.* Copy number variation detection and genotyping from exome sequence data. *Genome Res* (2012) 22, 1525-1532.

Ku, C. S., Naidoo, N. & Pawitan, Y. Revisiting Mendelian disorders through exome sequencing. *Hum Genet* (2011) 129, 351-370.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C. *et al.* Initial sequencing and analysis of the human genome. *Nature* (2001) 409, 860-921.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* (2007) 23, 2947-2948.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L. *et al.* The diploid genome sequence of an individual human. *PLoS Biol* (2007) 5, e254.

Lindner, T. H. & Hoffmann, K. easyLINKAGE: a PERL script for easy and automated two-/multi-point linkage analyses. *Bioinformatics* (2005) 21, 405-407.

Liu, S. G., Zhao, J. J., Zhuang, M. Y., Li, F. F., Zhang, Q. J. *et al.* Clinical and genetic study of SPG6 mutation in a Chinese family with hereditary spastic paraplegia. *J Neurol Sci* (2008) 266, 109-114.

Lupski, J. R., de Oca-Luna, R. M., Slaugenhaupt, S., Pentao, L., Guzzetta, V. *et al.* DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell* (1991) 66, 219-232.

Marian, A. J. Challenges in medical applications of whole exome/genome sequencing discoveries. *Trends Cardiovasc Med* (2012) 22, 219-223.

Martin, E., Schule, R., Smets, K., Rastetter, A., Boukhris, A. *et al.* Loss of function of glucocerebrosidase GBA2 is responsible for motor neuron defects in hereditary spastic paraplegia. *Am J Hum Genet* (2013) 92, 238-244.

Masciullo, M., Modoni, A., Tessa, A., Santorelli, F. M., Rizzo, V. *et al.* Novel SACS mutations in two unrelated Italian patients with spastic ataxia: clinico-diagnostic characterization and results of serial brain MRI studies. *Eur J Neurol* (2012) 19, e77-78.

Mavlyutov, T. A., Epstein, M. L., Verbny, Y. I., Huerta, M. S., Zaitoun, I. *et al.* Lack of sigma-1 receptor exacerbates ALS progression in mice. *Neuroscience* (2013) 240, 129-134.

McCorquodale, D. S., Montenegro, G., Peguero, A., Carlson, N., Speziani, F. *et al.* Mutation screening of mitofusin 2 in Charcot-Marie-Tooth disease type 2. *J Neurol* (2011) 258, 1234-1239.

Menkes, J. H., Philippart, M. & Clark, D. B. Hereditary Partial Agenesis of Corpus Callosum; Biochemical and Pathological Studies. *Arch Neurol* (1964) 11, 198-208.

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

Miller, S. A., Dykes, D. D. & Polesky, H. F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* (1988) 16, 1215.

Moore, B., Hu, H., Singleton, M., De La Vega, F. M., Reese, M. G. *et al.* Global analysis of disease-related DNA sequence variation in 10 healthy individuals: implications for whole genome-based clinical diagnostics. *Genet Med* (2011) 13, 210-217.

Mostacciuolo, M. L., Rampoldi, L., Righetti, E., Vazza, G., Schiavon, F. *et al.* Hereditary spastic paraplegia associated with peripheral neuropathy: a distinct clinical and genetic entity. *Neuromuscul Disord* (2000) 10, 497-502.

Mucaki, E. J., Shirley, B. C. & Rogan, P. K. Prediction of mutant mRNA splice isoforms by information theory-based exon definition. *Hum Mutat* (2013) 34, 557-565.

Neale, B. M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K. E. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* (2012) 485, 242-245.

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* (2009) 461, 272-276.

Ng, S. B., Bigham, A. W., Buckingham, K. J., Hannibal, M. C., McMillin, M. J. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* (2010) 42, 790-793.

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* (2010) 42, 30-35.

Oliveira, J., Soares-Silva, I., Fokkema, I., Goncalves, A., Cabral, A. *et al.* Novel synonymous substitution in POMGNT1 promotes exon skipping in a patient with congenital muscular dystrophy. *J Hum Genet* (2008) 53, 565-572.

Ortega-Roldan, J. L., Ossa, F. & Schnell, J. R. Characterization of the human sigma-1 receptor chaperone domain structure and binding immunoglobulin protein (BiP) interactions. *J Biol Chem* (2013) 288, 21448-21457.

Pabba, M. The essential roles of protein-protein interaction in sigma-1 receptor functions. *Front Cell Neurosci* (2013) 7, 50.

Pareyson, D., Scaioli, V. & Laura, M. Clinical and electrophysiological aspects of Charcot-Marie-Tooth disease. *Neuromolecular Med* (2006) 8, 3-22.

Pareyson, D. & Marchesi, C. Diagnosis, natural history, and management of Charcot-Marie-Tooth disease. *Lancet Neurol* (2009) 8, 654-667.

Parfitt, D. A., Michael, G. J., Vermeulen, E. G., Prodromou, N. V., Webb, T. R. *et al.* The ataxia protein sacsin is a functional co-chaperone that protects against polyglutamine-expanded ataxin-1. *Hum Mol Genet* (2009) 18, 1556-1565.

Parla, J. S., Iossifov, I., Grabill, I., Spector, M. S., Kramer, M. *et al.* A comparative analysis of exome capture. *Genome Biol* (2011) 12, R97.

Pierce, S. B., Walsh, T., Chisholm, K. M., Lee, M. K., Thornton, A. M. *et al.* Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault Syndrome. *Am J Hum Genet* (2010) 87, 282-288.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* (2010) 20, 110-121.

Prause, J., Goswami, A., Katona, I., Roos, A., Schnizler, M. *et al.* Altered localization, abnormal modification and loss of function of Sigma receptor-1 in amyotrophic lateral sclerosis. *Hum Mol Genet* (2013) 22, 1581-1600.

Pyle, A., Griffin, H., Yu-Wai-Man, P., Duff, J., Eglon, G. *et al.* Prominent sensorimotor neuropathy due to SACS mutations revealed by whole-exome sequencing. *Arch Neurol* (2012) 69, 1351-1354.

Reilly, M. M. & Shy, M. E. Diagnosis and new treatments in genetic neuropathies. *J Neurol Neurosurg Psychiatry* (2009) 80, 1304-1314.

Riva, N., Iannaccone, S., Corbo, M., Casellato, C., Sferrazza, B. *et al.* Motor nerve biopsy: clinical usefulness and histopathological criteria. *Ann Neurol* (2011) 69, 197-201.

Rossor, A. M., Kalmar, B., Greensmith, L. & Reilly, M. M. The distal hereditary motor neuropathies. *J Neurol Neurosurg Psychiatry* (2012) 83, 6-14.

Saxena, S. & Caroni, P. Selective neuronal vulnerability in neurodegenerative diseases: from stressor thresholds to degeneration. *Neuron* (2011) 71, 35-48.

Schulte, T., Miterski, B., Bornke, C., Przuntek, H., Epplen, J. T. *et al.* Neurophysiological findings in SPG4 patients differ from other types of spastic paraplegia. *Neurology* (2003) 60, 1529-1532.

Seelow, D., Schuelke, M., Hildebrandt, F. & Nurnberg, P. HomozygosityMapper--an interactive approach to homozygosity mapping. *Nucleic Acids Res* (2009) 37, W593-599.

Seth, P., Ganapathy, M. E., Conway, S. J., Bridges, C. D., Smith, S. B. *et al.* Expression pattern of the type 1 sigma receptor in the brain and identity of critical anionic amino acid residues in the ligand-binding domain of the receptor. *Biochim Biophys Acta* (2001) 1540, 59-67.

Shaw, P. J. Molecular and cellular pathways of neurodegeneration in motor neurone disease. *J Neurol Neurosurg Psychiatry* (2005) 76, 1046-1057.

Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat Biotechnol* (2008) 26, 1135-1145.

Shimazaki, H., Sakoe, K., Niijima, K., Nakano, I. & Takiyama, Y. An unusual case of a spasticity-lacking phenotype with a novel SACS mutation. *J Neurol Sci* (2007) 255, 87-89.

Shojaee, S., Sina, F., Banihosseini, S. S., Kazemi, M. H., Kalhor, R. *et al.* Genome-wide linkage analysis of a Parkinsonian-pyramidal syndrome pedigree by 500 K SNP arrays. *Am J Hum Genet* (2008) 82, 1375-1384.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* (2005) 15, 1034-1050.

Simon, D. B., Lu, Y., Choate, K. A., Velazquez, H., Al-Sabban, E. *et al.* Paracellin-1, a renal tight junction protein required for paracellular Mg2+ resorption. *Science* (1999) 285, 103-106.

Sironi, M., Menozzi, G., Riva, L., Cagliani, R., Comi, G. P. *et al.* Silencer elements as possible inhibitors of pseudoexon splicing. *Nucleic Acids Res* (2004) 32, 1783-1791.

Siskind, C. E., Panchal, S., Smith, C. O., Feely, S. M., Dalton, J. C. *et al.* A review of genetic counseling for Charcot Marie Tooth disease (CMT). *J Genet Couns* (2013) 22, 422-436.

Smith, P. J., Zhang, C., Wang, J., Chew, S. L., Zhang, M. Q. *et al.* An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum Mol Genet* (2006) 15, 2490-2508.

Su, H., Fan, W., Coskun, P. E., Vesa, J., Gold, J. A. *et al.* Mitochondrial dysfunction in CA1 hippocampal neurons of the UBE3A deficient mouse model for Angelman syndrome. *Neurosci Lett* (2011) 487, 129-133.

Sumner, C. J., d'Ydewalle, C., Wooley, J., Fawcett, K. A., Hernandez, D. *et al.* A dominant mutation in FBXO38 causes distal spinal muscular atrophy with calf predominance. *Am J Hum Genet* (2013) 93, 976-983.

Synofzik, M., Soehn, A. S., Gburek-Augustat, J., Schicks, J., Karle, K. N. *et al.* Autosomal recessive spastic ataxia of Charlevoix Saguenay (ARSACS): expanding the genetic, clinical and imaging spectrum. *Orphanet J Rare Dis* (2013) 8, 41.

Szigeti, K. & Lupski, J. R. Charcot-Marie-Tooth disease. *Eur J Hum Genet* (2009) 17, 703-710.

Takiyama, Y. Autosomal recessive spastic ataxia of Charlevoix-Saguenay. *Neuropathology* (2006) 26, 368-375.

Tanaka, H., Katoh, H. & Negishi, M. Pragmin, a novel effector of Rnd2 GTPase, stimulates RhoA activity. *J Biol Chem* (2006) 281, 10355-10364.

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* (2012) 337, 64-69.

Thiele, H. & Nurnberg, P. HaploPainter: a tool for drawing pedigrees with complex haplotypes. *Bioinformatics* (2005) 21, 1730-1732.

Thiffault, I., Dicaire, M. J., Tetreault, M., Huang, K. N., Demers-Lamarche, J. *et al.* Diversity of ARSACS mutations in French-Canadians. *Can J Neurol Sci* (2013) 40, 61-66.

Timmerman, V., Clowes, V. E. & Reid, E. Overlapping molecular pathological themes link Charcot-Marie-Tooth neuropathies and hereditary spastic paraplegias. *Exp Neurol* (2013) 246, 14-25.

Tsui, L. C., Buchwald, M., Barker, D., Braman, J. C., Knowlton, R. *et al.* Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science* (1985) 230, 1054-1057.

Vallat, J. M., Mathis, S. & Funalot, B. The various Charcot-Marie-Tooth diseases. *Curr Opin Neurol* (2013) 26, 473-480.

Vazza, G., Zortea, M., Boaretto, F., Micaglio, G. F., Sartori, V. *et al.* A new locus for autosomal recessive spastic paraplegia associated with mental retardation and distal motor neuropathy, SPG14, maps to chromosome 3q27-q28. *Am J Hum Genet* (2000) 67, 504-509.

Veiga-da-Cunha, M., Tyteca, D., Stroobant, V., Courtoy, P. J., Opperdoes, F. R. *et al.* Molecular identification of NAT8 as the enzyme that acetylates cysteine S-conjugates to mercapturic acids. *J Biol Chem* (2010) 285, 18888-18898.

Vierbuchen, T., Ostermeier, A., Pang, Z. P., Kokubu, Y., Sudhof, T. C. *et al.* Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* (2010) 463, 1035-1041.

Vucic, S., Kennerson, M., Zhu, D., Miedema, E., Kok, C. *et al.* CMT with pyramidal features. Charcot-Marie-Tooth. *Neurology* (2003) 60, 696-699.

Wan, J., Yourshaw, M., Mamsa, H., Rudnik-Schoneborn, S., Menezes, M. P. *et al.* Mutations in the RNA exosome component gene EXOSC3 cause pontocerebellar hypoplasia and spinal motor neuron degeneration. *Nat Genet* (2012) 44, 704-708.

Wang, J. L., Yang, X., Xia, K., Hu, Z. M., Weng, L. *et al.* TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. *Brain* (2010) 133, 3510-3518.

Wang, Z., Rolish, M. E., Yeo, G., Tung, V., Mawson, M. *et al.* Systematic identification and analysis of exonic splicing silencers. *Cell* (2004) 119, 831-845.

Weers, P. M., Abdullahi, W. E., Cabrera, J. M. & Hsu, T. C. Role of buried polar residues in helix bundle stability and lipid binding of apolipophorin III: destabilization by threonine 31. *Biochemistry* (2005) 44, 8810-8816.

Weers, P. M. & Ryan, R. O. Apolipophorin III: role model apolipoprotein. *Insect Biochem Mol Biol* (2006) 36, 231-240.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* (2008) 452, 872-876.

Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* (2004) 11, 377-394.

Ylikallio, E., Poyhonen, R., Zimon, M., De Vriendt, E., Hilander, T. *et al.* Deficiency of the E3 ubiquitin ligase TRIM2 in early-onset axonal neuropathy. *Hum Mol Genet* (2013) 22, 2975-2983.

Zuchner, S., Mersiyanova, I. V., Muglia, M., Bissar-Tadmouri, N., Rochelle, J. *et al.* Mutations in the mitochondrial GTPase mitofusin 2 cause Charcot-Marie-Tooth neuropathy type 2A. *Nat Genet* (2004) 36, 449-451.

Zuchner, S. & Vance, J. M. Molecular genetics of autosomal-dominant axonal Charcot-Marie-Tooth disease. *Neuromolecular Med* (2006) 8, 63-74.

http://www.omim.org/

http://genetics.bwh.harvard.edu/pph2

http://www.mutationtaster.org

http://sift.jcvi.org

http://genome. ucsc. edu/

http://www.ncbi.nlm.nih.gov/projects/SNP/

http://evs.gs.washington.edu/EVS/

http://browser.1000genomes.org/

https://genomics.med.miami.edu/

http://www.ebi.ac.uk/Tools/pfa/iprscan

http://protein.bio.unipd.it/homer/

http://www.pymol.org