SAPIENZA
UNIVERSITÀ DI ROMA

Analysis of non-coding DNA from Whole Exome Sequencing data

PhD school: Biology and Molecular Medicine

PhD course: Human Biology and Medical Genetics (XXXI cycle)

Curriculum: Medical Genetics

Candidate

Agnese Giovannetti

Advisor

Dr. Viviana Caputo

Coordinator

Prof. Antonio Pizzuti

Academic Year 2017-2018

Table of contents

Abstract

Next Generation Sequencing technologies have completely changed the way to study molecular bases underlying Rare Genetic Diseases (RGDs). Currently, sequencing of the exonic portion of the human genome – the exome (1%) – performed through Whole Exome Sequencing (WES) experiments represents the most used approach to discover molecular mechanisms underlying RGDs. To date, several tools have been developed to analyse and interpret data generated from WES. However, due to both technical and experimental limitations, its diagnostic rate is ~20-30%.

In this context, we evaluated whether WES data contain information on non-coding sequences, focusing on microRNAs (miRNAs). Comparative analysis of capture design and experimental coverage allowed to disclose that in WES data reside information related to miRNA sequences that are efficiently captured by most exome enrichment kits. We therefore analysed WES of a cohort of 259 individuals, including patients affected by several genetic diseases and their unaffected relatives, searching for variants in miRNAs and performing functional annotation. Sanger sequence validation confirms the reliable call of variants mapping in miRNA sequences.

To date, no dedicated tool is available to properly retrieve and analyse miRNAs from WES and WGS data. We therefore developed a tool, "AnnomiR", that allows to systematically analyse miRNA variants and miRNAs, providing functional annotation retrieved from several databases. This tool can be integrated in a standard workflow of analysis for WES and WGS data.

WES data contain a great amount of information that is generally discarded by commonly used workflow of analysis and that should be considered, as it could help in the comprehension of molecular mechanisms underlying RGDs. In this context, systematic study of miRNAs could help elucidating their role as disease-causative and phenotypic modifiers in a wide spectrum of human diseases, allowing to achieve a better characterisation of variability of the human genome related to these non-coding sequences.

1. Introduction

1.1 Rare Genetic Diseases' molecular bases identification: the advent of Next Generation Sequencing

Rare diseases are defined as diseases that affect fewer than 200,000 people in US[1] or less than 1 in 2,000 in Europe[2]. Therefore, these diseases, though individually rare, are collectively common. Rare diseases have been estimated to be 7,000, ~80% of which has genetic causes[3], prevalently, alterations of single genes[4]. Molecular bases of these Rare Genetic Diseases (RGDs), also called monogenic or Mendelian diseases, have been extensively studied, leading to the determination of more than 3,500 disease-gene associations[5].

First successes in the identification of disease-gene associations were obtained through a combination of linkage analysis, positional cloning and sequencing of candidate genes[6,7]. Linkage analysis is based on the observation that genes physically close on a chromosome co-segregate during meiosis[8]. In this approach, the sequencing of several affected individuals and controls from a set of families (or from the same one), using a group of DNA polymorphisms, allows to calculate the probability that two loci are genetically linked[6,8]. The comparison of linked regions obtained, with information on status of affected members is then useful to discriminate between regions presumably containing disease-causative mutations and regions not relevant in the physiopathology of the disease. Linkage analysis often represents the first step for positional cloning[9]. Starting from a previously identified candidate region, positional cloning is used to narrow this genomic region, with the intent to identify gene (or genes) in which disease-causative mutations could rely. Combined approach of linkage analysis followed by positional cloning allowed to identify several disease-gene associations, as in the case of *CFTR* for Cystic Fibrosis[10] (MIM: 219700) and *HTT* for Huntington disease[11] (MIM: 143100). While linkage analysis and positional cloning do not require any functional information on genes associated with RGDs, candidate-genes approach is based on Sanger sequencing of genes that seem to be involved in the disease investigated. These genes can be selected for several reasons: because they resemble genes associated with similar diseases, because their protein products seem to be correlated with the pathophysiology of the disease, or because they are located in a relevant region previously identified with other strategies (e.g. linkage analysis)[7]. Through candidate-gene approach

many disease-causative mutations were discovered as those in *p53* associated with Li-Fraumeni syndrome[12] (MIM: 151623).

However, several factors limit the power of these traditional methods as the availability of a small number of cases, the lack of *a priori* biological information, the reduced penetrance of a mutation, and locus heterogeneity[7,13].

Improvements in DNA sequencing, achieved through the introduction of Next Generation Sequencing (NGS) technologies in 2009, have allowed to overcome these limitations. Sequencing a genomic region of interest with a single-nucleotide resolution in a rapid and cost-effective way, NGS substantially changed the way to study RGDs, accelerating the pace of discovery of molecular bases underlying human diseases. From its first application in medical genetics - that led to the identification of disease-causative mutations in *DHODH* gene in patients affected by Miller syndrome[14] (MIM: 263750) – NGS has allowed to elucidate many other disease-gene associations[4,15].

NGS technologies used to sequence human DNA can sequence a specific panel of genes (targeted sequencing - TS), the coding portion of the human genome – the exome – (Whole Exome Sequencing) or the entire genome (Whole Genome Sequencing). NGS experiments produce a large amount of data, demanding several bioinformatics tools to detect and interpret variations identified. Therefore, one limiting factor in the application of these methods is represented by the analysis and the interpretation of the data, rather than their production. As the amount and the kind of variations identified strictly depend on the sequencing approach, all NGS strategies present advantages and limitations in terms of costs and data analysis. Consequently, the choice of the appropriate method is generally guided by *a priori* knowledge of molecular defects underlying the disease investigated (e.g. known disease-gene associations) and by hypothesis on kind of disease-causative mutations (e.g. sequence variations rather than chromosomal rearrangements).

TS is used to sequence either a panel of genes known (or predicted) to be associated with the investigated disease, or the entire set of genes known to be mutated in Mendelian diseases – the Mendeliome – composed by ~5000 genes[16]. This approach has the great advantage of identifying a small number of variations strictly related to the genes of interest, and of facilitating their interpretation. For these reasons, TS has been revealed a powerful tool at identifying disease-causative mutations in Mendelian cohorts[17]. Limitations of TS reside in

its partial ability to detect clinically relevant sequence variants, especially Copy Number Variants (CNVs); moreover, due to the high increase of disease-gene associations discoveries, gene panels require continuous updates. Therefore, TS may be inconclusive in cases in which disease-causative mutations are not identified, often requiring additional tests[16].

WES experiments perform sequencing of the coding portion of all known genes (1% of the human genome), allowing to overcome some of TS limitations. Conversely from TS, WES may be used not only to detect new disease-causative mutations but even to discover new disease-gene associations. Due to low costs and to the plenty of bioinformatics tools available for data analysis, WES has been widely used in medical genetics, leading to the comprehension of molecular bases of several RGDs[13]. However, several limitations, as the inability to assess non-coding variations, narrow WES successful rate, that is attested to be ~20-30%[16,18].

WGS has the advantage of sequencing the entire genome, allowing to detect all sequence and structural variations. Thus, compared to WES, WGS has a greater successful rate[15,16] even if its use is limited by the high costs and the lack of tools and abilities to deeply analyse and interpret data obtained. With the passing of these limitations, WGS will be widely used in the study of RGDs, leading to a better elucidation of molecular mechanisms underlying human diseases[4].

1.2 Whole Exome Sequencing

Among NGS technologies, WES is currently the most used approach in the study of molecular bases of RGDs[4,15]. Indeed, due to the observation that 85% of disease-causing mutations reside within protein-coding genes[19], and due to the accessible costs, WES has been largely applied in medical genetics, leading to the conversion of this approach from a research tool to a diagnostic one[4,15].

1.2.1 Whole Exome Sequencing experimental procedure

Exome experiments can be performed following different protocols, all requiring fundamental steps of exome capture and sequencing[13,20].

DNA fragmentation represents the starting point of exome capture and it can be performed through chemical, physical or enzymatic methods (Figure 1). Fragments obtained

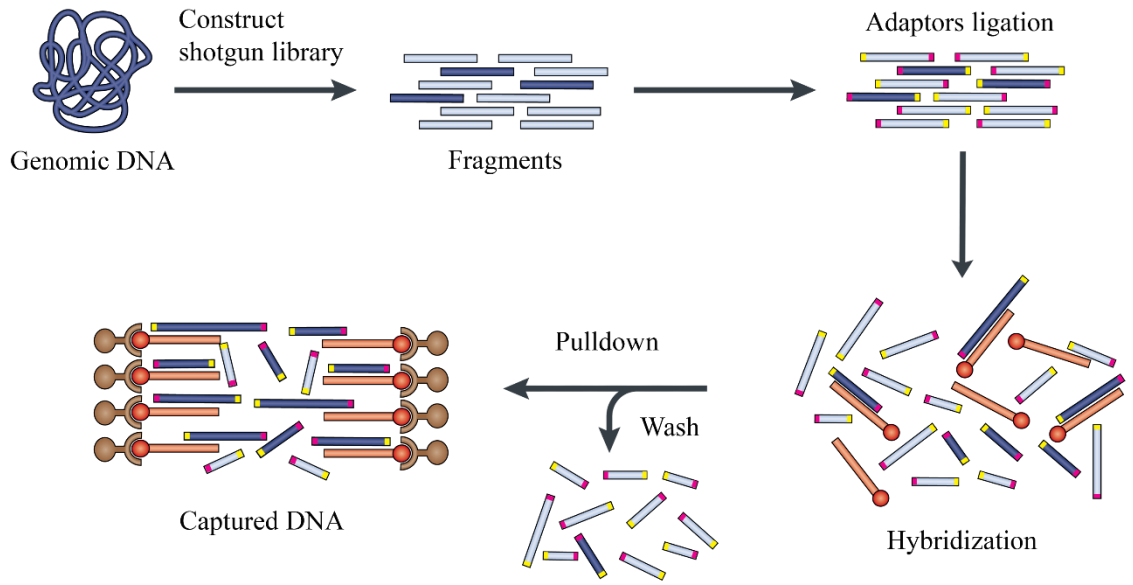

Figure 1. Library construction for WES experiment. Genomic DNA is fragmented through different systems to create a library. Fragments obtained are then ligated to adaptors (shown in yellow and light purple). After adaptors ligation, fragments are enriched for exonic sequences (dark blue) using an exome enrichment capture system constituted by RNA or DNA probes (orange sequences) that are biotinylated (red dots on orange sequences). Fragments hybridized, corresponding to exonic sequences, are then recovered through a biotin-streptavidin based pulldown, while the ones not ligated are washed away. **Figure adapted[13].**

are ligated to adaptors, to generate a library. Next, the library obtained is enriched for sequences corresponding to exons. Among strategies available to capture protein-coding sequences[21], the capture by hybridization approach in the aqueous-phase is the most used. In this case, the selection of exonic sequences occurs through the hybridization of the library with an exome enrichment kit constituted by DNA or RNA biotinylated baits complementary to sequences of interest. Recovery of hybridized fragments (corresponding to exonic sequences) is then performed through biotin-streptavidin-based pulldown[13].

Fragments recovered are successively amplified following technology used by the sequencer chose. Among amplification strategies[20], the most diffuse is the one used by Illumina platforms, represented by a solid-state amplification, specifically, a bridge-amplification (Figure 2). In this technique, DNA captured fragments are hybridized to a solid

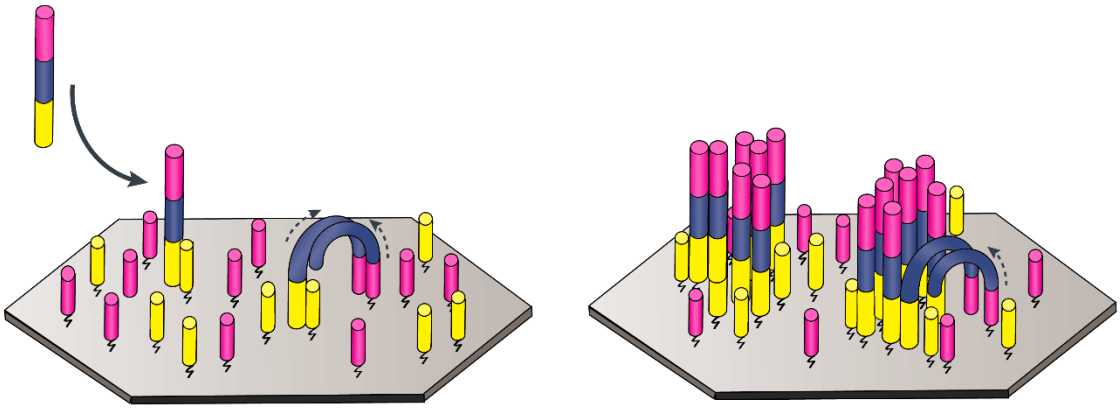

Figure 2. Solid-phase bridge amplification. In Illumina sequencers, library amplification is performed in a solid-phase. The solid surface used shows forward and reverse primers (yellow and light purple fragments) complementary to the adaptors ligated to the DNA fragments during library generation (Figure 1). DNA single-stranded fragments obtained from previous phase, are hybridized to the solid surface. Each DNA fragment binds to one of the primers on the solid surface and a polymerase is used to create the complementary sequence (figure on the left). Double-strand DNA molecules synthetized are denatured, and the two single-stranded DNA molecules resulting, fold over (figure on the right), binding nearby primers and being sequenced again. This process is also called cluster generation since it generates a cluster of identical fragments starting from the same DNA molecule. Cluster generated are then sequenced through a short-read sequencing process based on sequencing by ligation or sequencing by synthesis approach. **Figure adapted[20].**

surface in which forward and reverse primers, complementary to the adaptors on the fragments, are found. Each single-stranded DNA fragment binds to one of the primers and a polymerase creates the complementary sequence. Once that double-strand DNA molecule has been generated, it is denatured, and the two single-stranded DNA fragments obtained fold over, binding the nearby primers and encountering a new process of sequencing. This process is repeated many times, generating a cluster containing millions of copies of the starting DNA fragment.

Once that DNA fragments have been amplified, they are sequenced through short-read sequencing approaches, based on sequencing by ligation (SBL) or sequencing by synthesis (SBS)[20]. In SBL approaches, a probe sequence is labelled with a fluorophore and, when the probe hybridizes to a DNA fragment, it releases the fluorophore, allowing to identify the probe complementary to the sequence, through the emission spectrum generated. In SBS approaches, when a nucleotide is incorporated during extension of a DNA fragment, it releases a signal such as a fluorophore or a change in ionic concentration, that allows to identify the nucleotide[20].

1.2.1.1 Comparison of sequencing platforms and exome enrichment capture systems

Among short-reads platforms, SBL technique is mostly used by SOLiD (Thermo Fisher Scientific) and Complete Genomics (BGI) systems[20]. These platforms can generate reads very different in length, ranging from 75 bp for SOLiD to 28-100 bp for Complete Genomics[20]. Although these systems show high accuracy in base identification (~99,99%) as each base is probed multiple times, they present several limitations, as low sensitivity and specificity, since true variants are missed while few false variants are called[20]. The SBS technique is used by Illumina sequencers which generate reads of length up to 300 bp[20]. Although these platforms show a lower accuracy compared to SOLiD and Complete Genomics systems (> 99.5%)[20], they also show a higher sensitivity (even if false-positive rate is around 2.5%)[20]. Therefore, providing a wide range of sequencers[20], Illumina NGS platforms are currently the most used for short-reads sequencing.

To perform exome capture, several exome enrichment systems have been developed. Most used kits provided by Roche NimbleGen, Agilent Technologies and Illumina, show several differences in terms of target size and design. The dimensions of kits commercially available span from ~37Mb of Nextera Rapid Capture Exome (Illumina) to ~67Mb of SureSelect Clinical Research Exome V2 (Agilent Technologies). Differences observed in target size are mostly due to target design. Indeed, exome enrichment capture systems are designed considering gene sequences contained in several databases as RefSeq (NCBI Reference Sequence Database)[22], GENCODE[23] and CCDS (Consensus CDS)[24]. In addition to capture exonic regions, exome enrichment kits can also contain probes targeting protein non-coding sequences, as microRNAs (miRNAs). In this case, reference database used is represented by miRBase (the microRNA database)[25] which contains information on all miRNA sequences identified in more than 200 species. Moreover, exome enrichment kits may be available in expanded versions, with probes for sequences outside coding exons. This is the case of SureSelect Human All Exon V6+UTR (89Mb, Agilent Technologies) which target regions include 5′ and 3′ UTR regions, and of SeqCap EZ MedExome (Roche NimbleGen) that, combined with SeqCap EZ Mitochondrial Genome Design (Roche NimbleGen), allows to sequence the entire mitochondrial DNA.

1.2.2 Whole Exome Sequencing bioinformatics processing

WES experiments produce a large amount of data that can vary substantially, according to the experimental design (e.g. due to the exome enrichment kit used). Fundamental steps in the processing of WES data require the alignment of the reads to the reference genome, the identification of variants present in the WES analysed (i.e. variant calling) and the functional annotation of variants and genes identified (Figure 3).

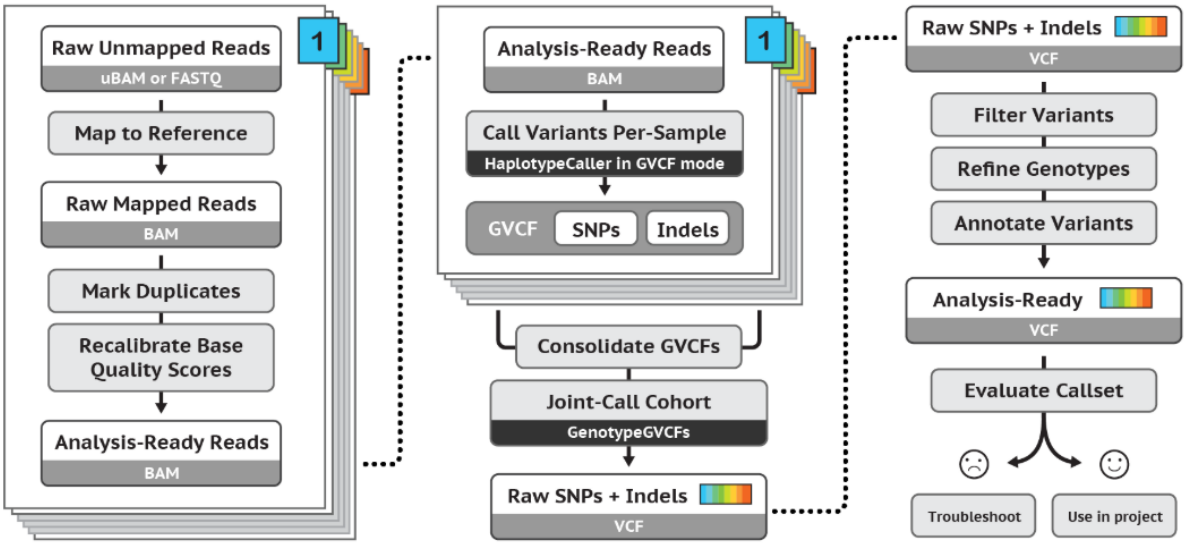

Figure 3. Genome Analysis Toolkit (GATK) pipeline for germline variants discovery. According to GATK pipeline, widely used for germline short variant discovery, raw data produced are aligned to reference genome. Raw mapped reads are analysed to identify and mark duplicated reads. Base quality scores are recalibrated, and germline variants are identified applying "HaplotypeCaller" algorithm. If multiple samples are available (e.g. in case of a trios), only one variant calling step (joining the three samples) should be performed. After variant calling step, raw single nucleotide variants and small insertions and deletions identified should be filtered, according to quality criteria. At this step, genotypes can be refined, and variants can be annotated. **Figure reprinted from GATK**[I]

A WES experiment generally produces from tens to hundreds of millions of reads stored in FASTQ files. Prior to the alignment step, reads are treated to remove the adaptors used during sequencing experiment, to obtain reads containing only sample DNA sequences. For adapter trimming, several tools can be used, as TrimGalore![II] and Trimmomatic[26]. File produced from this pre-processing step is still a FASTQ file. Trimmed reads are then aligned to the reference genome through tools as Burrows-Wheeler Aligner (BWA)[27] and NovoAlign[III]. Currently, the assembly hg38/GRCh38 represents the most recent update of the human reference genome, even if analyses can be also performed using previous genome version,

hg19/GRCh37. Indeed, for hg19/GRCh37 assembly, a plenty of tools has been developed during past years, while tools available for the most recent version of the human genome are, in large part, still under development. After the alignment step, the mark of PCR duplicates is required, since WES experiments produce a lot of duplicated reads, due to the clonal amplification, that are uninformative for variants detection. Therefore, through tools as MarkDuplicates by Picard[IV] or markdup by SAMtools[28], PCR duplicates are flagged and easily discarded. Next step is represented by base quality recalibration (Base Quality Score Recalibration - BQSR) generally performed through Genome Analysis Toolkit (GATK)[29]. BQSR estimates systematic errors made by the sequencer during base calling and, consequently, it adjusts overall base quality values. As variant calling algorithms highly rely on quality values assigned to each base call, BQSR is a fundamental step to get more accurate base qualities, which in turn improves the accuracy of variant calling. The file obtained from the steps of alignment to the reference genome, mark of PCR duplicates and BQSR is a BAM (Binary Alignment Map) file which dimensions can vary from 6 to 13 Gb[16]. Parameters as coverage, depth and unique mapped reads can be used, at this level, to evaluate the quality of data (Figure 4). Considering a genomic region, coverage refers to the extension of the effective capture of the region (expressed as a percentage), while depth is related to the number of reads that supports each base in the region (and it is expressed as a number). Unique mapped reads refer to the reads that, depleted from PCR duplicates, can be used to call variants.

Variant calling is the process in which WES data are analysed to identify variants. GATK HaplotypeCaller[29] is the most used tool to perform germline variant calling, identifying both single-nucleotide variants (SNVs) and small insertions and deletions (indels). Variants identified are reported in a VCF (Variant Call Format) file and they can be filtered, according to quality criteria, to retain only reliable ones.

Finally, variants and genes in which variants localise are annotated to evaluate their potential biological role. Functional annotations on variants may regard their effect on transcripts, their frequencies in population databases and information on already known

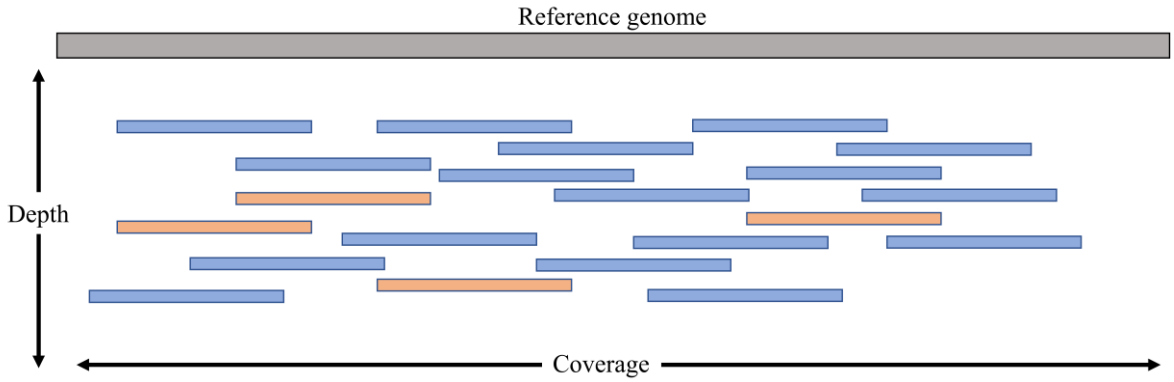

Figure 4. Coverage, depth and unique mapped reads. After the first step of pre-processing, BAM files can be evaluated considering parameters of coverage, depth and unique mapped reads. Coverage parameter expresses in percentage the extension of a region of interest covered by reads aligned. Depth parameter (also called depth of coverage) is calculated at a nucleotide level and refers to the number of reads that support a specific call. Coverage and depth can be combined to obtain summary statistics indicating the percentage of a region of interest covered at a defined depth (e.g. a region of interest can be covered at 90% with a depth of 20X, meaning that at least 20 reads cover the 90% of the region investigated). Finally, unique mapped reads parameter indicates reads, depleted from PCR duplicates (shown in orange), that should be considered to perform variant calling. Unique mapped reads parameter can be evaluated both as a number and as a percentage. Since duplicated reads introduce a bias in the evaluation of coverage of a region of interest (as they increase the number of reads mapping in the region), coverage and depth parameters should be calculated on unique mapped reads.

disease-causative mutations. Functional annotation on genes may involve gene function, gene pathways, gene ontology, phenotypes caused by homolog genes, and already known disease-gene associations.

To annotate variants effect on transcripts, several tools can be used as ANNOVAR[30], SnpEff[31], and VEP (Variant Effect Predictor)[32]. Using as a reference database a set of transcripts, as RefSeq[22], GENCODE[23], or Ensembl[33], these tools identify variants localisation in transcripts, evaluating possible functional consequences (e.g. whether variants alter an exonic sequence). Variants frequencies across several populations can be retrieved from databases as dbSNP (Database of Single Nucleotide Polymorphisms)[34], 1000 Genome[35] or the most recent gnomAD (genome Aggregation Database)[36] comprehensive of more than 120,000 WES and 15,000 WGS data. Other information on variants can be added considering their conservation across genomes, using software as phyloP[37] and GERP++[38]. Information related to already known disease-causative mutations can be added through databases as HGMD (The Human Gene Mutation Database)[39] and ClinVar[40]. Databases containing population frequencies, conservation scores and known disease-causative mutations (and others

information, Table 1) can be separately added or can be comprehensively annotated using metadatabases, as dbNSFP[48], which contain data for both variants and genes functional annotation.

Table 1. Functional annotation on variants and genes that can be integrated in analysis of WES data.

Tool	Annotation on	Purpose	Reference
ESP6500 (Exome Sequenincg Project v. 6500)	Variants	Reports population-specific variants frequencies	Exome Variant Server[41]
ExAC (Exome Aggregation Consortium)	Variants	Reports population-specific variants frequencies	Lek et al., 2016[36]
PhastCons	Variants	Identifies conserved sites scoring each substitution	Siepel et al., 2005[42]
InterPro	Variants	Provides information on protein domain in which the variant locates	Finn et al., 2017[43]
BioCarta	Genes	Provides information on gene pathways	Nishimura, 2001[44]
RVIS	Genes	Gives a score to genes in terms of whether they have more or less common functional genetic variations	Petrovski et al., 2013[45]
Expression Atlas	Genes	Provides information on genes and proteins expression across species and biological conditions	Petryszak et al., 2016[46]
HI	Genes	Estimates probability of genes haploinsufficiency	Huang et al., 2010[47]

Information on gene function can be retrieved from databases as UniProt[49]. To annotate gene pathways, single resources as KEGG (Kyoto Encyclopedia of Genes and Genomes)[50], or systems integrating databases, as ConsensusPathDB[51], can be used. Genes may also be analysed using information coming from databases that classify them relying on biological, molecular and cellular features, as GO (Gene Ontology)[52]. To analyse possible involvement of genes in the disease investigated, information may also come from homolog genes functions and phenotypes-homolog genes associations. To this aim, information contained in databases regarding mouse and zebrafish, as MGI (Mouse Genome Informatics)[53] and ZFIN (The Zebrafish Information Network)[54] respectively, can be used. Finally, information on genes already associated with diseases may be obtained from OMIM (Online Mendelian Inheritance in Man)[55] or HGMD[39] databases.

1.2.3 Whole Exome Sequencing data analysis

WES data analysis can be performed considering either the whole gene set or using an *in silico* panel of genes[16], which can be firstly analysed to facilitate variants interpretation, and that can be subsequently expanded. Furthermore, as 85% of disease-causative mutations in Mendelian diseases resides in protein coding-regions[19], pipelines used to analyse WES data generally analyse only SNVs and indels that fall in these regions. WES data analysis focuses on non-synonymous, non-sense, frameshift and on splice donor and acceptor variants. Then, to identify possible disease-causative mutations, several filters can be applied, based on biological observations and functional annotations.

Biological information may come from phenotype observed, may regard molecular analyses previously assessed on the patient (e.g. linkage analysis, SNP array, CGH array), may come from his familiar history (e.g. pedigree information) or from other non-related affected patients.

Functional annotations can be used to filter variants and genes. The most used filter relies on variants frequency. Indeed, as disease-causative mutations would be rare and therefore likely to be previously unidentified, population databases information is used to remove annotated variants with high frequency.

Furthermore, several strategies may be used to prioritise variants and genes, to identify those potentially related to the disease investigated. Among tools that predict variants potential deleterious effect, there are PolyPhen-2 (Polymorphism Phenotyping v2)[56] and SIFT (Sorting Intolerant From Tolerant)[57], which score only non-synonymous variants, and systems as CADD (Combined Annotation Dependent Depletion)[58] and DANN[59] that through a machine-learning system approach, trained on comparative genomics data, allow to assess the potentially damaging effect of all SNVs and indels. Furthermore, other tools can be used to predict variants interfering with splicing, as dbscSNV[60] or SPIDEX[61] which are respectively based on data coming from several databases and from RNA sequencing (RNA-seq) experiments.

For genes prioritisation, several strategies can be used, mostly relying on the identification of similarities between genes investigated and already known phenotype-gene associations. Among these tools, there are GeneDistiller[62] and the most recent Phenolyzer[63] which allows to better define phenotype investigated, considering phenotypic standardised

terms (taken from HPO – The Human Phenotype Ontology)[64], and therefore to discover more accurate associations of genes with similar phenotypes.

To interpret the clinical relevance of variants prioritised, tools as InterVar[65] or Sherloc[66] can be used. Applying ACMG-AMP (American College of Medical Genetics and Genomics - Association for Molecular Pathology) guidelines[67], these tools classify variants, allowing to reveal those definable as pathogenic according to standardised criteria. Moreover, as one of the big issues related to WES data analysis concerns secondary findings, these tools allow to assign a clinical significance also to variants that are not necessarily correlated with the disease investigated, improving patient management.

1.3 Limitations and potentialities of Whole Exome Sequencing

WES successful rate has been estimated to be ~20-30%[16,18], despite the number of tools available to analyse data and interpret variants. This can be due to a combination of biological, technical and analytical reasons that can limit the power of new disease-gene discovery (Table 2).

Although WES presents fundamental limitations, it can also provide a meaningful amount of information, usually discarded by commonly used workflow analyses[68], and that should be considered as it can have a relevant biological role in the phenotype investigated. Indeed, a recent study published by Bergant and colleagues showed that an extended exome analysis improved their diagnostic rate of ~4% in a cohort of more than 1000 cases[69].

Several tools have been developed to analyse from WES data information related to genome, as for Copy Number Variants (CNVs) and Regions of Homozygosity (ROHs), and for synonymous and non-coding variants. Thus, even if molecular cytogenetic methods (as

CGH and SNP array) and WGS experiments can be more accurate in the analysis of these genomic elements, they can also be analysed starting from WES data.

Table 2. Factors contributing to bottlenecks in the identification of new disease-causative mutations and new disease-gene associations. Table adapted[4].

Level of analysis	Possible issues
Clinical data	•non-specific clinical presentations (e.g., developmental delay and hypotonia) •ultra-rare and unrecognized genetic diseases •lack of ontology encompassing the complete spectrum of human phenotypes •insufficient utilization of ontologies or 3D facial-gestalt analysis in phenotyping •inconsistent multidisciplinary approaches to patient evaluation •inability to account for and compare age-specific disease presentations
Genomic data	•technical limitations of WES (e.g., copy-number variants and structural variation are not captured well) •lack of standardized technical and informatics approaches •incompleteness of population-specific control datasets
Data discovery and sharing	•lack of a widely adopted data-sharing framework •lack of common data-sharing standards •lack of a systematic way to record data-use conditions •lack of a privacy-preserving linkage system for each research participant
Genetic evidence	•siloed datasets •lack of and use of data-sharing infrastructure
Functional evidence	•lack of standardized and moderate-throughput analyses of variant impact •lack of biological insight into the function of most human genes
Novel disease mechanisms	•other mechanisms including tissue-specific mosaicism, methylation, and di- or oligogenic inheritance

1.4 MicroRNAs: biogenesis, function, and involvement in Rare Genetic Diseases

MicroRNAs (miRNAs), which have been shown to play an important role in RGDs[70], are small non-coding RNAs of ~22 nucleotides, widely expressed in all human tissues, that interfere with gene translation by targeting 3' untranslated regions (UTRs) of messenger RNAs (mRNAs)[71,72]. MiRNAs are mainly transcribed by RNA polymerase II, from long non-coding RNAs, intronic regions and, to a lesser extent, from exonic regions[72,73] (Figure 5). Several miRNA loci can be found near each other, therefore constituting a polycistronic transcription unit[72,73]. Transcription of miRNA genes generate primary miRNA (pri-miRNAs) transcripts that are further processed in the nucleus. Drosha and DGCR8 cleave the pri-miRNA leading to the formation of a precursor miRNA (pre-miRNA), ~70 bp long, which shows a 2-nucleotide overhang at 3' end. Exportin 5 recognises the pre-miRNA and exports it to the cytoplasm. Here another protein, Dicer, which acts as a 'molecular ruler', cleaves the

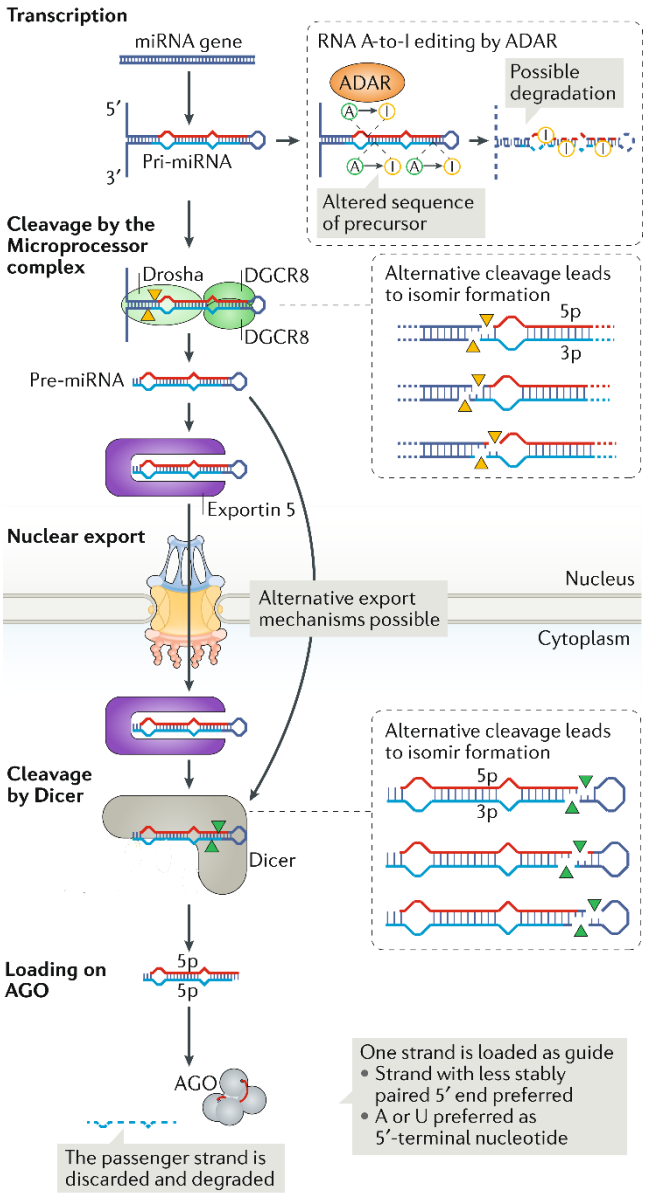

Figure 5. MicroRNA biogenesis. Figure adapted[72]

pre-miRNA generating a mature-miRNA duplex (composed of two mature miRNAs), with another typical 2-nucleotide 3'overhang[72]. One of the two mature miRNAs, the 'guide strand', is loaded into one of four AGO proteins (AGO 1-4) to form an effector complex called RNA-induced silencing complex (RISC)[73], while the other mature miRNA, the 'passenger strand', is discarded. Loading preference is given to the less stably 5' end[72]. Along with canonical mature miRNAs, there can be produced multiple isoforms, isomirs, that originate from pri-miRNA modifications, due to an RNA A-to-I editing process, or from different cleavages performed by Drosha and Dicer[71,72].

The RISC complex can act binding target sites for miRNAs which are generally located in 3′ UTRs of mRNAs and are highly complementary to miRNAs seed regions. Seed region of a miRNA is generally defined as the region spanning from second to eighth nucleotide at mature 5′ end. Moreover, miRNAs can present specific sequences motifs based on AGO proteins on which they are loaded. This is the case of miRNAs loaded into an AGO2 protein which tend to show an A or a U at 5′-terminal-nucleotide, since AGO2 protein prefers one of these two nucleotides as first base. Through seed region, mature miRNAs can bind one or more mRNAs and mediate gene silencing through translation repression and mRNA decay[72]. Although these two modes seem to be interconnected, it has been observed that from 66 to 90% of gene silencing events occur through mRNA decay[72].

Since one miRNA can target more genes and more miRNAs can interact with the same gene, deregulation of miRNAs function has been associated with several human diseases, particularly cancer[74], but even RGDs[70]. MiRNA variants associated with RGDs have been found in genes responsible for miRNA biogenesis, in miRNA target sites and in miRNA sequences[70].

Since multiple enzymes and cofactors participate in the biogenesis of miRNAs (e.g. Drosha and Dicer), pathogenic variants in these genes generally result in the reduced efficiency of miRNA processing, which can lead to human diseases[70]. Several disease-causative variants have been found in these genes, as in the case of *DICER1* whose heterozygous germline variants have been associated with Familial Pleuropulmonary Blastoma (PPB, MIM: 601200)[75].

Disease-causative variants in miRNA binding sites may function as regulatory elements through modifying miRNA binding affinity and/or specificity, leading to a deregulation of expression of target genes. Among variants identified in 3′ UTR binding sites, a variant in 3′UTR of gene *SLITRK1* was found in patients affected by Tourette syndrome (MIM: 137580)[76].

Pathogenic variants in mature miRNAs can alter miRNA processing and miRNA targeting, leading to the recognition of many novel and aberrant direct targets. Few disease-causing variants in miRNA sequences have been associated to Mendelian diseases so far, specifically in miR-96, miR-204 and miR-184.

In 2009 Mencía and colleagues analysed a Spanish family affected by non-syndromic progressive hearing loss (MIM: 613074), revealing the presence of two disease-causative variants in seed region of miR-96[77]. These variants significantly alter both miRNA biogenesis, leading to a reduced expression of mature miRNA, and miRNA targeting, bringing to an overexpression of several predicted miRNA target genes expressed in the inner ear (as *AQP5*, *ODF2*, *MYRIP* and *RYK*). The variants identified in miR-96 were therefore recognised as disease-causative for non-syndromic progressive hearing loss[77]. Moreover, another variant in miR-96 precursor sequence was found associated with the same phenotype, impairing both mature miRNAs processing and expression[78].

Another pathogenic variant in the seed region of a mature miRNA was found in miR-204 by Conte and colleagues in 2015 in patients showing retinal dystrophy associated with ocular coloboma[79]. Studying a five-generation family, a pathogenic variant in the seed region of miR-204 was identified. This variant alters miRNA targeting, through the loss of canonical gene targets and the creation of new ones. These alterations are responsible for an increase of retinal cell apoptosis, that lead to a reduced number of both cones and rods photoreceptor cells, causing a phenotype consistent with the one observed in the family[79].

In 2011, Hughes and colleagues, identified a disease-causative variant in miR-184 responsible for keratoconus and early-onset anterior polar cataracts in a large Irish family[80]. The mutant miR-184 fails to compete with miR-205 for overlapping target sites on the 3' UTRs of *INPPL1* and *ITGB4* genes, leading to their dysregulation. Although these target genes and miR-205 are expressed widely, miR-184 is highly expressed only in cornea and lens. Therefore, phenotype observed, due to miR-184 pathogenic variant, is restricted to these tissues. The same variant was also found in patients affected by EDICT syndrome (MIM: 614303) showing differences in keratoconus phenotype[81] compared with the family reported by Hughes and colleagues, thus supporting the hypothesis that other genetic modifiers can explain different corneal phenotype in these two families. Moreover, two new pathogenic variants in miR-184 were found in patients affected by isolated keratoconus. These variants reside in precursor sequence of miRNA and interfere with efficiency of processing[82].

1.5 Methods used to study microRNA sequences and expression

Since miRNA dysregulation may have substantial effects on gene silencing, miRNAs have been extensively studied, firstly focusing on their expression profiles. Studying miRNA expression variability can be very informative, elucidating biological processes in which miRNAs play a crucial role, as organismal development and establishment and maintenance of tissue differentiation[83].

Three major approaches are currently used to study miRNAs profiling: hybridization-based methods (e.g. DNA microarrays), quantitative reverse transcription PCR (qRT-PCR) and NGS approaches (RNA-seq)[83].

Hybridization-based methods, as microarrays, were among the first methods used to study simultaneously several miRNAs. These experiments are based on the reverse transcription of miRNAs, on their labelling (e.g. by fluorescence), and their subsequent hybridization on an array in which there are DNA complementary probes. Even if these systems have really low-costs, they do not allow to perform absolute quantification of miRNAs considered, as well as they cannot identify novel miRNAs[83].

Methods as qRT-PCR are based on the reverse transcription of miRNAs to cDNAs. Once obtained cDNAs, there are amplified through a qPCR with real-time monitoring of reaction product accumulation. Although these systems allow to obtain an absolute quantification of miRNAs amplified, these techniques do not allow to discover new miRNAs[83].

NGS experiments have allowed to completely change the way in which miRNA profiling is performed. Indeed, RNA-seq based on NGS technologies allows to simultaneously study a plenty of miRNAs. The great advantage of this technique, compared with microarrays and qRT-PCR, resides in its ability to investigate both miRNA expression profiles and their sequences, not only analysing already known miRNA sequences, but even investigating those completely new[83].

Due to this great advantage, RNA-seq has quickly become the most diffuse approach to analyse both miRNA sequences and their expression profiles in a sample.

1.5.1 MicroRNAs and Whole Exome Sequencing

Even if miRNA expression profiles can be only detected through the aforementioned methods, miRNA sequences can also be recovered from WES data. Indeed, exome enrichment capture systems may enclose probes to capture also miRNA sequences.

One of the evidences of the presence of miRNA information in WES data comes from the work of Carbonell and colleagues[84]. To study miRNA variability in the human genome, they collected 1,152 healthy individuals. For 60 of them a WES experiment was performed using SeqCap EZ Exome (Roche NimbleGen), while for the remaining 1,092, data were downloaded from 1000 Genome Project[84]. Since the exome enrichment kit chose has been designed considering 720 miRNAs (taken from miRBase v13), they analysed all the samples considering variants localised in these miRNAs[84].

WES also allowed to discover one of the disease-causative variants in miRNAs associated with RGDs, specifically in the case of miR-204[79]. Focusing on a candidate region evidenced by linkage analysis, the authors analysed WES data, discovering the variant reported[79].

Taken together these evidences suggest that information regarding miRNAs may reside in WES. However, currently, there are not dedicated tools to retrieve miRNA related information from WES data.

2. Aim of the thesis

The aim of this thesis was to investigate the presence of information related to miRNA sequences in WES data. To this purpose, we evaluated the ability of the most used exome enrichment capture systems commercially available to effectively capture miRNA sequences. Then, we developed a dedicated tool to retrieve, analyse and functionally annotate variants in miRNAs. To test our tool, we analysed WES data of a cohort of 259 individuals including patients affected by different genetic diseases.

3. Material and Methods

3.1 Whole Exome Sequencing and microRNAs: capture evaluation

We investigated which amount of miRNA-related information could be found in WES data. To this aim, we evaluated both theoretical and experimental coverage relative to miRNA sequences using several exome enrichment capture systems commercially available: SeqCap EZ Human Exome Library v3.0 (Roche NimbleGen), SeqCap EZ MedExome (Roche NimbleGen), Nextera Rapid Capture Exome (Illumina), SureSelect Human All Exon V4 (Agilent Technologies), SureSelect Clinical Research Exome (Agilent Technologies), SureSelect Clinical Research Exome V2 (Agilent Technologies), SureSelect Human All Exon V6 (Agilent Technologies).

Most exome capture systems are designed on hg19/GRCh37 assembly (except for SeqCap EZ MedExome, designed on hg38/GRCh38). Therefore, to evaluate miRNAs coverage, we considered miRNA "primary transcript" sequences, as reported in miRBase v20[25] (version designed on assembly hg19/GRCh37). To perform coverage analyses we used bedtools package (version 2.26), composed by a series of utilities that allow to perform several genomics analyses[85].

3.1.1 Theoretical coverage

As a first step, we evaluated whether exome enrichment capture systems contain probes to specifically capture miRNA sequences. To this aim, we compared genomics coordinates of target regions for each kit considered with miRNA "primary transcripts" defined in miRBase v20[25]. For SeqCap EZ MedExome we considered coordinates of capture regions based on assembly hg19/GRCh37 as furnished by Roche NimbleGen. We therefore evaluated miRNA sequences overlapping at least at 50% with target regions, using bedtools "intersect" tool (version 2.26)[85]. We chose this overlapping threshold since miRNA "primary transcript" sequences are ~80bp long while target regions are generally larger and, consequently, we expected that overlapping target regions would be able to capture miRNA full sequences. Histogram showing theoretical coverage of miRNA sequences was generated using "ggplot2"[86], a R library.

3.1.2 Experimental coverage

Then we evaluated whether exome baits would be effectively able to capture miRNA sequences. To this purpose, we selected 14 WES data captured through different exome enrichment capture systems (Table 3).

Table 3. WES cases considered for evaluation of miRNA sequences coverage in WES data. Table reports pre-processing steps performed to generate BAM files analysed, along with algorithms used and respective versions. BQSR stands for Base Quality Score Recalibration.

Case	Exome enrichment capture system	Alignment to reference genome	Removal of PCR duplicates	BQSR
1	SeqCap EZ Human Exome Library v3.0 (Roche NimbleGen)	BWA-MEM, version 0.7.10	Picard MarkDuplicates, version 1.119	Not performed before generating final BAM
2	SeqCap EZ Human Exome Library v3.0 (Roche NimbleGen)	BWA-MEM, version 0.7.10	Picard MarkDuplicates, version 1.119	Not performed before generating final BAM
3	SeqCap EZ MedExome (Roche NimbleGen)	BWA-MEM, version 0.7.12	Picard MarkDuplicates, version 2.3.0	Not performed before generating final BAM
4	SeqCap EZ MedExome (Roche NimbleGen)	BWA-MEM, version 0.7.10	Picard MarkDuplicates, version 1.119	Not performed before generating final BAM
5	Nextera Rapid Capture Exome (Illumina)	BWA-MEM, version 0.7.12	Picard MarkDuplicates, version 2.3.0	GATK PrintReads, version 3.7.0
6	Nextera Rapid Capture Exome (Illumina)	BWA-MEM, version 0.7.10	Picard MarkDuplicates, version 1.119	Not performed before generating final BAM
7	SureSelect Human All Exon V4 (Agilent Technologies)	BWA-MEM, version 0.7.10	Picard MarkDuplicates, version 1.119	Not performed before generating final BAM
8	SureSelect Human All Exon V4 (Agilent Technologies)	BWA-MEM, version 0.7.10	Picard MarkDuplicates, version 1.119	Not performed before generating final BAM
9	SureSelect Clinical Research Exome (Agilent Technologies)	BWA-MEM, version 0.7.12	Picard MarkDuplicates, version 1.119	Not performed before generating final BAM
10	SureSelect Clinical Research Exome (Agilent Technologies)	BWA-MEM, version 0.7.10	Picard MarkDuplicates, version 1.119	Not performed before generating final BAM
11	SureSelect Clinical Research Exome V2 (Agilent Technologies)	BWA-MEM, version 0.7.12	Picard MarkDuplicates, version 2.3.0	GATK PrintReads, version 3.7.0
12	SureSelect Clinical Research Exome V2 (Agilent Technologies)	BWA-MEM, version 0.7.12	Picard MarkDuplicates, version 2.3.0	GATK PrintReads, version 3.7.0
13	SureSelect Human All Exon V6 (Agilent Technologies)	BWA-MEM, version 0.7.12	Picard MarkDuplicates, version 2.3.0	GATK PrintReads, version 3.7.0
14	SureSelect Human All Exon V6 (Agilent Technologies)	BWA-MEM, version 0.7.12	Picard MarkDuplicates, version 2.3.0	GATK PrintReads, version 3.7.0

To perform coverage analysis, we considered WES BAM files. Briefly, raw exome data were aligned to the reference genome, assembly hg19/GRCh37, using BWA-MEM algorithm[87]. PCR duplicates were then removed using MarkDuplicates by Picard[IV]. In some cases, Base

Quality Score Recalibration was performed before generating final BAM file, through Genome Analysis Toolkit software (GATK)[29]. As WES were performed during a period of 4 years (2014-2018), versions of aforementioned algorithms may slightly differ (Table 3).

First, to assess the quality and homogeneity of the WES, we evaluated the individual "on target" coverage, comparing each WES data with target regions of its own exome enrichment capture system. Next, we evaluated the coverage of miRNA "primary transcript" sequences, reported in miRBase v20[25]. The evaluation of "on target" and miRNA sequences experimental coverage was performed using bedtools "coverage" tool (version 2.26)[85], without any overlapping threshold. Graphs showing relation between coverage and depth for target regions and miRNA sequences were generated using the R library, "ggplot2"[86].

3.1.3 Comparison with Whole Genome Sequencing data

Since WGS experiments do not rely on a procedure for selection and capture of target regions, we compared data on miRNAs coverage between WES and WGS experiments. To this aim, we analysed two cases, specifically case 1 and 2 (Table 3), for which, besides WES data, we also had WGS data. We computed experimental coverage of miRNA "primary transcript" sequences (reported in miRBase v20)[25] in WGS data, using bedtools "coverage" tool (version 2.26)[85]. To evaluate miRNAs coverage, we used WGS BAM files. Raw genome data were aligned to the reference genome, assembly hg19/GRCh37, using BWA-MEM algorithm (version 0.7.12)[87] and, as WGS performed were not PCR-free, PCR duplicates were removed using MarkDuplicates by Picard (version 1.119)[IV]. We therefore compared data on miRNAs coverage obtained from WGS analysis with data already generated from WES analysis.

As the number of WES and WGS considered was limited (2 WES and 2 WGS), we decided to extend our analyses on data available in the public database gnomAD (genome Aggregation Database version 2.0.2)[36], containing 123,136 WES and 15,496 WGS.

We therefore evaluated coverage of miRNA sequences in publicly available coverage data for gnomAD WES and WGS. These coverage data report, at a single nucleotide level, statistics on coverage, as mean and median, calculated evaluating all individuals sequenced. GnomAD WGS data contain coverage data for all the human genome, while gnomAD WES data have been processed considering only the exonic portion[36]. Therefore, to evaluate

whether gnomAD WES data have been processed in respect of miRNA sequences, we first assessed whether exome calling intervals contain miRNA "primary transcript" sequences, using bedtools "intersect" tool (version 2.26)[85], with an overlapping threshold of 100%.

Then, we analysed gnomAD WES and WGS coverage data, focusing on regions of miRNA "primary transcript" sequences. To select from coverage data only regions of interest, we used tabix (from HTSlib 1.9, Samtools[V]) and bedtools "intersect" tools (version 2.26)[85]. Next, we generated a coverage summary file for gnomAD WES and WGS data; specifically, we computed the same statistics provided by bedtools "coverage" tool on single BAM files, considering "mean" coverage values reported in gnomAD coverage data, using a R script (version 3.4.4)[88].

Graphs showing relation between coverage and depth for miRNA sequences in WES and WGS experiments were generated through the R library "ggplot2"[86].

3.2 Evaluation of microRNA variants in a cohort

Once we assessed that WES data contain information on miRNA sequences, we analysed miRNA variants in a cohort composed by 259 individuals sequenced through WES experiments. Individuals were sequenced and analysed in a collaboration with Genetics and Rare Diseases Research Division at Bambino Gesù Children's Hospital, Rome, Italy. Specifically, in the cohort considered, 110 patients affected by several RGDs were sequenced to characterise molecular bases of the observed phenotypes. Where possible, relatives, including parents, brothers and sisters were sequenced. Therefore, the cohort of WES resulted composed by 110 probands (including 11 pairs of siblings) and 149 relatives (unaffected parents and siblings).

WES experiments were performed in a homogeneous way: exomes were captured using the SureSelect Human All Exon V4 (Agilent Technologies) and were subsequently sequenced through the HiSeq2000 platform (Illumina).

Raw exome data were aligned to the reference genome hg19/GRCh37 through BWA-MEM algorithm (version 0.7.10)[87]. PCR duplicates were removed through MarkDuplicates by Picard (version 1.119)[IV]. Base Quality Score Recalibration was performed through GATK (version 3.3)[29]. To specifically identify germline variants localised in miRNA sequences, variant calling was performed through GATK Haplotype Caller (version 3.7)[29] using, as calling regions, miRNA "primary transcript" sequences with a padding of 50bp. When multiple

samples of the same family were available, variant calling was performed through a joint calling. Hard filtering on variants was performed applying following criteria:

1) Qual by Depth (QD) < 2.0,
2) Fisher Strand (FS) > 60.0,
3) Strand Odds Ratio (SOR) > 3.0,
4) Root Mean Square of Mapping Quality (MQ) < 40.0,
5) Mapping Quality Rank Sum Test (MQRankSum) < -12.5,
6) Read Pos Rank Sum Test (ReadPosRankSum) < -8.0,
7) QUAL parameter < 100.0.

Only variants passing these criteria were identified as good quality variants and flagged as "PASS" or "SnpCluster" if more than 3 variants were found in a range of 10 bp.

To analyse miRNA variants in the WES cohort, we considered only probands. When two patients were present in the same family, we randomly selected only one of them to eliminate bias due to the high amount of DNA shared between siblings. We therefore analysed miRNA variants on 99 probands.

To retrieve miRNA variants, we developed a script in Python (version 2.7.14)[89], that allowed to analyse VCF files searching for variants localising in miRNA sequences, as defined in miRBase v20[25]. Comprehensively, the database contains the genomics coordinates of 1871 miRNA "primary transcripts" and 2794 mature miRNAs. To better define miRNA variants location in miRNA sequences, we defined different miRNA regions, corresponding to the following substructures:

1) Seed regions: bases from two to eight at 5' of the mature miRNA;
2) Mature sequences: the rest of mature miRNAs out of seed regions;
3) Precursor regions: the regions out of mature miRNAs.

We annotated substructures closer to 5' end as "5p" and those closer to 3' end as "3p". When this information was not available, we calculated the distance of a miRNA substructure from both 5' and 3' end.

MiRNAs that reside on opposite strands, but in the same genomic trait, were analysed separately (e.g. miR3116-1 and miR3116-2, which genomics coordinates are respectively chr1:62544458-62544531 and chr1:62544461-62544528). Therefore, in these cases, eventual variants identified were annotated in respect of substructures defined for both miRNAs.

To analyse only reliable variants, miRNA SNVs flagged as "PASS" with an overall Allel Depth (AD) equal or greater than 10 were considered. MiRNA variants identified were therefore analysed considering their segregation: homozygous, heterozygous and, for 66 probands for which we had parents, "de novo" variants. Then, variants were analysed considering localisation in miRNA substructures, i.e. seed, mature and precursor regions and their distribution was normalised on length of single substructures.

3.2.1 Experimental validation

Next, we selected some miRNA variants, identified in the cohort, to be experimentally validated. Selected miRNA variants were amplified by PCR (GoTaq Flexi DNA Polymerase – Promega) and analysed using Sanger sequencing (ABI BigDye Terminator Sequencing Kit V.3.1, ABI Prism 3500 Genetic Analyzer). Details on experimental validations performed are reported in Table 4.

Table 4. Details on miRNA variants experimentally validated. Conditions related to PCR and Sanger sequencing parameters are reported along with primers used for both experiments.

MiRNA	Sequences (5'-3')	PCR parameters	Sequencing Reaction parameters
MIR146A	FW: TCATGAGTGCCAGGACTAGAC REV: TCTCACAGGAACTCACACTCC	95°C - 2 min, 95°C - 30 sec/ 62°C - 30 sec / 72°C - 40 sec 30 cycles 72°C - 5 min 4°C - 5 min	96°C - 1 min, 96°C - 15 sec/ 58°C - 5 sec / 60°C - 4 min 25 cycles 4°C - 5 min
MIR202	FW: CTGGACCACAGGTAAGACGAG REV: ACGTCCTCCCCAGACACTTC	95°C - 2 min, 95°C - 30 sec/ 62°C - 30 sec / 72°C - 40 sec 30 cycles 72°C - 5 min 4°C - 5 min	96°C - 1 min, 96°C - 15 sec/ 58°C - 5 sec / 60°C - 4 min 25 cycles 4°C - 5 min
MIR938	FW: TCATTCTGGCAGTGAACACTTC REV: GTTGGGATCACCACCAGTTCG	95°C - 2 min, 95°C - 30 sec/ 62°C - 30 sec / 72°C - 40 sec 30 cycles 72°C - 5 min 4°C - 5 min	96°C - 1 min, 96°C - 15 sec/ 58°C - 5 sec / 60°C - 4 min 25 cycles 4°C - 5 min
MIR4634	FW: ATGAAGGCGAATCGCAGCCTC REV: TCCACCCAGAACCTCTGGTC	95°C - 2 min, 95°C - 30 sec/ 62°C - 30 sec / 72°C - 40 sec 30 cycles 72°C - 5 min 4°C - 5 min	96°C - 1 min, 96°C - 15 sec/ 58°C - 5 sec / 60°C - 4 min 25 cycles 4°C - 5 min
MIRLET7C	FW: TCCTTGCCAAGCCCTTAGGTG REV: AGTGACAACCCATTAGAAATACC	95°C - 2 min, 95°C - 30 sec/ 62°C - 30 sec / 72°C - 40 sec 30 cycles 72°C - 5 min 4°C - 5 min	96°C - 1 min, 96°C - 15 sec/ 58°C - 5 sec / 60°C - 4 min 25 cycles 4°C - 5 min

3.2.2 Functional annotation

Finally, we added functional annotation regarding miRNA variants and miRNAs to better characterise their potential biological role. For variants, we annotated information using tools regarding variants frequencies (gnomAD (version 2.0.2)[36] and dbSNP 150[34]) and their potential deleterious effect (CADD v 1.4[58] and DANN[59]). For miRNAs functional annotation,

we added information relative to miRNAs associated with human diseases using data from HMDD v3.0 database (the Human microRNA Disease Database version 3.0)[90].

To understand how functional annotation could help in the elucidation of miRNA variants and miRNAs potentially related to RGDs, we first annotated miRNA variants already associated with Mendelian diseases and reported in Table 5. Subsequently, we annotated miRNA variants identified in our cohort.

Table 5. MiRNA variants associated with Mendelian diseases.

Genomic coordinate	MiRNA	OMIM phenotype	Reference
chr7:129414596 G/T	miR96	613074	Mencía et al., 2009[77]
chr7:129414597 C/T	miR96	613074	Mencía et al., 2009[77]
chr7:129414553 A/G	miR96	613074	Soldà et al., 2012[78]
chr9:73424964 G/A	miR204	616722	Conte et al., 2015[79]
chr15:79502186 C/T	miR184	614303	Hughes et al., 2011[80]; Iliff et al., 2012[81]
chr15:79502137 C/A	miR184	614303	Lechner et al., 2013[82]
chr15:79502132 A/G	miR184	614303	Lechner et al., 2013[82]

3.3 Development of a dedicated microRNAs analysis tool

To date, available systems that annotate WES data do not allow to properly analyse miRNA variants. We therefore decided to integrate the script used to identify miRNA variants, developing a dedicated tool, "AnnomiR" ("Annotation of miR"), to retrieve and annotate miRNA-related information from WES data. Starting from a VCF file, "AnnomiR" searches for variants localising in miRNAs, specifying their location, based on information contained in miRBase[25]. Furthermore, retrieving information from several databases (i.e. gnomAD[36], dbSNP[34], CADD[58], DANN[59] and HMDD[90]) downloaded locally, "AnnomiR" also performs functional annotation of miRNA variants and of miRNAs, annotating variants frequency and potential deleterious effect, and already known associations of a miRNA with human diseases.

4. Results

4.1 Whole Exome Sequencing and microRNAs: capture evaluation

WES sequencing experiments generally produce a large amount of data that require both technical and biological skills to be analysed. In the study of RGDs, several assumptions are made to discriminate potentially pathogenetic variants among all the thousands of identified variants. Since 85% of disease-causative mutations in Mendelian diseases resides in protein coding-regions[19], standard WES analyses generally focus only on protein-coding variants that alter coding-sequences (e.g. non-synonymous or splice sites). However, due to biological and technical issues, diagnostic rate of this approach is attested to be ~20-30%[16,18]. Nevertheless, WES data could contain other meaningful information that could be useful in the identification of the molecular bases underlying RGDs, and that is currently discarded.

In this context, we investigated whether WES data could contain information related to miRNA sequences, as reported from preliminary evidences[79,84]. To this aim, we evaluated both theoretical and experimental coverage of miRNA sequences using several exome enrichment capture systems commercially available.

4.1.1 Theoretical coverage

First, we assessed whether exome enrichment capture systems commonly used to perform WES experiments, contain target regions specifically designed to capture miRNA sequences. Among the exome enrichment capture systems analysed, three declare the presence of specific baits for miRNA sequences (Table 6).

Table 6. Exome enrichment capture systems included in this analysis. Table shows technical details on design of target regions.
*This version of miRBase has been designed on hg38/GRCh38 assembly.

Exome enrichment capture system	Designed on assembly	Target size (Mb)	Target for miRNAs sequences
SeqCap EZ Human Exome Library v3.0 (Roche NimbleGen)	hg19/GRCh37	64	Declared in technical sheet, designed on miRBase v16
SeqCap EZ MedExome (Roche NimbleGen)	hg38/GRCh38	47	Declared in technical sheet, designed on miRBase v21*
Nextera Rapid Capture Exome (Illumina)	hg19/GRCh37	37	Not declared
SureSelect Human All Exon V4 (Agilent Technologies)	hg19/GRCh37	51	Declared in technical sheet, designed on miRBase v17
SureSelect Clinical Research Exome (Agilent Technologies)	hg19/GRCh37	54	Not declared
SureSelect Clinical Research Exome V2 (Agilent Technologies)	hg19/GRCh37	67	Not declared
SureSelect Human All Exon V6 (Agilent Technologies)	hg19/GRCh37	60	Not declared

We therefore assessed whether exome kits analysed could contain probes specifically targeting miRNA sequences, even if not reported in technical sheets. To this aim, we calculated theoretical coverage of miRNAs, evaluating the overlap between target regions of exome enrichment capture systems considered and miRNA sequences. Results are reported in Figure 6 and Table 7. As it can be observed, even if in a variable quota, all exome kits analysed, present target regions specifically designed on miRNA sequences.

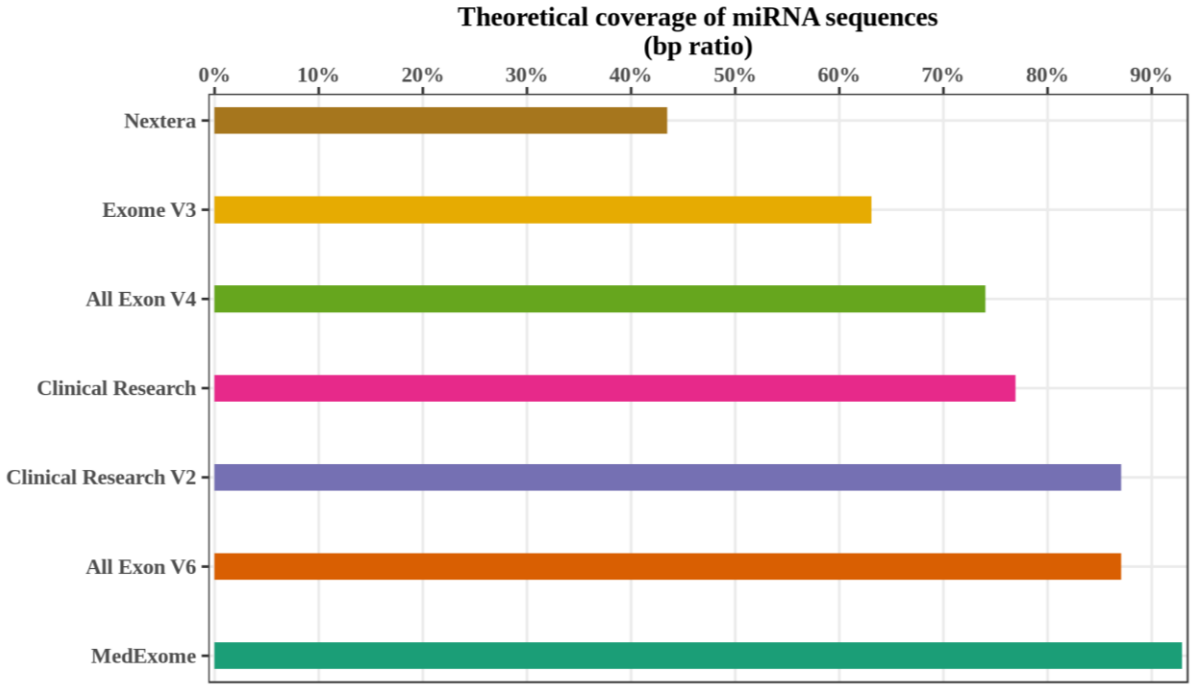

Figure 6. Theoretical coverage of miRNA regions in exome enrichment capture systems currently used. Histogram represents the theoretical coverage of miRNA "primary transcript" sequences as reported in miRBase v20[25] among several exome kits. Coverage has been evaluated as the ratio between the extension of miRNA sequences overlapping with target regions and the total extension of miRNA sequences (both evaluated in base pairs). For complete names of exome enrichment capture systems considered, see Table 7.

Table 7. MiRNAs specifically targeted by exome enrichment capture systems considered.

Exome enrichment capture system	Number of miRNA specifically targeted
Nextera Rapid Capture Exome (Illumina)	764
SeqCap EZ Human Exome Library v3.0 (Roche NimbleGen)	1154
SureSelect Human All Exon V4 (Agilent Technologies)	1383
SureSelect Clinical Research Exome (Agilent Technologies)	1443
SureSelect Clinical Research Exome V2 (Agilent Technologies)	1634
SureSelect Human All Exon V6 (Agilent Technologies)	1634
SeqCap EZ MedExome (Roche NimbleGen)	1743

4.1.2 Experimental coverage

Next, we evaluated the effective capture of miRNA sequences from WES data. To this aim we analysed a representative cohort composed by 14 WES captured through 7 exome enrichment capture systems (2 WES for each exome kit, Table 3).

First, we assessed the quality and homogeneity of WES data considered, calculating the "on target" coverage for each WES performed. Results reported in Figure 7 show the

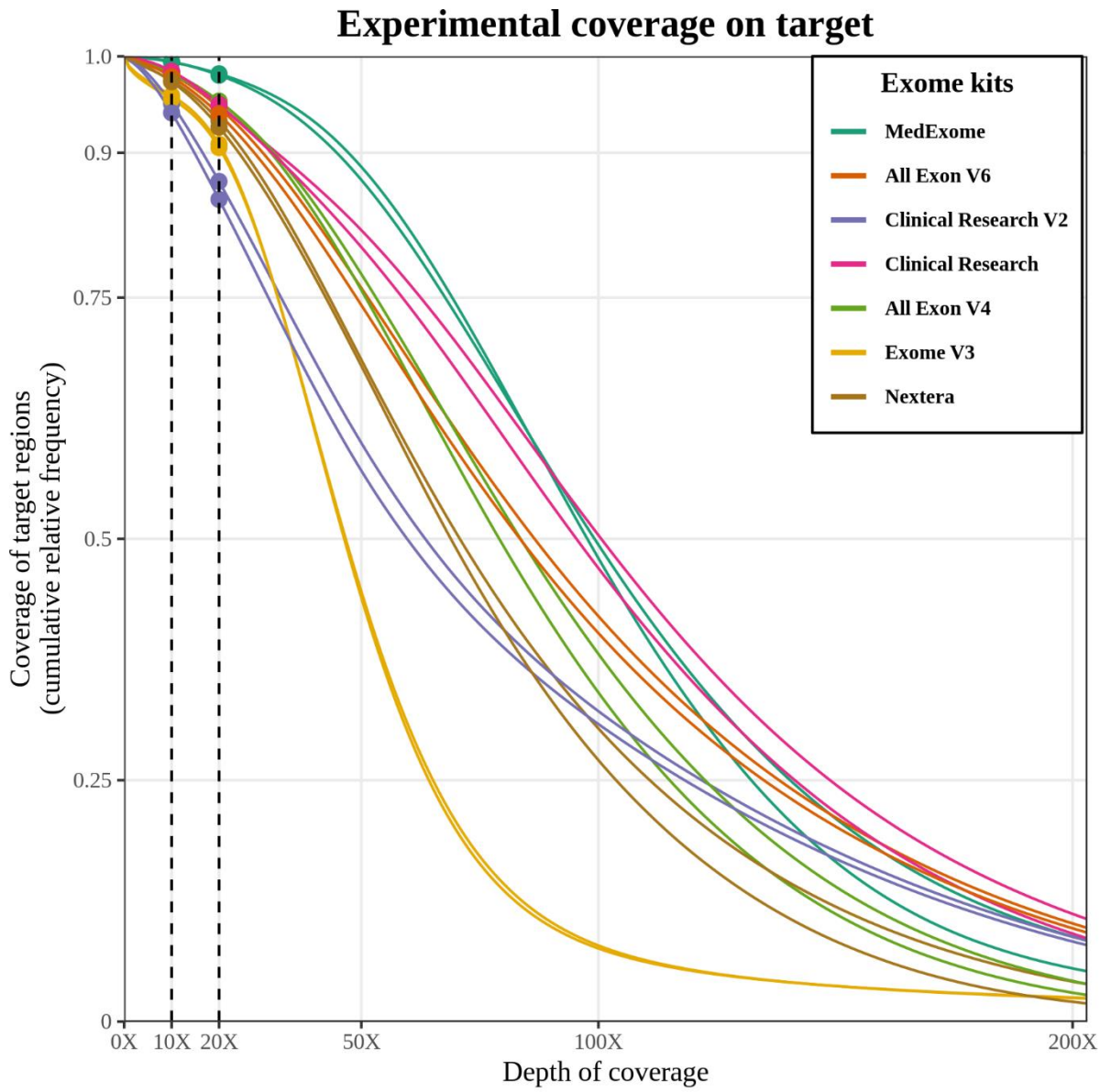

Figure 7. On target coverage of WES data considered. Figure reports the experimental "on target" coverage of each WES considered, obtained comparing each WES data with target regions of its own exome enrichment capture system. Values on y axis are calculated as cumulative relative frequency of target regions (measured in bp) effectively captured at a defined depth (reported on x axis). For complete names of exome enrichment capture systems considered, see Table 7.

uniformity of data analysed. As it can be observed, at a less stringent depth of 10X, all WES show a highly similar "on target" coverage, on average ~97% (ranging from 94% of SureSelect Clinical Research Exome V2 – Agilent Technologies to 99% of SeqCap EZ MedExome - Roche NimbleGen). At a more stringent depth of 20X, compatible with a highly reliable variant calling, "on target" coverage changes among WES data, but remains around 90% (from 86% obtained through SureSelect Clinical Research Exome V2 – Agilent Technologies to 98% reached through SeqCap EZ MedExome – Roche NimbleGen), attesting therefore the high quality of WES data.

Once we assessed the homogeneity of WES data, we evaluated experimental coverage of miRNA sequences. Results reported in Figure 8 show the high variability observed in the coverage of miRNA sequences. While differences among WES captured with same exome enrichment capture systems are almost no detectable (curves relative to same kits are almost overlapped), high variability can be encountered considering the different exome enrichment capture systems used to perform WES experiment. Indeed, at a depth of 10X, miRNAs coverage is 80% on average, spanning from 52% of Nextera Rapid Capture Exome (Illumina) to 96% of SeqCap EZ MedExome (Roche NimbleGen). At a depth of 20X, average miRNAs coverage is 73%, from 46% obtained through Nextera Rapid Capture Exome (Illumina) to 91% reached using SeqCap EZ MedExome (Roche NimbleGen).

Results obtained suggest that all exome enrichment capture systems considered can capture miRNA sequences, even if with a different efficiency. Overall, these results confirm that WES data contain information related to these non-coding species.

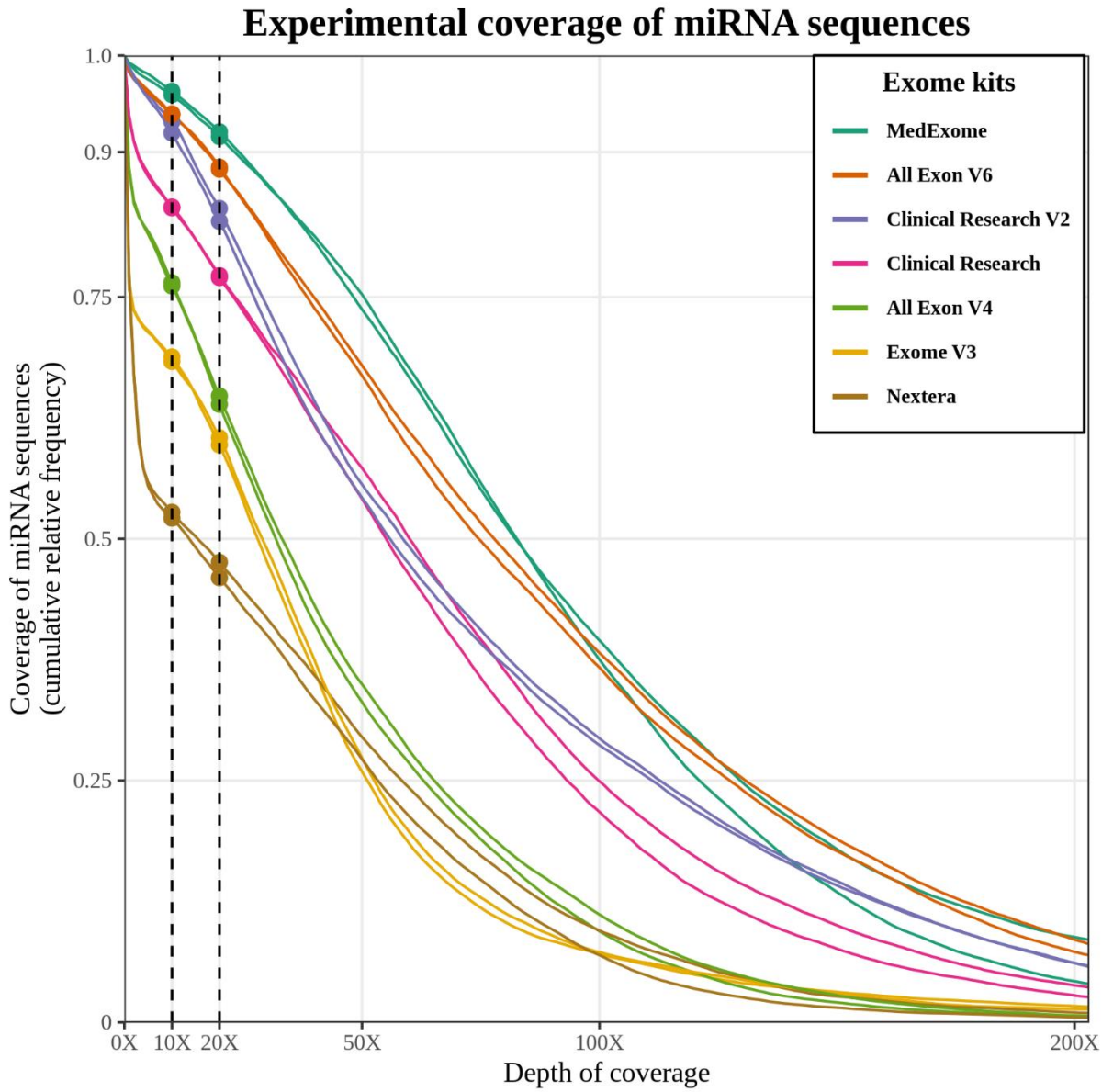

Figure 8. Experimental coverage of miRNA sequences. Figure reports the experimental coverage of miRNA sequences in each WES considered, obtained comparing each WES data with miRNA "primary transcript" sequences described in miRBase v20[25]. Values on y axis are calculated as cumulative relative frequency of miRNA sequences (measured in bp) effectively captured at a defined depth (reported on x axis). For complete names of exome enrichment capture systems considered, see Table 7.

4.1.2 Comparison with Whole Genome Sequencing data

Conversely from WES, WGS experiments do not rely on capture and enrichment of specific target regions, allowing to sequence all coding and non-coding sequences with a uniform coverage. Therefore, we analysed differences in miRNA sequences coverage between WES and WGS, comparing data coming from these two experiments. Specifically, we evaluated miRNA sequences coverage in cases 1 and 2 (Table 3), for which we performed both WES and WGS.

As it can be observed from Figure 9, WES and WGS present a different efficiency in the coverage of miRNA sequences, spanning from 68% to 98% at a depth of 10X for WES and WGS respectively. Differences increase if a depth of 20X is considered: while for WGS experiments a coverage of 93% can be observed, corresponding value for WES is 60%. Results obtained in the representative cohort suggest that WGS can sequence miRNA regions with a more uniform coverage at a greater depth compared to WES.

Since the number of WES and WGS analysed was very limited (2 WES and 2 WGS), we extended the comparison to publicly available database gnomAD[36], comprehending 123,136 WES and 15,496 WGS. Coverage data publicly available for gnomAD WES data are available at a single nucleotide level and contain statistics on coverage measured on all individuals sequenced. GnomAD WGS data contain data for all genomic coordinates of the human genome, while gnomAD WES data contain only data relative to exonic regions as defined by exome calling intervals provided by gnomAD database[36].

Therefore, we first assessed whether exome calling intervals used to process gnomAD WES, contain miRNA sequences, evaluating the overlap between exome calling intervals and miRNA sequences. We found that gnomAD WES have been analysed considering all miRNA "primary transcript" sequences contained in miRBase v20[25].

Next, we evaluated miRNA sequences coverage for gnomAD WES and WGS. Results reported in Figure 10 show differences between WES and WGS: at a depth of 10X, coverage is 53% for WES and 97% for WGS, while at a depth of 20X coverage values are respectively 48% and 93%.

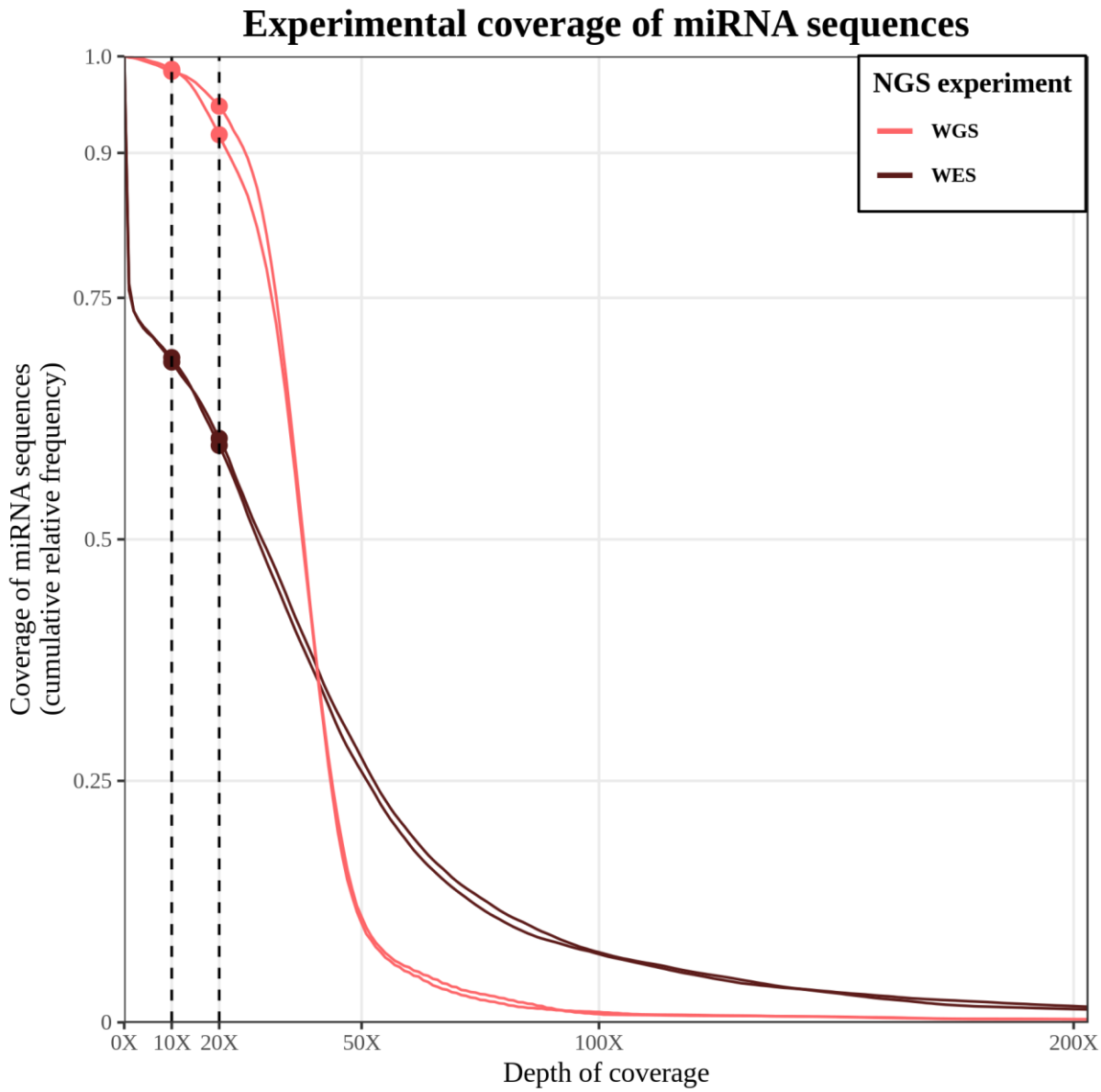

Figure 9. Experimental coverage of miRNA sequences in WES and WGS data in two cases. Figure reports the experimental coverage of miRNA "primary transcript" sequences (described in miRBase v20[25] between WES and WGS experiments performed for cases 1 and 2 (Table 3). Values on y axis are calculated as cumulative relative frequency of miRNA sequences (measured in bp) effectively captured at a defined depth (reported on x axis).

These results indicate that miRNA regions are fully considered to evaluate coverage in gnomAD WES, and that WGS show higher and more uniform coverage of miRNA regions.

Overall, comparison between WES and WGS confirms that WGS allow to obtain better results on sequencing of miRNA regions compared to WES. Nevertheless, results also confirm that WES contains information of miRNA sequences that is currently discarded by standard WES workflow of analysis.

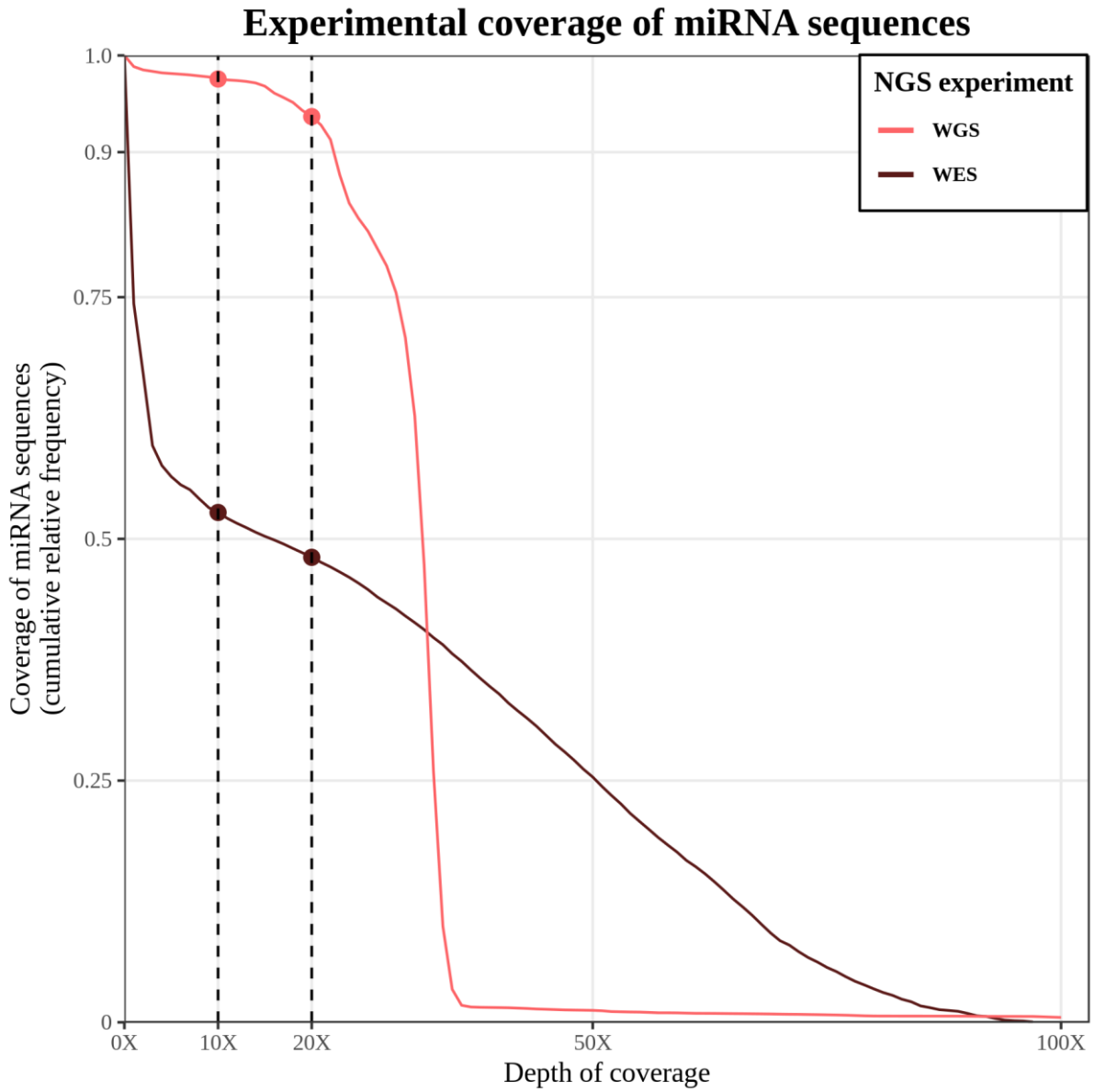

Figure 10. Experimental coverage of miRNA sequences in WES and WGS data in a publicly available cohort. Figure reports the experimental coverage of miRNA "primary transcript" sequences (described in miRBase v20[25]) between WES and WGS experiments performed respectively on 123,136 WES and 15,496 WGS contained in gnomAD[36]. Values on y axis are calculated as cumulative relative frequency of miRNA sequences (measured in bp) effectively captured at a defined depth (reported on x axis).

4.2 Evaluation of microRNA variants in a cohort

Once we assessed that WES data contain miRNA-related information, we analysed miRNA variants in a heterogeneous cohort of 259 individuals including 110 probands (11 pairs of siblings) and 149 unaffected relatives (parents and siblings), sequenced through WES experiments.

Based on information contained in miRBase[25], we developed a script to detect variants localising in miRNA sequences and retrieve respective information on miRNAs name and ids.

We also annotated miRNA variants localisation, considering the substructures in which variants reside (as reported in Figure 11) and their distance from the 5′ and 3′

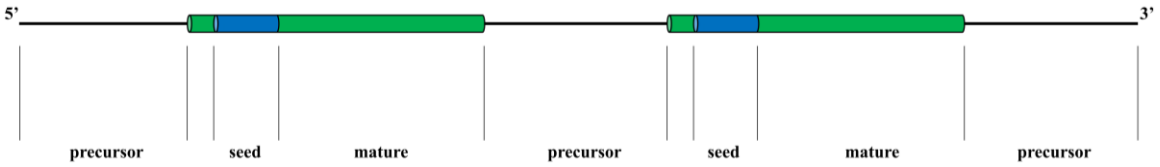

Figure 11. Schematic representation of miRNA substructures. We annotated miRNA variants considering following substructures: seed (shown in blue), mature (in green) and precursor regions (in black).

extremities. Therefore, miRNA variants closer to 5′ end, were annotated as "precursor-", "seed-", or "mature-" followed by the suffix "5p", while, for variants closer to 3′ end, suffix was "3p".

In cases in which a miRNA variant localised in two miRNAs in the same genomic region but on opposite strands, we annotated the single variant reporting information relative to both miRNAs. An example of the annotation performed is reported in Figure 12.

To analyse only reliable variants, we selected only miRNA high quality (i.e. "PASS") SNVs identified with an overall AD greater than 10.

On 99 probands analysed, we identified 555 miRNA SNVs. Of these variants, ~70% (385) were found exclusively in heterozygosis, while 1% (5 variants) in homozygosis or hemizygosis. The remaining 165 variants were found both in heterozygosis and homozygosis.

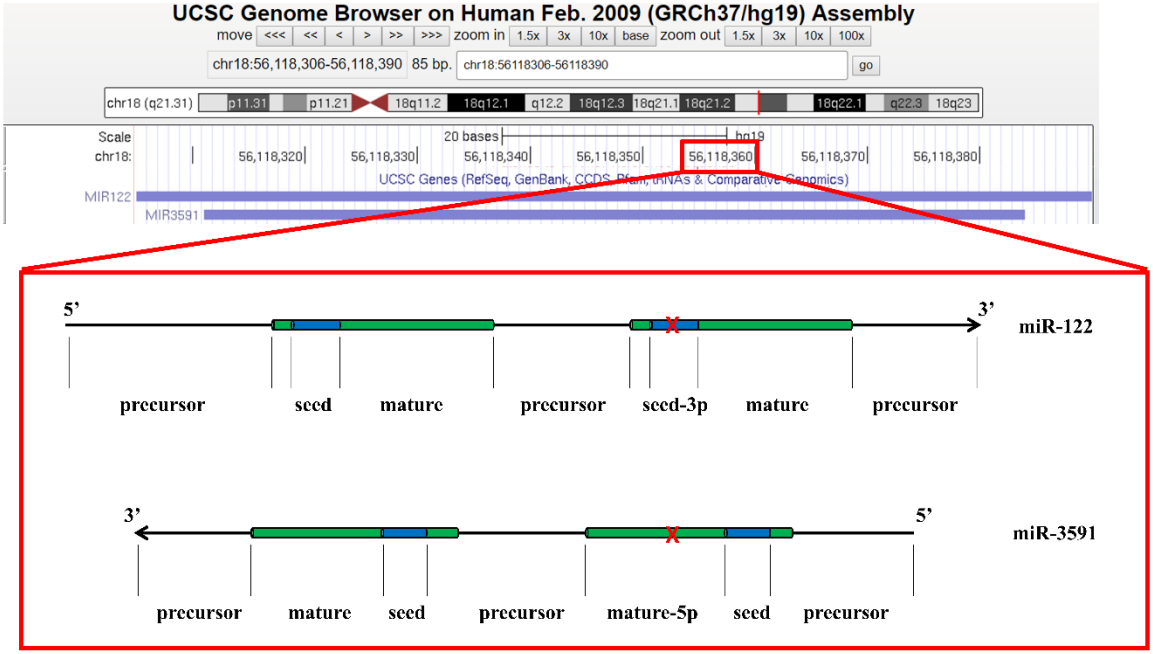

Figure 12. Annotation of MiRNA variants localising in two miRNAs on opposite strands. MiRNA variants localising in two miRNAs on opposite strands, but in the same genomic trait, were annotated using information on both miRNAs. The figure shows an example for variant chr18:56118360. As it can be observed, variant fell in both miR-122 and miR-3591 and we therefore annotated it as seed-3p on miR-122 and mature-5p on miR-3591.

For 66 of the 99 probands analysed we also had unaffected parents. We therefore analysed "de novo" variants identified in these cases, finding 49 variants. Of these variants, 7 were identified in two patients, therefore, overall, we found 42 "de novo" unique variants. We also analysed miRNA variants in respect of their localisation in miRNA substructures. We found 71 variants in seed regions, 109 in mature regions and 375 in precursor regions; their distribution is reported in Figure 13. Results obtained suggest that mature and seed regions present a low number of miRNA variants compared to precursor regions, probably reflecting a different sequence conservation.

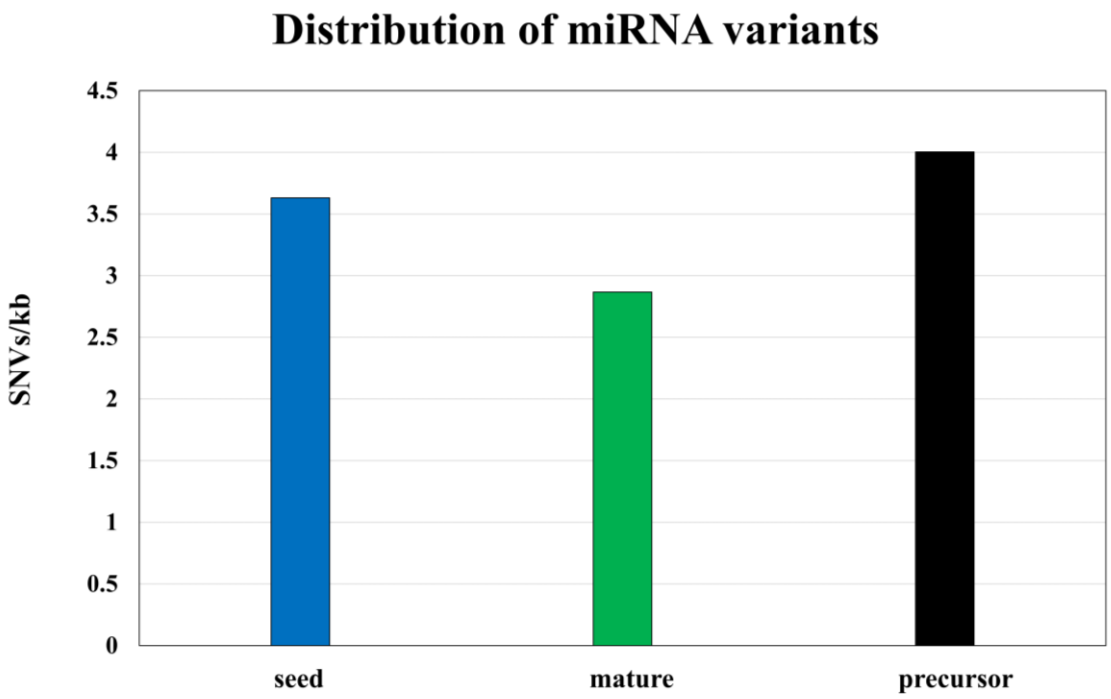

Figure 13. Distribution of miRNA variants found in our cohort on seed, mature and precursor regions. Distribution has been calculated normalising miRNA variants found to length of single substructures.

4.2.1 Experimental validation

Among miRNA SNVs identified we selected some of them to be experimentally validated. Due to the high identity of sequence observed in miRNAs, we expected that some of them could be false positive variants. We therefore tested some of variants identified. Results (Figure 14) show that all selected miRNA variants were confirmed, therefore indicating that our detection method could be considered a reliable system to identify variants lying in these regions.

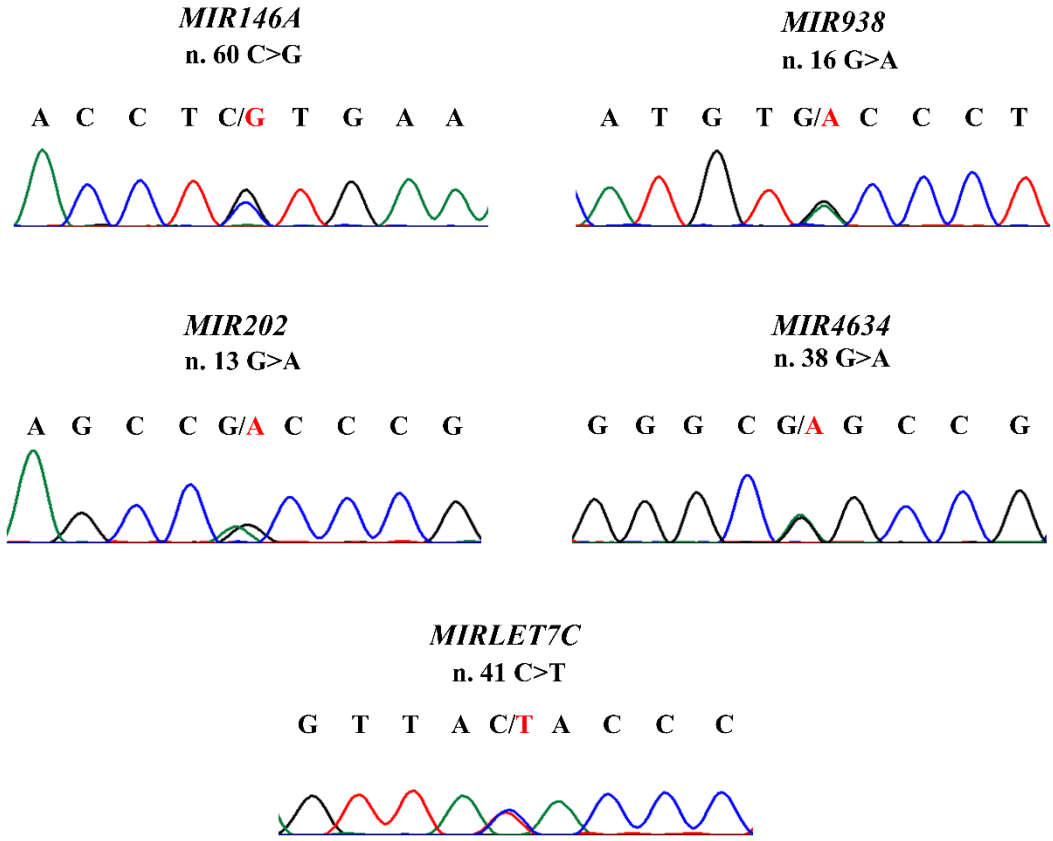

Figure 14. Sanger validation of miRNA variants selected.

4.2.2 Functional annotation

To better characterise the potential biological role of miRNA variants, we added functional annotation retrieved from several databases and tools. First, we annotated data on variants frequencies, using gnomAD[36] and dbSNP[34] databases. To predict potential deleterious effect of identified variants, we added information from scoring systems already used for analysis of NGS data that can score non-coding variants, i.e. CADD[58] and DANN[59] systems.

Finally, to predict potential involvement of miRNAs in human diseases, we used data contained in HMDD[90]. To test how functional annotation could help in the identification of miRNA variants and miRNAs potentially related to human diseases, we first annotated miRNA variants already associated to Mendelian diseases. Results are shown in Table 8. Five out of seven variants considered fell in seed regions while the other two localised in precursor regions, indicating that they could differently interfere with miRNA biogenesis and/or targeting. All variants analysed were not reported in public databases gnomAD[36] and dbSNP[34] or were annotated as rare (frequency ≤ 0.01%). CADD[58] and DANN[59] scores were respectively

greater than 10 and 0.8 (used as thresholds for non-coding variants), predicting therefore a potential deleterious effect of miRNA variants. Finally, information retrieved from HMDD[90] report the association of miRNAs with several diseases, allowing to hypothesize the involvement of these miRNAs in several biological pathways, along with those already known.

Overall results indicate that functional annotations allow to better characterise miRNA variants and miRNAs, leading to a proper investigation of their potential biological role.

As a second step, we performed a functional annotation on miRNA variants and miRNAs identified in our cohort, aimed at identifying candidate miRNAs potentially involved in the phenotypes observed in patients sequenced. Specifically, we are considering variants rare or not annotated in public databases and variants predicted to be potentially deleterious. Furthermore, variants and miRNAs are being analysed considering their potential implication in human diseases, trying to correlate them with available clinical information on patients.

Table 8. MiRNA variants annotation. Table reports miRNA variants and miRNAs with their respective functional annotations.

*HMDD annotations reported have been limited to human diseases already known to be associated with miRNAs analysed.

Genomic coordinate	MiRNA	MiRNA ID	Variant localisation	rs ID	gnomAD		dbSNP frequency	CADD	DANN	HMDD diseases*
					exomes	genomes				
chr7:129414596 G/T	miR96	MI0000098	seed-5p	.	.	.	.	22.4	0.9479	Sensorineural Hearing Loss
chr7:129414597 C/T	miR96	MI0000098	seed-5p	.	.	.	.	22.5	0.9485	Sensorineural Hearing Loss
chr7:129414553 A/G	miR96	MI0000098	seed-3p	rs546098287	0.002%	0.010%	0.0002%	22.3	0.8508	Sensorineural Hearing Loss
chr9:73424964 G/A	miR204	MI0000284	seed-5p	rs767146880	.	.	.	22	0.9227	Coloboma
chr15:79502186 C/T	miR184	MI0000481	seed-3p	.	.	.	.	22.5	0.9495	Cataract
chr15:79502137 C/A	miR184	MI0000481	precursor-5p	rs368718261	0.007%	0.010%	0.005%	14.78	0.9414	Cataract
chr15:79502132 A/G	miR184	MI0000481	precursor-5p	rs761900392	0.0004%	.	0.001%	14.32	0.8392	Cataract

4.3 Development of a dedicated microRNAs analysis tool

To date, there is not a dedicated tool to properly analyse miRNA-related information in WES data. In this context, we integrated the script used to identify miRNA variants and functional annotation provided for variants and miRNAs, developing a dedicated tool, "AnnomiR" ("Annotation of miR"). "AnnomiR" can analyse a VCF file searching for variants localising in miRNAs and specifying their location in miRNA substructures. "AnnomiR" also annotates miRNA variants, adding information on their frequencies and their potential deleterious effect, retrieved from several databases (gnomAD[36] and dbSNP[34], and CADD[58] and DANN[59] respectively) downloaded locally. Finally, "AnnomiR" annotates miRNAs considering information on miRNAs already known to be associated with human diseases using HMDD[90] database locally available.

"AnnomiR" can be easily integrated in a workflow for WES and WGS data processing, allowing to analyse miRNA regions, along with coding portion of the human genome, in a single step analysis.

5. Discussion

MicroRNAs (miRNAs) are small non-coding RNA that mediate gene silencing mostly recognising, by complementarity, 3'UTR of target mRNAs[71,72]. Expression of miRNAs is under tight control during development and it is subjected to a cell-specific regulation. Recently, miRNAs have been associated to several human diseases, comprehending both RGDs and complex traits[70].

Among known miRNAs dysregulation mechanisms, modifications of miRNAs expression profiles and sequence variants occurring in miRNAs or miRNAs-related genes have been disclosed. Alterations in miRNAs expression profiles can be due to several factors, such as changes in methylation of genes containing miRNA sequences, and responses to physiological and pathological stimuli, as steroid hormones or stress[91]. These modifications can significantly alter miRNAs expression, increasing or reducing miRNAs bioavailability, and have been so far associated with various diseases, mostly tumours[91]. Variants affecting miRNA biogenesis and function can occur in genes involved in miRNA machinery, in 3'UTR of target mRNAs and in miRNA sequences. Through the impairment of miRNA biogenesis and/or targeting, these sequence variants have been demonstrated to act both as disease-causative and phenotypic modifiers[92], in RGDs and multifactorial diseases[70].

In this context, comprehension of miRNAs alterations results crucial for the elucidation of molecular mechanisms that regulate onset and phenotypic variability underlying human diseases. Traditional methods as microarrays and quantitative reverse transcription PCR allow to study only expression profiles[83]. Information on miRNA sequences can be detected through Sanger sequencing and NGS approaches, such as Whole Genome Sequencing (WGS).

More recently, RNA sequencing (RNA-seq) experiments allowed to study simultaneously both miRNA expression profiles and sequences.

In this context, we evaluated exome enrichment capture designs and WES data of patients with different phenotypes in order to evaluate miRNA-related information.

As a first step, we evaluated theoretical coverage of miRNA sequences across most used exome enrichment capture systems. Then, we considered the experimental coverage of miRNA sequences, analysing a representative cohort of 14 WES captured through 7 exome enrichment capture systems. Results obtained from exome enrichment kit designs and WES

data analyses strongly suggest that all exome kits commercially available are designed to target miRNA sequences and that are able to efficiently capture these regions.

We also compared miRNAs coverage of WES with that of WGS, in 2 our cases and in a publicly available cohort (gnomAD database[36]). As expected, WGS can better sequence miRNA regions, showing a more uniform coverage, due to the fact that WGS experiments do not require a process of selection and enrichment of target regions. Data obtained from our 2 cases, are highly concordant with data obtained from sequencing of more than 100,000 individuals, contained in gnomAD database[36]. These results suggest that WES data contain valuable biological information that is usually non-considered using standard analysis workflow. Indeed, to date, no dedicated tool is available to retrieve information on miRNA sequences from WES and WGS experiments.

To retrieve variants localising in miRNAs, we developed a script based on information contained in miRBase[25]. We defined specific substructures of miRNAs: seed, mature and precursor regions. Discriminating miRNA variants based on their location could be helpful in elucidating their potential biological role as variants in miRNA precursors regions can alter miRNA biogenesis while variants in mature and seed regions can be associated with altered targeting of mRNAs.

We annotated a heterogeneous cohort of 259 WES, including patients affected by several genetic diseases and their unaffected relatives. The cohort represents an excellent system to study miRNA sequence variants as individuals were sequenced through the same exome enrichment capture system and sequencing platform. We specifically focused on probands available in this cohort, analysing 99 individuals. We identified 555 miRNA SNVs, retrieving information on their localisation in miRNA substructures. As already reported from previous studies[93], results obtained show that miRNA variants are more conserved in mature and seed regions compared to precursor regions (Figure 13). This confirms the crucial functional role of these regions. However, we cannot exclude that biological and technical factors could influence the analysis on miRNA variants distribution. First, according also to previous works[84,93], we are not considering a complete definition of miRNA substructures, e.g. loop regions, due to incomplete information available in miRBase[25]. Furthermore, results obtained are strictly related to exome enrichment capture system used to perform WES (i.e. SureSelect Human All Exon V4, Agilent Technologies) that include in its design ~1400

miRNAs (Table 7), therefore lacking ~500 miRNAs annotated in the version of miRBase used as reference (i.e. miRBase v20[25]). In addition, we cannot exclude that, besides their localisation, there could be other considerations to properly assess miRNA variants biological functions. An example could be represented by miRNAs structure conservation, that has been demonstrated to play a key role in regulation of miRNA biogenesis and/or function[94], and that it is not currently considered, due to the lack of tools that allow to systematically analyse it.

Since miRNAs are characterised by a high level of sequence identity, we evaluated whether called miRNA variants identified were, at least in part, false positives. Results obtained from experimental validations suggest instead that miRNA variants, identified through current calling variant algorithms (e.g. GATK[29]), are reliably called and, therefore, that these tools could be used to properly identify variants lying in these non-coding regions.

To characterise biological information on miRNA variants and miRNAs we added functional annotation. With the intent to understand whether information added could be helpful in discriminating miRNA candidates potentially related to RGDs, we first annotated miRNA variants already associated with Mendelian diseases. Functional annotation added on miRNA variants suggest that systems currently used to analyse and prioritise coding variants (variants frequencies and deleteriousness scoring systems) could be powerful for discriminating potentially biologically relevant miRNA variants.

The script we developed to identify miRNA variants and add functional annotation, "AnnomiR", could be integrated in a standard workflow of analysis for both WES and WGS data. "AnnomiR" could be used to integrate analysis of data coming from patients affected by several human diseases, not only by RGDs, as reported in this work, but even by complex traits, allowing to better elucidate miRNAs role in human diseases, and giving a more complete overview of variability of human genome.

6. Conclusions

Over the last few years, Next Generation Sequencing (NGS) technologies allowed to completely change the way to study Rare Genetic Diseases, accelerating the pace of discovery of their molecular bases.

Currently, sequencing of the exonic portion of the human genome – the exome (1%) – performed through Whole Exome Sequencing (WES) experiments represents the most used approach to characterise molecular mechanisms underlying RGDs. However, its diagnostic rate is attested to be ~20-30%. To date, several tools have been developed to analyse and interpret data generated from WES.

In this context, we evaluated whether WES data contain information on a non-coding portion of the human genome, i.e. microRNAs (miRNAs), since they have been demonstrated to play a key role in several human genetic diseases, acting both as disease-causative and phenotypic modifiers.

The development of a dedicated tool to identify and functionally annotate miRNA variants and miRNAs from WES and WGS will allow to analyse these regions from NGS data. We expect that systematic study of miRNAs will allow to elucidate their biological role in a wide spectrum of human diseases, leading to a better characterisation of the variability of the human genome related to these non-coding sequences.

7. Web sites

I. software.broadinstitute.org/gatk/best-practices/workflow?id=11145

II. bioinformatics.babraham.ac.uk/projects/trim_galore

III. novocraft.com/products/novoalign

IV. broadinstitute.github.io/picard

V. htslib.org

8. References

1. United States Congress. 2002. Rare Diseases Act of 2002. gpo.gov/fdsys/pkg/PLAW-107publ280/html/PLAW-107publ280.htm

2. The European Parliament and the Council of the European Union. 1999. Regulation (EC) No 141/2000 of the European parliament and of the council of 16 December 1999 on orphan medicinal products. ec.europa.eu/health//sites/health/files/files/eudralex/vol-1/reg_2000_141_cons-2009-07/reg_2000_141_cons-2009-07_en.pdf.

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

4. 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.

5. 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.

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

7. Gilissen C, Hoischen A, Brunner HG, Veltman JA. 2012. Disease gene identification strategies for exome sequencing. Eur J Hum Genet. *20*(5):490-497.

8. Pulst SM. 1999. Genetic linkage analysis. Arch Neurol. *56*(6):667-672.

9. Collins FS. 1995. Positional cloning moves from perditional to traditional. Nat Genet. *9*(4):347-350.

10. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL, Drumm ML, Iannuzzi MC, Collins FS, Tsui L. 1989. Identification of the cystic fibrosis gene:cloning and characterization of complementary DNA. Science. *245*(4922):1066-1073.

11. The Huntington's Disease Collaborative Research Group. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell. *72*(6):971-983.

12. Malkin D, Li FP, Strong LC, Fraumeni JF Jr, Nelson CE, Kim DH, Kassel J, Gryka MA, Bischoff FZ, Tainsky MA, Friend SH. 1990. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science. *250*(4985):1233-1238.

13. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. *12*(11):745-755.

14. 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.

15. 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 H, Hetrick K, Watkins L, Patterson KE, Reinier F, Blue E, Muzny D, Kircher M, Bilguvar K, López-Giráldez F, Sutton VR, Tabor HK, Leal SM, Gunel M, Mane S, Gibbs RA, Boerwinkle E, 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.

16. 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.

17. Sun Y, Man J, Wan Y, Pan G, Du L, Li L, Yang Y, Qiu L, Gao Q, Dan H, Mao L, Cheng Z, Fan C, Yu J, Lin M, Kristiansen K, Shen Y, Wei X. 2018. Targeted next-generation sequencing as a comprehensive test for Mendelian diseases: a cohort diagnostic study. Sci Rep. *8*(1):11646.

18. 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.

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

20. Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. *17*(6):333-351.

21. Samorodnitsky E, Jewell BM, Hagopian R, Miya J, Wing MR, Lyon E, Damodaran S, Bhatt D, Reeser JW, Datta J, Roychowdhury S. 2015. Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing. Hum Mutat. *36*(9):903-914.

22. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*(D1):D733-D745.

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

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

25. Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. *42*(Database issue):D68-D73.

26. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. *30*(15):2114-2220.

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

28. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. *25*(16):2078-2079.

29. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky 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.

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

31. 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.

32. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. Genome Biol. *17*(1):122.

33. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, To JK, Laird MR, Lavidas I, Liu Z, Loveland JE, Maurel T, McLaren W, Moore B, Mudge J, Murphy DN, Newman V, Nuhn M, Ogeh D, Ong CK, Parker A, Patricio M, Riat HS, Schuilenburg H, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Zadissa A, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Cunningham F, Yates A, Flicek P. 2018. Ensembl 2018. Nucleic Acids Res. *46*(D1):D754-D761.

34. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*(1):308-311.

35. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic variation. Nature. *526*(7571):68-74.

36. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R,

Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG; Exome Aggregation Consortium. 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature. *536*(7616):285-291.

37. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*(1):110-121.

38. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. *6*(12):e1001025.

39. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN. 2003. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat. *21*(6):577-581.

40. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, Karapetyan K, Katz K, Liu C, Maddipatla Z, Malheiro A, McDaniel K, Ovetsky M, Riley G, Zhou G, Holmes JB, Kattman BL, Maglott DR. 2018. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46*(D1):D1062-D1067.

41. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) [30(10, 2018) accessed].

42. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*(8):1034-1050.

43. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztányi Z, El-Gebali S, Fraser M, Gough J, Haft D, Holliday GL, Huang H, Huang X, Letunic I, Lopez R, Lu S, Marchler-Bauer A, Mi H, Mistry J, Natale DA, Necci M, Nuka G, Orengo CA, Park Y, Pesseat S, Piovesan D, Potter SC, Rawlings ND, Redaschi N, Richardson L, Rivoire C, Sangrador-Vegas A, Sigrist C, Sillitoe I, Smithers B, Squizzato S, Sutton G, Thanki N, Thomas PD, Tosatto SC, Wu CH, Xenarios I, Yeh LS, Young SY, Mitchell AL. 2017. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res. *45*(D1):D190-D199.

44. Nishimura D. 2001. BioCarta. Biotech Software & Internet Report. *2*(3): 117–120.

45. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. 2013. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. *9*(8):e1003709.

46. Petryszak R, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, Füllgrabe A, Fuentes AM, Jupp S, Koskinen S, Mannion O, Huerta L, Megy K, Snow C, Williams E, Barzine M, Hastings E, Weisser H, Wright J, Jaiswal P, Huber W, Choudhary J, Parkinson HE, Brazma A. 2016. Expression Atlas update--an integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res. *44*(D1):D746-D752.

47. Huang N, Lee I, Marcotte EM, Hurles ME. 2010. Characterising and predicting haploinsufficiency in the human genome. PLoS Genet. *6*(10):e1001154.

48. Liu X, Wu C, Li C, Boerwinkle E. 2016. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat. *37*(3):235-241.

49. The UniProt Consortium. 2018. UniProt: the universal protein knowledgebase. Nucleic Acids Res. *46*(5):2699.

50. Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*(1):27-30.

51. Kamburov A, Wierling C, Lehrach H, Herwig R. 2009. ConsensusPathDB--a database for integrating human functional interaction networks. Nucleic Acids Res. *37*(Database issue):D623-D628.

52. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. *25*(1):25-9.

53. Smith CL, Blake JA, Kadin JA, Richardson JE, Bult CJ; Mouse Genome Database Group. 2018. Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse. Nucleic Acids Res. *46*(D1):D836-D842.

54. Howe DG, Bradford YM, Conlin T, Eagle AE, Fashena D, Frazer K, Knight J, Mani P, Martin R, Moxon SA, Paddock H, Pich C, Ramachandran S, Ruef BJ, Ruzicka L, Schaper K, Shao X, Singer A, Sprunger B, Van Slyke CE, Westerfield M. 2013. ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. Nucleic Acids Res. *41*(Database issue):D854-D860.

55. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), {09/06/2018}. World Wide Web URL: https://omim.org/

56. Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. Chapter 7:Unit7.20.

57. Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*(13):3812-3814.

58. 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.

59. Quang D, Chen Y, Xie X. 2015. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. *31*(5):761-763.

60. Jian X, Boerwinkle E, Liu X. 2014. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. *42*(22):13534-13544.

61. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, Morris Q, Barash Y, Krainer AR, Jojic N, Scherer SW, Blencowe BJ, Frey BJ. 2015. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science. *347*(6218):1254806.

62. Seelow D, Schwarz JM, Schuelke M. 2008. GeneDistiller--distilling candidate genes from linkage intervals. PLoS One. *3*(12):e3874.

63. Yang H, Robinson PN, Wang K. 2015. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods. *12*(9):841-843.

64. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, Baynam G, Bello SM, Boerkoel CF, Boycott KM, Brudno M, Buske OJ, Chinnery PF, Cipriani V, Connell LE, Dawkins HJ, DeMare LE, Devereau AD, de Vries BB, Firth HV, Freson K, Greene D, Hamosh A, Helbig I, Hum C, Jähn JA, James R, Krause R, F Laulederkind SJ, Lochmüller H, Lyon GJ, Ogishima S, Olry A, Ouwehand WH, Pontikos N, Rath A, Schaefer F, Scott RH, Segal M, Sergouniotis PI, Sever R, Smith CL, Straub V, Thompson R, Turner C, Turro E, Veltman MW, Vulliamy T, Yu J, von Ziegenweidt J, Zankl A, Züchner S, Zemojtel T, Jacobsen JO, Groza T, Smedley D, Mungall CJ, Haendel M, Robinson PN. 2017. The Human Phenotype Ontology in 2017. Nucleic Acids Res. *45*(D1):D865-D876.

65. Li Q, Wang K. 2017. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. Am J Hum Genet. *100*(2):267-280.

66. Nykamp K, Anderson M, Powers M, Garcia J, Herrera B, Ho YY, Kobayashi Y, Patil N, Thusberg J, Westbrook M; Invitae Clinical Genomics Group, Topper S. 2017. Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. Genet Med. *19*(10):1105-1117.

67. 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.

68. Samuels DC, Han L, Li J, Quanghu S, Clark TA, Shyr Y, Guo Y. 2013. Finding the lost treasures in exome sequencing data. Trends Genet. *29*(10):593-599.

69. Bergant G, Maver A, Lovrecic L, Čuturilo G, Hodzic A, Peterlin B. 2018. Comprehensive use of extended exome analysis improves diagnostic yield in rare disease: a retrospective survey in 1,059 cases. Genet Med. *20*(3):303-312.

70. Kawahara Y. 2014. Human diseases caused by germline and somatic abnormalities in microRNA and microRNA-related genes. Congenit Anom (Kyoto). *54*(1):12-21.

71. Cammaerts S, Strazisar M, De Rijk P, Del Favero J. 2015. Genetic variants in microRNA genes: impact on microRNA expression, function, and disease. Front Genet. 6:186.

72. Gebert LFR, MacRae IJ. 2018. Regulation of microRNA function in animals. Nat Rev Mol Cell Biol. [Epub ahead of print]

73. Ha M, Kim VN. 2014. Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol. *15*(8):509-524.

74. Lin S, Gregory RI. 2015. MicroRNA biogenesis pathways in cancer. Nat Rev Cancer. *15*(6):321-333.

75. Hill DA, Ivanovich J, Priest JR, Gurnett CA, Dehner LP, Desruisseau D, Jarzembowski JA, Wikenheiser-Brokamp KA, Suarez BK, Whelan AJ, Williams G, Bracamontes D, Messinger Y, Goodfellow PJ. 2009. DICER1 mutations in familial pleuropulmonary blastoma. Science. *325*(5943):965.

76. Abelson JF, Kwan KY, O'Roak BJ, Baek DY, Stillman AA, Morgan TM, Mathews CA, Pauls DL, Rasin MR, Gunel M, Davis NR, Ercan-Sencicek AG, Guez DH, Spertus JA, Leckman JF, Dure LS 4th, Kurlan R, Singer HS, Gilbert DL, Farhi A, Louvi A, Lifton RP, Sestan N, State MW. 2005. Sequence variants in SLITRK1 are associated with Tourette's syndrome. Science. *310*(5746):317-320.

77. Mencía A, Modamio-Høybjør S, Redshaw N, Morín M, Mayo-Merino F, Olavarrieta L, Aguirre LA, del Castillo I, Steel KP, Dalmay T, Moreno F, Moreno-Pelayo MA. 2009. Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. Nat Genet. *41*(5):609-613.

78. Soldà G, Robusto M, Primignani P, Castorina P, Benzoni E, Cesarani A, Ambrosetti U, Asselta R, Duga S. 2012. A novel mutation within the MIR96 gene causes non-syndromic inherited hearing loss in an Italian family by altering pre-miRNA processing. Hum Mol Genet. *21*(3):577-585.

79. Conte I, Hadfield KD, Barbato S, Carrella S, Pizzo M, Bhat RS, Carissimo A, Karali M, Porter LF, Urquhart J, Hateley S, O'Sullivan J, Manson FD, Neuhauss SC, Banfi S, Black GC. 2015. MiR-204 is responsible for inherited retinal dystrophy associated with ocular coloboma. Proc Natl Acad Sci U S A. *112*(25):E3236-3245.

80. Hughes AE, Bradley DT, Campbell M, Lechner J, Dash DP, Simpson DA, Willoughby CE. 2011. Mutation altering the miR-184 seed region causes familial keratoconus with cataract. Am J Hum Genet. *89*(5):628-633.

81. Iliff BW, Riazuddin SA, Gottsch JD. 2012. A single-base substitution in the seed region of miR-184 causes EDICT syndrome. Invest Ophthalmol Vis Sci. *53*(1):348-53.

82. Lechner J, Bae HA, Guduric-Fuchs J, Rice A, Govindarajan G, Siddiqui S, Abi Farraj L, Yip SP, Yap M, Das M, Souzeau E, Coster D, Mills RA, Lindsay R, Phillips T, Mitchell P, Ali M, Inglehearn CF, Sundaresan P, Craig JE, Simpson DA, Burdon KP, Willoughby CE. 2013. Mutational analysis of MIR184 in sporadic keratoconus and myopia. Invest Ophthalmol Vis Sci. *54*(8):5266-5272.

83. Pritchard CC, Cheng HH, Tewari M. 2012. MicroRNA profiling: approaches and considerations. Nat Rev Genet. *13*(5):358-369.

84. Carbonell J, Alloza E, Arce P, Borrego S, Santoyo J, Ruiz-Ferrer M, Medina I, Jiménez-Almazán J, Méndez-Vidal C, González-Del Pozo M, Vela A, Bhattacharya SS, Antiñolo G, Dopazo J. 2012. A map of human microRNA variation uncovers unexpectedly high levels of variability. Genome Med. *4*(8):62.

85. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. *26*(6):841-842.

86. Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

87. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN]

88. R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

89. van Rossum G, de Boer J. 1991. Interactively Testing Remote Servers Using the Python Programming Language. CWI Quarterly. *4*(4):283–303.

90. Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, Zhou Y, Cui Q. 2018. HMDD v3.0: a database for experimentally supported human microRNA-disease associations. Nucleic Acids Res. [Epub ahead of print]

91. Gulyaeva LF, Kushlinskiy NE. 2016. Regulatory mechanisms of microRNA expression. J Transl Med. *14*(1):143

92. Bandiera S, Hatem E, Lyonnet S, Henrion-Caude A. 2010. MicroRNAs in diseases: from candidate to modifier genes. Clin Genet. *77*(4):306-313.

93. Torruella-Loran I, Laayouni H, Dobon B, Gallego A, Balcells I, Garcia-Ramallo E, Espinosa-Parrilla Y. 2016. MicroRNA Genetic Variation: From Population Analysis to

Functional Implications of Three Allele Variants Associated with Cancer. Hum Mutat. *37*(10):1060-1073.

94. Fernandez N, Cordiner RA, Young RS, Hug N, Macias S, Cáceres JF. 2017. Genetic variation and RNA structure regulate microRNA biogenesis. Nat Commun. *8*:15114.