



SAPIENZA
UNIVERSITÀ DI ROMA

SAPIENZA UNIVERSITY OF ROME

SCHOOL OF BIOLOGY AND MOLECULAR MEDICINE

PhD THESIS

**“NEXT GENERATION SEQUENCING APPROACHES IN RARE DISEASES:
THE STUDY OF FOUR DIFFERENT FAMILIES”**

Human Biology and Medical Genetics PhD course
Medical Genetics curriculum
XXXI cycle

Coordinator: Prof. Antonio Pizzuti

Tutor: Dr. Viviana Caputo

Candidate: Maria Luce Genovesi

Academic Year 2017-2018

INDEX

ABSTRACT	p. 3
1. INTRODUCTION	p. 4
1.1 Mendelian disorders and genetic tests	p. 4
1.2 Next generation sequencing (NGS) technology	p. 6
1.3 Bioinformatic pipeline	p. 10
1.4 Gene identification approaches	p. 12
1.5 NGS approaches	p. 13
1.5.1 Targeted sequencing (TS)	p. 13
1.5.2 Clinical exome sequencing (CES)	p. 15
1.5.3 Whole exome sequencing (WES)	p. 15
1.5.4 Whole genome sequencing (WGS)	p. 18
1.6 NGS approaches and their diagnostic rate	p. 19
1.7 NGS guidelines	p. 20
1.7.1 The American College of Medical Genetics and Genomics (ACMG) guidelines	p. 20
1.7.2 The Italian Society of Human Genetics (SIGU) guidelines	p. 22
2. AIM OF THE STUDY	p. 25
3. MATERIALS AND METHODS	p. 26
3.1 Subjects selection	p. 26
3.2 DNA extraction	p. 27
3.3 NGS techniques: whole exome and clinical exome sequencing	p. 28
3.4 Data analysis	p. 29
3.5 Selection of candidate variants (filtering and prioritization)	p. 31
3.6 Variants validation	p. 32
3.7 Modeling of the nucleotidyltransferase domain of FKTN (family D)	p. 34

4. RESULTS	p. 35
4.1 Family A	p. 35
4.2 Family B	p. 41
4.3 Family C	p. 45
4.4 Family D	p. 49
4.4.1 Modeling of the nucleotidyltransferase domain of FKTN	p. 52
5. DISCUSSION	p. 55
5.1 Family A	p. 56
5.2 Family B	p. 60
5.3 Families C and D	p. 63
6. CONCLUSIONS	p. 70
BIBLIOGRAPHY	p. 71
SITOGRAPHY	p. 84

ABSTRACT

The main purpose of this PhD project was to study the molecular bases of rare Mendelian diseases with Next Generation Sequencing approaches.

To this aim, we enrolled at Umberto I General Hospital and Sapienza University of Rome four different families with a phenotype with a supposed genetic cause, in order to find the causative gene/genes. Clinical exome sequencing or whole exome sequencing was performed on selected subjects of each family. The supposed mode of inheritance defined the selection and the number of individuals to sequence, as well as the analytical approach to use. Sequencing data were analysed through a dedicated bioinformatic pipeline; variants were then filtered and prioritized according to several parameters, specific for each case.

The selected variant/variants were validated through Sanger sequencing on the proband and on the other family members, to study their segregation in the family.

The functional link between the candidate variant/variants and the phenotype was investigated, retrieving information from literature and online resources.

In the four studied families the different approaches allowed us to identify the molecular causes of each disorder, with consequences on diagnosis, prognosis and genetic counselling.

1. INTRODUCTION

1.1 Mendelian disorders and genetic tests

Mendelian or monogenic diseases are caused by mutation in one gene. For this kind of disorders alternative genotypes fall into distinct and discrete phenotypes (Antonarakis and Beckmann, 2006). They are usually inherited in one of several patterns, depending on the location of the gene and whether one or two normal copies of the gene are needed for the disease phenotype to manifest: the expression of the mutated allele with respect to the normal one can be dominant, co-dominant or recessive; the five basic modes of inheritance for single-gene diseases are autosomal dominant, autosomal recessive, X-linked dominant, X-linked recessive and mitochondrial. To date, Mendelian diseases are estimated to be about 7,000 (Boycott et al., 2017). Clinically recognized Mendelian phenotypes compose a substantial fraction (~0.4% of live births) of known human diseases. If all congenital anomalies are included, ~8% of live births have a genetic disorder recognizable by early adulthood (Chong et al., 2015).

Of approximately ~19,000 protein-coding genes predicted to exist in the human genome: variants causing Mendelian phenotypes have been identified in ~2,937 (~15.5%); genes underlying ~643 Mendelian phenotypes (~3.38%) have been mapped but not yet identified; loss of function variants in up to ~30% of genes (~5,960) could result in embryonic lethality in humans; for a minimum of ~52% of genes (~10,330), the impact in humans has not yet been determined. Collectively, ~16,063 genes remain candidates for Mendelian phenotypes (Chong et al., 2015; Figure 1).

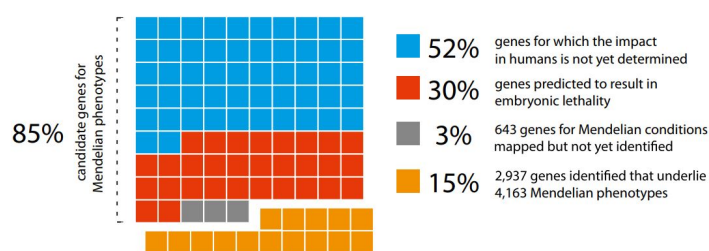


Figure 1. Relationship between human protein-coding genes and Mendelian phenotypes (Chong et al., 2015).

Causative genetic variants can range in size from substitution, deletion or duplication of a single base pair to structural variants and to altered copy numbers of an entire chromosome (aneuploidy); in some cases, the penetrance of the disorder can be incomplete and the expressivity of individual

features can be variable. A proportion of variability in genetic diseases can also be attributed to locus heterogeneity and allelic heterogeneity; other important sources of variability include genetic variants at one or more other loci (modifiers) and environmental factors (Wright et al., 2018).

In the past, the identification of Mendelian disease genes was carried out by linkage mapping and Sanger sequencing of candidate genes, which were selected because they reminded of genes associated with similar diseases, because the predicted protein function seemed relevant to the physiology of the disease or because a positional mapping approach pointed to these genes in a genomic region (Gilissen et al., 2012).

Traditionally, there are two kinds of clinical genetic tests: high-resolution molecular single gene tests by Sanger sequencing and low-resolution genome-wide cytogenetic tests. The first ones are useful for the diagnosis of that conditions caused by just one or few genes, as cystic fibrosis or Duchenne muscular dystrophy; the second kind of tests can be used to diagnose aneuploidies and chromosome rearrangements (G-banded karyotype) or smaller structural variants (microarray) (Wright et al., 2018).

Next generation sequencing (NGS) has revolutionized medical genetics through the high-throughput massively parallel sequencing (Figure 2): it is accelerating the research about rare-genetic diseases and it is facilitating clinical diagnosis and personalized medicine. In the last decade the capacity of NGS technology has increased, leading a throughput several orders of magnitude higher than Sanger sequencing (Goodwin et al., 2016), and its costs have come down considerably, facilitating the translation of sequencing from a research technology to a clinical tool.



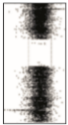
	Light microscope	G-banded karyotype	Microarray	Whole-exome sequence	Whole-genome sequence
Appearance				CGGATGATTACCCGTT G.....GCTC TAGCTAGCTATA....	CGGATGATTACCCGTT GATATAGCTCTCGCTC GCTCTAGCTAGCTATA GGCTATGGGTGGGGCC
Resolution	Entire chromosome	5–10 Mb	50–100 kb	1 bp	1 bp
Number of loci probed	N/A	~500	~0.05–2 million	~50 million	3 billion
Variants detected	Aneuploidy, polyploidy	Variants >5 Mb	Copy number variants	Coding regions	Majority of variants
Variants per person	0 or 1	0 or 1	10–100s	~20,000	4–5 million
Diagnostic yield	Low	—————→			High
Incidental findings	Low	—————→			High

Figure 2. Genome-wide assays used in clinical genetics: from traditional methods to whole genome sequencing.

As the resolution of the test increases, the number of detectable variants, the diagnostic yield and likelihood of detecting incidental findings and variants of uncertain clinical significance increase too (Wright et al., 2018).

For a long time, a clinician first exhausted a battery of medical tests and then he turned to genetic testing only if the previous ones did not yield a definitive diagnosis or if there was a need to assess recurrence risk. Even positive genetic test results did not often change management of the patients. However, the introduction of NGS technology in the clinics and the increased knowledge in genetics let the clinicians begin altering the placement of genetic testing in the evaluation of their patients, saving time and money in identifying an aetiology.

1.2 Next generation sequencing (NGS) technology

Traditionally, NGS experiments have been performed using short-read sequencing (SRS) that produces reads from 100 to 400 bp in length, depending on the technology. SRS is based on library preparation by random fragmentation of input DNA, adapter ligation, amplification and massively parallel sequencing of adapter-ligated fragments (Caspar et al., 2018).

There are two categories of short-read sequencing approaches: sequencing by ligation (SBL) and sequencing by synthesis (SBS). In SBL approaches (SOLiD and Complete Genomics), a labelled probe and anchor sequences hybridize to a DNA fragment and are ligated to an adjacent oligonucleotide through a DNA ligase. After ligation, the template is imaged and the emission spectrum of the fluorophore indicates which base or bases are complementary to a specific position inside the probe. The removal of the anchor–probe complex allows to regenerate the ligation site: a new cycle can begin (Goodwin et al., 2016; Figure 3).

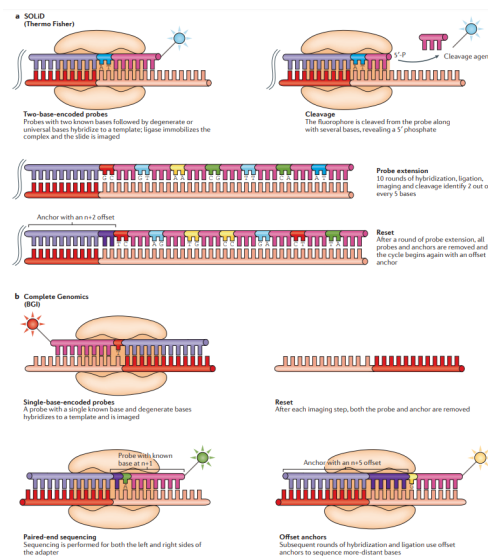


Figure 3. Sequencing by ligation methods: SOLiD (a) and Complete Genomics (b) (Goodwin et al., 2016).

SBS approaches can be classified either as cyclic reversible termination (CRT) or as single-nucleotide addition (SNA). In these approaches a polymerase incorporates a nucleotide into an elongating strand producing a signal, such as a fluorophore or a change in ionic concentration. CRT approaches (Illumina, Qiagen; Figure 4) use similar terminator molecules to those used in Sanger sequencing, in which the ribose 3'-OH group is blocked, preventing elongation. These molecules are individually labelled. After the incorporation of a single dNTP, unbound dNTPs are removed and the surface is imaged to identify which dNTP was incorporated at each cluster. The fluorophore and blocking group can then be removed and a new cycle can begin (Goodwin et al., 2016).

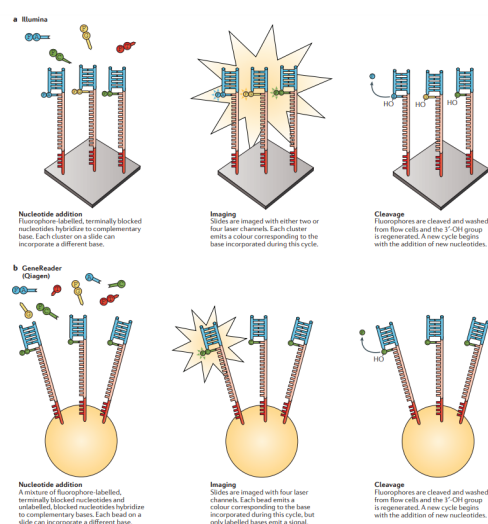


Figure 4. Sequencing by synthesis (cyclic reversible termination approaches): Illumina (a) and Qiagen (b) (Goodwin et al., 2016).

The Illumina technology reaches its maximum throughput with the HiSeq X Ten System, a set of 10 HiSeq X platforms, which generates tens of thousands of high-quality and high-coverage genome sequences, breaking the \$1000 barrier for 30× coverage of a human genome.

SNA approaches (454, Ion Torrent; Figure 5) rely on a single signal to mark the incorporation of a dNTP into an elongating strand. For this reason, each nucleotide has to be added individually. The 454 technology is based on pyrosequencing: when a dNTP is incorporated, an enzymatic cascade occurs, resulting in a bioluminescence signal; the Ion Torrent platform detects a change in pH: when a dNTP is incorporated, H⁺ ions are released (Goodwin et al., 2016).

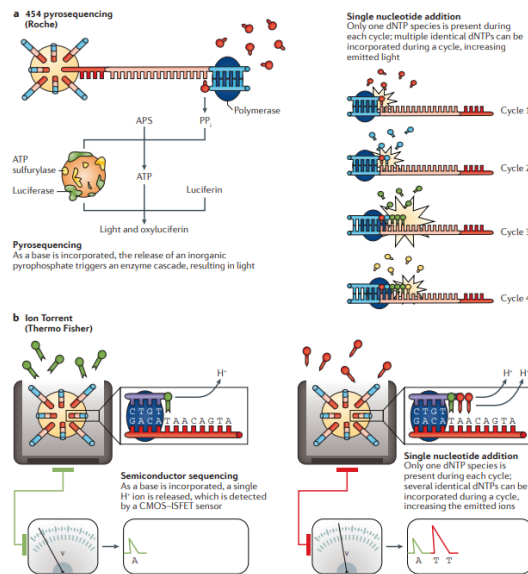


Figure 5. Sequencing by synthesis (single-nucleotide addition approaches): 454 pyrosequencing (a) and Ion Torrent (b) (Goodwin et al., 2016).

In both SBL and SBS approaches DNA is clonally amplified on a solid surface: thousands of identical copies of a DNA fragment in a defined area allow the signal to be distinguished from background noise. Different strategies can be used to generate clonal template populations: emulsion PCR [454 (Roche), SOLiD (Thermo Fisher), GeneReader (Qiagen), Ion Torrent (Thermo Fisher)], solid-phase bridge amplification (Illumina), in-solution DNA nanoball generation [Complete Genomics (BGI)]. Millions of individual SBL or SBS reaction centres are created, each with its own clonal DNA template and, in this way, millions of DNA molecules are sequenced in parallel (Goodwin et al., 2016).

The main advantages of SRS are high-throughput, low per base cost and low raw-read error rate; the main disadvantage is the short-read length, which leads to reads alignment difficulties, that can cause misalignments, false-positive and false-negative variant calling (Caspar et al., 2018).

These limitations can be overcome using long-read sequencing (LRS) or third-generation sequencing, which is a single-molecule sequencing PCR-free: the long reads facilitate unambiguous alignment to a reference genome through their increased ability to span large, complex, repetitive or homologous regions. However, LRS is not yet routinely applied due to its significantly lower throughput and higher per sample sequencing cost; furthermore, it has a high raw error rate of ~10%. These errors can be minimized by increasing read depth and reading the template multiple times (Caspar et al., 2018).

There are two main long-read technologies: single-molecule real-time sequencing approaches [PacBio and Oxford Nanopore Technologies (ONT); Figure 6] and synthetic approaches (Illumina synthetic long-read sequencing platform, 10X Genomics emulsion-based system; Figure 7), that

construct *in silico* long reads based on short-read technologies. Pacific Biosciences (PacBio) instrument has a flow cell with thousands of picolitre wells with transparent bottoms, called zero-mode waveguides (ZMW), in which there is the DNA polymerase. There is a single circular molecule template per well and when the labelled nucleotide momentarily pauses during incorporation at the bottom of the ZMW, dNTP incorporation is continuously visualized with a laser and a camera system that records the colour and duration of emitted light. Each template is sequenced multiple times as a function of its length as the polymerase repeatedly traverses the circular molecule. Oxford Nanopore sequencers directly detect the DNA composition of a native single strand DNA molecule, which passes through a protein pore modifying the current that passes through the pore. These sequencers have flow cells with thousands of pores (Goodwin et al., 2016).

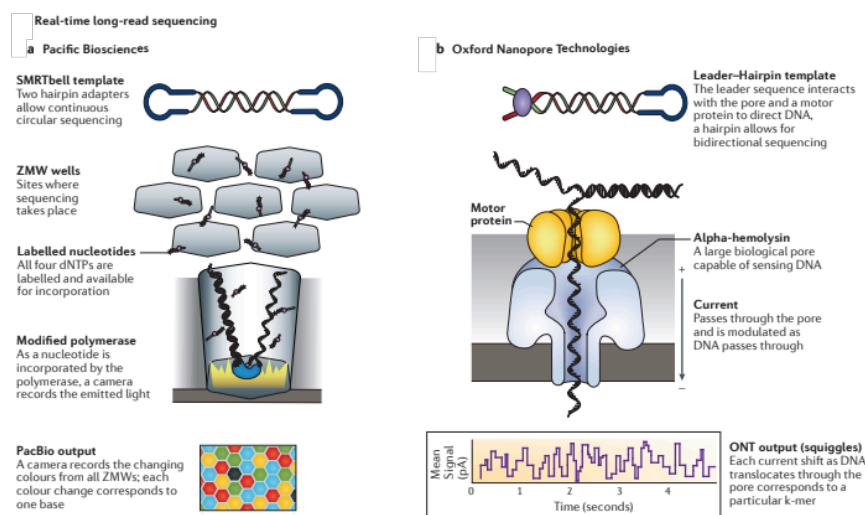


Figure 6. Real-time long-read sequencing approaches: Single-molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio; a) and Oxford Nanopore Technologies (ONT; b) (Goodwin et al., 2016).

Synthetic long-read technology is based on a system of barcoding to associate fragments sequenced on existing short-read sequencers. The Illumina system partitions DNA into a microtitre plate and does not require specialized instrumentation; the 10X Genomics instruments use emulsion to partition DNA and require the use of a microfluidic instrument to perform pre-sequencing reactions (Goodwin et al., 2016).

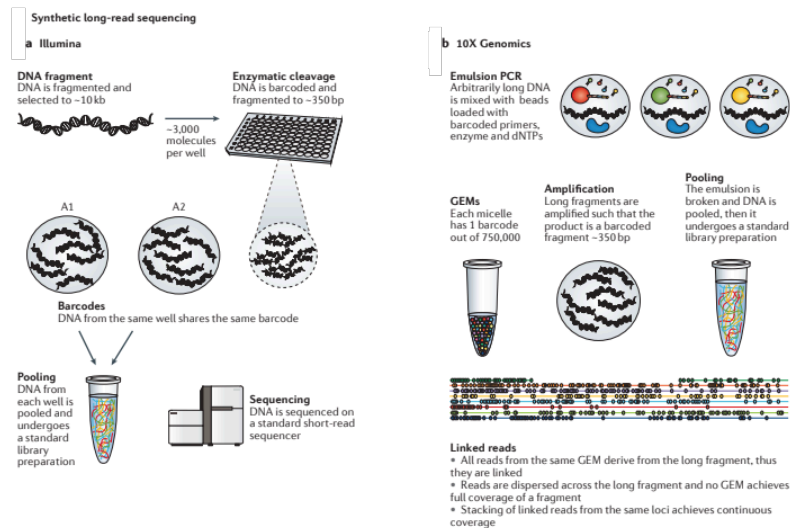


Figure 7. Synthetic long-read sequencing approaches: Illumina (a) and 10X Genomics' emulsion-based sequencing (b) (Goodwin et al., 2016).

1.3 Bioinformatic pipeline

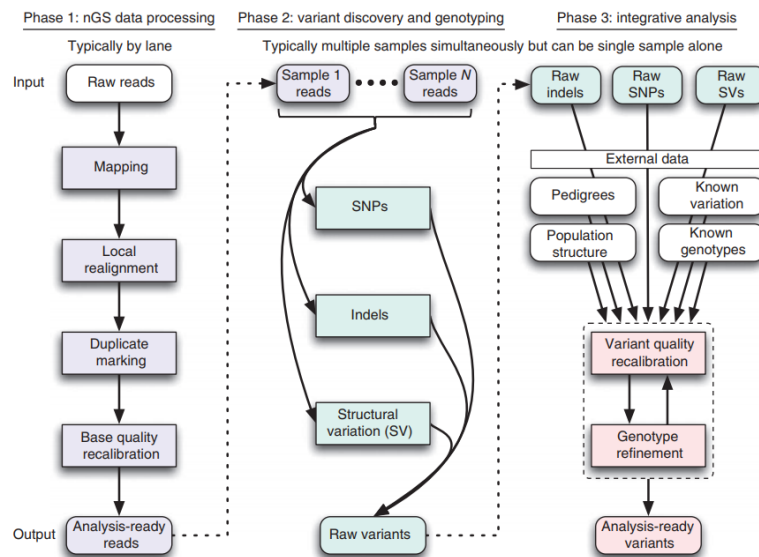


Figure 8. Framework for variant discovery and genotyping from next-generation sequencing data (DePristo et al., 2011).

After completing the sequencing run, raw reads data (FASTQ files) have to be processed (Figure 8): the first analysis step is to evaluate the quality of raw reads and to remove, trim or correct reads that do not meet the defined standards. During this step also adaptor sequences are removed (DePristo et al., 2011). The second step consists in the alignment of the reads to the reference genome, which is given in a FASTA format. Currently, there are two main sources for the human reference genome assembly: the University of Santa Cruz (UCSC) and the Genome Reference Consortium (GRC).

Both provide several versions of the human genome, at the moment versions hg19 and hg38 for the UCSC assembly and GRCh37 and GRCh38 for the GRC one. Both human assemblies are identical but differ with regards to their nomenclature (Pabinger et al., 2014). After the alignment, the duplicate reads are removed: indeed, during library amplification and sequencing process the same DNA molecules can be sequenced several times and the multiple reads can interfere with variant calling statistics. The output is a SAM (sequence alignment/map) file containing all retained reads (Pabinger et al., 2014). This format is commonly used to store next-generation sequencing alignments. SAM files can be easily converted to the BAM (binary alignment/map) format, that is a binary representation of the SAM. Initial alignments are refined by local realignment and then there is the step of base recalibration that assigns a well-calibrated probability to each base call (Pabinger et al., 2014).

Two important parameters to evaluate a NGS experiment are coverage and depth: the empirical per-base coverage represents the exact number of times that a base in the reference genome is covered by a high-quality aligned read from a given sequencing experiment; redundancy of coverage is also called the depth or the depth of coverage (Sims et al., 2014).

The following step is variant calling, which consists in the identification of the DNA sequence variations relative to the reference genome. Variations that can be recognized are single-nucleotide variants (SNVs) and small insertion-deletions (Indels). The output file of this analysis is a Variant Calling File (VCF) (Van der Auwera et al., 2013). Then there is the functional annotation of the variants and the genes that harbour them (Pabinger et al., 2014), a process that places mutations identified by the variant calling step into their biological context (Salgado et al., 2016). The main objective of the annotation step is to gather substantial information at the variant and at the gene levels. At the variant level it includes data quality, genomic position, genotype, frequency in the general population, impact at the mRNA and protein levels, conservation of the affected protein residues among species, variant pathogenicity prediction and reported associations with diseases. At the gene level it includes the function of the gene, tissue expression pattern, involvement in pathways and in phenotypes/diseases (Salgado et al., 2016). Accurate annotation of variants is important to understand their functional effects and to select the most promising candidate pathogenic mutations. Accurate annotation of genes is critical to understand the functional associations of the genes with pathways in normal and disease states (Chakravorty and Hegde, 2017). Finally, the variants are filtered based on quality criteria and prioritized, according to the specific disease, on the basis of pedigree information and the mode of inheritance, the localization of the variant, the mutation type, the frequency of the variant, the predicted impact of the variant on protein function and structure, the functional evidences and the evolutionary conservation of variant

nucleotide (Salgado et al., 2016). The aim of these two last steps is to combine different criteria to identify potentially candidate variants (Salgado et al., 2016). There are two options to proceed with the prioritization of the variants: one is to employ a semiautomatic prioritization system, which is a useful approach for example when the phenotype is clearly described using the proper phenotype ontology such as the Human Phenotype Ontology (HPO terms); the second one is to adopt a manual prioritization procedure based on expert knowledge about disease phenotype and gene functions. This approach can be greatly facilitated by the use of filtration tools (Salgado et al., 2016).

1.4 Gene identification approaches

When a rare phenotype is recurring in a family, the likelihood of a monogenic rare disease is high. The mode of inheritance influences the selection and the number of individuals to sequence, as well as the analytical approach to use (Boycott et al., 2013; Figure 9).

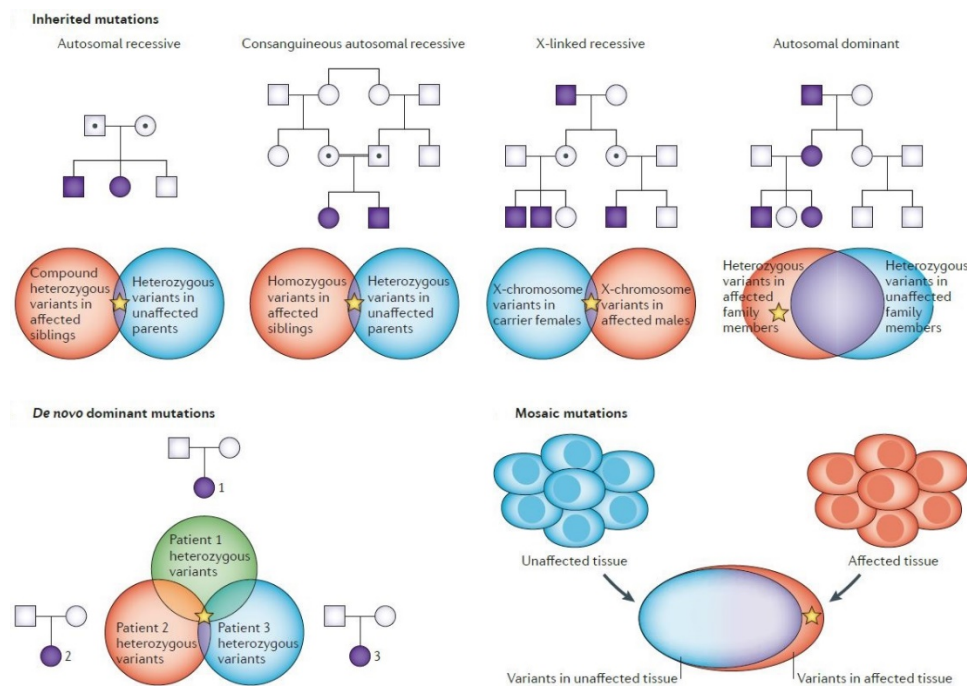


Figure 9. Gene identification approaches for different categories of rare diseases (Boycott et al., 2013).

For autosomal recessive disorders, sib pair analysis is often needed to reduce the number of gene variants to one or few candidates. For this kind of diseases, compound heterozygous or homozygous variants should be searched in affected siblings and heterozygous variants in their unaffected parents. For X-linked recessive diseases, the favoured strategy is to analyse the two most remotely related male family members, looking for X-chromosome variants in carrier females and

in affected males. For autosomal dominant disorders, the approach is to analyse the two most remotely related family members, looking for shared heterozygous variants, which are absent in unaffected family members. Analysis of whole-exome sequencing data from unaffected parents and the affected child (*trio* analysis) is extremely useful for *de novo* variants, which are in a heterozygous state in the proband and absent in unaffected parents; comparison of these heterozygous variants between different families in which the probands have the same phenotype generally reduces these variants to a single candidate gene. The comparison of sequence data from a patient's affected and unaffected tissues is frequently sufficient to identify mosaic disease-causing mutations (Boycott et al., 2013).

1.5 NGS approaches

The NGS technology allows to sequence a specific subset of genes (targeted sequencing, TS; clinical exome sequencing, CES), the exome (whole exome sequencing, WES) or the entire genome (whole genome sequencing, WGS) in a matter of hours to days, depending on the protocol and the platform used.

The management of NGS data, the lack of understanding of the impact of most genetic variants on human health and disease and the amount of secondary findings which can be found during a genetic test are some of the parameters that a clinician has to consider before assigning a NGS test to patients.

1.5.1 Targeted sequencing (TS)

Targeted sequencing (TS) or gene panel sequencing allows to enrich only the coding regions of genes of interest for a specific disease or a diagnostic category (Table 1).

Disease area	Disease type	Genes
Cancer	Hereditary cancers (for example, breast, colon and ovarian)	10–50
Cardiac diseases	Cardiomyopathies	50–70
	Arrhythmias (for example, long QT syndrome)	10–30
	Aortopathies (for example, Marfan's syndrome)	10
Immune disorders	Severe combined immunodeficiency syndrome	18
	Periodic fever	7
Neurological, neuromuscular and metabolic disorders	Ataxia	40
	Cellular energetics, metabolism	656
	Congenital disorders of glycosylation	23–28
	Dementia (for example, Parkinson's disease and Alzheimer's disease)	32
	Developmental delay, autism, intellectual disability	30–150
	Epilepsy	53–130
	Hereditary neuropathy	34
	Microcephaly	11
	Mitochondrial disorders	37–450
	Muscular dystrophy	12–45
	Sensory disorders	Eye disease (for example, retinitis pigmentosa)
Hearing loss and related syndromes		23–72
Other	Rasopathies (for example, Noonan's syndrome)	10
	Pulmonary disorders (for example, cystic fibrosis)	12–40
	Short stature	12

Table 1. Clinically available disease-targeted tests (Rehm, 2013).

This strategy has some advantages:

- it is cheaper than the other ones;
- panels can have a much higher or often complete coverage of the genes they contain, because the gaps can be filled with supplemental Sanger sequencing and other complementary technologies (Rehm, 2013);
- a targeted approach also allows for a deeper coverage of all phenotype-specific genes, providing a greater confidence in the variants detection (Jamuar and Tan, 2015). In order to use TS for clinical diagnostics, high-quality data are essential, i.e. not the mean on-target depth, but ideally all nucleotides are seen at a minimal read depth of 20x–40x (Weiss et al., 2013). For this reason, a targeted approach is more efficient to reveal mosaic mutations than WES;
- panels usually are used in laboratories with an extensive clinical experience with a particular disease and its causative genes, so these laboratories may be better able to prioritize variants in those genes and to understand their clinical significance (Rehm, 2013);
- targeted sequencing minimizes the problems of incidental findings (Rehm, 2013);
- depending on the enrichment strategy and the platform used, several hundred target genes can be sequenced for multiple patients in the same run. Data can also be analysed within a relatively short processing time (de Koning et al., 2015);
- the size of the data files generated by panels is small and it is possible to store not only the variant files but also the FASTQ files for longer periods (Weiss et al., 2013).

However, this strategy has also some disadvantages:

- as the other NGS approaches, it is prone to sequencing artefacts and Sanger sequencing of candidate variants is always recommended (Jamuar and Tan, 2015);
- panels have to be continuously updated when new genes are identified (Jamuar and Tan, 2015).

For these reasons many laboratories have now shifted to performing WES and limiting the analysis to genes associated with phenotype and filling up the gaps with Sanger sequencing (*in silico* gene panels). Although this strategy is more expensive, it allows for re-analyse the data when new genes are discovered (Jamuar and Tan, 2015).

1.5.2 Clinical exome sequencing (CES)

Clinical exome sequencing is a technology that allows to sequence all the genes associated with diseases and it is being applied to a wide range of clinical presentations that require a broad search for causal variants across the spectrum of genetically heterogeneous Mendelian disorders (Lee H et al., 2014).

The TruSight One Sequencing Panel (2014) provides comprehensive coverage of about 4,800 disease-associated genes, while the TruSight One Expanded Sequencing Panel (2017) targets ~1,900 additional genes with recent disease associations in the scientific literature (<https://www.illumina.com/products/by-type/clinical-research-products/trusight-one.html>).

The analysis can be initially limited to only those genes that are relevant to the patient's phenotype, but then it can be extended to a much broader gene set or even to the entire disease-associated exome (Rehm, 2013). This approach is simple as a disease-targeted test and it enables the laboratory to minimize test development and validation efforts (Rehm, 2013).

1.5.3 Whole exome sequencing (WES)

Whole exome sequencing is currently the most used approach for the discovery of those rare-disease-causing genes that conventional approaches have failed to identify. It is estimated, in fact, that 85% of the disease-causing mutations is located in coding and functional regions of the genome. For this reason, sequencing of the complete coding regions (exome) can uncover the causes of a large number of rare genetic disorders as well as predisposing variants of common diseases and cancers.

The initial proof-of-concept for using WES in rare-disease research came with the identification of

genes responsible for the dominant Freeman–Sheldon syndrome (OMIM #193700; Ng et al., 2009), recessive Miller syndrome (OMIM #263750; Ng et al., 2010) and dominant Schinzel–Giedion syndrome (OMIM #269150; Hoischen et al., 2010), respectively *MYH3*, *DHODH* and *SETBP1*. Then the discovery of disease-causing genes using WES has increased rapidly (Boycott et al., 2013; Figure 10).

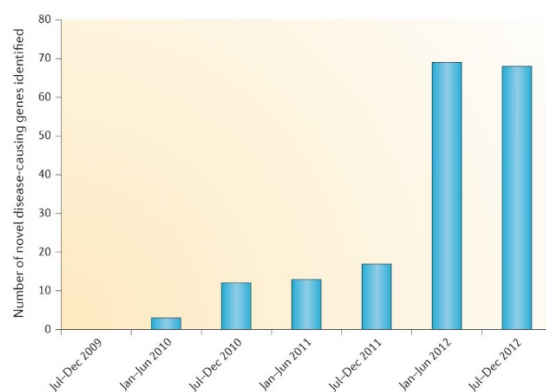


Figure 10. Rate of discovery of novel rare-disease-causing genes using whole-exome sequencing (Boycott et al., 2013).

WES has been also used to identify the causative variants in several heterogeneous conditions, such as hearing loss, intellectual disabilities, autism spectrum disorders and retinitis pigmentosa (Rabbani et al., 2014). WES has been successfully deployed in the clinics too, appearing as the most cost-effective NGS approach (Jamuar and Tan, 2015). It is probably the most efficient technique for identifying *de novo* mutations in a parents-patient *trio* approach for heterogeneous disorders with very large numbers of putative causative genes (de Koning et al., 2015).

Although WES is supposed to cover all the protein-coding regions of the genome, the overall coverage depends on the enrichment strategy used and it tends to be between 85-95% only. The reasons include poorly performing capture probes due to high GC content, sequence homology and repetitive sequences (Jamuar and Tan, 2015). Coverage is also heterogeneous probably because of the hybridization/capture and PCR-amplification steps required for the preparation of sequencing libraries for WES (Krebschull and Zador, 2015).

Coverage of almost each nucleotide of interest is of major importance for the application of NGS technology in clinical diagnostics and sequence depth is, therefore, an important quality parameter in NGS applications. At a mean on-target read depth of 20x, which is commonly used in WES studies for diagnosing rare disorders for instance, one would miss 5–15% of the heterozygous and 1–4% of the homozygous single nucleotide variants in the targeted regions (Meynert et al., 2013).

An exome approach produces terabytes of data that demand major storage capacity (Weiss et al.,

2013). On average, ~60,000 to 100,000 variants are detected during a WES experiment. These variants can be classified into pathogenic, benign and variants of uncertain significance (VUS). Pathogenic variants are those that adversely alter protein function and have either been reported previously in other affected individuals or have been shown to affect protein function in cellular or animal models. Benign variants or polymorphisms exist in a significant proportion of the population and account for the majority of the variants detected through NGS testing. VUS are variants that could possibly affect protein function based on *in silico* prediction, but they either have not been described in other individuals (affected or unaffected) or do not have any functional analysis in other model systems (Jamuar and Tan, 2015).

Using WES, a single pathogenic variant can be detected about 20–36% of the time; in other cases, it is possible to either find multiple candidate variants or no candidate variant. If multiple candidate variants were detected, segregation analysis and/or functional analysis would help to determine the molecular aetiology. If no variants were detected, it would be possible that the causal variant is in a poorly covered region or outside protein-coding regions (Jamuar and Tan, 2015). WES is prone to sequencing artefacts and Sanger sequencing of candidate variants is always recommended (Jamuar and Tan, 2015).

WES is not a useful approach for the identification of copy number variants (CNVs), due to the non-contiguous nature of the captured exons and to the extension of most CNVs beyond the regions covered by the exome kit (Belkadi et al., 2015). However, numerous methods have been developed to detect CNVs from exome sequencing data, like ExomeDepth (Plagnol et al., 2012), ExomeCopy (Love et al., 2011),XHMM (Fromer et al., 2012), cn.MOPS (Klambauer et al., 2012), ExomeCNV (Sathirapongsasuti et al., 2011), CoNVEX (Amarasinghe et al., 2013), EXCAVATOR (Magi et al., 2013), CoNIFER (Krumm et al., 2012), CANOES (Backenroth et al., 2014), CODEX (Jiang et al., 2015) and many others. Now there are also kits able to enrich for CNVs: one example is OneSeq Target Enrichment (Agilent, Santa Clara, USA), which consists of baits designed to detect CNVs and Loss of Heterozygosity (LOH) genome-wide down to 1 Mb and 10 Mb resolution, respectively. In addition, OneSeq includes user-defined baits for any Agilent exome, gene or custom panel or custom regions for Single Nucleotide Variants and Indels calling.

Furthermore, during a WES experiment, secondary or incidental findings (IFs) could be detected. They can be defined as pathogenic or likely pathogenic alterations in genes that are not apparently relevant to the diagnostic indication for which the sequencing test was ordered. Different guidelines about incidental findings are followed in different countries. In the United States, for example, patients have always to be advised before the test that secondary findings may be detected and laboratories should report on incidental findings detected in a minimum set of 56 genes, selected by

the American College of Medical Genetics and Genomics (ACMG) (Green et al., 2013).

1.5.4 Whole genome sequencing (WGS)

Whole genome sequencing allows the most continuous sequence coverage and it permits to identify sequence variants throughout the genome. Because of the complexity and the greater cost of WGS, WES is currently the most used approach, even though WGS has a better diagnostic yield based on overall variant calling sensitivity and efficiency (lower coverage depth required for similar sensitivity), lack of bias and uniformity of coverage, features that have a great importance in a clinical setting where reliability and reproducibility of results are crucial (Lelieveld et al., 2015). Furthermore, a higher coverage increases the sensitivity for detecting copy number variants (Medvedev et al., 2010; Szatkiewicz et al., 2013) and the lack of allele biases improves the detection of somatic variations (Lelieveld et al., 2015).

WGS is also slightly but significantly more powerful than WES for detecting variants in the regions covered by the exome kit, particularly for SNVs. WGS is prone to sequence artefacts too and Sanger sequencing of candidate variants is always recommended (Jamuar and Tan, 2015). In addition, WGS is certainly more appropriate for detecting CNVs because it covers all breakpoints and detects variations in RNA- and protein-coding exome regions not covered by the exome (Belkadi et al., 2015).

However, the costs (WGS currently costs two to three times as much as WES, but most of the costs of WGS are directly related to sequencing whereas WES costs are mainly due to the capture kit; Belkadi et al., 2015) and analysis time (4 million variants are identified) still seriously limit implementation in routine diagnostics (Sun et al., 2015), as the problems related to incidental findings and data storage.

To date, intronic, intergenic and regulatory sequence variants are difficult or impossible to interpret, but in the future they will add a superior value to WGS data and probably *in silico* WGS-based gene panels will be used in routine diagnostics.

1.6 NGS approaches and their diagnostic rate

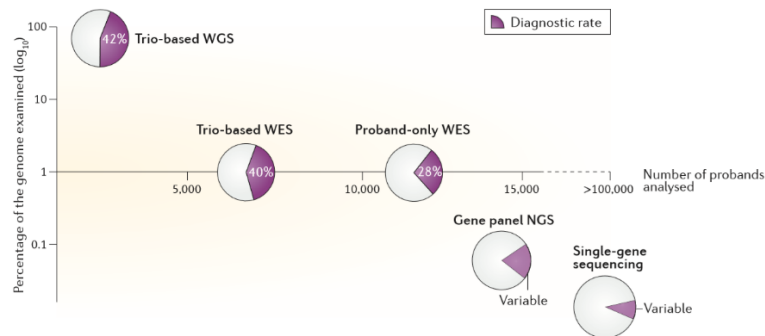


Figure 11. Balance between the diagnostic potential of a sequencing strategy and its feasibility and cost (Wright et al., 2018).

There is a balance between the diagnostic potential of a sequencing strategy and its feasibility and cost (Figure 11). *Trio*-based WGS is the approach with the highest diagnostic yield, but it is also the most demanding and expensive one regarding the informatics approach. Since 85% of the disease-causing mutations is located in coding and functional regions of the genome, the use of WES approach only slightly lowers the diagnostic yield (for example in severe intellectual disability from ~42% for WGS to ~40% for WES), but hugely reduces the cost. Moreover, although moving from a family *trio*-based approach to a proband-only approach reduces costs and practical problems, it substantially reduces also the diagnostic yield (to ~28%), because *de novo* status or phase can not be directly assigned to determine from which parent the variants were inherited. Gene panels and single genes sequencing are the most common approaches, but the rate of diagnosis varies considerably depending on the patient's phenotype (Wright et al., 2018).

Testing a single gene or a small number of genes may be preferable when the disease is phenotypically and/or genetically homogeneous; for phenotypically and/or genetically heterogeneous conditions, many hundreds of genes may need to be tested through NGS technologies. Genetic heterogeneity, in fact, increases as phenotypic specificity decreases: the less specific the phenotype associated with a disease is, the more likely it is to be caused by variants in many different genes (Wright et al., 2018; Figure 12).

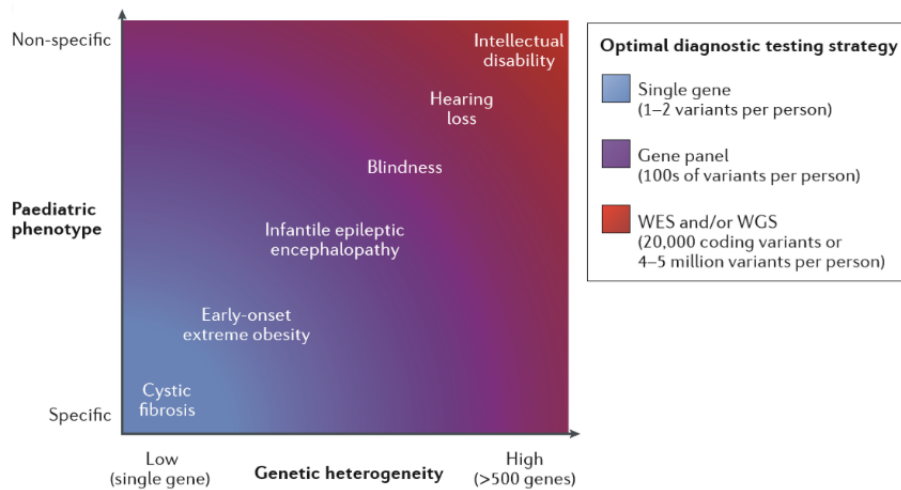


Figure 12. Genetic heterogeneity increases as phenotypic specificity decreases (Wright et al., 2018).

1.7 NGS guidelines

1.7.1 The American College of Medical Genetics and Genomics (ACMG) guidelines

ACMG recommends that the terms “mutation” and “polymorphism”, which have been used widely leading to confusion because of the incorrect assumption of their pathogenic and benign effect respectively, should be replaced by the term “variant” with the following modifiers:

- pathogenic (class V): the sequence variation has been previously reported and recognized as causative of the disorder;
- likely pathogenic (class IV): the sequence variation has not been previously reported, but it is inside a known disease gene;
- uncertain significance (VUS; class III): the sequence variation is unknown or expected to be related to a clinical presentation;
- likely benign (class II): the sequence variation has not been previously reported, but it is probably not causative of the pathology;
- benign (class I): the sequence variation has been already reported and documented as neutral variant (Di Resta et al., 2018).

For a given variant, the user has to select criteria based on the evidence observed; then the criteria are combined according to some scoring rules that allow to classify the variant (Table 2). When a variant does not fulfill criteria or the evidence for benign and pathogenic is conflicting, it defaults to uncertain significance (Richards et al., 2015).

	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			

Table 2. The criteria that allow to classify the variants are organized by the type of evidence as well as their strength for a benign (left side) or a pathogenic (right side) assertion (Richards et al., 2015).

When a laboratory finds a variant in a gene without a validated association to the patient's phenotype, it is a gene of uncertain significance (GUS). It can occur when a gene has never been associated with any patient phenotype or when the gene has been associated with a different phenotype from that under consideration. Additional evidence would be required to support the association of the gene to the disease (Richards et al., 2015).

The standard gene variant nomenclature maintained and versioned by the Human Genome Variation Society (HGVS) should be used. Laboratories should note the version used in their test methods; clinical reports should include sequence references to ensure unambiguous naming of the variant at the DNA, RNA and protein levels. Only specific exceptions to the HGVS nomenclature are supported (Richards et al., 2015).

ACMG recommends also to clinicians to report incidental findings in some relevant genes associated with a set of disorders, on the bases of clinical validity and utility. No technology can be

used to measure the size of tandem repeats accurately: for this reason, some disorders are not included. The laboratories have to actively look for the specified kinds of variant in the genes listed in the recommendations. Clinicians have to contextualize any incidental finding for the patient in light of personal and family history, physical examination and other relevant findings. The variants that have to be reported are those fitting two categories: “sequence variation is previously reported and is a recognized cause of the disorder” (Known Pathogenic, KP), “sequence variation is previously unreported and is of the type which is expected to cause the disorder” (Expected Pathogenic, EP). Laboratories should not ensure a depth of coverage for these genes equivalent to molecular testing for primary indication (Green et al., 2013).

Incidental variants should be reported for any clinical sequencing conducted on a constitutional (but not tumor) tissue and incidental variants should be reported regardless of the age of the patient (Green et al., 2013).

The clinician has to provide a comprehensive pre- and post-test counselling to the patient (Green et al., 2013). Whenever clinical sequencing is ordered, the clinician should discuss with the patient the possibility of incidental findings and the laboratory should report incidental findings in the genes listed in the recommendations without reference to patient’s preferences. However, the patient has the right to decline clinical sequencing if he judges the risks of eventual incidental findings to outweigh the benefits of the test (Green et al., 2013).

1.7.2 The Italian Society of Human Genetics (SIGU) guidelines

Patient’s phenotypic characterization is crucial for the choice of the molecular test and for the analysis of the identified variants. The Italian Society of Human Genetics (SIGU) recommends that different approaches should be applied in different situations (Documento Commissione SIGU-NGS, 2016; www.sigu.net):

- in case of phenotype characterized by a low genetic heterogeneity and caused by small genes, it is better to use conventional molecular methods;
- in case of phenotype characterized by a high genetic heterogeneity or caused by very long genes, it is suggested to use NGS platforms, in order to decrease time and costs;
- in case of phenotype characterized by a high genetic heterogeneity, but caused in most cases by mutations in the same genes, targeted sequencing (gene panels) is suggested;

- targeted sequencing is also recommended in case of mosaic mutations;
- in case of phenotype characterized by an increasing genetic heterogeneity, it is suggested to perform WES and limit the analysis to genes associated with phenotype (*in silico* gene panels);
- when a diagnostic hypothesis can not be formulated it is suggested to perform WES, looking for known or new genes.

Genetic counselling is always recommended before and after the test (Documento Commissione SIGU-NGS, 2016; www.sigu.net):

- before the test, the proband has to subscribe an informed consent, which explains exhaustively what kind of results he can obtain from the test (IFs, VUS, information about parents) and allows him to decide which results he agrees to know. IFs can be: deleterious variants, which have an immediate clinical utility; known or presumed-deleterious variants, that, despite being reliably associated with a disease or relevant trait, are not medically actionable; variants which have no known medical relevance and do not have a clinical utility (Berg et al., 2011). The proband has to be supported in the decision to whether receive or not information about the variants of the first two categories; variants of the third category should not be communicated;
- a post-test counselling is necessary for the communication of the results and eventually of the IFs. Only those incidental findings decided by the patient should be communicated. However, sometimes the patient's right to decide on his own could not coincide with the principles of medical deontology. It can happen when IFs are related to diseases for which it is possible to realize therapeutic or preventive measures or in the case of a disease involving also relatives. At this stage, the clinician can also propose further investigations, if it deemed necessary;
- in the case of oncological test, the possibility to find variants associated with cancer predisposition has to be discussed during the counselling and a specific informed consent is used.

Quality requirements are specified (Documento Commissione SIGU-NGS, 2016; www.sigu.net):

- it is important to use a standardized terminology to describe the patient's phenotype (HPO terms) as it allows to share information accurately;
- it is important to standardize parameters and minimum requirements for each test within different laboratories;
- the medical report should be composed of only one page, with some technical attachments;

- if the test is an *in silico* panel, one attachment should include the analysed gene subset with the average coverage of each gene; the limits of the panel and the bioinformatic pipeline have to be specified together with any supplementary technique used;
- positive results have to be confirmed with Sanger sequencing and it is necessary to verify the coverage of the region through a specific browser, like Integrative Genomics Viewer (IGV);
- minimum parameters of coverage have to be established.

The big amount of data produced by NGS techniques causes a lot of storage problems. However, laboratories should conserve for a long time at least the most important files of the NGS workflow: the FASTQ file, which contains the reads produced by the sequencer; the BAM file, which contains the reads aligned to the reference genome; the VCF file, which contains the variants compared to the reference sequence. An important feature of NGS tests, in fact, is the temporary nature of their results, related to the development of the scientific knowledge: for this reason, they should be re-evaluated periodically (Documento Commissione SIGU-NGS, 2016; www.sigu.net).

2. AIM OF THE STUDY

The main purpose of this PhD project was to study Mendelian diseases with different Next Generation Sequencing approaches, selecting the most appropriate NGS technology and analysis workflow to investigate the molecular bases of four different genetic disorders, in order to improve their diagnosis and prognosis and to support genetic counselling.

3. MATERIALS AND METHODS

3.1 Subjects selection

In this work we enrolled at Umberto I General Hospital and Sapienza University of Rome four different families in which a phenotype with a supposed genetic cause was recurrent, in order to identify the causative gene/genes with the most appropriate NGS technology and data analysis approach. Informed consents for DNA storage and genetic analyses were obtained for each subject or his parents; permission to publish photographs was given for all subjects reported in this work.

Family A: the index patient (III:6) was a 16-year-old girl from a small town in central Italy, who was referred to the Department of Oral and Maxillo Facial Sciences of Sapienza University of Rome. The patient showed a bilateral absence of permanent maxillary canines and anamnestic analysis suggested the presence of several family members also affected by canine anomalies. Clinical standardized assessment, including panoramic radiographs, oral photographs and anamnestic data, was performed on fourteen members of the family by a trained orthodontist. Exome enrichment and massively parallel sequencing were performed on the genomic DNA of subjects III:1, III:4 and III:6 (pedigree in Figure 13).

Family B: the proband (III:2) was a 5-year-old boy affected by an isolated form of brachydactyly with features of type A1 (OMIM #112500) and type C (OMIM #113100), as his maternal grandfather (I:1), while his mother (II:2) showed a very mild phenotype. The proband was also referred to the medical geneticist because of his short stature. Detailed information on pedigree, anamnesis, clinical assessment and radiographs were collected for all subjects. Exome enrichment and massively parallel sequencing were performed on the genomic DNA of subjects I:1 and III:2 (pedigree in Figure 21).

Family C: the proband (II:3) was a 4-year-old girl affected by corpus callosum hypoplasia, daughter of a healthy Italian mother and a healthy Chinese father. The mother underwent amniocentesis and CGH-array during prenatal period, with negative results. The proband had a healthy sister (II:1). We analysed also the fetal DNA from a previously interrupted pregnancy: the fetus (II:2) was a female and showed corpus callosum agenesis and other severe malformations. Before the voluntary interruption of pregnancy, a CGH-array was performed identifying a *de novo*

microduplication 17q12 (31,635,490-33,323,002) x3 of uncertain significance.

Clinical exome sequencing was performed on the genomic DNA of the *trio*, I:1, I:2 and II:3 (pedigree in Figure 30).

Family D: the proband (II:2) was a fetus with corpus callosum agenesis and other severe malformations. SNP-array was performed before the voluntary interruption of pregnancy and revealed a degree of homozygosity of 1% in the fetus, excluding parental consanguinity. A previous pregnancy was interrupted because of a male fetus (II:1) with the Dandy-Walker syndrome (OMIM %220200) and hydrocephalous. Exome enrichment and massively parallel sequencing were performed on the genomic DNA of II:2 (pedigree in Figure 36).

3.2 DNA extraction

The DNA was extracted from circulating leukocytes, saliva, buccal mucosa cells or hair bulbs using the Genra Puregene Blood Kit (Qiagen, Hilden, Germany).

To extract DNA from circulating leukocytes the protocol “DNA purification from whole blood or bone marrow using the Genra Puregene Blood Kit” was used. Erythrocytes were lysed and discarded; then, leukocytes were lysed with an anionic detergent in the presence of a DNA stabilizer, to limit the activity of DNases. RNA was then removed by treatment with a RNA digesting enzyme. Other contaminants, such as proteins, were removed by salt precipitation. The genomic DNA was then recovered by precipitation with isopropanol and the pellet washed using 70% ethanol. Finally, the genomic DNA was dissolved in hydration solution (1 mM EDTA, 10 mM Tris·Cl pH 7.5).

To extract DNA from saliva the protocol “DNA purification from body fluid using the Genra Puregene Blood Kit” was used. The cells were lysed with an anionic detergent in the presence of a DNA stabilizer, to limit the activity of DNases. Puregene Proteinase K was added in order to digest contaminating proteins. RNA was then removed by treatment with a RNA digesting enzyme. Other contaminants, such as proteins, were removed by salt precipitation. The genomic DNA was then recovered by precipitation with isopropanol and the pellet washed using 70% ethanol. Finally, the genomic DNA was dissolved in hydration solution (1 mM EDTA, 10 mM Tris·Cl pH 7.5).

To extract DNA from buccal mucosa cells collected through a buccal brush the protocol “DNA

purification from a buccal brush using the Gentra Puregene Blood Kit” was used. The buccal brush was placed inside an anionic detergent in the presence of a DNA stabilizer, to limit the activity of DNases. Puregene Proteinase K was added in order to digest contaminating proteins. The brush was removed and the solution was centrifugated to recover as much liquid as possible. RNA was then removed by treatment with a RNA digesting enzyme. Other contaminants, such as proteins, were removed by salt precipitation. The genomic DNA was then recovered by precipitation with isopropanol and the pellet washed using 70% ethanol. Finally, the genomic DNA was dissolved in hydration solution (1 mM EDTA, 10 mM Tris·Cl pH 7.5).

To extract DNA from hair, the hair bulbs were cut and dissolved in an anionic detergent in the presence of a DNA stabilizer, to limit the activity of DNases. Puregene Proteinase K was added in order to digest contaminating proteins. RNA was then removed by treatment with a RNA digesting enzyme. Other contaminants, such as proteins, were removed by salt precipitation. The genomic DNA was then recovered by precipitation with isopropanol and the pellet washed using 70% ethanol. Finally, the genomic DNA was dissolved in hydration solution (1 mM EDTA, 10 mM Tris·Cl pH 7.5).

To accurately assess sample quantity and quality, the extracted DNA was quantified through the NanoDrop spectrophotometer. A 260/280 ratio of ~ 1.8 is generally accepted as “pure” for DNA; if the ratio is appreciably lower, it may indicate the presence of protein, phenol or other contaminants that absorb strongly at or near 280 nm. The 260/230 value for a “pure” DNA is often higher than the respective 260/280 value and is commonly in the range of 1.8 – 2.2; if the ratio is appreciably lower, it may indicate the presence of contaminants which absorb at 230 nm, as EDTA, carbohydrates and phenol.

3.3 NGS techniques: whole exome and clinical exome sequencing

The extracted genomic DNA was processed for the sequencing experiment through different steps:

- **library preparation:** the genomic DNA was enzymatically or physically fragmented and an *in vitro* shotgun library was constructed through the ligation with adaptors to the ends of the fragments and amplified;
- **targeted enrichment:** targeted enrichment was performed by an hybridization capture approach: the fragments were hybridized to biotinylated baits in the presence of blocking oligonucleotides that were complementary to the adaptors; the hybridized fragments were recovered by biotin-streptavidin-based pulldown;

- **indexing and pooling:** the targeted enrichment was followed by an adapter ligation with specific barcodes; all the samples were pooled together and the libraries were sequenced in parallel.

The whole exome or the clinical exome of the selected subjects were sequenced using different technologies and kits for the targeted enrichment:

Family	Sequenced subject	NGS approach	NGS platform	Exome capture kit
Family A	III:1	Whole Exome Sequencing	HiSeq 2000 (Illumina)	SeqCap EZ Human Exome Kit V.3.0 (Roche, Basel, Switzerland)
Family A	III:4	Whole Exome Sequencing	HiSeq 2000 (Illumina)	SeqCap EZ Human Exome Kit V.3.0 (Roche, Basel, Switzerland)
Family A	III:6	Whole Exome Sequencing	Complete Genomics	Complete Genomics (Mountain View, CA, USA)
Family B	I:1	Whole Exome Sequencing	NextSeq (Illumina)	SureSelect Clinical Research Exome (Agilent, Santa Clara, CA, USA)
Family B	III:2	Whole Exome Sequencing	NextSeq (Illumina)	SureSelect Clinical Research Exome (Agilent, Santa Clara, CA, USA)
Family C	I:1	Clinical Exome Sequencing	Miseq (Illumina)	Trusight one sequencing panel (Illumina, San Diego, CA, USA)
Family C	I:2	Clinical Exome Sequencing	Miseq (Illumina)	Trusight one sequencing panel (Illumina, San Diego, CA, USA)
Family C	II:3	Clinical Exome Sequencing	Miseq (Illumina)	Trusight one sequencing panel (Illumina, San Diego, CA, USA)
Family D	II:2	Whole Exome Sequencing	Ion Proton™ System (Applied Biosystems)	Ion TargetSeq™ Exome Enrichment Kit (Applied Biosystems, Foster City, CA, USA)

Table 3. NGS platforms and kits for the targeted enrichment used for whole exome or clinical exome sequencing.

The whole exome sequencing of III:1 and III:4 of the family A was carried out by BGI (Shenzen, China) and the exome of III:6 by Complete Genomics (Mountain View, CA, United States), on DNA extracted from circulating leukocytes; the whole exome sequencing of I:1 and III:2 of the family B was performed at Casa Sollievo della Sofferenza Hospital (Foggia, Italy) on DNA extracted from circulating leukocytes; the clinical exome sequencing of I:1, I:2 and II:3 of the family C was performed at the CSS-Mendel Institute (Rome, Italy) on DNA extracted from circulating leukocytes; the whole exome sequencing of II:2 of the family D was performed at CRIBI (Centro di Ricerca Interdipartimentale per le Biotecnologie Innovative) Genomics (Padua, Italy) on DNA extracted from amniotic fluid cells (technical details about sequencing in Table 3).

3.4 Data analysis

Sequencing data were processed through several steps: at first the removal of the sequences of the adaptors and the alignment of the reads to the reference genome (UCSC GRCh37/hg19) were performed through the Burrows-Wheeler Aligner (Li and Durbin, 2009), a software package for mapping low-divergent sequences against a large reference genome, in order to determine the exact position of each read on the human genome. As a following step, the duplicate reads were labelled through Picard's MarkDuplicates (<http://broadinstitute.github.io/picard>). Duplicates arose from

artefacts during PCR amplification (PCR duplicates) or resulted from a single amplification cluster, incorrectly detected as multiple clusters by the optical sensor of the sequencing instrument (optical duplicates). Because all the duplicate reads were sampled from the same DNA molecule, they gave an uneven representation of that molecule compared to the others and they biased the SNV calling. Picard's MarkDuplicates identified duplicates as reads that mapped with identical coordinates and orientations. After the removal of duplicate reads, Genome Analysis Toolkit (GATK) (McKenna et al., 2010), a collection of command-line tools for analysing high-throughput sequencing data, performed at first the base recalibration step through GATK Base Recalibrator, which assigned an error to each base. After that, GATK HaplotypeCaller performed the variant calling, which consisted in the identification of the DNA sequence variations relative to the reference genome (Single Nucleotide Variations or SNVs and small Indels). Finally, functional annotation of variants was performed using ANNOVAR (Wang K et al., 2010) and SnpEff (Cingolani et al., 2012) tools, to annotate SNVs and small Indels and to analyse their functional consequence on transcripts and proteins and their frequency in population database (1000 Genomes Project, dbSNP, ExAC, gnomAD). As a transcript reference dataset we used RefSeq. We added further information from different online resources on variants and genes, related to clinical information (ClinVar, OMIM), *in-silico* pathogenicity predictions (e.g., Eigen, CADD, DANN, PolyPhen-2, SIFT), conservation (e.g., PhyloP and PhastCons), functional descriptions of genes, gene expression and gene interaction information (e.g., Gene Ontology, KEGG pathways, tissue specific gene expression and Variation Intolerance Score).

Sequencing data were analysed depending on the NGS technology used; different versions of the tools were used:

Family	Sequenced subject	Alignment tool	Variant calling tool	Annotation tool
Family A	III:1	Burrows-Wheeler Aligner (BWA V.0.7.12)	GATK's HaplotypeCaller (V.3.5)	SnpEff (V.4.2) and dbNSFP (V.2.9)
Family A	III:4	Burrows-Wheeler Aligner (BWA V.0.7.12)	GATK's HaplotypeCaller (V.3.5)	SnpEff (V.4.2) and dbNSFP (V.2.9)
Family A	III:6	Complete Genomics software	Complete Genomics software	Complete Genomics software
Family B	I:1	Burrows-Wheeler Aligner (BWA V.0.7.12)	GATK's HaplotypeCaller (V.3.7)	ANNOVAR (July 2017 release) and dbNSFP (V.3.5a)
Family B	III:2	Burrows-Wheeler Aligner (BWA V.0.7.12)	GATK's HaplotypeCaller (V.3.7)	ANNOVAR (July 2017 release) and dbNSFP (V.3.5a)
Family C	I:1	Burrows-Wheeler Aligner (BWA V.0.7.12)	GATK's HaplotypeCaller (V.3.7)	ANNOVAR (July 2017 release) and dbNSFP (V.3.5a)
Family C	I:2	Burrows-Wheeler Aligner (BWA V.0.7.12)	GATK's HaplotypeCaller (V.3.7)	ANNOVAR (July 2017 release) and dbNSFP (V.3.5a)
Family C	II:3	Burrows-Wheeler Aligner (BWA V.0.7.12)	GATK's HaplotypeCaller (V.3.7)	ANNOVAR (July 2017 release) and dbNSFP (V.3.5a)
Family D	II:2	Torrent Suite Software (package V.5.2.1)	Torrent Suite Software (package V.5.2.1)	ANNOVAR (July 2017 release) and dbNSFP (V.3.5a)

Table 4. Tools for NGS data analysis.

3.5 Selection of candidate variants (filtering and prioritization)

At first, the variants were filtered based on quality criteria and their effect: only high-quality variants and those with an effect on the coding sequence and splice site regions were retained. Then variants were prioritized, according to the specific disease, on the basis of pedigree information and the mode of inheritance, the localization of the variant, the mutation type, the frequency of the variant, the predicted impact of the variant on protein function and structure, the functional evidences, the evolutionary conservation of variant nucleotide and the annotation in databases.

Different parameters and thresholds were used for each phenotype:

Family A: we prioritized variants using a public database (ExAC V.0.3.1) to retain novel and annotated changes with an unknown frequency or having a minor allele frequency (MAF) $\leq 5\%$ and occurring with a frequency $\leq 10\%$ in an in-house database, which includes approximately 600 exomes. Then we analysed the functional impact of variants by Combined Annotation Dependent Depletion (CADD) (V.1.3), a tool for scoring the deleteriousness of DNA variants, using as threshold a value of 10 (Kircher et al., 2014). We drew up a list of known genes for isolated and syndromic phenotypes characterized by hypodontia and/or related dental anomalies. We used search terms such as “hypodontia”, “primary failure of tooth eruption”, “selective tooth agenesis”, “oligodontia”, “anodontia” and “agenesis of permanent teeth” to retrieve information from literature (PubMed), mutation database (HGMD-Human Genome Mutation Database, <http://www.hgmd.cf.ac.uk/ac/>) and phenotype databases (OMIM-Online Mendelian Inheritance in Man, <https://www.omim.org>; HPO-Human Phenotype Ontology, <http://human-phenotype-ontology.github.io>). Then, we analysed the WES data in order to identify and prioritize variants segregating according to different inheritance patterns, and matching at least one of the following criteria: known causative variants, variants in known genes, variants in genes functionally related to teeth development and variants predicted deleterious using CADD scoring system. Finally, we analysed the potentially causative variants in terms of gene function, gene expression, animal models, and phenotype, retrieving information from several databases, *i.e.* OMIM-Online Mendelian Inheritance in Man (<https://www.omim.org>), HPO-Human Phenotype Ontology (<http://human-phenotype-ontology.github.io>), MGI-Mouse Genome Informatics (<http://www.informatics.jax.org>), ZFIN-Zebrafish Information Network (<https://zfin.org>), and literature (PubMed).

Family B: we prioritized the variants to retain only those with a frequency $\leq 4\%$ in gnomAD

database (V.2.0) and with a CADD (V.1.3) score ≥ 10 (Kircher et al., 2014), predicted to have a high functional impact on the protein. We drew up a list of known genes for isolated and syndromic brachydactylies; then, we prioritized the variants hypothesising an X-linked transmission or an autosomal dominant transmission with variable expressivity and matching at least one of the following criteria: known causative variants, variants in known genes and variants predicted deleterious using CADD scoring system.

Family C: we prioritized the variants to retain only those with a frequency $\leq 3\%$ in gnomAD database (V.2.0) and those with a CADD score (V.1.3) ≥ 10 (Kircher et al., 2014). We drew up a list of known genes for corpus callosum dysgenesis; then we prioritized the variants, looking for a *de novo* heterozygous variant or homozygous/compound heterozygous variants and variants matching at least one of the following criteria: known causative variants, variants in known genes, variants in genes functionally related to brain development and variants predicted deleterious using CADD scoring system. We further analysed potentially causative variants on the basis of different parameters, as gene function and gene expression.

Family D: we prioritized the variants to retain only those with a frequency $\leq 3\%$ in gnomAD database (V.2.0) and those with a CADD score (V.1.3) ≥ 10 (Kircher et al., 2014). We drew up a list of known genes for corpus callosum dysgenesis; then, we prioritized the variants, looking for an autosomal recessive or a X-linked variant and variants matching at least one of the following criteria: known causative variants, variants in known genes, variants in genes functionally related to brain development and variants predicted deleterious using CADD scoring system. We further analysed potentially causative variants on the basis of different parameters, as gene function and gene expression.

3.6 Variants validation

Sanger sequencing was used to validate selected candidate variants and to perform segregation analyses.

Variants were PCR-amplified by using GoTaq G2 Flexi DNA polymerase (Promega, Madison, WI, USA) and custom primers (Tables 5, 6 and 7).

Family	Gene	Forward primer	Reverse primer	Annealing temperature (C°)
Family A	<i>EDARADD</i>	5'-TAAAGGATGTAAGAAATGAATGC-3'	5'-AGGAAGCACAATTCGTCATAG-3'	59
Family A	<i>COL5A1</i>	5'-ACCCTCTGAGGCTGCGTG-3'	5'-TCTGGGATCAAGACAGCTGC-3'	59
Family A	<i>RSPO4</i>	5'-GCTACCAAGCCTGGAGCTAC-3'	5'-TGCACTCTGCTGATCCCTTG-3'	60
Family A	<i>T</i>	5'-AGACTAATGGATTGTTTTCTACAG-3'	5'-AGAATCTAAATAAAGACCAACC-3'	60
Family A	<i>NELLI</i>	5'-TGTATCACTGCAGGCTATTAAC-3'	5'-TGAGAGAGGTAATAGGATATGC-3'	60
Family B	<i>GDF5</i>	5'-CTGGATACGAGAGCATTCCAC-3'	5'-CTCCCTTGGCCCTGGCATTG-3'	62
Family C	<i>ARX</i>	5'-GCTGCTGGAGGACGAAGAAG-3'	5'-CAGCTGGTAGCTGGTGAACG-3'	62
Family D	<i>FKTN</i>	5'-TGCCACAGAAAGGTTCTAGC-3'	5'-AGGAAATCAATAGATCCTTGCTG-3'	62

Table 5. Primers used to amplify selected variants.

Quantity/concentration	
DNA	50-100 ng
Gotaq G2 Flexi DNA polymerase (5U/μl)	0.65 U
Gotaq Flexi Buffer 5X	1X
MgCl ₂ solution 25 mM	1.5 mM
dNTPs 2.5 mM	0.13 mM
Forward primer 25 μM	0.6 μM
Reverse primer 25 μM	0.6 μM
ddH ₂ O	up to final volume of 25 μl

Table 6. PCR protocol with GoTaq G2 Flexi DNA polymerase.

Temperature	Time	Cycle
95°C	2'	1
95°C	30"	30
60°C	30"	30
72°C	40"	30
72°C	5'	1
4°C	5'	1

Table 7. PCR cycling parameters.

The amplicons were checked through 2% agarose gel electrophoresis and purified using MSB Spin PCRapace (Stratec Molecular, Berlin, Germany).

Sanger sequencing was performed by using the ABI BigDye Terminator Sequencing Kit (V.3.1) (Applied Biosystems, Foster City, CA, USA) as per the manufacturer's protocol (Tables 8 and 9).

Quantity/concentration	
DNA	15 ng / 100 bp of purified PCR product
Big dye 3.1 Master Mix 2.5X	0.16X
Buffer Big dye Terminator 5X	0.5X
Forward/reverse primer 2.5 μM	0.16X
ddH ₂ O	up to final volume of 20 μl

Table 8. Sanger sequencing protocol with Big Dye 3.1.

Temperature	Time	Cycle
96°C	1'	1
96°C	15"	25
54°C	5"	25
60°C	4'	25
4°C	5'	1

Table 9. Sequencing reaction cycling parameters.

Dye removal from cycle-sequencing reactions was carried on with MSB Spin PCRapace (Strattec Molecular, Berlin, Germany). Sanger sequencing was performed with automated capillary sequencers: the ABI Prism 3500 Genetic Analyzers (Applied Biosystems, Foster City, CA, USA) or the 3130/3130xl Genetic Analyzers (Applied Biosystems, Foster City, CA, USA).

Sequence electropherograms were analysed by using ChromasPro (V.1.7.5; Technelysium Pty Ltd, Brisbane, Australia).

3.7 Modeling of the nucleotidyltransferase domain of FKTN (family D)

The *in silico* modeling of the nucleotidyltransferase domain of human FKTN (hFKTN) protein (NP_001073270.1) was made through the Phyre software, in collaboration with Professor Alessandro Paiardini, Sapienza University of Rome.

4. RESULTS

4.1 Family A

The family was composed of three generations (Figure 13): the index patient (III:6) showed a bilateral absence of permanent maxillary canines; the other members of the family showed different maxillary canine anomalies, including canine agenesis, either monolateral or bilateral, canine impaction and canine ectopic eruption, phenotypes that seem to be different manifestations of the same disorder (Figure 14).

We performed WES of three cousins of the third generation, one for each branch of the family.

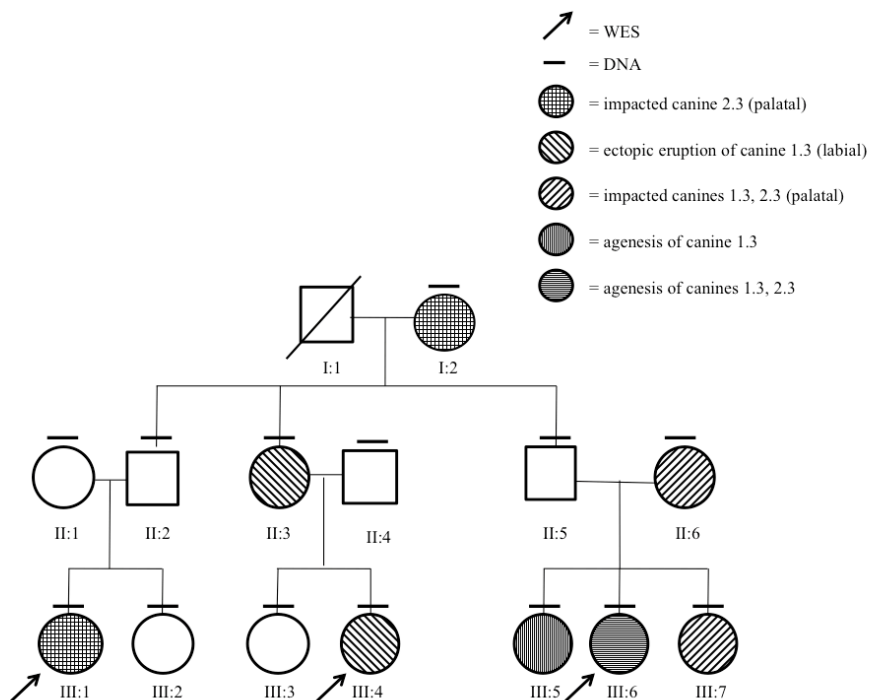


Figure 13. Pedigree of the family with canine anomalies; black lines indicate individuals for whom DNA was available for the molecular analyses; the arrows indicate individuals who underwent whole exome sequencing (WES).

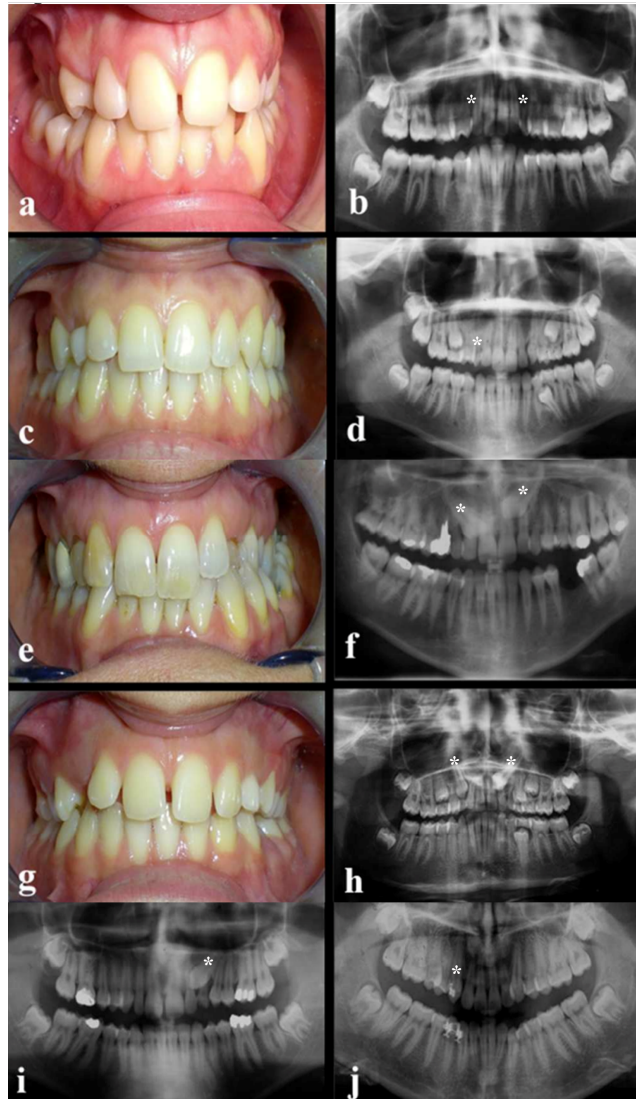


Figure 14. Clinical photographs and panoramic radiographs of dentitions of six affected family members. Subject III:6 (a-b): the index patient shows congenital absence of 1.3 and 2.3 and the persistence of the left upper deciduous canine. Panoramic radiograph shows the agenesis of permanent maxillary canines and root resorption of left deciduous canine. All third molars are present. Subject III:5 (c-d): agenesis of 1.3. Panoramic radiograph shows the persistence of the right upper deciduous canine with root resorption; 3.5 anomalous radicular distal tip is observed. Subject II:6 (e-f): palatal bilateral maxillary impacted canines and persistence of the upper deciduous canines are observed. Subject III:7 (g-h): palatal bilateral maxillary impacted canines and persistence of the upper deciduous canines are observed. Subject III:1 (i): impacted canine 2.3; persistence of the left upper deciduous canine and 1.2 microdontic lateral incisor are shown. Subject III:4 (j): ectopic eruption 1.3. Panoramic radiograph shows the orthodontic treatment. Mild crowding of maxillary arch can also be observed. The asterisks indicate the missing, impacted or ectopically erupting permanent maxillary canines (Barbato et al., 2018).

We obtained the following whole exome sequencing data output:

Subject	Coverage	Depth
III:1	86%	50x
III:4	86%	49x
III:6	91%	165x

Table 10. Whole exome sequencing data output.

We assumed that the phenotype in this family fit an autosomal dominant segregation model with incomplete penetrance in males and variable expressivity in the first two branches; we supposed a more complex segregation pattern in the third branch, due to a possible contribution of both paternal and maternal origin.

After bioinformatic analysis of WES data, we obtained 64,852, 66,389 and 103,815 total variants in III:1, III:4 and III:6, respectively.

Then, we filtered the exome variants using several criteria (see Material and Methods section) and we obtained: 13,144, 13,213 and 12,494 non synonymous/frameshift/splicing (-8/+3) variants in III:1, III:4 and III:6, respectively; 1,735, 1,708 and 2,345 variants with unknown ExAC frequency or MAF \leq 5% in III:1, III:4 and III:6, respectively; 1,119, 1,071 and 2,061 variants with a frequency \leq 10% (58/587) in the in-house database in III:1, III:4 and III:6, respectively; 648, 680 and 1,234 variants with combined annotation dependent depletion (CADD) score \geq 10 in III:1, III:4 and III:6, respectively.

Variants, either shared or not by the three affected cousins, were filtered and prioritized. To this aim, as a first step, we created a list of 96 genes involved in dental anomalies, using information from several databases (OMIM-Online Mendelian Inheritance in Man, HGMD-Human Genome Mutation Database, HPO-Human Phenotype Ontology, PubMed) and we analysed single WES data looking for candidate variants in genes previously associated with dental anomalies; in a second step of the analysis, we prioritized candidate genes on the basis of their involvement in teeth development and the sharing among the three cousins.

Using the previous described criteria we selected the following variants:

Gene	Chr	Genomic position	Genbank accession number	Nucleotide substitution	Aminoacid substitution	Variant ID	ExAC frequency	in-house database frequency	CADD score
<i>EDARADD</i>	1	236645609	NM_145861.2	c.308C > T	p.Ser103Phe	rs114632254	2.1%	8.5%	27.8
<i>COL5A1</i>	9	137642654	NM_000093.4	c.1588G > A	p.Gly530Ser	rs61735045	3.6%	8.7%	24.9
<i>RSPO4</i>	20	947909	NM_001029871.3	c.317G > A	p.Arg106Gln	rs6140807	0.9%	2.2%	22.8
<i>T</i>	6	166574346	NM_003181.3	c.1013C > T	p.Ala338Val	rs117097130	0.5%	0.8%	20.1
<i>NELI</i>	11	20968970	NM_001288713.1	c.1244G > A	p.Arg415His	rs141323787	0.4%	0.6%	26.2

Table 11. Candidate variants identified through WES approach.

The first step of the analysis led to the identification in subject III:6 (bilateral canine agenesis) of two missense variants in *EDARADD* and *COL5A1*, previously associated with tooth agenesis and a syndromic phenotype including dental anomalies, respectively.

The *EDARADD* variant (NM_145861.2: c.308C>T; NP_665860.2: p.Ser103Phe; rs114632254) was found also in her sister (III:5; monolateral canine agenesis) and her mother (II:6; bilateral canine maxillary inclusion) (Figures 15 and 20).

EDARADD
c.308C>T; p.Ser103Phe

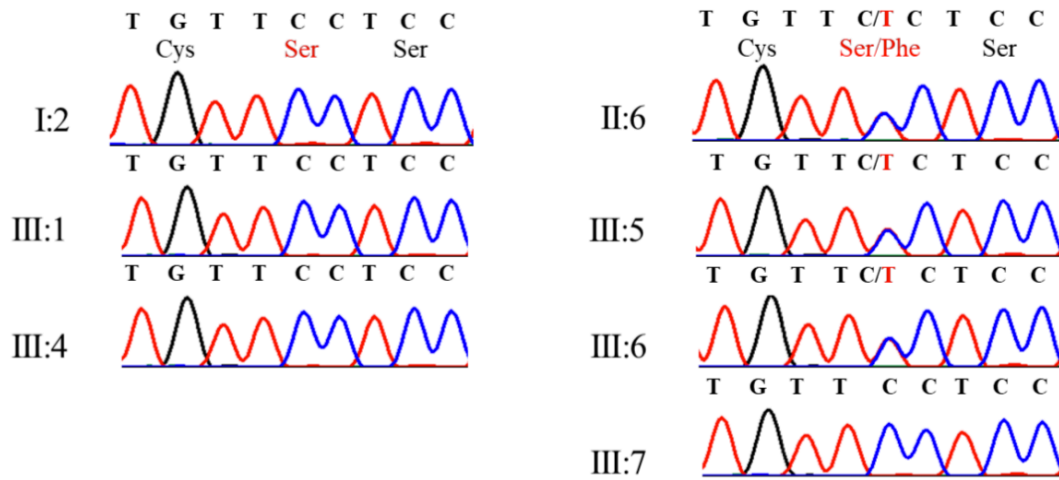


Figure 15. Electropherograms showing genotypes of patients and unaffected individuals for *EDARADD* variant (Barbato et al., 2018).

The *COL5A1* variant (NM_000093.4: c.1588G>A; NP_000084.3: p.Gly530Ser; rs61735045) was found in subjects III:5 (monolateral canine agenesis), III:6 (bilateral canine agenesis) and III:2 (unaffected); it segregated from the paternal grandmother (I:2; monolateral upper left palatal impacted canine) (Figure 16 and 20).

COL5A1
c.1588G>A; p.Gly530Ser

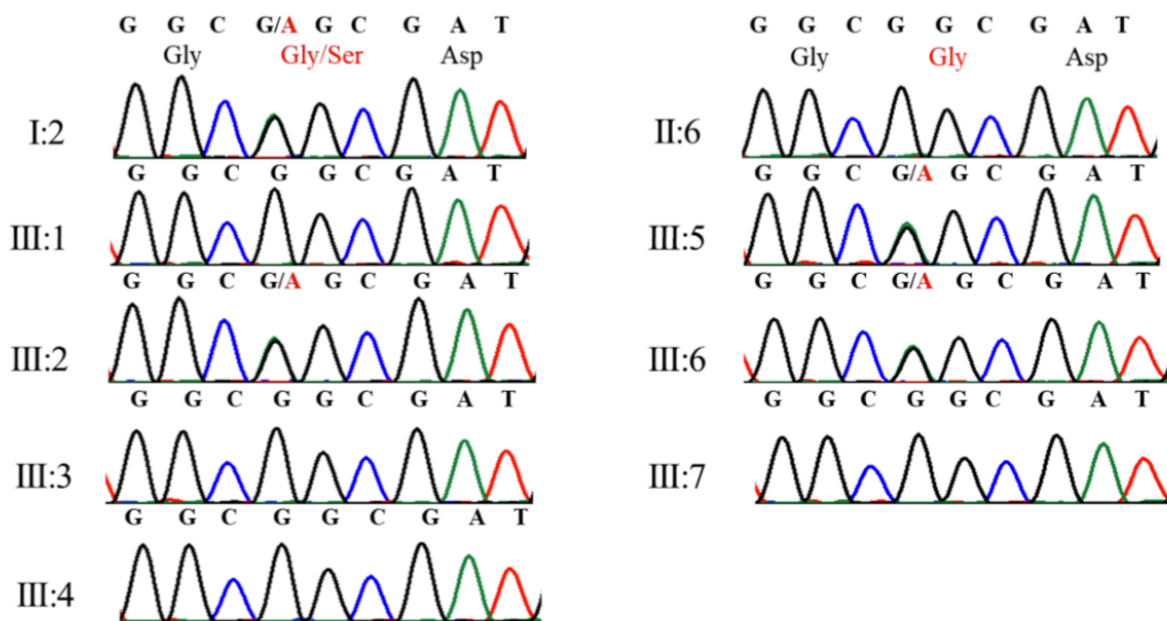


Figure 16. Electropherograms showing genotypes of patients and unaffected individuals for *COL5A1* variant (Barbato et al., 2018).

The second step of the analysis didn't lead to the identification of interesting variants in genes functionally related to teeth development, shared by subjects III:1, III:4 and III:6. We therefore focused on variants shared by subjects III:1 (monolateral upper left palatal impacted maxillary canine and right lateral incisor microdontia) and III:4 (monolateral upper right ectopic labial eruption of maxillary canine) and we found three missense variants in *RSPO4* (NM_001029871.3: c.317G>A; NP_001025042.2: p.Arg106Gln; rs6140807), *T* (NM_003181.3: c.1013C>T; NP_003172.1: p.Ala338Val; rs117097130) and *NELL1* (NM_001288713.1: c.1244G>A; NP_001275642.1: p.Arg415His; rs141323787) genes.

The *RSPO4* variant was found in subjects III:1 (monolateral upper left palatal impacted canine), III:4 (monolateral upper right ectopic labial eruption of canine), I:2 (monolateral upper left palatal impacted canine) and III:3 (unaffected) (Figures 17 and 20).

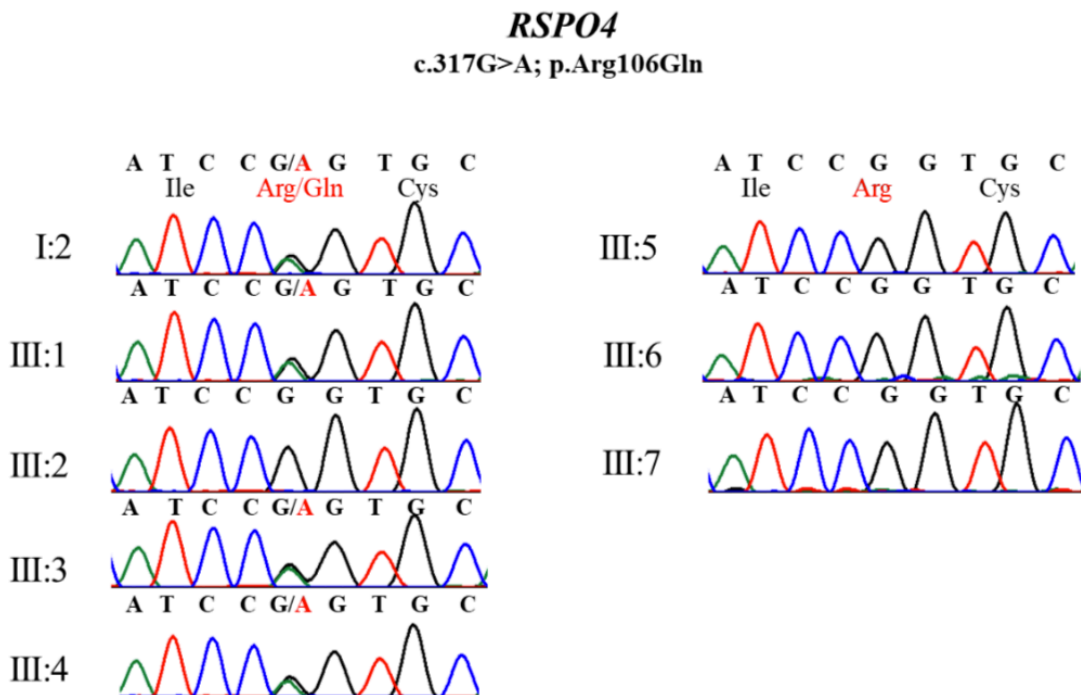


Figure 17. Electropherograms showing genotypes of patients and unaffected individuals for *RSPO4* variant (Barbato et al., 2018).

The *T* variant was found in subjects III:1 (monolateral upper left palatal impacted canine), III:4 (monolateral upper right ectopic labial eruption of canine), I:2 (monolateral upper left palatal impacted canine), III:2 (unaffected) and III:3 (unaffected) (Figures 18 and 20).

T
c.1013C>T; p.Ala338Val

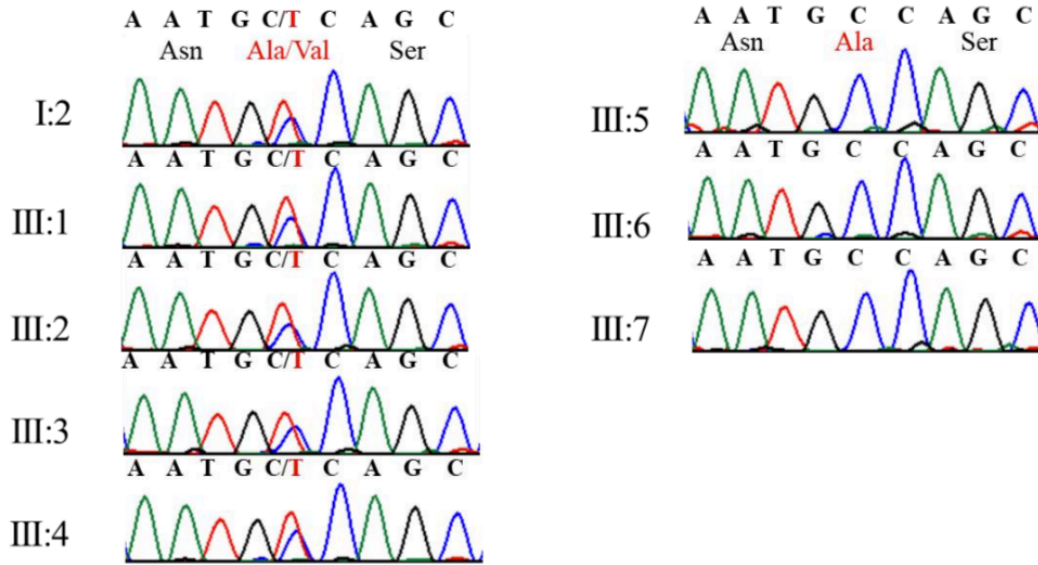


Figure 18. Electropherograms showing genotypes of patients and unaffected individuals for *T* variant (Barbato et al., 2018).

The *NELLI* variant was found in subjects III:1 (monolateral upper left palatal impacted canine), III:4 (monolateral upper right ectopic labial eruption of canine), I:2 (monolateral upper left palatal impacted canine), III:2 (unaffected) and III:3 (unaffected) (Figures 19 and 20).

NELLI
c.1244G>A; p.Arg415His

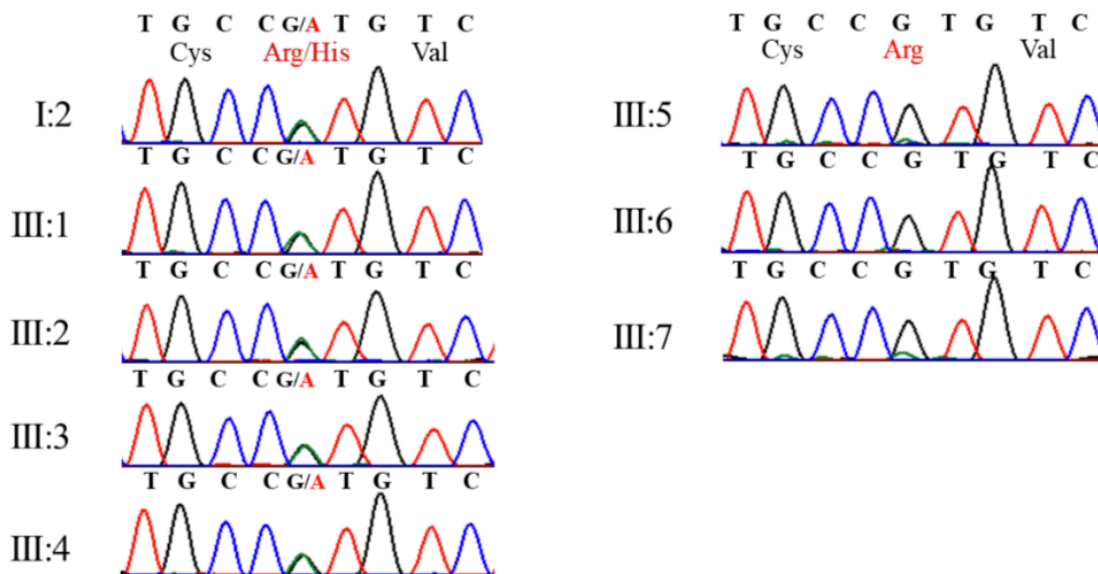


Figure 19. Electropherograms showing genotypes of patients and unaffected individuals for *NELLI* variant (Barbato et al., 2018).

Segregation analysis of all the analysed variants was summarized in the following genealogic tree:

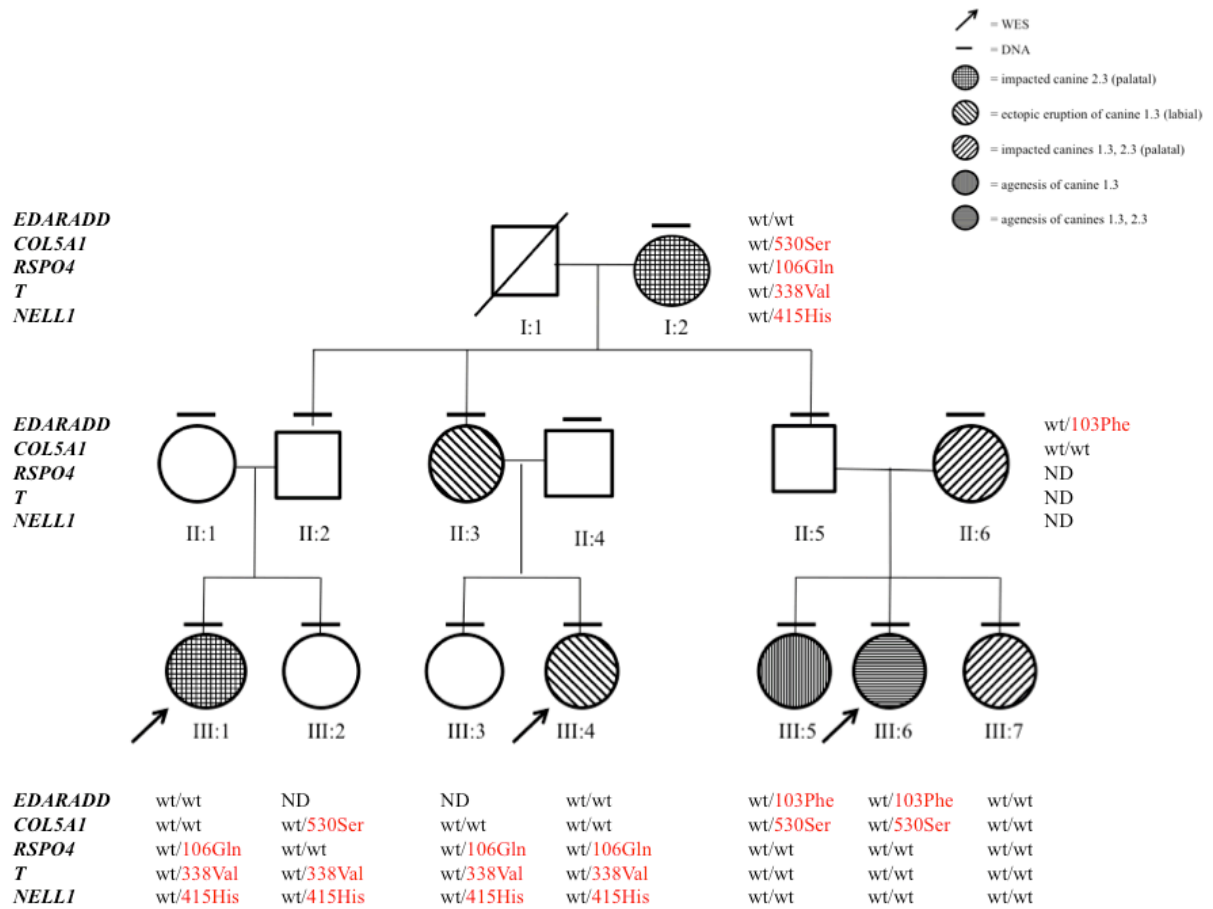


Figure 20. Variants identified in *EDARADD*, *COL5A1*, *RSPO4*, *T* and *NELL1* genes in analysed family members; aminoacidic substitutions are reported for all tested variants; “wt” indicates wild type allele; “ND” indicates genotypes that have not been experimentally determined.

4.2 Family B

The family was composed of three generations (Figure 21): a proband (III:2) affected by an isolated form of brachydactyly with features of type A1 (OMIM #112500) and type C (OMIM #113100) (Figures 22 and 23), as his maternal grandfather (Figures 24 and 25), and his mother with a very mild phenotype of the hands (Figures 26 and 27).

We analysed the pedigree and performed WES of the proband and the grandfather.

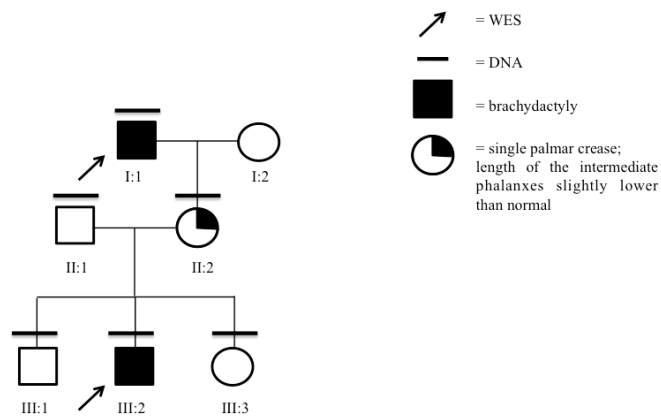


Figure 21. Pedigree of the family with brachydactyly; black lines indicate individuals for whom DNA was available for the molecular analyses; the arrows indicate individuals who underwent whole exome sequencing (WES).

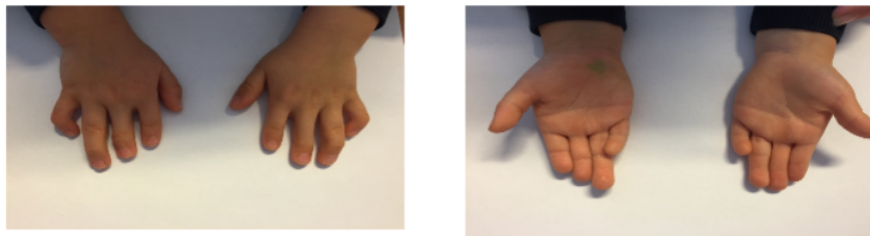


Figure 22. Clinical photographs of the proband (III:2), affected by an isolated form of brachydactyly with features of type A1 and type C.



Figure 23. Radiograph of the left hand of the proband (III:2), affected by an isolated form of brachydactyly with features of type A1 and type C. The arrows show the absence of an intermediate phalanx in the index, the presence of an anomalous intermediate phalanx in the ring finger and the delayed ossification of the carpal bones.

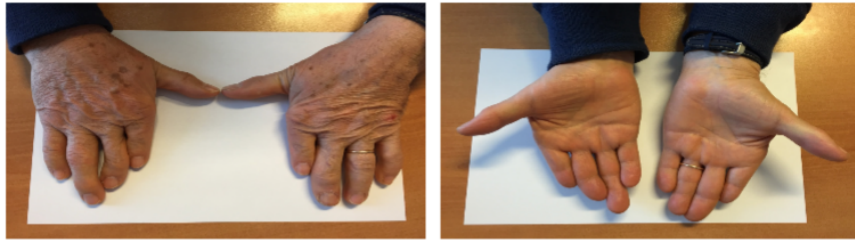


Figure 24. Clinical photographs of I:1, affected by an isolated form of brachydactyly with features of type A1 and type C.



Figure 25. Radiograph of the hands of I:1, affected by an isolated form of brachydactyly with features of type A1 and type C. The arrows show the subluxation of the metacarpophalangeal joint of the index, the presence of an intermediate phalanx of reduced size of the ring finger especially on the left and the absence of the intermediate phalanx of the index of the right hand.



Figure 26. Clinical photographs of II:2, who shows a single palmar crease and the length of the intermediate phalanges slightly lower than normal.



Figure 27. Radiograph of the left hand of II:2, who shows a single palmar crease and the length of the intermediate phalanges slightly lower than normal. The arrows show the shortening of intermediate phalanges in particular of the second and fifth fingers.

We obtained the following whole exome sequencing data output:

Subject	Coverage	Depth
I:1	77%	52x
III:2	80%	82x

Table 12. Whole exome sequencing data output.

After bioinformatic analysis of WES data, we obtained 61,840 total variants.

We then filtered the exome variants using several criteria (see Material and Methods section) and we obtained: 13,579 high-quality variants with an effect on the coding sequence and splice site regions; 1,644 variants with frequency $\leq 4\%$ (the most frequent brachydactylies have a frequency of 2%) in gnomAD database or with unknown frequency; 1,077 variants with a CADD (Combined Annotation Dependent Depletion V.1.3) score ≥ 10 , predicted to have a high functional impact on the protein. We then prioritized the variants hypothesising an X-linked transmission, obtaining 11 variants, or an autosomal dominant transmission with variable expressivity, obtaining 261 variants. Then we focused on genes known to cause isolated or syndromic brachydactylies and we selected as the most interesting a frameshift variant in *GDF5* (NM_000557.4: c.157dupC; NP_000548.2: p.Leu53Profs*41; rs778834209) that had a gnomAD frequency of 0.000866%; the CADD scoring system predicted a high functional impact (22.8).

Sanger sequencing confirmed the presence of the heterozygous variant in *GDF5* in the proband and the grandfather and disclosed its presence also in the mother; the other analysed family members resulted not carrier of the variant (Figure 28).

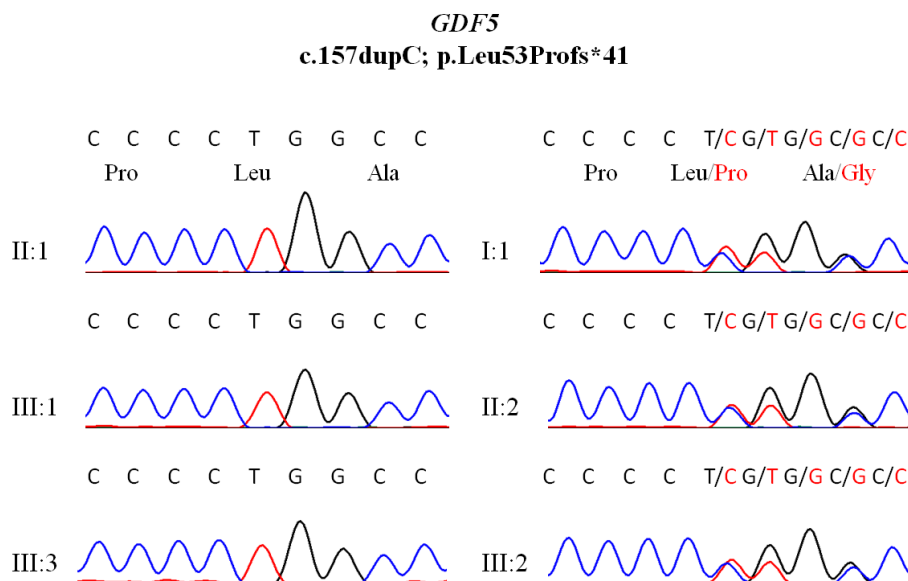


Figure 28. Electropherograms showing genotypes of patients and unaffected individuals for *GDF5* variant.

Segregation analysis of the analysed variant was summarized in the following genealogic tree:

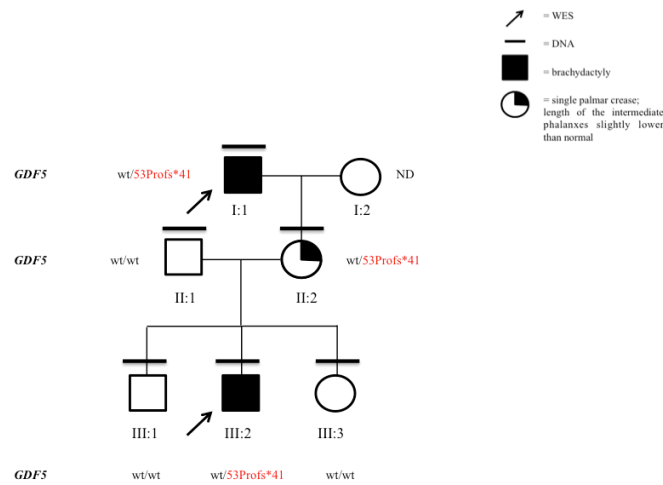


Figure 29. Variant identified in the *GDF5* gene in analysed family members; the aminoacidic substitution is reported; “wt” indicates wild type allele; “ND” indicates the genotype that has not been experimentally determined.

4.3 Family C

The family was composed of two generations (Figure 30): the proband (II:3), daughter of a healthy Italian mother and a healthy Chinese father, was affected by corpus callosum hypoplasia (Figure 31), discovered first through ultrasound and after confirmed through fetal magnetic resonance. Fetal karyotype and CGH-array, performed after amniocentesis, gave negative results. The proband has a healthy sister; the fetus from a previous interrupted pregnancy was female and showed corpus callosum agenesis and other severe malformations. Before the voluntary interruption of pregnancy, a CGH-array was performed identifying a *de novo* microduplication 17q12 (31,635,490-33,323,002) x3 of uncertain significance.

We performed clinical exome sequencing of the proband and her parents (*trio*).

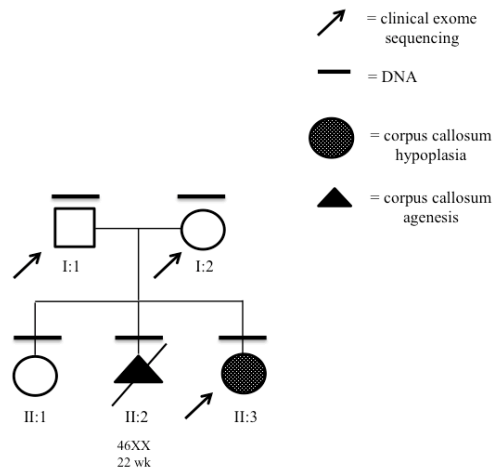


Figure 30. Pedigree of the family with corpus callosum dysgenesis; black lines indicate individuals for whom DNA was available for the molecular analyses; the arrows indicate individuals who underwent clinical exome sequencing (CES).

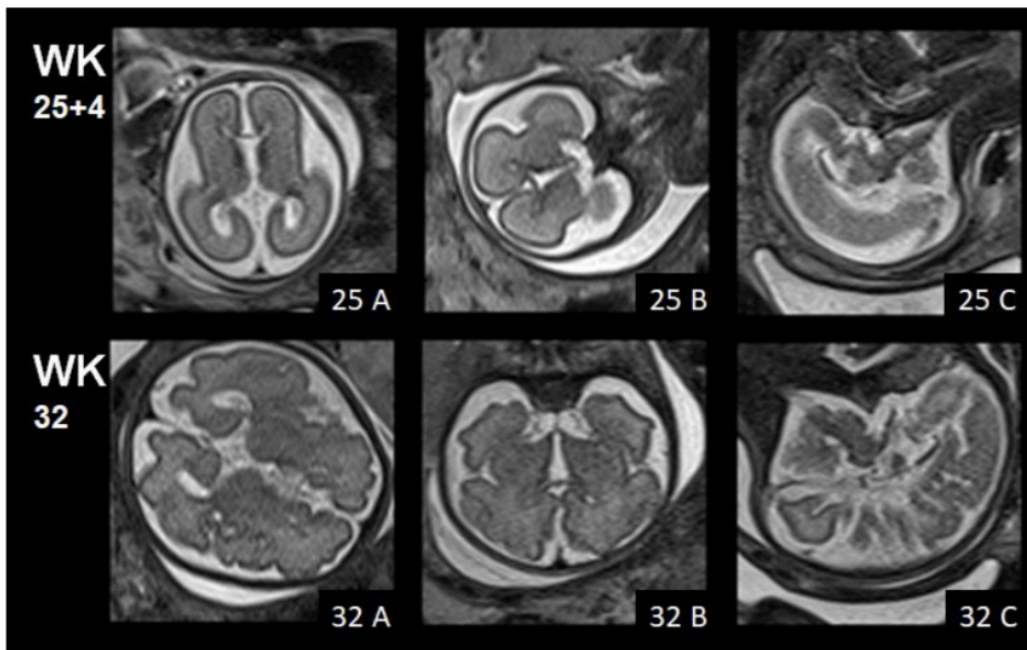


Figure 31. Fetal magnetic resonance of II:3 at 25+4 and 32 weeks. Sagittal (A), coronal (B) and transversal (C) section.

We obtained the following clinical exome sequencing data output:

Subject	Coverage	Depth
I:1	96%	96x
I:2	96%	85x
II:3	97%	109x

Table 13. Clinical exome sequencing data output.

After bioinformatic analysis of CES data, we obtained 42,062 total variants in the *trio*.

We filtered the clinical exome variants using several criteria (see Material and Methods section) and

we obtained: 4,317 high-quality variants in the proband with an effect on the coding sequence and splice site regions; 552 variants with a frequency $\leq 3\%$ or with unknown frequency in gnomAD database; 335 variants with a CADD score ≥ 10 .

Then we prioritized the variants, obtaining 7 *de novo* heterozygous, 3 homozygous and 31 compound heterozygous. We also performed a prioritization step based on the phenotype through the Phenolyzer tool. Potentially causative variants were further analysed based on different parameters, as gene function and gene expression.

We found a candidate *de novo* nonsense variant in *ARX* (NM_139058.2: c.922G>T; NP_620689.1: p.Glu308*), which was not reported in gnomAD database; it was recently annotated in ClinVar (ID 522170) as pathogenic; the CADD scoring system predicted a high functional impact (36).

Sanger sequencing confirmed the presence of the variant in the proband and disclosed the presence of the same variant also in the fetus (Figure 32).

***ARX* (BLOOD-AMNIOTIC FLUID)
c.922G>T; p.Glu308***

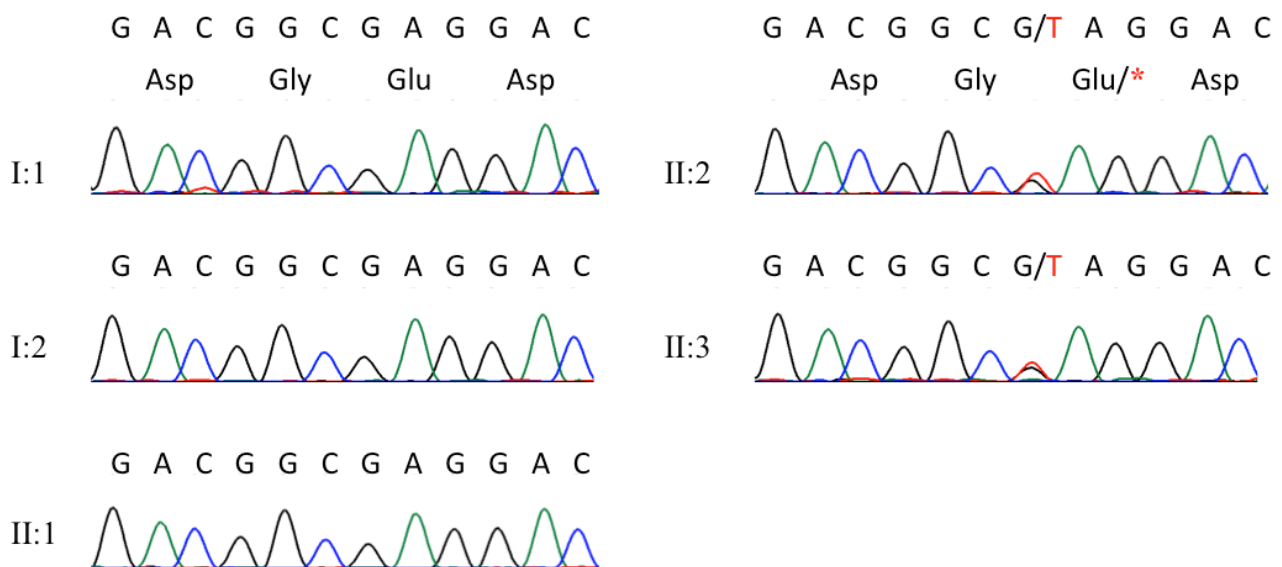


Figure 32. Electropherograms showing genotypes of patients and unaffected individuals for *ARX* variant; these sequences were obtained from DNA extracted from blood (I:1, I:2, II:1 and II:3) and from amniotic fluid (II:2).

This result led to suppose a gonadal or gonosomal mosaicism in one of the parents for the causative *ARX* variant. For this reason we performed Sanger sequencing also on DNA extracted from saliva and from hair bulbs (Figures 33 and 34) that did not disclose the presence of the variant in the parents.

***ARX* (SALIVA)
c.922G>T; p.Glu308***

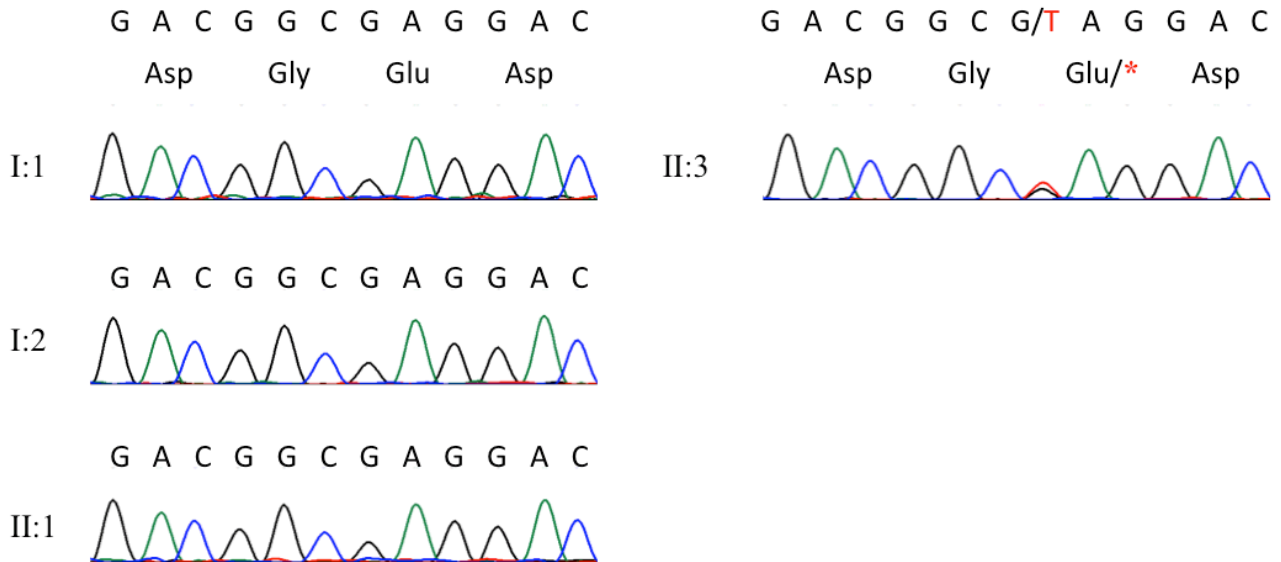


Figure 33. Electropherograms showing genotypes of patients and unaffected individuals for *ARX* variant; these sequences were obtained from DNA extracted from saliva.

***ARX* (HAIR)
c.922G>T; p.Glu308***

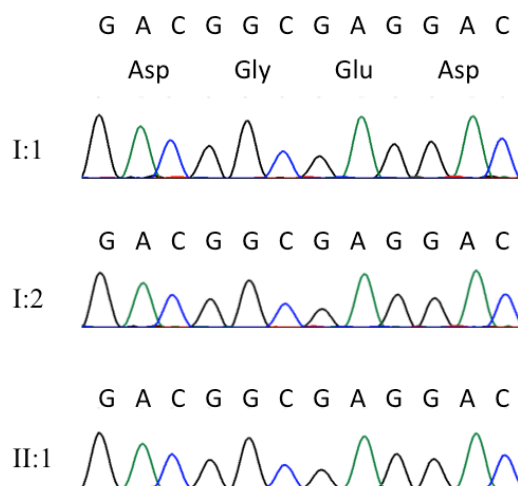


Figure 34. Electropherograms showing genotypes of unaffected individuals for *ARX* variant; these sequences were obtained from DNA extracted from hair bulbs.

Sequence analysis results in different tissues were summarized in the following genealogic tree:

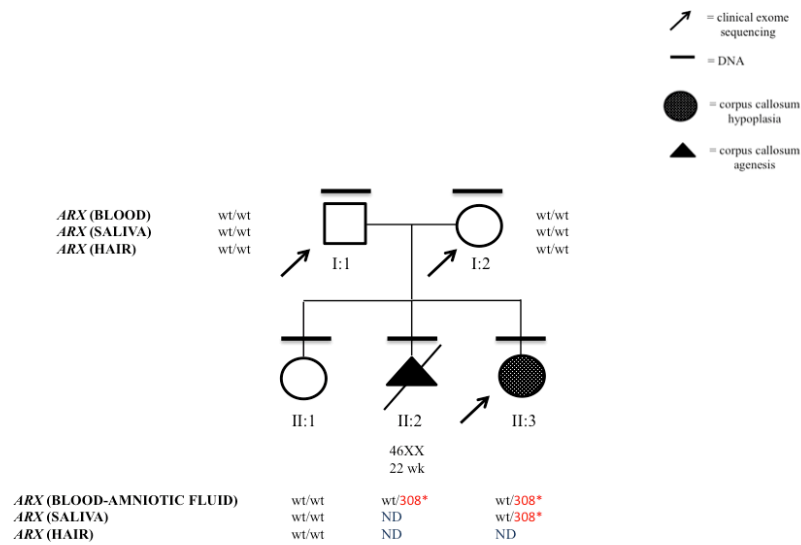


Figure 35. Variant identified in the *ARX* gene in different tissues of analysed family members; the aminoacidic substitution is reported; “wt” indicates wild type allele; “ND” indicates genotypes that have not been experimentally determined.

4.4 Family D

The family was composed of two generations (Figure 36): the proband (II:2) was a fetus with corpus callosum agenesis and other severe malformations (Figure 37). Fetal karyotype and SNP-array, performed after amniocentesis, gave negative results; SNP-array data demonstrated homozygosity levels of 1% in the fetus, excluding parental consanguinity. Corpus callosum agenesis was discovered first through ultrasound and after confirmed through fetal magnetic resonance. A previous pregnancy was interrupted because of a male fetus with the Dandy-Walker syndrome (OMIM %220200) and hydrocephalous.

We analysed the pedigree and performed WES of the proband (II:2).

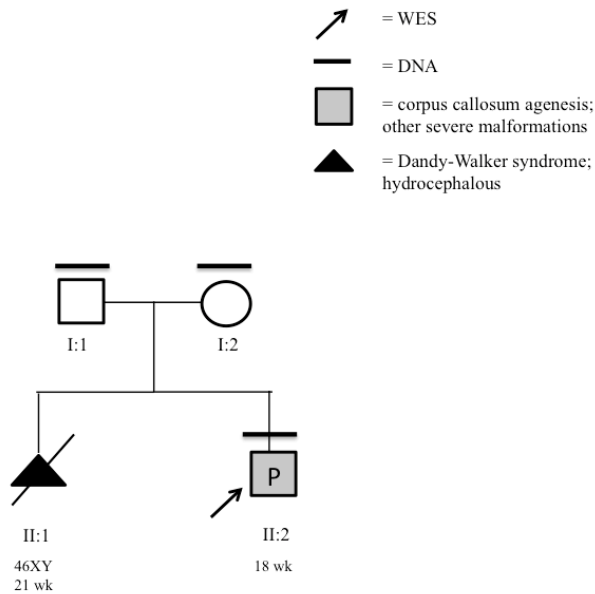


Figure 36. Pedigree of the family with corpus callosum dysgenesis; black lines indicate individuals for whom DNA was available for the molecular analyses; the arrow indicates the individual who underwent whole exome sequencing (WES).

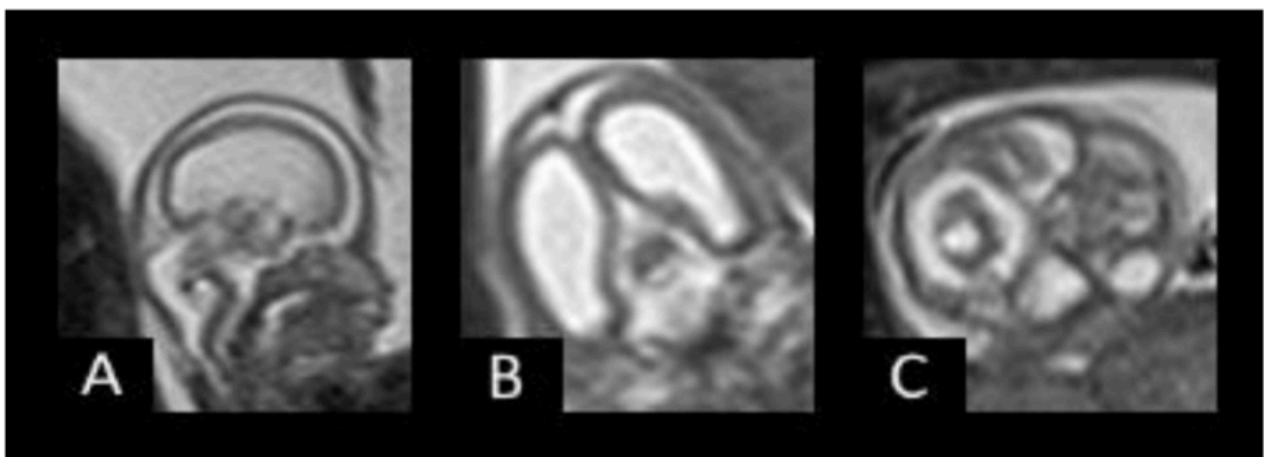


Figure 37. Fetal magnetic resonance of II:2. Sagittal (A), coronal (B) and transversal (C) section.

After bioinformatic analysis of WES data, we obtained 46,610 total variants.

We filtered the exome variants using several criteria (see Material and Methods section) and we obtained: 11,301 high-quality variants with an effect on the coding sequence and splice site regions; 1,132 variants with a frequency $\leq 3\%$ or with unknown frequency in gnomAD database; 713 variants with a CADD (Combined Annotation Dependent Depletion V.1.3) score ≥ 10 . The phenotype, in fact, is very rare and severe and we supposed it has to be caused by a mutation with a high functional impact on the protein.

Then we prioritized the variants hypothesising an autosomal recessive transmission, obtaining 42 variants, or an X-linked transmission, obtaining 21 variants, or an autosomal dominant

transmission, obtaining 658 variants. We also performed a prioritization step based on the phenotype through the Phenolyzer tool. Potentially causative variants were further analysed on the bases of different parameters: we studied gene function, gene expression and animal models that we retrieved from several databases, as OMIM-Online Mendelian Inheritance in Man, HPO-Human Phenotype Ontology and MGI-Mouse Genome Informatics.

We selected as the most interesting a homozygous variant in *FKTN* (NM_006731.2: c.898G>A; NP_006722.2: p.Gly300Arg; rs909129168), which had no frequency in gnomAD database; the CADD scoring system predicted a high functional impact (34).

Sanger sequencing confirmed the presence of the homozygous variant in the proband; both the parents resulted heterozygous (Figure 38).

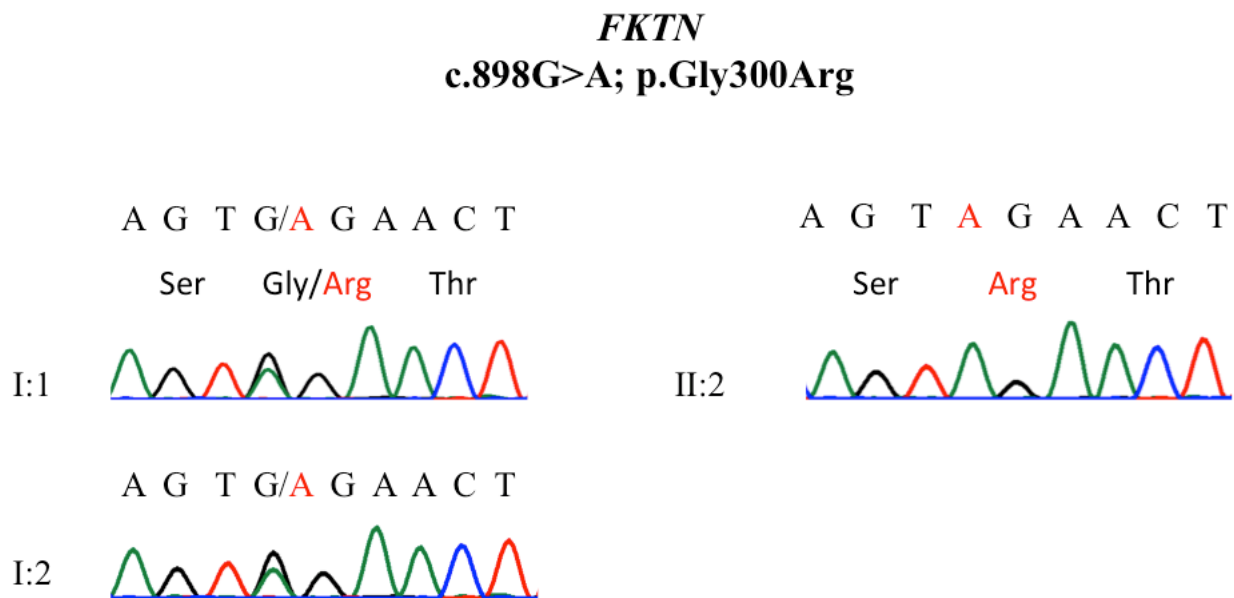


Figure 38. Electropherograms showing genotypes of patients and unaffected individuals for *FKTN* variant.

Segregation analysis of the analysed variant was summarized in the following genealogic tree:

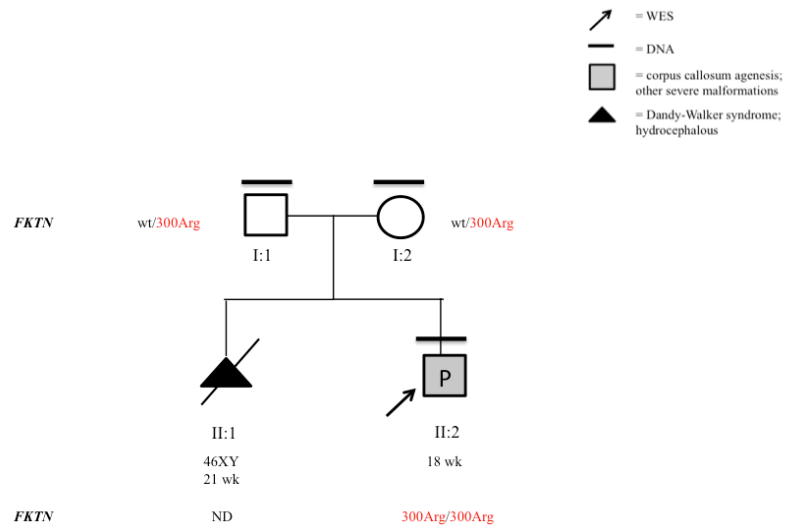


Figure 39. Variant identified in the *FKTN* gene in analysed family members; the aminoacidic substitution is reported; “wt” indicates wild type allele; “ND” indicates the genotype that has not been experimentally determined.

4.4.1 Modeling of the nucleotidyltransferase domain of FKTN

In order to test the pathogenicity of the variant, a structural modeling of FKTN was performed. Domain analysis of human FKTN (hFKTN) protein (NP_001073270.1) showed that Gly300 is part of a nucleotidyltransferase (NT) fold comprising residues 278-411 of hFKTN. Modeling of residues 278-411 of hFKTN (based on PDB Codes 4WQL, nucleotidyltransferase ANT(2'')-Ia and 4FO1, adenylyltransferase LnuA) showed that this domain follows the $\alpha 1\text{-}\beta 1\text{-}\alpha 2\text{-}\beta 2\text{-X-}\beta 4\text{-}\beta 3$ NT fold topology and that binding of nucleoside triphosphate groups (NTGs) is stabilized by two Mg^{2+} ions in the active site. Due to similarities in the active sites of LnuA, ANT(2'')-Ia and hFKTN, it was possible to identify the putative nucleotide binding site in hFKTN (Figures 40 and 41).

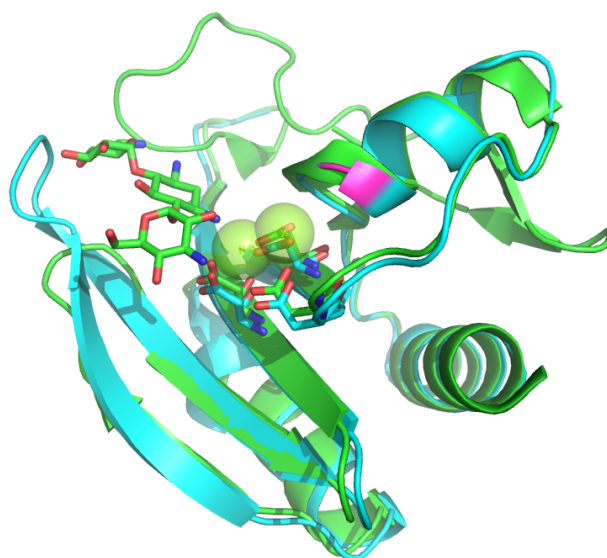


Figure 40. Comparison between the model of hFKTN (residues 278-411; cyan) and its structural template 4WQL (green). The two Mg²⁺ ions are displayed as spheres. Conserved Asp residues coordinating the ions are shown as sticks. The aminoglycoside Kanamycin is also shown as green sticks. The position of Gly300 is shown in pink.

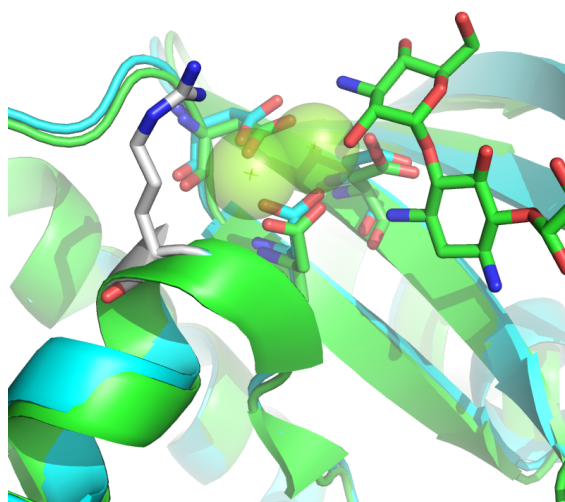


Figure 41. Comparison between the model of hFKTN (residues 278-411; cyan) and its structural template 4WQL (green). The two Mg²⁺ ions are displayed as spheres. Conserved Asp residues coordinating the ions are shown as sticks. The aminoglycoside Kanamycin is also shown as green sticks. The mutation of Gly300 with Arg is shown in white sticks. The position of the Arg residue corresponds to the predicted binding site of ATP.

This comparison identified a cleft just beside the active site and sugar binding site in which the Gly300 residue resides.

The analysis of the active site of hFKTN was highly consistent with a catalytic mechanism involving a nucleophilic attack on the α -phosphate of the NTGs by the substrate 2''-OH of sugar moieties, for subsequent transfer to α -dystroglycan. Replacing Gly300 with an Arg residue was therefore predicted to have steric hindrance effects that prevent the NTGs from binding the cleft,

abolishing the activation of the sugar moiety. The latter activity is required for the glycosylation of α -dystroglycan in skeletal muscle.

Thus, Gly300Arg is likely responsible for a reduced activation of the sugar moieties, which in turn impairs transfer to α -dystroglycan.

5. DISCUSSION

We studied the molecular bases of four different phenotypes with a supposed genetic cause through different NGS technologies and data analysis approaches.

The selection of the experimental strategy, the number of subjects to sequence and the data analysis approach were dictated by some considerations on the diagnostic potential of each sequencing strategy and its feasibility and cost.

Indeed, some criteria can significantly influence the choice of which NGS test has to be performed. They include the diagnostic rate, the possibility to re-evaluate the NGS data periodically, the management of NGS data, the functional interpretation of coding and non coding variants and the number of secondary findings. Moreover, there are some features that influence the choice and that depend on each specific case, e.g. the supposed mode of inheritance, the available samples and the information about the phenotype object of the study. For this reason, a deep and as much as accurate as possible knowledge about the potentiality of each NGS approach is essential to solve the molecular bases of a phenotypic picture. In the studied cases, we selected a different NGS strategy (i.e. whole exome or clinical exome sequencing) and different family members to sequence (i.e. the most distant family members, *trio* or singleton) according to considerations related to the observed phenotype and the hypothesized segregation pattern. The adopted experimental and data analysis strategies allowed to identify the molecular bases of phenotypes involving different systems (i.e. teeth, limbs and central nervous system) and belonging to different clinical pictures.

To study the molecular bases of the complex phenotypes regarding canine agenesis and eruption anomalies in the family A, we used a WES approach on three first degree cousins. We selected an exome sequencing strategy as the pedigree analysis suggested a significant genetic component underlying the phenotype but with a complex segregation pattern. Different data analyses, based on different shared genetic causes, allowed to identify several candidate variants potentially involved in the pathogenetic mechanisms.

To find the cause of the isolated brachydactyly observed in family B, we used a WES approach on the proband and his grandfather, looking for shared variants in the exome. We selected this strategy as isolated brachidactylies are a group of very heterogeneous limb anomalies, with different and not completely characterized molecular bases.

To find the cause of the corpus callosum anomaly observed in the proband of family C, we chose a *trio* based approach, as the segregation pattern of the causative variant was not known. We performed a clinical exome sequencing, using an enrichment kit that included 171 on a total of 180

genes reported in literature as causative of corpus callosum malformations.

To investigate the molecular bases of the recurrent phenotype observed in the family D, we performed WES only of the proband, because the phenotype suggested a recessive disorder. The malformation was at first defined as an isolated Dandy-Walker malformation, a rare anomaly for which molecular causes have not yet been recognized.

5.1 Family A

Dental agenesis is one of the most common human dental abnormalities: its prevalence, excluding third molars, ranges between 0.15% and 16.2% (Rakhshan, 2015). This condition may be classified as “oligodontia”, which consists of the absence of more than six teeth (excluding third molars), or “hypodontia”, which consists of the absence of one to six teeth.

It may occur either as isolated condition or in syndromic phenotypes and in both familial and sporadic cases (Nieminen, 2009); both genetic and environmental factors are supposed to contribute to its pathogenesis.

Several genes of few cell signaling pathways have been associated with isolated hypodontia, that seems to be transmitted as a dominant or recessive trait: *PAX9* (Stockton et al., 2000), *EDA* (Tao et al., 2006), *MSX1* (Vastardis et al., 1996), *AXIN2*, *EDARADD* (Bergendal et al., 2011), *LRP6* (Massink et al., 2015), *WNT10A* (Kantaputra and Sripathomsawat, 2011), *GREM2* (Kantaputra et al., 2015), *BMP4*, *BMP2* (Mu et al., 2012), *WNT10B* (Yu et al., 2016), *PTHIR* (Decker et al., 2008), *EDAR* (Arte et al., 2013) and *SMOC2* (Alfawaz et al., 2013); mutations in *WNT10A* have been also reported as associated with the isolated agenesis of the permanent maxillary canine (Kantaputra et al., 2014).

Another human dental abnormality is tooth impaction: the maxillary permanent canine is the second most frequent impacted tooth after the third molar, with a reported prevalence of 1-2% (Sajjani, 2015). This could be due to a perturbation of tooth eruption and several contributing factors have been suggested including localized, systemic and genetic causes (Becker and Chaushu, 2015; Leonardi et al., 2003; Leonardi et al., 2009; Lombardo et al., 2007; Mercuri et al., 2013; Peck et al., 1994) (Figure 42).

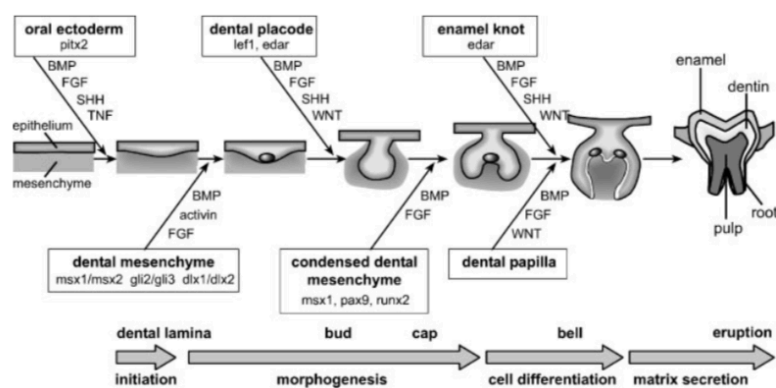


Figure 42. Tooth development is regulated by conserved signaling pathways (FGF, BMP, SHH, WNT, TNF). The signals mediate interactions between the oral ectoderm and mesenchyme and regulate the expression of key transcription factors (shown in the boxes) (Thesleff, 2006).

The peculiar aspect of the studied family is the occurrence of several members with a variable phenotype that specifically involved maxillary canines, including agenesis, inclusion and ectopic eruption, without any other dental or extra-dental associated feature. The only exception is represented by the subject III:1, who shows a microdontic lateral incisor (1.2) placed on the opposite side of the impacted maxillary canine (2.3). Interesting to note, the ectopic labial eruption of maxillary canine observed in patients II:3 and III:4 is not due to inadequate arch space permitting the exclusion of the mechanical effect as the cause of eruption disorder.

The exclusively female expression of the phenotypes, which appears to be characterized by incomplete penetrance in males, could be related to the strong difference in the prevalence of maxillary canine impaction between females and males (Ericson and Kuroi, 1986), which could be caused by an earlier dentition development in the former (Rutledge and Hartsfield, 2010); however, it might be also due to the skewed proportion of females in the family tree (11 females vs 4 males). Canine anomalies are regarded as complex traits, but this uncommon family suggests a significant genetic contribution to these phenotypes: they might either represent different manifestations of the same basic disorder or they might share, in part, a molecular basis. In the first two branches of the family four members with impacted or ectopic erupted canines are present: a transmission of the trait according to an autosomal dominant segregation model, with incomplete penetrance and variable expressivity, is suggested by the evidence that ectopic eruption may be a manifestation of the impaction (Peck et al., 1994) and that both buccally and palatally displaced canines share similar etiologies (Sajnani and King, 2012). A more complex segregation pattern could occur in the third branch of the family, due to a possible concomitant paternal and maternal contribution.

To date, only very few studies report Next Generation Sequencing strategies to study orodontal pathologies, in particular canine anomalies (Massink et al., 2015; Ockeloen et al., 2016; Prasad et al., 2016; Salvi et al., 2016; Yamaguchi et al., 2017).

WES approach on three first degree cousins allowed to identify a heterozygous mutation in the *EDARADD* gene (NM_145861.2: c.308C>T; NP_665860.2: p.Ser103Phe; rs114632254) in two subjects with the most severe phenotype (i.e. canine agenesis). The variant was transmitted by the mother, who is not related to the family and presents a phenotype of bilateral impacted canines. *EDARADD* codes for a protein that interacts with EDAR, a death domain receptor required for hair, teeth and the development of other ectodermal derivatives. *EDARADD* is expressed in epithelial cells during the formation of hair follicles and teeth. Indeed, mutations in *EDARADD* have been associated with ectodermal dysplasias (OMIM #614940, #614941), characterized by defective development of hair, teeth and eccrine sweat glands (van der Hout et al., 2008). p.Ser103Phe variant has been previously associated with isolated oligodontia, including canine teeth, and is predicted to be functionally relevant (Arte et al., 2013; Bergendal et al., 2011; Salvi et al., 2016). *EDARADD* mutation carriers in this family did not show symptoms of ectodermal origin other than teeth dysgenesis. This finding suggests the involvement of the EDA signaling pathway in isolated oligodontia and confirms the considerable variation in clinical expression of *EDARADD* mutations. The most severe phenotype of subjects III:5 and III:6 could be caused by the concurrent contribution of paternally inherited DNA variants at other loci. In fact, the association of the p.Ser103Phe *EDARADD* variant with other mutations in genes related to teeth development has been already reported in isolated oligodontia, suggesting additive effects of other pathways, e.g. the WNT signaling pathway (Arte et al., 2013; Salvi et al., 2016).

Consistent with this hypothesis, a *COL5A1* variant (NM_000093.4: c.1588G>A; NP_000084.3: p.Gly530Ser; rs61735045) was identified in subjects with canine agenesis (III:5 and III:6), segregating from the paternal affected grandmother (I:2). *COL5A1* codes for a component of type V collagen particularly expressed in tendons. The variant localizes to the “interrupted collagenous region” of the protein and is predicted to interfere with the correct folding of the COL2 domain (Symoens et al., 2011). Haploinsufficiency of *COL5A1* is a cause of Ehlers-Danlos syndrome (OMIM #130000), where dental anomalies, including hypodontia of permanent teeth and delayed eruption, are a common feature. Previously described cases suggest that this substitution causes a recessive form of Ehlers-Danlos syndrome (Giunta et al., 2002). The Gly530Ser change was previously reported as a disease modifying variant in Ehlers-Danlos syndrome type 1 (Steinmann and Giunta, 2000) and also as a biallelic causative mutation, with heterozygous parents showing subtle clinical signs (Giunta et al., 2002). None of the mutation carriers in the analysed family showed symptoms of connective disorder. However, it should be noted that the clinical significance of this variant has not been definitely established as there are conflicting interpretations of its pathogenicity (<https://www.ncbi.nlm.nih.gov/clinvar/variation/38863>).

According to a supposed dominant segregation model with incomplete penetrance and variable expressivity, several shared candidate variants in genes functionally related to tooth morphogenesis were identified in the two analysed subjects affected by canine eruption anomalies (III:1 and III:4). Potential harmful missense variants were identified in *RSPO4* (NM_001029871.3: c.317G>A; NP_001025042.2: p.Arg106Gln; rs6140807), *T* (NM_003181.3: c.1013C>T; NP_003172.1: p.Ala338Val; rs117097130) and *NELLI* (NM_001288713.1: c.1244G>A; NP_001275642.1: p.Arg415His; rs141323787) genes.

RSPO4 codes for a member of the R-spondin protein family that has essential roles in vertebrate development and is expressed in the dental papilla (Pemberton et al., 2007). *RSPO4* is a secreted protein that may be involved in activation of WNT/ β -catenin signaling pathway. It is mutated in inherited anonychia (OMIM #206800), a rare autosomal recessive condition characterized by the absence or severe hypoplasia of nails. The p.Arg106Gln variant is positioned in the Furin-like repeat two domain of the protein, which is required for β -catenin stabilization (Ishii et al., 2008).

The *T* gene codes for Brachyury, a transcription factor placed downstream of the WNT/ β -catenin signaling pathway (Arnold et al., 2000) that binds to a specific DNA element, the palindromic T-site. It binds through a region in its N-terminus, called the T-box, and effects transcription of genes required for mesoderm formation and differentiation. Murine models demonstrated that this gene is involved in establishing notochord cell identity and differentiation, and in the organization of the axial development (Herrmann, 1992). Diseases associated with *T* include Sacral Agenesis with Vertebral Anomalies (OMIM #615709) and Neural Tube Defects (OMIM #182940). The p.Ala338Val substitution lies in the second transactivation domain (Kispert et al., 1995) and has been previously associated with variable vertebral phenotypes, including multiple regional vertebral segmentation defects (Ghebranius et al., 2008).

NELLI codes for a cytoplasmic protein that contains epidermal growth factor (EGF)-like repeats: Nel-like molecule-1 (Nell-1) is a secreted heterotrimeric protein that plays an important role in osteoblast differentiation, bone formation and regeneration; it may be involved in cell growth regulation and differentiation. It is a protein strongly expressed in neural tissue, encoding epidermal growth factor-like domains suggesting its specificity for the craniofacial region. The expression patterns during tooth development suggest that it plays an important role in tooth morphogenesis (Tang et al., 2013).

The WES data and segregation analyses of the family pointed to two different signaling pathways as responsible for the dental phenotypes, one of them (i.e. EDA) for the canine agenesis, and the other (i.e. WNT) for the less severe canine eruption anomalies.

Further functional research on the mechanisms through which EDA and WNT pathways regulate

tooth morphogenesis and eruption is warranted to shed light on the pathogenesis of tooth agenesis.

5.2 Family B

The term brachydactyly is derived from the ancient Greek (brachy-: short; dactylos: digit) and indicates shortening of the hands and/or feet digits due to a lack or an abnormal development of phalanges, metacarpals, or both. It may occur as an isolated trait or as part of a syndrome (Temtamy and Aglan, 2008). According to the skeletal involvement, the isolated brachydactyly forms have been categorized in the groups A–E (Figure 43), including several subgroups. However, there is a considerable phenotypic overlap.

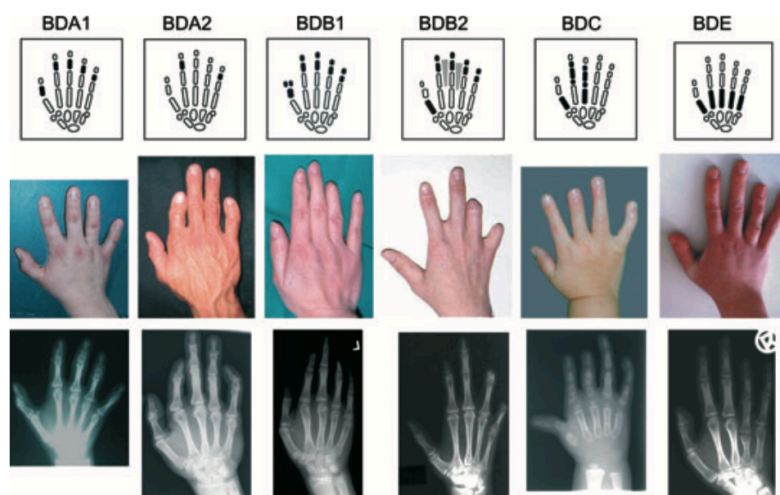


Figure 43. Clinical features of non-syndromic brachydactylies. Schematics showing bone and tissue involvement are shown on top. Middle row shows typical clinical features of hands. The corresponding X-ray figures are shown below (modified from Mundlos, 2009).

The isolated forms usually occur as autosomal dominant traits and show variable expressivity and incomplete penetrance (Mundlos, 2009). Brachydactyly may be also associated with other hand malformations, such as syndactyly, polydactyly, reduction defects or symphalangism (Temtamy and Aglan, 2008); it can also occur as part of a syndrome and in bone dysplasias, as hand foot genital syndrome (HFGS; OMIM #140000), Robinow syndrome (RS; OMIM #268310), Acromesomelic chondrodysplasia with genital anomalies (OMIM #609441), Grebe type acromesomelic dysplasia (OMIM #200700), Hunter-Thompson type acromesomelic dysplasia (OMIM #201250), Du Pan type acromesomelic dysplasia (OMIM #228900) and Feingold or oculodigitoesophagoduodenal syndrome (OMIM #164280) (Mundlos, 2009). Many brachydactylies as well as the acromesomelic dysplasias are caused by mutation in various components of the BMP pathway and its modulators

(Mundlos, 2009).

Despite many descriptions of these phenotypes in literature, there are no unanimously definite clinical or radiological criteria to assess the type of brachydactyly affecting a patient and the relative weight of the single features is not determined.

The family we studied is composed of three generations: the proband is a 5-year-old boy affected by an isolated form of brachydactyly with features of type A1 (OMIM #112500) and type C (OMIM #113100), characterized by long proximal phalanx of finger, delayed ossification of carpal bones, short metacarpal, aplasia/hypoplasia of the middle phalanges of the hand, abnormality of the distal and the middle phalanges of the toes and clinodactyly of the 5th finger, as his maternal grandfather, who shows an isolated brachydactyly with the same characteristics; both of them have only one ring finger of the normal length; the proband was also referred to the medical geneticist because of his short stature, which is a relatively common but inconstant feature in some isolated or syndromic brachydactylies. His mother shows a mild phenotype: single palmar crease; length of the intermediate phalanxes slightly lower than normal.

To find the cause of the isolated brachydactyly in this family, we used a WES approach on the proband (III:2) and his grandfather (I:1). Through this approach we identified a heterozygous variant in the *GDF5* gene (NM_000557.4: c.157dupC; NP_000548.2: p.Leu53Profs*41; rs778834209).

GDF5 gene is predominantly expressed in long bones during embryonic development; in the adult, it is expressed in fibroblasts and salivary glands. This gene encodes for growth and differentiation factor 5, a secreted ligand of the TGF- β (transforming growth factor-beta) superfamily of proteins; ligands of this family bind various TGF- β receptors leading to recruitment and activation of SMAD family transcription factors that regulate gene expression.

The biological role of GDF-5 *in vivo* became first apparent from the genetic analysis of the *brachypodism* mice (*bp*), in which the length and the number of bones in the limbs were altered (Storm et al., 1994). These studies also led to the discovery of GDF-6 and -7 (Storm et al., 1994).

Loss of function mutations in *GDF5* reduce the signaling through the BMPRII receptor and result in loss of an element (brachydactyly), whereas an increase in signaling either through activating mutations in *GDF5* or through loss of function mutations in the inhibitor Noggin results in joint fusion (sympalangism) (Mundlos, 2009; Figure 44). *GDF5*-induced isolated brachydactylies are characterized by the relative preservation of the ring finger: it could be due to the highest p-SMAD activity in this finger, which permits a residual signaling via BMPs when *GDF5* is mutated (Al-Qattan, 2014).

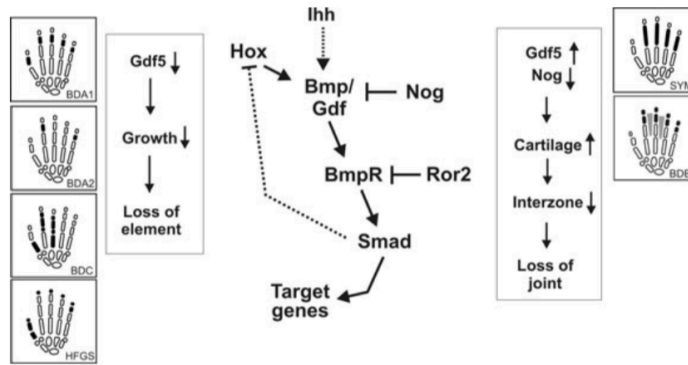


Figure 44. Schematic diagram of BMP pathway and its modulators that control digit and joint development (Mundlos, 2009).

A GDF-5 monomer consists of an N-terminal signal peptide domain, a prodomain and the C-terminal mature part containing six highly conserved cysteine residues forming the cystine knot motif, whereas the seventh cysteine connects two monomers via an intermolecular disulfide bond. Single nucleotide polymorphisms (SNPs) in 5' and 3' UTR regions of the gene are associated with osteoarthritis (Valdes et al., 2012); mutations within the prodomain are typically frameshift mutations that prevent the production of an active protein and may be subclassified into two subgroups according to zygosity; mutations within the active (mature) domain and within the cleavage area (between the prodomain and active domain) are usually amino acid substitutions that can impair GDF5 function in many ways, causing a variety of phenotypes (Al-Qattan et al., 2015; Figure 45).

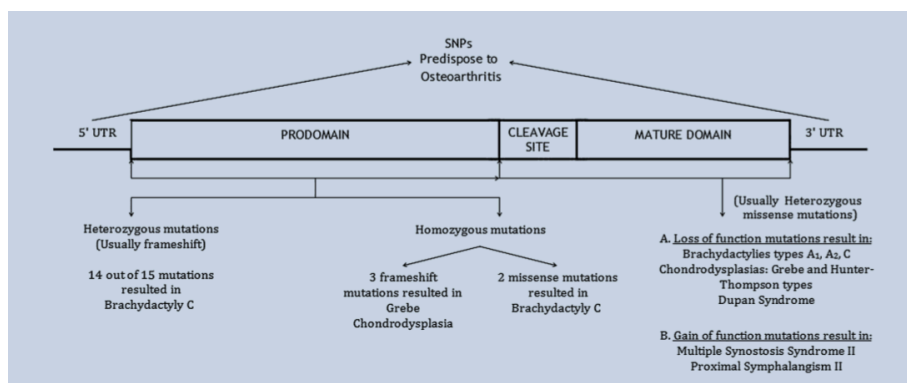


Figure 45. Genotype-phenotype correlations in the GDF5 mutational spectrum (Al-Qattan et al., 2015).

Mutations along the whole sequence of *GDF5* are known to cause in a heterozygous or homozygous state Brachydactyly type A1, C (OMIM #615072) (Byrnes et al., 2010), Brachydactyly type A2 (OMIM #112600) (Kjaer et al., 2006; Plöger et al., 2008; Schwaerzer et al., 2012; Seemann et al., 2005), Brachydactyly type C (OMIM #113100) (Al-Qattan et al., 2015; Everman et

al., 2002; Galjaard et al., 2001; Polinkovsky et al., 1997; Savarirayan et al., 2003; Schwabe et al., 2004; Seo et al., 2013; Stange et al., 2014; Stange et al., 2015; Stavropoulos et al., 2016; Travieso-Suárez et al., 2018; Ullah et al., 2018; Uyguner et al., 2014; Yang et al., 2008), Multiple synostoses syndrome 2 (OMIM #610017) (Dawson et al., 2006; Schwaerzer et al., 2012; Seemann et al., 2009), Symphalangism, proximal, 1B (OMIM #615298) (Wang X et al., 2006; Yang et al., 2008), Grebe chondrodysplasia (OMIM #200700) (Al-Yahyaee et al., 2003; Basit et al., 2008; Costa et al., 1998; Faiyaz-Ul-Haque et al., 2002a; Faiyaz-Ul-Haque et al., 2008; Martinez-Garcia et al., 2016; Mumtaz et al., 2015; Umair et al., 2017), acromesomelic dysplasia Hunter-Thompson type (OMIM #201250) (Thomas et al., 1996) and Du Pan syndrome (OMIM #228900) (Douzgou et al., 2008; Faiyaz-Ul-Haque et al., 2002b; Szczaluba et al., 2005).

p.Leu53Profs*41 is a frameshift duplication located in the prodomain of GDF5 protein, which creates a premature stop codon 41 codons downstream of duplication, resulting in a truncated protein product comprising only of 92 amino acids. It results in loss of function of the GDF5 protein (Umair et al., 2017).

The variant has been already associated with brachydactyly type C in a heterozygous state (Everman et al., 2002) and with Grebe chondrodysplasia in a homozygous state (Umair et al., 2017). The identification of this variant was important for genetic counselling as it could be hypothesized that it is causative of a mild phenotype, i.e. limb anomaly, in heterozygous state, but also of a very severe phenotype, characterized by severe abnormality of the limbs and limb joints, in a homozygous state.

5.3 Families C and D

The corpus callosum (CC) is the largest white matter tract in the human brain, containing about 200 million axons that connect the left and right cerebral hemispheres (Aboitiz and Montiel, 2003). It is one of the five main cerebral commissures, which are bundles of nerve fibres that cross the midline of the human brain at the level of their origin. The others are the anterior, posterior, hippocampal and habenular commissures (Palmer and Mowat, 2014). The corpus callosum starts to develop at approximately 8 weeks of gestational age in humans and it is completely developed at 18-19 weeks, although further maturation and growth continue into postnatal life (Craven et al., 2015; Figure 46).

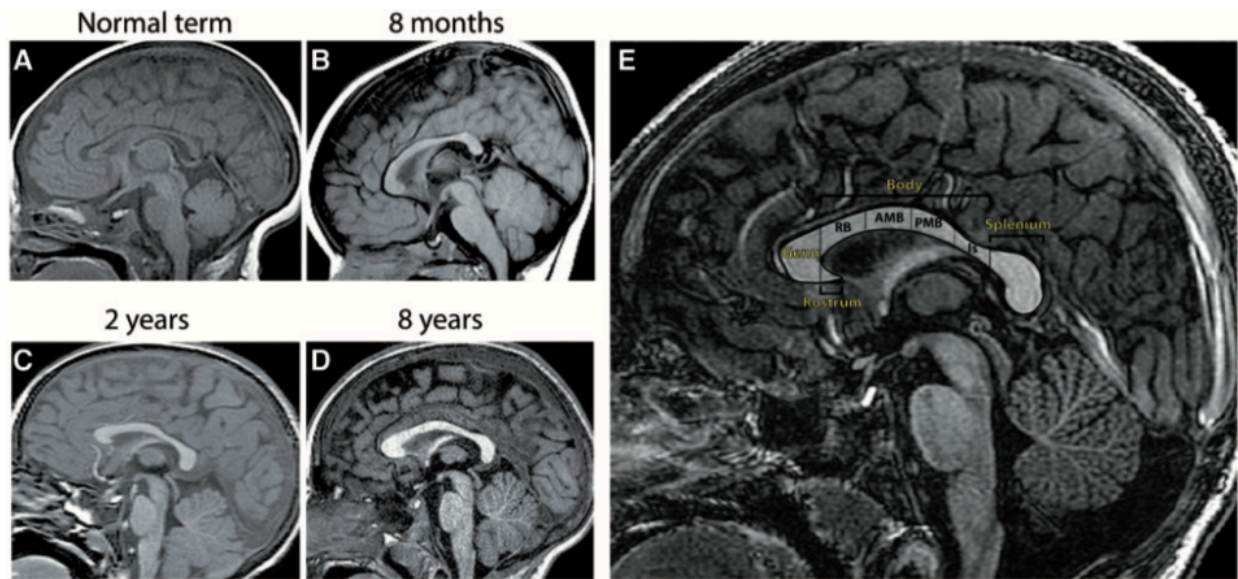


Figure 46. T1-weighted sagittal MRI scans showing the structure of the normal human corpus callosum in the full-term infant (A), 8-month-old (B), 2-year-old (C), 8-year-old (D) and adult (E) (Edwards et al., 2014).

Functionally, the corpus callosum permits not only the information transfer between cerebral hemispheres, but also the inhibition of concurrent activity in the contralateral hemisphere (Ozyüncü et al., 2014). Defects of the corpus callosum development can lead to different form of dysgenesis: this term refers to the CC being present but malformed in some way. Agenesis of the corpus callosum (ACC) is one of the most common congenital brain anomalies, with an estimated prevalence ranging from 1.8 per 10,000 in the general population to 230–600 per 10,000 in children with neurodevelopmental disabilities (D'Antonio et al., 2016). It results from failure of commissuration and implies that the entire structure has failed to form.

ACC is a clinically and genetically heterogeneous condition, which can be observed either as an isolated phenotype or as a manifestation in the context of an autosomal-dominant, autosomal-recessive or X-linked syndrome, as, for example, Apert syndrome (OMIM #101200), Miller–Dieker syndrome (OMIM #247200), Mowat–Wilson syndrome (OMIM #235730), Rubinstein–Taybi syndrome (OMIM #180849), Joubert syndrome (OMIM #608629), Walker–Warburg syndrome (OMIM #236670) and Opitz GBBB syndrome (OMIM #300000). Several causative mutations have been identified so far. In addition, ACC has been observed in constitutional trisomies as well as in some chromosomal rearrangements (Schell-Apacik et al., 2008).

Hypoplasia of the corpus callosum consists, instead, in the partial failure of the corpus callosum development: the CC is thinner than normal, but with a regular anterior-posterior extent (Craven et al., 2015). This phenotype can be isolated or associated with a syndrome. Also other abnormalities of the CC, as its hyperplasia, have been noted in a variety of neurodevelopmental conditions

(Palmer and Mowat, 2014).

The family C is composed of two generations: the proband is a 4-year-old girl affected by corpus callosum hypoplasia and growth and speech delays, daughter of a healthy Italian mother and a healthy Chinese father; the proband has a healthy sister; the fetus from a previous interrupted pregnancy was female and showed corpus callosum agenesis and other severe malformations.

To find the cause of these recurrent corpus callosum phenotypes in the family, we performed a *trio* based clinical exome sequencing approach and we identified a supposed *de novo* nonsense variant in *ARX* gene (NM_139058.2: c.922G>T; NP_620689.1: p.Glu308*) in the proband.

ARX gene is on the X chromosome and encodes a transcription factor, which contains one homeobox and one aristaless domain. It is highly expressed in gonads and brain, but also in pancreas and skeletal muscle. *ARX* is a homeobox-containing gene critical for early development and formation of a normal brain (Kitamura et al., 2002; Ohira et al., 2002; Strømme et al., 2002). It plays a vital role in telencephalic development, specifically in tangential migration and differentiation of GABAergic and cholinergic neurons (Colasante et al., 2009; Colombo et al., 2007; Friocourt et al., 2008; Kitamura et al., 2002; Lee K et al., 2014).

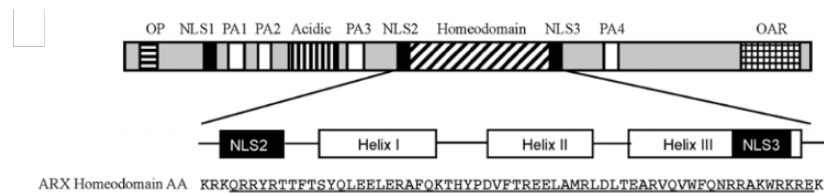


Figure 47. Schematic representation of the human *ARX* protein (Mattiske et al., 2017).

ARX contains multiple domains that include four polyalanine (pA) tracts (Figure 47), the first two of which are frequently expanded by mutations (Fullston et al., 2011).

ARX is known to cause a broad spectrum of CNS disorders, including Epileptic Encephalopathy, Early Infantile, 1 (EIEE1; OMIM #308350), hydranencephaly or lissencephaly with abnormal genitalia (OMIM #300215), Partington syndrome (OMIM #309510), Proud syndrome (OMIM #300004) and intellectual disability (OMIM #300419).

Mutations in this gene typically result in families with affected males across multiple generations transmitted via (usually) asymptomatic carrier females. However, there is an increasing prevalence of reported mutations in *ARX* that result in lissencephaly with abnormal genitalia (XLAG) in male patients and in variable and milder forms in females (Ekşioğlu et al., 2011; Kato et al., 2004; Kitamura et al., 2002; Marsh et al., 2009; Scheffer et al., 2002; Strømme et al., 2002). In these

familial cases the mutations are either missense mutations of residues in the homeodomain or nonsense/deletion mutations resulting in a loss-of-function of the ARX homeodomain and/or aristaless domain activity. A smaller number of *de novo* cases also result in truncation and loss of ARX function (Bettella et al., 2013; Kwong et al., 2015; Marsh et al., 2009; Wallerstein et al., 2008), but the type and the location of mutations in affected females are restricted compared with those in affected males. Across these cases, there is a consistent phenotype of intellectual disability (ID) and/or developmental delay, infantile seizures and hypotonia/dystonia/ataxia (Mattiske et al., 2017; Figure 48).

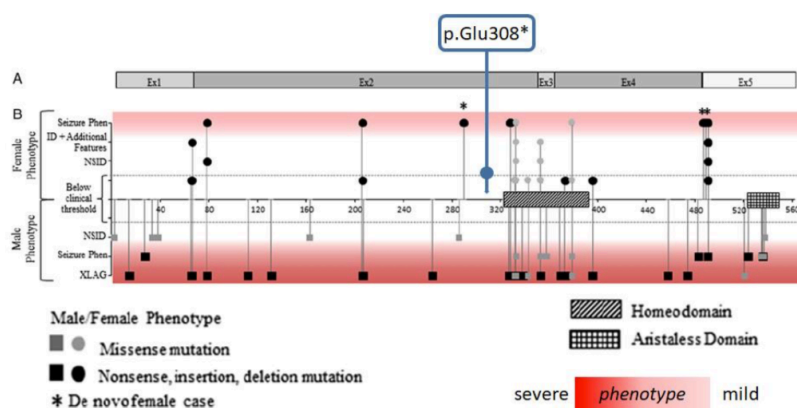


Figure 48. Identified ARX mutations in females and males leading to a range of phenotypes (modified from Mattiske et al., 2017).

The nonsense p.Glu308* variant leads to the production of a truncated protein, lacking in the homeobox domain, and is predicted to be very deleterious using CADD scoring system. It is not reported in gnomAD; it was recently annotated in ClinVar (ID 522170) as pathogenic. Segregation analysis disclosed the presence of the same variant also in the fetus of a previous pregnancy with a more severe phenotype, which could be supposedly due to the concomitant presence of a *de novo* microduplication of a region on chromosome 17. The presence of the same mutation on both the affected fetus but not in the healthy parents suggests a gonadal or gonosomal mosaicism in one of the parents. The identification of this variant was important for genetic counselling as there is an increased recurrence risk for the couple to have a child with the same disorder. The identification of the causative mutation is of importance also for the proband not only for clinical prognosis but also as it allows to properly calculate the risk to transmit the mutation, which is associated with different clinical outcomes depending on the sex.

The family D is composed of two generations: the proband (II:2) is a fetus with corpus callosum agenesis and other severe malformations; a previous pregnancy was interrupted because of a male

fetus with the Dandy-Walker syndrome (OMIM %220200) and hydrocephalous.

To investigate the molecular bases of the phenotype in this family, we performed WES of the fetus and we identified a homozygous variant in *FKTN* gene (NM_006731.2: c.898G>A; NP_006722.2: p.Gly300Arg; rs909129168).

FKTN gene is ubiquitously expressed, but in particular in the central nervous system. This gene encodes a putative transmembrane protein that is localized to the cis-Golgi compartment, where it may be involved in the glycosylation of α -dystroglycan in skeletal muscle (Figure 49); it is a glycosyltransferase with a role also in brain development.

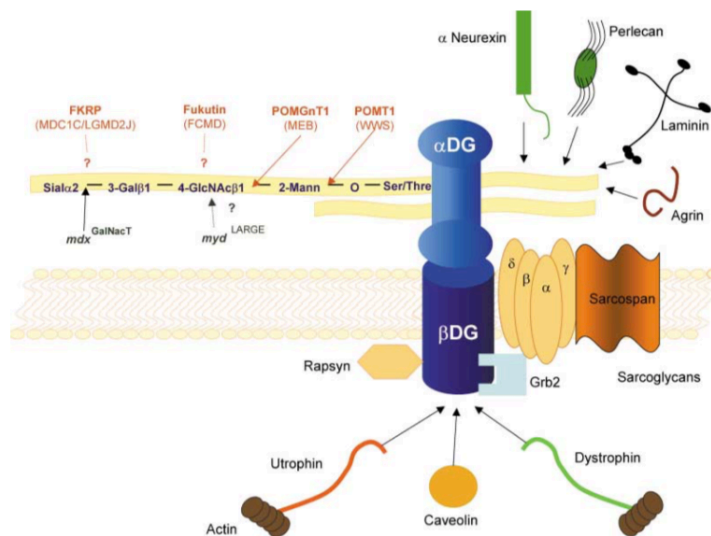


Figure 49. O-mannosyl-Linked Glycosylation of DG in a Complex of Dystrophin-Associated Proteins (Montanaro and Carbonetto, 2003).

The dystrophin-glycoprotein complex (DGC) is a multisubunit complex that connects the extracellular matrix components (ECM) to the cytoskeletal matrix of muscle fiber cells and maintains muscle integrity. Mutations in this complex are associated with muscular dystrophy. Although the role of dystroglycan has been explored mainly in the context of muscle, recently a novel role for dystroglycan inside the CNS has been demonstrated and thus provides potential insights into the brain abnormalities associated with some forms of muscular dystrophy, as the α -Dystroglycanopathies (α -DGP), a group of muscular dystrophy characterized by abnormal glycosylation of α -dystroglycan (α -DG) (Montanaro and Carbonetto, 2003).

Secondary dystroglycanopathies are caused by mutations in different genes, coding for putative or actual glycosyltransferases: *POMT1*, *POMT2*, *POMGNT1*, *LARGE*, *FKTN* and genes coding for fukutin-related proteins (FKRPs). Irrespective of the gene defect, the characteristic and diagnostic feature of all of these conditions is hypoglycosylation of α -dystroglycan, which led to the term

hindrance effects that abolish the activation of the sugar moiety. The latter activity is required for the glycosylation of α -dystroglycan in skeletal muscle. Thus, Gly300Arg is likely responsible for a reduced activation of the sugar moieties, which in turn impairs transfer to α -dystroglycan.

Concurrently to whole exome data analysis, a fetal MRI (fMRI) performed at 18 weeks showed a marked dilatation of the ventricular system and a diffuse compression on cerebral parenchyma. The corpus callosum was absent and the cerebellum was markedly hypoplastic, with a severe reduction of the vermis. Differently from what was suggested by ultrasound analysis, the association of all these anomalies was consistent with the diagnosis of a more severe neurological phenotype, compatible with Muscular Dystrophy-Dystroglycanopathy Type A, confirming molecular results. The couple decided to terminate the pregnancy: the fetopsy confirmed the diagnosis.

The identification of this variant was important for genetic counselling as it allowed to properly redefine the clinical diagnosis, with implications on recurrence risk for the couple and on reproductive choices.

6. CONCLUSIONS

The main purpose of this PhD project was to find the most appropriate NGS technology and approach to study the molecular bases of different genetic diseases.

The different choices regarding the NGS approach and the number of individuals to sequence were therefore dictated by the pedigree structure and the specific phenotype.

These results demonstrate how the proper and accurate definition of the phenotype and the evaluation of diagnostic potential, feasibility and cost of each NGS approach can result in the efficient identification of new variants/genes underlying rare phenotypes. Applying NGS technologies also in the clinics will improve the diagnosis, the genetic counselling and, potentially, at least in some cases, the treatment of genetic diseases.

The lack of a diagnosis can have considerable adverse effects for patients and their families: failure to identify potential treatments, failure to recognize the risk of recurrence in subsequent pregnancies and failure to provide anticipatory guidance and prognosis. NGS may improve patient health outcomes and facilitate the more efficient use of health-care resources: the “diagnostic odyssey”, to which patients are often subjected without receiving a diagnosis, has implications for societal medical expenditures, with unsuccessful attempts consuming limited resources. The application of NGS to clinical diagnosis raises a lot of challenges, but it is revolutionizing medical genetics and leading us to the personalized medicine.

BIBLIOGRAPHY

- Aboitiz F, Montiel J (2003). One hundred million years of interhemispheric communication: the history of the corpus callosum. *Braz J Med Biol Res.* 36 (4): 409-420.
- Alfawaz S, Fong F, Plagnol V, Wong FS, Fearne J, Kellsell DP (2013). Recessive oligodontia linked to a homozygous loss-of-function mutation in the SMOC2 gene. *Arch Oral Biol.* 58 (5): 462-466.
- Al-Qattan MM (2014). Embryology of familial (non-syndromic) brachydactyly of the hand. *J Hand Surg Eur Vol.* 39 (9): 926-933.
- Al-Qattan MM, Al-Motairi MI, Al Balwi MA (2015). Two novel homozygous missense mutations in the GDF5 gene cause brachydactyly type C. *Am J Med Genet A.* 167 (7): 1621-1626.
- Al-Yahyaee SA, Al-Kindi MN, Habbal O, Kumar DS (2003). Clinical and molecular analysis of Grebe acromesomelic dysplasia in an Omani family. *Am J Med Genet A.* 121A (1): 9-14.
- Amarasinghe KC, Li J, Halgamuge SK (2013). CoNVEX: copy number variation estimation in exome sequencing data using HMM. *BMC Bioinformatics.* 14 Suppl 2: S2.
- Antonarakis SE, Beckmann JS (2006). Mendelian disorders deserve more attention. *Nat Rev Genet.* 7 (4): 277-282.
- Arnold SJ, Stappert J, Bauer A, Kispert A, Herrmann BG, Kemler R (2000). Brachyury is a target gene of the Wnt/beta-catenin signaling pathway. *Mech Dev.* 91 (1-2): 249-258.
- Arte S, Parmanen S, Pirinen S, Alaluusua S, Nieminen P (2013). Candidate gene analysis of tooth agenesis identifies novel mutations in six genes and suggests significant role for WNT and EDA signaling and allele combinations. *PLoS One.* 8 (8): e73705.
- Backenroth D, Homsy J, Murillo LR, Glessner J, Lin E, Brueckner M, Lifton R, Goldmuntz E, Chung WK, Shen Y (2014). CANOES: detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Res.* 42 (12): e97.
- Barbato E, Traversa A, Guarnieri R, Giovannetti A, Genovesi ML, Magliozzi MR, Paolacci S, Ciolfi A, Pizzi S, Di Giorgio R, Tartaglia M, Pizzuti A, Caputo V (2018). Whole exome sequencing in an Italian family with isolated maxillary canine agenesis and canine eruption anomalies. *Arch Oral Biol.* 91: 96-102.
- Basit S, Naqvi SK, Wasif N, Ali G, Ansar M, Ahmad W (2008). A novel insertion mutation in the cartilage-derived morphogenetic protein-1 (CDMP1) gene underlies Grebe-type chondrodysplasia in a consanguineous Pakistani family. *BMC Med Genet.* 9:102.
- Becker A, Chaushu S (2015). Etiology of maxillary canine impaction: a review. *Am J Orthod Dentofacial Orthop.* 148 (4): 557-567.
- Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, Shang L, Boisson B, Casanova JL, Abel L (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for

- detecting exome variants. *Proc Natl Acad Sci U S A*. 112 (17): 5473-5478.
- Berg JS, Khoury MJ, Evans JP (2011). Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time. *Genet Med*. 13 (6): 499-504.
- Bergendal B, Klar J, Stecksén-Blicks C, Norderyd J, Dahl N (2011). Isolated oligodontia associated with mutations in EDARADD, AXIN2, MSX1, and PAX9 genes. *Am J Med Genet A*. 155A (7): 1616-1622.
- Bettella E, Di Rosa G, Polli R, Leonardi E, Tortorella G, Sartori S, Murgia A (2013). Early-onset epileptic encephalopathy in a girl carrying a truncating mutation of the ARX gene: rethinking the ARX phenotype in females. *Clin Genet*. 84 (1): 82-85.
- Boycott KM, Rath A, Chong JX, Hartley T, Alkuraya FS, Baynam G, Brookes AJ, Brudno M, Carracedo A, den Dunnen JT, Dyke SOM, Estivill X, Goldblatt J, Gonthier C, Groft SC, Gut I, Hamosh A, Hieter P, Höhn S, Hurles ME, Kaufmann P, Knoppers BM, Krischer JP, Macek M Jr, Matthijs G, Olry A, Parker S, Paschall J, Philippakis AA, Rehm HL, Robinson PN, Sham PC, Stefanov R, Taruscio D, Unni D, Vanstone MR, Zhang F, Brunner H, Bamshad MJ, Lochmüller H (2017). International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *Am J Hum Genet*. 100 (5): 695-705.
- Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet*. 14 (10): 681-691.
- Byrnes AM, Racacho L, Nikkel SM, Xiao F, MacDonald H, Underhill TM, Bulman DE (2010). Mutations in GDF5 presenting as semidominant brachydactyly A1. *Hum Mutat*. 31 (10): 1155-1162.
- Caspar SM, Dubacher N, Kopps AM, Meienberg J, Henggeler C, Matyas G (2018). Clinical sequencing: From raw data to diagnosis with lifetime value. *Clin Genet*. 93 (3): 508-519.
- Chakravorty S, Hegde M (2017). Gene and Variant Annotation for Mendelian Disorders in the Era of Advanced Sequencing Technologies. *Annu Rev Genomics Hum Genet*. 18: 229-256.
- Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, Harrell TM, McMillin MJ, Wiszniewski W, Gambin T, Coban Akdemir ZH, Doheny K, Scott AF, Avramopoulos D, Chakravarti A, Hoover-Fong J, Mathews D, Witmer PD, Ling , Hetrick K, Watkins L, Patterson KE, Reinier F, Blue E, Muzny , Kircher M, Bilguvar K, López-Giráldez F, Sutton VR, Tabor HK, Leal SM, Gunel M, Mane S, Gibbs RA, Boerwinkle , Hamosh A, Shendure J, Lupski JR, Lifton RP, Valle D, Nickerson DA; Centers for Mendelian Genomics, Bamshad MJ (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet*. 97 (2): 199-215.
- Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*. 6 (2): 80-92.
- Colasante G, Sessa A, Crispi S, Calogero R, Mansouri A, Collombat P, Broccoli V (2009). Arx acts as a regional key selector gene in the ventral telencephalon mainly through its transcriptional repression activity. *Dev Biol*. 334 (1): 59-71.

- Colombo E, Collombat P, Colasante G, Bianchi M, Long J, Mansouri A, Rubenstein JL, Broccoli V (2007). Inactivation of *Arx*, the murine ortholog of the X-linked lissencephaly with ambiguous genitalia gene, leads to severe disorganization of the ventral telencephalon with impaired neuronal migration and differentiation. *J Neurosci.* 27 (17): 4786-4798.
- Costa T, Ramsby G, Cassia F, Peters KR, Soares J, Correa J, Quelce-Salgado A, Tsipouras P (1998). Grebe syndrome: clinical and radiographic findings in affected individuals and heterozygous carriers. *Am J Med Genet.* 75 (5): 523-529.
- Craven I, Bradburn MJ, Griffiths PD (2015). Antenatal diagnosis of agenesis of the corpus callosum. *Clin Radiol.* 70 (3): 248-253.
- D'Antonio F, Pagani G, Familiari A, Khalil A, Sagies TL, Malinger G, Leibovitz Z, Garel C, Moutard ML, Pilu G, Bhide A, Acharya G, Leombroni M, Manzoli L, Papageorghiou A, Prefumo F (2016). Outcomes Associated With Isolated Agenesis of the Corpus Callosum: A Meta-analysis. *Pediatrics.* 138 (3).
- Dawson K, Seeman P, Sebald E, King L, Edwards M, Williams J 3rd, Mundlos S, Krakow D (2006). *GDF5* is a second locus for multiple-synostosis syndrome. *Am J Hum Genet.* 78 (4): 708-712.
- Decker E, Stellzig-Eisenhauer A, Fiebig BS, Rau C, Kress W, Saar K, Rüschenhoff F, Hubner N, Grimm T, Weber BH (2008). *PTHR1* loss-of-function mutations in familial, nonsyndromic primary failure of tooth eruption. *Am J Hum Genet.* 83 (6): 781-786.
- de Koning TJ, Jongbloed JD, Sikkema-Raddatz B, Sinke RJ (2015). Targeted next-generation sequencing panels for monogenetic disorders in clinical diagnostics: the opportunities and challenges. *Expert Rev Mol Diagn.* 15 (1): 61-70.
- DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 43 (5): 491-498.
- Di Resta C, Galbiati S, Carrera P, Ferrari M (2018). Next-generation sequencing approach for the diagnosis of human diseases: open challenges and new opportunities. *EJIFCC.* 29 (1): 4-14.
- Douzgou S, Lehmann K, Mingarelli R, Mundlos S, Dallapiccola B (2008). Compound heterozygosity for *GDF5* in Du Pan type chondrodysplasia. *Am J Med Genet A.* 146A (16): 2116-2121.
- Edwards TJ, Sherr EH, Barkovich AJ, Richards LJ (2014). Clinical, genetic and imaging findings identify new causes for corpus callosum development syndromes. *Brain.* 137 (Pt 6): 1579-1613.
- Ekşioğlu YZ, Pong AW, Takeoka M (2011). A novel mutation in the aristaless domain of the *ARX* gene leads to Ohtahara syndrome, global developmental delay, and ambiguous genitalia in males and neuropsychiatric disorders in females. *Epilepsia.* 52 (5): 984-992.
- Ericson S, Kurol J (1986). Radiographic assessment of maxillary canine eruption in children with clinical signs of eruption disturbance. *Eur J Orthod.* 8 (3): 133-140.

- Everman DB, Bartels CF, Yang Y, Yanamandra N, Goodman FR, Mendoza-Londono JR, Savarirayan R, White SM, Graham JM Jr, Gale RP, Svarch E, Newman WG, Kleckers AR, Francomano CA, Govindaiah V, Singh L, Morrison S, Thomas JT, Warman ML (2002). The mutational spectrum of brachydactyly type C. *Am J Med Genet.* 112 (3): 291-296.
- Faiyaz-Ul-Haque M, Ahmad W, Wahab A, Haque S, Azim AC, Zaidi SH, Teebi AS, Ahmad M, Cohn DH, Siddique T, Tsui LC (2002a). Frameshift mutation in the cartilage-derived morphogenetic protein 1 (CDMP1) gene and severe acromesomelic chondrodysplasia resembling Grebe-type chondrodysplasia. *Am J Med Genet.* 111 (1): 31-37.
- Faiyaz-Ul-Haque M, Ahmad W, Zaidi SH, Haque S, Teebi AS, Ahmad M, Cohn DH, Tsui LC (2002b). Mutation in the cartilage-derived morphogenetic protein-1 (CDMP1) gene in a kindred affected with fibular hypoplasia and complex brachydactyly (DuPan syndrome). *Clin Genet.* 61 (6): 454-458.
- Faiyaz-Ul-Haque M, Faqeih EA, Al-Zaidan H, Al-Shammary A, Zaidi SH (2008). Grebe-type chondrodysplasia: a novel missense mutation in a conserved cysteine of the growth differentiation factor 5. *J Bone Miner Metab.* 26 (6): 648-652.
- Friocourt G, Kanatani S, Tabata H, Yozu M, Takahashi T, Antypa M, Raguénès O, Chelly J, Férec C, Nakajima K, Parnavelas JG (2008). Cell-autonomous roles of ARX in cell proliferation and neuronal migration during corticogenesis. *J Neurosci.* 28 (22): 5794-5805.
- Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, Handsaker RE, McCarroll S a, O'Donovan MC, Owen MJ, Kirov G, Sullivan PF, Hultman CM, Sklar P, Purcell SM (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet.* 91 (4): 597-607.
- Fullston T, Finnis M, Hackett A, Hodgson B, Brueton L, Baynam G, Norman A, Reish O, Shoubridge C, Gecz J (2011). Screening and cell-based assessment of mutations in the Aristaless-related homeobox (ARX) gene. *Clin Genet.* 80 (6): 510-522.
- Galjaard RJ, van der Ham LI, Posch NA, Dijkstra PF, Oostra BA, Hovius SE, Timmenga EJ, Sonneveld GJ, Hoogeboom AJ, Heutink P (2001). Differences in complexity of isolated brachydactyly type C cannot be attributed to locus heterogeneity alone. *Am J Med Genet.* 98 (3): 256-262.
- Ghebranious N, Blank RD, Raggio CL, Staubli J, McPherson E, Ivacic L, Rasmussen K, Jacobsen FS, Faciszewski T, Burmester JK, Pauli RM, Boachie-Adjei O, Glurich I, Giampietro PF (2008). A missense T (Brachyury) mutation contributes to vertebral malformations. *J Bone Miner Res.* 23 (10): 1576-1583.
- Gilissen C, Hoischen A, Brunner HG, Veltman JA (2012). Disease gene identification strategies for exome sequencing. *Eur J Hum Genet.* 20 (5): 490-497.
- Giunta C, Nuytinck L, Raghunath M, Hausser I, De Paepe A, Steinmann B (2002). Homozygous Gly530Ser substitution in COL5A1 causes mild classical Ehlers-Danlos syndrome. *Am J Med Genet.* 109 (4): 284-290.
- Godfrey C, Clement E, Mein R, Brockington M, Smith J, Talim B, Straub V, Robb S, Quinlivan R, Feng L, Jimenez-Mallebrera C, Mercuri E, Manzur AY, Kinali M, Torelli S, Brown SC, Sewry CA,

- Bushby K, Topaloglu H, North K, Abbs S, Muntoni F (2007). Refining genotype phenotype correlations in muscular dystrophies with defective glycosylation of dystroglycan. *Brain*. 130 (Pt 10): 2725-2735.
- Goodwin S, McPherson JD, McCombie WR (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 17 (6): 333-351.
- Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, McGuire AL, Nussbaum RL, O'Daniel JM, Ormond KE, Rehm HL, Watson MS, Williams MS, Biesecker LG; American College of Medical Genetics and Genomics (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med*. 15 (7): 565-574.
- Herrmann BG (1992). Action of the Brachyury gene in mouse embryogenesis. *Ciba Found Symp*; 165: 78-86; discussion 86-91.
- Hoischen A, van Bon BW, Gilissen C, Arts P, van Lier B, Steehouwer M, de Vries P, de Reuver R, Wieskamp N, Mortier G, Devriendt K, Amorim MZ, Revencu N, Kidd A, Barbosa M, Turner A, Smith J, Oley C, Henderson A, Hayes IM, Thompson EM, Brunner HG, de Vries BB, Veltman JA (2010). *De novo* mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat Genet*. 42 (6): 483-485.
- Ishii Y, Wajid M, Bazzi H, Fantauzzo KA, Barber AG, Blaydon DC, Nam JS, Yoon JK, Kelsell DP, Christiano AM (2008). Mutations in R-spondin 4 (RSPO4) underlie inherited anonychia. *J Invest Dermatol*. 128 (4): 867-870.
- Jamuar SS, Tan EC (2015). Clinical application of next-generation sequencing for Mendelian diseases. *Hum Genomics*. 9:10.
- Jiang Y, Oldridge DA, Diskin SJ, Zhang NR (2015). CODEX: A normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res*. 43 (6): e39.
- Kantaputra PN, Kaewgahya M, Hatsadaloi A, Vogel P, Kawasaki K, Ohazama A, Ketudat Cairns JR (2015). GREMLIN 2 Mutations and Dental Anomalies. *J Dent Res*. 94 (12): 1646-1652.
- Kantaputra P, Kaewgahya M, Kantaputra W (2014). WNT10A mutations also associated with agenesis of the maxillary permanent canines, a separate entity. *Am J Med Genet A*. 164A (2): 360-363.
- Kantaputra P, Sripathomsawat W (2011). WNT10A and isolated hypodontia. *Am J Med Genet A*. 155A (5): 1119-1122.
- Kato M, Das S, Petras K, Kitamura K, Morohashi K, Abuelo DN, Barr M, Bonneau D, Brady AF, Carpenter NJ, Ciperio KL, Frisone F, Fukuda T, Guerrini R, Iida E, Itoh M, Lewanda AF, Nanba Y, Oka A, Proud VK, Saugier-veber P, Schelley SL, Selicorni A, Shaner R, Silengo M, Stewart F, Sugiyama N, Toyama J, Toutain A, Vargas AL, Yanazawa M, Zackai EH, Dobyns WB (2004). Mutations of ARX are associated with striking pleiotropy and consistent genotype-phenotype correlation. *Hum Mutat*. 23 (2): 147-159.
- Kebschull JM, Zador AM (2015). Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res*. 43 (21): e143.

- Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet.* 46 (3): 310-315.
- Kispert A, Koschorz B, Herrmann BG (1995). The T protein encoded by Brachyury is a tissue-specific transcription factor. *EMBO J.* 14 (19): 4763-4772.
- Kitamura K, Yanazawa M, Sugiyama N, Miura H, Iizuka-Kogo A, Kusaka M, Omichi K, Suzuki R, Kato-Fukui Y, Kamiirisa K, Matsuo M, Kamijo S, Kasahara M, Yoshioka H, Ogata T, Fukuda T, Kondo I, Kato M, Dobyns WB, Yokoyama M, Morohashi K (2002). Mutation of ARX causes abnormal development of forebrain and testes in mice and X-linked lissencephaly with abnormal genitalia in humans. *Nat Genet.* 32 (3): 359-369.
- Kjaer KW, Eiberg H, Hansen L, van der Hagen CB, Rosendahl K, Tommerup N, Mundlos S (2006). A mutation in the receptor binding site of GDF5 causes Mohr-Wriedt brachydactyly type A2. *J Med Genet.* 43 (3): 225-231.
- Klambauer G, Schwarzbauer K, Mayr A, Clevert D-A, Mitterecker A, Bodenhofer U, Hochreiter S (2012). cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* 40 (9): e69.
- Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, Quinlan AR, Nickerson D a, Eichler EE (2012). Copy number variation detection and genotyping from exome sequence data. *Genome Res.* 22 (8): 1525-1532.
- Kwong AK, Ho AC, Fung CW, Wong VC (2015). Analysis of mutations in 7 genes associated with neuronal excitability and synaptic transmission in a cohort of children with non-syndromic infantile epileptic encephalopathy. *PLoS One.* 10 (5): e0126446.
- Lee H, Deignan JL, Dorrani N, Strom SP, Kantarci S, Quintero-Rivera F, Das K, Toy T, Harry B, Yourshaw M, Fox M, Fogel BL, Martinez-Agosto JA, Wong DA, Chang VY, Shieh PB, Palmer CG, Dipple KM, Grody WW, Vilain E, Nelson SF (2014). Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA.* 312 (18): 1880-1887.
- Lee K, Mattiske T, Kitamura K, Gecz J, Shoubridge C (2014). Reduced polyalanine-expanded Arx mutant protein in developing mouse subpallium alters Lmo1 transcriptional regulation. *Hum Mol Genet.* 23 (4): 1084-1094.
- Lelieveld SH, Spielmann M, Mundlos S, Veltman JA, Gilissen C (2015). Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Hum Mutat.* 36 (8): 815-822.
- Leonardi R, Barbato E, Vichi M, Caltabiano M (2009). Skeletal anomalies and normal variants in patients with palatally displaced canines. *Angle Orthod.* 79 (4): 727-732.
- Leonardi R, Peck S, Caltabiano M, Barbato E (2003). Palatally displaced canine anomaly in monozygotic twins. *Angle Orthod.* 73 (4): 466-470.
- Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 25 (14): 1754-1760.
- Lombardo C, Barbato E, Leonardi R (2007). Bilateral maxillary canines agenesis: a case report and a literature review. *Eur J Paediatr Dent.* 8 (1): 38-41.

- Love MI, Myšičková A, Sun R, Kalscheuer V, Vingron M, Haas SA (2011). Modeling Read Counts for CNV Detection in Exome Sequencing Data. *Stat Appl Genet Mol Biol.* 10 (1).
- Magi A, Tattini L, Cifola I, D'Aurizio R, Benelli M, Mangano E, Battaglia C, Bonora E, Kurg A, Seri M, Magini P, Giusti B, Romeo G, Pippucci T, De Bellis G, Abbate R, Gensini GF (2013). EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol.* 14 (10): R120.
- Marsh E, Fulp C, Gomez E, Nasrallah I, Minarcik J, Sudi J, Christian SL, Mancini G, Labosky P, Dobyns W, Brooks-Kayal A, Golden JA (2009). Targeted loss of Arx results in a developmental epilepsy mouse model and recapitulates the human phenotype in heterozygous females. *Brain.* 132 (Pt 6): 1563-1576.
- Martinez-Garcia M, Garcia-Canto E, Fenollar-Cortes M, Aytes AP, Trujillo-Tiebas MJ (2016). Characterization of an acromesomelic dysplasia, Grebe type case: novel mutation affecting the recognition motif at the processing site of GDF5. *J Bone Miner Metab.* 34 (5): 599-603.
- Massink MP, Créton MA, Spanevello F, Fennis WM, Cune MS, Savelberg SM, Nijman IJ, Maurice MM, van den Boogaard MJ, van Haaften G (2015). Loss-of-Function Mutations in the WNT Co-receptor LRP6 Cause Autosomal-Dominant Oligodontia. *Am J Hum Genet.* 97 (4): 621-626.
- Mattiske T, Moey C, Vissers LE, Thorne N, Georgeson P, Bakshi M, Shoubridge C (2017). An Emerging Female Phenotype with Loss-of-Function Mutations in the Aristaless-Related Homeodomain Transcription Factor ARX. *Hum Mutat.* 38 (5): 548-555.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20 (9): 1297-1303.
- Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M (2010). Detecting copy number variation with mated short reads. *Genome Res.* 20 (11): 1613-1622.
- Mercuri E, Cassetta M, Cavallini C, Vicari D, Leonardi R, Barbato E (2013). Dental anomalies and clinical features in patients with maxillary canine impaction. *Angle Orthod.* 83 (1): 22-28.
- Meynert AM, Bicknell LS, Hurler ME, Jackson AP, Taylor MS (2013). Quantifying single nucleotide variant detection sensitivity in exome sequencing. *BMC Bioinformatics.* 14: 195.
- Montanaro F, Carbonetto S (2003). Targeting dystroglycan in the brain. *Neuron.* 37 (2): 193-196.
- Mu Y, Xu Z, Contreras CI, McDaniel JS, Donly KJ, Chen S (2012). Phenotype characterization and sequence analysis of BMP2 and BMP4 variants in two Mexican families with oligodontia. *Genet Mol Res.* 11 (4): 4110-4120.
- Mumtaz S, Riaz HF, Touseef M, Basit S, Faiyaz UI Haque M, Malik S (2015). Recurrent mutation in CDMP1 in a family with Grebe chondrodysplasia: broadening the phenotypic manifestation of syndrome in Pakistani population. *Pak J Med Sci.* 31 (6): 1542-1544.
- Mundlos S (2009). The brachydactylies: a molecular disease family. *Clin Genet.* 76 (2): 123-136.
- Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ (2010). Exome sequencing identifies the cause of a

mendelian disorder. *Nat Genet.* 42 (1): 30-35.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature.* 461 (7261): 272-276.

Nieminen P (2009). Genetic basis of tooth agenesis. *J Exp Zool B Mol Dev Evol.* 312B (4): 320-342.

Ockeloen CW, Khandelwal KD, Dreesen K, Ludwig KU, Sullivan R, van Rooij IALM, Thonissen M, Swinnen S, Phan M, Conte F, Ishorst N, Gilissen C, RoaFuentes L, van de Vorst M, Henkes A, Steehouwer M, van Beusekom E, Bloemen M, Vankeirsbilck B, Bergé S, Hens G, Schoenaers J, Poorten VV, Roosenboom J, Verdonck A, Devriendt K, Roeleveldt N, Jhangiani SN, Vissers LELM, Lupski JR, de Ligt J, Von den Hoff JW, Pfundt R, Brunner HG, Zhou H, Dixon J, Mangold E, van Bokhoven H, Dixon MJ, Kleefstra T, Hoischen A, Carels CEL (2016). Novel mutations in LRP6 highlight the role of WNT signaling in tooth agenesis. *Genet Med.* 18 (11): 1158-1162.

Ohira R, Zhang YH, Guo W, Dipple K, Shih SL, Doerr J, Huang BL, Fu LJ, Abu-Khalil A, Geschwind D, McCabe ER (2002). Human ARX gene: genomic characterization and expression. *Mol Genet Metab.* 77 (1-2): 179-188.

Ozyüncü O, Yazıcıoğlu A, Turğal M (2014). Antenatal diagnosis and outcome of agenesis of corpus callosum: A retrospective review of 33 cases. *J Turk Ger Gynecol Assoc.* 15 (1): 18-21.

Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform.* 15 (2): 256-278.

Palmer EE, Mowat D (2014). Agenesis of the corpus callosum: a clinical approach to diagnosis. *Am J Med Genet C Semin Med Genet.* 166C (2): 184-197.

Peck S, Peck L, Kataja M (1994). The palatally displaced canine as a dental anomaly of genetic origin. *Angle Orthod.* 64 (4): 249-256. Review.

Pemberton TJ, Li FY, Oka S, Mendoza-Fandino GA, Hsu YH, Bringas P Jr, Chai Y, Snead ML, Mehrian-Shai R, Patel PI (2007). Identification of novel genes expressed during mouse tooth development by microarray gene expression analysis. *Dev Dyn.* 236 (8): 2245-2257.

Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, Wood NW, Hambleton S, Burns SO, Thrasher AJ, Kumararatne D, Doffinger R, Nejentsev S (2012). A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics.* 28 (21): 2747-2754.

Plöger F, Seemann P, Schmidt-von Kegler M, Lehmann K, Seidel J, Kjaer KW, Pohl J, Mundlos S (2008). Brachydactyly type A2 associated with a defect in proGDF5 processing. *Hum Mol Genet.* 17 (9): 1222-1233.

Polinkovsky A, Robin NH, Thomas JT, Irons M, Lynn A, Goodman FR, Reardon W, Kant SG, Brunner HG, van der Burgt I, Chitayat D, McGaughan J, Donnai D, Luyten FP, Warman ML (1997). Mutations in CDMP1 cause autosomal dominant brachydactyly type C. *Nat Genet.* 17 (1): 18-19.

Prasad MK, Geoffroy V, Vicaire S, Jost B, Dumas M, Le Gras S, Switala M, Gasse B, Laugel-Haushalter V, Paschaki M, Leheup B, Droz D, Dalstein A, Loing A, Grollemund B, Muller-Bolla M, Lopez-Cazaux S, Minoux M, Jung S, Obry F, Vogt V, Davideau JL, Davit-Beal T, Kaiser AS, Moog U, Richard B, Morrier JJ, Duprez JP, Odent S, Bailleul-Forestier I, Rousset MM, Merametdijan L, Toutain A, Joseph C, Giuliano F, Dahlet JC, Courval A, El Alloussi M, Laouina S, Soskin S, Guffon N, Dieux A, Doray B, Feierabend S, Ginglinger E, Fournier B, de la Dure Molla M, Alembik Y, Tardieu C, Clauss F, Berdal A, Stoetzel C, Manière MC, Dollfus H, Bloch-Zupan A (2016). A targeted next-generation sequencing assay for the molecular diagnosis of genetic disorders with orodental involvement. *J Med Genet.* 53 (2): 98-110.

Rabbani B, Tekin M, Mahdieh N (2014). The promise of whole-exome sequencing in medical genetics. *J Hum Genet.* 59 (1): 5-15.

Rakhshan V (2015). Congenitally missing teeth (hypodontia): A review of the literature concerning the etiology, prevalence, risk factors, patterns and treatment. *Dent Res J (Isfahan).* 12 (1): 1-13.

Rehm HL (2013). Disease-targeted sequencing: a cornerstone in the clinic. *Nat Rev Genet.* 14 (4): 295-300.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL; ACMG Laboratory Quality Assurance Committee (2015). 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 (5): 405-424.

Riisager M, Duno M, Hansen FJ, Krag TO, Vissing CR, Vissing J (2013). A new mutation of the fukutin gene causing late-onset limb girdle muscular dystrophy. *Neuromuscul Disord.* 23 (7): 562-567.

Rutledge MS, Hartsfield JK Jr (2010). Genetic factors in the etiology of palatally displaced canines. *Seminars in Orthodontics.* 16 (3), 165–171.

Sajnani AK (2015). Permanent maxillary canines- review of eruption pattern and local etiological factors leading to impaction. *J Investig Clin Dent.* 6 (1): 1-7.

Sajnani AK, King NM (2012). The sequential hypothesis of impaction of maxillary canine - a hypothesis based on clinical and radiographic findings. *J Craniomaxillofac Surg.* 40 (8): e375-85.

Salgado D, Bellgard MI, Desvignes JP, Bérout C (2016). How to Identify Pathogenic Mutations among All Those Variations: Variant Annotation and Filtration in the Genome Sequencing Era. *Hum Mutat.* 37 (12): 1272-1282.

Salvi A, Giacomuzzi E, Bardellini E, Amadori F, Ferrari L, De Petro G, Borsani G, Majorana A (2016). Mutation analysis by direct and whole exome sequencing in familial and sporadic tooth agenesis. *Int J Mol Med.* 38 (5): 1338-1348.

Sathirapongsasuti JF, Lee H, Horst B a J, Brunner G, Cochran AJ, Binder S, Quackenbush J, Nelson SF (2011). Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics.* 27 (19): 2648-2654.

Savarirayan R, White SM, Goodman FR, Graham JM Jr, Delatycki MB, Lachman RS, Rimoin DL, Everman DB, Warman ML (2003). Broad phenotypic spectrum caused by an identical heterozygous CDMP-1 mutation in three unrelated families. *Am J Med Genet A*. 117A (2): 136-142.

Scheffer IE, Wallace RH, Phillips FL, Hewson P, Reardon K, Parasivam G, Stromme P, Berkovic SF, Gecz J, Mulley JC (2002). X-linked myoclonic epilepsy with spasticity and intellectual disability: mutation in the homeobox gene ARX. *Neurology*. 59 (3): 348-356.

Schell-Apacik CC, Wagner K, Bihler M, Ertl-Wagner B, Heinrich U, Klopocki E, Kalscheuer VM, Muenke M, von Voss H (2008). Agenesis and dysgenesis of the corpus callosum: clinical, genetic and neuroimaging findings in a series of 41 patients. *Am J Med Genet A*. 146A (19): 2501-2511.

Schwabe GC, Türkmen S, Leschik G, Palanduz S, Stöver B, Goecke TO, Mundlos S (2004). Brachydactyly type C caused by a homozygous missense mutation in the prodomain of CDMP1. *Am J Med Genet A*. 124A (4): 356-363.

Schwaerzer GK, Hiepen C, Schrewe H, Nickel J, Ploeger F, Sebald W, Mueller T, Knaus P (2012). New insights into the molecular mechanism of multiple synostoses syndrome (SYNS): mutation within the GDF5 knuckle epitope causes noggin-resistance. *J Bone Miner Res*. 27 (2): 429-442.

Seemann P, Brehm A, König J, Reissner C, Stricker S, Kuss P, Haupt J, Renninger S, Nickel J, Sebald W, Groppe JC, Plöger F, Pohl J, Schmidt-von Kegler M, Walther M, Gassner I, Rusu C, Janecke AR, Dathe K, Mundlos S (2009). Mutations in GDF5 reveal a key residue mediating BMP inhibition by NOGGIN. *PLoS Genet*. 5 (11): e1000747.

Seemann P, Schwappacher R, Kjaer KW, Krakow D, Lehmann K, Dawson K, Stricker S, Pohl J, Plöger F, Staub E, Nickel J, Sebald W, Knaus P, Mundlos S (2005). Activating and deactivating mutations in the receptor interaction site of GDF5 cause symphalangism or brachydactyly type A2. *J Clin Invest*. 115 (9): 2373-2381.

Seo SH, Park MJ, Kim SH, Kim OH, Park S, Cho SI, Park SS, Seong MW (2013). Identification of a GDF5 mutation in a Korean patient with brachydactyly type C without foot involvement. *Ann Lab Med*. 33 (2): 150-152.

Sims D, Sudbery I, Illott NE, Heger A, Ponting CP (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet*. 15 (2): 121-132.

Stange K, Ott CE, Schmidt-von Kegler M, Gillesen-Kaesbach G, Mundlos S, Dathe K, Seemann P (2015). Brachydactyly Type C patient with compound heterozygosity for p.Gly319Val and p.Ile358Thr variants in the GDF5 proregion: benign variants or mutations? *J Hum Genet*. 60 (8): 419-425.

Stange K, Thieme T, Hertel K, Kuhfahl S, Janecke AR, Piza-Katzer H, Penttinen M, Hietala M, Dathe K, Mundlos S, Schwarz E, Seemann P (2014). Molecular analysis of two novel missense mutations in the GDF5 proregion that reduce protein activity and are associated with brachydactyly type C. *J Mol Biol*. 426 (19): 3221-3231.

Stavropoulos DJ, Merico D, Jobling R, Bowdin S, Monfared N, Thiruvahindrapuram B, Nalpathamkalam T, Pellecchia G, Yuen RKC, Szego MJ, Hayeems RZ, Shaul RZ, Brudno M, Girdea M, Frey B, Alipanahi B, Ahmed S, Babul-Hirji R, Porras RB, Carter MT, Chad L, Chaudhry

- A, Chitayat D, Doust SJ, Cytrynbaum C, Dupuis L, Ejaz R, Fishman L, Guerin A, Hashemi B, Helal M, Hewson S, Inbar-Feigenberg M, Kannu P, Karp N, Kim R, Kronick J, Liston E, MacDonald H, Mercimek-Mahmutoglu S, Mendoza-Londono R, Nasr E, Nimmo G, Parkinson N, Quercia N, Raiman J, Roifman M, Schulze A, Shugar A, Shuman C, Sinajon P, Siriwardena K, Weksberg R, Yoon G, Carew C, Erickson R, Leach RA, Klein R, Ray PN, Meyn MS, Scherer SW, Cohn RD, Marshall CR (2016). Whole Genome Sequencing Expands Diagnostic Utility and Improves Clinical Management in Pediatric Medicine. *NPJ Genom Med.* 1. pii: 15012.
- Steinmann B, Giunta C (2000). The devil of the one letter code and the Ehlers-Danlos syndrome: corrigendum. *Am J Med Genet.* 93 (4): 342.
- Stockton DW, Das P, Goldenberg M, D'Souza RN, Patel PI (2000). Mutation of PAX9 is associated with oligodontia. *Nat Genet.* 24 (1): 18-19.
- Storm EE, Huynh TV, Copeland NG, Jenkins NA, Kingsley DM, Lee SJ (1994). Limb alterations in brachypodism mice due to mutations in a new member of the TGF beta-superfamily. *Nature.* 368 (6472): 639-643.
- Strømme P, Mangelsdorf ME, Scheffer IE, Gécz J (2002). Infantile spasms, dystonia, and other X-linked phenotypes caused by mutations in Aristaless related homeobox gene, ARX. *Brain Dev.* 24 (5): 266-268.
- Sun Y, Ruivenkamp CA, Hoffer MJ, Vrijenhoek T, Kriek M, van Asperen CJ, den Dunnen JT, Santen GW (2015). Next-generation diagnostics: gene panel, exome, or whole genome? *Hum Mutat.* 36 (6): 648-655.
- Symoens S, Renard M, Bonod-Bidaud C, Syx D, Vaganay E, Malfait F, Ricard-Blum S, Kessler E, Van Laer L, Coucke P, Ruggiero F, De Paepe A (2011). Identification of binding partners interacting with the $\alpha 1$ -N-propeptide of type V collagen. *Biochem J.* 433 (2): 371-381.
- Szatkiewicz JP, Wang W, Sullivan PF, Wang W, Sun W (2013). Improving detection of copy-number variation by simultaneous bias correction and read-depth segmentation. *Nucleic Acids Res.* 41 (3): 1519-1532.
- Szczaluba K, Hilbert K, Obersztyn E, Zabel B, Mazurczak T, Kozłowski K (2005). Du Pan syndrome phenotype caused by heterozygous pathogenic mutations in CDMP1 gene. *Am J Med Genet A.* 138 (4): 379-383.
- Tang R, Wang Q, Du J, Yang P, Wang X (2013). Expression and localization of Nell-1 during murine molar development. *J Mol Histol.* 44 (2): 175-181.
- Tao R, Jin B, Guo SZ, Qing W, Feng GY, Brooks DG, Liu L, Xu J, Li T, Yan Y, He L (2006). A novel missense mutation of the EDA gene in a Mongolian family with congenital hypodontia. *J Hum Genet.* 51 (5): 498-502.
- Temtam SA, Aglan MS (2008). Brachydactyly. *Orphanet J Rare Dis.* 3: 15.
- Thesleff I (2006). The genetic basis of tooth development and dental defects. *Am J Med Genet A.* 140 (23): 2530-2535.
- Thomas JT, Lin K, Nandedkar M, Camargo M, Cervenka J, Luyten FP (1996). A human

chondrodysplasia due to a mutation in a TGF-beta superfamily member. *Nat Genet.* 12 (3): 315-317.

Travieso-Suárez L, Pereda A, Pozo-Román J, Pérez de Nanclares G, Argente J (2018). Brachydactyly type C due to a nonsense mutation in the GDF5 gene. *An Pediatr (Barc).* 88 (2): 107-109.

Ullah A, Umair M, Hussain S, Jan A, Ahmad W (2018). Sequence variants in GDF5 and TRPS1 underlie brachydactyly and tricho-rhino-phalangeal syndrome type III. *Pediatr Int.* 60 (3): 304-306.

Umair M, Rafique A, Ullah A, Ahmad F, Ali RH, Nasir A, Ansar M, Ahmad W (2017). Novel homozygous sequence variants in the GDF5 gene underlie acromesomelic dysplasia type-grebe in consanguineous families. *Congenit Anom (Kyoto).* 57 (2): 45-51.

Uyguner ZO, Kocaoğlu M, Toksoy G, Basaran S, Kayserili H (2014). Novel indel Mutation in the GDF5 Gene Is Associated with Brachydactyly Type C in a Four-Generation Turkish Family. *Mol Syndromol.* 5 (2): 81-86.

Valdes AM, Doherty S, Muir KR, Zhang W, Maciewicz RA, Wheeler M, Arden N, Cooper C, Doherty M (2012). Genetic contribution to radiographic severity in osteoarthritis of the knee. *Ann Rheum Dis.* 71 (9): 1537-1540.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics.* 43: 11.10.1-33.

van der Hout AH, Oudesluijs GG, Venema A, Verheij JB, Mol BG, Rump P, Brunner HG, Vos YJ, van Essen AJ (2008). Mutation screening of the Ectodysplasin-A receptor gene EDAR in hypohidrotic ectodermal dysplasia. *Eur J Hum Genet.* 16 (6): 673-679.

Vastardis H, Karimbux N, Guthua SW, Seidman JG, Seidman CE (1996). A human MSX1 homeodomain missense mutation causes selective tooth agenesis. *Nat Genet.* 13 (4): 417-421.

Waite A, Brown SC, Blake DJ (2012). The dystrophin-glycoprotein complex in brain development and disease. *Trends Neurosci.* 35 (8): 487-496.

Wallerstein R, Sugalski R, Cohn L, Jawetz R, Friez M (2008). Expansion of the ARX spectrum. *Clin Neurol Neurosurg.* 110 (6): 631-634.

Wang K, Li M, Hakonarson H (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38 (16): e164.

Wang X, Xiao F, Yang Q, Liang B, Tang Z, Jiang L, Zhu Q, Chang W, Jiang J, Jiang C, Ren X, Liu JY, Wang QK, Liu M (2006). A novel mutation in GDF5 causes autosomal dominant symphalangism in two Chinese families. *Am J Med Genet A.* 140A (17): 1846-1853.

Weiss MM, Van der Zwaag B, Jongbloed JD, Vogel MJ, Brüggewirth HT, Lekanne Deprez RH, Mook O, Ruivenkamp CA, van Slegtenhorst MA, van den Wijngaard A, Waisfisz Q, Nelen MR, van der Stoep N (2013). Best practice guidelines for the use of next-generation sequencing applications in genome diagnostics: a national collaborative study of Dutch genome diagnostic

laboratories. *Hum Mutat.* 34 (10): 1313-1321.

Wright CF, FitzPatrick DR, Firth HV (2018). Paediatric genomics: diagnosing rare disease in children. *Nat Rev Genet.* 19 (5): 253-268.

Yamaguchi T, Hosomichi K, Yano K, Kim YI, Nakaoka H, Kimura R, Otsuka H, Nonaka N, Haga S, Takahashi M, Shiota T, Kikkawa Y, Yamada A, Kamijo R, Park SB, Nakamura M, Maki K, Inoue I (2017). Comprehensive genetic exploration of selective tooth agenesis of mandibular incisors by exome sequencing. *Hum Genome Var.* 4: 17005.

Yang W, Cao L, Liu W, Jiang L, Sun M, Zhang D, Wang S, Lo WH, Luo Y, Zhang X (2008). Novel point mutations in GDF5 associated with two distinct limb malformations in Chinese: brachydactyly type C and proximal symphalangism. *J Hum Genet.* 53 (4): 368-374.

Yis U, Uyanik G, Heck PB, Smitka M, Nobel H, Ebinger F, Dirik E, Feng L, Kurul SH, Brocke K, Unalp A, Özer E, Cakmakci H, Sewry C, Cirak S, Muntoni F, Hehr U, Morris-Rosendahl DJ (2011). Fukutin mutations in non-Japanese patients with congenital muscular dystrophy: less severe mutations predominate in patients with a non-Walker-Warburg phenotype. *Neuromuscul Disord.* 21 (1): 20-30.

Yu P, Yang W, Han D, Wang X, Guo S, Li J, Li F, Zhang X, Wong SW, Bai B, Liu Y, Du J, Sun ZS, Shi S, Feng H, Cai T (2016). Mutations in WNT10B Are Identified in Individuals with Oligodontia. *Am J Hum Genet.* 99 (1): 195-201.

SITOGRAPHY

Documento Commissione SIGU-NGS (2016): Il sequenziamento del DNA di nuova generazione: indicazioni per l'impiego clinico: <https://www.sigu.net/show/documenti/5/1/linee%20guida?page=1>

HGMD-Human Genome Mutation Database: <http://www.hgmd.cf.ac.uk/ac/>.

HPO-Human Phenotype Ontology: <http://human-phenotype-ontology.github.io>.

MGI-Mouse Genome Informatics: <http://www.informatics.jax.org>.

OMIM-Online Mendelian Inheritance in Man: <https://www.omim.org>.

Picard's MarkDuplicates: <http://broadinstitute.github.io/picard>.

TruSight One Sequencing Panel: <https://www.illumina.com/products/by-type/clinical-research-products/trusight-one.html>.

Veritas Genetics: <https://www.veritasgenetics.com/press-releases>.

ZFIN-Zebrafish Information Network: <https://zfin.org>.