

UNIVERSITÀ DEGLI STUDI DI GENOVA
SCUOLA DI SCIENZE MEDICHE E FARMACEUTICHE



TESI DI LAUREA MAGISTRALE IN MEDICINA E CHIRURGIA

Dipartimento di Neuroscienze, Riabilitazione, Oftalmologia, Genetica e Scienze Materno-Infantili (DINOEMI)

**“WHOLE-EXOME SEQUENCING IN A TERTIARY PEDIATRIC
NEUROLOGY CENTER: A PILOT STUDY”**

Relatore
PROF. VINCENZO SALPIETRO

Candidato
ELISA CALÌ

Anno accademico 2019-2020

INDEX

1. INTRODUCTION	5
1.1 SPECIFIC AIMS OF THIS STUDY	5
1.2 DNA SEQUENCING TECHNOLOGIES	6
1.2.1 HISTORY OF DNA SEQUENCING	6
1.2.2 SANGER SEQUENCING OR FIRST GENERATION SEQUENCING	7
1.2.3 NEXT GENERATION SEQUENCING	11
1.2.4 SGS WORKFLOW.....	13
1.2.5 SANGER VS SGS.....	19
1.2.6 TARGETED SEQUENCING	20
1.2.7 GENOME OR EXOME?	21
1.2.8 WHOLE EXOME SEQUENCING	23
1.2.9 HOW NEXT-GENERATION SEQUENCING IS CHANGING THE FIELD OF CLINICAL GENETICS AND NEUROGENETICS	26
1.2.10 IMPACT OF NGS ON GENETIC COUNSELING	26
1.3 DEEP PHENOTYPING	29
1.3.1 DEFINITION OF PHENOTYPE.....	29
1.3.2 HUMAN PHENOTYPE ONTOLOGY IN DEEP PHENOTYPING	30
1.4 “DEEP LEARNING”: THE USE OF MACHINE LEARNING IN THE FIELD OF GENETICS	35
1.5 REVERSE PHENOTYPING	36
1.6 HEREDITARY NEUROLOGICAL DISORDERS	38
1.6.1 CLASSIFICATION	39
2. WHOLE EXOME SEQUENCING IN A TERTIARY PEDIATRIC NEUROLOGY CENTRE: A PILOT STUDY	41
2.1 STUDY DESIGN	41
2.1.2 INCLUSION CRITERIA	42
2.1.3 EXCLUSION CRITERIA	42
2.2 MATERIALS AND METHODS	43
2.2.1 PHENOTYPE DOCUMENTATION	43

2.2.2	<i>GENOTYPE DOCUMENTATION</i>	44
2.2.3	<i>DATA ANALYSIS: ALIGNMENT</i>	45
2.2.4	<i>DATA ANALYSIS: VARIANT CALLING AND ANNOTATION</i>	45
2.2.5	<i>REVERSE PHENOTYPING</i>	49
2.3	RESULTS	50
2.3.1	<i>STUDY PARTECIPANTS</i>	50
2.3.2	<i>RESULTS OF PHENOTYPING</i>	50
2.3.4	<i>RESULTS OF GENOTYPING</i>	52
2.4	DISCUSSION	58
2.4.1	<i>CASES OF PARTICULAR INTEREST: PATHOGENIC VARIANTS IN KNOWN DISEASE GENE</i>	58
2.4.2	<i>CASES OF PARTICULAR INTEREST: VARIANTS IN A CANDIDATE GENE</i>	63
3.	CONCLUSION AND IMPLICATIONS FOR CLINICAL CARE.	68
	LIMITATIONS OF THE STUDY.	70
4.	REFERENCES	71
5.	WEB RESOURCES AND TOOLS	85
6.	FIGURES	86
7.	TABLES	87
8.	ACKNOWLEDGMENTS.	88

ABSTRACT

Background: Neurogenetic disorders are a clinically heterogeneous group of diseases with frequent infantile and childhood onset. Monogenic causes can be identified in a large proportion of these disorders, particularly with the advent of whole-exome sequencing (WES), improving the diagnostic yield in genetically undetermined patients. In this study we used WES in a pediatric tertiary care center (Pediatric Neurology and Muscular Diseases Unit, “G. Gaslini” Institute) to reach a molecular diagnosis in children with complex neurodevelopmental conditions.

Methods: WES was performed in 45 children diagnosed with rare and genetically undetermined neurodevelopmental conditions. DNA was extracted from peripheral blood and sequenced with NextSeq 500. Raw data were analyzed according to a custom pipeline developed at “G. Gaslini” Institute and based on Burrows-Wheeler Alignment (BWA), Genome Analysis Toolkit (GATK), and ANNOVAR. The variants were classified following the updated guidelines from American College of Medical Genetics (ACMG). Candidate variants were validated through Sanger sequencing and filtered according to family segregation, population genetics, and phenotype prediction programs.

Results: In 12/45 affected children, pathogenic/likely pathogenic variants have been identified in causative genes, achieving a diagnostic yield of 34,2%. In 5/45 affected children, WES led to the identification of variants of unknown significance (VUS), and in 5/45 a children a variant was found in a novel (candidate) gene. WES data from 10/45 children are currently being analyzed.

Conclusion: This pilot study confirms the high diagnostic potential of WES, with a good detection rate compared with the results obtained in similar studies. WES is an effective diagnostic tool which facilitate the identification of disease-causing variants in known genes as well as novel disease genes. The results support the importance of accurate phenotyping and highlight how genetic diagnosis can significantly improve clinical management and etiologically targeted treatments.

1. INTRODUCTION

1.1 Specific aims of this study

Whole exome sequencing (WES) is the targeted sequencing of the subset of the human genome that is protein coding.

The use of next-generation DNA sequencing (NGS) has become more and more common over the last 15 years, reducing the cost of DNA sequencing and offering a wide analysis of our DNAs. The WES can provide a map of all coding variations present in an individual human genome, and it is a powerful and cost-effective new tool for dissecting the genetic basis of Mendelian disorders, some of which have proved to be intractable to conventional gene-discovery strategies.¹

Thanks to this new approach, the field of personalized medicine is widening: deep phenotyping is one essential step in the genetic workflow and it has improved with the use of some particular tools (HPO, OMIM, DECIPHER).

In my thesis we performed Whole Exome Sequencing on a cohort of children with undiagnosed neurological illness presumed to be of genetic origin. The aims of this study are :

- Testing the utility of Whole Exome Sequencing in a pediatric tertiary care center and evaluating the impact on research settings.
- Achieving a definitive genetic diagnosis in patients with complex medical condition to improve their clinical management
- Creating a practical procedure for phenotyping and genotyping that would achieve a high rate of diagnosis for unspecific neurological illnesses of childhood.
- Understanding the value of detailed phenotypic characterization (deep phenotyping and HPO codes) in order to define new diseases and to improve the knowledge of the relationship between genetics and disease phenotypes.
- Dissect the opportunities for the future: treatment options, ethical issues, practical considerations.

1.2 DNA sequencing technologies

1.2.1 HISTORY OF DNA SEQUENCING

The first attempts in sequencing the structure of the DNA were made by Sanger (Figure 1). He was very determined to find a way to understand the chemical structures of important biological molecules. Insulin was the first protein to have its structure revealed², in the early 50s, by Sanger itself. The roots of DNA sequencing dates back in the 1970s, when Sanger and colleagues³ and Maxam and Gilbert⁴ developed methods to sequence DNA by chain termination and by chemical cleavage, respectively.

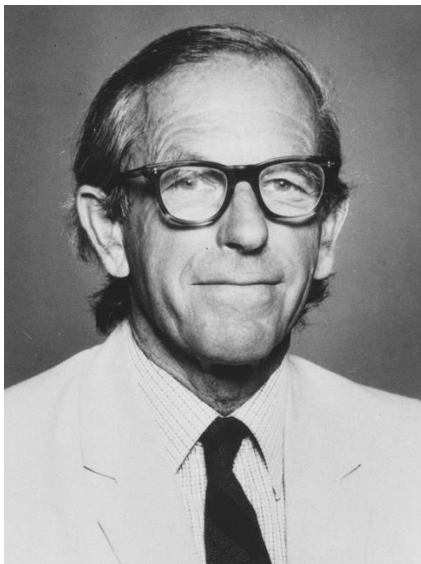
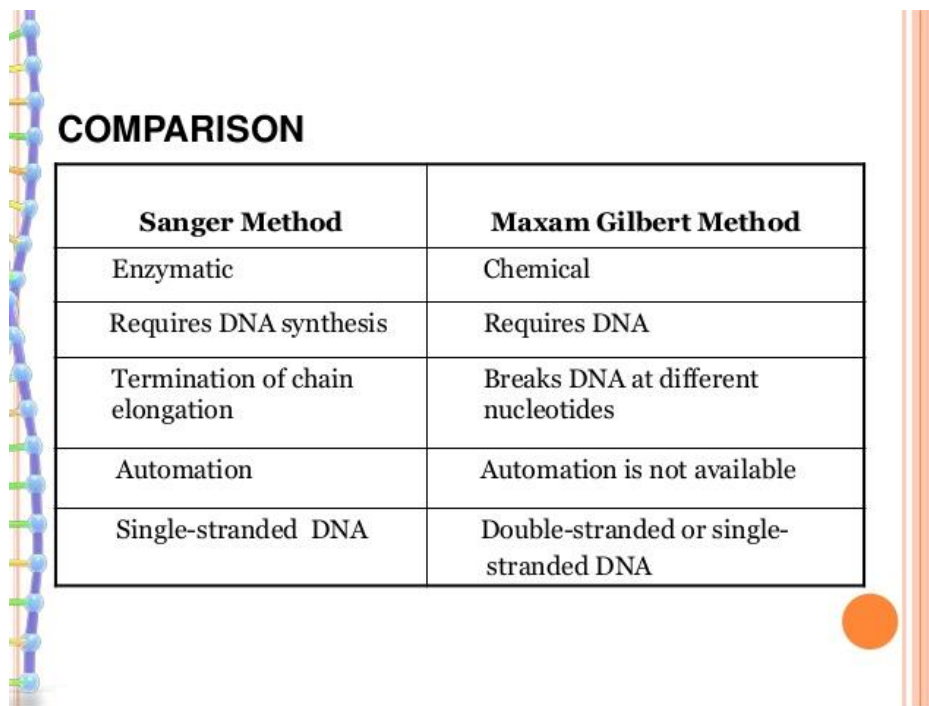


Figure 1 Frederick Sanger⁵

The Maxam-Gilbert method is based on chemical modification of a terminally labeled DNA-restriction fragment and the consequent partial cleavage backbone at sites adjacent to the modified nucleotides.

The technique developed by Sanger and colleagues required less handling of toxic chemicals and radioisotopes than Maxam and Gilbert's, and as a result it became the prevailing DNA sequencing method for the next 30 years (Figure 2). Since the

1990s, DNA sequencing has almost exclusively been carried out with capillary-based, semi-automated implementations of the Sanger biochemistry.^{6,7}



COMPARISON

Sanger Method	Maxam Gilbert Method
Enzymatic	Chemical
Requires DNA synthesis	Requires DNA
Termination of chain elongation	Breaks DNA at different nucleotides
Automation	Automation is not available
Single-stranded DNA	Double-stranded or single-stranded DNA

Figure 2 Sanger Method vs. Maxam Gilbert Method⁸

1.2.2 SANGER SEQUENCING OR FIRST GENERATION SEQUENCING

Sanger sequencing is the process of selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication.

In the first step target DNA is denatured and annealed to an oligonucleotide primer. The single stranded primer is then extended by DNA polymerase using a mixture of deoxynucleotide triphosphates (normal dNTPs) and chain-terminating dideoxynucleotide triphosphates (ddNTPs). Each round of primer extension is randomly terminated by the incorporation of radioactively or fluorescently labeled ddNTPs, whether the sequence is determined in sequencing gels or automated sequencing machines, respectively. The peculiarity of the ddNTPs is to lack the 3' OH group to which the next dNTP of the growing DNA chain is added.

Thus, no more nucleotides can be added, resulting in termination of the growing DNA and falling off of the DNA polymerase. This is called “chain termination event” (Figure 3).

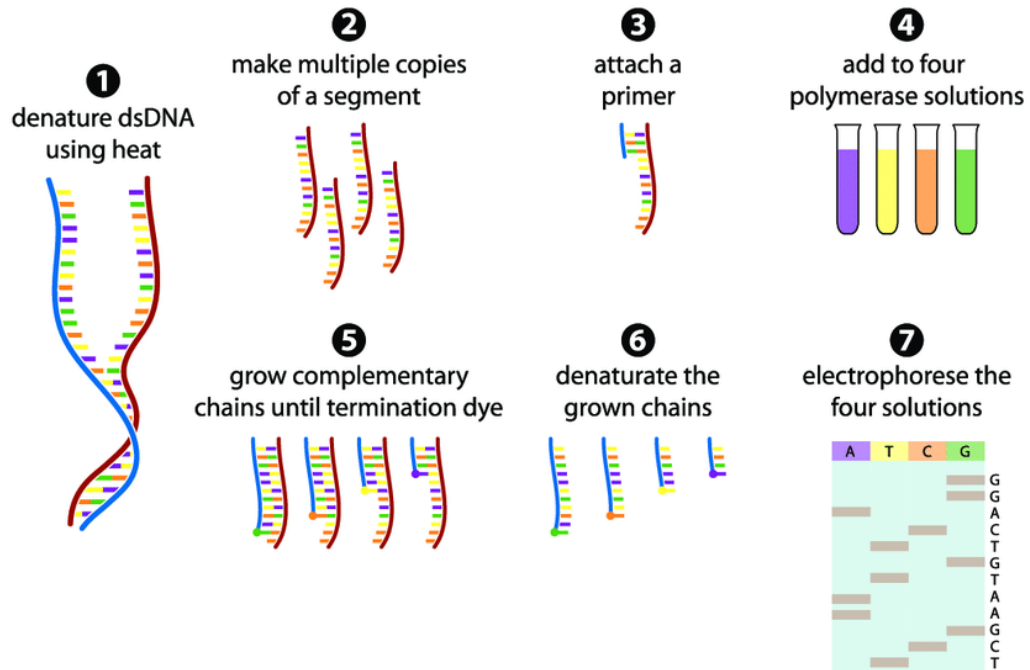


Figure 3 Sanger sequencing method in 7 steps. Courtesy of Michael G Gauthier⁹

The newly synthesized DNA chains will be a mixture of lengths, depending on how long the chain was when a ddNTP was randomly incorporated.¹⁰ The products can be then separated and run on a polyacrylamide/urea gel, that is dried onto chromatography paper and exposed to X-ray film (autoradiograph, Figure 4).³

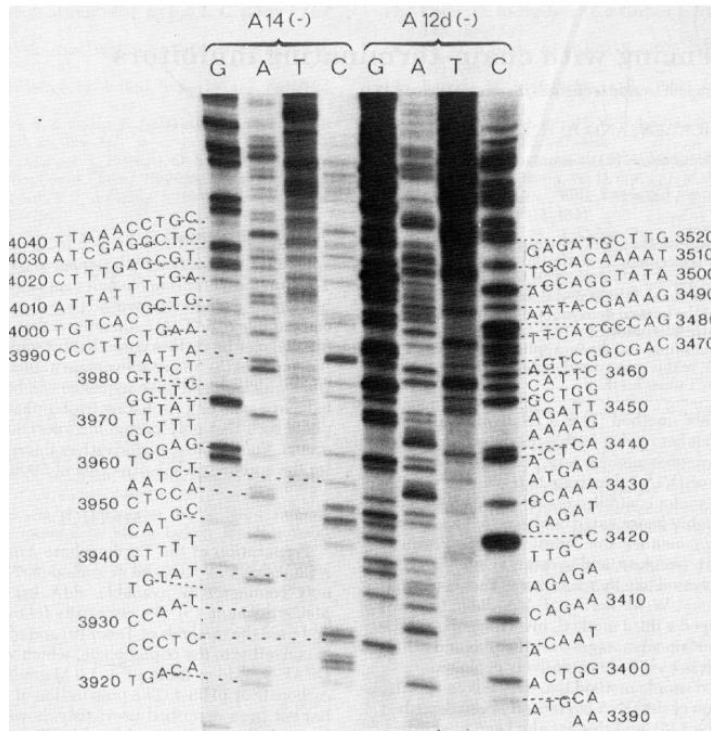


Figure 4 Autoradiograph ³

However, most sequencing is now automated. In this case, each ddNTP is labeled with a different fluorescent marker, and every of them runs in the same lane. A machine reads the lane with a laser, the products are detected and the fluorescence intensity translated into a data “peak”, creating a chromatogram (Figure 5)

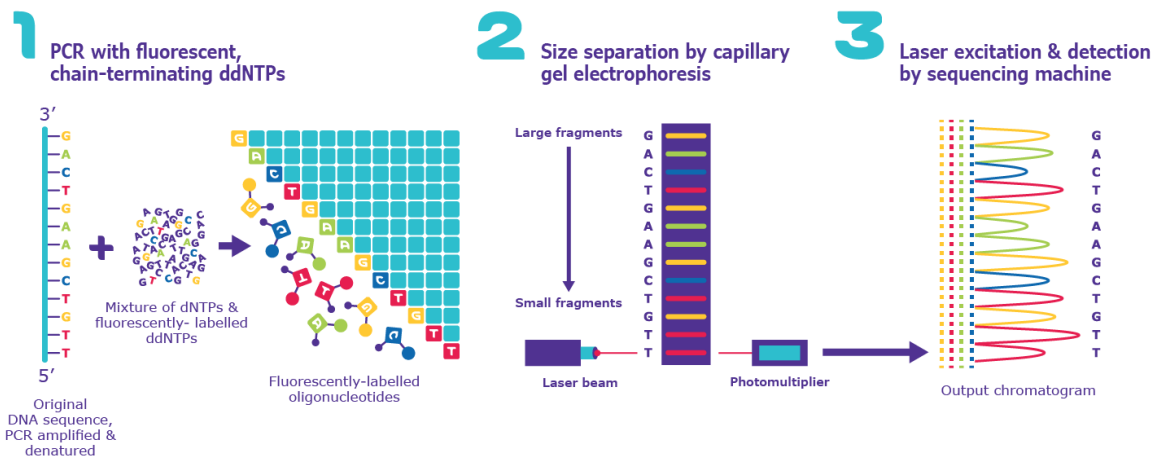


Figure 5 Steps of automated Sanger sequencing. Courtesy of www.sigmaldrich.com¹¹

To date, the most significant innovations in Sanger sequencing have been:

- the development of fluorescent (terminator) dyes
- the use of thermal-cycle sequencing and thermostable polymerases
- software developments to interpret and analyze the sequences.

The leader in automated Sanger sequencing is Applied Biosystems (AB). The current commercialized AB sequencers all utilize fluorescent dyes and capillary electrophoresis (CE). The machines differ in capacity, from 4 capillaries (SeqStudio Genetic Analyzer), to 8–24 (3500 Series Genetic Analyzer), to 48–96 (3700 Series Genetic Analyzer) for DNA sequencing or fragment analysis protocols.¹²

1.2.3 NEXT GENERATION SEQUENCING

The term next generation sequencing describes “highly parallel or high-output sequencing methods that produce data at or beyond the genome scale”¹³.

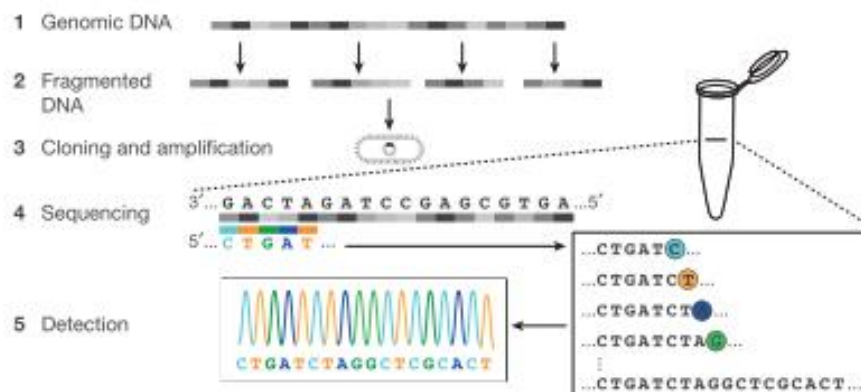
The word “next” relates to an important step forward in the development of DNA sequencing technologies. As with the suggestion that in the future there will be a new technology considered as the next step, many authors prefer to use the term second generation (or third generation as well) . For some authors, the use of the AB sequencing automated machine technology as a development of the original Sanger sequencing¹² is considered as a second generation technology.

Thanks to the advances made in the Sanger sequencing, with the completion of the Human Genome Project, the first human genome sequence has been identified in 2004.¹⁴ However, this project required vast amounts of time and resources, thus the National Human Genome Research Institute (NHGRI) started a funding program aimed to reduce the cost of human genome¹⁵ For this reason, many other technologies have emerged in order to satisfy the need of faster, higher throughput sequencing of large genomes at the lower cost possible (Figure 6).

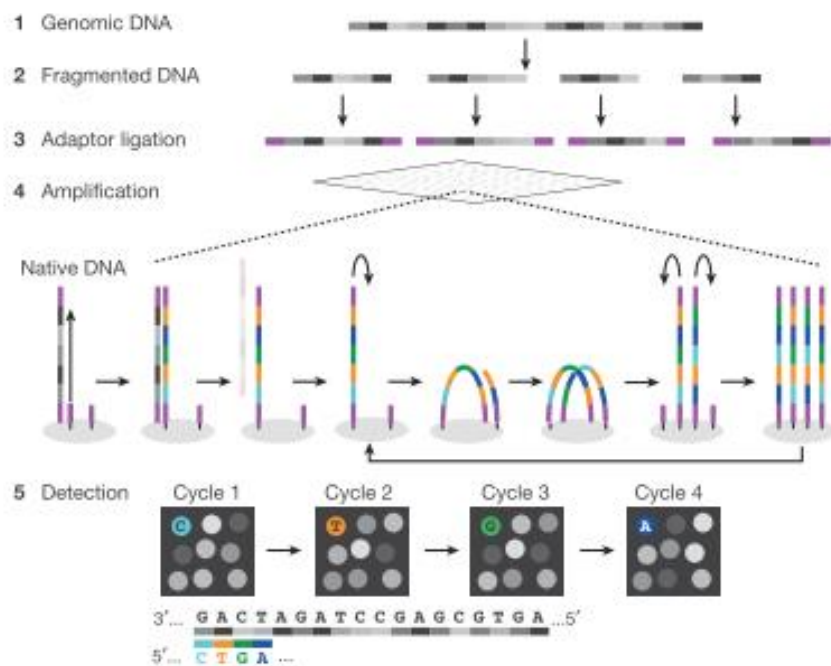
Second generation sequencing (SGS) methods can be grouped into two major categories, sequencing by ligation and sequencing by synthesis (SBS), mainly classified into four major sequencing platforms: Roche/454 launched in 2005, Illumina/Solexa in 2006 , ABI/Solid in 2007 and Ion Torrent in 2011. Illumina is by far the most important.

Third generation sequencing can be divided into three main categories: sequencing by synthesis, nanopore sequencing technologies and microscopy-based approaches.

First generation sequencing (Sanger)



Second generation sequencing (massively parallel)



Third generation sequencing (Real-time, single molecule)

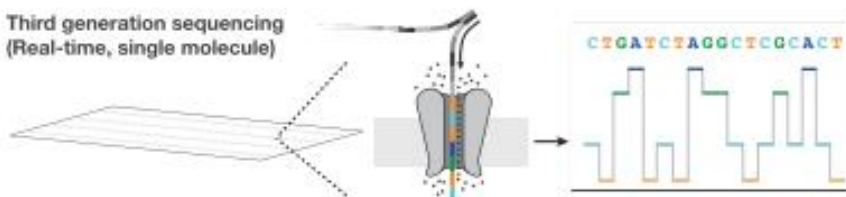


Figure 6 DNA sequencing technologies. ¹⁶

1.2.4 SGS WORKFLOW

In Second Generation Sequencing, instead of using one tube per reaction, it is created a complex library of DNA templates, each of them will be then amplified and ready to be sequenced.

The first step is to create the library. The source could be genomic DNA, immunoprecipitated DNA, reverse-transcribed RNA or cDNA. All of them has to be converted into DNA small molecules.

The next important step is to prepare the library: DNA is fragmented, terminal overhangs are repaired and platform specific synthetic oligonucleotides are attached to the ends of the fragments to facilitate sequencing reactions, working as universal priming sites. (Figure 7)

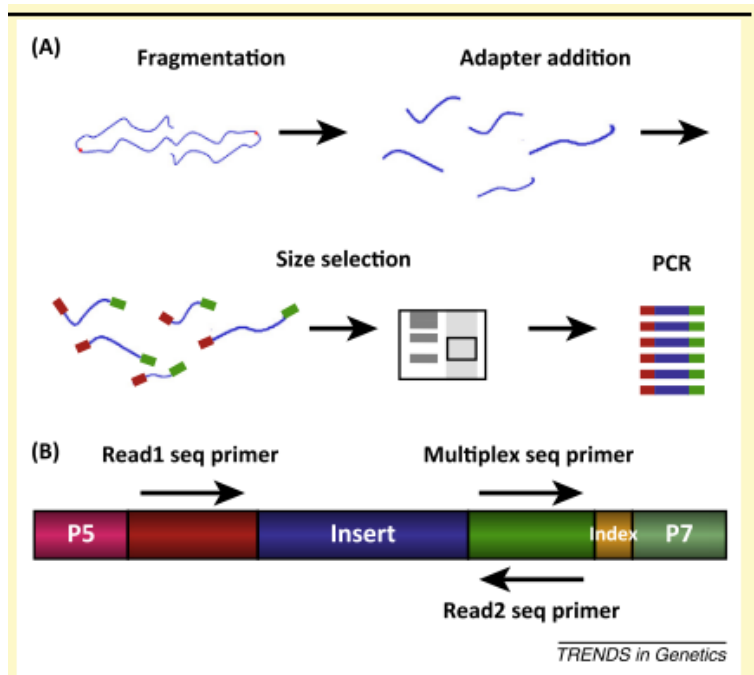


Figure 7 Typical NGS library preparation workflow. ¹⁷

Commonly the fragments are enriched for specific genes of interests (targeted sequencing) or for all coding regions (Whole exome sequencing).¹⁸

The templates are then amplified by either water in oil bead based PCR (Roche 454) or solid surface (generally a glass slide) bridge amplification (Illumina) (Figure 8)

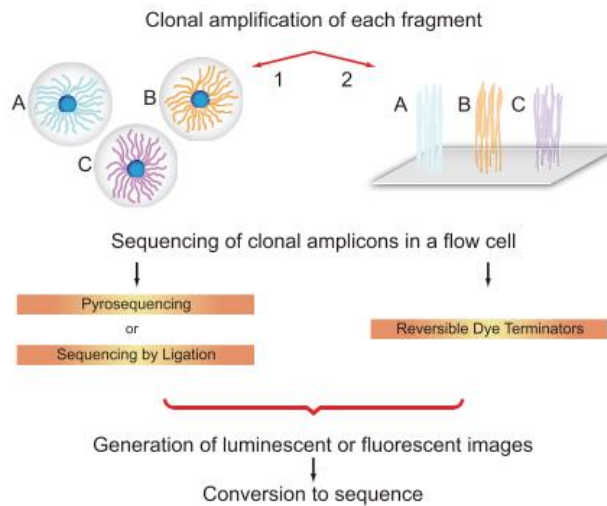


Figure 8 Amplification of the templates¹⁹

The last step includes the sequencing reaction, in which luminescent or fluorescent images are generated and processed into sequence short reads (35-400 bp) to be analyzed. The raw sequencing data consists of large computer text files (tens to hundreds of gigabytes), generally FASTQ files, containing several million reads.

The data analysis consists of two processes: the alignment or mapping, to determine the position of the reads, and the variant calling, to compare each nucleotide with its counterpart in the reference genome.

ALIGNMENT

The most crucial step of most NGS analysis pipelines is to map reads to sequences to the reference genome. First, it is identified a small set of places in the reference sequence where the sequence is most likely to accurately align to. Then accurate alignment algorithms are run on the subset of possible mapping locations.

After mapping the sequence FASTQ file to the reference genome, you will end up with a SAM or BAM alignment file. SAM stands for Sequence Alignment/Map

format, while a BAM file is the binary version of a SAM file (saving storage and faster manipulation).

VARIANT CALLING.

Each difference from the reference (mismatch, insertion or gap) is called variant.¹⁸

A variant call is a conclusion that there is a nucleotide difference vs. some reference at a given position in an individual genome or transcriptome, often referred to as a Single Nucleotide Polymorphism (SNP). There are, however, other variants, known as structural variants (SVs), that are large genomic alterations, where large is typically (and somewhat arbitrarily) defined as encompassing at least 50 bp. Copy number variations (CNVs) are a particular subtype of SVs mainly represented by deletions and duplications.²⁰⁻²²

The call is usually accompanied by an estimate of variant frequency and some measure of confidence.

Similar to other steps in this workflow (Figure 9), there are a number of tools available for variant calling, that can predict the pathogenicity of variants and can therefore allow variants that are predicted to be benign to be removed. Genome Analysis Toolkit(GATK) is available for SNVs, while CNVnator12 is for structural variants.^{23,24}

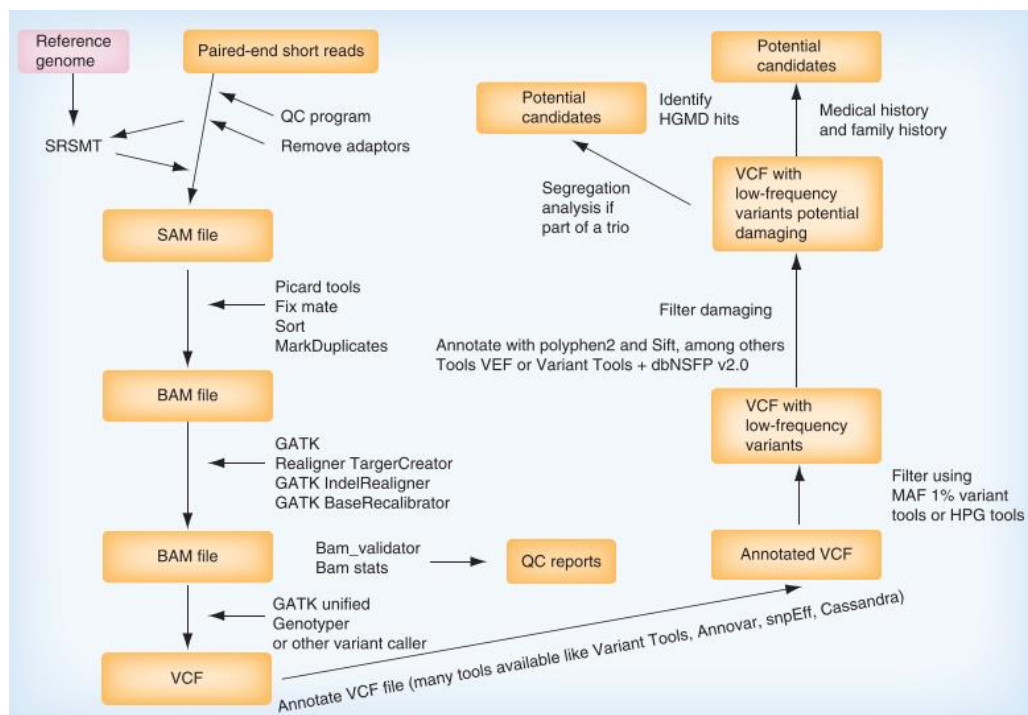


Figure 9 Variant calling and annotation workflow.²²

FILTERING AND ANNOTATION.

One of the main hurdles to overcome in NGS is the sheer number of variants that are present compared to the reference sequence. Based on literature, there can be anywhere from 20-30.000 variants in a single exon sequence. Filtering these results further requires a set of assumptions about which variants are more likely to be deleterious.

Thus, after variant calls are generated, researchers need to understand the functional content within the data and therefore perform prioritization analysis on all variants for functional follow-up on selected variants. For this reason, the variants have to be annotated according to their potential effects on genes and transcripts; this requires the translation of variant-describing semantics in the Variant Call Format (VCF)²⁵, which reflects the chromosomal coordinates of each variant into gene-based variant annotations.²⁶²⁷

One of the most important annotation tools is ANNOVAR(ANNOtate VARIation). It supports three different types of annotations: gene-based, region-based and filter based; gene based annotations tell the functional impact on known genes, region-based focus on the relationship with different specific genomic regions, and filter based annotations gives information of the variant, such as frequency, prediction scores, that can be used to filter the non-deleterious variants(Figure 10).²⁸

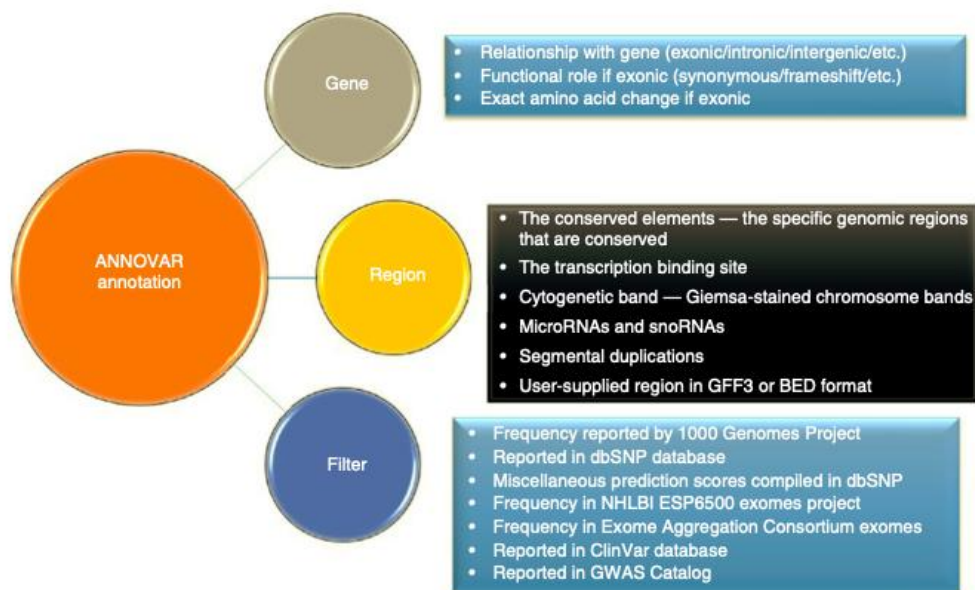


Figure 10 ANNOVAR different types of annotations.²⁸

There are some strategies that can help reduce the number of variants:

- Filtering for rare variants: variants reported to be common in the general population are not likely to be responsible for Mendelian disease. Such variants can be found in databases such as dbSNP, the 1000 Genome Project, and in-house exome databases. This approach will remove the most common variants by setting the MAF, maximum and minor allele frequency, to 1,0% and 0,1% respectively.
- Prediction of their deleteriousness and pathogenicity: a greater weight may be given to nonsense and frameshift mutations, as they are predicted to result in a loss of protein function and are heavily enriched among disease-causal variation. If possible, functional studies can be performed on tissue samples to confirm the physiologic effects of the mutation, such as reduced enzyme activity.
- Filtering based on cross-reference gene databases: once variants of interest have been identified, it can also be useful to determine whether the gene is one that is conserved across evolution, and therefore a more functionally important gene, using the UCSC Genome Browser. The mutation(s) of greatest interest can then be confirmed using Sanger sequencing, particularly if the read coverage is relatively low.
- Filtering by mode of inheritance: if there is enough medical history data to theorize a mode of inheritance, or an etiological diagnosis can be made from the phenotype of the patient, this can provide another filter by which to narrow candidate genes.
- Filtering by pedigree information: for Mendelian disorders, the use of pedigree information can substantially narrow the genomic search space for candidate causal alleles.
- Clinical evaluation: in the event that multiple unrelated individuals with the same phenotype are available for sequencing, comparison of their common variants can be extremely useful as a filter.(Figure 11)

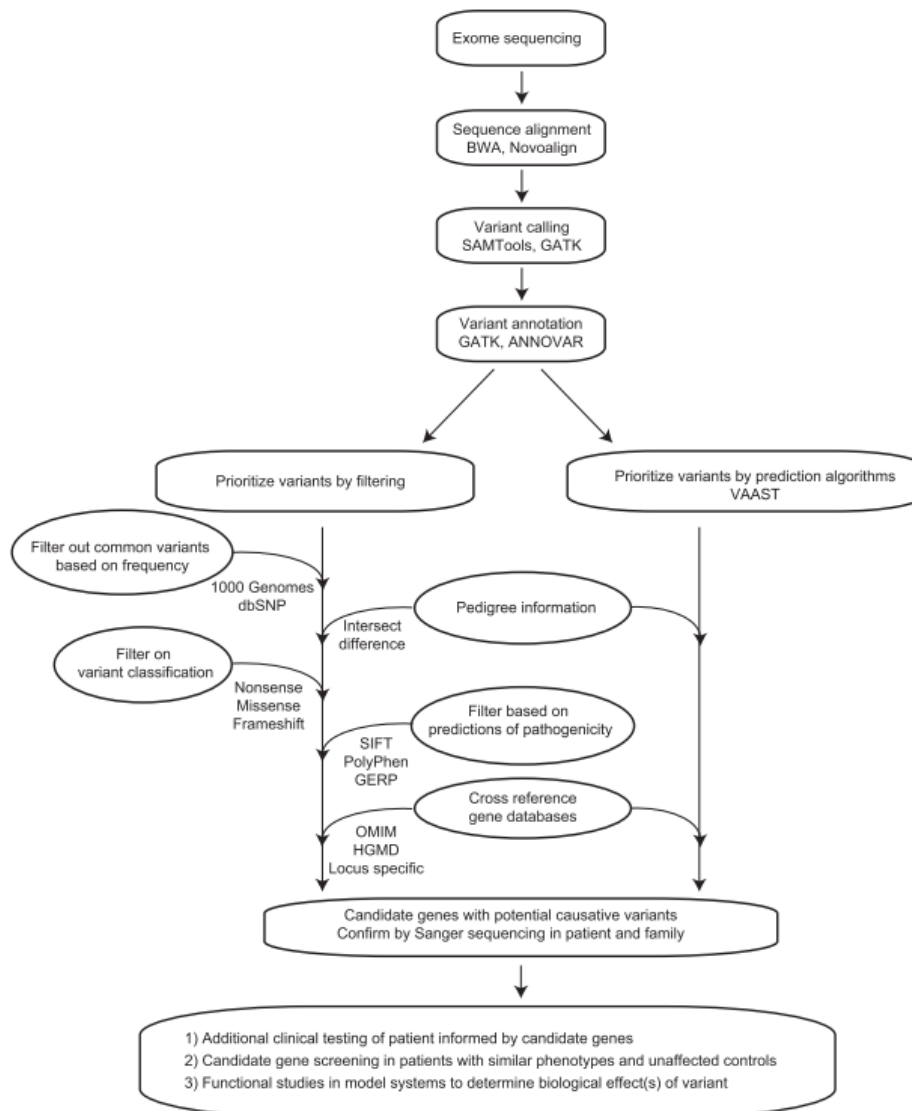


Figure 11 Variant filtering pipeline. ²⁹

The manual examination of the filtered candidates by a geneticist represent the last step of the interpretation process. This is performed through the visual inspection of detailed annotations, especially OMIM, and review of the literature. At the end of the interpretation process, the variants of interest are classified according to the ACMG guidelines³⁰, that distinguish:

- Pathogenic variants (V class)
- Likely pathogenic variants (IV class)
- Variant of unknown significance (VUS, III class)
- Likely benign (II class)
- Benign variants (I class).

1.2.5 SANGER VS SGS

Compared to Sanger sequencing, SGS:

- a. Offers the possibility to generate massive volume of data vastly increasing the output and capturing a broader spectrum of mutations. The spectrum of DNA variation in a human genome comprises small base changes (substitutions), insertions and deletions of DNA, large genomic deletions of exons or whole genes and rearrangements such as inversions and translocations. Traditional Sanger sequencing is restricted to the discovery of substitutions and small insertions and deletions. For the remaining mutations, dedicated assays(FISH, CGH, microarrays) are performed. NGS sequencing can derive these data directly, obviating the need for dedicated assays while harvesting the full spectrum of genomic variation in a single experiment. ³¹
- b. reads more than one billion short reads in single run (short-read technology)
- c. offers faster and cheaper sequencing: the preparation of samples in NGS requires less than 1 hour. Sanger requires, on the other hand, many processes that consume days or weeks (depending on the genomic size)³²

However, the reads of NGS are largely shorter in length than the ones from original Sanger methods with relatively higher sequencing errors in the reads.³³ Distinguishing real variants from background noise can be a challenge, particularly when there isn't a high depth of coverage.

Besides, NGS has limitations in regions which sequence map erroneously due to extreme guanine/cytosine content or repeat architecture (i.e. expansions in fragile x syndrome or Huntington disease) ³¹, with no difference in the enrichment method used.³⁴

Lastly, although SGS is cheaper and faster in comparison to traditional Sanger, it is not affordable to all labs and remain time-consuming in many of its aspects (i.e. the SGS technologies generally require PCR amplification step which is a long procedure in execution time and expansive in sequencing price and can also bias the procedure).³⁵

Third generation sequencing (TGS)³⁶ fulfills the above gaps of NGS technologies and allows direct sequencing of single molecules.

Single molecule sequencing (SMS)³⁷ technologies have the ability to sequence with longer read lengths keeping the cost and time lower without compromising the quality. They can be grouped into three main categories: sequencing by synthesis, nanopore sequencing technologies and microscopy-based approaches.

Two technologies are currently dominating: Pacific Biosciences' (PacBio) single-molecule real-time (SMRT) sequencing and Oxford Nanopore Technologies' (ONT) nanopore sequencing.^{38,39}

1.2.6 TARGETED SEQUENCING

There are three NGS approaches to improve diagnostics for heterogeneous diseases: targeted sequencing enriched of a set of genes, whole exome sequencing (WES), whole genome sequencing (WGS).⁴⁰

Targeted sequencing^{34,41} consists in subset of genes, called gene panel, useful in discovering the point mutations, insertion or deletions, gene rearrangements and variations occurring in copy number (Table 1).

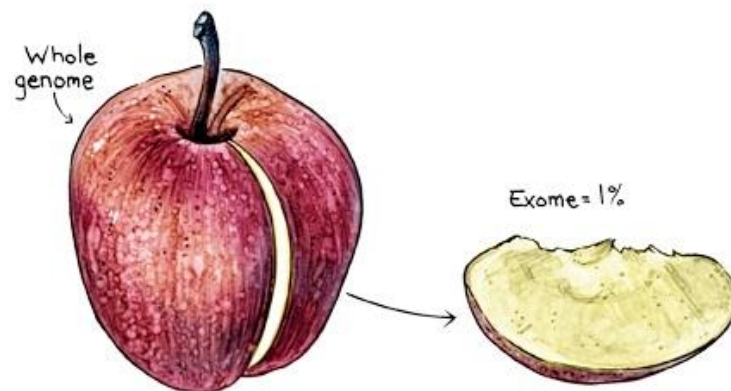
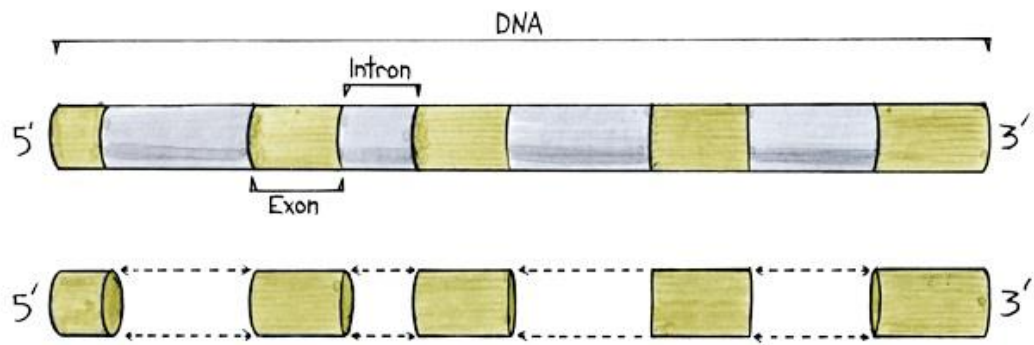
Table 1 Advantages of targeted sequencing. Schnekenberg, R. P. & Németh, A. H. Next-generation sequencing in childhood disorders. *Arch. Dis. Child.* 99, 284–290 (2014).¹⁸

Advantages of targeted sequencing
<ul style="list-style-type: none">▶ possibility of customization and optimization of the target regions;▶ more affordable benchtop sequencers can be used;▶ higher average depth of coverage;▶ simpler Information Technology (IT) infrastructure for data processing and analysis;▶ fewer variants to interpret;▶ possibly shorter turnaround time in a diagnostic setting.▶ novel, relevant genes can simply be added to the gene panel by a small modification of the bioinformatic analysis pipeline⁴²

Targeted NGS of a disease-specific gene panel has an equal or better quality than Sanger Sequencing and it can therefore be reliably implemented as a stand-alone diagnostic test.^{43,44}

1.2.7 GENOME OR EXOME?

The human genome is composed of roughly 3 billion nucleotide base pairs arranged into approximately 30,000 genes. Each gene contains protein-coding and non-coding regions. The exome comprises all coding regions: the exons contain information for the construction of the amino-acid sequence of the protein. Non-coding regions include introns and the 3'-5' regions of the gene (Figure 12). Most variation between humans occurs in the non-coding DNA regions and in degenerate positions in amino acid codons that do not change the intended identity of the amino acid.



Copyright © 2012 University of Washington

Figure 12 The exome is only 1% of the entire genome. Courtesy of University of Washington.

Humans vary on average ever 1 out of 100 nucleotides and most of these variations occur frequently in the population with little or no effect on protein function. As such, they are called polymorphisms. Mutations in the genetic sequence are more likely to have deleterious effects if they result in a shift of the reading frame, non-synonymous substitution of one amino acid for another (particularly amino acids with vastly different chemical properties), insertion of a premature stop codon resulting in a truncation of the protein product, or loss of a stop codon.

Of course the best and more direct approach to detect all variations would be sequencing the whole genome, providing the widest coverage and allowing precise calling of structural variants⁴⁵. However, despite the coding regions are only 1-1,5% of the human genome, this portion houses approximately 85% of disease-causing mutations.⁴⁶ Also, the interpretation of the functional effects of a mutation in a non-coding area is very difficult. For this reasons, the best choice is to focus on the exome.⁴⁷

1.2.8 WHOLE EXOME SEQUENCING

Whole exome sequencing is a sequencing strategy that isolates the protein-coding portion, which represent 2-3% of the genome.

The workflow for WES is the same as SGS (Figure 13).

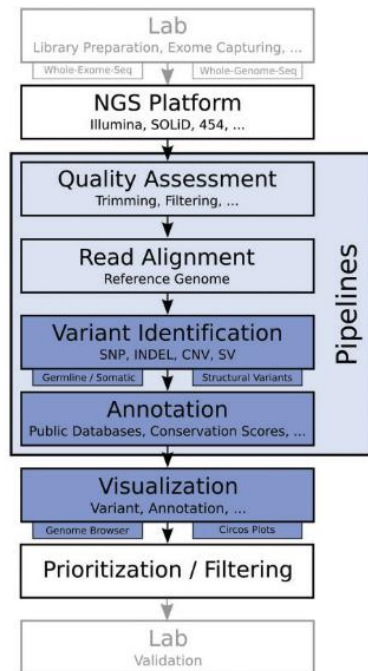


Figure 13 WES workflow.⁴⁸

After library preparation, sequencing, alignment and variant calling⁴⁹, the variants are identified. Detected mutations are annotated to understand the biological relevance, and they can also be filtered and prioritized.

The main purpose of WES is the mutational analysis: detection of SNV, small insertion/deletions and CNVs. Our individual genomes contain variants that can protect or increase susceptibility to a certain disease.⁵⁰ Our exome contains an average of 20,000 SNV, and 95% of them are already known.⁵¹

The challenge of WES is to identify the causal alleles, considering that, filtering and prioritizing the variants by quality criteria, mode of inheritance, pedigree

structure information, phenotype evaluation and locus heterogeneity, the number of causal variants vastly decreases(Figure 14).⁵²

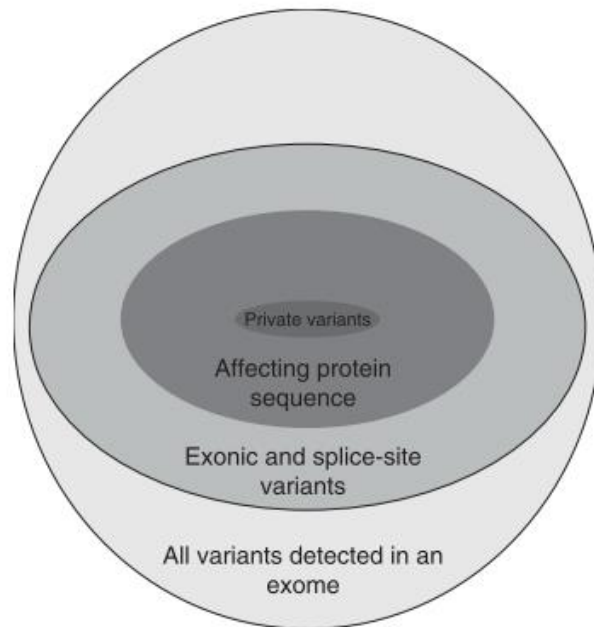


Figure 14 How variant filtering in WES reduces the number of variants.⁵²

After individuating the possible causal genes, it is important to confirm the findings by Sanger sequencing⁴² or targeted resequencing, especially when there are relevant genes that are found to be poorly covered by WES⁵³.

Recently, exon focus microarrays have been created to complement WES analysis, since there are single exons or small CNV that are beyond the detection limit of exome sequencing.²¹

Whole exome sequencing has been increasingly used for rapid discover of new genes for mendelian disorders, with a reported diagnostic rate of 30%.^{51,54-61}

This rate is set to rise, since lack of diagnosis can derive, other than the presence of non-coding mutations, from errors in the variant prioritization, imprecise phenotyping or incomplete gene databases.⁶² In this context a detailed reanalysis can improve the yield, when supplemented with additional family data and with more relaxed variant filtering parameters.⁶³⁻⁶⁵

Considering that potentially causal variants are identified, it can be used in families too small to use linkage, and it can offer diagnosis also in single probands.¹ However, the diagnostic yield is significantly better in trios compared to proband-only testing.⁶⁶

Whereas WES has a lower diagnostic yield in the adult⁶⁷ population, it is particularly efficient in pediatric neurology: in neurodevelopmental disorders it experimented an overall diagnostic rate from 35% up to 49%, and it is considered a useful tool to identify de novo mutations causing intellectual disability or autism and to characterize dystonia or ataxia.⁶⁸⁻⁷⁷

WES has been evaluated also in the field of inherited neuropathies, including CMT, with a diagnostic rate of 24-38%, it is particularly useful in undiagnosed peripheral neuropathies but it could become prevalent taking over IPN gene panels.⁷⁸⁻⁸²

The challenge, though, is still real in diagnosing epilepsies, on account of the high heterogeneity.^{83,84}

It is the perfect bridge-solution between WGS and gene panels: the perfect tool when gene panels have failed and in atypical phenotype presentation.^{18,85}

Although it is often considered at the end of the diagnostic odyssey⁸⁶, in some cases it is most cost effective to directly perform WES as first diagnostic tool.⁶⁰

1.2.9 HOW NEXT-GENERATION SEQUENCING IS CHANGING THE FIELD OF CLINICAL GENETICS AND NEUROGENETICS

There are three main applications of NGS:

1. identification of causative genes in Mendelian disorders (germline mutations)
2. identification of candidate genes in complex diseases for further functional studies and
3. identification of constitutional mutations (somatic mutations).⁴⁸

The NGS consistently improved the diagnostic yield in genetics and childhood neurology.⁸⁷⁻⁸⁹

Through applications of NGS, genotyping chips and comparative genome hybridization array (CGH), chromosomal structural variation or gene copy number variations (CNVs) have been identified for many hereditary neurological disorders. The rapid technological advance in the field is leading quickly to personalized genomic medicine⁹⁰. A major achievement in this area is the possibility, having a deeper knowledge on the functional effects of a mutation, of giving a more accurate therapy, or even target therapy.

1.2.10 IMPACT OF NGS ON GENETIC COUNSELING

DIAGNOSTIC ODYSSEY

WES will likely be first used for those patients of a geneticist who have eluded a diagnosis through all other testing avenues. For the patient, this can mean years or even decades of knowing that they have a condition but not having any information on the name or the advantage of medical literature to guide them in anticipating what they might expect in the future, reproductive impact, or treatment options specifically for that condition.

INFORMED CONSENT

Given the popularity of genomic technology in the media, it is always important to assess the patient's level of knowledge, concerns, and expectations about testing. With whole-exome sequencing in particular, patients may have an unrealistically high expectation of a test that "looks at all of the genes" to deliver an answer or diagnosis. The limitations of current knowledge and testing should therefore be addressed in the informed consent process. WES is a complex test for patients to understand, and even after a thorough explanation by a genetic counselor or researcher patients can have a difficult time understanding.

THE RETURN OF THE RESULTS

One of the main issues is the ambiguity around how to interpret the results, whether report back incidental findings and which ones⁹¹. The ACGM considers as incidental findings (IFs) the "results of a deliberate search of pathogenic or likely path alterations in genes that are not apparently relevant to a diagnostic indication".⁹² To overcome this particular problem, some suggest that the detailed analysis should be limited to genes more likely to be relevant to the disease phenotype under investigation^{93,94}, while others suggest to report the IFs included on a "minimum list"⁹².

Furthermore, also ethical issues have arisen concerning results, i.e. findings of a pathology that is untreatable or of uncertain significance.

WHAT HAPPEN NEXT?

If WES allows to achieve a diagnosis for a rare disorder there are many benefits for the patient and family in both the short and long term. However, only 30% of patients that go through the entire process receive a diagnosis. As WES is entering the clinical practice, genetic counselors will play an important role in helping patients navigate the process and understand the impact of the results on their lives.

Before undertaking WES with a patient, the physician or genetic counselor needs to consider all of these aspects carefully so that they can navigate the process to the best advantage of the patient (Figure 15). For this reason, in the years geneticists and clinicians tried to create guidelines for the application of next-generation sequencing.⁹³

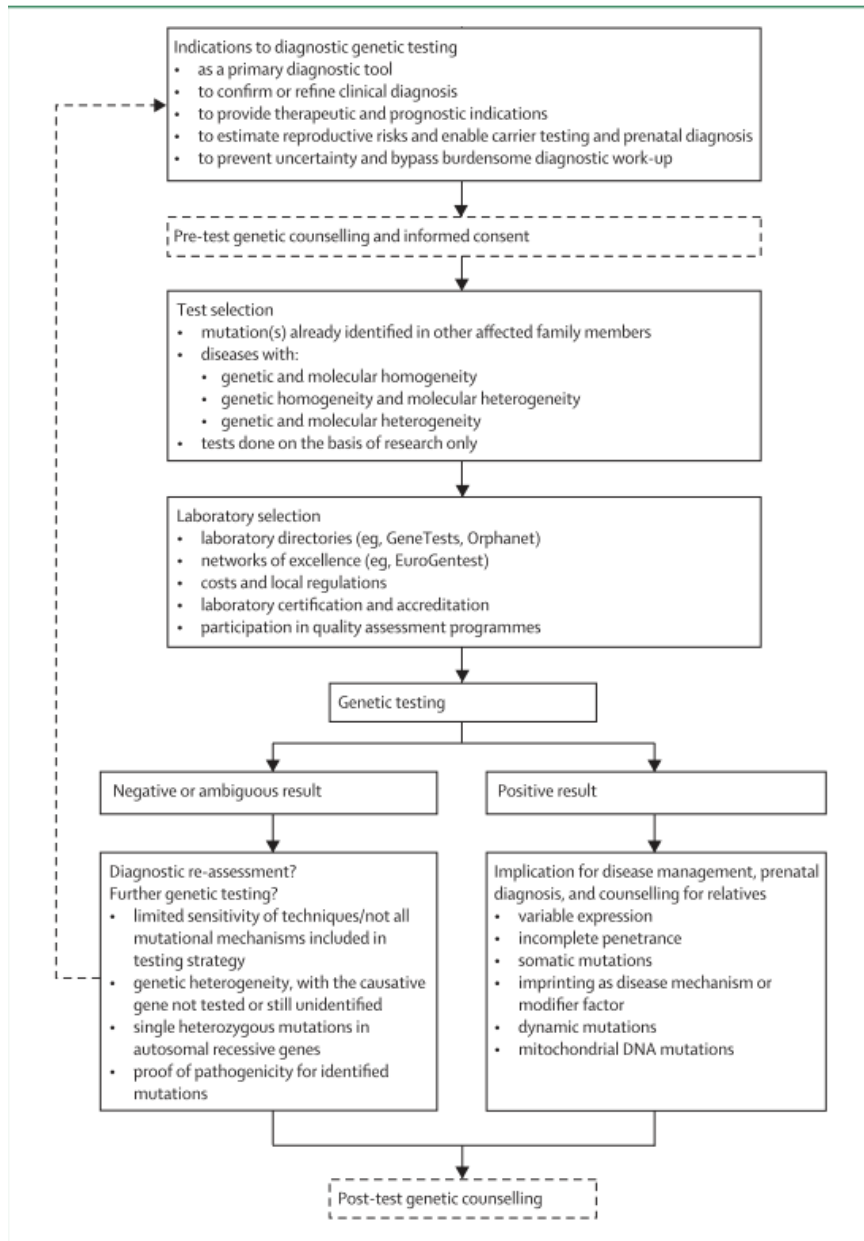


Figure 15 Decision-making flowchart for diagnostic genetic testing in pediatric neurology center, Valente, E. M., Ferraris, A. & Dallapiccola, B. Genetic testing for pediatric neurological disorders. Lancet Neurol. 7, 1113–1126 (2008)⁹⁵

1.3 Deep phenotyping

Precise phenotype analysis has a central role in the mapping of disease genes. It can substantially improve the interpretation of NGS results. For instance, the NGS provides plausible candidate variants, but the diagnosis will require the consequences of these variants to be studied and compared to the clinical findings.

1.3.1 DEFINITION OF PHENOTYPE

The term ‘phenotype’ was originally coined by Wilhelm Johannsen in 1909, together with the term “genotype” to denote presumably distinct realities. One modern definition reads: “The observable structural and functional characteristics of an organism determined by its genotype and modulated by its environment”.⁹⁶ Phenotypes are defined as observable characteristics of organisms.⁹⁷ In clinical domains, the word “phenotype” has the specific meaning of deviation from normal morphology, physiology, or behavior.⁹⁸

The study of the phenotype is essential for our understanding of the physiology and pathophysiology of cellular networks, suggesting groups of genes that work together and their relation with biological process activities, which disruption could lead to the clinical findings.

Therefore, without a good clinical description of the affected individuals, the relevance of the molecular data for diagnosis and treatment is lessened.^{99–101} One of the most important aspects in phenotype data analysis is to precisely measure the similarity between phenotypes. The best way to make full use of phenotypic information is to benefit of a computational approach. To date, many databases have been created, such as OMIM¹⁰², Orphanet¹⁰³ and DECIPHER¹⁰⁴.

1.3.2 HUMAN PHENOTYPE ONTOLOGY IN DEEP PHENOTYPING

Deep phenotyping is a key step in the field of personalized medicine and precision medicine. It has been defined as “the precise and comprehensive analysis of phenotypic abnormalities in which the individual components of the phenotype are observed and described”.⁹⁸ One important role in deep phenotype analysis is given to the computational analysis of scientific and clinical phenotypes narrated in the literature. This approach has gained increasing attention thanks to Human Phenotype Ontology (HPO), a widely used ontology resource, which provides a standardized vocabulary of phenotypic abnormalities encountered in human diseases.¹⁰⁵ An ontology is a philosophical discipline which purpose is to understand how things in the world are divided into categories and how these categories are related together.¹⁰⁶

In computer science, the word ontology is used with a related meaning to describe a knowledge-based structured, automated representation of the entities within a certain domain in fields such as science, government, industry, and healthcare, in which each entity represent a term of the ontology.¹⁰⁷

Phenotype terms in HPO, as in many other ontologies, are organized in a directed acyclic graph (DAG), made of nodes and edges (also called links), in which the edges are one-way and go from one node to another. The nodes of the DAG, also called terms of the ontology, correspond to the concept of the domain. In HPO, as in GO, terms closer to the root are more general than their descendant terms, so that the specificity increases moving toward to lower levels.¹⁰⁸

In the majority of ontologies, the true-path rule applies, also known as annotation propagation rule¹⁰¹: entities annotated with a specific term are also implicitly annotated to all the “parent” terms (the more general ones). Terms in HPO are linked to parent terms by “is a” relationships (or subclass relationships), meaning that they represent one of the subclasses of an ancestor term, but they may also have multiple parents, allowing different phenotypic aspects to be explored.

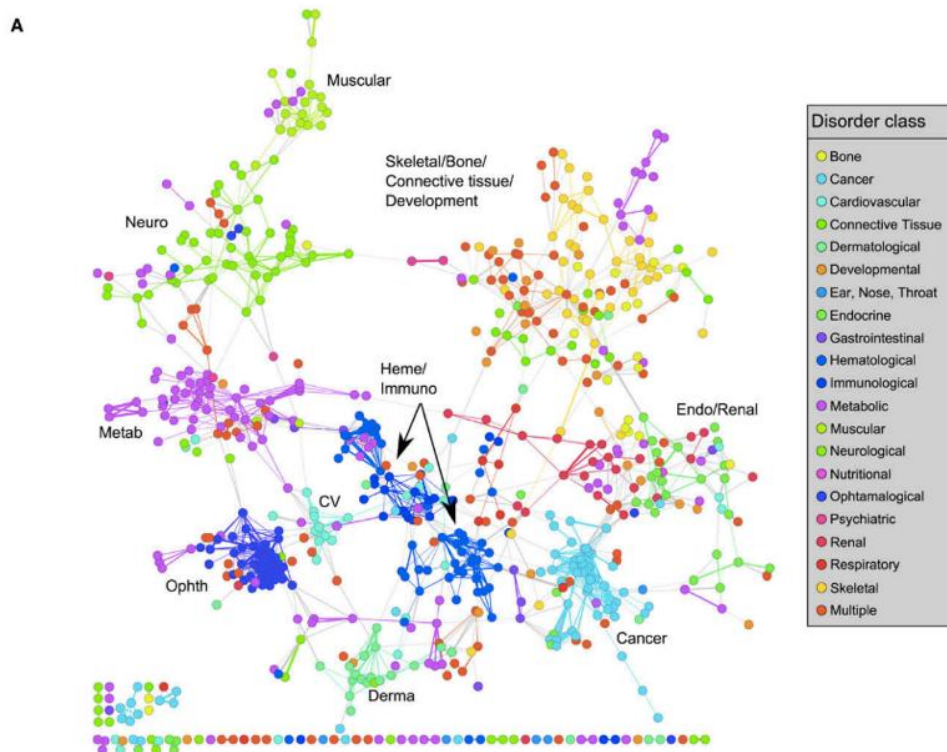


Figure 16 Human Phenotype Ontology network.¹⁰⁵

The HPO project was initiated in 2007 and published in 2008, with the purpose of integrating phenotype information, encountered in human monogenic diseases, across scientific databases. The HPO was originally developed using data from OMIM, downloading omim.txt files from the database. It was then constructed with OBO-Edit¹⁵¹⁰⁹, where synonyms were merged in order to define terms and semantic links were created between them to create an ontology(Figure 16).¹⁰⁵

The HPO is now also using definitions from Elements of Morphology^{110,111} and is also part of the Orphanet database content.

At the moment, the majority of frequency annotations are derived from Orphanet, but a growing number is based on the manual annotation efforts by the HPO team. The main components in which the HPO is constructed are the phenotype vocabulary, the disease annotations and the algorithm that works on them.¹¹²

It is organized as five independent subontologies that cover different categories: ¹¹³

- mode of inheritance, to define diseases according to Mendelian or non-Mendelian inheritance modes,
- onset and clinical course
- phenotypic abnormality, which is the largest and more detailed one
- clinical modifier, as triggering factors, location, severity
- frequency.

Each class of the HPO has a unique identifier (e.g. HP:0001263), a label and a list of synonyms. 65% of them have a precise definition written by clinical experts.¹¹⁴ The class does not represent an individual disease entity but, rather, the phenotypic abnormality associated with it. Besides, the vocabulary is not based on quantitative information, such as “glucose level of 130 mg/dl”, but qualitative information, like “decrease or increase in a certain entity”.

The purpose of the HPO project is to create a common vocabulary that can be used to link mutation data, locus-specific databases^{115,116} and genotype-phenotype databases, in order to help data inter-change between clinical researchers¹¹⁷.

However, recognizing the right phenotype is a real challenge, due to the highly lexical and syntactic variability. The semantic structure of the HPO enables researchers to calculate, via computational methods, the similarity between different terms ^{118–123}.

In 2009, Kohler and his group tried to measure phenotypic similarities between the queries and the diseases annotated with the HPO, implementing the algorithm with Phenomizer. They created a statistical model, giving a p value for each term, describing the probability of having a similar or higher score choosing the same number of query terms.

In clinical practice, the main problems are the “noise” and the “imprecision”. “Noise” relates to the presence of terms describing clinical features unrelated to the underlying disorder. The “imprecision” is the undetailed description of a term, due

to the lack of awareness of the precise terminology or to the inadequacy of clinical or laboratory investigations. It seems that these two factors cause a decrease in performance that is less relevant in ontological approaches than in diagnostic algorithms (i.e. the imprecision causes less errors because of the ability of the ontology to recognize in the imprecise term a similar meaning to the term in the database)¹¹⁹. This study also underlines the importance of adding the right term, because it can make one or few diagnoses more significant. The wrong term can instead distract from the right diagnosis, including diseases that are completely offset.

Most of current phenotype similarity measurement are based on the following metrics: balance between frequency and specificity, information content (IC) of the term, disease-driven density of a subset of terms, and gene or disease annotations.^{120,123} In the evaluation of gene/disease annotation, you have to consider four factors: size of annotation set, evidence code of annotations, quality of annotations and coverage. In particular, these last two are the most influencing the performance, as the reduction of coverage and quality causes a decrease in semantic similarity as well.¹²²

Also, many studies have presented methods to enrich the HPO vocabulary with new synonyms. Some methods are based on the lexical properties of the terms, other on the hierarchical structure of the ontology. Regarding the lexical properties, the purpose is to learn new names to generate synonyms for the descendant terms. However, since ontologies are not created to be the lexical basis for name recognition systems, the performance in this case is lower than required.¹²⁴ It is, indeed, important to include the hierarchical relationships, that are able to predict associations between genes and abnormal phenotypes with competitive results.¹²⁵ The initial focus of the HPO was on rare, mendelian disease, and now there are many annotations for CNV and common diseases. It is always combined with NGS techniques to support the diagnosis and has been shown to improve the ability of NGS-based methods to identify the candidate genes.^{126,127}

Recently, it has also been proved that using HPO could be a promising approach for increasing the prediction of disease-associated lncRNAs.¹¹⁸

There are, however, some weaknesses: particularly, some disease-categories are not accurate enough, and it is challenging to describe a patient with the right phenotype, i.e. in patients with epilepsy. It is hoped that the phenotype data will achieve dimensionality as a result of the amplification of HPO modifiers.

It is hoped that the HPO will provide, through a wider vocabulary, a more and more unified basis for clinical research, implementing the accuracy of medical genetics. The current version (version 1.2 releases/2017– 2-14) contains approximately 12,000 terms, and it is available online at <https://hpo.jax.org>.¹¹²

1.4 “Deep learning”: the use of machine learning in the field of genetics

Deep phenotyping is one of the essential steps in what is called “deep medicine”. The following step is the deep learning, that consists in the use of machine learning as a tool to accelerate the diagnosis of mendelian disorders.

There are over 4 millions variants in a typical genome¹²⁸, some of which are very rare in reference databases of control individuals.^{129,130} Therefore, the clinical interpretation of exomes can be very time consuming, considering that identifying the causal mutation requires from 40 up to 100 hours of expert analysis time.⁴⁰ A number of databases that curate gene-disease associations have been created, and they are routinely used in WES, along with several variant annotation tools.^{102,103,131}

The application of computational analysis can help and suggest expanded phenotypes and improve the field of personalized medicine, even if in some case the clinician evaluation of variants outperformed these computational approaches.^{27,132,133}

Most tools use a machine learning classifier that compares the information about a patient’s phenotype and genotype to its knowledgebase, in order to prioritize the candidate causal genes and rank the genes for their likelihood of being causative.^{134–136}

Phenotypic information is not only limited to symptoms included in the Human Phenotype Ontology (HPO), but may also comprise clinical diagnoses and the suspected mode of inheritance. The clinical symptoms can be entered via HPO, OMIM and Orphanet, and the gene informations can be integrated with data from HMGD¹³⁷, Gene Ontology¹³⁸, HGNC and UniProt^{139,140}, in a Variant Call Format²⁵ file. These tools are growing interest in the last years as they can improve the diagnostic yield and lower the time of the WES.^{134,141–148}

There is, however, much work to do for them to become reliable.

1.5 Reverse phenotyping

Reverse phenotyping is a new approach where phenotypes are refined based on genetic marker data. A good clinician knows that accurate clinical diagnosis depends on 3 key elements:

- systematic collection of clinical data, through history and examination;
- careful ruling out of a set of ‘clinical leads’, through directed examination and laboratory evaluations (differential diagnosis);
- returning to the patient to confirm the diagnosis.

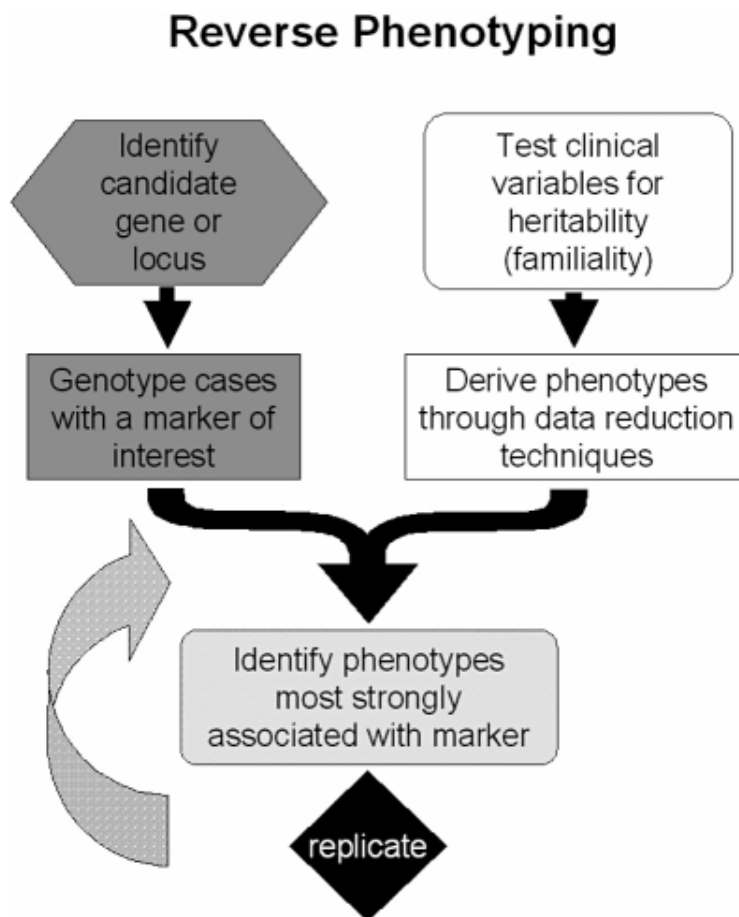


Figure 17 Workflow for reverse phenotyping.¹⁴⁹

Genetic studies of complex diseases generally progress from phenotype to genotype to analysis. This is the so-called 'forward genetics' method. Such a unidirectional approach often fails. The growing availability of large data sets and massive genotyping now makes it possible to consider another approach, which might be called 'reverse phenotyping.' In this approach, the genetic marker data is used to drive, or form the basis of, new phenotype definitions. The goal is to define phenotypic groupings that are distinguished by higher rates of allele-sharing (linkage data) or more deviant allele frequencies (association data) than are seen in the traditional diagnostic categories (figure 17).

Reverse phenotyping can be thought of as a 2-step process:

1. formulation of a hypothesis as to which clinical features predict which genotype(s)
2. testing of this hypothesis in a second sample.

There remain many challenges in the identification of genes underlying complex human diseases. As the molecular and statistical tools reach an unprecedented level of sophistication, the next frontier may well lie at this kind of approach.¹⁴⁹

1.6 Hereditary neurological disorders

In my thesis I will perform WES on the DNA of children with undiagnosed neurological illnesses, with suspect genetic etiology.

Hereditary neurological disorders (HNDs) are a clinically heterogeneous group of diseases, quite common in childhood. The aetiology is variable, and only some of the genes have been defined. It's important for the right characterization of the pathology not only to know the causal genes, but also to know the several phenotypes that can define the same disease. Indeed, recognizing the clinical features is essential for genetic testing results interpretation and genetic consultation.

According to Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), the pediatric age includes five age stages of neurodevelopment: neonatal, infancy, toddler, childhood, adolescence.¹⁵⁰

HNDs are a group of genetic diseases affecting the neurological system, that can occur at any stage of this classification, but most commonly diagnosed during certain ages. For example, congenital malformations could be identified during preterm and neonatal periods, while developmental delay is often discovered in infants.^{151,152}

In the clinical classification it is outstanding the value of the right approach, that consists in the early identification of clinical signs that refer to abnormalities specific to a particular disease and in the evaluation of their age onset, in the study of the mode of inheritance and in the interpretation of extra-neural symptoms, whether they're linked or not to the same genetic disorder.¹⁵³

1.6.1 CLASSIFICATION

Hereditary neurological disorders can be classified in:¹⁵²

- a. Movement disorders, generally characterized by impaired movements control, ataxia and/or spasticity¹⁵⁴
- b. Developmental disorders and neuropsychiatric disorders: include several clinical features, most of which are developmental delay, intellectual disability and autistic behavior. Developmental delay refers to a delay in the achievement of motor or mental milestones in the domains of development of a child, including motor skills, speech and language, cognitive skills, social and emotional skills, and describe children with less than 5 years of age. The intellectual disability is a subnormal intellectual functioning that originates during the developmental period, previously referred to as mental retardation. Autistic behavior is characterized by persistent deficits in social interaction and communication as well as markedly restricted repertoire of activity and interests, as well as repetitive patterns of behavior.¹¹³
- c. Neuron peripheral disorders: a group of conditions in which the peripheral nervous system is damaged. In children most neuropathies are symmetrical, distal and with mixed features.¹⁵⁵
- d. Epilepsy: a central nervous system disorder characterized by recurrent epileptic seizures unprovoked by any immediately identifiable cause. An epileptic seizure is due to an abnormal and excessive discharge of a set of neurons in the brain.^{156,157} This may cause seizures or periods of unusual behavior, sensations, and sometimes loss of awareness. Seizures can affect any brain process, causing confusion, loss of consciousness, staring spell, jerking movements of the arms and legs, fear, anxiety or déjà vu.¹⁵⁸ Seizures can be classified as either focal or generalized, based on how the abnormal brain activity begins. The classification of seizures is very complex and detailed and it is continuously changing throughout the years as new aspects are discovered.^{159–161}

- e. Neuromuscular disorders: Myotonic dystrophy (MD), Duchenne muscular dystrophy (DMD) and Becker muscular dystrophy (BMS) , Spinal muscular atrophies(SMAs) and mitochondrial diseases.

To date, it is believed that the molecular basis of only about half of Mendelian diseases have been discovered and that the other half awaits elucidation. ^{162–164}

Many of the Mendelian diseases still waiting to be discovered are very rare or difficult to diagnose clinically, but it is also difficult to make the right diagnosis purely based on sequencing technologies, due to the high clinical heterogeneity(Figure 18).

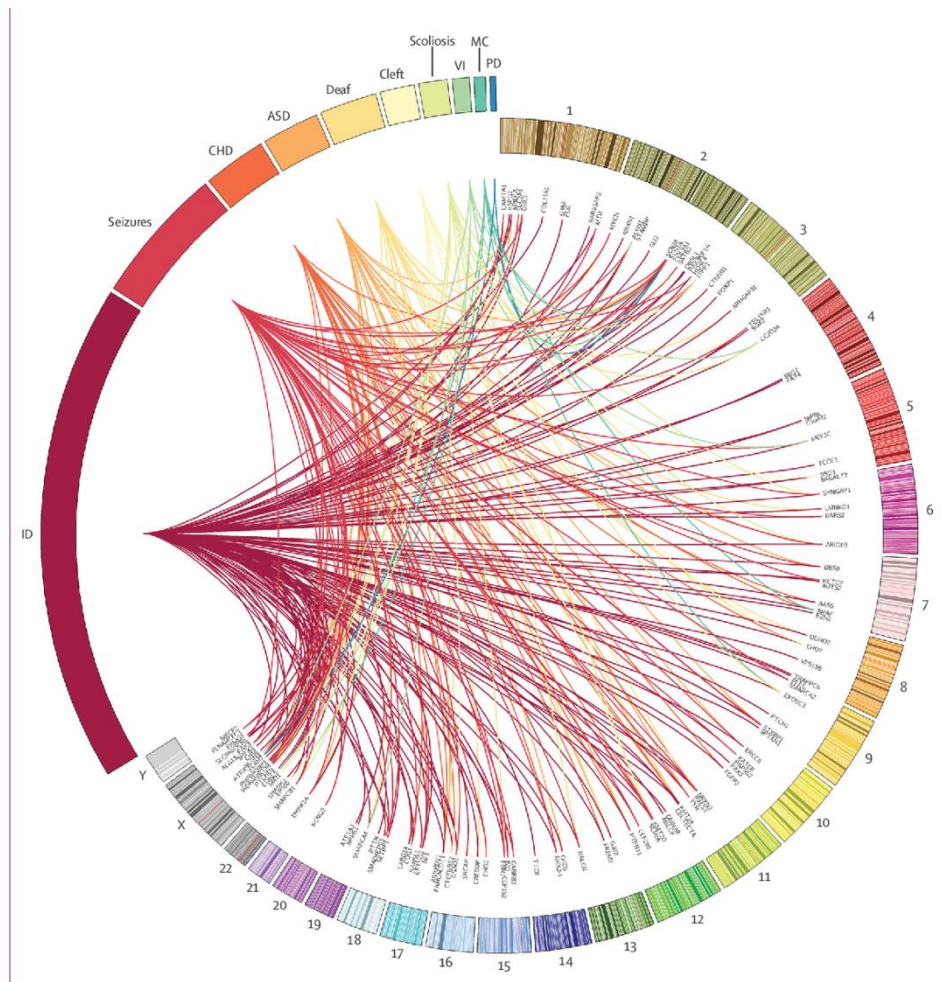


Figure 18 Clinical and genetic heterogeneity¹⁶⁵

2. WHOLE EXOME SEQUENCING IN A TERTIARY PEDIATRIC NEUROLOGY CENTRE: A PILOT STUDY

2.1 Study design

In this study we tested whole exome sequencing in a pediatric tertiary center (Pediatric Neurology and Muscular Diseases Unit, “G. Gaslini” Institute). Over a period of 18 months, a total of 45 consecutive children with rare and genetically undetermined medical conditions with variable neurological impairment underwent a standardized assessment program in the context of a complex neuropediatric diagnostic work-up (Figure 19).

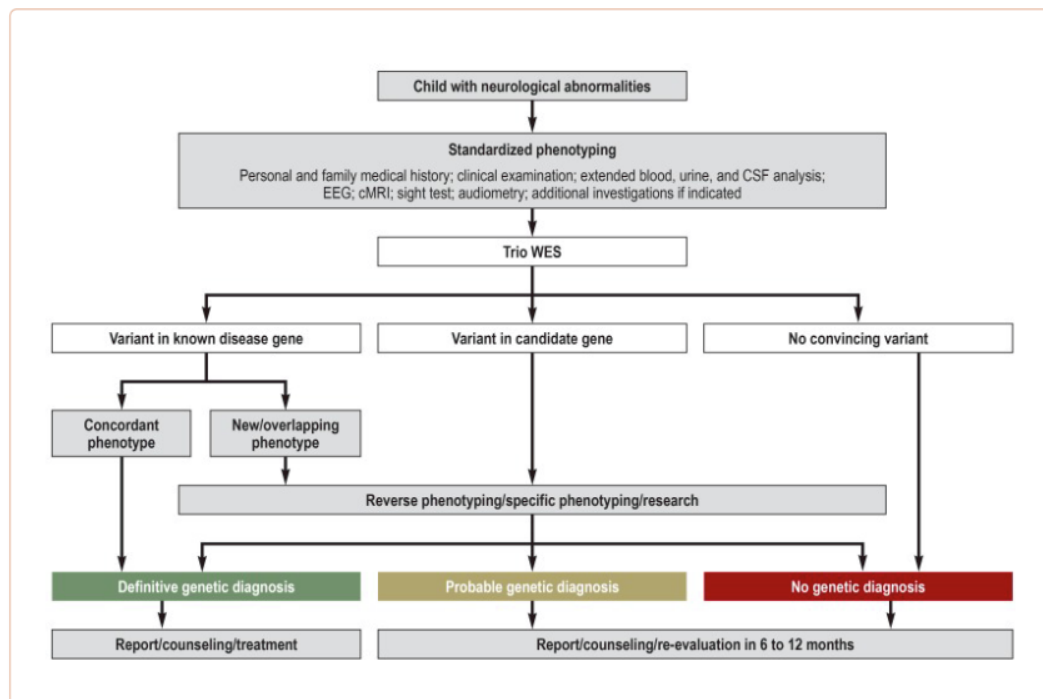


Figure 19 Neuropediatric diagnostic work-up¹⁶⁶

2.1.2 INCLUSION CRITERIA

- Neurological symptoms (e.g., global developmental delay, ataxia, seizures)
- Suspected genetic etiology
- No specific provisional diagnosis
- Signed consent from the parents, following appropriate explanation, for exhaustive investigations including trio WES

2.1.3 EXCLUSION CRITERIA

- Signs of serious perinatal complications, infection, injury, other exogenic factors

2.2 Materials and methods

2.2.1 PHENOTYPE DOCUMENTATION

The clinical assessment included detailed questioning about personal medical history, familiar history, clinical evaluation by a neuropsychiatrist and a clinical geneticist, investigation of blood and urine, diagnostic imaging and EEG. Depending on the clinical finding, other tests were added. The relevant clinical findings were reported using Human Phenotype Ontology.

We created, for each patient, a PowerPoint presentation including:

- the name, the familiar history, photos, molecular investigations and the **core phenotype**, in the first page;
- personal history and important radiological findings
- The HPO codes in the last page (Figure 20)

CORE PHENOTYPE	HPO CODES
Global developmental delay	HP:0001263
Generalized hypotonia	HP:0001290
Abnormality of brain morphology	HP:0012443

Figure 20 Example of the deep phenotyping: core phenotype and the HPO codes

2.2.2 GENOTYPE DOCUMENTATION

Written informed consent was obtained for all individuals and their relatives, after which DNA was extracted. The samples were obtained by venipuncture, taking 2-4 ml of blood in EDTA tubes, or by scraping the inside of the cheeks with a cytobrush for buccal epithelial cells. The samples were stored at room temperature and dispatched within 24-48h. We performed whole exome sequencing (WES) by using the Agilent SureSelect QXT Clinical Research Exome (Agilent Technologies, Santa Clara, CA, USA) that provides a 54 Mb target, including an enhanced coverage of disease-relevant targets from HGMD, OMIM and ClinVar. DNA was enriched with Nextera Rapid Capture Exomes Kit. Enriched DNA was validated and quantified by microfluidic analysis and libraries were sequenced using the NextSeq500 system (Illumina Inc., San Diego, CA, USA), covering at least 2x150 nt.. The average exome coverage of the target bases at 20X was > 85%.

2.2.3 DATA ANALYSIS: ALIGNMENT

Data were processed and filtered with established pipelines at the academic laboratories involved in the study. The sequence reads were aligned against the reference genome “human genome assembly hg19” (UCSC Genome Browser) with the aid of the Burrows-Wheeler Aligner (BWA), which is a software package for mapping low-divergent sequences against a large reference genome.

The alignment process consists of two steps:

- a) Indexing the reference genome: it allows the aligner to quickly find potential alignment sites for query sequences in a genome, which saves time during alignment.
- b) Aligning the reads to the reference genome

2.2.4 DATA ANALYSIS: VARIANT CALLING AND ANNOTATION

After the mapping, variant calling process was performed: single-nucleotide variants and short insertion or deletion variants were identified by Haplotype Caller of GATK (v3.3.0).¹⁶⁷

Table 2 Evidence of pathogenicity of filtered variants

18

Previous reports of the mutation in curated mutation or literature databases (eg. Human Genome Mutation Database, Online Mendelian Inheritance in Man)
Allele frequency data (eg, deposition in dbSNP, 1000 Genomes Project or Exome Variant Server): the more common an allele the less likely it is to be causal in a rare disease.
Literature support (eg, animal models).
Absence in ethnically matched controls.
Co-segregation with the disease in a family.
Identification of a de novo variant in a sporadic condition.
Evolutionary conservation (nucleotide and amino acid residue).
Large physicochemical distance in a missense amino acid
In silico prediction of effect on splicing.
In silico prediction of deleteriousness.

Variants were annotated with ANNOVAR to assign frequencies in large scale variants datasets (1000Genomes, ExAC, gnomAD) and potential impact on protein function.

Closer attention was paid to genetic variants which were very rarely identified in the population (minor alleles frequency [MAF]; occurrence of the rarer allele in the population <0.01%) and which, according to several different bioinformatic prediction algorithms, impact negatively on gene function (functionally relevant variants).

The disease relevance of a given identified variant was evaluated with the aid of a number of variables (Table 2): MAF, assessment by bioinformatic prediction programs, comparison with databases (e.g., Database of Exome Aggregation Consortium, ExAC [e1]; Online Mendelian Inheritance of Man, OMIM [e2]), and current knowledge of the coded protein and its function.

A variant in a gene was classified as pathogenic, or causing disease, if:

- It affected a known disease gene (the association with disease had been demonstrated by published clinical–genetic and/or functional data)
- The very same variant had previously been described as causing disease, or the variant was comparable in type with known disease-causing variants in the same gene.
- There were few or none differences in phenotype between our patient and the published patients with causative mutations in the same gene

A variant in a gene was classified as probably causing disease if:

- It affected a candidate gene
- More than one prediction program (Polyphen 2, SIFT and CADD) classified it as pathogenic.
- In the case of a de novo mutation, it was not listed in the ExAC database

A gene was classified as a candidate gene if:

- Based on the WES results, it was the only one of the patient's genes in which a rare variant was found
- Based on the WES results and bioinformatic prediction algorithms it was classified as probably causing disease
- Previously published data pointed to disease association in humans or there was contact to other study groups who had also detected variants in the same gene in patients with overlapping symptoms. ¹⁶⁶

The variants of interest are classified according to the ACMG guidelines³⁰ (Figure 21), that distinguish:

- Pathogenic variants (V class)
- Likely pathogenic variants (IV class)
- Variant of unknown significance (VUS, III class)
- Likely benign (II class)
- Benign variants (I class).

	Benign			Pathogenic		
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder BA1/BS1 OR observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in <i>trans</i> with a dominant variant BP2 Observed in <i>cis</i> with a pathogenic variant BP2		For recessive disorders, detected in <i>trans</i> with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

Figure 21 ACMG guidelines for variant classification.³⁰

In all cases, variants that did not adhere to the following criteria were excluded from further analysis: (1) allele balance of >0.70 , (2) QUAL of <20 , (3) QD of <5 and (4) coverage of $<20\times$.

The filtered variants were confirmed by the conventional Sanger sequencing according to the standard methods.

2.2.5 REVERSE PHENOTYPING

In many cases the team ordered additional specific clinical investigations (reverse phenotyping) to enable more detailed assessment of a genetic variant. The results of reverse phenotyping were then interpreted by the assembled team.

2.3 Results

2.3.1 STUDY PARTICIPANTS

Forty-five probands were analyzed through WES, with a male to female ratio of 1.14. The median age of the forty-five unrelated children at study inclusion was 10,32 years (mean age 10,13). Twenty-nine of them (66%) had already undergone genetic testing (chromosome analysis, array-based comparative genomic hybridization, and/or single gene sequencing). In eleven of the twenty-nine cases (38%) the results were abnormal but of uncertain significance.

2.3.2 RESULTS OF PHENOTYPING

The first symptoms were observed at a median age of 12 months (mean age of 3,2 years).

A median of 5,82 years (3 months to 21 years) elapsed between occurrence of the first symptoms and inclusion in the study (Figure 22)

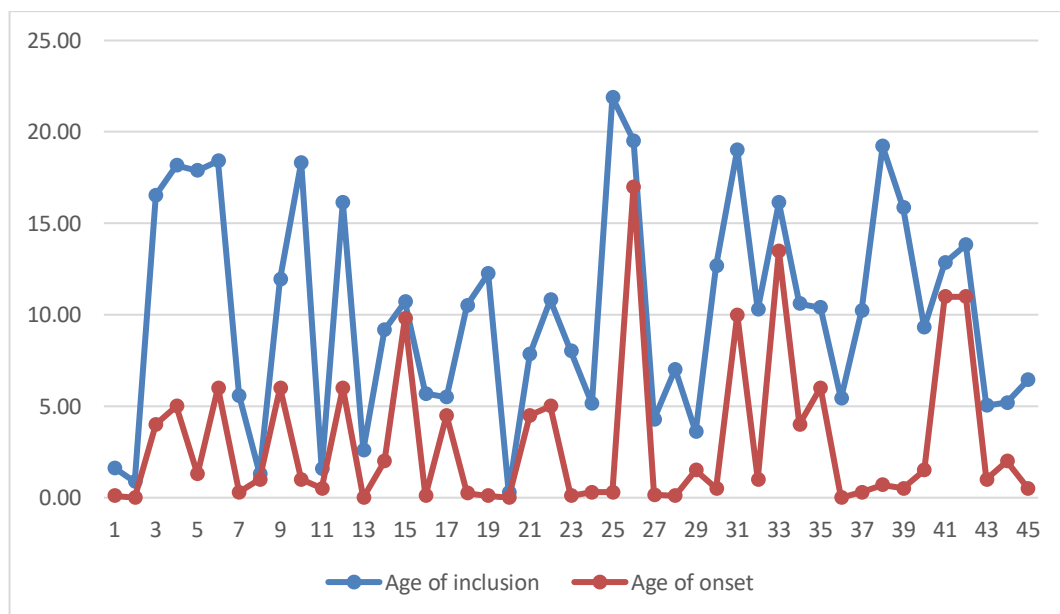


Figure 22 Age of onset of the symptoms and age of inclusion in the study.

In line with the inclusion criteria, all forty-five patients had neurological symptoms at the time of entry into the study. These comprised global developmental delay, intellectual disability, language impairment and/or delay in 55% of the cases (Figure 23).

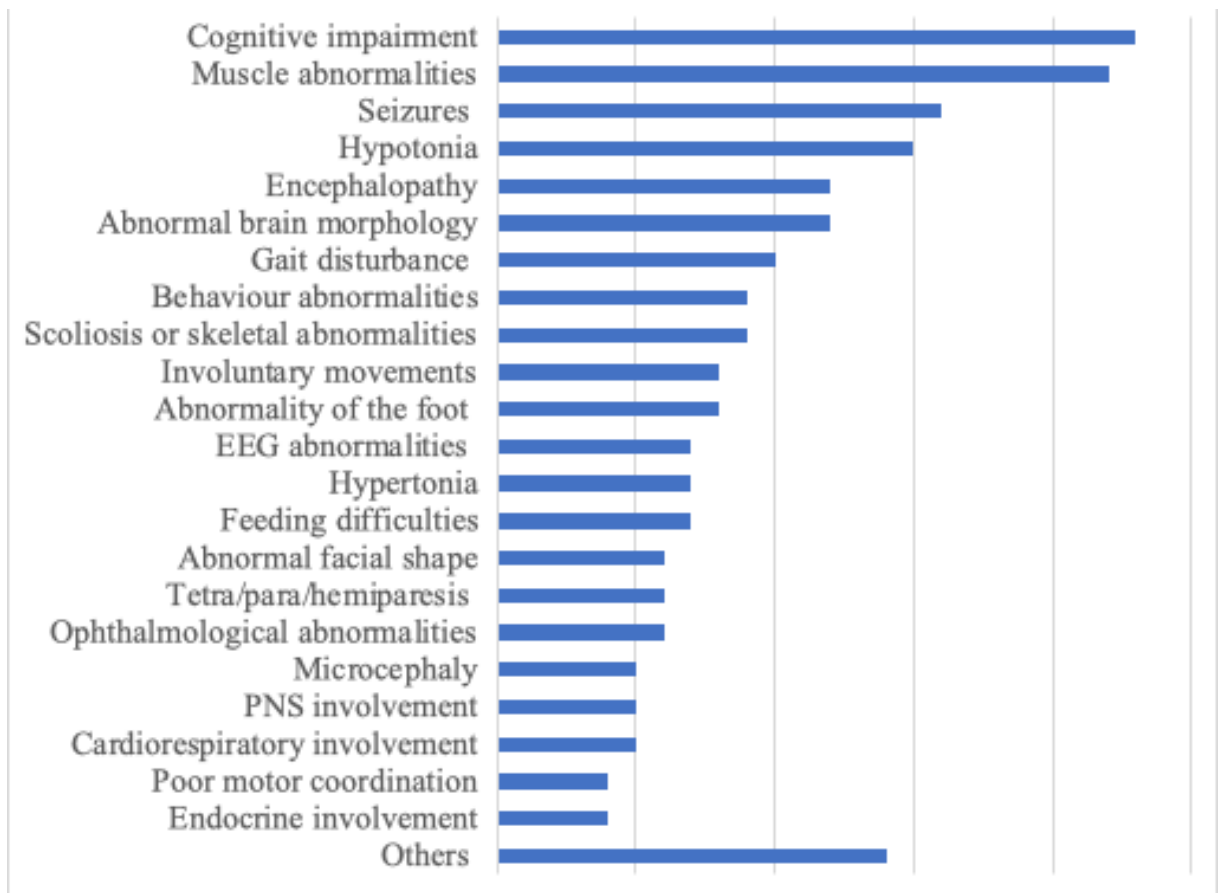


Figure 23 Phenotypic characterization: of the study group according to Human Phenotype Ontology (HPO)

2.3.4 RESULTS OF GENOTYPING

Out of 45 affected individuals, 35 have been analyzed, while the others are still under investigation. Pathogenic or likely pathogenic variants explaining the phenotypes have been identified in 12 patients, with a diagnostic rate of 34,2%. The classification of the ACMG guidelines revealed: 9 class V variants and 3 class IV variants. In 5 affected individuals, WES led to the identification of VUS or likely benign, in others 5 individuals variants in a new gene not known in literature have been identified (Table 3 and Figure 24). In the remaining 13 affected individuals, there were variants in weaker candidate genes or no variant at all. In these cases variants will be periodically reanalyzed or, in some cases, Whole Genome Sequencing (WGS) will be performed.

Table 3 Findings of Whole-Exome Sequencing

	Number of patients
Findings	
Mutation in a known disease gene	12(34%)
Variant in a candidate gene	5 (14%)
Variant of uncertain significance	5 (15%)
No convincing variant	13 (37%)

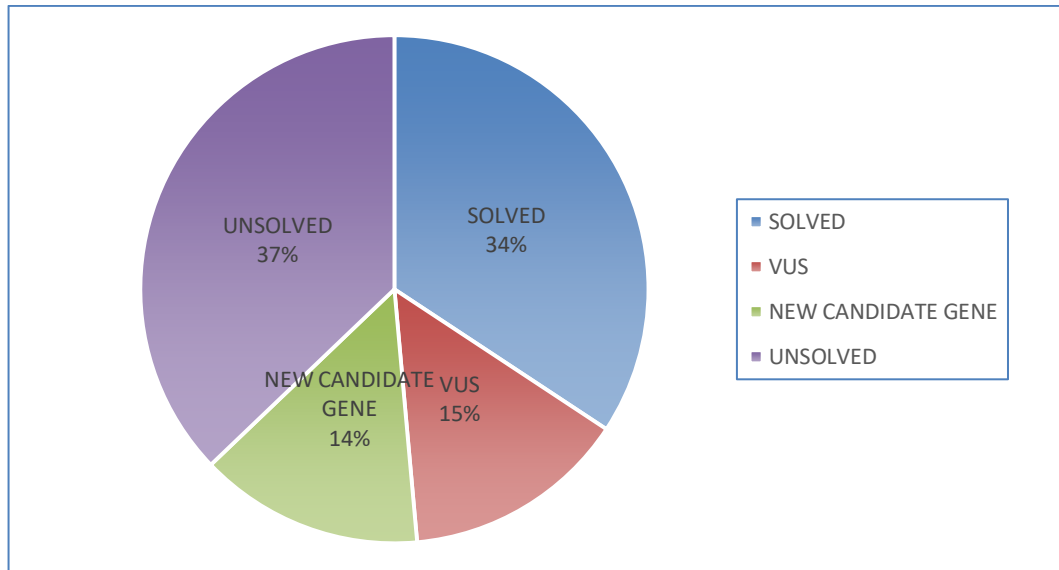


Figure 24 Pie chart graph showing the results of the study.

De novo variants were found in 9 individuals, while homozygous, X-chromosomal and biallelic variants were respectively identified in 1 proband. With regard to the molecular consequences of the variants, we identified: 6 missense, 4 nonsense and 1 splicing mutation. In the biallelic mutation, we identified 1 missense and 1 nonsense variant (Table 4).

The missense variants were found to be disease causing in both dominant (e.g. KCNQ2-related epileptic encephalopathy #613720) and recessive disorders (e.g. ZC4H2-related X-linked neurodevelopmental disorder affecting the central and peripheral nervous systems).

The nonsense variants were found to be disease-causing in dominant disorders (e.g. MYH6-related hypertrophic cardiomyopathy with encephalomyopathy-#613251).

The splicing mutation was a splice acceptor variant (c.317-2A>G) identified in a female patient with TMEM70-related encephalopathy, where the variant is known as a founder mutation (-#614052)

The biallelic mutation was identified in a male patient with PEX13-related neurodevelopmental disorder and encephalopathy. It consisted of two different variants: a nonsense variant (p.L91Ter chr2:61031897-CTT-C) and a missense variant (p.R294W chr2:61045818-C-T) (Table 5).

Table 4 Characterization of the mutation in a known disease gene.

	Number of patients
Characterization of the mutation in a known disease gene.	
Inheritance	
De novo	9 (75%)
Autosomal recessive	1 (8,3%)
X-chromosomal	1 (8,3%)
Biallelic	1 (8,3%)
Mutation type	
Missense	6 (50%)
Nonsense	4 (33%)
Splice site/ acceptor	1 (8%)
Biallelic missense/nonsense variants	1(8%)

Table 5 Genotype-phenotype correlations found in diagnosed patients with pathogenic variants in a known disease gene.

Patient	Consanguinity of parents	Family members with similar symptoms	Core phenotype	disease gene	variant	Mutation nucleotide	Mutation protein	Inheritance	Associated disease	OMIM disease number
<i>Mutations in a known disease gene</i>										
#1	No	No	Myopathy Shoulder girdle muscle weakness Proximal amyotrophy	ANG	Missense	c.407C>T	p.Pro136Leu	De novo	ALS9	#611895
#2	No	No	Global developmental delay, seizures, hypotonia, arthrogriphosis, camptodactily, congenital vertical talus	ZC4H2	Missense	c.593G>A	p.Arg198Gln	X-chromosomal recessive	WRWF	#314580
#3	No	No	Muscle weakness and atrophy, gait disturbance, myopathy, skeletal dysplasia and ligamentous laxity.	FIG4	Missense	c.122T>C	p.Ile41Thr	De novo	CMT4J	#611228

#4	No	No	Encephalopathy, global developmental delay, hypotonia, seizures.	KCNQ2	Missense	c.1678C>T	p.Arg560Trp	De novo	EIEE7	#613720
#5	No	No	Encephalopathy, developmental regression, delayed speech.	PEX13	Nonsense Missense	5q35.2-35.3 c.880C>T	p.L91Ter p.Arg294Trp	Biallelic mutation	PBD11A	#614883
#6	No	No	Global developmental and growth delay, myoclonic jerks, behavioural abnormalities, gait disturbance, EEG discharges	ADCY5	Missense	c.1253G>A	p.Arg418Gln	De novo	FDFM	#606703
#7	Born in the same small village.	No	Encephalopathy, hypotonia, skeletal muscle atrophy.	TMEM70	Splice site	c.317-2A>G	Splice acceptor	Autosomal recessive	MC5DN2	#614052
#8	No	No	Global developmental delay, microcephaly, seizures, language impairment, tetraparesis.	NCOR1	Nonsense	c.6827_6828del	p.G2276Vfs*7	De novo	ASD	#209850
#9	No	No	Seizures, spastic paraparesis.	CHD2	Nonsense			De novo	EEOC	#615369

#10	No	No	Concentric hypertrophic cardiomyopathy, tetraparesis, mitochondrial encephalopathy, hypertriglyceridemia.	MYH6	Nonsense	c.23383233 C>A	p.E1885Ter	De novo	CMH14	#613251
#11	No, but grandparents born in the same village of 5000 inhabitants.	No	Encephalopathy, motor developmental delay, gait disturbance with lower limb spasticity.	GCH1	Missense	c.551G>A	p.Arg184His	De novo	DRD	#128230
#12	No	No	Autistic behaviour, impaired social interactions, developmental regression.	PLAA	Missense	c.2383C>A	p.Leu795Met	De novo	NDMSBA (our mutation is less serious)	#617527

ALS: amyotrophic lateral sclerosis 9, **WRWF:** Wieacker-Wolf syndrome. **CMT:** Charcot Marie Tooth. **EIEE7:** Epileptic encephalopathy early infantile 7. **PBD11A:** peroxisome biogenesis disorder 11A (Zellweger). **FDFM:** Familial dyskinesia with facial myokymia. **MC5DN2:** Mitochondrial complex V deficit, nuclear type 2. **ASD:** autism spectrum disorder. **EEOC:** epileptic encephalopathy childhood onset. **CMH14:** Cardiomyopathy familial hypertrophic ,14. **DRD:** dopa-responsive dystonia. **NDMSBA:** neurodevelopmental disorder with progressive microcephaly, seizures and behavioral abnormalities.

2.4 Discussion

The use of WES in our center allowed the identification of several pathogenic variants, improving the diagnostic yield comparing to other tests.

Indeed, 66% of the enrolled patients had already been studied with the most common genetic techniques (CGH-array, NGS panel testing or single gene testing) and the results were negative or of uncertain significance.

Thanks to the deep phenotyping and accurate selection, a high diagnostic yield was achieved in our study. The diagnosed diseases included rare type of epileptic encephalopathies, neurodegenerative conditions, neurodevelopmental disorders, PNS disorders, movement disorders, myocardial diseases and metabolic conditions.

Several pathogenic variants in different disease-causing genes have been identified. Among these genes, out of the variants identified in disease causing genes and in candidate genes, some were of particular interest because of the peculiar function of the encoded protein.

2.4.1 CASES OF PARTICULAR INTEREST: PATHOGENIC VARIANTS IN KNOWN DISEASE GENE

Patient #1 is a male patient, 12 years old, that was suspected with a limb-girdle dystrophy. The phenotype includes shoulder girdle muscle weakness and proximal amyotrophy. Targeted sequencing for limb-girdle dystrophy showed negative results. WES led to identification of a heterozygous variant c.407C>T p.Pro136Leu in the ANG gene, that encodes angiogenin. This variant was described as pathogenic in 2007 confirmed by functional studies, and reported to be causing a rare form of Amyotrophic Lateral Sclerosis(ALS).¹⁶⁸

ALS is a progressive neurodegenerative disease that typically presents in the fifth to sixth decades of life with upper and lower motor neuron signs. Initially, there are symptoms that include distal muscle weakness and wasting, increased muscle tone with hyperreflexia, and at times diaphragmatic and/or bulbar weakness. This case is of particular interest for the precociousness of the onset. Moreover, it unravels ethical issues for the future management of the patient, being SLA untreatable right now.

Patient #5 is a male patient, 8 years old, affected with leukoencephalopathy, developmental regression (since 5 years old) and hearing impairment. WES led to the identification of two different variants:

1. The first variant is a nonsense variant (ch2: 61031897-CTT-C p.L91Ter)
2. The second is a missense variant (chr2:61045818-C-T p.Arg294Trp). The variant is located in the SH3 domain (amino acid 276-334), which is a highly conserved portion of the protein. For this reason this mutation can interfere with other functions of the protein and with other related molecules.

Both variants, according to a biallelic mechanism, affect PEX13, a gene that encodes peroxisome biogenesis factor-13, a peroxisomal membrane protein that acts as an essential docking factor for the import of peroxisomal matrix proteins.¹⁶⁹ PEX13 is related with Zellweger syndrome, an autosomal recessive multiple congenital anomaly syndrome resulting from disordered peroxisome biogenesis, generally characterized by hypotonia, seizures, feeding difficulties. Our patient seems to be affected with an incomplete spectrum of the syndrome.

The peculiarity of this case is the presence of a biallelic mutation. Indeed, it was the combination of the two variants that led to the dysfunction of the gene and to the related disease. The variants will be studied with functional studies and accurately described in future works.

Patient #6 is a female patient, 5 years old, affected with a neurodevelopmental disorder. The core phenotype is characterized by global developmental delay, growth delay, involuntary movements as myoclonic jerks, behavioural abnormalities with hyperactivity and seizures. In this family we identified a *de novo* missense variant (Chr3:123352463-C-TNM_183357.2 c.1253G>A p.Arg418Gln) in ADCY5 gene, related with neurodevelopmental and hyperkinetic movement disorders. The very same variant (p.Arg418Gln) was already described as pathogenic in a previous work (Chen DH et al., 2015) and has been reported once in ClinVar(VCV000218354.1). The rarity of the variant makes this case interesting, and future works will be published on its pathogenicity.

Patient #11 is a male patient, 5 years old, affected with encephalopathy, delayed gross motor development, lower limb spasticity and gait disturbance. In WES data analysis a heterozygous variant was found in GCH1 gene, which disruption is related to a rare form of dopa-responsive dystonia.

The autosomal recessive dystonia is also known as Segawa syndrome: there are 2 main phenotypes: one is a severe complex encephalopathy apparent in the perinatal period, and the other shows a less severe course with onset in the first year of life of a progressive hypokinetic-rigid syndrome and generalized dystonia. The less severe type shows a better response to levodopa compared to the more severe type.¹⁷⁰

Autosomal dominant dopa-responsive dystonia (DRD) is characterized by generalized dystonia, diurnal fluctuation of symptoms, and a dramatic therapeutic response to L-dopa. In autosomal dominant form, dystonia involved at first only the lower limbs, and with the progression of the disease all limbs were involved.¹⁷¹

The variant found in our patient DNA (c.551G>A p.Arg184His) is reported as pathogenic on ClinVar, and one previous work has been published with the same variant.

Our patient has less serious impairment compared with others described in literature (probably related to the heterozygous mutation, less damaging than the homozygous).

This is a case where reverse phenotyping is mandatory. Mutations in GCH1 lead to a form of dystonia in which the particularity is the fluctuation of symptoms during the day, that has to be evaluated in our patient.

Another peculiarity of this disease is to be nearly completely ameliorated by treatment with levodopa. This finding has important implications in the clinical course of the patient, considering that L-DOPA administration could completely improve the dystonic posturing and the gait disturbance. Early treatment can prevent morbidity and contracture formation, and may also reduce the motor and intellectual developmental delay.

Patient #12 is a female patient, 5 years old, affected with intellectual disability (developmental regression), autistic behavior with impaired social interactions and language impairment. We identified a missense variant (c.2383C>A p.Leu795Met) in PLAA gene. PLAA mutation cause a autosomal recessive neurodevelopmental disorder with progressive microcephaly, spasticity and brain abnormality.¹⁷²

PLAA plays a particular role in the turnover of synaptic membrane proteins. Physiologically, the neurotransmitter release of synaptic vesicles is mediated by SNARE protein (e.g. SNAP-25, syntaxin 1). This proteins create a complex to mediate the fusion of the vesicle with the plasmatic membrane. When the process is completed, PLAA interacts with p97, creating a complex that is essential for the disassembling of the SNARE complex(Figure 25).¹⁷³

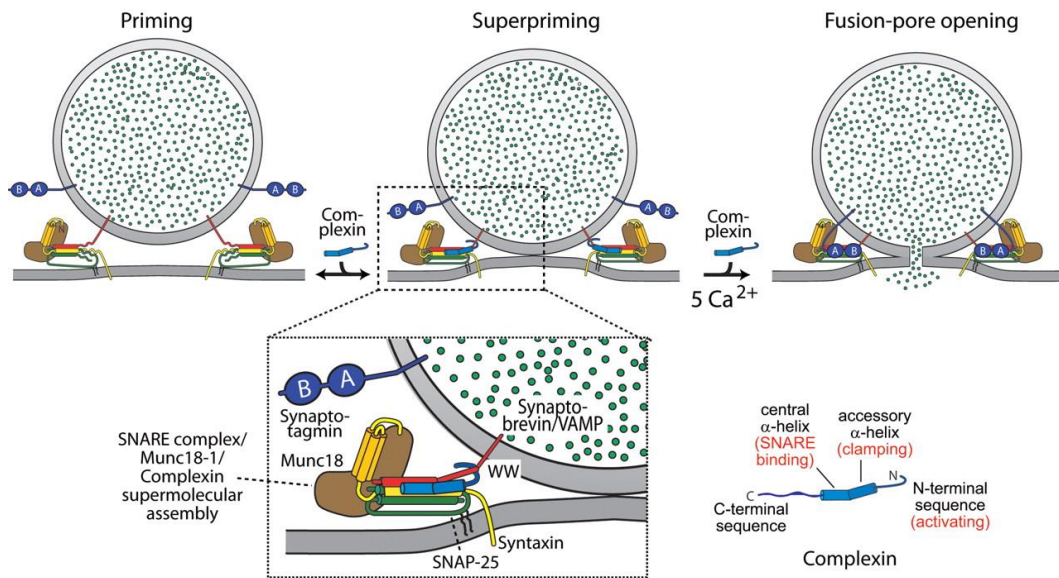


Figure 25 Release of the synaptic vesicles: SNARE complex assembly.
 Courtesy of Maximoy and Tang.¹⁷⁴

This particular variant is located in the 795 residue of the protein, that is the last amino acid of PUL domain, which is the domain that interacts with p97 (Figure 26)

PLAA-related neurodevelopmental/neurodegenerative disorder

ARTICLE

PLAA Mutations Cause a Lethal Infantile Epileptic Encephalopathy by Disrupting Ubiquitin-Mediated Endolysosomal Degradation of Synaptic Proteins

(AR Neurodevelopmental disorder with progressive microcephaly, spasticity, and brain anomalies MIM- 617527)

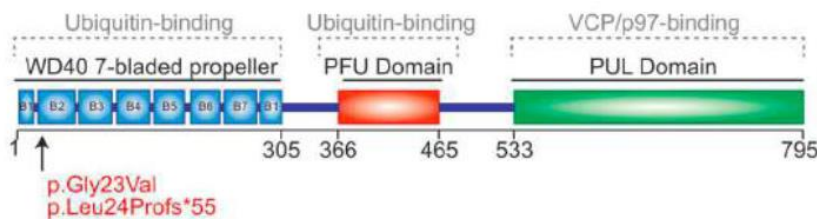


Figure 26 Article showing the interaction of p-97 with the PUL domain.

Our patient has a de novo heterozygous variant, that can explain the less serious phenotype compared to the literature.

2.4.2 CASES OF PARTICULAR INTEREST: VARIANTS IN A CANDIDATE GENE

Patient #13 is a female patient, 19 years old, with an onset of seizures at 12 months, than diagnosed as epileptic encephalopathy with abnormality of brain morphology (cerebral atrophy, white matter abnormalities). She is affected with intellectual disability, cardiomyopathy, gait disturbance with spasticity.

During WES investigation, a variant was found in DENND5 gene.

Human DENN domain proteins

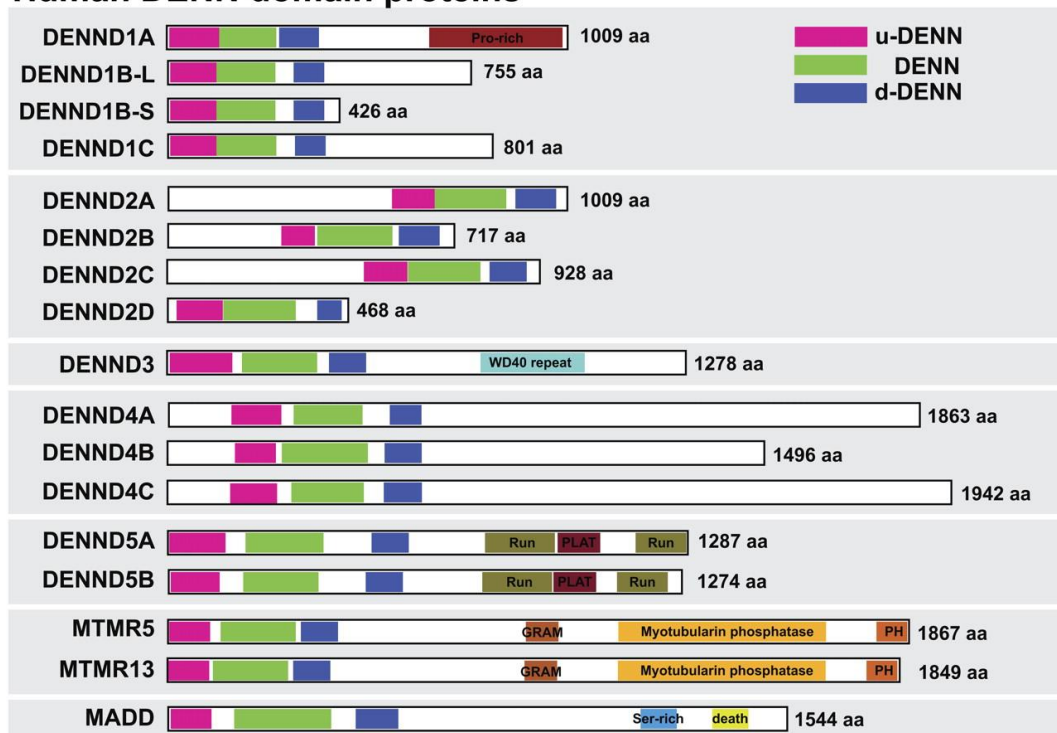


Figure 27 Human DENN domain proteins. Courtesy of Yoshimura et al. ¹⁷⁵

DENN domain-containing protein 5 controls membrane trafficking events and activates RAB GTPases. A study found that both DENND5 showed GEF activity toward RAB39 (Figure 27).¹⁷⁵

RAB proteins, such as RAB39B, are small GTPases involved in the regulation of vesicular trafficking between membrane compartments.

Rab39 exists in two isoforms:

- 39a: controls endocytic pathway and autophagy
- 39b: controls vesicular recycling.

Mutations in RAB39b are responsible for intellectual disability associated with autism, epilepsy and macrocephaly.¹⁷⁶

Our variant (c.3083C>T p.Arg1028Glx) is located in DENND5 gene, that has not been reported in literature as disease-causing gene. This can be a candidate novel disease gene, considering that DENND5 disruption could lead to impaired activation or interaction with RAB39 GTPases leading to intellectual disability and epilepsy (as described in literature for RAB39 mutations and DENND5A disruption).¹⁷⁷

Functional studies will be performed (GDP releasing assays, autophagy essays, knockdown in cultured neurons) to confirm its pathogenicity.

Patient #14 is a 5 years old male patient, affected with global developmental delay and seizures, hypotonia, duplicated renal collecting system and right hemihypertrophy. A nonsense homozygous variant was found in SCL22A18AS (p.Glu177Ter). Human chromosomal band 11p15.5 contains genes involved in development of several pediatric and adult tumors and in Beckwith-Wiedemann syndrome (BWS). BWS is a pediatric syndrome characterized by overgrowth (hemihypertrophy), hypoglycemia, kidney malformations, predisposition to tumors, macroglossia, umbilical hernia, exomphalos. The genes involved, CDKN1C, KCNQ1OT1; IGF2, H19, ICR1, SCL22A18AS, are included in the BWS region, the chromosomal band 11p14.5-11p15.5.



Figure 28 Phenotypic characterization of patient #14. The right hip is almost twice the left one.

Our patient presented an incomplete spectrum of the syndrome, with only overgrowth limited to the right part of skeletal system, connective tissue and internal organs, and a duplicated collecting system in the right kidney (Figure 28). The idea is that SCL22A18AS could be implicated in particular with the hemihypertrophy and the tissue overgrowth, and that each gene in the region has its specific function (e.g. IGF2 and hyperinsulinism). Only the mutation of a larger portion of the region 11p14.5-11p15.5 may lead to the complete phenotype of BWS. This discovery could have a great impact in the understanding of the pathophysiology of the disease.

Patient #15 is an 18 years old male with diagnosed muscular dystrophy, intellectual disability, autism spectrum disorder with behavioral impairment and epilepsy. WES led to identification of the variant c.239-2_239delAGAGinsT in KRBOX4 gene. This gene is described in only one work and appears to be related to an X-chromosomal disorder with intellectual disability and autism spectrum disorder. Functional studies will be performed and we will obtain more information from the additional family members in order to better understand the pathogenicity. (Table 6)

Table 6: Phenotype-genotype relationship of patient with variants found in novel candidate genes.

Patient	Consanguinity of parents	Family members with similar symptoms	Core phenotype	disease gene	variant	Mutation nucleotide	Mutation protein	Inheritance
#13	No	No	Epileptic encephalopathy, cerebral atrophy, white matter abnormalities. Intellectual disability, cardiomyopath, gait disturbance with spasticity.	<i>DENND5</i>	Missense	c.3083C>T	p.Arg1028Glx	De novo
#14	No	No	Global developmental delay, epilepsy, hypotonia, duplicated renal collecting system, right hemihypertrophy	<i>SCL22A18AS</i>	Nonsense		p.Glu177Ter	Autosomal recessive
#15	No	No	Muscular dystrophy, intellectual disability, autism spectrum disorder with behavioral impairment and epilepsy	<i>KRBOX4</i>	Splice acceptor	c.239-2_239delAGAinsT		X-chromosomal dominant, de novo.

3. CONCLUSION AND IMPLICATIONS FOR CLINICAL CARE.

The study supports the use of WES in clinical setting in order to improve the diagnostic rate in pediatric patients with rare and genetically undetermined medical and neurological conditions. The achieved diagnosis of at least 34,2% underlines the efficiency of exome sequencing in the pediatric cohort, reducing the diagnostic odyssey for many patients whose diagnosis was not established by other standard genetic techniques.

The advantage of WES is to sequence the coding regions of all human genes, that are 1% of the entire genome but include 85% of the known mutations. For this reason, it allows the identification of disease-causing variants in known genes as well as novel candidate genes. The results achieved in this study line up with the results obtained in previous studies.^{68,166} Moreover, our study showed better results than other studies involving complex and sporadic conditions.¹⁷⁸

A high diagnostic rate can be achieved combining the exome sequencing data with a detailed phenotyping. This study supports the importance of deep phenotyping using a standardized language created by the Human Phenotype Ontology, in order to create a large international network in which clinicians and geneticists from all over the world can confront the core phenotype and the genotype-phenotype correlation.

It is essential to obtain a good cooperation and data sharing from different parts of the world: this can help better understand the physiopathology of many disease and find novel disease genes.

Although WES has proved to be a powerful tool, the interpretation is not so simple, even with the development of specific guidelines (ACMG). Most identified variants are still classified as VUS, variants of uncertain significance, lacking evident pathogenicity but non clearly recognizable as polymorphisms. Functional studies

can be performed in those variants classified as VUS but with strong correlation with the phenotype. However, most clinicians don't take action due to the lack of clinical relevance. The advantage of WES is that it offers the possibility to reanalyze the data at different time points, according to changes in the clinical status of the patients and/or advancement in scientific knowledge.⁵⁴

Additionally, to clarify the VUS, it can be useful to sequence the DNA of other family members, to trace the segregation of the variant. Sanger sequencing is also used to perform segregation analysis to provide further evidence of pathogenicity and in case of novel candidate genes.

The next step in the diagnostic odyssey, in case of negative results, can be the whole genome sequencing (WGS).

A final consideration should be made on the impact of the molecular diagnosis achieved through WES on the management of the patients.

First of all, several studies have confirmed that specific treatment strategies can be adopted according to the peculiar pathophysiology on the basis of WES data, ranging from rare metabolic disorders to epileptic encephalopathies.^{68,179}

In our study, we identified in a patient a mutation causing a dopa-responsive dystonia: the treatment with L-DOPA could completely meliorate the dystonic posturing, consistently improving the quality of life of the patient.

Second of all, the molecular diagnosis plays a relevant role in the estimation of the recurrence risk for other family members, sometimes allowing an earlier diagnosis. However, the identification of genes related with neurodegenerative or untreatable disorders open the doors to ethical issues, particularly in a pediatric cohort.

In conclusion, WES currently represent the most versatile and cost effective application of NGS in clinical practice. It continuously improve both clinical diagnosis of rare genetic conditions and scientific research. Genetic diagnosis can significantly influence clinical management and potentially improve etiologically-targeted treatments. For this reason, clinicians and geneticists need to understand the perceptions and psychosocial impacts that the test has, to achieve the fullest degree of awareness, sensitivity and efficacy as possible.

Limitations of the study.

One of the limitations of the study has been the relatively small number of patients. Furthermore, our study lacked in the analysis and detection of CNV or structural variants. Although WES can detect CNVs, it is not the technique of choice for this kind of variants, being both CGHarrays and whole-genome-sequencing (WGS) a more valid choice. For this reason, for those with negative results, WGS could be performed to detect CNVs and to analyze those non-coding regions not sequenced by WES.

4. REFERENCES

1. Sheerin, U. The the Use of Next of Generation Parkinson ' s Sequencing Technologies to Dissect A etiologies disease and Dystonia. 1–225 (2014).
2. SANGER, F. The terminal peptides of insulin. *Biochem. J.* **45**, 563–574 (1949).
3. Sanger, F., Nicklen, S. & Coulson, A. . DNA sequencing with chain-terminating. *Proc Natl Acad Sci USA* **74**, 5463–5467 (1977).
4. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. 1977. *Biotechnology* **24**, 99–103 (1980).
5. <https://aulascienze.scuola.zanichelli.it/come-te-lo-spiego/2018/04/18/dal-metodo-sanger-a-oggi-40-anni-di-sequenziamento-del-dna/>. Dal metodo Sanger ad oggi: 40 anni di sequenziamento.
6. Swerdlow, H. & Gesteland, R. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res.* **18**, 1415–1419 (1990).
7. Kaiser, R. J. *et al.* Specific-primer-directed DNA sequencing using automated fluorescence detection. *Nucleic Acids Res.* **17**, 6087–6102 (1989).
8. DNA Sequencing process. <https://www.slideshare.net/MonsurAhmedShafiq/dna-sequencing-process>.
9. Gauthier, M. G. Simulation of polymer translocation through small channels. (2007).
10. Alberts, B. *et al.* Molecular Biology of the Cell. in 575–576 (2012).
11. Sanger Sequencing Steps | DNA Sequencing | Sigma-Aldrich. <https://www.sigmaaldrich.com/technical-documents/articles/biology/sanger-sequencing.html>.
12. Slatko, B. E., Gardner, A. F. & Ausubel, F. M. Overview of Next-Generation Sequencing Technologies. *Curr. Protoc. Mol. Biol.* **122**, 1–15 (2018).
13. Levy, S. E. & Myers, R. M. Advancements in Next-Generation Sequencing. *Annu. Rev. Genomics Hum. Genet.* **17**, 95–115 (2016).
14. F.S. Collins, E.S. Lander, J. Rogers, *et al.* International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).

15. Schloss, J. A. How to get genomes at one ten-thousandth the cost. *Nat. Biotechnol.* **26**, 1113–1115 (2008).
16. Shendure, J. *et al.* DNA sequencing at 40: Past, present and future. *Nature* **550**, (2017).
17. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **30**, 418–26 (2014).
18. Schnekenberg, R. P. & Németh, A. H. Next-generation sequencing in childhood disorders. *Arch. Dis. Child.* **99**, 284–290 (2014).
19. Voelkerding, K. V., Dames, S. & Durtschi, J. D. Next generation sequencing for clinical diagnostics-principles and application to targeted resequencing for hypertrophic cardiomyopathy: A Paper from the 2009 William Beaumont Hospital Symposium on Molecular Pathology. *J. Mol. Diagnostics* **12**, 539–551 (2010).
20. Mefford, H. C. *et al.* Rare copy number variants are an important cause of epileptic encephalopathies. *Ann. Neurol.* **70**, 974–985 (2011).
21. Gambin, T. *et al.* Identification of novel candidate disease genes from de novo exonic copy number variants. *Genome Med.* **9**, 1–15 (2017).
22. Li, Y., Chen, W., Liu, E. Y. & Zhou, Y. H. Single Nucleotide Polymorphism (SNP) Detection and Genotype Calling from Massively Parallel Sequencing (MPS) Data. *Stat. Biosci.* **5**, 3–25 (2013).
23. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* (2010) doi:10.1101/gr.107524.110.
24. Jones, J. D. G., & Dangl, J. L. (2006). The plant immune system. *Nature*, 444(7117), 323–329. <https://doi.org/10.1038/nature05286> *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)*. (2012) doi:10.4161/fly.19695.
25. Unknown. The Variant Call Format (VCF) Version 4 . 2 Specification. *Online Resour.* 1–28 (2015).
26. Westerfield, L. the Use of Whole Exome Sequencing To Detect Novel Genetic Disorders: Two Cases and an Assessment of the Technology. (2011).

27. Smedley, D. & Robinson, P. N. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Med.* **7**, 1–11 (2015).
28. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* (2010) doi:10.1093/nar/gkq603.
29. Coonrod, E. M., Margraf, R. L. & Voelkerding, K. V. Translating exome sequencing from research to clinical diagnostics. *Clin. Chem. Lab. Med.* **50**, 1161–1168 (2012).
30. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, (2015).
31. Behjati, S. & Tarpey, P. S. What is next generation sequencing? *Arch. Dis. Child. Educ. Pract. Ed.* **98**, 236–238 (2013).
32. Raza, K. & Ahmad, S. Recent advancement in next-generation sequencing techniques and its computational analysis. *Int. J. Bioinform. Res. Appl.* **15**, 191–220 (2019).
33. Chen, G. & Shi, T. L. Next-generation sequencing technologies for personalized medicine: Promising but challenging. *Sci. China Life Sci.* **56**, 101–103 (2013).
34. Valencia, C. A. *et al.* Assessment of target enrichment platforms using massively parallel sequencing for the mutation detection for congenital muscular dystrophy. *J. Mol. Diagnostics* **14**, 233–246 (2012).
35. Van Dijk, E. L., Jaszczyszyn, Y. & Thermes, C. Library preparation methods for next-generation sequencing: Tone down the bias. *Exp. Cell Res.* **322**, 12–20 (2014).
36. Schadt, E. E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. *Hum. Mol. Genet.* **19**, 227–240 (2010).
37. Pushkarev, D., Neff, N. F. & Quake, S. R. Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* **27**, 847–850 (2009).
38. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read

- sequencing data analysis. *Genome Biol.* **21**, 1–16 (2020).
39. Ammar, R., Paton, T. A., Torti, D., Shlien, A. & Bader, G. D. Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. *F1000Research* **4**, (2015).
 40. Dewey, F. E. *et al.* Clinical interpretation and implications of Whole Genome Sequencing. **311**, 1035–1045 (2015).
 41. Lemke, J. R. *et al.* Targeted next generation sequencing as a diagnostic tool in epileptic disorders. *Epilepsia* **53**, 1387–1398 (2012).
 42. Sun, Y. *et al.* Next generation diagnostics: gene panel, exome, or whole genome? 1–30 (2014) doi:10.1002/humu.22783.This.
 43. Sikkema-Raddatz, B. *et al.* Targeted Next-Generation Sequencing can Replace Sanger Sequencing in Clinical Diagnostics. *Hum. Mutat.* **34**, 1035–1042 (2013).
 44. Ankala, A. *et al.* A comprehensive genomic approach for neuromuscular diseases gives a high diagnostic yield. *Ann. Neurol.* **77**, 206–214 (2015).
 45. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, (2014).
 46. Dixon-Salazar, T. J. *et al.* Exome Sequencing Can Improve Diagnosis and Alter Patient Management. *Sci. Transl. Med.* **4**, (2012).
 47. Gonzalez-Garay, M. L. The road from next-generation sequencing to personalized medicine. *Per. Med.* **11**, 523–544 (2014).
 48. Pabinger, S. *et al.* A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.* **15**, 256–278 (2014).
 49. Ulintz, P. J., Wu, W. & Gates, C. M. *Bioinformatics Analysis of Whole Exome Sequencing Data. Methods in Molecular Biology* vol. 1881 (2019).
 50. Majewski, J., Schwartzenuber, J., Lalonde, E., Montpetit, A. & Jabado, N. What can exome sequencing do for you? *J. Med. Genet.* **48**, 580–589 (2011).
 51. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).
 52. Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Disease gene identification strategies for exome sequencing. *Eur. J. Hum. Genet.* **20**, 490–497 (2012).

53. Hamilton, A. *et al.* Concordance between whole-exome sequencing and clinical sanger sequencing: Implications for patient care. *Mol. Genet. Genomic Med.* **4**, 504–512 (2016).
54. Yang, Y. *et al.* Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.* **369**, 1502–1511 (2013).
55. Iglesias, A. *et al.* The usefulness of whole-exome sequencing in routine clinical practice. *Genet. Med.* **16**, 922–931 (2014).
56. Farwell, K. D. *et al.* Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: Results from 500 unselected families with undiagnosed genetic conditions. *Genet. Med.* **17**, 578–586 (2015).
57. Lee, H. *et al.* Clinical exome sequencing for genetic identification of rare mendelian disorders. *JAMA - J. Am. Med. Assoc.* **312**, 1880–1887 (2014).
58. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42**, 30–35 (2010).
59. Stark, Z. *et al.* A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. *Genet. Med.* **18**, 1090–1096 (2016).
60. Valencia, C. A. *et al.* Clinical Impact and Cost-Effectiveness of Whole Exome Sequencing as a Diagnostic Tool: A Pediatric Center’s Experience. *Front. Pediatr.* **3**, (2015).
61. Yang, Y. *et al.* Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA - J. Am. Med. Assoc.* **312**, 1870–1879 (2014).
62. Wenger, A. M., Guturu, H., Bernstein, J. A. & Bejerano, G. Systematic reanalysis of clinical exome data yields additional diagnoses: Implications for providers. *Genet. Med.* **19**, 209–214 (2017).
63. Eldomery, M. K. *et al.* Lessons learned from additional research analyses of unsolved clinical exome cases. *Genome Med.* **9**, 1–15 (2017).
64. Shamseldin, H. E. *et al.* Increasing the sensitivity of clinical exome sequencing through improved filtration strategy. *Genet. Med.* **19**, 593–598 (2017).

65. Koboldt, D. C. *et al.* Exome-based mapping and variant prioritization for inherited mendelian disorders. *Am. J. Hum. Genet.* **94**, 373–384 (2014).
66. Retterer, K. *et al.* Clinical application of whole-exome sequencing across clinical indications. *Genet Med* **18**, (2016).
67. Posey, J. E. *et al.* Molecular diagnostic experience of Whole Exome Sequencing in Adult Patients. **18**, 678–685 (2016).
68. Srivastava, S. *et al.* Clinical whole exome sequencing in child neurology practice. *Ann. Neurol.* **76**, 473–483 (2014).
69. Kuperberg, M. *et al.* Utility of Whole Exome Sequencing for Genetic Diagnosis of Previously Undiagnosed Pediatric Neurology Patients. *J. Child Neurol.* **31**, 1534–1539 (2016).
70. Thevenon, J. *et al.* Diagnostic odyssey in severe neurodevelopmental disorders: Toward clinical whole-exome sequencing as a first-line diagnostic test. *Clin. Genet.* **89**, 700–707 (2016).
71. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: An exome sequencing study. *Lancet* **380**, 1674–1682 (2012).
72. Tarailo-Graovac, M. *et al.* Exome sequencing and the management of neurometabolic disorders. *N. Engl. J. Med.* **374**, 2246–2255 (2016).
73. De Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).
74. O’Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.* **43**, 585–589 (2011).
75. Willsey, a J. *et al.* De novo mutations revealed by whole exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2013).
76. Pyle, A. *et al.* Exome sequencing in undiagnosed inherited and sporadic ataxias. *Brain* **138**, 276–283 (2015).
77. Chen, W. J. *et al.* Exome sequencing identifies truncating mutations in PRRT2 that cause paroxysmal kinesigenic dyskinesia. *Nat. Genet.* **43**, 1252–1255 (2011).
78. Choi, B. O. *et al.* Exome sequencing is an efficient tool for genetic screening of Charcot-Marie-Tooth Disease. *Hum. Mutat.* **33**, 1610–1615 (2012).

79. Drew, A. P. *et al.* Improved inherited peripheral neuropathy genetic diagnosis by whole-exome sequencing. *Mol. Genet. Genomic Med.* **3**, 143–154 (2015).
80. Walsh, M. *et al.* Diagnostic and cost utility of whole exome sequencing in peripheral neuropathy. *Ann. Clin. Transl. Neurol.* **4**, 318–325 (2017).
81. Klein, C. J. *et al.* Application of whole exome sequencing in undiagnosed inherited polyneuropathies. *J. Neurol. Neurosurg. Psychiatry* **85**, 1265–1272 (2014).
82. Hartley, T. *et al.* Whole-exome sequencing is a valuable diagnostic tool for inherited peripheral neuropathies: Outcomes from a cohort of 50 families. *Clin. Genet.* **93**, 301–309 (2018).
83. Heinzen, E. L. *et al.* Exome sequencing followed by large-scale genotyping fails to identify single rare variants of large effect in idiopathic generalized epilepsy. *Am. J. Hum. Genet.* **91**, 293–302 (2012).
84. Patel, J. *et al.* Diagnostic yield of genetic testing in epileptic encephalopathy in childhood. *J. Neurol. Sci.* **357**, e31 (2015).
85. Sawyer, S. L. *et al.* Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: Time to address gaps in care. *Clin. Genet.* **89**, 275–284 (2016).
86. Shashi, V. *et al.* The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. *Genet. Med.* **16**, 176–182 (2014).
87. Németh, A. H. *et al.* Next generation sequencing for molecular diagnosis of neurological disorders using ataxias as a model. *Brain* **136**, 3106–3118 (2013).
88. Calvo, S. E. *et al.* Molecular diagnosis of infantile mitochondrial disease with targeted next-generation sequencing. *Sci. Transl. Med.* **4**, (2012).
89. Kilpinen, H. & Barrett, J. C. How next-generation sequencing is transforming complex disease genetics. *Trends Genet.* **29**, 23–30 (2013).
90. Biesecker, L. G. & Green, R. C. Diagnostic clinical genome and exome sequencing. *N. Engl. J. Med.* **370**, 2418–2425 (2014).
91. Dorschner, M. O. *et al.* Actionable, pathogenic incidental findings in 1,000

- participants' exomes. *Am. J. Hum. Genet.* **93**, 631–640 (2013).
92. Green, R. C. *et al.* ACMG Recommendations for Reporting of Incidental Findings in Clinical Exome and Genome Sequencing. *Genet Med* **15**, 565–574 (2013).
 93. Matthijs, G. *et al.* Guidelines for diagnostic next-generation sequencing. *Eur. J. Hum. Genet.* **24**, 2–5 (2016).
 94. Wright, C. F. *et al.* Policy challenges of clinical genome sequencing. *BMJ* **347**, 1–6 (2013).
 95. Valente, E. M., Ferraris, A. & Dallapiccola, B. Genetic testing for paediatric neurological disorders. *Lancet Neurol.* **7**, 1113–1126 (2008).
 96. Rice, J. P., Saccone, N. L. & Rasmussen, E. 6 Definition of the phenotype. *Advances in Genetics* (2001) doi:10.1016/s0065-2660(01)42015-3.
 97. Shen, F., Wang, L. & Liu, H. Using human phenotype ontology for phenotypic analysis of clinical notes. *Stud. Health Technol. Inform.* **245**, 1285 (2017).
 98. Robinson, P. N. Deep phenotyping for precision medicine. *Hum. Mutat.* **33**, 777–780 (2012).
 99. Köhler, S., Bauer, S., Horn, D. & Robinson, P. N. Walking the Interactome for Prioritization of Candidate Disease Genes. *Am. J. Hum. Genet.* **82**, 949–958 (2008).
 100. Goh, K. Il *et al.* The human disease network. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 8685–8690 (2007).
 101. Robinson, P., Krawitz, P. & Mundlos, S. Strategies for exome and genome sequence data analysis in disease-gene discovery projects. *Clin. Genet.* **80**, 127–132 (2011).
 102. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–D798 (2015).
 103. Pavan, S. *et al.* Clinical practice guidelines for rare diseases: The orphanet database. *PLoS One* **12**, 1–14 (2017).
 104. Bragin, E. *et al.* DECIPHER: Database for the interpretation of phenotype-

- linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res.* **42**, 993–1000 (2014).
105. Robinson, P. N. *et al.* The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *Am. J. Hum. Genet.* **83**, 610–615 (2008).
 106. Smith, B. & Munn, K. *Applied Ontology*.
 107. Devitt, M. & Hanley, R. *The Blackwell Guide to the Philosophy of Language. The Blackwell Guide to the Philosophy of Language* (2008). doi:10.1002/9780470757031.
 108. Ashburner, M. *et al.* The Gene Ontology Consortium, Michael Ashburner¹, Catherine A. Ball³, Judith A. Blake⁴, David Botstein³, Heather Butler¹, J. Michael Cherry³, Allan P. Davis⁴, Kara Dolinski³, Selina S. Dwight³, Janan T. Eppig⁴, Midori A. Harris³, David P. Hill⁴, Laurie Is. *Nat. Genet.* **25**, 25–29 (2000).
 109. Day-Richter, J. *et al.* OBO-Edit - An ontology editor for biologists. *Bioinformatics* **23**, 2198–2200 (2007).
 110. Allanson, J. E., Cunniff, C., Hoyme, H. E., Mcgaughran, J. & Neri, G. Morphology of head and face. *Am. J. Med. Genet.* 6–28 (2009) doi:10.1002/ajmg.a.32612.Elements.
 111. Biesecker, L. G. An introduction to standardized clinical nomenclature for dysmorphic features: The Elements of Morphology project. *BMC Med.* **8**, (2010).
 112. Köhler, S. *et al.* The human phenotype ontology in 2017. *Nucleic Acids Res.* **45**, D865–D876 (2017).
 113. Human Phenotype Ontology. <https://hpo.jax.org/app/>.
 114. Köhler, S. *et al.* The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* **42**, 966–974 (2014).
 115. Claustres, M., Horaitis, O., Vanevski, M. & Cotton, R. G. H. Time for a unified system of mutation description and reporting: A review of locus-specific mutation databases. *Genome Res.* **12**, 680–688 (2002).
 116. Cotton, R. G. H. *et al.* Recommendations for locus-specific databases and

- their curation. *Hum. Mutat.* **29**, 2–5 (2008).
117. Robinson, P. N. & Mundlos, S. The Human Phenotype Ontology. *Clin. Genet.* **77**, 525–534 (2010).
 118. Le, D. H. & Dao, L. T. M. Annotating Diseases Using Human Phenotype Ontology Improves Prediction of Disease-Associated Long Non-coding RNAs. *J. Mol. Biol.* **430**, 2219–2230 (2018).
 119. Köhler, S. *et al.* Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. *Am. J. Hum. Genet.* **85**, 457–464 (2009).
 120. Xue, H., Peng, J. & Shang, X. Predicting disease-related phenotypes using an integrated phenotype similarity measurement based on HPO. *BMC Syst. Biol.* **13**, 1–12 (2019).
 121. Gong, X., Jiang, J., Duan, Z. & Lu, H. A new method to measure the semantic similarity from query phenotypic abnormalities to diseases based on the human phenotype ontology. *BMC Bioinformatics* **19**, (2018).
 122. Peng, J., Li, Q. & Shang, X. Investigations on factors influencing HPO-based semantic similarity calculation. *J. Biomed. Semantics* **8**, (2017).
 123. Groza, T. *et al.* The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease. *Am. J. Hum. Genet.* **97**, 111–124 (2015).
 124. Taboada, M., Rodriguez, H., Gudivada, R. C. & Martinez, D. A new synonym-substitution method to enrich the human phenotype ontology. *BMC Bioinformatics* **18**, 1–12 (2017).
 125. Notaro, M., Schubach, M., Robinson, P. N. & Valentini, G. Prediction of Human Phenotype Ontology terms by means of hierarchical ensemble methods. *BMC Bioinformatics* **18**, 1–18 (2017).
 126. Smedley, D. *et al.* Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat. Protoc.* (2015) doi:10.1038/nprot.2015.124.
 127. Robinson, P. N. *et al.* Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res.* **24**, 340–348 (2014).
 128. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
 129. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1 , 092

- human genomes. *Nature* **491**, 56–65 (2013).
130. Taylor JC, Martin HC, Lise S, Broxholme J, Cazier JB, Rimmer A, Kanapin A, Lunter G, Fiddy S, Allan C, Aricescu AR, Attar M, Babbs C, Becq J, Beeson D, Bento C, Bignell P, Blair E, Buckle VJ, Bull K, Cais O, Cario H, Chapel H, Copley RR, Cornall R, Craft, M. G. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders _ Nature Genetics _ Nature Publishing Group. *Nat Genet* **47**, 717–726 (2015).
 131. Stenson, P. D. *et al.* The Human Gene Mutation Database: Building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.* **133**, 1–9 (2014).
 132. Stark, Z. *et al.* A clinically driven variant prioritization framework outperforms purely computational approaches for the diagnostic analysis of singleton WES data. *Eur. J. Hum. Genet.* **25**, 1268–1272 (2017).
 133. Zhu, X. *et al.* Whole-exome sequencing in undiagnosed genetic diseases: Interpreting 119 trios. *Genet. Med.* **17**, 774–781 (2015).
 134. Birgmeier, J. *et al.* AMELIE 2 speeds up Mendelian diagnosis by matching patient phenotype & genotype to primary literature. (2019).
 135. Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. Mutationtaster2: Mutation prediction for the deep-sequencing age. *Nat. Methods* **11**, 361–362 (2014).
 136. Seelow, D., Schwarz, J. M. & Schuelke, M. Genedistiller - Distilling candidate genes from linkage intervals. *PLoS One* **3**, (2008).
 137. Stenson, P. D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.* **136**, 665–677 (2017).
 138. Carbon, S. *et al.* Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium. *Nucleic Acids Res.* **45**, D331–D338 (2017).
 139. Gray, K. A., Yates, B., Seal, R. L., Wright, M. W. & Bruford, E. A. Genenames.org: The HGNC resources in 2015. *Nucleic Acids Res.* **43**,

D1079–D1085 (2015).

140. Bateman, A. *et al.* UniProt: A hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).
141. Birgmeier, J. *et al.* AMELIE accelerates Mendelian patient diagnosis directly from the primary literature. *bioRxiv* 171322 (2017) doi:10.1101/171322.
142. Singleton, M. V. *et al.* Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *Am. J. Hum. Genet.* **94**, 599–610 (2014).
143. Schubach, M., Siragusa, E., Zemojtel, T. & Buske, O. J. the Exomiser. **10**, 2004–2015 (2017).
144. Yang, H., Robinson, P. N. & Wang, K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Am. J. Physiol. - Lung Cell. Mol. Physiol.* **268**, 841–843 (1995).
145. Javed, A., Agrawal, S. & Ng, P. C. Phen-gen: Combining phenotype and genotype to analyze rare disorders. *Nat. Methods* **11**, 935–937 (2014).
146. Sifrim, A. *et al.* EXtasy: Variant prioritization by genomic data fusion. *Nat. Methods* **10**, 1083–1086 (2013).
147. Fujiwara, T., Yamamoto, Y., Kim, J. D., Buske, O. & Takagi, T. PubCaseFinder: A Case-Report-Based, Phenotype-Driven Differential-Diagnosis System for Rare Diseases. *Am. J. Hum. Genet.* **103**, 389–399 (2018).
148. Wu, C. *et al.* Rapid and accurate interpretation of clinical exomes using Phenoxome: a computational phenotype-driven approach. *Eur. J. Hum. Genet.* **27**, 612–620 (2019).
149. Schulze, T. G. & McMahon, F. J. Defining the phenotype in human genetic studies: Forward genetics and reverse phenotyping. *Hum. Hered.* **58**, 131–138 (2004).
150. Williams, K. *et al.* Standard 6: Age groups for pediatric trials. *Pediatrics* **129**, (2012).
151. Contopoulos-Ioannidis, D. G. *et al.* Empirical evaluation of age groups and age-subgroup analyses in pediatric randomized trials and pediatric meta-analyses. *Pediatrics* **129**, (2012).

152. Huang, Y., Yu, S., Wu, Z. & Tang, B. Genetics of hereditary neurological disorders in children. *Transl. Pediatr.* **3**, 108–10819 (2014).
153. Piña-Garza. *Fenichel ' S Clinical Pediatric Neurology: A Signs and symptom approach.* (2013).
154. Schulz, J. B. Hereditäre Bewegungsstörungen. *Bundesgesundheitsblatt - Gesundheitsforsch. - Gesundheitsschutz* **50**, 1524–1530 (2007).
155. Wilmschurst, J. M. & Ouvrier, R. Hereditary peripheral neuropathies of childhood: An overview for clinicians. *Neuromuscul. Disord.* **21**, 763–775 (2011).
156. Sander, J. W. a S. & Shorvon, S. D. Epidemiology of the epilepsies Methodological issues. *J. Neurol. Neurosurg. Psychiatry* **61**, 433–443 (1996).
157. Roberts, R. The Epilepsies The diagnosis and management of the epilepsies in adults and children in primary and secondary care. (2012).
158. Collins, R. C. Epilepsy- Symptoms and causes- Mayo Clinic. *Neurosciences Research Program bulletin* vol. 14 498–501 (1976).
159. Penry, K. Proposal for Revised Clinical and Electroencephalographic Classification of Epileptic Seizures. *Epilepsia* **22**, 489–501 (1981).
160. Neurologia e psichiatria dello sviluppo - Emilio Franzoni, Martino Ruggieri - Google Libri.
https://books.google.it/books?id=V0Dn_UgKW9AC&printsec=frontcover&hl=it#v=onepage&q&f=false.
161. Pamplona, M. M. *et al.* Proposal for Revised Classification of Epilepsies and Epileptic Syndromes. *Epilepsia* **30**, 389–399 (1989).
162. Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. Rare-disease genetics in the era of next-generation sequencing: Discovery to translation. *Nat. Rev. Genet.* **14**, 681–691 (2013).
163. Chong, J. X. *et al.* The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *Am J Hum Genet* **97**, (2015).
164. Amberger, J., Bocchini, C. & Hamosh, A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Hum. Mutat.* **32**, 564–567 (2011).

165. Wright, C. F. *et al.* Genetic diagnosis of developmental disorders in the DDD study: A scalable analysis of genome-wide research data. *Lancet* **385**, 1305–1314 (2015).
166. Mahler, E. A. file:///Users/Eli/Downloads/Thevenon_et_a.-2016-C. pd. *et al.* Exom-Sequenzierung bei Kindern. *Dtsch. Arztebl. Int.* **116**, 197–204 (2019).
167. Rauer, C. & Finamore, N. Accelerating Genomics Research with OpenCL and FPGAs. *ALTERA Corp.* (2016).
168. Wu, D. *et al.* Angiogenin loss-of-function mutations in. **62**, 609–617 (2009).
169. Liu, Y. *et al.* PEX13 is mutated in complementation group 13 of the peroxisome- biogenesis disorders. *Am. J. Hum. Genet.* (1999) doi:10.1086/302534.
170. Segawa, M., Nomura, Y. & Nishiyama, N. Dopa-responsive dystonia. in *Handbook of Dystonia* (2006). doi:10.3109/9781841848525.015.
171. Wijemanne, S. & Jankovic, J. Dopa-responsive dystonia - Clinical and genetic heterogeneity. *Nature Reviews Neurology* (2015) doi:10.1038/nrneuro.2015.86.
172. Hall, E. A. *et al.* PLAA Mutations Cause a Lethal Infantile Epileptic Encephalopathy by Disrupting Ubiquitin-Mediated Endolysosomal Degradation of Synaptic Proteins. *Am. J. Hum. Genet.* (2017) doi:10.1016/j.ajhg.2017.03.008.
173. Zhang, X. *et al.* Structure of the AAA ATPase p97. *Mol. Cell* (2000) doi:10.1016/S1097-2765(00)00143-X.
174. Tang, J. Complexins. in *Encyclopedia of Neuroscience* (2009). doi:10.1016/B978-008045046-9.01372-3.
175. Yoshimura, S. I., Gerondopoulos, A., Linford, A., Rigden, D. J. & Barr, F. A. Family-wide characterization of the DENN domain Rab GDP-GTP exchange factors. *J. Cell Biol.* (2010) doi:10.1083/jcb.201008051.
176. Giannandrea, M. *et al.* Mutations in the Small GTPase Gene RAB39B Are Responsible for X-linked Mental Retardation Associated with Autism, Epilepsy, and Macrocephaly. *Am. J. Hum. Genet.* (2010) doi:10.1016/j.ajhg.2010.01.011.
177. Han, C. *et al.* Epileptic Encephalopathy Caused by Mutations in the Guanine

- Nucleotide Exchange Factor DENND5A. *Am. J. Hum. Genet.* (2016)
doi:10.1016/j.ajhg.2016.10.006.
178. Neveling, K. *et al.* A Post-Hoc Comparison of the Utility of Sanger Sequencing and Exome Sequencing for the Diagnosis of Heterogeneous Diseases. *Hum. Mutat.* **34**, 1721–1726 (2013).
179. Saunders, C. J. *et al.* Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci. Transl. Med.* **4**, (2012).

5. WEB RESOURCES AND TOOLS

Human Phenotype Ontology. <https://hpo.jax.org/app/>

GeneCards

OMIM <http://omim.org>

Orphanet <https://www.orpha.net>

M. Haeussler, Download, convert and process the full text of scientific articles:
<https://github.com/maximilianh/pubMunch3>

6. FIGURES

FIGURE 1 FREDERICH SANGER ⁵	6
FIGURE 2 SANGER MEHOD VS. MAXAM GILBERT METHOD ⁸	7
FIGURE 3 SANGER SEQUENCING METHOD IN 7 STEPS. COURTESY OF MICHAEL G GAUTHIER ⁹	8
FIGURE 4 AUTORADIOGRAPH ³	9
FIGURE 5 STEPS OF AUTOMATED SANGER SEQUENCING. COURTESY OF WWW.SIGMAALDRICH.COM ¹¹	10
FIGURE 6 DNA SEQUENCING TECHNOLOGIES. ¹⁶	12
FIGURE 7 TYPICAL NGS LIBRARY PREPARATION WORKFLOW. ¹⁷	13
FIGURE 8 AMPLIFICATION OF THE TEMPLATES ¹⁹	14
FIGURE 9 VARIANT CALLING AND ANNOTATION WORKFLOW. ²²	15
FIGURE 10 ANNOVAR DIFFERENT TYPES OF ANNOTATIONS. ²⁸	16
FIGURE 11 VARIANT FILTERING PIPELINE. ²⁹	18
FIGURE 12 THE EXOME IS ONLY 1% OF THE ENTIRE GENOME. COURTESY OF UNIVERSITY OF WASHINGTON.	22
FIGURE 13 WES WORKFLOW. ⁴⁸	23
FIGURE 14 HOW VARIANT FILTERING IN WES REDUCES THE NUMBER OF VARIANTS. ⁵²	24
FIGURE 15 DECISION-MAKING FLOWCHART FOR DIAGNOSTIC GENETIC TESTING IN PEDIATRIC NEUROLOGY CENTER, VALENTE, E. M., FERRARIS, A. & DALLAPICCOLA, B. GENETIC TESTING FOR PEDIATRIC NEUROLOGICAL DISORDERS. LANCET NEUROL. 7, 1113–1126 (2008) ⁹⁵	28
FIGURE 16 HUMAN PHENOTYPE ONTOLOGY NETWORK. ¹⁰⁵	31
FIGURE 17 WORKFLOW FOR REVERSE PHENOTYPING. ¹⁴⁹	36
FIGURE 18 CLINICAL AND GENETIC HETEROGENEITY ¹⁶⁵	40
FIGURE 19 NEURO-PEDIATRIC DIAGNOSTIC WORK-UP ¹⁶⁶	41
FIGURE 20 EXAMPLE OF THE DEEP PHENOTYPING: CORE PHENOTYPE AND THE HPO CODES	43
FIGURE 21 ACMG GUIDELINES FOR VARIANT CLASSIFICATION. ³⁰	48
FIGURE 22 AGE OF ONSET OF THE SYMPTOMS AND AGE OF INCLUSION IN THE STUDY.	50
FIGURE 23 PHENOTIPIC CHARACTERIZATION: OF THE STUDY GROUP ACCORDING TO HUMAN PHENOTYPE ONTOLOGY (HPO)	51
FIGURE 24 PIE CHART GRAPH SHOWING THE RESULTS OF THE STUDY.	53
FIGURE 25 RELEASE OF THE SYNAPTIC VESICLES: SNARE COMPLEX ASSEMBLY. COURTESY OF MAXIMOY AND TANG. ¹⁷⁴	62
FIGURE 26 ARTICLE SHOWING THE INTERACTION OF P-97 WITH THE PUL DOMAIN.	62
FIGURE 27 HUMAN DENN DOMAIN PROTEINS. COURTESY OF YOSHIMURA ET AL. ¹⁷⁵	63
FIGURE 28 PHENOTIPIC CHARACTERIZATION OF PATIENT #14. THE RIGHT HIP IS ALMOST TWICE THE LEFT ONE.	65

7. TABLES

TABLE 1 ADVANTAGES OF TARGETED SEQUENCING. SCHNEKENBERG, R. P. & NÉMETH, A. H. NEXT-GENERATION SEQUENCING IN CHILDHOOD DISORDERS. <i>ARCH. DIS. CHILD.</i> 99, 284–290 (2014). ¹⁸	21
TABLE 2 EVIDENCE OF PATHOGENICITY OF FILTERED VARIANTS ¹⁸	46
TABLE 3 FINDINGS OF WHOLE-EXOME SEQUENCING	52
TABLE 4 CHARACTERIZATION OF THE MUTATION IN A KNOWN DISEASE GENE.	54
TABLE 5 GENOTYPE-PHENOTYPE CORRELATIONS FOUND IN DIAGNOSED PATIENTS WITH PATHOGENIC VARIANTS IN A KNOWN DISEASE GENE.	55
TABLE 6: PHENOTYPE-GENOTYPE RELATIONSHIP OF PATIENT WITH VARIANTS FOUND IN NOVEL CANDIDATE GENES.	67

8. ACKNOWLEDGMENTS.

Desidero, in primis, ringraziare il mio relatore Prof. Vincenzo Salpietro, per le conoscenze trasmesse, per i suoi indispensabili consigli, per l'entusiasmo e per le grandi opportunità che mi sta fornendo.

Ringrazio infinitamente i miei genitori che mi hanno sempre sostenuta, economicamente e moralmente, accompagnandomi lungo questo percorso in ogni mia decisione e difficoltà.

A mia sorella Silvia, grazie per essere sempre stata presente anche a chilometri di distanza, per aver ascoltato i miei sfoghi strappandomi un sorriso.

Alla nonna Iole, la mia fan numero uno, alle nostre interminabili chiamate alle 11 di sera quando, finito di studiare, mi hai rallegrato con le tue storie di vita quotidiana.

A Fede, il mio fidanzato, il mio migliore amico, il mio compagno di studi e di vita, con te ho condiviso i momenti più belli di questi 6 anni: abbiamo affrontato insieme passo dopo passo questo cammino, festeggiando insieme ogni vittoria. Tante volte abbiamo fantasticato sui sogni e sui traguardi futuri. Questa è l'ennesima vittoria che festeggiamo insieme, il primo di tanti traguardi che taglieremo insieme. Grazie per esserci.

A Fra , insieme a me dal primo fino all'ultimo giorno di università. Grazie per le belle serate, per l'ottimismo e la spensieratezza, per avermi assecondato nella mia pazzia, memorizzando le cose più assurde e inutili di ogni materia, grazie a te questi 6 anni sono stati più leggeri.

Infine, un grazie generale al resto della mia famiglia, zii e cugini, e a tutti i miei amici, tutti voi avete contribuito a rendere questi anni memorabili.