SAPIENZA
Università di Roma

Dipartimento di Medicina Sperimentale e Patologia
Dottorato di Ricerca in Genetica Medica

EXOME SEQUENCING IN MENDELIAN DISEASES

PhD student:
Valentina Di Pierro

Tutor:
Prof. **Marco Tartaglia**

XXV ciclo

INDEX

INTRODUCTION

In 1977 Fred Sanger published the method for a rapid determination of DNA sequence, which would go on to transform biology as a whole by providing a tool for deciphering complete genes and later entire genomes [1]. Subsequently, the introduction of reduced handling of toxic chemicals and radioisotopes rapidly made Sanger sequencing the only DNA sequencing method used for the next 30 years.

The paradigm of DNA sequencing changed with the advent of the first forms of Next-generation sequencing (NGS) that greatly reduced the necessary reaction volume while dramatically extended the number of sequencing reactions [2, 3]. NGS technologies constitute various strategies that rely on a combination of template preparation, sequencing and imaging, and genome alignment and assembly methods, which make it possible to process hundreds of thousands to millions of DNA templates in parallel, resulting in a low cost per base of generated sequence and a throughput on the gigabase scale [4].

Primarily through linkage mapping and candidate gene resequencing, loci underlying about one-half to one-third of all known or suspected Mendelian disorders have been discovered [5]. Genome-wide linkage analysis followed by positional cloning have been very successful in identifying causal variants for Mendelian disorders because of the perfect segregation of the causal variant with the disease-phenotype according to Mendelian inheritance patterns, due to complete or almost-complete penetrance of the mutation. On the other hand, homozygosity mapping has been a more powerful and effective approach to study recessive disorders in consanguineous families [6]. However, for those disorders that are not amendable to these two conventional approaches, their causal variants remain elusive. Several factors limit the power of traditional gene-discovery strategies: for example, the availability of only a small number of cases or families to study, reduced penetrance, locus heterogeneity and substantially diminished reproductive fitness [7].

The development of methods for coupling targeted capture and massively parallel DNA sequencing has made it possible to determine cost-effectively nearly all of the coding variation present in an individual human genome, a process termed exome sequencing. This technique has become a powerful new approach for identifying genes that underlie Mendelian disorders in circumstances in which conventional approaches have failed [8]. Even where conventional approaches are eventually expected to succeed – for example, in homozygosity mapping – exome sequencing provides a means for accelerating discovery [9].

Despite the fundamental limitation that exome sequencing does not currently assess the impact of non-coding alleles, it is a well-justified strategy for discovering rare alleles underlying Mendelian phenotypes. The exome represents for several reasons a highly enriched subset of the genome in which to search for variants with large effect sizes. First, positional cloning studies focused on protein-coding sequences have, when adequately powered, proved to be highly successful at identifying variants for monogenic diseases. Second, most alleles that are known to underlie Mendelian disorders disrupt protein-coding sequences. Third, a large fraction of rare, protein–altering variants are predicted to have functional consequences and/or to be deleterious [10].

Thus, since protein coding genes constitute only approximately 1% of the human genome, but harbor 85% of the mutations with large effects on disease-related traits, efficient strategies for selectively sequencing complete coding regions - the whole exome - have the potential to contribute to the understanding of rare and common human diseases.

NEXT-GENERATION TECHNOLOGIES

Sequencing technologies include a number of methods that are grouped broadly as template preparation, sequencing and imaging, and data analysis.

1. Template preparation: target-enrichment strategies

Template preparation requires robust methods that produce a representative, non-biased source of nucleic acid material from the genome under investigation. Generally it involves randomly breaking genomic DNA into smaller sizes from which fragments templates or mate-pairs are created. While the fragment library is prepared by randomly shearing genomic DNA, mate-pair templates are obtained cutting the circularized sheared DNA selected for a given size.

Since 2007, there has been tremendous progress in the development of diverse technologies for capturing arbitrary subsets of mammalian genome at a scale commensurate with that of massive parallel sequencing [11]. Current techniques for targeted enrichment can be categorized according to the nature of their core reaction principle.

Polymerase-mediated capture

Although all capture methods use polymerase to amplify captured fragment, in this case polymerase chain reaction (PCR) is directed toward the targeted regions of interest by conducting multiple long-range PCRs in parallel, a limited number of standard multiplex PCRs or highly multiplexed PCR methods that amplify very large numbers of short fragments.

A strikingly elegant application is the micro-droplet PCR technology developed by RainDance, where each microdroplet can be made to contain a single primer pair along with genomic DNA and other reagents. The entire population of droplets represents thousands of distinct primer pairs and is subject to termal cycler, after which this emulsion is broken and products are recovered.

Hybrid capture

The hybrid capture principle is based upon the hybridization of selected fragments of DNA or RNA representing the target region against a shotgun library of DNA fragments from the genome to be enriched. Two alternative strategies are used to perform the hybrid capture: reactions on a solid support and in solution.

- Solid-phase hybridization methods generally utilize probes complementary to the sequences of interest affixed to a solid support, such as microarrays or filters. The total DNA is applied to the probes, where the desired fragments hybridize. The non-targeted fragments are subsequently washed away and the enriched DNA is eluted for sequencing. Even if it is quicker than PCR-based approaches, it is necessary to start library preparation with a large amount of DNA (around 10-15 μg) in order to obtained a suitable DNA library prepped.
- Liquid-phase hybridization is similar to solid phase, but in this method the probes are not attached to a solid matrix, but instead are biotinylated. Following hybridization, the biotinylated probes – with the complementary desired genomic DNA – are bound to magnetic streptavidin beads and are separated from the undesired DNA by washing. Whereas an on-array target enrichment uses a vast excess of DNA library over probes, solution capture has an excess of probes over template, which drives the hybridization reaction further to completion using a smaller quantity of sequencing library.

Molecular inversion probes (MIP)

In the MIP technique single-stranded oligonucleotides, consisting of common linker flanked by targeted sequences, anneal to their target sequence and become circularized by a ligase. Uncircularized species are digested by exonucleases to reduce background, while circularized species are PCR amplified via primers directed at the common linker.

All three major targeted enrichment techniques differ in terms of sample library preparation workflow enabling sequencing on any of the current NGS instruments. Enrichment by hybrid selection relies on short fragment library preparations which are generated before hybridization. In contrast, enrichment by PCR is performed directly on genomic DNA and thereafter the library primers are added. Enrichment by circularization offers the easiest library preparation for NGS because the sequencing primers can be added to the circularization probe, thus eliminating the need for any further preparation steps.

A series of metrics need to be considered in order to evaluate the performance of each target enrichment approach (Table 1):

- sensitivity, or the percentage of the target bases that are represented by one or more sequence reads;
- specificity, or the percentage of sequences that map to the intended targets;
- uniformity, or the variability in sequence coverage across target regions;
- reproducibility, or how closely results obtained from replicate experiments correlate;
- cost;
- ease of use;
- amount of DNA required per experiment, or per megabase of target.

	PCR	**On-array hybrid capture**	**In-solution hybrid capture**	**MIP**
Cost	High	Medium	Low	Low (>100 samples)
Ease of use	Low	Medium	High	High
Mass DNA	8 μg (5 kb amplicons)	10-15 μg (30 Mb target)	1-3 μg (50 Mb target)	200 ng
Sensitivity	> 99.5%	98.6%	> 99.5%	> 98%
Specificity	93%	70-80%	80-90%	> 98%
Uniformity	80%	60%	60-70%	60%
Reproducibility	100%	95%	96%	92%

Table 1. Performance of target-enrichment methods *(Adapted from Mamanova L et al. Nat Methods 7:111-118).*

A major obstacle for targeted enrichment is posed by repeating elements, including interspersed and tandem repeats as well as elements – such as pseudogenes – located within and outside the region of interest. Furthermore, at extreme values, the GC content of the target region has a considerable impact on the evenness and efficiency of the enrichment, affecting in particular the analysis of the 5'-UTR/promoter region and the first exon of genes. Therefore, expectations regarding the outcome of the experiment require the evaluation of the appropriate enrichment method [12].

1.1 Enrichment workflow in exome sequencing

Among the methods described for targeted capture, only the hybrid capture can handle large target regions, therefore in the last years the target of the human exome has largely converged on the in-solution capture by hybridization approach (Figure 1).

Genomic DNA is randomly sheared and 1-3 μg are used to construct an *in vitro* shotgun library, whose fragments are flanked by adaptors. Next, the library is enriched for sequences corresponding to exons: the fragments are hybridized to biotinylated DNA or RNA baits in the presence of blocking oligonucleotides that are complementary to the adaptors. Recovery of the hybridized fragments by biotin-streptavidin based pull-down is followed by amplification and massively parallel sequencing of the enriched, amplified library and the mapping and calling of candidate causal variants. It is possible to sequence more than one sample in a single sequencing lane introducing barcodes, that allow sample indexing, during the initial library construction or post-capture amplification. Key performance parameters include the degree of enrichment, the uniformity with which targets are captured and the molecular complexity of the enriched library.

In the workflow for exome sequencing there are some critical parameters with a large influence over the outcome of a target-enrichment experiment, including the fragment size, the PCR amplification and the pre-hybridization cleanup.

There are a variety of methods available to fragment nucleic acids, but the mechanical shearing remains the method of choice for achieving high sensitivity and unbiased results. Using a sonicator it is possible to generate a sufficiently narrow fragment-size distribution that the size-selection step can be omitted. It is very important to control the fragment size since longer fragments are captured with lower specificity than shorter ones because they contain a higher proportion of off-target sequence. There is also a lower size limit to

fragments for efficient capture, but in practice the minimum size is determined by the length of the wished final sequences. Longer reads would be expected to map to the reference sequence with lower ambiguity than shorter ones and can reduce the overrepresentation toward the ends of capture probes [13].

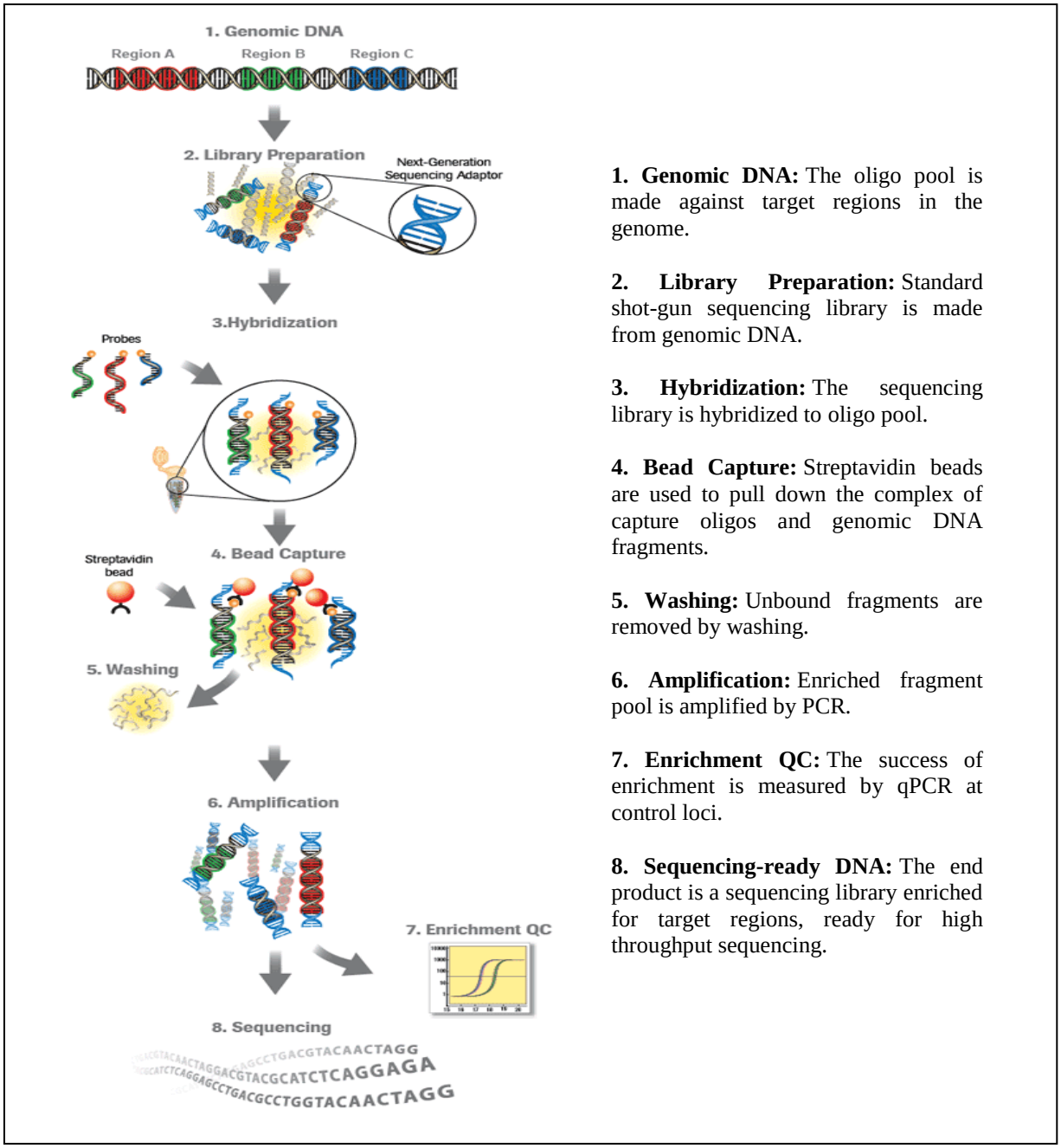


Figure 1. Protocol of the in-solution based method for target enrichment.*(Adapted from Roche NimbleGen brochure)*

Most imaging systems have not been designed to detect single fluorescent events, so amplified templates are required. But it has been noted a negative influence of PCR amplification on the uniformity of enrichment: performing 18 cycles of PCR amplification of libraries both before and after hybridization can introduce severe bias toward neutral G+C content in the resulting sequences. So it is desirable to keep PCR amplification to a minimum and only performing it after hybridization. However, an amplification-free library preparation tends to lack robustness [14].

Salt concentration is an important factor in determining the specificity and efficiency of hybridization. Therefore, it is convenient to use solid-phase reversible immobilization (SPRI) beads, to which nucleic acids can bind reversibly and capture DNA can be eluted in water [15].

One particular challenge for applying exome sequencing has been how best to define the set of target that constitute the exome. Considerable uncertainty remains regarding which sequences of the human genome are truly protein coding.

There are currently three major exome enrichment platforms: Agilent's SureSelect Human All Exon Kits, Roche/Nimblegen's SeqCap EZ Exome Library and Illumina's TruSeq Exome Enrichment. The technologies differ in their target choice, bait lengths and density, and molecule used for capture (DNA for Nimblegen and Illumina, RNA for Agilent). There are substantial differences in the density of oligonucleotide baits between the three platforms. Nimblegen contains overlapping baits that cover the bases it targets multiple times, making it the highest density platform of the three. Agilent baits reside immediately adjacent to one another across the target exon intervals. Illumina relies on paired-end reads to extend outside the bait sequences and fill the gaps. The exome enrichment platforms also have different target regions. Numerous databases of mRNA coding sequences exist, including RefSeq, coding and untranslated region (UTR) [16], UCSC KnownGenes [17] and Ensembl, total and coding sequence (CDC) [18]. They contain different numbers of noncoding RNA genes, and the start and end positions of some transcripts differ between them. Each commercial platform targets particular exomic segments based on combinations of the available databases. A large number of bases (29.45 Mb) are targeted by all three platforms. Nonetheless, each platform does not target specific regions: the majority of the Illumina-specific 27.73 Mb targets UTR regions; Nimblegen covers a greater portions of miRNAs; Agilent better covers Ensembl genes (Figure2).

Moreover, input genomic DNA ranges from 1 μg (Illumina) to 3 μg (Nimblegen and Agilent). The total procedure time before sequencing and the pre- and post-hybridization PCR cycles vary across platforms [19].

Nevertheless, all existing targets have limitations since current capture probes can only target exons that have been identified so far; efficiency of capture probes varies considerably and some sequences fail to be targeted by capture probe design altogether.

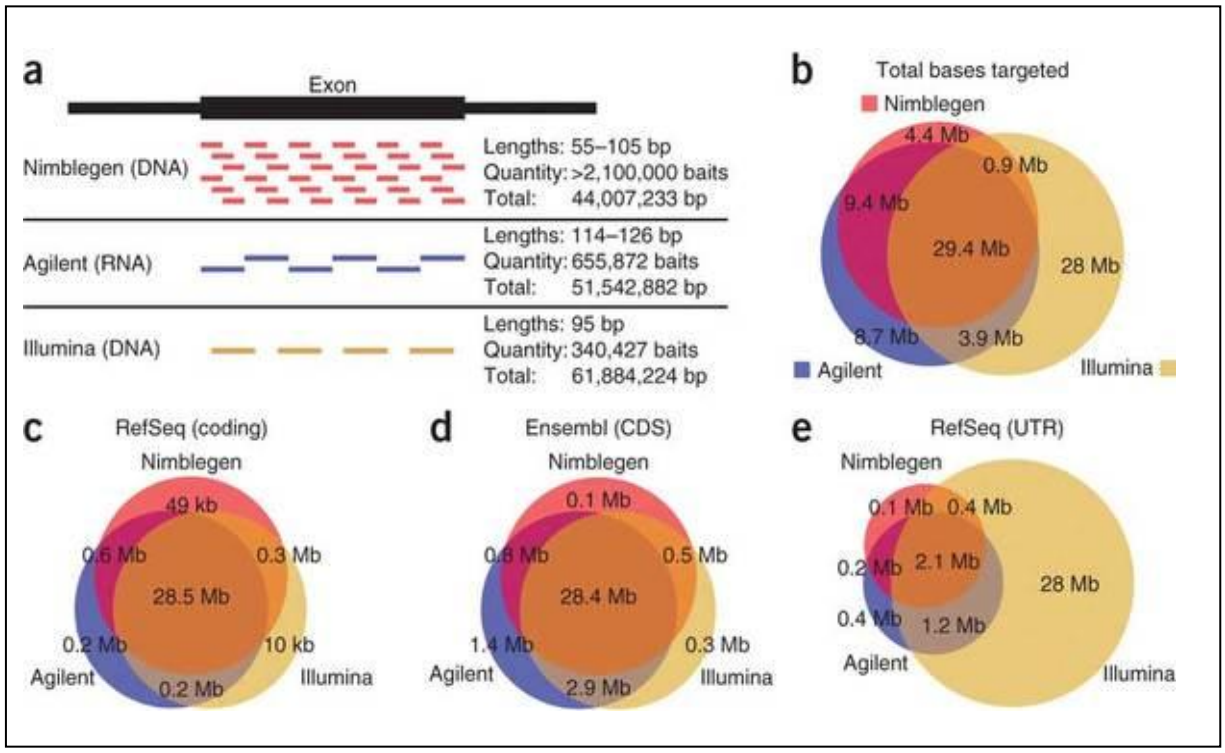


Figure 2. (**a**) Bait design details for each commercial platform. (**b**) Venn diagram showing the overlap of targeted genome regions for all three platforms. (**c, d e**) Venn diagram showing coverage of RefSeq coding exons, Ensembl CDS and RefSeq UTR exons respectively, and overlap between platforms. (*From Clark MJ, et al. Nat Biotechnol 29:908-914)*

2. Sequencing and imaging

DNA sequencing-by-synthesis (SBS) technology has been incorporated in several next-generation DNA sequencing systems with significant performance. A common SBS strategy is to use DNA polymerase or ligase enzymes to extend many DNA strands in parallel. Nucleotides or short oligonucleotides are provided either one at time or modified with identifying tags, so that the base type of the incorporated nucleotide or oligonucleotide can be determined as extension proceeds.

SBS strategies may be categorized as either single molecule-based (involving the sequencing of a single molecule) or ensemble based (involving the sequencing of multiple identical copies of a DNA molecule, amplified together on isolated surfaces or beads). There are fundamental differences in sequencing clonally amplified and single-molecule templates. Clonal amplification results in a population of identical templates, each of which has undergone the sequencing reaction. Upon imaging, the observed signal is a consensus of the nucleotides or probes added to the identical templates for a given cycle. A downside of ensemble-based SBS architectures is that sample preparation passes the analyte molecules through a single-molecule stage, only then to re-amplify them. The amplification from single molecules makes the process sensitive to amplification errors and products must be strictly isolated to reduce contamination of other libraries under construction. All SBS schemes involve the use of surface-bound components that provide a mean for parallel synthesis of DNA molecules and a structure for optimizing imaging and for flowing-in substrates and removing products. The challenges of surface chemistry include providing surfaces compatible with enzymatic processing of nucleotides along with the DNA, eliminating stray sticking of dye molecules and maximizing the density of SBS features over the surface. For this reason, structures are typically coated with a hydrophilic, functional surface layer, designed to tightly or covalently bind the molecules of interest – to minimized loss of molecules and thus signal – but to be inert to binding other materials during sequencing.

SBS is based on the stepwise enzymatic synthesis of DNA complementary to the template DNA to be sequenced. To obtain reads using ensemble-based SBS, repeated additions must be accommodated with virtually the total stepwise yields, so that chain extension is synchronous over all the molecules of the sample. This includes both the enzymatic addition and any subsequent chemical, enzymatic or photolytic, steps that may be needed to unblock

the substrate or remove the dye for the next addition. Several polymerase-based SBS schemes require using blocking groups that allow the addition of a single nucleotide at a time [20].

These technical challenges have been accommodated to varying degrees in commercially available systems, based on different sequencing and imaging strategies (Table 2). At the core of most next-generation sequencing methods is the use of dye-labeled modified nucleotides: ideally, these nucleotides are incorporated specifically, cleaved efficiently during or following fluorescent imaging, and extended as modified or natural bases in ensuing cycles [21, 22]

Cyclic reversible termination (CRT)

As the name implies, CRT uses reversible terminators in a cyclic methods that comprises nucleotide incorporation, fluorescence imaging and cleavage [23]. In the first step, a DNA polymerase, bound to the primed template, adds or incorporates just one fluorescently modified nucleotide, which represents the complement of the template base. The termination of the DNA synthesis after the addition of a single nucleotide is an important feature of the application. Following incorporation, the remaining unincorporated nucleotides are washed away. Imaging is then performed to determine the identity of the incorporated nucleotide and it is followed by a cleavage step, which removes the terminating/inhibiting group and the fluorescent dye. Additional washing is then performed before starting the next incorporation step. The key to the CRT method is the reversible terminator, of which there are two types: 3' blocked and 3' unblocked.

Currently, the Illumina sequencing platforms dominate the whole-exome sequencing market: it uses the clonally amplified template method, coupled with the four-color CRT strategy [24]. The four colors are detected by total internal reflection fluorescent imaging using two lasers. The slide is partitioned into eight channels, which allows independent samples to be run simultaneously (Figure 3).

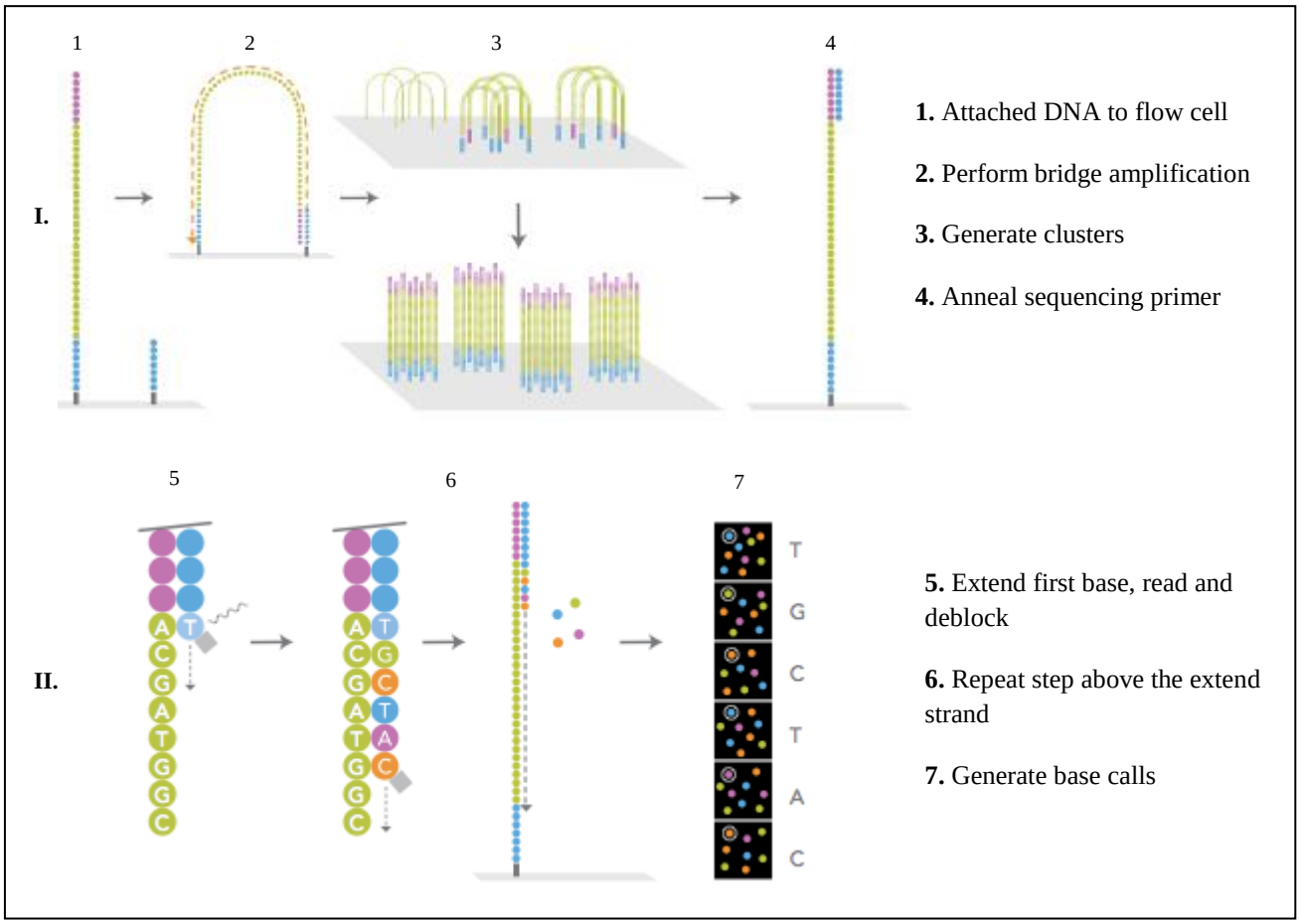


Figure 3. Outline of the Illumina sequencing platform. (**I**) Attached DNA fragments form "bridge" molecules which are amplified via an isothermal amplification process, leading to a cluster of identical fragments that are subsequently denatured for sequencing primer annealing. (**II**) Amplified DNA fragments are subjected to sequencing-by-synthesis using 3'unblocked labeled nucleotides. (*Adapted from the Illumina Genome Analyzer brochure).*

Sequencing by ligation (SBL)

In its simplest form, a fluorescently labeled probe hybridizes to its complementary sequence adjacent to the primed template. DNA ligase is then added to join the dye-labeled probe to the primer. Non-ligated probes are washed away, followed by fluorescence imaging to determine the identity of the ligated probe. The cycle can be repeated either by using cleavable probes to remove the fluorescent dye and regenerate a 5' $-PO_4$ group for subsequent ligation cycles or by removing and hybridizing a new primer to the template.

Applied Biosystems has commercialized its SBL platform called SOLiD (support oligonucleotide ligation detection), characterized by a colour-space imaging system that is a linear sequence of colour calls from the ligation round [25].

Pyrosequencing

Pyrosequencing is a non-electrophoretic, bioluminescence method that measures the release of inorganic pyrophosphate by proportionally converting it into visible light using a series of enzymatic reactions. Unlike other sequencing approach that use modified nucleotides to terminate DNA synthesis, the pyosequencing method manipulates DNA polymerase by the single addition of a dNTP in limiting amounts. The order and intensity of the light peaks are recorded as flowgrams, which reveal the underlying DNA sequencing.

The first platform using pyrosequencing was commercialized by Roche/454, which, unlike other instruments that produce shorter read lengths, does not require the run to be doubled for the sequencing of mate-pair templates. For homopolymeric repeats of up of six nucleotides, the number of DNTs added is directly proportional to the light of signal [26].

Differences in chemistries and raw data collection require individualized data processing pipelines and hinder combining output from different next-generation platforms. The Illumina sequencing instruments generate base-specific signal intensities, with basic algorithms needed to determine the most likely template-directed base being incorporated and the output is readily obtained as simple base sequence. In contrast, the Roche GS FLX adds only one type of nucleotide at a time, allowing multiple base incorporations across mononucleotide stretches in a single cycle, resulting in a signal proportional to the number of bases incorporated. The resulting flowgram can be readily converted to bases, but with some uncertainty surrounding the length of long mononucleotides repeats. SOLiD uses dibase encoding, whereby two adjacent template bases at a time are interrogated by the incoming labeled oligonucleotide destined for ligation. The sequencing output is encoded not as single bases but as numbers from 0 to 3 representing four possible dinucleotides [27].

	Illumina Genome Analyzer/HiSeq	Applied Biosystems SOLiD	Roche/454 Genome Sequencer FLX
NGS chemistry	Reversible terminator	Sequencing by ligation	Pyrosequencing
Total output data	20 Gb	20 Gb	0.5 Gb
Read length	Up to 150 bp	Up to 75 bp	Up to 700 bp
Base calling	Nucleotide space	Color space	Flow space
Error profile	Substitutions, underrepresentation of AT-rich and GC-rich regions	Substitutions, underrepresentation of AT-rich and GC-rich regions	Insertions and deletions, especially in long stretches (>6) of the same nucleotide

Table 2. Comparison of next-generation sequencing platforms used in exome sequencing**.**

3. Data analysis

Compared to Sanger sequencing, next-generation sequencing produces much more sequences, but of much shorter length and inferior quality; this has a tremendous impact on how the resulting readouts have to be processed in a downstream analysis. The short length of the DNA sequences imposes computational challenges for the detection of specific variations.

As mentioned above, an image-capturing device records the light signals generated by the synthesis or ligation processes at the newly generated strands. After acquisition of the image data, these recorded signals have to be converted into nucleotide bases. Furthermore, statistical models provides a measure of certainty of each base call in addition to the nucleotide itself. These statistical models base their error estimate on information such as signal intensities from the recorded image, the number of the sequencing cycle and distances to other sequence colonies. These certainties are usually expressed as Phred-like quality scores, which represent the decadic logarithm of the expected error probability of the base call:

$$Q_{phred} = -10 \log_{10} P(\text{error})$$

The preliminary step of base calling is usually automatically performed by the sequencing platform itself and it is specific since each sequencing platform has to solve challenges unique to the underlying methodology.

3.1 The SNP calling pipeline

In a common SNP calling pipeline the base calling is followed by an initial quality control of the generated reads, succeeded by the alignment of the reads to a reference sequence and a post-processing of the alignment. While these steps are shared by nearly all NGS applications, the remaining steps – quality score recalibration, SNP calling and filtering of SNP candidates - are more specific to the SNP calling pipeline (Figure 4).

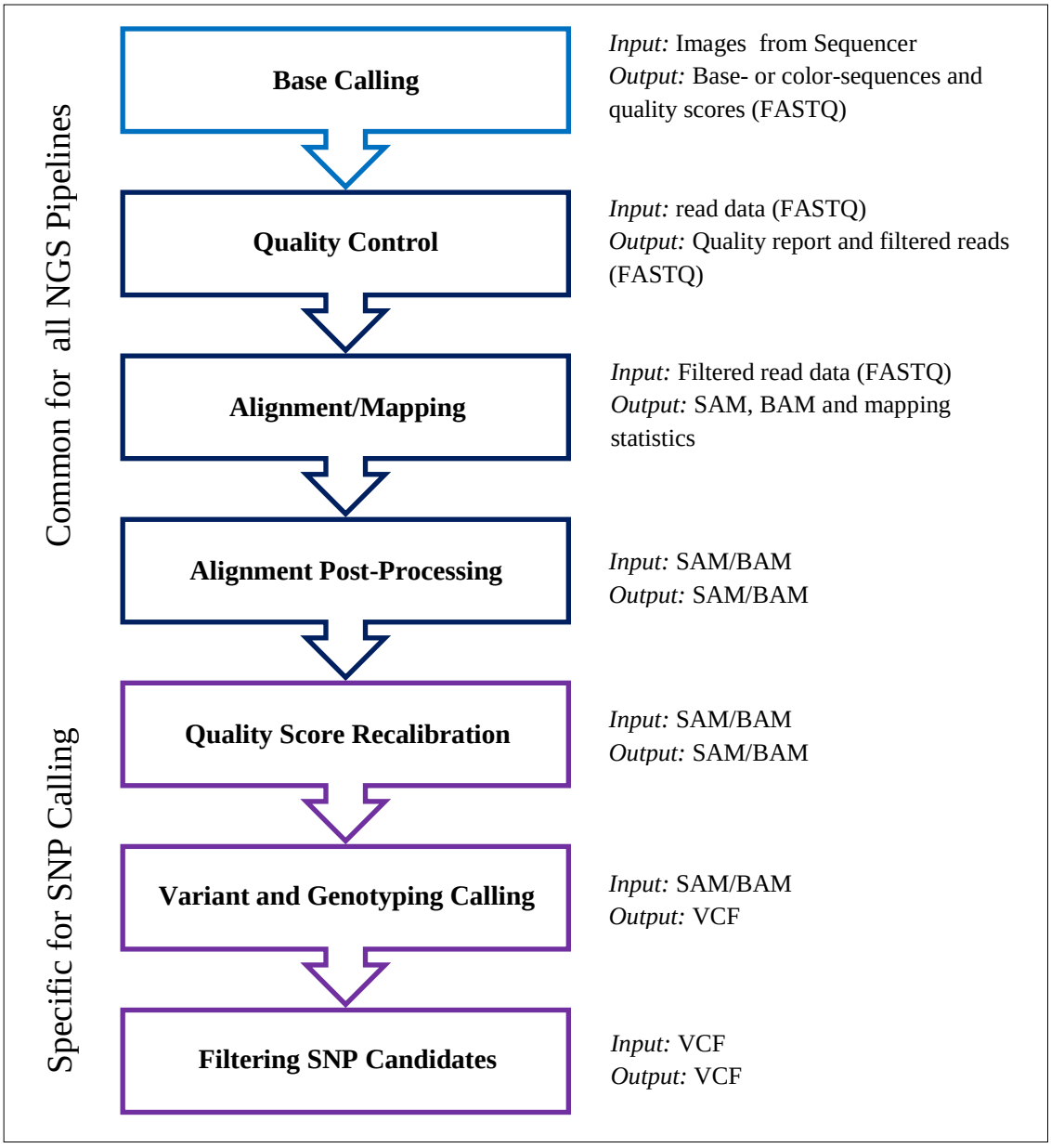


Figure 4. Workflow of the SNP calling pipeline.

Step 1: Quality control

Most platforms provide the read data directly in a flat file format such as FASTQ [28] or at least provide tools for conversion of the native output format into the quasi-standard FASTQ. The distribution of the quality scores at each position is one of the most interesting quality parameters for the overall quality of the run. Regarding the quality of the raw reads, there are noticeable difference between platforms. Illumina reads, for instance, undergo a quality control by the manufacturer's software. In case of the SOLiD platforms, no quality control is provided: it relies on the fact that reads of insufficient quality will not align to the reference sequence.

Step 2: Alignment/mapping

The next step in the processing pipeline for almost all applications involves alignment and assembly. There are two fundamental considerations when designing alignment and assembly algorithms for sequence analysis: first, the amount of data produced; second, the techniques produce data with different error profiles which must be addressed at the algorithmic level to obtain the maximum information from the data. A central challenge to the analysis of these data is sequence alignment whereby sequence reads must be compared to a reference. Alignment programs normally follow a multistep procedure to accurately map sequences. Using heuristic techniques in the first step, effort are made to quickly identify a small set of places in the reference sequence where the best mapping is most likely. Once the smaller subset of possible mapping locations has been identified, slower and more accurate alignment algorithms are run on this limited subset [29] (Figure 5).

Alignment is the process of mapping the reads to a reference sequence. Two approaches are commonly used: many sequence alignment software tools apply the lossless Burrows-Wheeler transform (BWT) for efficient data compression [30]; other algorithms rely on hashing to accelerate the alignment step. While the use of hashing allows quick access to the information on the location of subsequences in the reference sequence, the clear advantage of the BWT-based algorithms is the processing speed, as they are much faster at the same sensitivity level [31]. BWT algorithms typically create a suffix array from the BWT transformed sequence, rather than from the original sequence. First, the sequence order of the reference genome is modified using the BWT, a reversible process that reorders the genome

grouping together in the data structure the sequences that appear multiple times. Next, the final index is created and it is used for rapid read placement on the genome.

On the other hand, assembly starts from aligned reads to reconstructed the original DNA sequence computationally, which generates large, continuous regions of DNA sequence. Many alignment software provide tools to perform the assembly after the read alignment. *De novo* assembly is the suggested approach for reads mapping in regions prone to rearrangements, rapidly evolving, or where the reference genome might not be informative.

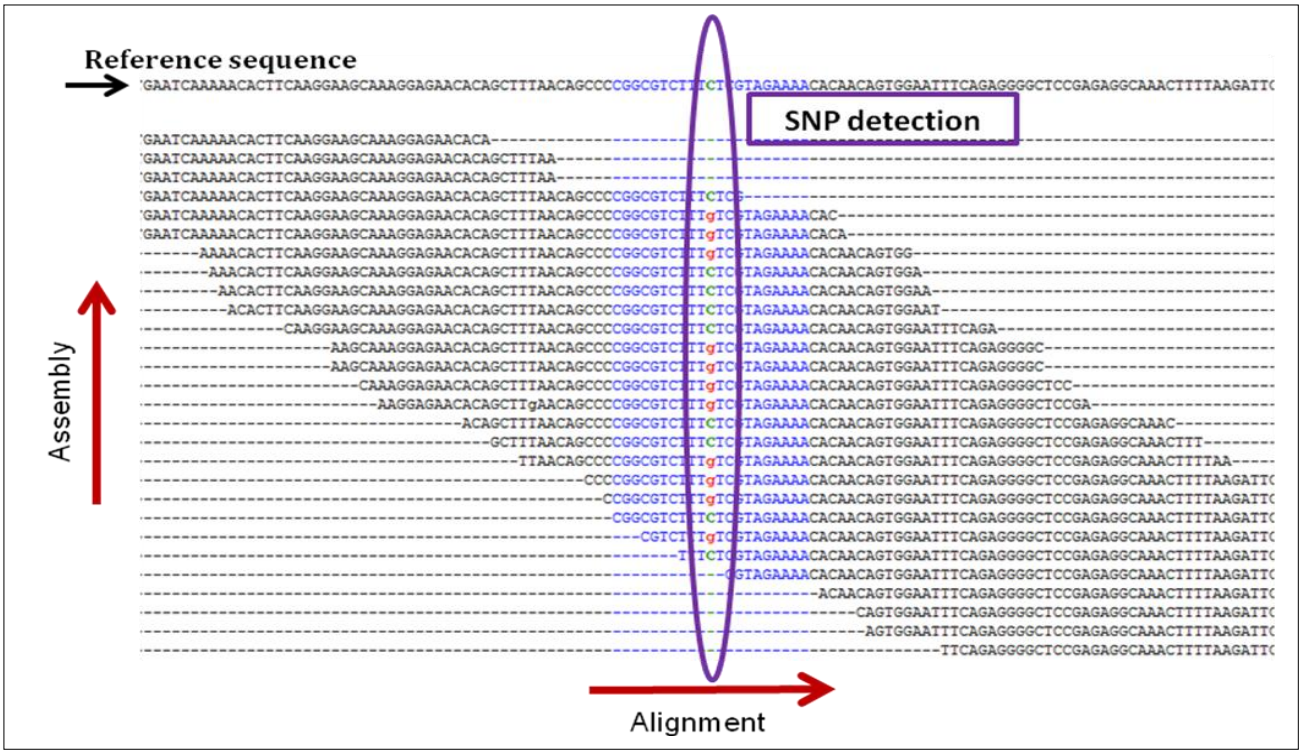


Figure 5. Two of the most fundamental computational issues in the context of sequencing analysis: alignment and assembly. Alignment is the process of determining the most likely source within the genome sequence for the observed DNA sequencing read. Assembly leads to the generation of large, continuous regions of DNA sequence. A resequencing application requires reads that can be accurately mapped in such a way that both nucleotide and structural variation can be reliably assessed.

In general, the choice of alignment tool and the corresponding settings significantly affect the outcome. This holds especially true for SNP calling, as wrongly aligned reads may result in artificial deviations from the reference. These deviations in turn may falsely be classified as SNPs in the downstream processing.

Once the reads have been aligned to the reference, many algorithms allow to store the results in the sequence alignment/map (SAM) format [32]. Briefly, the SAM format stores information about each aligned read, in particular, the position of the reference contig, the

orientation of the read, quality of the alignment and potential further alignment possibilities of the read. The SAM format, and its binary version the BAM format, are by now a quasi-standard for storing the result of the alignment step.
After the mapping step, it is advisable to check the alignment again by generating a mapping statistic, which is the computing fraction of reads that was successfully mapped to the reference, the fraction of reads that was rightly paired and the distribution of the insert size.

Step 3: Alignment post-processing

Prior to the actual variant calling, the algorithms require the alignments to be sorted with respect to their chromosomal position. Next, since the PCR used for amplifying the library and adding adapters may introduce artifacts, in form of reads or read pairs starting at exactly the same position and having the same length, it is common practice to remove or simply mark such PCR artifacts. The next post-processing step is the removal of all non-unique alignments, i.e. reads with more than one optimal alignment, since in these cases it cannot be determined from which site the read really originates. Then, it is common to realign reads around small indels, since difference in resolving small insertions and deletions may cause artificial SNPs in the downstream analysis.

Step 4: Quality score recalibration

The first software to provide recalibration of quality scores was SOAPsnp [33]: the approach exploits sites in the reference genome without any reported SNPs. On these sites SOAPsnp computes the empirical mismatch rate as an estimate for the true base quality. For a given machine provided quality score, sequencing cycle (in other terms, the position of the base in the read) and the substitution type, it calculates the average mismatch rate with respect to the reference, then used as the recalibrated quality score. Based on a similar concept, the GATK software [34] also provides a recalibration function: first, bases are grouped with respect to several features, as raw quality and dinucleotide content; second, for each category, the empirical mismatch rate is computed and used to correct the raw quality score.

Step 5: Variant and genotype calling

Recent SNP calling approaches integrate several sources of information within probabilistic framework. This procedure facilitates SNP calls in medium to low coverage regions and provide a way to quantifying uncertainty about the variant call. One major advantage of the statistical framework is the use of prior probabilities for a SNP at a given position. These prior probabilities can be derived from databases listing of confirmed SNPs or by carrying out SNP calling in multiple individuals at the same time [35].

Furthermore, the field of computational methods for discovering structural variation on NGS data is still an open bioinformatics challenge. The copy number variations (CNVs) discovery methods operate following a framework that allows detecting anomalous patterns, then calls the related variants using mainly four different approaches: read-pair methods, read-depth methods, split read approaches and *de novo* assembly [36].

Step 6: Filtering SNP candidates

Filtering is an essential step in reducing the number of false-positive SNP calls. Typically applied filters check for deviations from the Hardy-Weinberg equilibrium, minimum and maximum read depth, adjacency to indels and strand bias. While filtering might also remove real SNPs from the candidate list, it is an important tool for minimizing SNP calling artifacts.

Most SNP calling tools have the option to generate the data in the VCF format [37], which records for each identified SNP candidate basic information, such as the chromosomal position, the reference base, the identified alternative base or bases in case of triallelic SNPs. Furthermore, information on the quality of the SNP call as well as the amount of sequence data available for the call are stored.

Tools for automated variant annotation have been developed. Most of them offer a command line interface for annotating variants from different species. Basally, they rely on information available via the UCSC genome browser and offer in addition precomputed scores for predicting the likely functional consequences of non-synonymous amino acid exchanges and they allow to easily filter already known SNPs using information from public controls databases. The outcome of a SNP pipeline is not set in a stone. A recommended strategy is the use of different aligners and SNP callers for generating independent SNP candidates, since the most reliable are those appearing in more than one setting [38].

NOVEL DISEASE-GENE DISCOVERY

A key challenge of using exome sequencing to find novel causal genes for Mendelian diseases is how to identify disease-related alleles among the background of non-pathogenic polymorphism and sequencing errors.

The SNP calling process on the whole exome data generates on average 20,000 variants as compared with the genomic reference sequence, although the numbers have varied between different studies, which presumably reflects differences in the technologies and analysis strategies [39] . Thus, the post-processing and interpreting the generated huge amount of data are now the substantial challenges in the next-generation era.

More than 95% of these variants are already known as polymorphism in human populations. Strategies for finding causal alleles against this background vary, as they do for traditional approaches to gene discovery, depending on factors such as: the mode of inheritance of a trait; the pedigree or population structure; whether a phenotype arises owing to *de novo* or inherited variants; and the extend of locus heterogeneity for a trait. Such factors also influenced both the sample size needed to provide adequate power to detect trait-associated alleles and the selection of the most successful analytical framework.

1. Filtering for rare variants

The sequencing of only a modest number of affected individuals and then applying discrete filtering to the data to reduce the number of candidate genes is an important advantage that exome sequencing has over conventional approaches (Figure 6). In fact, this strategy alone can be exceptionally powerful for very rare Mendelian disorders. Rare diseases, by definition, have an individual incidence of less than 1/2000 in the population [40] and it is expected that mutations causing them will be at correspondingly rare frequencies, and most likely private to affected subjects. This is especially true for mutations that are highly penetrant, that are not expected to be found in the population at large and, hence, will not be seen in genome-wide scans for variants nor in polymorphism repositories.

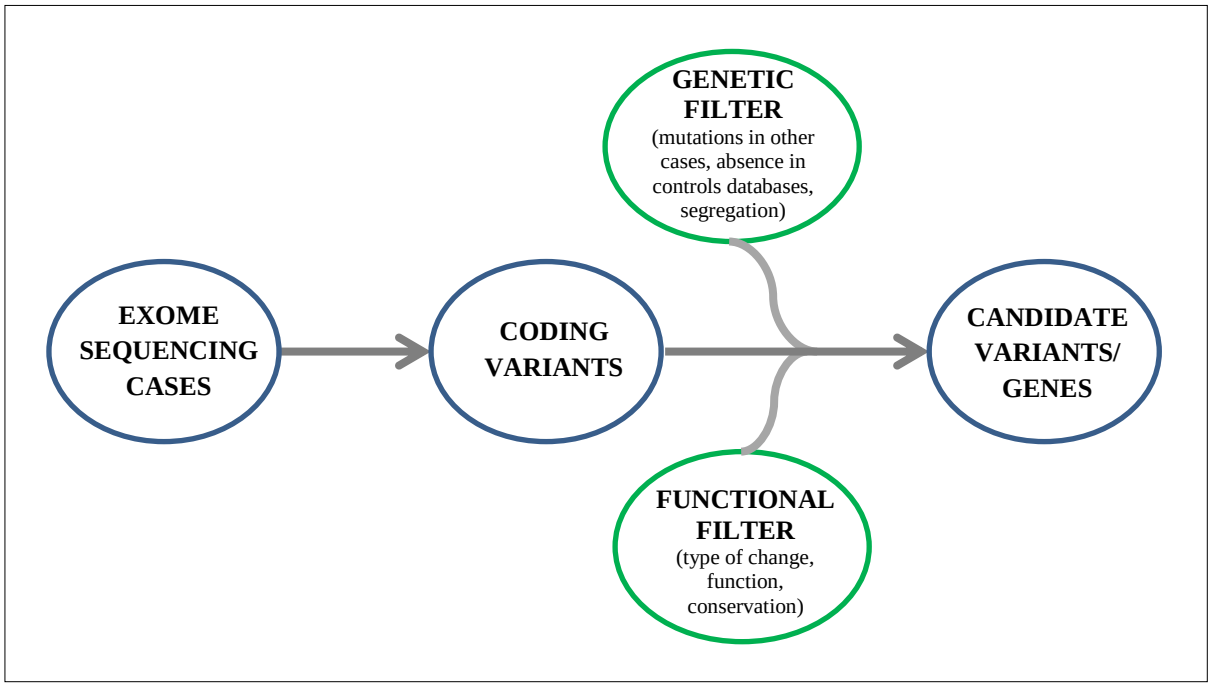


Figure 6. Filtering strategy for the analysis of Mendelian disorder through exome sequencing.

Therefore, novelty is assessed by filtering the variants against a set of polymorphisms that are available in public databases, such as dbSNP (http://www.ncbi.nlm.nih.gov/SNP) and 1000 Genome Project (http://www.1000genomes.org) and those found in a set of unaffected controls. This discrete filtering step is used to reduce the list of candidate genes by assuming that any allele found in the "filter set" cannot be causative. This approach is powerful because only a small fraction of the variants identified in an individual exome is novel.

Fundamental in this method is the assumption that the filter set contains no alleles from individuals with the phenotype being studied. This assumption can be problematic for two reasons. First, dbSNP is contaminated with an appreciable number of pathogenic alleles [41, 42]. Second, as the number of sequenced exomes and genomes increases, the filtering of observed alleles in a manner that is independent of their minor allele frequency (MAF) runs the risk of eliminating truly pathogenic alleles that are segregating in the general population at low frequency. This risk is especially relevant for recessive disorders, in which carrier status will not result in a phenotype that might otherwise exclude an individual from a control population. A lower MAF cutoff of 0.1% is helpful for dominant disorders, as the estimated

prevalence of the condition (generally well below 0.1%) provides an upper bound on the MAF.

2. Filtering based on function, effect and conservation

Candidate alleles can be further stratified on the basis of their predicted impact or deleteriousness, by giving greater weight to non-synonymous variants, frameshifts, stop codons and disruptions of canonical splice sites. However this is an oversimplification that is insensitive to causal alleles that do not directly alter protein-coding sequences. The main rationale given for this filter is that these kind of variants tend to be of larger effect than non-coding variants and also because it is difficult to predict the effects of non-coding and synonymous variants with any certainty. As such, in order to reduce noise when analyzing possible disease-causing variants, non-coding and synonymous variants are often ignored or greatly down-weighted. For some disorders, it is possible to filter variants even further, by focusing only on those that are loss-of-function (as nonsense and frameshift). Since there are only limited number of such mutations in any genome (<50), the candidate list is shortened very quickly [43].

Additionally, candidate alleles can be stratified by existing biological pathway or its interactions with genes or proteins that are known to cause a similar phenotype.

Variants can also be ranked by potential effect on protein structure and function, and conservation scores using quantitative estimates which exploit the observation that regions of genes and genomes in which mutations are deleterious tend to show high sequence conservation as a result of purifying selection. Sites that have experienced purifying selection can be identified by quantifying rates of mammalian evolution at the nucleotide level. Computational prediction of functional SNP effect often employs evolutionary conservation as well as physicochemicals properties of the affected amino acid in the protein. Implementations of this strategy include phyloP and Genetic Evolutionary Rate Profiling (GERP) [44] to predict the impact of potential causal variants that are either coding or non-coding. Approaches that stratify non-synonymous alleles – for example, Sorting Intolerant From Tolerant (SIFT) [45], Polymorphism Phenotyping v2 (PolyPhen2) [46] – also explore the predicted changes in proteins caused by specific amino acids substitutions. All of these strategies enrich for functional sites at which observed variants are more likely to affect phenotype, but have limited specificity and sensitivity [47]. Thus, these rankings are normally used in conjunction with other strategies and not as stand-alone filter.

3. Effect of mode of inheritance and pedigree information

The mode of inheritance of a monogenic disorder strongly influences both the experimental design (for example the number and selection of the most informative cases to be sequenced in multiplex families) and the analytical approach. Intuitively, discrete filtering should be more efficient for recessive disorders – since they require sequencing of fewer cases – than for dominant cases, because the genome of any given individual has around 50-fold fewer genes with two rather than one novel protein-altering alleles per gene [10]. This conclusion is supported by simulation studies and by the greater rate at which exome sequencing is identifying genes for recessive relative to dominant disorders.

For Mendelian phenotypes the use of pedigree information can substantially narrow the genomic search space for candidate causal alleles (Figure 7). Which individuals are the most informative ones to sequence depends on the frequency of a disease-causing allele and the nature of the relation-ship between individuals. For every rare alleles, the probability of identity-by-descent given identity-by-state is high even among distantly related individuals. In the absence of mapping data, sequencing the two most distantly related individuals with the phenotype of interest can substantially restrict the genome search space. When mapping data are available, the most efficient strategy is to sequence a pair of affected individuals whose overlapping haplotype produce the smallest genomic region. If the haplotype shared by all affected subjects is sufficiently short that the candidate interval is unlikely to include multiple candidate causal alleles, then a single individual may be sequenced. For consanguineous pedigrees in which a recessive mode of inheritance is suspected, sequencing just the one person with the smallest region (or regions) of homozygosity, as determined by the genome-wide genotyping data, should be sufficient.

Exome sequencing of parent-child trios is a highly effective approach for identifying *de novo* coding mutations, as a multiple *de novo* events occurring within a specific gene (or within a gene family or a pathway) is an extremely rare event [48]. This study design may be particularly applicable to gene discovery in disorders for which most cases are sporadic and when a dominant mode of inheritance is suspected or substantial locus heterogeneity is expected.

4. Filtering using tests of association

For identifying likely causal variants, an alternatively strategy to discrete filtering is to apply tests of association. The use of two-sample tests that compare cases (unrelated individuals with the same Mendelian phenotype) to a set of controls can either eliminate some of the problems of discrete filtering or provide estimates of the sample size needed for adequate power in the presence of complicating factors, such as genetic heterogeneity. As long as false positive are equally probable both in cases and in controls, the expected number of variants in any gene will be the same both in cases and in controls under any distribution of mutations. When genetic heterogeneity is known to be present, as indicated for example by the presence of complementing groups of mutations, or suspected, this information can be taken into account when performing power calculations to ensure that enough individuals are included in the study.

Furthermore, the growing number of well-documented exome data sets available will allow for the use of thousands of control chromosomes, which can increase the power to detect causal alleles, even when the number of available cases is limited.

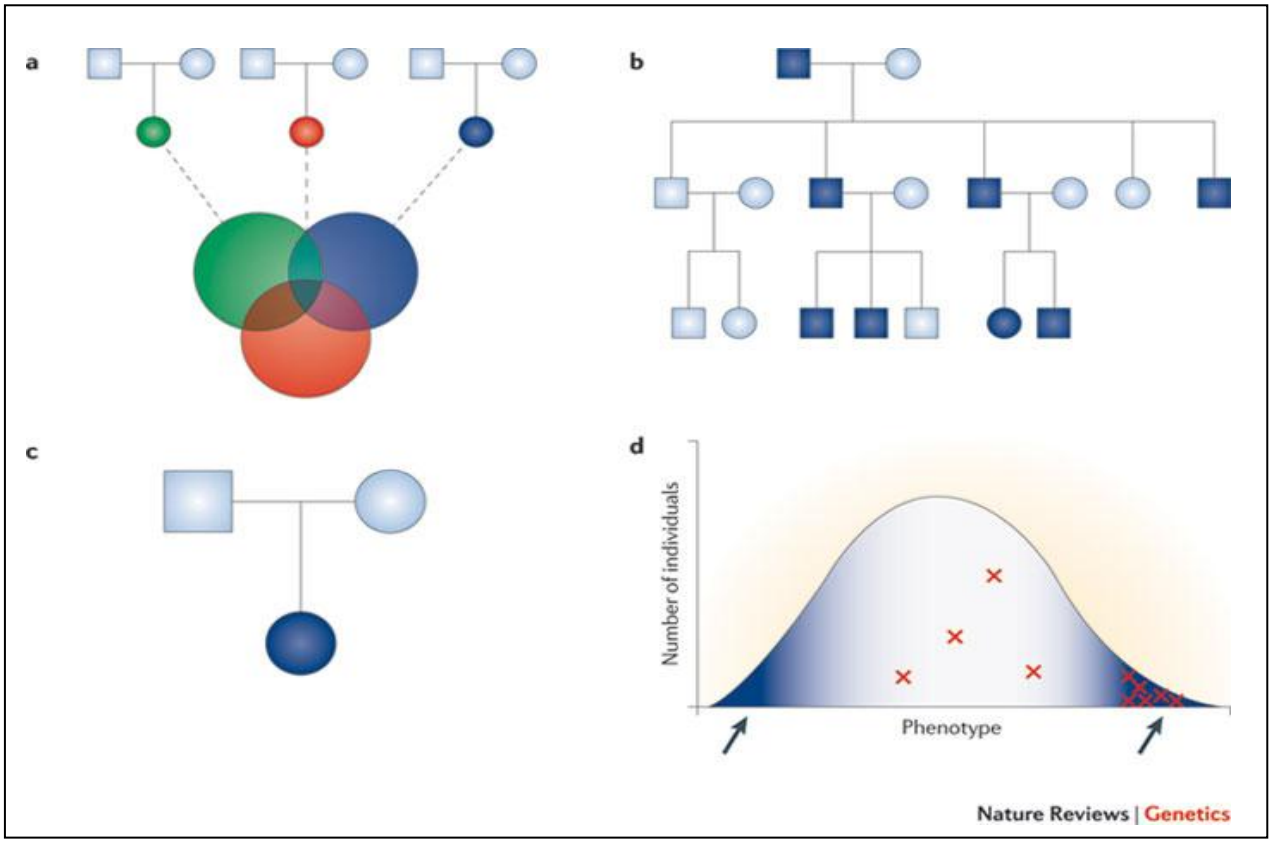


Figure 7. Strategies for finding disease-causing rare variants using exome sequencing. (**a**) Sequencing and filtering across multiple unrelated, affected individuals (**b**) Sequencing and filtering among multiple affected individuals in a pedigree. (**c**) Sequencing parent-child trios for identifying *de novo* mutations. (**d**) Sampling and

comparing the extremes of a distribution for a quantitative phenotype *(From Bamshad MJ et al. Nature Rev Genet 12:745-755).*

Considering that exome sequencing has been proved to be a powerful and cost-effective new tool for dissecting the genetic basis of diseases, I have applied different strategies to sift through variants in order to determine causal mutations and candidate genes in Mendelian disorders, such as six sporadic cases of Noonan syndrome, a family with Teebi syndrome, a consanguineous family showing agammaglobulinemia, and monogenic forms of complex diseases, as autosomal-dominant atrial septal defect or common variable immunodeficiency. My focus is to show some of the experimental and analytical options for employing massively parallel sequencing as a tool for disease-gene discovery and to highlight the key challenges in using this approach both within and without the paradigm of conventional approaches.
Moreover, I was involved in the development of a new consensus-calling and SNP-detection method from high-throughput sequencing data.

MATERIALS AND METHODS

SUBJECTS

I analyzed a total of six sporadic cases and four families affected by different Mendelian diseases (Table 3).

Disease	Model	# individuals genotyped	#individuals sequenced
Noonan syndrome	Autosomal dominant	6 parents-child trios	6 parents-child trios
Teebi syndrome	Autosomal dominant	6 affected, 3 unaffected and 1 obligated carrier	3 affected
Atrial Septal Defect	Autosomal dominant	-	4 affected
Common Variable Immunodeficiency	Autosomal dominant	9 affected and 8 unaffected	3 affected
Agammaglobulinemia	Autosomal recessive	2 affected siblings	1 affected

Table 3. Individuals included in genotyping and sequencing analysis.

For exome sequencing, I selected 6 individuals of European ancestry with Noonan syndrome and their unaffected parents (Figure 8). Even if Noonan syndrome features were present in all patients studied, each of them had unique and divergent manifestations collected in the clinical heterogeneity which characterizes the disorder.

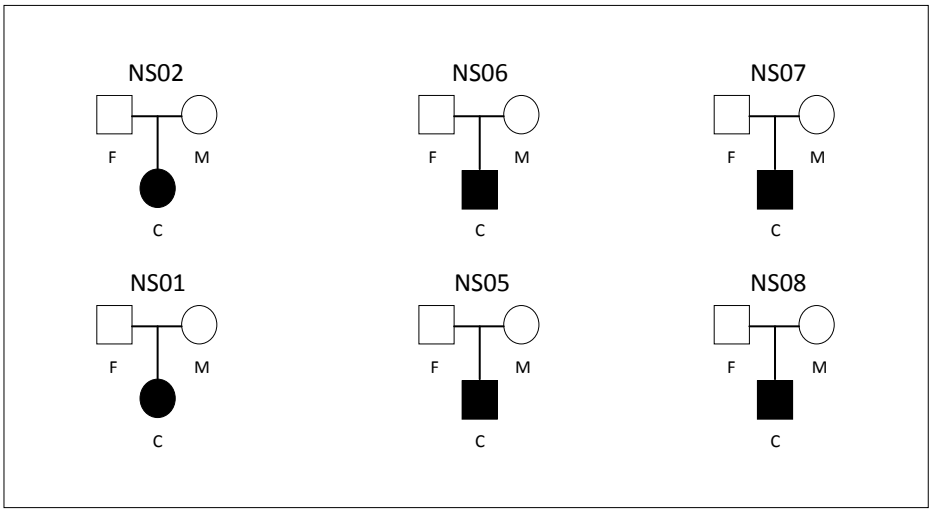


Figure 8 Parents-child trios with Noonan syndrome. F: father; M: mother; C: child.

The second set of analyses was of a moderate-sized multi-generational kindred with several cases of Teebi syndrome (Figure 9), with a substantial phenotypic heterogeneity: all affecteds had ptosis, but they showed various grades of other facial dysmorphism and cardiac defects. The individual II-1 seemed to lack strong symptoms, but her son and the grand-daughter were affected, having probably inherited the causative mutation from her. Thus, I strongly suspected incomplete penetrance of the causative allele in this family. All 10 subjects whose DNAs were available, both affected and unaffected, were included in the linkage study (II.1, III.1, III.2, III.3, III.5, III.7, III.8, IV.1, IV.9, IV.10), while 3 affected individuals (III.5, IV.1 and IV.10) were sequenced by the exome approach.

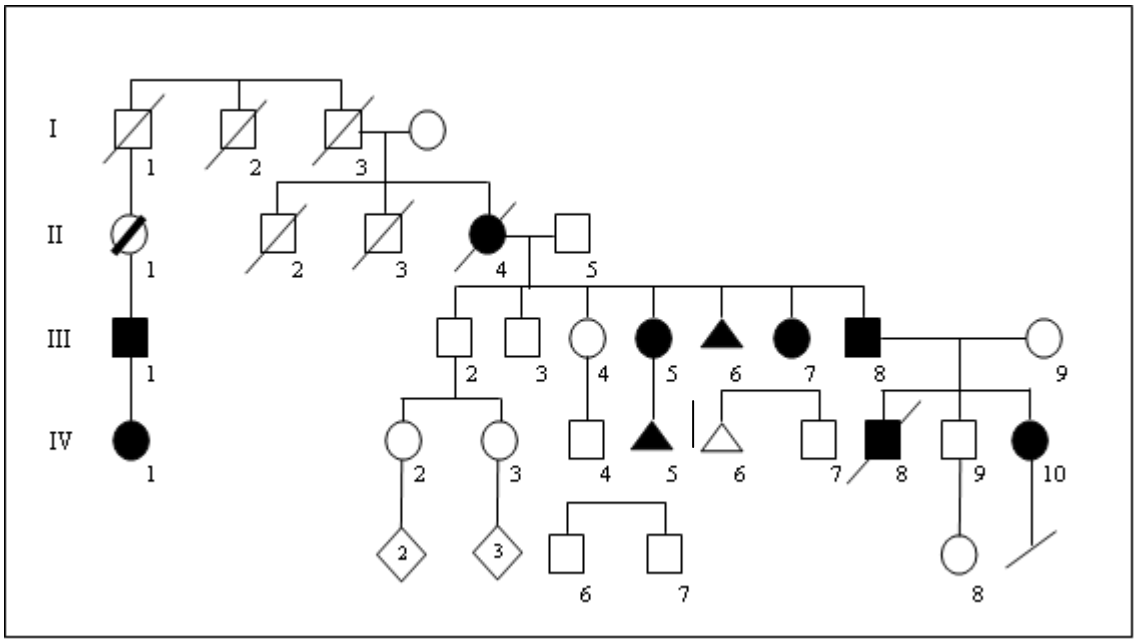


Figure 9. Pedigree of the studied family with Teebi syndrome.

Then, I studied an American family that was strongly suggestive for dominant transmission of atrial septal defect (ASD) (Figure 10). It was a five-generation family comprising 11 affected and 23 unaffected individuals including 2 obligated carriers (III.1 and IV.6). Phenotypic data were collected from review of medical records, phone interviews and photographs. Genome-wide linkage analysis was previously performed using microsatellite markers in all available subjects. This approach allowed the identification of a single candidate genomic interval on chromosome 15q23-q24.3 (Data not reported). The chosen strategy in this family was to

sequenced the whole exome of 4 affected individuals in the family (IV.20, V.4, V.10 and V.12) to give priority to the variants within the linkage interval.

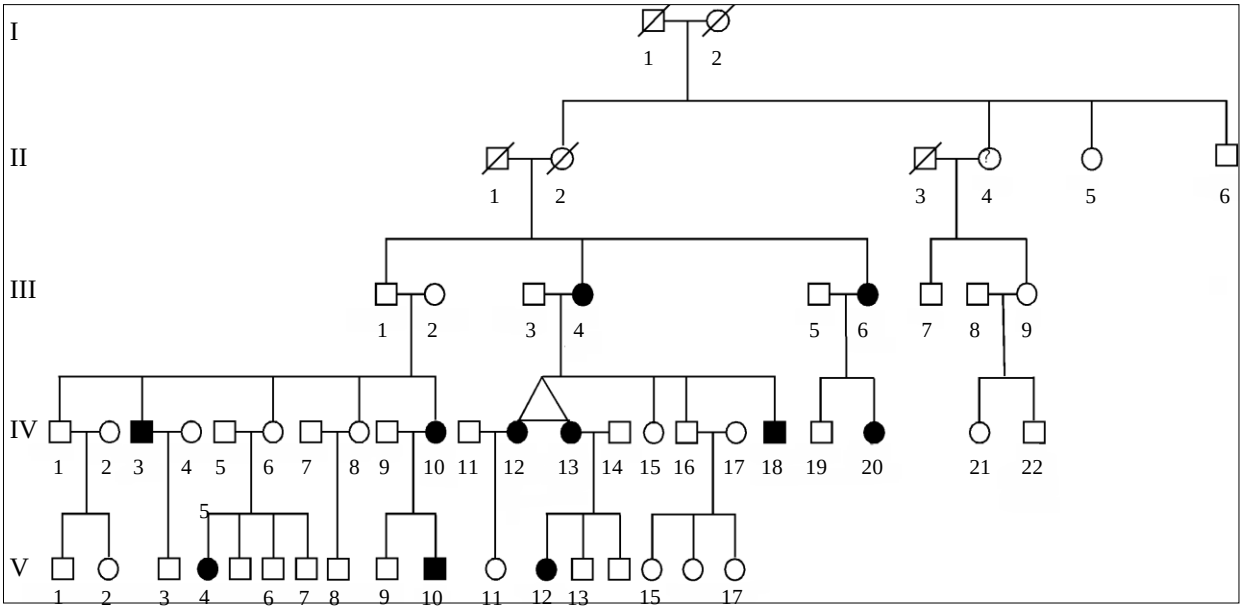


Figure 10. Pedigree of the family affected by ASD. The pedigree structure suggests the incomplete penetrance of the disease.

In order to identify the gene responsible for common variable immunodeficiency (CVID) in a large family with autosomal dominant inheritance of the disease (Figure 11), the DNA extracted from 8 unaffected and 9 affected members of the family was used to perform genome-wide genotyping (II.2, III.2, III.7, III.10, IV.2, IV.10, V.1, V.3, III.4, III.6, III.11, IV.1, IV.6, IV.7, IV.8, IV.11, V.2), while the exome data were obtained and analyzed in 3 affected subjects (III.7, IV.2 and IV.10).

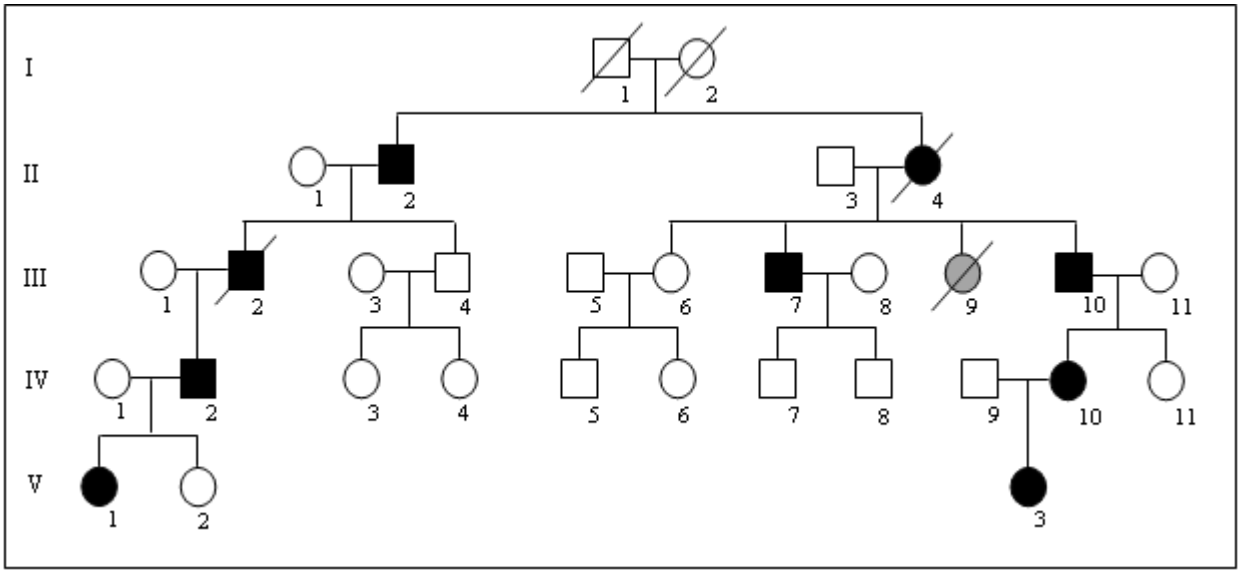


Figure 11. Pedigree of a large Italian family with CVID.

Finally, my study included a consanguineous Italian family, composed by two affected siblings from a first cousins marriage affected by an autosomal-recessive form of agammaglobulinemia (Figure 12). Of the two affected sisters, both genotyped, only one was evaluated by exome sequencing.

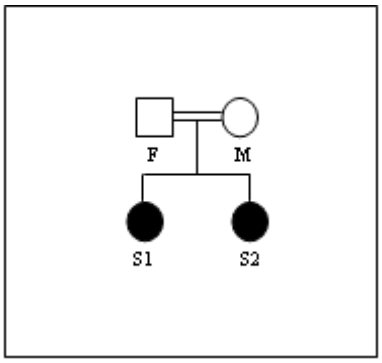


Figure 12. The consanguineous family with an autosomal-recessive form of agammaglobulinemia.

Genomic DNA was extracted from peripheral blood lymphocytes of each participating individual using Gentra Systems Puregene DNA purification kit (Qiagen), obtaining samples of high quality with an OD 260/280 ratio ranging from 1.8 to 2.0. The concentration and the quality parameters of extracted DNAs were measured by a spectrophotometer (NanoDrop ND-1000, NanoDrop Technologies).

WHOLE-EXOME SEQUENCING AND BIOINFORMATICS

1. Targeted capture and massively parallel sequencing

Genomic DNA samples were processed by in-solution hybridization using initially the NimbleGen SeqCap EZ Exome Library v2.0 method and, as the designed improve to capture additional exons and previously unannotaded genes, the v3.0 was substituted. The Illumina GAII and HISeq 2000 platforms were used to obtained paired end 76 bp and 101 bp sequencing reads as technology progressed.

In order to achieve informative exome sequencing results, DNA samples followed a multistep library preparation procedure:

- *Fragmentation*: 3 µg of DNA from each designed participating individual was firstly sheared by sonication (Covaris, MA, USA) at fixed conditions in order to obtain the majority of fragments of 200-300 bp in length:

  - duty cycle: 10%
  - intensity: 10
  - cycles per burst: 200
  - time: 250 s
  - temperature: 25°C

- *End-repair and A-tailing*: the fragmentation step produced double-stranded DNA with a mixture of blunt ends, recessed 3' and 5'ends, with and without phosphate moiety. To uniform the genomic fragments, they were treated with the NEBnext End Repair Enzyme mix to generate 5'-phosphorilated blunt ends, and Klenow Fragment (3'→5' exo$^-$), an N-terminal truncation of DNA Polymerase I which retains polymerase activity but lacks 5'→3' exonuclease activity, with dA-Tailing buffer to add a dAMP to the 3' end at a blunt DNA fragment. Addition of a single A nucleotide to the 3' ends of fragments deterred concatemerization of templates and increased the efficiency of the next step

- *Adapter ligation*: adapter oligonucleotides – linker 1 and 2 – were annealed and ligated to the fragment ends using a Quick T4 DNA Ligase (NEBNext^{TM} DNA sample prep master mix set 1, New England BioLabs Inc.). Adapter ligation to fragments must be as efficient as possible, but at the same time, ligation of adapter to one another must be suppressed: adapter dimers will also generate clusters that can be sequenced and will reduce the total proportion of desired sequence obtained from a run

- *Size selection*: in order to give a DNA library with a particular insert size range and allow the removing of majority of adapter dimers, the adaptor-ligated templates were fractionated by agarose gel electrophoresis and fragments of the desired size were excised and gel purified on Qiagen purification columns
- *PCR*: extracted DNA was amplified using the primers complementary to the previously ligated linkers by ligation-mediated (LM) PCR for 12 cycles to selectively enrich for properly ligated template strands, to generate enough DNA for accurate quantification and to add oligonucleotide sequences to the template strands that allow hybridization to the flow cell surface. The amplified product was then purified using the QIAquick PCR Purification kit (Qiagen) according to the manufacturer's recommendations
- *Library hybridization, hybrid capture selection and amplification*: each library was hybridized for target enrichment, followed by washing, elution from magnetic beads and additional LM PCR
- *Quality assessment*: the Agilent 2100 Bioanalyzer was used to assess quality, quantity and size range of the enriched libraries, that were diluted to a working concentration of 10 nM
- *Cluster amplification*: paired-end cluster generation on the flow cells was performed on the Cluster Station (Illumina Inc. CA USA). During this step, single DNA fragments with Illumina supplied adapter sequences ligated at both ends (the template) were attached to the surface of the oligonucleotidide-coated flow cell and amplified to form a surface-bound colony (the cluster). The result was a heterogeneous population of clusters, with each cluster consisting of many identical copies of the original template molecule
- *Sequencing-by-synthesis*: hybridized flow cells containing the captured, purified and clonally amplified libraries targeting the exome were, finally, transferred from the Cluster Station to the sequencing platform for 76 or 101 cycles of nucleotide incorporation, imaging and cleavage.

2. Next-generation sequencing data analysis

Bioinformatic analysis of Next-generation sequencing data include all the processes needed to extract biological information from the sequenced reads.

For each of the sequenced sample, I performed the following steps (Table 4):

1. Reads quality check. Reads generated with the sequencing technology were checked using the FastQC program (http://www.bioinformatics.babraham.ac.uk/publications.html)
2. Alignment. Raw reads were aligned with BWA [49] against the reference genome hg19, and a bam file containing the aligned reads was generated with SAMtools [50]
3. Local realignment around indels. Aligned reads from every bam file were realigned locally with the GATK package (www.broadinstitute.org) [34], in order to transform regions with misalignments into clean reads containing a consensus indel suitable for variant discovery
4. Duplicate reads removal. PCR duplicate reads were removed with the Picardtools MarkDuplicates utility (picartools.sourceforge.net)
5. Quality score recalibration. Base quality scores of aligned reads from every bam file were recalibrated with the GATK package, in order to obtain more accurate base quality scores
6. Alignment and Coverage metrics. Metrics of every sample's alignment were collected from bam files with SAMtools, in order to check whether the alignment process gave a consistent output. Metrics of the coverage obtained on the targeted exome were calculated using the GATK package, in order to obtain all the available information about the reliability of the variant calling results
7. Variant calling. Variant positions with respect to the reference sequence hg19 were called in the targeted exome in order to obtain a vcf file with all the sample's variants
8. Variant filtering. A filter tag was assigned to every variant in the vcf file in order to discriminate between good and bad quality variants
9. Variant annotation. All the variants have been annotated according to NCBI RefGene (www.ncbi.nlm.nih.gov) and UCSC KnownGene (genome.ucsc.edu) databases.

Step	Analysis	Tool	Output
1	Reads quality control	FastQC	Assessment of the quality of the reads
2	Alignment to hg19	BWA	Reads aligned to the genome
3	Local realignment	GATK	Artifacts due to indel misalignment solved
4	Duplicates marking	Picardtools	PCR duplicated reads removed
5	Base quality recalibration	GATK	Biases in base quality solved
6	Final aligment statistics	GATK	Assessment of the quality of the alignment
7	Variant calling	GATK	Variants discovered
8	Variant filtering	GATK	Filtered variants based on quality criteria
9	Variant annotation	Annovar	Variants annotated against RefSeq
10	Gene prioritization	Manual	A list of most likely candidate genes

Table 4. Summary of the data analysis pipeline with the tools used and a short description of analyses performed and output obtained.

2.1 Step 1: Reads quality check with FastQC

Before analyzing NGS reads, some simple quality control checks of the raw data are needed. Most sequencers generate a QC report as part of their analysis pipeline, but this is usually only focused on identifying problems which were generated by the sequencer itself. FastQC provides a QC report which can spot problems originated either in the sequencer or in the starting library material.

Reads generated by NGS sequencers are usually stored in FastQ format, which contains information about sequence content and quality. Paired-end reads are stored in two different files, the one containing all the '+' strand reads, the other containing all the '-' strand reads. I checked the quality of '+' and '-' strand reads contained in the FastQ files with FastQC.

The analysis in FastQC is performed by a series of analysis modules. For the most of these modules, FastQC reports a warning or a failure message if some problem is encountered across the reads.

Basic Statistics

The Basic Statistics module generates some simple composition statistics for the file analyzed, such as the filename, the filetype (base-space or color-space), the encoding (which ASCII encoding of quality values is used) the count of the total/filtered sequences (in Casava

mode, sequences flagged to be filtered are removed from all analyses), the sequence length and the %GC.

Per Base Sequence Quality

This module shows an overview of the range of quality values across all bases at each position in the FastQ file.

A warning is issued if the lower quartile for any base is less than 10, or if the median for any base is less than 25. It raises a failure if the lower quartile for any base is less than 5 or if the median for any base is less than 20.

Per Sequence Quality Scores

The per sequence quality score report allows to see if a subset of sequences have universally low quality values. If a significant proportion of the sequences in a run have overall low quality, this could indicate some kind of systematic problem - possibly with just part of the run (for example one end of a flow cell). A warning is raised if the most frequently observed mean quality is below 27 - this equates to a 0.2% error rate. An error is raised if the most frequently observed mean quality is below 20 - this equates to a 1% error rate.

Per Base Sequence Content

Per Base Sequence Content plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called.

In a random library it is expected that there would be little to no difference between the different bases of a sequence run, so this module issues a warning if the difference between A and T, or G and C is greater than 10% in any position. This module fails if the difference between A and T, or G and C is greater than 20% in any position.

Per Base GC Content

Per Base GC Content plots out the GC content of each base position in a file. Since in a random library it is expected no difference between the bases of a sequence run, the overall GC content should reflect the GC content of the underlying genome.

This module issues a warning it the GC content of any base strays more than 5% from the mean GC content; it fails if the GC content of any base strays more than 10% from the mean GC content.

Per Sequence GC Content

This module measures the GC content across the whole length of each sequence in a file and compares it to a modeled normal distribution of GC content. An unusually shaped distribution could indicate a contaminated library or some other kinds of biased subset. A normal distribution which is shifted indicates some systematic bias which is independent of base position. If there is a systematic bias which creates a shifted normal distribution then this won't be flagged as an error by the module since it doesn't know what the genome GC content should be. A warning is raised if the sum of the deviations from the normal distribution represents more than 15% of the reads, while it indicates a failure if the sum of the deviations from the normal distribution represents more than 30% of the reads.

Per Base N Content

If a sequencer is unable to make a base call with sufficient confidence then it will normally substitute an N rather than a conventional base call. This module plots out the percentage of base calls at each position for which an N was called. This module raises a warning if any position shows an N content of >5% or raises an error if any position shows an N content of >20%.

Sequence Length Distribution

Some high throughput sequencers generate sequence fragments of uniform length, but others can contain reads of wildly varying lengths. In my case, I did not used this module since a fixed read length is used by Illumina platform.

Duplicate Sequences

In a diverse library most sequences occur only once in the final set. A low level of duplication may indicate a very high level of coverage of the target sequence, but a high level of duplication is more likely to indicate some kind of enrichment bias (e.g. PCR over amplification).

This module counts the degree of duplication for every sequence in the set and creates a plot showing the relative number of sequences with different degrees of duplication. There is a warning if non-unique sequences make up more than 20% of the total and an error if non-unique sequences make up more than 50% of the total.

Overrepresented Sequences

A normal high-throughput library contains a diverse set of sequences, with no individual sequence making up a tiny fraction of the whole. Finding that a single sequence is very overrepresented in the set either means that it is highly biologically significant, or indicates that the library is contaminated, or not as diverse as expected.

This module lists all of the sequence which make up more than 0.1% of the total and it issues an error if any sequence is found to represent more than 1% of the total.

Overrepresented Kmers

The analysis of overrepresented sequences will spot an increase in any exactly duplicated sequences, but it does not work with very long sequences with poor sequence quality, where random sequencing errors dramatically reduces the counts for exactly duplicated sequences, and with a partial sequence which is appearing at a variety of places because it has not be seen either by the per base content plot or the duplicate sequence analysis.

This module counts the enrichment of every 5-mer within the sequence library. It calculates an expected level at which this k-mer should have been seen based on the base content of the library as a whole and then uses the actual count to calculate an observed/expected ratio for that k-mer. In addition to reporting a list of hits, it draws a graph for the top 6 hits to show the pattern of enrichment of that Kmer across the length of the reads. Any k-mer showing more than a 3 fold overall enrichment or a 5 fold enrichment at any given base position will be reported by this module. To allow this module to run in a reasonable time only 20% of the whole library is analyzed and the results are extrapolated to the rest of the library. There is an error issued if any k-mer is enriched more than 10 fold at any individual base position.

2.2 Step 2: Alignment to the reference genome hg19 with BWA and visualization of the alignments using IGV

Burrows-Wheeler Aligner (BWA, http://bio-bwa.sourceforge.net/) is a fast light-weighted tool that aligns relatively short sequences (queries) to a sequence database (target), such as the human reference genome. It implements an algorithm designed for short queries up to ~200bp with low error rate (<3%). It does gapped global alignment, supports paired-end reads, and is one of the fastest short read alignment algorithms to date while also visiting suboptimal hits. BWA allows alignment of paired-end reads in two steps:

1. The paired reads are aligned to the reference genome independently (their coordinates are mapped to the reference genome) with the command:

```
bwa aln [-n maxDiff] [-o maxGapO] [-e maxGapE] [-d nDelTail] [-i nIndelEnd] [-k maxSeedDiff] [-l seedLen] [-t nThrds] [-cRN] [-M misMsc] [-O gapOsc] [-E gapEsc] [-q trimQual] <in.db.fasta> <in.query.fq> > <out.sai>
```

The 'aln' command is used for each of the two fastq files ('+' and '-' strands). The options of this first command (in square parentheses) can be used to control the way BWA map the single-end reads to the reference genome. Of all these possible options, I changed only the '-q' one, trimQual. This option is useful for trimming reads which have poor quality. Trimming allows to keep the “good” part of a read (usually the first part) while discarding the “bad one” which is rich of poor quality bases. To do this, trimQual calculates the maximum argument of the Phred-like quality score function along every single read and discards all the bases after the one at which the function reaches its maximum. The trimming value was set to 20.
The 'aln' command requires two input files: <in.db.fasta> and <in.query.fq>. The first is the reference genome sequence in fasta format, the second is the fastq file. The output of this command is in the binary 'sai' format, designed for BWA use only.

2. Alignments in the SAM format given paired-end reads are generated with the command:

```
bwa sampe [-a maxInsSize] [-o maxOcc] [-n maxHitPaired] [-N maxHitDis] [-P] <in.db.fasta> <in1.sai> <in2.sai> <in1.fq> <in2.fq> > <out.sam>
```

The 'sampe' command requires as input files the two 'fastq' files, the two 'sai' files and again the reference genome in fasta format. BWA outputs the final alignment in the SAM (Sequence Alignment/Map) format.

SAM file can be finally compressed into binary BAM format by SAMtools (http://samtools.sourceforge.net/). The general command is the following:

```
samtools view -bt ref_list.txt -o aln.bam aln.sam.gz
```

The BAM file is the final product of the alignment process. It contains all the data needed to localize a pair of reads on the reference genome and provides detailed information about the accuracy of the paired-end read mapping.

Alignments stored in the BAM file can be visualized using the Integrative Genomics Viewer (IGV) tool (www.broadinstitute.org/igv/). The IGV is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations. Visualization of the aligned reads can give a general overview of the alignment outcome. Biases such as poor quality of reads or high percentage of duplicate reads are easily highlighted by direct visualization. Also, visualization can be useful when looking at the variants, since it gives an idea of how much a variant is reliable of it is affected by classical biases that make it a likely false-positive signal.

2.3 Steps 3, 4, 5: Local indel realignment, duplicate removal and base quality score recalibration with GATK

Once a BAM file is generated, three key processes are used:

- *Local realignment around indels.* Reads that align on the edges of indels often get mapped with mismatching bases that might look like evidence for SNPs. I used GATK to look for the most consistent placement of the reads with respect to the indel in order to clean up these artifacts. The local realignment tool is designed to consume one or more BAM files and to locally realign reads such that the number of mismatching bases is minimized across all the reads. In general, a large percent of regions requiring local realignment are due to the presence of an insertion or deletion (indels) in the individual's genome with respect to the reference genome. Such alignment artifacts result in many bases mismatching the reference near the misalignment, which are easily mistaken as SNPs. Moreover, since read mapping algorithms operate on each read independently, it is impossible to place reads on the reference genome such that mismatches are minimized across all reads. Consequently, even when some reads are correctly mapped with indels, reads covering the indel near just the start or end of the read are often incorrectly mapped with respect the true indel, also requiring realignment. Local realignment serves to transform regions with misalignments due to indels into clean reads containing a consensus indel suitable for standard variant discovery approaches. Following local

realignment, the GATK tool Unified Genotyper can be used to sensitively and specifically identify indels.

There are two steps to the realignment process:

1. Determining (small) suspicious intervals which are likely in need of realignment (RealignerTargetCreator tool)

2. Running the realigner over those intervals (IndelRealigner). The following general command is used to run the indel realigner:

```
java -Xmx4g -jar GenomeAnalysisTK.jar
   -I input.bam
   -R ref.fasta
   -T IndelRealigner
   -targetIntervals intervalListFromRTC.intervals
   -o realignedBam.bam
   [-known /path/to/indels.vcf]
```

- *Mark Duplicates.* Duplicately sequenced molecules shouldn't be counted as additional evidence for or against a putative variant. By marking these reads as duplicates the algorithms in the GATK know to ignore them. GATK exploits the MarkDuplicates utility of Picard tools (http://picard.sourceforge.net/command-line-overview.shtml#MarkDuplicates) with the command:

```
java -jar MarkDuplicates.jar I= input.bam O= output.bam
METRICS_FILE= output.metrics.markdup.stats
REMOVE_DUPLICATES=true CREATE_INDEX=true VALIDATION_STRINGENCY=
LENIENT
```

- *Base quality score recalibration.* The per-base estimate of error known as the base quality score is the foundation upon which all statistically calling algorithms are based. It has been found that the estimates provided by the sequencing machines are often inaccurate, and worse, biased. Through recalibration an empirically accurate error model is assigned to the bases to create an analysis-ready bam file. After recalibration, the quality scores in the QUAL field in each read in the output BAM are more accurate in that the reported quality score is closer to its actual probability of mismatching the reference genome. Moreover, the recalibration tool attempts to correct for variation in quality with machine cycle and sequence context. This process is accomplished by analyzing the covariation among several features of a base, for example, the reported quality score, the

position within the read, the preceding and current nucleotide (sequencing chemistry effect) observed by the sequencing machine. These covariates are then subsequently applied through a piecewise tabular correction to recalibrate the quality scores of all reads in a BAM file.

The general command line used for this end is:

```
java -jar GenomeAnalysisTK.jar \
-T PrintReads \
-R reference.fasta \
-I input.bam \
-BQSR recalibration_report.grp \
-o output.bam
```

2.4 Step 6: Alignment and coverage metrics

With the GATK Depth of Coverage utility, I performed metric calculations of the NGS exome sequencing experiment yields. Version 3.0 of Depth of Coverage is a coverage profiler for a (possibly multi-sample) BAM file. It uses a granular histogram that can be user-specified to present useful aggregate coverage data. It reports the following metrics over the entire .bam file:

- Total, mean, median, and quartiles for each partition type: aggregate
- Total, mean, median, and quartiles for each partition type: for each interval
- A series of histograms of the number of bases covered to Y depth for each partition type
- A matrix of counts of the number of intervals for which at least Y samples and/or read groups had a median coverage of at least X
- A matrix of counts of the number of bases that were covered to at least X depth, in at least Y groups (e.g. # of loci with ≥15x coverage for ≥12 samples)
- A matrix of proportions of the number of bases that were covered to at least X depth, in at least Y groups (e.g. proportion of loci with ≥18x coverage for ≥15 libraries)

The general command line used for this end is:

```
 java GenomeAnalysisTK.jar -T DepthOfCoverage -R ucsc.hg19.fasta -I
input.realign.markdup.recal.srt.bam -L target.bed -o coverage.stats
-omitBaseOutput -ct 5 -ct 10 -ct 20
```

2.5 Steps 7, 8: Variant calling and quality filtering with GATK Unified Genotyper

The GATK Unified Genotyper is a multiple-sample, technology-aware SNP and indel caller. It uses a Bayesian genotype likelihood model to estimate simultaneously the most likely genotypes and allele frequency in a population of N samples, emitting an accurate posterior probability of there being a segregating variant allele at each locus as well as for the genotype of each sample. The system can either emit just the variant sites or complete genotypes (which includes homozygous reference calls) satisfying some phred-scaled confidence value. The genotyper can make accurate calls on both single sample data and multi-sample data. The input is represented by the read data from which to make variant calls, while the output is a raw, unfiltered, highly sensitive callset in VCF format.
Generic command for variant calling is the following:

```
 java -jar GenomeAnalysisTK.jar
-R resources/Homo_sapiens_assembly19.fasta
-T UnifiedGenotyper
-I sample1.bam [-I sample2.bam ...]
--dbsnp dbSNP.vcf
-o snps.raw.vcf
-stand_call_conf [50.0]
-stand_emit_conf 10.0
-dcov [50 for 4x, 200 for >30x WGS or Whole exome]
[-L targets.interval_list]
```

The above command calls all of the samples in the provided BAM files [-I arguments] and produce a VCF file with sites and genotypes for the samples. Several arguments have parameters that should be chosen based on the average coverage per sample in the data. This command can get a dbSNP data as input, as well as a target list file to restrict search to those regions that are within the target.

After performing local realignment around indels and base quality score recalibration using GATK, I called Single Nucleotide Variants (SNVs) and small insertions/deletions (indels) using GATK Unified Genotyper, and filtered out the variants by quality using GATK VariantFiltrationWalker with the following parameters:

```
--clusterWindowSize 10
--filterExpression "MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)"
--filterExpression "QUAL<30||QD<5.0||HRun>5||SB>-0.10"
```

for SNVs and

```
--filterExpression "MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)"
--filterExpression "SB >=-1.0"
--filterExpression "QUAL<10"
```

for indels.

The resulting Variant Call Format file has the following general format with information fixed fields:

- CHROM (chromosome): an identifier from the reference genome (Alphanumeric String, Required). All entries for a specific CHROM should form a contiguous block within the VCF file
- POS (position): the reference position, with the first base having position one (Integer, Required). Positions are sorted numerically, in increasing order, within each reference sequence CHROM
- ID: semi-colon separated list of unique identifiers where available (Alphanumeric String). If this is a dbSNP variant it is better to use the rs numbers
- REF (reference bases): each base must be one of A, C, G, T and N in uppercase, with multiple bases permitted (String, Required). The value in the POS field refers to the position of the first base in the string. For InDels, the reference string must include the base before the event, which must be reflected in the POS field
- ALT: comma separated list of alternate non-reference alleles called on at least one of the samples (Alphanumeric String; no whitespace, commas, or angle-brackets are permitted in the ID String itself). Options are base strings made up of the bases A, C, G, T and N, or an angle-bracketed ID String (”<ID>”). If there are no alternative alleles, then the missing value should be used

- QUAL: phred-scaled quality score for the assertion made in ALT (Numeric). High QUAL scores indicate high confidence calls. This field is permitted to be a floating point to enable higher resolution for low confidence calls if desired
- FILTER: PASS if this position has passed all filters, i.e. a call is made at this position (Alphanumeric String)
- INFO: additional information (Alphanumeric String)
- If genotype information is present, then the same types of data must be present for all samples. First a FORMAT field is given specifying the data types and order. This is followed by one field per sample, with the colon-separated data in this field corresponding to the types specified in the format. The first sub-field must always be the genotype (GT).

VCF records use a single general system for representing genetic variation data composed of:
- Allele: representing single genetic haplotypes
- Genotype: an assignment of alleles for each chromosome of a single named sample at a particular locus
- VCF record: a record holding all segregating alleles at a locus (as well as genotypes for multiple individuals containing alleles at that locus). VCF records use a simple haplotype representation for REF and ALT alleles to describe variant haplotypes at a locus. ALT haplotypes are constructed from the REF haplotype by taking the REF allele bases at the POS in the reference genotype and replacing them with the ALT bases.

2.6 Step 9: Variant annotation with ANNOVAR

Variants were annotated from the VCF file, against the NCBI RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/) and UCSC KnownGene (http://genome.ucsc.edu/) databases with ANNOVAR (www.annovar.org) [51]. ANNOVAR is an efficient software tool to utilize update-to-date information to functionally annotate genetic variants detected from diverse genomes (including human genome hg18, hg19). Given a list of variants with chromosome, start position, end position, reference nucleotide and observed nucleotides, ANNOVAR can perform:

1. Gene-based annotation: identify whether SNPs or indels cause protein coding changes and the amino acids that are affected.

2. Region-based annotations: identify variants in specific genomic regions, including conserved regions among 44 species, predicted transcription factor binding sites, segmental duplication regions, GWAS hits, database of genomic variants, DNAse I hypersensitivity sites, ENCODE H3K4Me1/H3K4Me3/H3K27Ac/CTCF sites, ChIP-Seq peaks or RNA-Seq peaks

3. Filter-based annotation: identify variants that are reported in dbSNP, or identify the subset of common SNPs (MAF>1%) in the 1000 Genome Project, or identify subset of non-synonymous SNPs with SIFT score>0.05, or find intergenic variants with GERP++ score>2, or many other annotations on specific mutations.

SUMMARIZE_ANNOVAR.pl is a script within the ANNOVAR package: given a list of variants from whole-exome, it generates an Excel-compatible file with gene annotation, amino acid change annotation, SIFT scores, PolyPhen scores, LRT scores, MutationTaster scores, PhyloP conservation scores, GERP++ conservation scores, dbSNP identifiers, 1000 Genomes Project allele frequencies, NHLBI-ESP 5400 exome project allele frequencies and other information. I used SUMMARIZE_ANNOVAR to annotate variants in the exomes I analyzed. Information given by this annotation process is in table format and is like the following:

Column	Name	Description
1	Func	Impact at the gene level
2	Gene	Gene name
3	ExonicFunc	Impact at the coding level
4	AAChange	Amino acid change
5	Conserved	Consevation of the genomic element (phastCons LOD)
6	SegDup	Score for segmental duplication (0-1)
7	ESP 5400	Frequency in ESV
8	1000G_ALL	Frequency in 1000 genomes (Feb.2012)
9	dbSNP135	Presence in dbSNP135

10	SIFT	SIFT score (the closest to 0 the most deleterious)
11	PolyPhen2	Polyphen2 score (the closest to 1 the most deleterious)
12	LJB_PhyloP	PhyloP score (the closest to 1 the most conserved)
13	LJB_MutationTaster	MT score (the closest to 1 the most deleterious)
14	LJB_LRT	LRT score (the closest to 1 the most deleterious)
15-END	Additional fields	Various sequencing metrics

2.7 Step 10: Gene prioritization

In order to identify disease-related alleles among the background of non-pathogenic polymorphisms and sequencing errors I applied a discrete filtering approach:

- *Novelty:* predicting that the causative variant would be rare and therefore likely to be previously unidentified in public databases or control sequencing data, novelty was assessed by filtering the variants against a set of polymorphisms that are available in public databases, i.e. dbSNP build 135 (http://www.ncbi.nlm.nih.gov/snp), 1000 Genome Project (http://www.1000genomes.org) and Exome Variant Server Project (http://evs.gs.washington.edu/EVS)
- *Stratification:* candidate alleles were stratified on the basis of their predicted impact or damage. To distinguish potentially pathogenic mutations from other variants, I focused on nonsynonymous variants, splice acceptor and donor site mutations and short coding indels. Additionally, candidate alleles were stratified by existing biological or functional information about a gene or on the basis of the sequence conservation
- *Intrafamilial segregation* (Teebi syndrome, agammaglobulinemia, ASD and CVID cases): the candidate variants were tested by Sanger sequencing for segregation among other members of the family, whose DNAs were available
- *De novo origin:* in Noonan trios I considered only those variants present in the proband and not in the unaffected parents.

3. Mutation validation

Sanger sequencing of PCR amplicons from genomic DNA was used to confirm the presence of variants in the candidate genes identified via exome sequencing and to screen the candidate genes in additional cases. All primers were designed using the informatic tool Oligo 6 (Molecular Biology Insights, Inc).

GENOME-WIDE GENOTYPING FOR LINKAGE ANALYSIS AND HOMOZYGOSITY MAPPING

1. SNP chip array

10 individuals from Teebi family, the two affected siblings in the consanguineous kindred with agammaglobulinemia, 17 from the CVID pedigree and each Noonan-trio were genotyped with the Affymetrix® Genome-Wide Human SNP Array 6.0 Nsp/Sty Assay (Santa Clara, CA, USA). This platform uses over 900,000 SNP probes and 900,000 2-naphthalenecarbonitrile (NPCN) probes with a median spacing of 7.0 kb. A total amount of 500 ng of genomic DNA for each sample was used for the experiment. The protocol was divided in stages:

- *Genomic DNA preparation:* each sample was diluted to 50 ng/µl using reduced EDTA TE buffer and 5 µl were used for each of the two following digestion steps
- *Sty restriction enzyme digestion, ligation and PCR:* the genomic DNA was digested with the restriction enzyme Sty I, it was ligated to a common adaptor with T4 DNA ligase and the template underwent PCR using TITANIUM^{TM} Taq DNA polymerase. The corrected amplification was confirmed by running 3 µl of each PCR product on a 2% TAE agarose gel
- *Nsp restriction enzyme digestion, ligation and PCR:* as for Sty I, 5 µl of genomic DNA were digested with Nsp I, ligated and amplified
- *PCR product purification with AMPure XP beads*
- *Quantification:* the templates had an OD between 0.9 and 1.2, which was equivalent to a final PCR product concentration of 4.5 to 6.0 µg/µl

- *Fragmentation:* the purified PCR products were fragmented using Fragmentation reagents (DNAse I) and the result was checked by running 1.5 µl of each reaction on a 4% TAE agarose gel
- *Labeling:* the fragments were end-labeled using terminal deoxynucleotidyl transferase
- *Target hybridization:* after denaturation, each sample was load onto a Genome-Wide Human SNP Array 6.0 – one sample per array. The arrays were then placed into a hybridization oven preheated to 50°C for 18 hours
- *Washing and staining arrays:* I used a three-stage protocol for mapping arrays: (1) a streptavidin phycoerythin (SAPE) step; (2) an antibody amplification step; (3) a final stain with SAPE.

Finally, I used Affymetrix GeneChip® Command Console (AGCC) to operate the fluidics station and the scanner. The genotype calls of each individual were determined by the Birdseed genotyping calling algorithm, embedded in the Affymetrix Genotyping Console 2.0 software. The number of samples used to determine the genotype calls varied depending on the examination.

2. Linkage data analysis

I performed linkage data analysis using PedStats [52] and Merlin [53] softwares. PedStats allows to verify that the information about individual relatedness and SNP marker alleles is correct. General command line for PedStats is:

```
pedstats -p input.ped -d input.dat
```

PedStats provides statistics about the family structure, the marker genotypes and allele frequencies. Markers showing non Mendelian inheritance, inconsistent genotypes and deviation from Hardy-Weinberg equilibrium can be excluded from downstream analysis.
Moreover, dense set of marker SNPs - as the ones I used - likely introduce additional biases in linkage analysis. In fact, linkage analysis assumes linkage equilibrium between adjacent markers. Linkage disequilibrium (that is, the non-random association of marker alleles on the same haplotype) can inflate LOD score, which is the statistical measure of linkage between a marker and the trait (the disease). Linkage disequilibrium is a genetic force that act in a range of up to 500 Kb. To eliminate the effect of linkage disequilibrium, I used custom scripts in

Perl language in order to select a sparser set of markers (inter-marker distance > 500 kb). Then, I converted the original files to formats which were suitable for linkage analysis.
The program I used for linkage analysis was Merlin. Merlin allows parametric linkage analysis with the following command:

```
merlin -d parametric.dat -p parametric.ped -m parametric.map --model
parametric.model
```

Input files are a 'ped' file containing the pedigree and marker allele information, a 'map' file containing the chromosome map and 'dat' file containing the list of markers used. A model file is used to specify the mode of inheritance parameters. I used Merlin to calculate both single-point and multi-point LOD scores. I set model parameters for autosomal dominant disorders in the 'parametric.model' file. I set phenocopy rate - the probability of being affected if carrier of two wild-type alleles, P(affected/aa) - equal to 0.001 and penetrance - the probability of being affected if carrier of at least a mutated allele A, P(affected/aA) and P(affected/AA) - equal to 0.99.

2.1 Fine mapping

To better define the critical region identified in the CVID family, 10 dinucleotide repeats were typed in the available samples: D3S1262, D3S3570, D3S3600, D3S1580, D3S1294, D3S1314, D3S2747, D3S2418, D3S2748, D3S1262. Primers for all markers were prepared by labeling the forward primers with FAM fluorescent dye. PCR were carry out under standard conditions. Amplification product were loaded in an automated sequencer AB3730 and allele size defined using the Genescan software.

3. CNVs analysis

Affymetrix 6.0 genotyping data sets from the six probands with Noonan syndrome and their parents were also analyzed for CNVs using a commercial software package (Partek Genomics Suite; Partek Incorporated, St. Louis, MO). Raw intensity signals data in form of .CEL files were normalized, then a CNV was defined as a genomic region of consistent copy number variation spanning ten or more adjacent SNPs. Moreover, the CNV calls were required to have LOD scores (probability of the segment being the stated copy number versus the copy

number of the flanking region) ≥5 for inclusion. In order to validate de novo CNVs found in Noonan cases, quantitative PCR was performed using the TaqMan® PCR kit with a standard protocol, in conjunction with the 7900HT Real-time PCR system from Applied Biosystems (ABI, Foster City, California, USA). Primers and probes (probe set) were designed using Primer Express® software from ABI. Test probe sets were designed to be within each respective CNV to be validated.

4. Homozygosity mapping analysis

An alternative to linkage analysis in consanguineous families is homozygosity mapping. Homozygosity mapping is not a statistical measure, but rather an observation of the homozygous identical-by-descent regions in an individual genome. I performed homozygosity mapping with the Runs of Homozygosity utility implemented in PLINK [54]. This algorithm takes a window of X SNPs and slides this across the genome. At each window position determine whether this window looks 'homozygous' enough (i.e. allowing for some number of hets or missing calls). Then, for each SNP, calculate the proportion of 'homozygous' windows that overlap that position.

The exact window size and thresholds, relative to the SNP density and expected size of homozygous segments, is obviously important: sensible default values are supplied for the context of dense SNP maps. In general, this approach will ensure that otherwise long runs of homozygosity are not broken by the occasional heterozygote.

The general command is the following:

```
plink –bfile mydata –homozyg
```

And to define the 'window' that slides across the genome I used the options:

(to define the sliding window)

```
--homozyg-window-kb 5000
--homozyg-window-snp 50
```

(to set the number of heterozygotes allowed in a window)

```
--homozyg-window-het 1
```

(to set the number of missing calls allowed in window)

```
--homozyg-window-missing 5
```

(to define the proportion of overlapping windows that must be called homozygous to define any given SNP as 'in a homozygous segment')

```
--homozyg-window-threshold 0.05
```

While to define the final segments that are called as homozygous or not, I used:

```
--homozyg-snp 100
--homozyg-kb 1000
```

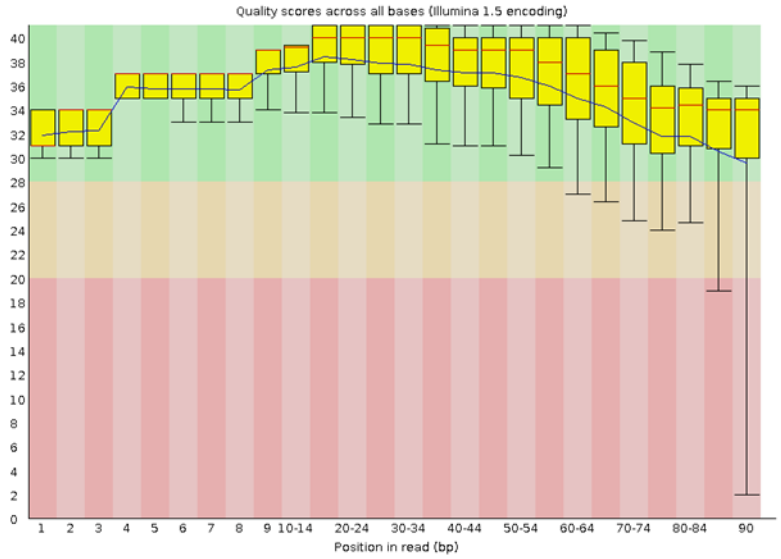
And to specify the required minimum density

```
--homozyg-density 50
```

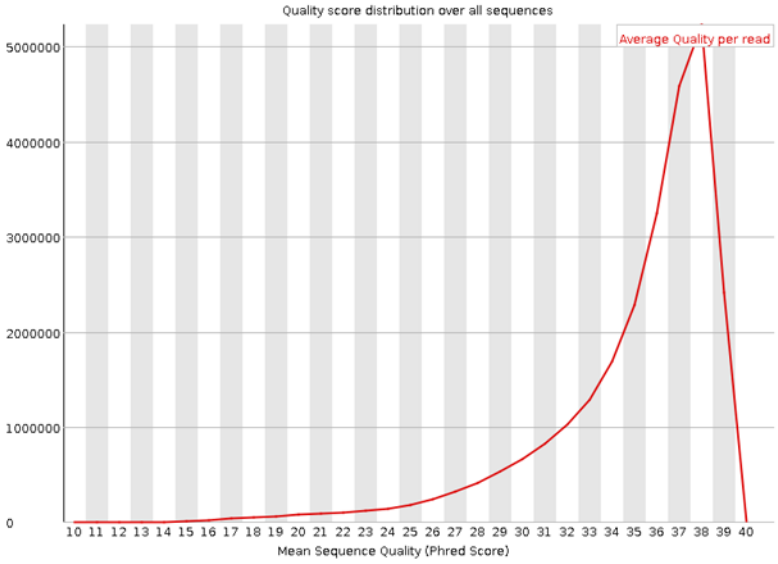
RESULTS AND DISCUSSION

The total set of sequenced exomes comprised 29 individuals. Before aligning the sequenced reads to the reference genome, all the samples were evaluated with FastQC. All the statistics passed the quality check (Figure 13). Using Nimblegen SeqCap EZ Exome as enrichment platform, I obtained a mean total of 7.9 Gb of sequence per individual as paired-end 76 or 101 bp reads from one lane of an Illumina Genome Analyzer II and HiSeq 2000 and a mean coverage of more than 100 fold, with 68.13% of bases mapping to the targeted exome (Table 5). The enrichment of reads that unambiguously mapped to regions outside the targeted bait intervals was approximately 9.3%, a percentage that correlated strongly with the general enrichment trend. Reads generated by hybrid selection tend to extend into sequences beyond the target region and the longer fragment library is, the more of these "near-target" sequences will be recovered. Moreover, off-target regions may be enriched if there is high sequence similarity between those regions and bait regions. In fact, a higher fraction of off-target reads mapped to repeat elements and segmental duplications than did on-target reads.

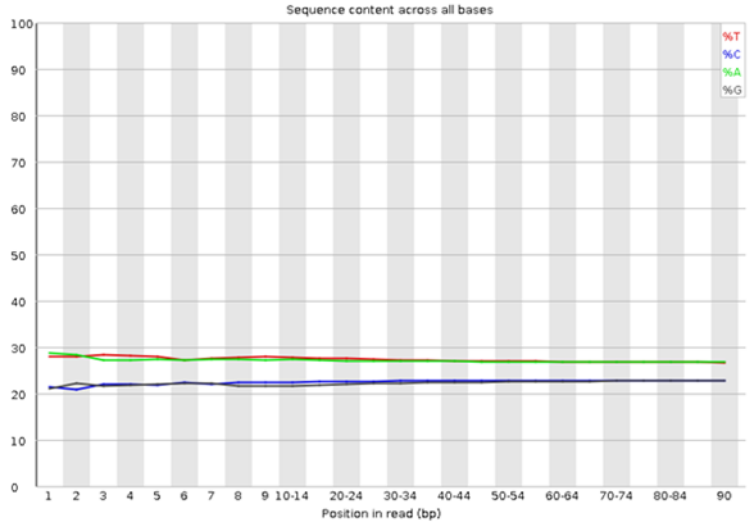
BWA mapped 99% of reads to human DNA; 98.6% of the targeted bases were covered at least once, and 96.8% showed a coverage greater than 10 folds. A major consideration in data analysis is the coverage needed to reliably identify sequence variants, which depend on multiple factors, such as the nature of the region of interest, the method used for targeted enrichment. I discarded reads with identical start and end sites. In fact, one technical artifact of capture-sequencing procedures is the generation of duplicate DNA sequencing reads that represent the repeated sequencing of copies of the same molecule. Detection of the duplicate reads by computational analysis is not trivial and generally relies on observation of the alignment positions. The presence of duplicate clones significantly influences the randomness of the sequencing process: some regions will have an unexpected high depth. This can also result in large frequency differences between the two alleles of an heterozygous site.



(**A**) Per base quality

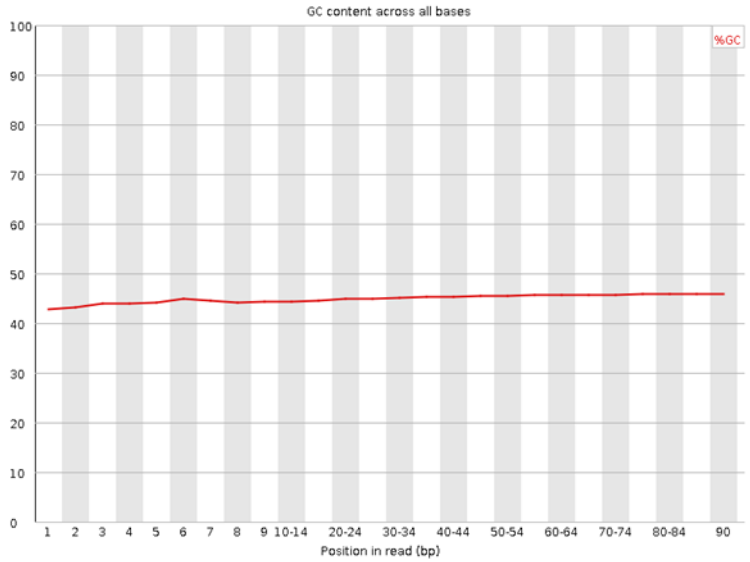


(**B**) Per sequence quality

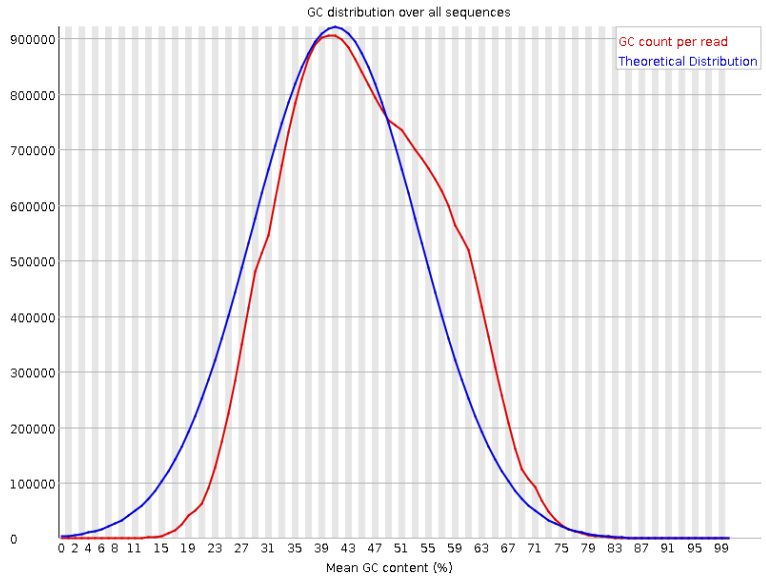


(**C**) Per base sequence content

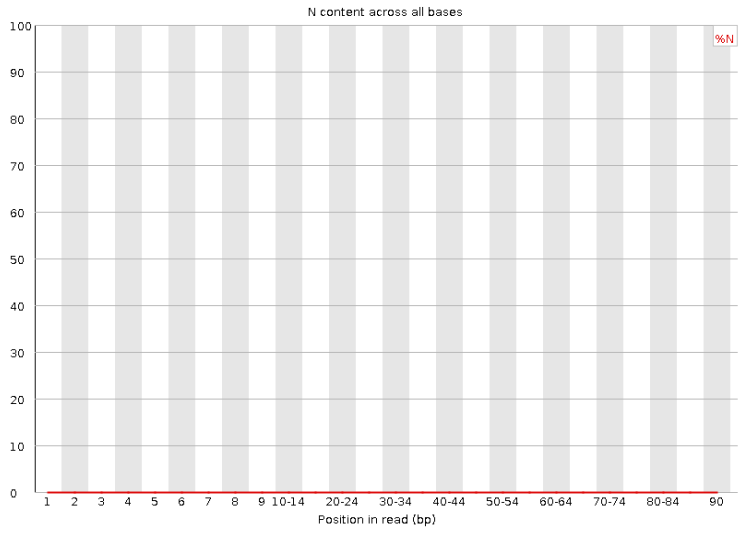
(Continued to the next page)



(**D**) Per base GC content

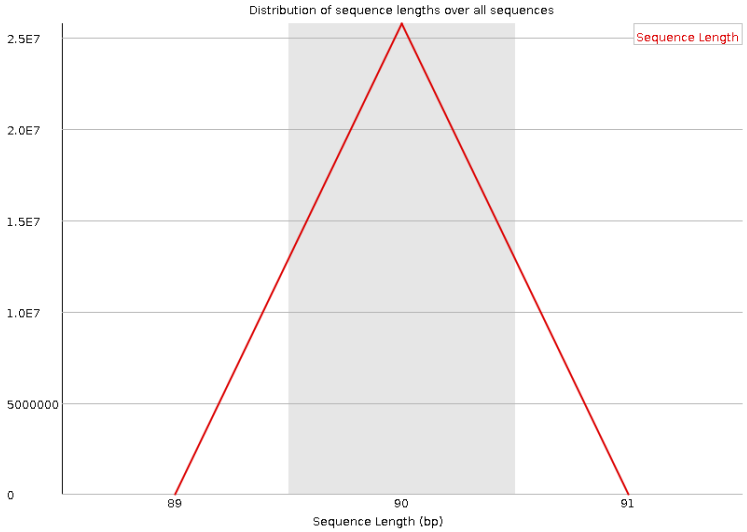


(**E**) Per sequence GC content

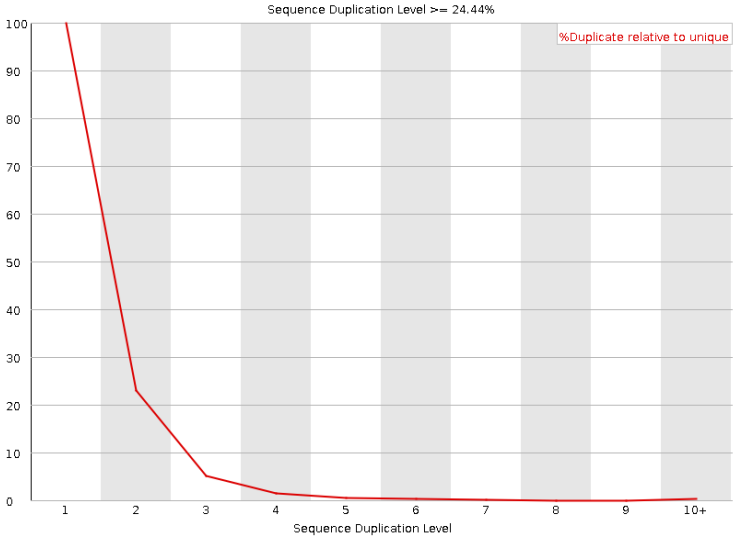


(**F**) Per base N content

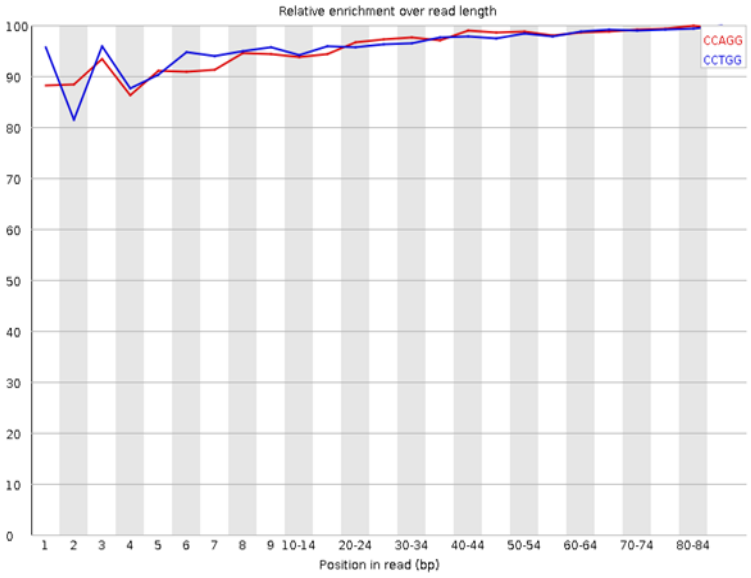
(Continued to the next page)



(**G**) Sequence lenght distribution



(**H**) Duplication level



(**I**) Kmer profiles

Figure 13. Graphical outcome of the FastQC analysis for sample S1.

Although enrichment efficiency is a function of read-depth, it does not necessary correlate with the ability to identify variants. SNPs represent the most numerous sequence variations in the human exome [55] and their accurate and comprehensive identification is a major goal of exome sequencing. To evaluate the generic SNP detection performance, I called variants in each dataset using GATK and, when possible, I compared the results analyzing the same samples with the Affymetrix 6.0 SNP Chip. Heterozygous positions in the chip were compared to the genotype calls in the exome sequencing data with a Phred-based quality score ≥30: the concordance rate was 99.5%. All nonconcordant genotype calls were calls of homozygous reference and SNP chips also have their own error rates that could account for some of the discordances.

Feature	Average
Total Gb content	7.9 Gb
Mean coverage	100.1X
% on target	68.13%
Paired reads duplicate	6.8%
Mean error rate	0.56%

Table 5. Summary of original exome sequencing data obtained on average per sample.

Identification of disease-causing gene among the variants generated by exome sequencing requires the separation of candidates with high pathogenic potential from variants that have a low-probability for disease causation. Biological source of low-interest variants include both common and rare population variation. On the other hand, high throughput sequencing techniques also generates unimportant variants in the form of genotype false positives. Errors can arise from biases in the library construction or errant polymerase reactions [56], difficulty making genotype calls at the end of short reads, loss of synchrony among DNA sequencing reactions within a cluster [57] or manufacturer/platform-specific mechanistic problems [58]. The errors occurred during sequencing often are not random; this is especially true for low quality bases and those near the 3'-end of reads. Illumina sequencing technology uses two lasers to excite the dye attached to each of the four nucleotides: A and C use the same laser, while G and T use another laser. As result, A-C and G-T substitution errors are significantly overrepresented.

Another problem is the misalignment of short reads to a reference sequence: since NGS read lengths are quite short, there is a higher probability of incorrectly mapping reads to loci with minimally divergent sequences, thus creating incorrect SNP calls [33]. Moreover, the reference sequence itself may be an additional source of variants: for some base positions, the reference sequence specifies a minor allele in most large human populations. Such biases occur because of the limited number of individuals on which the original reference sequence was based, sequencing an alignment errors. As a result, the NCBI human genome reference sequence includes minor or unique variants and, possibly, disease-causing mutations [59].

A transition mutation involves a change between two purines or two pyrimidines, while a transversion mutation involves a change from pyrimidine to purine or *vice versa*. This makes a transversion event twice as favourable as a transition event for any random mutation. Hence, the transition/transversions ratio (T_i/T_v) is a critical parameter for systematic errors in the sequencing technology, alignment artifacts and data processing failures. In my study, there was a slight increase in G→A/C→T transitions and slight decrease in non-G→C/C→G transversions because in the Nimblegen enrichment platform a larger percent of its target bases are in coding regions, which have a higher GC content and therefore different nucleotide substitution rates from the rest of the genome [60]. The T_i/T_v of total variants ranged from 2.53 to 2.67 and was slightly lower than estimates of 2.8 from the exome based on 1000 Genome data [61].

Among the sequenced samples, I found a mean of 19,279 exomic variations from the reference sequence per subject (Table 6). These variants included 18,871 coding SNPs, of which there were 8,918 missense variants, 85 premature termination codons, 79 canonical splice site variants and 9,789 synonymous substitutions. Splice sites included the 20 bp within exon boundaries (10 bp intronic and 10 bp exonic for each exon boundary). Small insertions and deletions (indels) were detected at a frequency of 12.5-14.5% that of SNPs. As expected [62], the frequency of indels present in the protein coding segments was much lower than in the total covered regions, which contains introns, UTRs and intergenic sequences. There was a strong bias toward indels of a size equal to multiple of three bases in coding regions. This pattern was presumably due to selective pressure against deleterious frameshift mutations in the coding regions.

Variant type	Mean number of coding variants
• *Total variants*	
Missense	8918
Nonsense	85
Synonymous	9789
Splice	79
Indels	408
Total	19279
• *Novel variants*	
Missense	162
Nonsense	2
Synonymous	103
Splice	3
Indels	122
Total	389
• *Non-novel variants*	
Missense	8756
Nonsense	83
Synonymous	9686
Splice	76
Indels	286
Total	18887

Table 6. Mean outcomes from exome analysis.

As thousands of single nucleotide variants and short indels have been detected in the sequencing of the exomes, I applied multiple robust filtering criteria to discern the causal variants (Table 7). I mainly focused on nonsynonymous variants, splice acceptor and donor site mutations and short coding insertions or deletions (indels), anticipating that synonymous variants would be far less likely to be pathogenic. Nonsynonymous mutations and splice-site disruptions are often assumed to be deleterious, but have a broad range of potential fitness effects to be evaluated.

Since I was studying rare pathogenic conditions, I also predicted that the putative causal variants would be as rare and therefore likely to be previously unidentified in public databases or control sequencing data, such as dbSNP, 1000 Genomes project and Exome Variant Server. However a caveat to note is that phenotypic information is not always available for the samples used in these data sets, and it is possible that pathogenic mutations are present in

them. Particularly, in case of recessive mutations there is a chance that a normal carrier could have been genotyped and the recessive disease-causing mutation deposited in the data base. Thus, in the agammaglobulinemia study I envisaged this possibility.

As homozygosity mapping in the consanguineous family, integration with linkage data had greatly facilitated the discovery of candidate genes in the families analyzed because it narrowed down the searching space. Moreover, sequencing multiple individuals from one family allowed me to hypothesize that in autosomal-dominant pattern – as in Teebi syndrome, ASD and CVID - all affected members should share the same causal variants.

To account for genetic and phenotypic heterogeneity in Noonan syndrome approach, I applied a less stringent strategy looking for candidate genes not only seen in all or in a subset of affected individuals, but also arisen *de novo,* unique to each case.

Mendelian disorder and sample	Variant filtering methodology and analysis strategies	Major results
• ***Exome sequencing of unrelated individuals***		
Noonan syndrome Six unrelated parents-child trios	- Focused primarly on heterozygous nonsynonymous coding variants - Removed presumably common variants - To allow for genetic heterogeneity, looked for *de novo* variants, unique to each trio	- in trio NS01 identified a nonsense mutations in *FOXC2*, associated to LD - in trio NS07 identified a del16p11.2 already described as pathogenetic - in trio NS06 identified a candidate variant in *PPP1R26* to be screened in a larger cohort of patients
• ***Exome sequencing in families with linkage analysis***		
Teebi syndrome Three affected individuals	- Performed linkage analysis using multiple individuals from the pedigree and found 5 regions with significant positive LOD score - Focused primarily on exomic and flanking intronic variants within the putative linkage regions	- Resulted in a list of 4 novel shared variants: 1 intronic, 1 in an evolving pseudogene and the other 2 already found in public databases
ASD Four affected individuals	- Considered primarly all nonsynonymous coding variants - Removed presumably common variants - Focused on variants in the linkage candidate region - Restricted the variants to those shared among all samples in heterozygous state, according to the inheritance pattern	- Identified a missense mutation in *SCAMP2* - Confirmed the cosegragation of the candidate variant with the phenotype, allowing incomplete penetrance
CVID Three affected individuals	- Considered primarly all nonsynonymous coding variants - Removed presumably common variants - Focused on variants in the linkage candidate region - Restricted the variants to those shared among the affected individuals in heterozygous state, according to the inheritance pattern	Not found any rare heterozygous variants in the linkage region shared among the affected members of the family
• ***Exome sequencing coupled with homozygosity mapping***		
Agammaglobulinemia One individual from a consanguineous family	- Focused on variants in the six homozygous segments - Considered primarly all nonsynonymous coding variants - Included the novel and low-frequency (<0.001) variants already reported in control databases	- Identified a homozygous variants in *BNIP1* - Confirmed the inheritance of each allele in the affected siblings from each parent

Table 7. Summary of the variant filtering methodology, analysis strategy and major results according to different study designs.

NOONAN SYNDROME

Noonan syndrome (NS, OMIM 163950) is a pleiomorphic and genetically heterogeneous autosomal dominant disorder predominantly characterized by distinctive facial dysmorphism, congenital heart defects, postnatally reduced growth, ectodermal and skeletal defects and variable cognitive deficits.

NS is caused by germline mutations that affect components of the RAS-MAPK pathway, accounting for approximately 70% of affected individuals [63]. To date, all mutations have had complete penetrance. As with most human autosomal dominant disorders, 1/3-1/2 of cases arise through *de novo* mutations.

In the Noonan project, I evaluated the use of Next-generation sequencing to provide genetic diagnosis using six parent-child trios in which the child had Noonan syndrome due to unexplained molecular mutations. Importantly, the patients were chosen to be representative of the 30% of patients with clinical diagnosis of NS but without mutations in the known genes previously associated to the disorder and they were not selected for phenotypic homogeneity. The rationale for this approach was the ability to reduce the number of candidate variants that we expected to detect in the affected individual to only those that were not observed in either parent. For autosomal dominant disorders, this strategy can discover *de novo* coding variants, as neither the parent is predicted to have a mutation that causes a fully penetrant dominant disorder [64].

Parental and proband alignments were examined for all potential *de novo* variants. The majority was ruled out for one of the following reasons: low coverage in the parents (<10X), variant was visibly present in parental alignments by not called by GATK, alignments presented multiple mismatches in same read or the variant was at the very end of reads in probands and/or parents.

After the filtering procedure, I found a possible candidate variant in one case, a causative mutation in *FOXC2* and a chromosomal deletion already associated to other diseases in two distinct probands, and no candidate gene containing previously unknown coding, novel, *de novo*, heterozygous variants in the remaining three trios.

The putative candidate variant in trio NS06 was a *de novo* Serine to Leucine substitution at position 1161 in one of the regulatory subunits of the protein phosphatase 1. PPP1R26 shows ubiquitous expression, which is consistent with the multisystemic defects in individuals with Noonan syndrome. Relatively little is known about the function of the encoded protein,

except that it seems to inhibit the phosphatase activity of protein phosphatase 1 (PP1) complexes, positively regulating cell proliferation [65]. A serine phosphorylation site was predict at this position by several *in silico* predictor tools, such as NetPhos 2.0 Server (http://www.cbs.dtu.dk/services/NetPhos/).

Even if the functional role of the variant in the Noonan pathology do not appear so clear from what we know about the PPP1R26 protein, since it has been proven that in a given trio the overall rate of *de novo* germline mutations is very low [66], it would be worth to further screen the candidate gene in a cohort of cases with clinical diagnosis of Noonan syndrome without mutations in known genes already associated to the disorder. In fact, the finding of independent *de novo* mutations in the same gene among even a small number of affected subjects would constitute compelling evidence of disease causation.

More interesting, a *de novo* non-synonymous mutation was identified in *FOXC2* in trio NS01, resulting in a premature truncated protein at amino acid position 99 (Y99X), confirmed by Sanger sequencing. This variant has already been reported to be causative in Lymphedema distichiasis syndrome (LD, OMIM 153400), a clinically heterogeneous and rare developmental disorder [67]. LD is characterized by lymphedema of the limbs with variable age of onset and double rows of eyelashes (distichiasis). There is a wide variation of associated secondary features including cleft palate and extradural cystis. Some of the patients with a well characterized FOXC2 lesion have complex congenital heart disease and ptosis, suggesting additional overlap with Noonan syndrome [68].

At the first clinical inspection, the proband presented the typical features of Noonan syndrome including facial dysmorphism (hypertelorism, epicanthic folds and downward slanting palpebral fissures, low-set posteriorly rotated ears and a low posterior hairline), webbed neck, *pectus carinatum* superiorly and *pectus excavatum* inferiorly and lymphedema of the limb. After identification of the mutation in FOXC2 suggested the diagnosis, a follow-up clinical evaluation revealed also mild distichiasis, further confirming the molecular indication for the diagnosis of LD. The phenotypical differentiation between Noonan and LD syndrome is quite difficult, as in both conditions a large variety of common signs, as webbed neck, lymphedema, ophthalmic and heart deformities, are known. However, distichiasis is only found in LD. Therefore, my data suggest that in all cases with Noonan-like features and lymphedema, lacking in a proven direct check for distichiasis, LD should always be considered as a differential diagnosis. For this reason, I screened 6 additional Noonan patients

with lymphedema, but they did not show any mutations in the entire coding sequence of *FOXC2*.

A parallel SNP-array analysis on the NS07 trio revealed a *de novo* 500-kb deletion on chromosome 16p11.2 in the affected proband. This alteration was first described in 2010 [69] and a more accurate clinical investigation on the affected offspring revealed an effectively Noonan-like phenotype, which perfectly fit the main symptoms associated to the deletion. Although prenatal history, facial dysmorphism, post natal sucking difficulty, obstructive hypertrophic cardiomyopathy and Chiari malformation type 1 alone were suggestive for NS, it emerged that together with other peculiar symptoms, such as macrocephaly, behavior disorders (attention deficit, speech delay) and syringomyelia, the phenotype was strictly due to the deletion, as already reported [70].

I found causal or candidate variants in ½ patients and in the two cases (NS01 and NS07) where we found a clear cause of the condition, this conclusion depended on the knowledge of the involvement of the variants in other Mendelian diseases. In particular, the FOXC2 case together with that of the 16p11.2 deletion illustrate how that 30% of Noonan patients lacking in the molecular diagnosis could include a broader range of Mendelian conditions and shows how a wrong diagnosis can easily be corrected by whole-exome sequencing.

There are many reasons for having missed the causal variants in half of the trios. One potential explanation includes the fact that NS is genetically heterogeneous and it is hard to identify an heterozygous *de novo* mutation with only few affected individuals: heterogeneity clearly increases the number of candidate genes that must be considered.

On the other hand, it is possible that a mutation was located in a poorly covered exon and thus escaped detection. In fact, I achieved a coverage of the targeted region for up to about 70% for each sequenced exome. If the mutations were located in poorly covered exons, the candidate gene would falsely be removed from further consideration.

Beside the fail in detecting the causal variants because of missing sequence or annotation, another important factor is that we don't have a comprehensive understanding of the function of most genes; it is also possible that causal variants may exert their effects through more complex inheritance patterns. Alternative models of inheritance include:

1) X-linked inheritance: two of the three affected probands sequenced were male, so I looked for recessive variants on chromosome X passed from the mother to the child;

2) Autosomal recessive model:

- simple: I looked for homozygous variants in each child inherited from both parents

- compound heterozygosity, which refers to the presence of two different heterozygous variants occurring in two positions in the homologous chromosomes. There are two possible combinations:

- the child inherited a variant from each parent
- one variant was inherited from a carrier parent and the other one arised *de novo.*

Analyses founded on these alternative inheritance models still failed to identify causative variants in the three remaining trios.

TEEBI SYNDROME

Teebi syndrome (TS, OMIM 145420) is a rare condition first described in 1987 [71] characterized by hypertelorysm with a facial appearance that can resemble craniofrontonasal dysplasia, limb anomalies, urogenital anomalies, cardiac defects and umbilical hernias. The facial features are striking and include a prominent forehead, pronounced ocular hypertelorism, heavy and broad eyebrows and a high and broad nasal bridge.

Autosomal dominant inheritance has been established on the basis of male-to-male transmission, an equal number of affected males and females and an absence of more severely affected males compared to females. In cases where chromosome analysis has been performed, no cytogenetic abnormality has been identified and the molecular genetic basis for TS remains unknown [72].

I analyzed a moderate-sized multi-generational kindred with several cases of TS with a substantial phenotypic heterogeneity and a strongly suspected incomplete penetrance of the causative allele. I sequenced the exomes of three individuals with clinical diagnosis of TS, who were as far apart as possible in the pedigree in order to minimize the number of shared variants among the exomes.

Again, I focused my analysis primarily on nonsynonymous variants, splice-site mutations and coding indels shared by all three samples because the disease is autosomal dominant, rare and they are relatives, so it is likely that their causative mutation was inherited from a common ancestor.

There were only three missense mutations shared in the three exomes (Table 8), all confirmed by direct sequencing. Then, I checked the segregation of two of the variants codifying for functional proteins among all members of the family whose DNA was available. I found that

the variant in *RBM28* was heterozygous in all affecteds, while among the three unaffected individuals analyzed, only one showed the variation, in accordance with the incomplete penetrance model suggested from the pedigree structure. But the *RBM28* variant was also been reported at low frequency in control databases, strongly suggesting that it was a rare polymorphism. The variant in *RSPH1* was reported to be homozygous in two cases, thus its causative role in the disease was excluded. In fact, it was already found in the control population at a frequency higher than 9%.

Gene	Position (hg19)	Effect	Function
RBM28	Chr7:127,950,437-127,983,962	missense p.M243L	Nucleolar component of the spliceosomal small nuclear ribonucleoprotein (snRNP) complexes
RSPH1	Chr21:43,892,597-43,916,401	missense p.G248R	Male meiotic metaphase chromosome-associated acid protein
TYW1B	Chr7:72,023,729-72,298,813	missense p.G27S	This locus appears to be an evolving pseudogene, but may still be functional in some members of the population

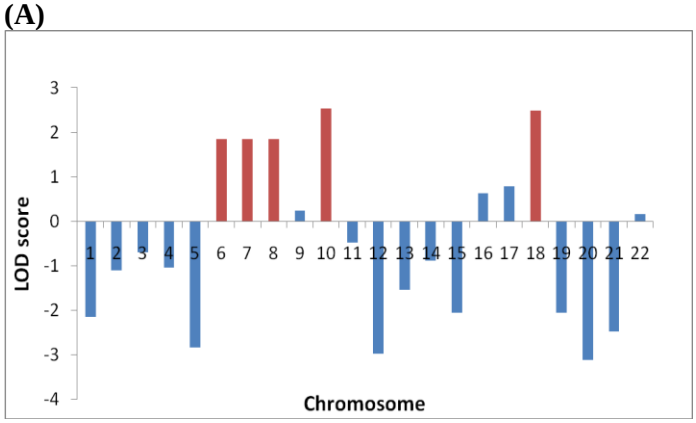
Table 8. Final candidate SNPs in Teebi syndrome study.

The exome sequencing data for the Teebi family were analyzed in parallel at Yale University by R. Lifton's group. This analysis indicated an intronic variation in *PTPRZ1* as the only novel heterozygous variants shared among all three affected individuals. This variant was not a good disease causative mutation because it was intronic, localized 36 bp upstream of exon 14, far from the splice-site region. Of interest, however, this variant resided on chromosome 7, about 6 Mb away from the variant in *RBM28*, showing the same pattern of segregation among the family. Taken together, these results indicated possible genetic linkage of TS to this genomic locus. In order to delineate the extension of the putative linkage region, as a first step I used the exome data for known SNPs as linkage markers; in this way I was able to delineate a region in which all three affecteds shared at least one allele for each marker.

As validation, I performed linkage analysis using multiple individuals from the pedigree (7 affecteds and 3 unaffecteds). After genotyping with a set of about 900,000 SNPs, multipoint linkage analysis was performed using the Merlin linkage software assuming a rare susceptibility allele (frequency 0.0001) and a dominant model with reduced penetrance (set at

0.8). Setting the minimal inter-marker distance at 0.5 cM, I obtained a map of 200,000 SNPs. Allelic frequencies were those found in the general population. Five candidate regions with positive LOD scores were found. In particular, those on chromosomes 10 and 18 reached scores higher than 2, evidence for suggestive linkage. Of note, I confirmed the shared region on chromosome 7 because it overlapped almost perfectly with that delineated by exome data (Figure 14).

(A)



Chr	Max LOD	Mb tot	Region
6	1.85	3.66	6p25
7	1.85	12.9	7q31-32
8	1.85	8.73	8p23-22
10	2.53	16.14	10q21-22
18	2.48	1.73	18q22.3

(B)

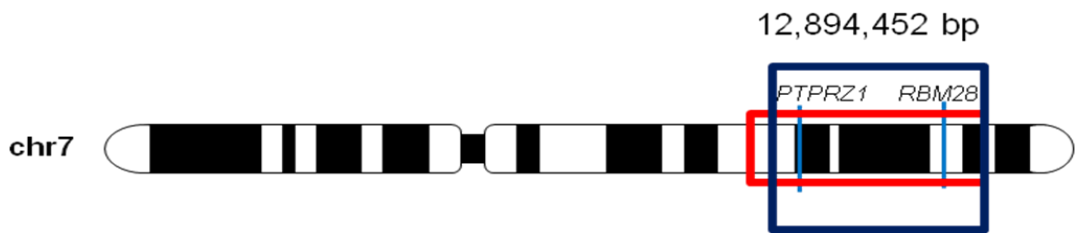


Figure 14. (**A**) Maximum LOD scores across the genome. (**B**) Extension of the putative linkage region on chromosome 7 according to exome processed data (red) and classical linkage analysis (blue).

No causative mutation in the linkage intervals was identified by exome sequencing, even after deeper analysis of the candidate regions. Several reasons may explain the failure to identify the gene responsible for Teebi syndrome in this family, including the inability to capture regulatory or evolutionary conserved sequences in non-coding regions and, most significantly, incomplete capture or inadequate sequencing depth of all target sequences.

ATRIAL SEPTAL DEFECT

Atrial septal defect (ASD) is one of the most common types of congenital heart diseases, accounting for about 33% of all congenital cardiovascular deformities, and is associated with a significant increase in the morbidity and mortality of affected individuals. ASD is clinically classified into 5 types by whether they involve other structures of the heart, but in general it is defined by an anatomically deficient interatrial septum allowing blood to flow directly between the left and right atria. Although the aberrant development of the atrial septum is implicated in a heterogeneous and complex biological process associated with environmental and genetic risk factors, accumulating evidence indicates that genetic defects play important roles in the pathogenesis of the disease. ASD is genetically heterogeneous: mutations in transcription factors, such as GATA4, GATA6, NKX2, TBX5 and TBX20, and cardiac structural proteins, such as MYH6 and ACTC1, were identified in familial cases. Nevertheless, the molecular etiology responsible for ASD in most affected individuals remains to be identified [73].

I used a combined strategy of exome sequencing and linkage analysis to identify a novel ASD causative gene in an American four-generation ASD family. Sequencing multiple affected individuals from one family allowed to hypothesize that all affected individuals should share the same causal variant, as ASD was inherited in an autosomal-dominant pattern in this family.

To localize the disease-causing gene, after exclusion of known ASD loci, previously a genome-wide linkage analysis using multiple individuals in the family was performed, which led to the identification of a single shared region on chromosome 15 (q23-q24.3) with a LOD score of 2.8 ($\theta = 0.00$). This locus was found to span about 6 Mb of genomic DNA and include 52 reference genes. Then, I sequenced the whole exome of 4 affected individuals in the family.

Although this study applied almost similar variant filtering strategies as with the Teebi analysis, comparison of the exome data among the four cases to find the shared variant in the linkage region was sufficient to identify *SCAMP2* as the sole candidate gene containing a new nonsynonymous variant (p.R292Q) predicted to be possibly damaging with a score of 0.818 (sensitivity 0.84; specificity 0.93) by PolyPhen-2 v2.1.

Genetic scanning of the family members available displayed that the gene variant we found was in the heterozygous state in all affected and among five unaffected subjects, two of

which obligated carriers, in accordance with the incomplete penetrant model suggested from the pedigree structure.

SCAMP2 encodes a 329-amino acid secretory carrier membrane protein which shows the highest expression in heart, placenta, and skeletal muscle [74]. It is not well characterized on the functional side, but it seems to act as a recycling carrier to the cell surface in post-Golgi recycling pathways.

As a further test of the significance of the variant found, I amplified and Sanger sequenced all 9 coding exons and associated splice sites of *SCAMP2* in an additional cohort of 54 individuals with a similar phenotype, but unfortunately the screening did not revealed any pathogenic mutations within the coding region of the gene.

This study was in partnership with the Pediatric Cardiac Genomic Consortium (PCGC), a group of clinical research teams collaborating to identify genetic causes of human congenital heart disease and to relate genetic variants present in the congenital heart disease patient population to clinical outcomes, which is currently involved in ongoing experiments to verify if the candidate genetic variant may cause the disease and in a mutational screening on a larger cohort of patients, since the need for functional work will decreases as the same variant or different nonsynonymous variants are shown to occur in multiple patients with similar presentations.

Functional analyses of the candidate variant may ultimately establish *SCAMP2* as a new gene responsible for dominantly inherited ASD. To date, there is not yet sufficient arguments on the functional effects of *SCAMP2* mutations in congenital heart defects and further work will be required to determine the molecular mechanism by which SCAMP2 contributes to the formation of aberrations in the atrial septum.

COMMON VARIABLE IMMUNODEFICIENCY

Common variable immunodeficiency (CVID) is a clinically and genetically heterogeneous group of disorders characterized by antibody deficiency, hypogammaglobulinemia, recurrent bacterial infections, and an inability to mount an antibody response to antigen. The defect results from a failure of B-cell differentiation and impaired secretion of immunoglobulins. CVID represents the most common form of primary immunodeficiency disorders and is the most common form of primary antibody deficiency [75]. The heterogeneity of the CVID phenotypes suggests a complex etiology, which is likely to include different monogenic

defects, as well as the combined effects of several susceptibility alleles together with environmental factors, including chronic infections. The majority of CVID cases are sporadic; approximately 20% are familial with a predominance of autosomal dominant over recessive patterns of inheritance [76]. Rare autosomal recessive mutations in ICOS, BAFF-R, CD19, CD20, CD81 coding genes have been recently reported and mutations in the TACI gene – *TNFRSF13B* - have been found in about 15% of cases. However, the underlying genetic defects remain unknown in the majority of cases [77].

In order to identify genetic variants responsible for CVID, genome-wide linkage scan was performed in a five-generation family with autosomal dominant inheritance of the disease.

Linkage analysis revealed a single shared region on chromosome 3q27.2-29 flanked by SNP markers rs62291969 (proximally) and rs3219697 (distally) with a maximum LOD score of 3.9 at marker rs3221020, while no other chromosomal loci achieved a LOD score higher than 2.0. The linkage result was further confirmed using ten additional microsatellite markers. The highest probability haplotype in the CVID pedigree was reconstructed using the Haplotyping option of Merlin and two key recombinant events were identified: recombination between D3S2748 and D3S1265 in individual IV-11 set the telomeric border on D3S1265; another recombination event was present between D3S3570 and D3S3600 in patient II-2, mapping the disease centromeric border to marker D3S3570 and affected descendents III-2, IV-2 and V-1 of this individual inherited this recombinant chromosome (Figure 15). According to these two key recombinations, a particular single haplotype, comprising identical alleles spanned by microsatellite markers D3S3570 and D3S1265 was identified to cosegregate with the disease phenotype in all examined affected family members with CVID. This locus was found to span 9.2 Mb of genomic DNA and included 47 genes.

One of the genes in the CVID locus is *BCL6*, mapping in the linkage interval between markers D3S3600 and D3S1580. BCL6 is involved in the gene regulatory network that controls the B cell terminal differentiation in the germinal center [78]. Since affected individuals may suffer from a block of the germinal center development, BCL6 might be a good candidate for CVID in our family. The analysis of the BCL6 gene was performed by direct sequencing of the 10 exons and exon/intron junctions on genomic DNA from two patients belonging to the family with recurrence of CVID. No causative alterations in the coding regions or in the splicing sites have been identified except a synonymous variant (p.N387N) reported at high frequency in the control population.

Then, I performed exome sequencing in three affected individuals whose overlapping haplotype produces the smallest shared genomic region.

After a three-step filtering procedure to extract heterozygous variants located in the linkage region and with low frequency (<0.01) in the general population, I did not find any candidate disease-causing allele shared among the three affected subjects analyzed. Several different reasons for failing in the identification of a strong disease candidate mutation have been mentioned above in this chapter. Here, linkage analysis yielded a single convincing signal on chromosome 3, thus allowing the narrowing of candidate variants to only those that are located within the 9.2 Mb linkage interval. As in other whole-exome sequencing projects in autosomal dominant families, it could be expected to identify only a single candidate gene in the LOD score region. However, the fact that no mutation was detected highlighted that some causative variants could escape identification by exome sequencing simply because they lied outside the coding regions. In those projects in which the search for disease-related genes can be limited to a specific region, whole genome sequencing should be considered the strategy of choice to capture non-coding variations that might underlie human disorders.

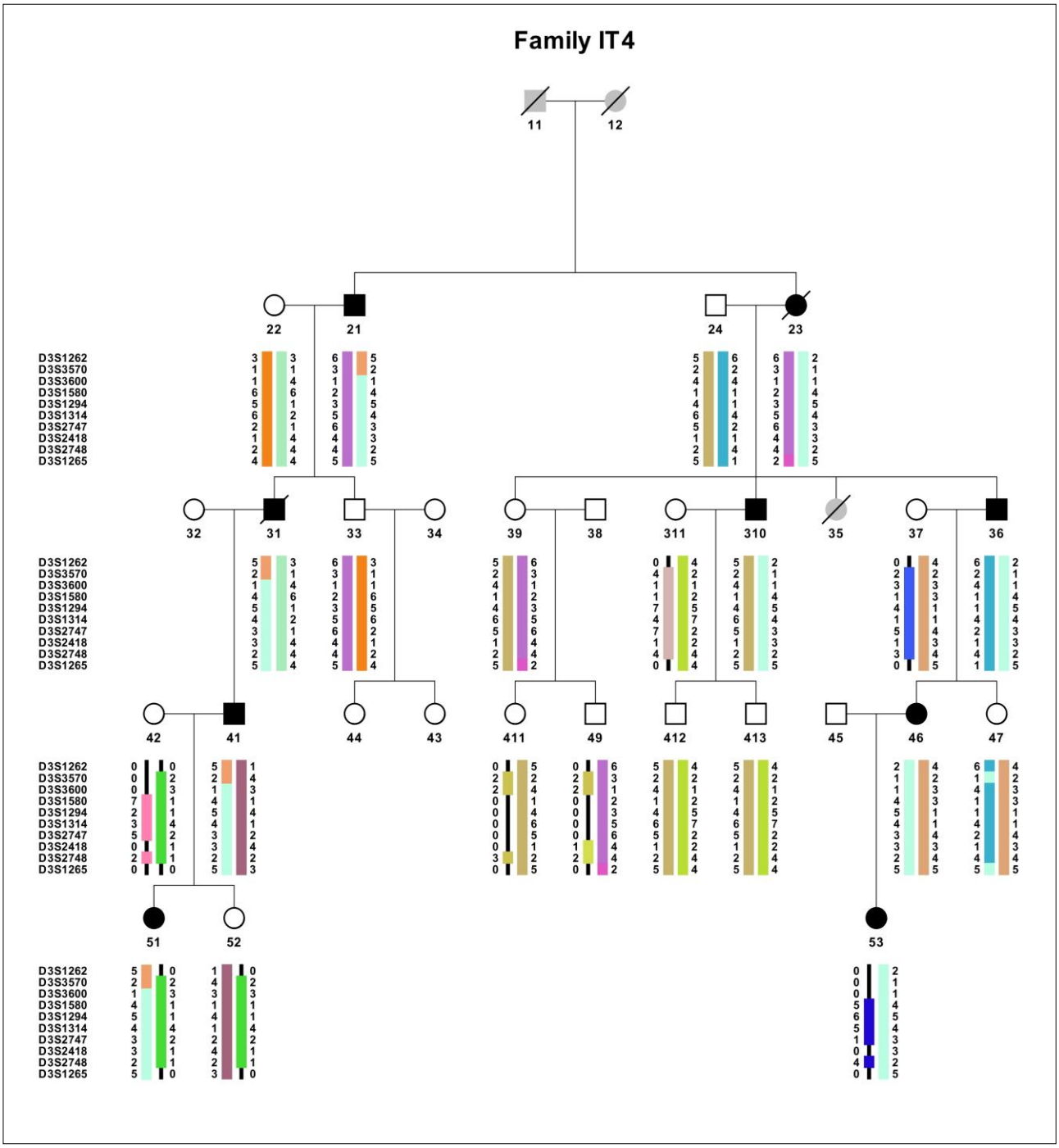


Figure 15. Pedigree and haplotype reconstruction of the CVID family. The light blue bar indicates the haplotype assumed to carry the disease allele (markers D3S3600, D3S1580, D3S1294, D3S1314, D3S2747, D3S2418, D3S2748, alleles 1-4-5-4-3-3-2). Recombinant chromosomes are colored accordingly to their chromosomes of origin.

AUTOSOMAL-RECESSIVE AGAMMAGLOBULINEMIA

Agammaglobulinemia is a rare congenital immunodeficiency characterized by absence of circulating B cells and low or absent serum immunoglobulin levels. In 86% of cases, childhood-onset agammaglobulinemia is an X-linked condition (XLA) affecting male offspring. XLA is caused by mutations in *BTK*, which encodes a signaling molecule downstream of the B cell antigen receptor. In the remaining patients affected by the autosomal recessive form of the disease, mutations in the components of the pre-B cell receptor, such as μ heavy chain, lambda5, Igα or downstream signaling elements have been described, but there are also patients without a definitive genetic diagnosis [79].

Since exome sequencing has also been swiftly integrated with homozygosity mapping to accelerate the investigation of recessive disorders in consanguineous families [80], I combined homozygosity mapping and whole exome sequencing in a consanguineous Italian family, composed by two affected siblings from a first cousins marriage. A two-step approach have been applied: the first step used the genome wide SNP genotyping to identify autozygous regions to narrow down the search space for possible loci; the second step examined exome sequences to identify genetic variations at base-pair resolution.

Genome-wide SNP genotyping was performed in the affected sisters using Affymetrix 6.0 array, defining 6 homozygosity regions longer than 2 Mb in chromosome 2, 5, 6, 8, 11 and 22 (Table 9).

In order to identify the responsible variant, genomic DNA of one patient was evaluated by exome sequencing. The subsequent comparison with the homozygosity mapping results identify a total of 7 variants in the candidate regions. Among them, after filtering out variants already reported in SNP databases at a frequency higher than 0.005 and variants with un uncorrected segregation pattern in the family, there was only one homozygous nonsynonymous variant that was previously unreported on chromosome 5, located in exon 2 of *BNIP1* and representing a likely candidate (Table 10). The mutation was confirmed to be homozygous in both affected subjects and to be heterozygous in both parents using Sanger sequencing. *In silico* analysis of the p.R55H substitution was performed with different bioinformatics tools: both SIFT and PolyPhen2 predicted that the mutation is deleterious. Moreover, multiple alignment of homologs from several species showed that the arginine in that position was a highly conserved residue.

SNP_start	SNP_end	Pos_start	Pos_end	kb	SNPs
• **Chromosome 2**					
rs11894732	rs4075737	50,684,575	66,530,723	15846	3774
rs4973472	rs12478296	232,550,160	243,048,760	10498	2487
• **Chromosome 5**					
rs11749351	rs1423115	37,865,461	55,641,683	14436	2950
rs4973472	rs12478296	167,749,878	176,040,365	4290	1358
• **Chromosome 6**					
rs1264344	rs13201350	30,800,577	44,033,168	13232	3364
• **Chromosome 8**					
rs7815277	rs9324551	131,960,656	142,562,774	10602	3283
• **Chromosome 11**					
rs4930358	rs4980619	66,133,311	70,439,839	4306	598
• **Chromosome 22**					
rs1006015	rs3761430	17,722,536	31,538,002	13790	3018

Table 9 Results of the homozygosity mapping.

Little is known about the exact function of the encoded protein in human biology, except that BNIP1 is a member of the BCL2/adenovirus E1B 19 kd-interacting protein (BNIP) family. It interacts with the E1B 19 kDa protein, which protects cells from virally-induced cell death. The encoded protein also interacts with E1B 19 kDa-like sequences of BCL2, another apoptotic protector. In addition, this protein is involved in vesicle transport into the endoplasmic reticulum and autophagy.

A further suggestion to the involvement of *BNIP1* in the pathogenesis of agammaglobulinemia came from the recent association of mutations in *LRBA* (LPS-responsive vesicle trafficking, beach and anchor containing) with hypogammaglobulinemia [81]. As BNIP1, LRBA is more expressed in cancer cells and it has been suggested that the protein acts as a positive regulator of cell survival by promoting proliferation and by preventing apoptosis; in the published work, the authors showed that individuals with

homozygous *LRBA* mutations have severe defects in B cell development, due to increased cell death.

Therefore, the next step will consist in the determination whether mutations in this gene might account for additional cases of agammaglobulinemia. In fact, a major problem in studying rare diseases is that one can never be entirely certain that a given gene is the sought after disease-gene until a second unrelated individual or family are described with a mutation in the same gene and a comparable phenotype. Since the mutation is predict to be very rare, I will primarily focus on products of consanguineous unions; then, I will use population-matched controls, i.e. unaffected individuals coming from the same geographical area, as filter for common regional variants.

Gene	Position	Effect	EVS freq	1000G freq	dbSNP ID	Sanger confirmation
HJURP	Chr2:234,749,399	missense H676R	-	-	-	Siblings homozygous Father homozygous
C7	Chr5:40,977,888	missense S308W	-	-	-	Siblings homozygous Mother homozygous
BNIP1	Chr5:172,573,948	missense R55H	-	-	-	Siblings homozygous Parents heterozygous
DAAM2	Chr6:39,865,007	missense I856T	0.0041	0.0014	rs61748650	Siblings homozygous Parents homozygous
CUL7	Chr6:43,021,587	missense G4S	0.0003	-	-	Siblings homozygous Mother homozygous
MICAL3	Chr22:18,364,065	missense R749Q	0.0004	-	-	Siblings homozygous Mother homozygous
CLTCL1	Chr22:19,263,266	missense V44F	0.0055	0.0023	rs34869740	Siblings homozygous Parents homozygous

Table 10. Variant filtering results in the consanguineous family with agammaglobulinemia.

A NEW SNP CALLING PIPELINE

Although methods for calling single nucleotides substitution are maturing, there is considerable room for improving efficient bioinformatics algorithms, necessary for analyzing the next generation sequencing data. Thus, I contributed to the development of a new tool for variant detection in deep sequencing datasets: MiST, the Mount Sinai SNP toolkit [82].

Several software pipelines that analyze the data from NGS are currently available: broadly, most approaches involve mapping the sequences to the reference genome to generate BAM/SAM files. These aligment files are subsequently analyzed to infer variants and SNPs [83].

In contrast, MiST - a new variant calling platform building on a previously published tool, Geoseq [84] - closely mimics the experimental technique, using exonic sequences as bait to identify sequences that can potentially map to the exon. A subsequent fine mapping step, which aligns the selected reads against the exon, permits a more sensitive and accurate identification of SNPs and variants. This approach reduces the computational complexity and allows for more sensitive mapping.

Our SNP calling pipeline is a six-step framework capable of discovering high-quality variation using diverse sequencing machines and experimental designs (Figure 16):

1. *Reference sequence pre-processing*. The exons of RefSeq mRNAs from the latest build of the genome are retrieved from the UCSC genome browser (http://genome.ucsc.edu/), augmented with other exons that are on the target list for the capture kit. Exons with overlapping genomic intervals are merged into super-exons. Each exon is extended by 70 bp on the 5' and 3' end to capture reads that can reach into introns (the length of extension depends on the average insert sizes, which is determined by the experimental protocol).

   We indexed sequence sets to facilitate rapid mapping, using Suffix-array implementation and BLAST [85]: the suffix-array index is used to rapidly find exact matches; the BLAST index is used to search for map locations with mismatches.
2. *FASTQ pre-processing*. Sequences containing ambiguous bases (usually marked as Ns) or with stretches of poor sequence quality (any 16-nt window with an average quality score less than 10) are removed. If either of the reads in a read-pair fails the quality check, the pair is discarded. We do not consider the quality scores beyond this step because the quality values are inaccurate and difficult to correct as they are a

moving target with manufacturers upgrading their algorithms for quality score calculations.

3. *Sequence retrieval and filtering*. Using Geoseq, tiles of a selected word size from the reference exon are used to retrieve matching reads from each of the subsets. Matching reads are then validated by identifying the approximate locus for each read-pair and keeping only those pairs that originate within a window (determined by the insert size) of the ends of the target exon.

   Matching reads are then validated by identifying the approximate locus for each read-pair and keeping only those read-pairs that originate within a window (determined by the insert size) of the ends of the target exon. To determine the approximate location, up to four non-overlapping tiles (subsequences) with the same length as the Geoseq word size are chosen from each read in a pair, then the suffix-array index of the reference genome is used to retrieve all potential position for each tile, while tiles containing stretches of mononucleotide or dinucleotide repeats are discarded. Moreover, tiles with greater than five potential mappings to the genome are discarded. For the remaining tiles, a 500-nt window around each of the mapping positions is marked as potential origin of the read-pair. If the selected window does not contain the end of the target exon, or if multiple windows share the same maximum match-number assigned to the read-pairs, the read-pairs are respectively discarded.

   The surviving read-pairs are split into their constituent reads and placed consecutively in a FASTA file in preparation for alignment.

4. *Alignment of reads to genomic fragments*. Each read is aligned against the reference sequence for the genomic region containing the exon using BLAST, to create an accurate alignment.

5. *Variant calling*. Using Sanger sequencing of the data from trios (parents and child), we found that the false positives predominated when the coverage was below 15 and/or the minor allele frequency dropped below 0.2. Based on this experience, we require variants to have a coverage greater than 15X and a mutation (mismatch) frequency higher than 0.2. Then, since the sequencing is independent of the strand, we expect approximately equal contributions from the two strands towards any variant call. As the depth of sequencing increases, the skew in contributions from the two strands becomes an indicator of a potential error.

   Clonal reads are identified and removed at this stage.

6. *Variant annotation*. We identify variants in our list that occur in dbSNP, 1000 Genomes or private collections to highlight the novel variants likely involved in rare diseases gene discovery. A potential difficulty in this process is that various databases reference different versions of the genomes. Thus, in order to be consistent, the variants are identified by their flanking sequence, 10 nt on each side. Variants and their host genes with known disease phenotype associations in dbGAP [86], OMIM (www.omim.org), PhenCode [87], SNPedia [88], and PharmGKB [89] are annotated. Missense variations are checked against the Polyphen predictions made on the UniProt protein database (www.uniprot.org) for severity of the mutation. Variants are also checked for their effect on protein function using a local installation of SNAP [90] and SIFT [91].

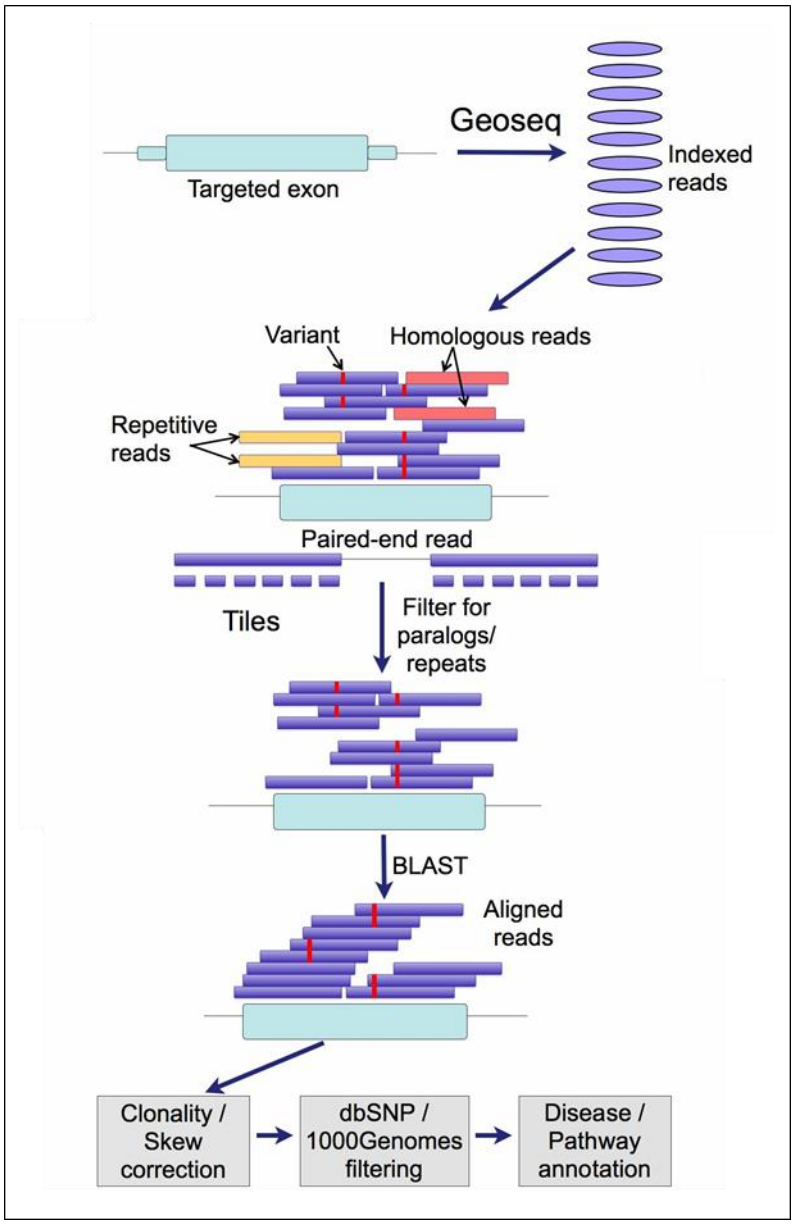


Figure 16. The MiST pipeline is based on an exon-centric approach to process NGS data. Paralogous and clonal reads are removed before calculation of coverage and calling variants.

The quality of this method has been evaluated processing by exome sequencing data from the same sample with GATK as well as our pipeline in order to compare and contrast the two approaches, after having genotyped it with a Human Exome SNP array. The GATK pipeline was the one implemented at the Yale genome center. We considered several features in GATK and MiST:

- Total number of variants. MiST called more variants than the GATK pipeline, but also had more in common with dbSNP and the SNP array (Figure 17).

- Coverage on calls. We expected lower coverage on calls, on average, from MiST, compared to GATK, due to stringent handling of clonal/paralogous reads and this was confirmed by the distribution of coverage across variants common to both platforms.
- Transitions (T_i) versus transversions (T_v). The ratio T_i/T_v is expected to be 2.0 for neutral SNPs. In coding regions this ratio has been empirically shown to be closer to 3.0 [92]. The majority of the variant calls are common to the two programs, which does not allow for major differences in these measures between them.

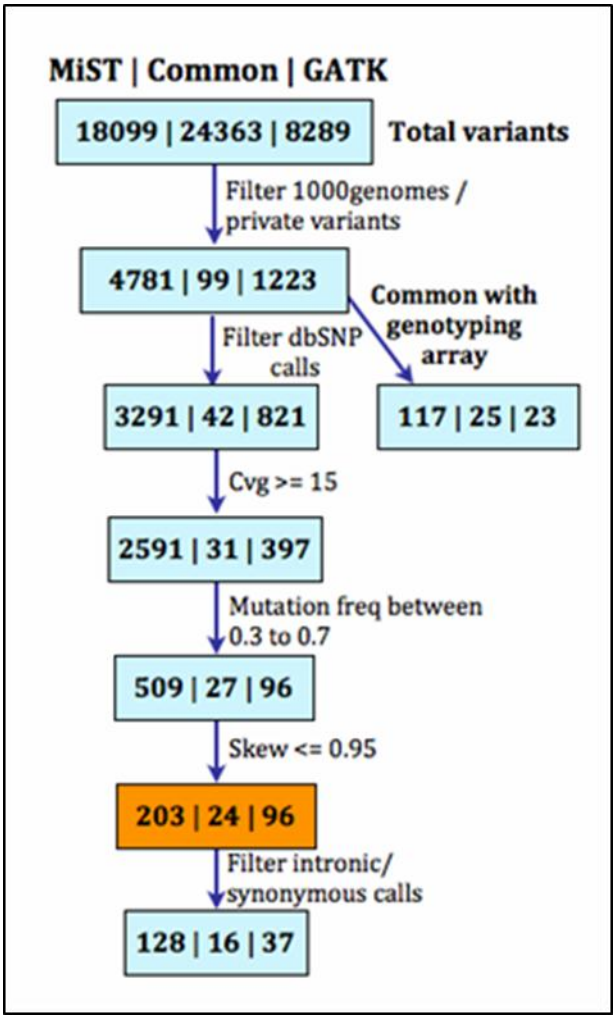


Figure 17. Comparison of MIST and GATK. Each box has three set of numbers, which refer to variant calls unique to MiST, common to both platforms and unique to GATK, respectively. After the filtering procedure, there were 37 calls private to GATK that were not called by MiST due to the exclusive use of RefSeq by MiST, which misses potentially valid exons, but avoids the noise arising from many spurious exons.

We have proven as MiST is an efficient and sensitive platform for variant detection from deep sequencing, comparable to the more-commonly used programs: it can work on data derived from a variety of platforms and techniques and works well on both single and paired-end data from whole-exome capture and sequencing. Since MiST is highly configurable, it allow handling changes in the experimental protocol, such as insert sizes, to keep pace with the throughput of sequencing technologies.

REFERENCES

Introduction

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. 1977 *Proc Natl Acad Sci USA* 74:5463-5467

2. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenguer ML, Jaryie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, VolkmerGA, Wang SH, Weiner MP, Yu P, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. 2005 *Nature* 437:376-380

3. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. Accurate multiplex polony sequencing of an evolved bacterial genome. 2005 *Science* 309:1728-1732

4. Shendure J, Ji H. Next-generation DNA sequencing. 2008 *Nat Biotechnol* 26:1135-1145

5. McKusick VA. Mendelian inheritance in man and its online version, OMIM. 2007 *Am J Hum Genet* 80:588-604

6. Harville HM, Held S, Diaz-Font A, Davis EE, Diplas BH, Lewis RA, Borochowitz ZU, Zhou W, Chaki M, MacDonald J, Kayserili H, Beales PL, Katsanis N, Otto E, Hildebrandt F. Identification of 11 novel mutations in eight BBS genes by high-resolution homozygosity mapping. 2010 *J Med Genet* 47:262-267

7. Antonarakis SE, Beckmann JS. Mendelian disorders deserve more attention. 2006 *Nature Rev Genet* 7:277-282

8. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. Exome sequencing identifies the cause of a Mendelian disorder. 2010 *Nature Genet* 42:30-35

9. Bilgüvar K, Oztürk AK, Louvi A, Kwan KY, Choi M, Tatli B, Yalnizoğlu D, Tüysüz B, Cağlayan AO, Gökben S, Kaymakçalan H, Barak T, Bakircioğlu M, Yasuno K, Ho W, Sanders S, Zhu Y, Yilmaz S, Dinçer A, Johnson MH, Bronen RA, Koçer N, Per H, Mane S, Pamir MN, Yalçinkaya C, Kumandaş S, Topçu M, Ozmen M, Sestan N, Lifton RP, State MW, Günel M. Whole-exomesequencing identifies recessive *WDR62* mutations in severe brain malformations. 2010 *Nature* 467:207-210

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

11. Teer JK, Mullikin JC. Exome sequencing: the sweet spot before whole genomes. *2010 Hum Mol Genet*. 19:R145-151

12. Mertes F, ElSharawy A, Sauer S, van Helvoort, JMLM, van der Zaag PJ, Franke A, Nilsson M, Lehrach H, Brookes AJ. Targeted enrichment of genomic DNA regions for next-generation sequencing. 2011 *Brief Funct Genomics* 10:374-86

13. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C. Solution hybrid selection with ultra-long oligonucleotidesfor massively parallel targeted sequencing. 2009 *Nat Biotechnol* 27:182-189

14. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. Target-enrichment strategies for next-generation sequencing. 2010 *Nat Methods* 7:111-118

15. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ. A large genome center's improvements to the Illumina sequencing system. 2008 *Nat Methods* 5:1005-1010

16. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features genome annotation policy. 2012 *Nucleic Acids Res* 40:D130-D135

17. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Raney BJ, Pohl A, Malladi VS, Li CH, Lee BT,

Learned K, Kirkup V, Hsu F, Heitner S, Harte RA, Haeussler M, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Dreszer TR, Diekhans M, Cline MS, Clawson H, Barber GP, Haussler D, Kent WJ. The UCSC Genome Browser database: extensions and updates 2013. 2012 *Nucleic Acids Res* Epub ahead of print

18. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovcova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernández-Suarez XM, Harrow J, Herrero J, Hubbard TJ, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadissa A, Searle SM. Ensembl. 2012 *Nucleic Acids Res* 40:D84-90

19. Clark MJ, Chen R, Lam HYK, Karczewski KJ, Chen R, Euskirchen G, Butte AJ, Snyder M. Performance comparison of exome DNA sequencing technologies. 2011 *Nat Biotechnol* 29:908-914

20. Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC, Vezenov DV. The challenges of sequencing by synthesis. 2009 *Nat Biotechnol* 27:1013-1023

21. Metzker ML. Sequencing technologies – the next generation. 2010 *Nat Rev Genet* 11:31-46

22. Ansorge WJ. Next-generation DNA sequencing techniques. 2009 *New Biotechnol* 25:195-203

23. Metzker ML. Emerging technologies in DNA sequencing. *Genome Res* 15:1767-1776

24. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou

A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. Accurate whole human genome sequencing using reversible terminator chemistry. 2008 *Nature* 456:53-59

25. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, Sidow A, Fire A, Johnson SM. A high resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. 2008 *Genome Res* 18:1051-1063

26. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage

KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. 2005 *Nature* 437:376-380

27. Mcpherson JD. Next-generation gap. 2009 *Nat Methods* 6:S2-S5

28. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. 2009 *Nucleic Acid Res* 38:1767-1771

29. Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. 2009 *Nat Methods* 6:S6-S12

30. Burrows M, Wheeler DJ. A block-sorting lossless data compression algorithm. 1994 *Technical Report Digital Equipment Corporation, Palo Alto*

31. Torri F, Dinov ID, Zamanyan A, Hobel S, Genco A, Petrosyan P, Clark AP, Liu Z, Eggert P, Pierce J, Knowles JA, Ames J, Kesselman C, Toga AW, Potkin SG, Vawter MP, Macciardi F. Next generation sequence analysis and computational genomics using graphical pipeline workflows. 2012 *Genes (Basel)* 3:545-575

32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. 2009 *Bioinformatics* 25:2078-2079

33. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K. SNP detection for massively parallel whole-genome resequencing. 2009 *Genome Res* 19:1124-1132

34. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing. 2010 *Genome Res* 20:1297-1303

35. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. 2011 *Nat Rev Genet* 12:443-451

36. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. 2011 *Nat Rev Genet* 12:363-376

37. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. The variant call format and VCFtools. 2011 *Bioinformatics* 27:2156-2158

38. Altmann A, Weber P, Bader D, Preuss M, Binder EB, Muller-Myhsok B. A beginners guide to SNP calling from High-throughput DNA-sequencing data. 2012 *Hum Genet* 131:1541-1554

39. Robinson PN, Krawitz P, Mundlos S. Strategies for exome and genome sequence data analysis in disease-gene discovery projects. 2011 *Clin Genet* 80:127-32

40. EURORDIS: Rare Diseases Europe. http://www.eurordis.org/about-rare-diseases

41. Won HH, Kim HJ, Kim JW. Cataloging coding sequence variations in human genome databases. 2008 *PLoS One* 3:e3575.

42. Day IN. dbSNP in the detail and copy number complexities. 2010 *Hum Mutat* 31:2-4

43. Ng SB, Nickerson DA, Bamshad MJ, Shendure J. Massively parallel sequencing and rare disease. 2010 *Hum Mol Genet* 19:R119-R124

44. Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, Shendure J, Nickerson DA. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. 2010 *Nat Methods* 7:250-251

45. Kumar PH, Ng S. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. 2009 *Nat Protoc* 4:1073-1081

46. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. 2010 *Nat Methods* 7:248-249

47. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. 2009 *Genome Res* 19:1553-1561

48. Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. 2000 *Genetics* 156:297-304

49. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. 2010 *Bioinformatics* 26:589-595

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

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

52. Wigginton JE and Abecasis GR PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data 2005 *Bioinformatics* 21:3445-3447

53. Abecasis GR, Cherny SS, Cookson WO and Cardon LR. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. 2002 *Nat Genet* 30:97-101

54. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. 2007 *Am J Hum Genet.* 81:559-575

55. The 1000 Genome Project Consortium. A map of human genome variation from population-scale sequencing. 2010 *Nature* 467:1061-1073

56. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe Db, Nusbaum C, Gnirke A. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. 2011 *Genome Biol* 12:R18.

57. Ledergerber C, Dessimoz. Base-calling for next-generation sequencing platforms. 2011 *Brief Bioinform* 12:489-497

58. Meacham F, Boffelli D, Dhahbi J, Martin DIK, Singer M, Pachter L. Identification and correction of systematic error in high-throughput sequence data. 2011 *BMC Bioinformatics* 12:451.

59. Snyder M, Du J, Gerstain M. Personal genome sequencing: current approaches and challenges. 2010 *Genes Dev* 24:423-431.

60. Zhang Z, Gerstein M. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. 2003 *Nucleic Acids Res* 31:5338-5348

61. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. 2011 *Nat Genet* 43:491-498

62. Taylor MS, Ponting CP, Copley RR. Occurrenceand consenquences of coding sequence insertions and deletions in mammalian genomes. 2004 *Genome Res* 14:555-566

63. Tartaglia M, Zampino G, Gelb BD. Noonan syndrome: clinical aspects and molecular pathogenesis. 2010 *Mol Syndromol* 1:2-26

64. Kuhlenbaumer G, Hullmann J, Appenzeller S. Novel genomic techniques open new avenues in the analysis of monogenic disorders. 2011 *Hum Mutat* 32:144-151

65. Yang L, Zhao J, Lu W, Li Y, Du X, Ning T, Lu G, Ke Y. KIAA0649, a 1A6/DRIM-interacting protein with the oncogenic potential. 2005 *BBRC* 334:884-890

66. Veltman JA, Brunner HG. De novo mutations in human genetic disease. 2012 *Nature Rev* 13:565-575

67. Fang J, Dagenais SL, Erickson RP, Arlt MF, Glynn MW, Gorski JL, Seaver LH, Glover TW. Mutations in *FOXC2* (*MFH-1*), a forkhead family transcription factor, are responsible for the hereditary Lymphedema-Distichiasis syndrome. 2000 *AJHG* 67:1382-1388

68. Bahuau M, Houdayer C, Tredano M, Soupre V, Couderc R, Vazquez MP. *FOXC2* truncating mutation in distichiasis, lymphedema, and cleft palate. 2002 *Clin Genet* 62:470-473

69. Shinawi M, Liu P, Kang SH, Shen J, Belmont JW, Scott DA, Probst FJ, Craigen WJ, Graham BH, Pursley A, Clark G, Lee J, Proud M, Stocco A, Rodriguez DL, Kozel BA, Sparagana S, Roeder ER, McGrew SG, Kurczynski TW, Allison LJ, Amato S, Savage S, Patel A, Stankiewicz P, Beaudet AL, Cheung SW, Lupski JR. Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy and abnormal head size. 2010 *J Med Genet* 47:332-341

70. Schaaf CP, Goin-Kochel RP, Nowell KP, Hunter JV, Aleck KA, Cox S, Patel A, Bacino CA, Shinawi M. Expanding the clinical spectrum of the 16p11.2 chromosomal rearrangements: three patients with syringomyelia. 2011 *Eur J Hum Genet* 19:152-6

71. Teebi AS. New autosomal dominant syndrome resembling craniofrontonasal dysplasia. 1987 *Am J Med Genet* 28:581-591

72. Koenig R. Teebi hypertelorism syndrome. 2003 *Clin dysmorphol* 12:187-189

73. Liu YL, Wang J, Zheng JH, Bai K, Liu ZM, Wang XZ, Liu X, Yang YQ. Involvement of a novel *GATA4* mutation in atrial septal defects. 2011 *Int J Mol Med* 28:17-23

74. Singleton DR, Wu TT, Castle JD. Three mammalian SCAMPs (secretory carrier membrane proteins) are highly related products of distinct genes having similar subcellular distributions. 1997 *J. Cell Sci*. 110: 2099-2107

75. Conley ME, Dobbs AK, Farmer DM, Kilic S, Paris K, Grigoriadou S, Coustan-Smith E, Howard V, Campana D. Primary B cell immunodeficiencies: comparisons and contrasts. 2009 *Annu Rev Immun* 27:199-227

76. Notarangelo L, Casanova JL, Conley ME, Chapel H, Fischer A, Puck J, Roifman C, Seger R, Geha RS. Primary immunodeficiency diseases: an update from the International Union of Immunological Societies Primary Immunodeficiency Diseases Classification Committee Meeting in Budapest, 2005. 2006 *J Allergy Clin Immunol* 117:883-96

77. van de Ven AA, Compeer EB, van Montfrans JM, Boes M. B-cell defects in common variable immunodeficiency: BCR signaling, protein clustering and hardwired gene mutations. 2011 *Crit Rev Immunol* 31:85-98

78. Basso K, Dalla-Favera R. BCL6: master regulator of the germinal center reaction and key oncogene in B cell lymphomagenesis. 2010 *Adv Immunol* 105:193-210

79. Lougaris V, Ferrari S, Plebani A. Ig beta deficiency in humans. 2008 *Curr Opin Allergy Clin Immunol* 8:515-519

80. Walsh T, Shahin H, Elkan-Miller T, Lee MK, Thornton AM, Roeb W, Abu Rayyan A, Loulus S, Avraham KB, King MC, Kanaan M. Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of Nonsyndromic Hearing Loss DFNB82. 2010 *AJHG* 87:90-94

81. Lopez-Herrera G, Tampella G, Pan-Hammarstrom Q, HerHolz P, Trujillo-Vargas CM, Phadwal K, Simon AK, Moutschen M, Etzioni A, Mory A, Srugo I, Melamed D, Hultenby K, Liu C, Baronio M, Vitali M, Philippet P, Dideberg V, Aghamohammadi A, Rezaei N, Enright V, Du L, Salzer U, Eibel H, Pfeifer D, Veelken H, Strauss H, Lougaris V, Plebani A, Gertz EM, Schaffer AA, Hammarstrom, Grimbacher B. Deletrious mutations in *LRBA* are associated with a syndrome of Immune Deficiency and Autoimmunity. 2012 *AJHG* 90:986-1001

82. **Di Pierro V**, Subramanian SL, Shah H, Jayaprakash A, Weisberger I, George A, Gelb BD, Sachidanandam R. MiST: a new approach to variant-detection in deep sequencing datasets. *Submitted*

83. Pattnaik S,Vaidyanathan DG, Pooja S, Deepak S, Panda B. Customisation of the exome data analysis pipeline using a combinatorial approach. 2012 *PLoS ONE* 7:e30080

84. Gurtowski J, Cancio A, Shah H, Levovitz C, George A, Sachidanandam R. Geoseq: a tool for dissecting deep-sequencing datasets. 2010 *BMC Bioinformatics* 11:506

85. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. 1990 *J Mol Biol* 215:403-410

86. Mailman MD, Feolo M, Jin Y, Kimura M, Tryja K, Bagoutdinov R, Hao L, Kiang A, Pashall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D,

Yashenko E, Graeff A, Ostell J, Sherry ST. The NCBI dbGaP database of genotypes and phenotypes. 2007 *Nat Genet* 39:1181-1186

87. Giardine B, Riemer C, Hefferon T, Thomas D, Hsu F, Zielenki J, Sang Y, Elnitski L, Cutting G, Trumbower H, Kern A, Kuhn R, Patrinos GP, Hughes J, Higgs D, Chui D, Scriver C, Phommarinh M, Patnaik SK, Blumenfeld O, Gottlieb B, Vihinen M, Vliaho J, Kent J, Miller W, Hardison RC. PhenCode: connecting ENCODE data with mutations and phenotype. 2007 *Hum Mutat* 28:554-562

88. Cariaso M, Lennon G. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. 2012 *Nucleic Acids Res* 40:D1308-1312

89. McDonagh EM, Whirl-Carrillo M, Garten Y, Altman RB, Klein TE. From pharmacogenomics knowledge acquisition to clinical applications: the pharmGKB as a clinical pharmacogenomics biomarker resource. 2011 *Biomark Med* 5:795-806

90. Bromberg Y, Yachdav G, Rost B. SNAP predicts effect of mutations on protein function. 2008 *Bioinformatics* 24:2397-2398

91. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. 2003 *Nucleic Acids Res* 31:3812-3814

92. Y. Freudenberg-Hua, J. Freudenberg, N. Kluck, S. Cichon, P. Propping, M. M. Nthen, Single nucleotide variation analysis in 65 candidate genes for CNS disorders in a representative sample of the european population. 2003 *Genome Res* 13:2271–2276