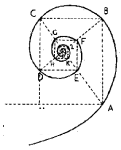
UNIVERSITÀ DEGLI STUDI DI MILANO

Dottorato in Medicina Molecolare e Traslazionale



Ciclo XXVI
Anno Accademico 2012/2013

Settore BIO/10

Next-Generation Sequencing Approach for Identification of Candidate Genes in Arrhythmogenic Diseases

Dottorando: Alessandro PIETRELLI
Matricola N° R09193

TUTORE: Prof. Cristina BATTAGLIA

CO-TUTORE: Dott Gianluca DE BELLIS

DIRETTORE DEL DOTTORATO: Prof. Mario CLERICI

ABSTRACT

Recent advances in genome sequencing technologies provided unexpected opportunities to characterize individual genomic landscape and identify mutations relevant for diagnosis and therapy in clinics. Specifically, whole-exome sequencing for complex disease (such as tumor/normal matched sample) and target resequencing for Mendelian disease, using next-generation sequencing (NGS) technologies, are gaining popularity in the human genetics community due to the moderate costs and the huge quantity of information provided by each experiment. However, NGS data analysis still remains the crucial bottleneck in this approach because of the great amount of data containing millions of potential disease-causing variants and difficulties involving the integration of different sources.

Here, we describe the application of a bioinformatics analysis pipeline for NGS data in two case studies about rare cardiac diseases. The case 1 is focused on the target sequencing of 158 candidate genes in 91 patients affected by Brugada Syndrome (BrS). To date the clinical phenotype is associated with mutations in the SCN5A gene but explain only the 30% of the BrS cases. Therefore we selected a panel of genes previously associated to cases of arrhythmogenic disorders in literature and we analysed them in a cohort of BrS patients, which were negative for known *SCN5A* mutations. We found 98 novel genetic variations and 60 clinical rs belonging to 70 genes. In particular we found 13 genes significantly mutated in our cohort compared to healthy controls in 1000 Genomes data that were not previously associated to BrS phenotype.

In case study 2 we performed a whole-exome sequencing experiment of trio family where the child is affected by a severe cardiac disease with unclear diagnosis. We developed a specific bioinformatics pipeline to filter out the germline mutations. We found six genes with novel deleterious mutations that were homozygous in the affected child, and heterozygous in both parents. Among the six mutated genes the *TRDN* and *UNC45A* genes

were already associated to cardiac dysfunctions in literature. Functional studies will be performed to evaluate the involvement of the mutated genes in the disease onset.

In conclusion, we developed an automatic and versatile pipeline to analyse NGS data coming from whole-exome sequencing and target sequencing strategies. In addition, we integrated in the pipeline several public variation databases to evaluate and interpret the candidate mutations. The mutations found were validated by Sanger sequencing to evaluate the strength of the pipeline filters.

SOMMARIO

I recenti progressi nelle tecnologie di sequenziamento del genoma umano hanno offerto opportunità inaspettate per la caratterizzazione del patrimonio genetico di un individuo e per l'identificazione di mutazioni rilevanti per la diagnosi e la terapia in clinica. In particolare, il sequenziamento dell'esoma per malattie complesse (come il confronto a coppia campione tumorale / campione normale) e il risequenziamento di regioni *target* applicato alle malattie genetiche di tipo Mendeliano, mediante sequenziamento di ultima generazione di tecnologie (NGS), stanno diventando molto popolari nella comunità di genetica umana grazie ai costi moderati e alla quantità di informazione fornita in ogni esperimento. Tuttavia, l'analisi dei dati NGS rimane ancora il collo di bottiglia in questo approccio a causa della grande quantità di potenziali varianti causative e della difficoltà di integrare i dati con fonti diverse.

Qui, descriviamo l'applicazione di workflow bioinformatico per l'analisi di dati NGS in due casi clinici riguardanti malattie cardiache rare. Il primo caso è focalizzato sul risequenziamento di 158 geni candidati in 91 pazienti affetti da Sindrome di Brugada (BrS). Ad oggi il fenotipo clinico è associato a mutazioni del gene *SCN5A* ma spiega solo il 30% dei casi BrS. Pertanto, abbiamo selezionato un gruppo di geni precedentemente associato a casi di disturbi aritmogenici in letteratura e li abbiamo analizzati in una coorte di pazienti affetti da BrS che non presentavano mutazioni conosciute in *SCN5A*. In totale, abbiamo trovato 98 varianti genetiche (SNP/INDEL) e 60 rs clinici appartenenti a 70 geni del nostro pannello. In particolare, abbiamo trovato 13 geni mutati in modo significativo nella nostra coorte rispetto ai controlli sani ricavati dal progetto 1000 Genomes. Questi geni non sono mai stati in precedenza associati ad un fenotipo di tipo BrS.

Nel secondo studio abbiamo condotto un esperimento di sequenziamento sull'intero esoma di un trio (genitori e figlio) in cui il bambino è affetto da una grave malattia cardiaca con diagnosi non chiara. Abbiamo sviluppato

uno specifico workflow di analisi bioinformatica per filtrare le mutazioni germinali. Abbiamo trovato sei geni con nuove mutazioni deleterie presenti in stato di omozigosi nel bambino affetto ed in eterozisi in entrambi i genitori. Tra i sei geni mutati, i geni *TRDN* ed *UNC45A* erano già associati a disfunzioni cardiache in letteratura. Gli studi funzionali saranno eseguiti per valutare il coinvolgimento dei geni mutati nell'insorgenza della malattia.
In conclusione, abbiamo sviluppato una *pipeline* automatica e versatile per analizzare i dati NGS provenienti da esperimenti di sequenziamento dell'intero esoma e da strategie di sequenziamento di tipo *target*. Inoltre, abbiamo integrato i dati mutazionali con diverse banche dati pubbliche per valutare e interpretare le mutazioni candidate. Le mutazioni trovate sono state inoltre convalidate dalla metodica di sequenziamento Sanger per valutare la forza dei filtri utilizzati nella *pipeline* di analisi.

TABLE OF CONTENTS

FIGURE INDEX

1. INTRODUCTION

1.1. Next-Generation sequencing technology

1.1.1. The advent of next-generation sequencing technology

Over the past five years, the massively parallel DNA sequencing technology has become widely available thanks to the previous innovative development in sequencing strategies [1, 2]. The advent of next-generation sequencing (NGS) has rebuilt the meaning of DNA sequencing by processing millions of DNA fragment in parallel resulting a very low cost per base and a throughput on the gigabase (Gb) scale [3] (*Figure 1*). The enormous volume of data cheaply produced by (NGS) technology has been the drawing power for shifting from automated Sanger sequencing and changing the way of thinking the genome research and clinical genetics [4].

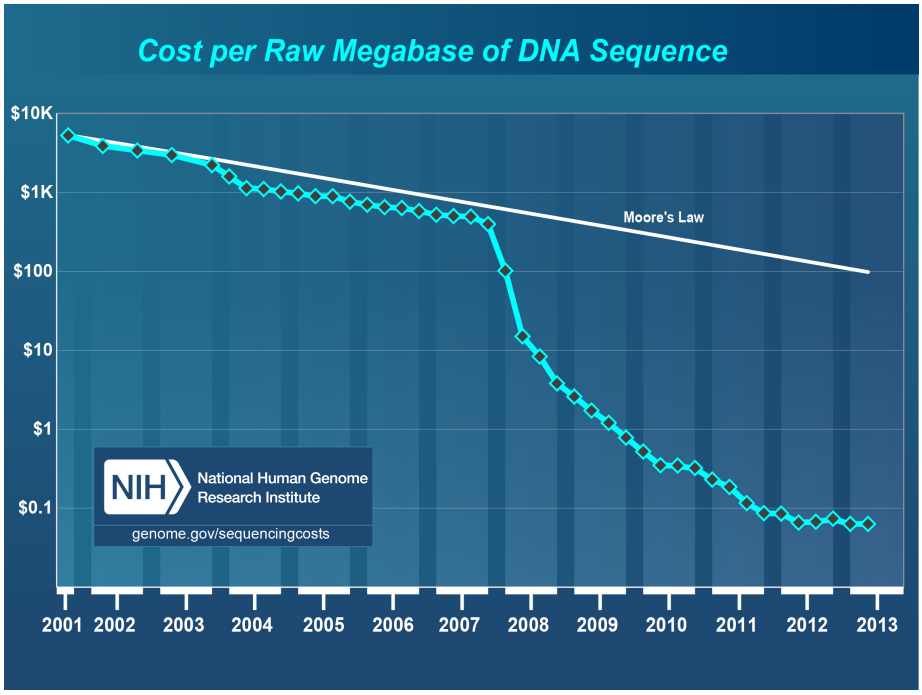


Figure 1 ***Sequencing costs trend****. The rapid development of the NGS technology marked in 2008, the free-fall of the sequencing costs. (http://www.genome.gov/sequencingcosts/)*

The powerful and flexible nature of the NGS technology led to the development of a broad range of applications allowing researchers to ask biological questions that were unimaginable just a few years ago. The broadest application of NGS may be the resequencing of human genomes to enhance our understanding of how genetic differences affect health and disease.

Despite a variety of NGS features allowed several companies to develop high-throughput sequencer machines, the general NGS workflow is well-defined and is divided into three steps classified by template preparation, sequencing and imaging [4]. The specific combination of protocols is unique for each technology and determines the type of data produced from each platform.

Three methods were strongly developed and standardized for NGS called (1) clonal amplification, (2) sequencing by synthesis and (3) real-time sequencing [4] (*Figure 2*).

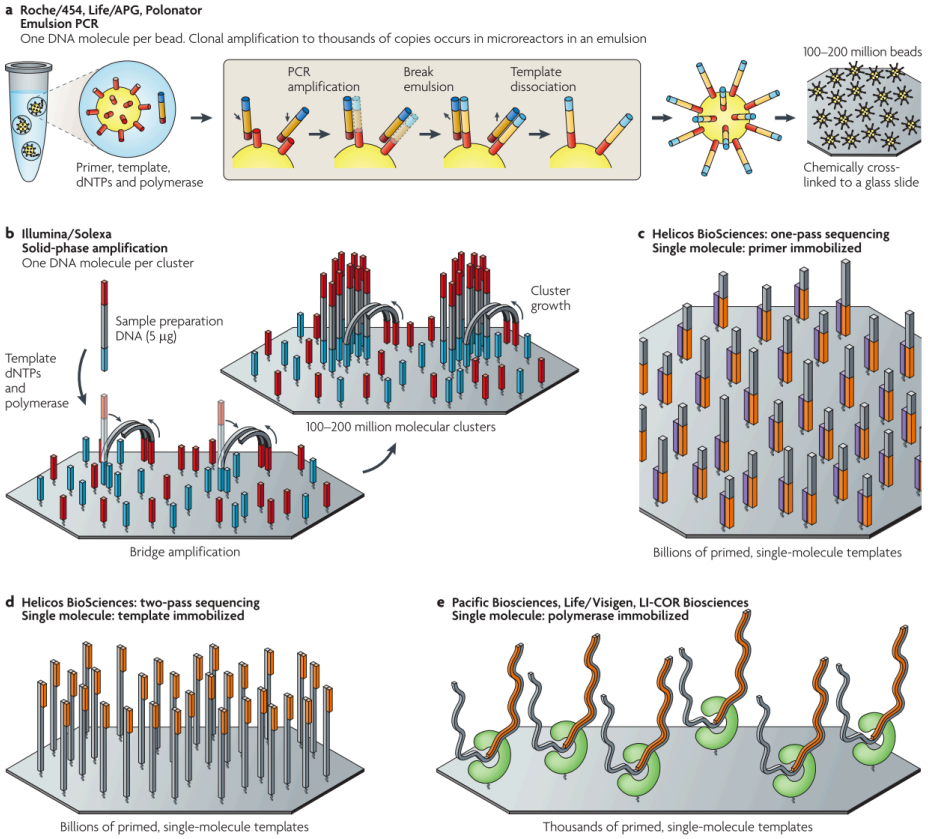


Figure 2 ***NGS sequencing strategies.*** *Representation of available template immobilization strategies adopted by sequencing companies.* [4]

The first two methods involve template amplification. In detail the clonal amplification is based on emulsion PCR (emPCR) where bead–DNA complexes are encapsulated into single aqueous droplets in oil emulsion and PCR amplification creates thousands of copies of the same template within each droplet. Afterward each amplified bead are immobilized on a solid plate (PicoTiterPlate) and ready for sequencing and imaging. This sequencing method is on the basis of the Roche/454 platforms (*Figure 2a*) The sequencing by synthesis does not require a PCR step but the template amplification is performed directly on a solid support. Illumina and Helicos BioSciences platforms use this method with some modifications between each other (*Figure 2b,c,d*). The most recent method and the one with no

template generation is the real-time sequencing which is capable to directly sequence DNA fragment by involving a DNA polymerase immobilized on a solid support and the continuous incorporation of dye-labelled nucleotides during DNA synthesis (*Figure 2e*).

1.1.2. Illumina sequencing technology workflow

The whole NGS workflow consists on a three-step process divided into template preparation (or sample preparation), sequencing and imaging. The Illumina sequencing strategy is based on the concept of sequencing by synthesis to produce millions of 'short-read' (from 36 to 86 nucleotides of length) simultaneously (*Figure 3*) [5]. The first step is the creation of a DNA library by and adding universal adapters by ligation to sample DNA fragments. Afterward, the process involves using a microfluidic cluster station to add these fragments to the surface of a glass flowcell thanks to the hybridization with complementary oligos to the surface. Each flowcell is divided into eight lanes where the interior surfaces have covalently attached oligos complementary to the specific adapters that are ligated onto the library fragments. The hybridization of the library fragment on the flowcell is followed by a subsequent incubation, called cluster generation, that amplifies the fragments in a discrete area or 'cluster' thanks to a PCR reaction called 'bridge PCR' [6] directly on the flowcell surface. The flowcell is placed within the sequencer where each cluster is supplied with polymerase and four differentially labeled fluorescent nucleotides that have their 3'-OH chemically inactivated to ensure that only a single base is incorporated per cycle [5]. Every cycle is composed by (1) base incorporation, (2) imaging step to discriminate the incorporated nucleotide at each cluster by laser light excitation and (3) a chemical step that removes the fluorescent group and release the 3' end for the next base incorporation cycle (*Figure 3*).

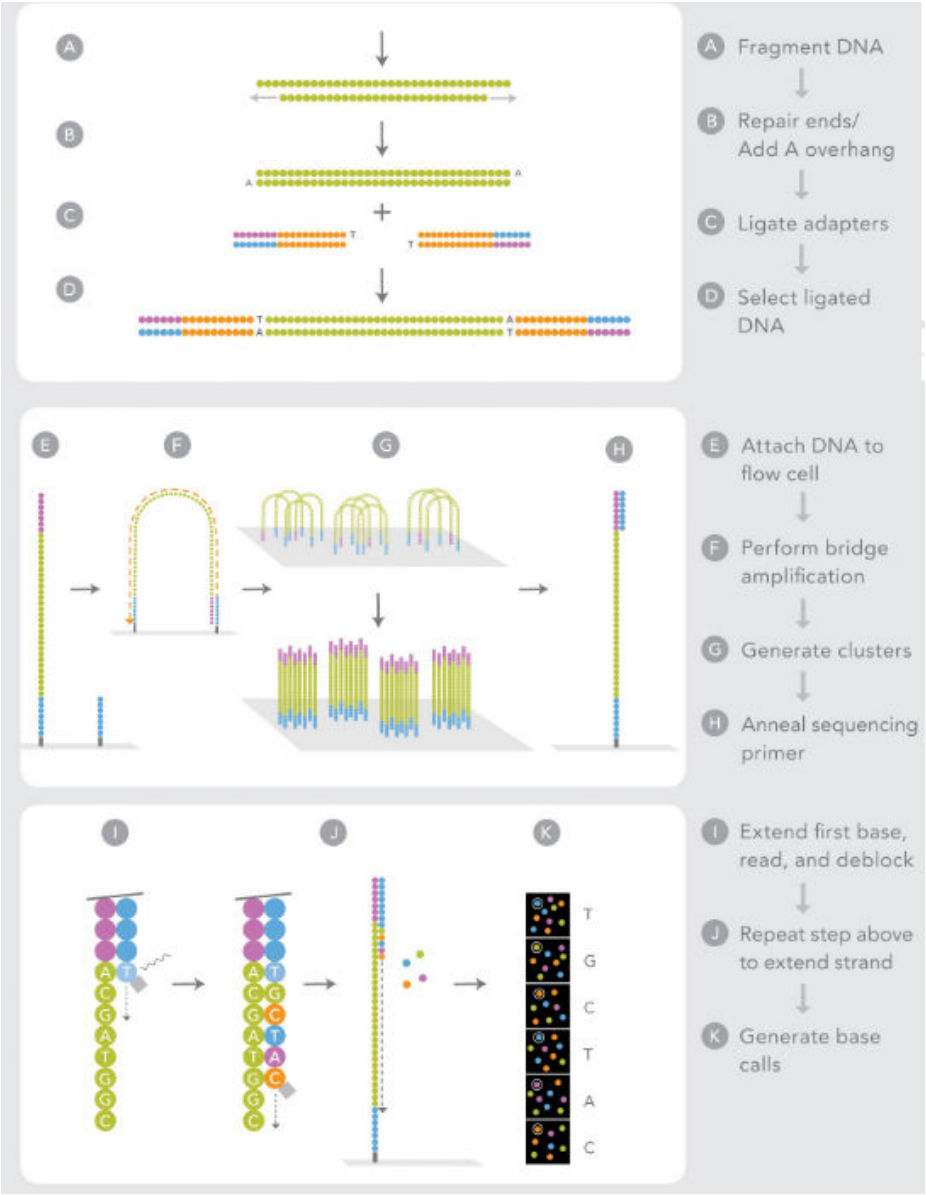


Figure 3 ***Illumina Genome Anlayzer sequencing workflow*** *(modified from Illumina website)*

The sequencing run turnaround varies depending on the number of cycles, which corresponds to the read length, between 3 and 10 days (36 to 85 base pair read length). The images, corresponding to the sequence of each cluster, are translated into text file as FastQ format [7] containing the reads

produced in the sequencing run. Every run produces from 50 to 60 millions of reads.

1.1.3. Capture technology

Despite the sequencing costs are constantly falling, the whole-genome sequencing approach is not yet feasible for large number of samples, especially in clinical and biomedical research, because the cost and time taken are still too great [3]. Consequently, substantial efforts have been made to develop several 'target-enrichment' methods in which genomic regions of interest are selectively captured before sequencing. Before the advent of NGS there were methods for target-enrichment based on PCR and molecular inversion probes (MIP). Both methods reveal great sensitivity in enrichment, however the difficulty of use grows with the number of genes in the target to enrich [3].

A widely used method in NGS for target-enrichment is the DNA capture based on hybridization of the genomic region of interest with complementary oligonucleotides (probes) properly designed (*Figure 4*).

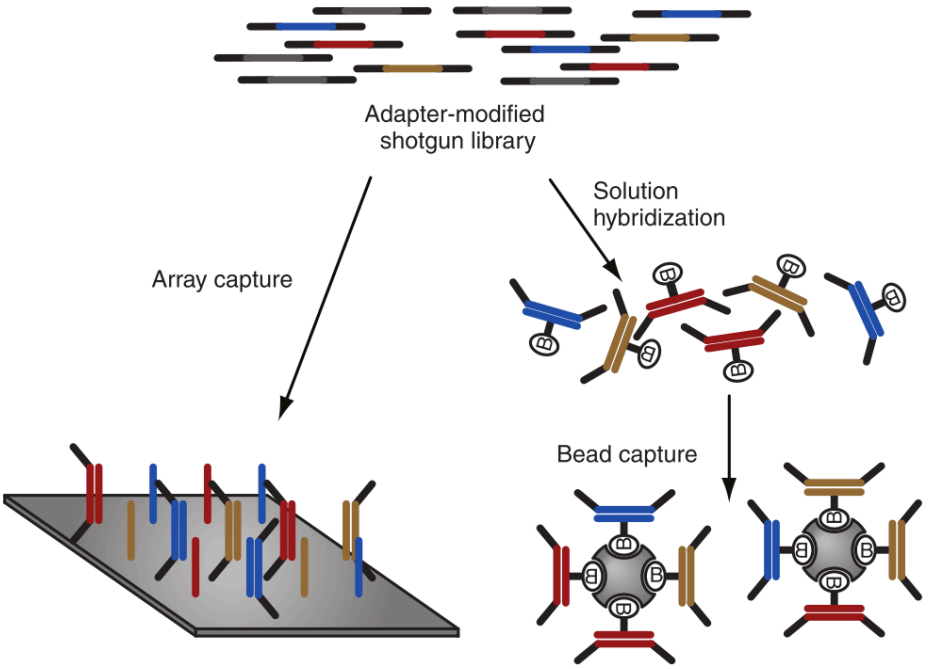


Figure 4 ***Capture enrichment technology****. Two main strategies to select specific DNA target sequence: array-based capture (left) and solution-based capture (right). Modified from* [3]

The capture selection has overcome the other existing method for its ease to use and for the scalability power [3]. Firstly this methodology was applied to a solid array with immobilized probe (on-array capture) [8, 9], but working with microarray support requires expensive hardware. Moreover the starting DNA amount required for library preparation is relatively large (around 10–15 μg) which is not easy to collect in particular experiment such as sorting cells [3]. To overcome some of the disadvantages present in the on-array capture it has been also developed a solution-based target-enrichment strategy. This methodology has the same principle of hybridization but the probes are in solution and are in excess number allowing much less DNA amount for library preparation compared to other methods [10].

This technology allows capturing the entire 'exome' (all the coding exons in the human genome) to a particular genomic region (from 20 kilobases to 5 megabases) with the 'target' design.

1.1.4. Agilent solution-based target enrichment

The Agilent target enrichment is based on the hybridization of complementary probes with genomic DNA region. Basically the workflow consists in three steps, which consider hybridization, bead capture and sequencing (*Figure 5*)

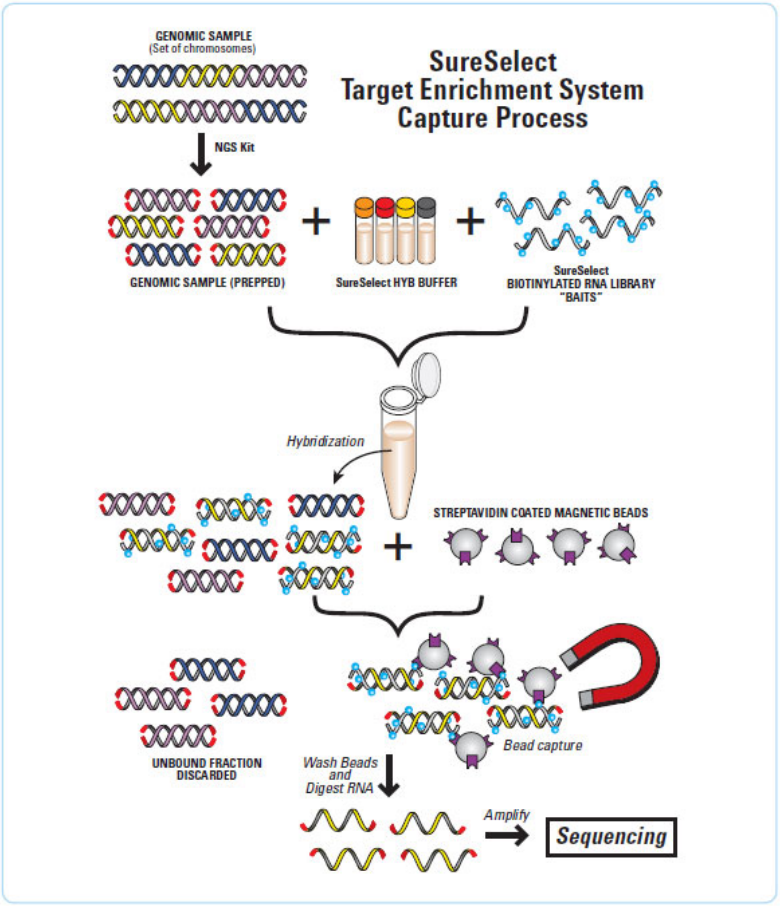


Figure 5 ***Agilent solution-based capture workflow.*** *Briefly the genomic DNA sample is prepared as NGS library and hybridized with the specific Agilent*

biotinylated probes. The targeted regions are selected using streptavidin coated magnetic beads which link only the hybridized probes with the complementary DNA sample. The enriched library is amplified and ready to sequence. (http://goo.gl/nLLyby)

Briefly the genomic DNA is shared and the adaptors ligation is performed to obtain a DNA library. The hybridization step is performed between the library and probes (120 nucleotide length) complementary to the genomic region to capture. Magnetic beads will select only those DNA strands that have been recognized by a probe and those with no probe recognition will be wash away. The enriched library is built and ready for sequencing.

Agilent developed several standard kits for the exome capture for a few organisms (human, mouse). In addition to the standard kits, it is also possible to create a custom target region to capture thanks to the Agilent SureSelect site (https://earray.chem.agilent.com/earray/). Based on genomic region of interest the site computes the complementary probes to cover the target range. The computed probes will be synthetize and the custom hybridization solution will be ready for use.

1.2. Bioinformatics analysis

As NGS has become a popular, whole-genome sequencing and whole-exome sequencing have proven to be valuable methods for the discovery of the genetic causes of rare and complex diseases [11]. Nevertheless, the great amount of data and the number of simple nucleotide variations (SNVs), including single-nucleotide polymorphisms (SNPs) and small insertions and deletions (INDELs) divert the problem to the computational analysis and data management which are the real bottleneck of the entire NGS workflow [12].

On average, whole-exome sequencing identifies from 12,000 to 20,000 variants in coding regions [13, 14], of which ≈90% are found in publicly

available databases [15]. To get meaningful biological results, each step of the analysis workflow needs to be carefully considered, and specific tools need to be considerate based on different experimental setups [16].
The bioinformatics analysis process for NGS data is divided in different steps involving alignment, variation discovery and annotation (**primary analysis**) as well as the use of tools for gene prioritization and mutation pathogenicity prediction (**secondary analysis**). The final aim is to select potentially driver mutations related to a given disease.

1.2.1. Computational tools and analysis pipelines

The bioinformatics analysis of NGS data is a complex process divided into different steps, which consist of a big collection of programs and databases and involve the management of huge amount of heterogeneous data [16] (*Figure 6*). The **primary analysis** of a typical whole-exome/target sequencing project is divided into three main steps: (1) quality filtering (2) read alignment, (3) variant identification or SNV call and (4) variant annotation. Generally those steps are recursive and, in the last year, several groups developed its own pipeline in order to standardize the results and speed-up the whole process by using different combination of tools for each single analysis step [17–19].

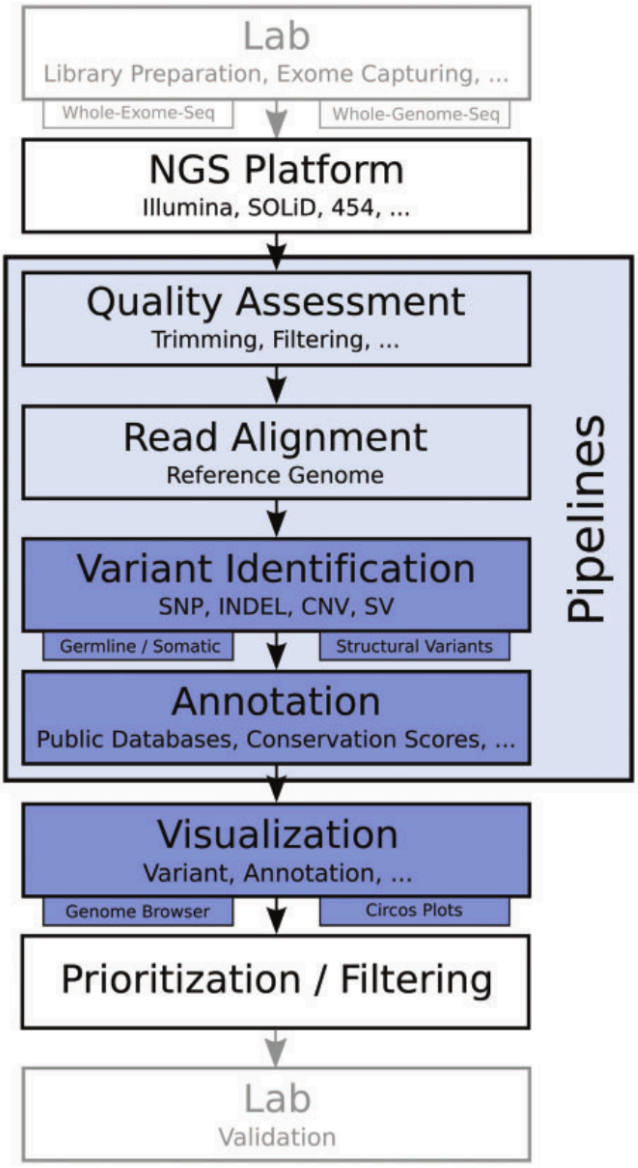


Figure 6 ***Schematic data analysis workflow for NGS experiment.*** *After library preparation and sequencing the raw reads are ready for bioinformatics analysis. The reads pass through different quality filters before the read alignment and the successively variant identification. The mutations found are annotated to infer the biological relevance and prioritized to select a candidate set of deleterious variations.* [16]

1.2.1.1. Quality filtering

The very first step, after completing the sequencing run, is the evaluation of the read quality produced by the high-throughput sequencer. Generally all the sequencing platforms are prone to base-calling error, contamination or sequence artefacts [16]. Quality check becomes a crucial step to evaluate the raw reads and try to eliminate or correct the errors introduced by the sequencer itself.

Several tools have been developed to perform different stages of quality assessment and to produce valuable summary reports. For example FastQC (http://goo.gl/vuP5aj) is multi-platform stand-alone software, which have the possibility to visualize the raw reads statistics with intuitive plots. Other similar tools to FastQC, specific for short-read coming from Illumina technology, are htSeqTools [20] and SolexaQA [21] that are also able to process the data and discard reads with low quality value or trim them where the base-call quality is poor.

1.2.1.2. Read alignment

After the quality step, the reads are ready for to be mapped against the reference genome [22]. In case of human sample, the reference genome assembly is available in two versions: the one curated by the University of Santa Cruz (UCSC) which is hosting the ENCODE data [23] and the one available from the Genome Reference Consortium (GRC), which focuses on creating reference assemblies (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc).

This specific task is the hardest step in light of computational resources, considering the whole data analysis process. Therefore the alignment of millions of reads back to a reference genome required an improvement and a development of new specific tools which are able to rapidly solve this issue [24]. In the last five years a plethora of aligner have been developed [25] which are mainly based on the construction of data structures named 'index' in order to compress the genome reference or the raw reads and

speed-up the alignment. The two main algorithm categories used for the majority of the aligner are (1) the hash table and (2) the suffix tree [24].

The hash table is a relatively old concept that was originally applied on the BLAST [26] software and successively adopted by few fast aligner such as MAQ [27], SOAP [28] and SeqMap [29] and following a seed-and-extension paradigm. All those software are based on the 'seed searching', as in BLAST program. Once the reads and the reference are indexed in a hash table, a portion of each read (seed) will be aligned with the reference sequence. The seed searching allows also mismatch, which improve the speed, and once the seed match between the read and the reference genome is found, it extends the partial match in order to places the read in the correct genomic location [24].

The other algorithm, based on the suffix-tree, also follows the concept of seed-and-extension paradigm but the reads and reference indexes are built in a suffix-tree data structure (*Figure 7*).

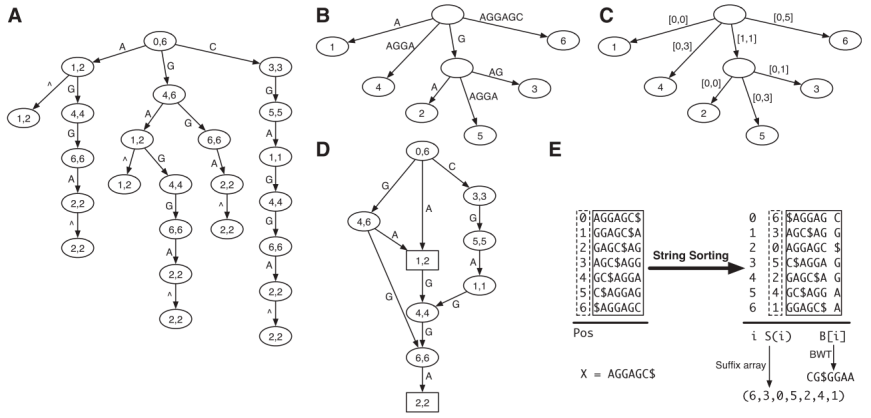


Figure 7 ***Suffix-tree algorithm representations****. The different bubble trees (A,B,C,D) represent the organization of prefix tree of the AGGAGC example string. The table E shows the construction for the suffix array and Burrows-Wheeler transform of AGGAGC string.* [24]

The suffix-tree structure allows a fast string matching by handling the 'suffix' (strings representing the entire reads/reference variability) as index data [24]. Link between the index and the suffix data is obtained by the

Burrows-Wheeler Transform (BWT) [30]. This data structure form the basis of recently developed softwares such as BOWTIE [31], BWA [32] and SOAP2 [33].

All the software cited before perform a fast and accurate mapping of the reads against the reference genome by producing a file in SAM/BAM [34] format. This format has been introduced by 1000Genomes Project [35] and rapidly became the standard format for storing the alignment data. To manipulate those particular files the 1000Genomes Project also developed a tool suite called *samtools* [34] that allows management and filtering in a very efficient way.

The SAM/BAM alignment data are the input files for the next step in the analysis process that will be the core task of the primary analysis that is the variant identification.

1.2.1.3. Variant call

After read alignment and once the SAM/BAM file is produced the next step is the identification of those position that differs from the reference sequence and could be recognized as variations (Figure *8*).

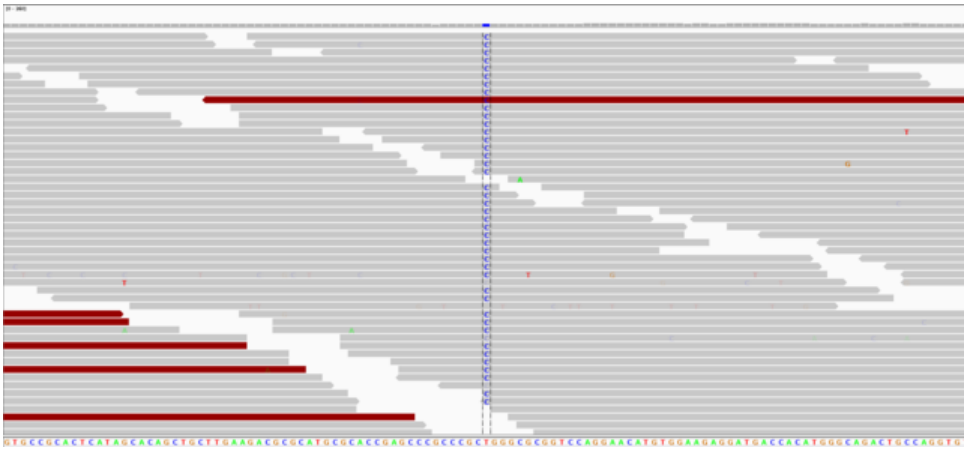


Figure 8 ***Visualization of single nucleotide variant found in read alignment****.*

The variation identification, or SNV calling, is performed by specific tools that can be divided into two categories: (1) simple allele count and (2) probabilistic method [22].

The simple allele count method has been the first algorithm implemented for variant calling [22]. It is simply based on the count of the alleles in each position of the genome and the genotype is inferred by the number of non-reference allele found. If the proportion of non-reference allele is between 20% and 80% will be called a heterozygous genotype, otherwise a homozygous genotype would be called [22]. This method is the simplest way to call variations and infer a genotype but is very dependent on the heuristic filters imposed by the users such as minimum base quality score (Phred score), read depth threshold and read mapping quality [22, 36]. This method has been embraced by several commercial software such as Roche GsMapper, CLC Genomics workbench and DNASTAR LaserGene software [22].

Most recently, newer software adopts the probabilistic method as engine of their tool. This method used the base quality score to calculate a posterior probability for each genotype call [27, 33]. This calculation dramatically improve the accuracy of genotype caller by giving a quality score for the genotype status of each variations [22]. Several software are based on this probabilistic method [33, 34] but one of the latest and most sophisticated software is GATK [37]. Its characteristic is the improvement of the input BAM file with a recalibration tool (BaseRecalibrator walker) which recalculate the base quality score of each aligned read by using a model on known variation site [37]. This particular procedure slightly improves the overall quality of calls and is one of the standard step for data preparation in the guidelines of the GATK software (http://www.broadinstitute.org/gatk/guide/best-practices).

As for the alignment, the variant calls are stored into a text file, which follows the standard of 1000Genomes data called VCF (variant call format)

(http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41). Basically it is a column-based file in which nine columns contain the specific information for each variant call including chromosomal location, reference and alternative allele, variation quality and the depth of the read supporting the alternative allele compared to the number of read supporting the reference allele (*Figure 9*).

```
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4
```

Figure 9 ***Example of VCF file.***

Thanks to the information given by this file and a suite of tools called *vcftools* [38] the end-user can easily filter-out those record with low variants quality score or variants that does not satisfy the user heuristic filters (such as read depth) indeed producing a high-quality VCF containing the variants of each sample sequenced.

The next step to evaluate the significance and the possible function of each variation called is the annotation step.

1.2.1.4. Annotation

Once the VCF file containing the quality-passed call has been produced, the annotation step gives, to variants, additional information regarding the possibility to predict their functional impact [22]. Generally, the annotation is performed by linking the variation data to existing public databases. This process became automatic thanks to software able to rapidly aggregate information coming from different databases to the VCF file such as ANNOVAR [39] and SnpEFF [40]. Basically those programs supply two type of annotation including information about the type of variants relative to known transcripts and functional implication by using database for pathogenicity prediction or conservation score.

Indeed each variant can be classified as intronic or coding variant. For single-nucleotide coding variants, can be predicted if the variation corresponds to a synonymous, missense or nonsense mutation relative to the protein coding sequence.

This basic annotation can be improved by using several resources such as public variants database and pathogenicity prediction tools, which will be discussed in the next paragraphs.

1.2.1.5. Public variation databases

The number of variations found in a typical exome sequencing experiment are approximately 20,000 variants [14] and searching for few candidate variants could be a tricky part. The use of public variations database has become necessary to annotate and filter out those variants that are commonly found within the human population. The three most used databases in NGS data analysis are the following:

1.2.1.5.1. dbSNP

The Single Nucleotide Polymorphism Database (dbSNP) is a free public database for genetic variation, including both polymorphisms and variations associated to diseases, across different species [41]. In particular for human population, the dbSNP has been created to support research in several applications such as genetic population studies, investigations into evolutionary relationships and pharmacogenomics [41]. Over the last five years the dbSNP repository had rapidly growing thanks to the different consortia and large sequencing project (Figure 10) [42].

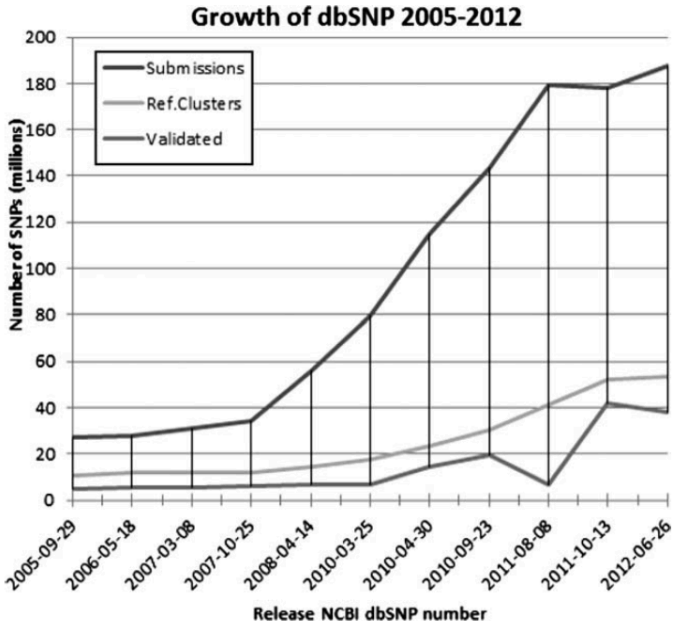


Figure 10 ***Growth of dbSNP repository.*** *dbSNP entries increasing trend from 2005 to 2012. Since the advent of NGS technology in 2008, the variance of the submission (bold black line) and the variations validated (light grey line) is dramatically growing.* [42]

In the NGS era, the use of this database during the bioinformatics analysis became a standard procedure to filter out all the common variants, classified as polymorphisms to narrow down the candidate variations, especially in recessive disorders [43].

Recently, the dbSNP also contains particular variations that are classified as clinically associated to diseases [41]. This information is particularly useful in exome sequencing project to select candidate genes reporting those 'precious' variations.

1.2.1.5.2. 1000 Genomes Project

Started in 2008, the 1000 Genomes Project is an international study with the purpose to sequence at least a thousand people genomes from around the world to create the most detailed and medically useful picture to date of human genetic variation [35]. To date, more than one thousand genomes

have been sequenced by using various NGS methodology and platforms (phase 1). The final purpose is to sequence 2500 genomes. The sample dataset already sequenced, is composed by 1092 healthy individuals coming from 13 populations (http://www.1000genomes.org/).

The phase 1 variation data produced by this huge study are completely available for further investigation. As for the dbSNP, the 1000 Genomes data are used to filter out common variations in NGS experiments. Furthermore, thanks to the allele frequency information calculated for each variant, it gives the possibility to refine the filter of common variants by looking at low frequency variants (>1%) within the 1000 Genomes population and also to evaluate the frequency difference in large dataset NGS project.

1.2.1.5.3. Exome Sequencing Project (ESP)

Another variation database, which has been recently created from the Grand Opportunity (GO) Exome Sequencing Project data, is the Exome Sequencing Project database (ESP). This database is pretty different from the other previously described, because the samples sequenced and analysed are affected by some disorders related to heart, lung and blood (https://esp.gs.washington.edu/drupal/). Moreover only the exome of those samples will be sequenced to catalogue the whole coding part of our genome. Many sequencing centres and hospitals are collaborating to sequence the exome of 8000 affected individuals. The last release (20 June 2012) including 6500 individuals exome data, is available for investigation at http://evs.gs.washington.edu/EVS/ website but the medical record for each patients will not be released.

Recent studies involving cardiac diseases [44–46] used this database to validate their findings by comparing the variation frequencies found in ESP. Although the medical information for the ESP individuals will not be released, this resource could help the clinicians and researchers to interpret and prioritize candidate genes in a cohort of affected samples.

1.2.2. Secondary analysis

Although the development of variation database and the information quality available are growing constantly, the functional and biological significance of the vast number of mutations in the human genome are still unknown [47]. Testing the effect of single-nucleotide mutations required functional study at protein-level, which would be extremely expensive and time-consuming for researchers [48].

Filtering procedures and *in-silico* prediction methods can ease the choice of the candidate genes and they can try to fill the gap between the mutation analysis and the disease.

1.2.2.1. Filtering methods

Recent genome/exome sequencing studies [14, 35, 49] have demonstrated that disease-causing genes are associated to missense (protein coding variations) mutations [48]. Approximately 50 to 75% of variants can be removed by focusing only on non-synonymous (protein-altering) changes [50, 51]. A filtering step is an efficient method to select a list of possible candidate variations.

A typical variant filtration workflow includes several steps to progressively narrow down the number of variations and select only the most pathogenic ones (*Figure 11*). The possible candidate variants are selected on the assumption that (1) the causing mutations alters the protein sequence, (2) should be extremely rare within the population (3) the disease-causing variants should be present in the affected samples and (4) every affected individuals should carry the candidate mutations [52].

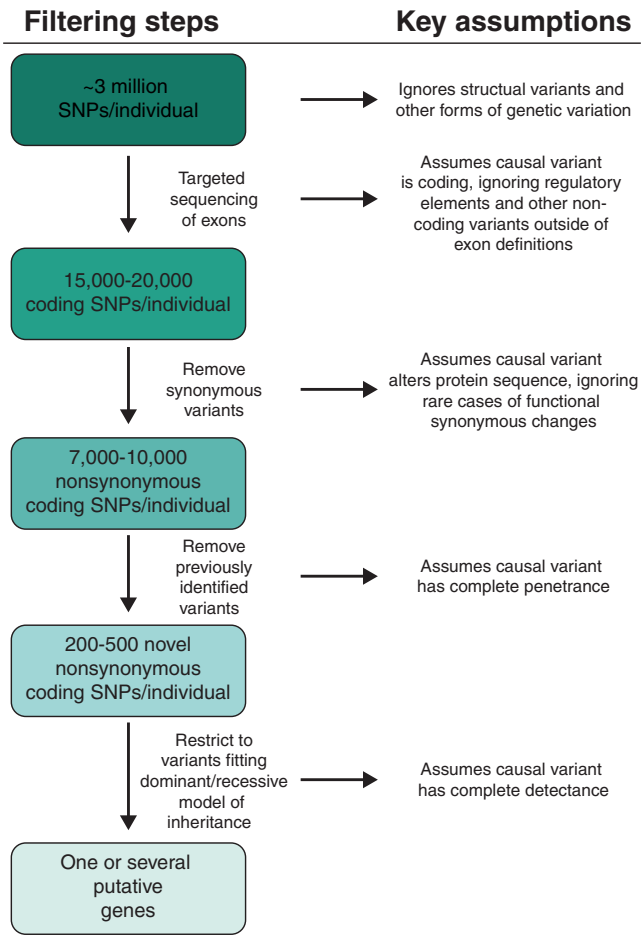


Figure 11 ***Filtration workflow in exome sequencing experiment*** [52]

The application of those heuristic filters on the NGS variants is extremely efficient in exome sequencing experiment for Mendelian disorders [14, 49]. Indeed it is highly proficient the synergistic combination between the variation databases to exclude harmless variation (polymorphisms and synonymous variants) and complete knowledge of the disease to fit the possible candidate variations in a correct inheritance model.

Nevertheless many experiments sometimes missing the control samples or the disease picture is pretty complex (as for the tumor samples). In this

case bioinformatics helps the researchers with specific tools for pathogenicity mutation prediction to depict a list of candidate mutations.

1.2.2.2. In-silico prediction tools

To further investigate the possible effect of non-synonymous variants, several bioinformatics tools have been developed [53]. Despite the huge number of software developed, the mostly used prediction tools in NGS secondary data analysis are based on (1) the sequence conservation among species and (2) the variant localization within the protein structure.

It has been demonstrated [54, 55] that the fraction of disease-causing missense mutation is over-represented in highly conserved region. This is relatively clear because the more conserved is the position the more fundamental it is for the protein function. GERP++ [56] and PhastCons [57] are the most used software to evaluate the genome-wide sequence conservation. Even is they developed two different algorithms, they are able to calculate a conservation score generated by multi alignment data across known species and phylogenetic tree information (*Figure 12*).

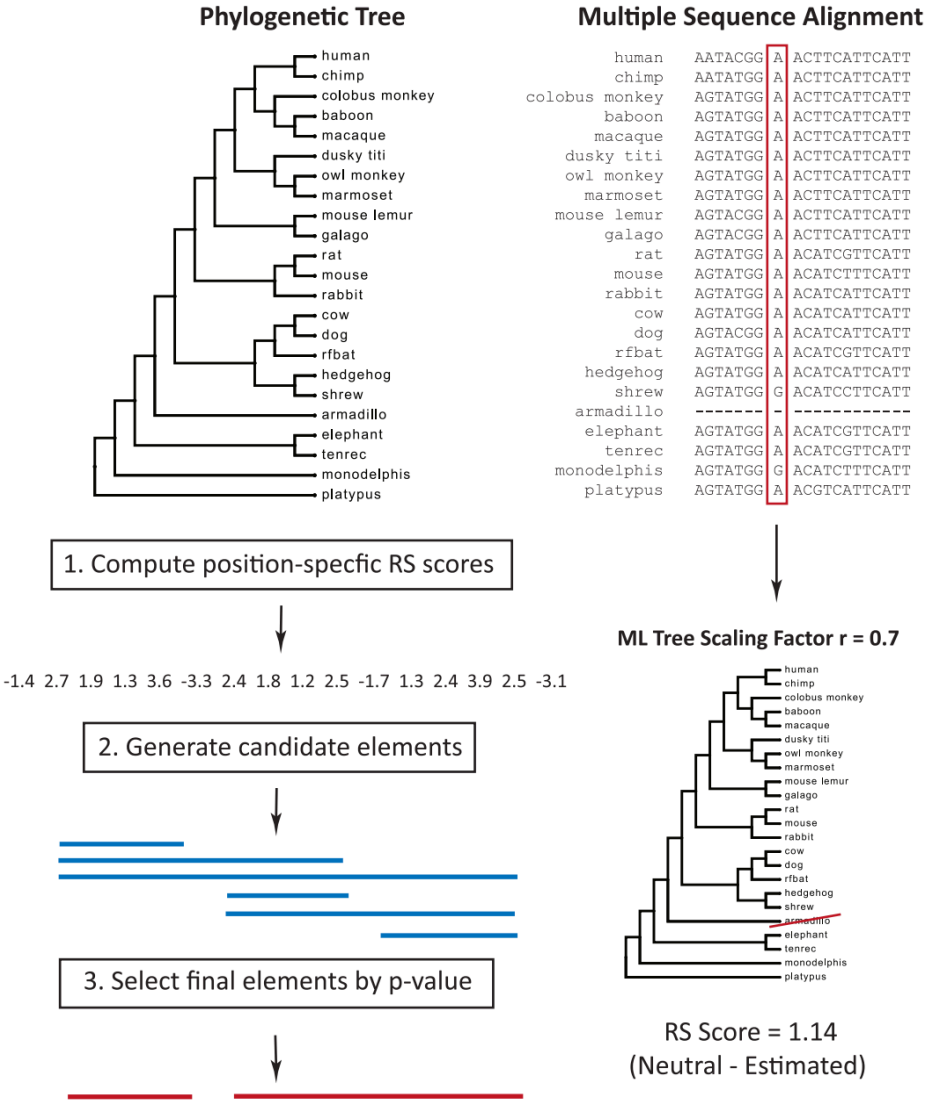


Figure 12 ***Calculation of conservation score workflow performed by GERP***

Other bioinformatics tools predict the pathogenic role of a missense variant by looking at the amino acid change type and the variant localization within the protein structure. Huge number of tools have been developed [53] but the most used in NGS for pathogenic prediction are SIFT [58] and PolyPhen2 [59]. SIFT calculates the probability of an amino acid change by

using a homology-based algorithm. It considers also the evolutionary conservation of each amino acid and predicts whether the substitution found would be tolerated or not. Polyphen2 is a Bayesian classifier that calculates the damage probability of a mutation on the basis of sequence homology and structure feature, where available.
Several data analysis pipelines implemented all those tools to facilitate the use of multiple conservation and prediction tools automatically [17, 18, 60]. Those pipelines used a pre-computed SNP classification for each prediction tool.

1.3. Clinical Sequencing

Application of NGS technologies to Mendelian and complex diseases has proven to be an effective alternative to single-gene tests in research for establishing a new genetic basis of diseases [11, 14, 61, 62]. In particular the development of capture technologies for target sequencing, instead of whole-genome sequencing, dramatically reduced the sample preparation time and the overall costs of the experiment The scalability power introduced by capture technologies defined two experiment categories based on the complexity of the disease: the whole-exome sequencing for complex traits and large gene-panel target sequencing for routine diagnostic (Figure 13).
The whole-exome sequencing would be particularly efficient for very few cases (<10 patients) where the standard clinical analysis (radiographic features, biopsy findings, karyotyping or single-gene testing) defects in order to find novel associated genes. On the other hand, the target sequencing of many candidate genes (from 50 to 250) in a large cohort of patients is now feasible and with a relatively low-cost per sample [63].

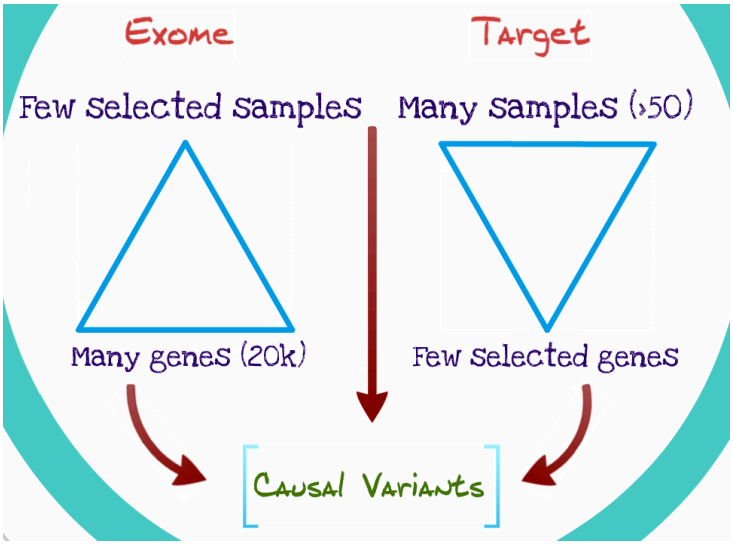


Figure 13 ***Whole-exome sequencing and target sequencing in clinical diagnostic.*** *Generally in diagnostics the use of NGS technology is applied following two main strategies: Whole-exome sequencing for few selected samples with a severe disease to discover novel causing genes where standard diagnostics procedure fails and (2) target sequencing of candidate genes to produce a molecular diagnostic platform to reduce the cost per sample.*

The application of this technology in clinical laboratories would be the perfect combination to unlock complex disease pictures and to provide anticipatory guidance and prognosis where other diagnostic testing has not been definitive [64].

Even if conceptually the NGS application is going to be attractive, the implementation as clinical laboratory standard will raise practical challenges. In particular three main aspects have to be considered: (1) the data management and analysis, (2) the reliability of the method and (3) the standard procedures for analysis. The huge amount of data and the intensive analysis process requires a bioinformatics facility core to store, analyse the data and interpret the results. Discussion regarding the analysis metrics such as read mapping, minimum depth for variant call sensitivity should be considered for standard guidelines. Moreover

technical limitations regarding the capture system should be taken into account for accuracy, sensitivity and precision in clinical practice [65].

The continuous improvement in capture technology and bioinformatics tools will rapidly revolutionize the concept of diagnosis in clinical field.

1.4. Brugada syndrome

Brugada Syndrome (BrS [MIM #601144]) is an inherited autosomal cardiac arrhythmogenic disorder with a prevalence of 1:5000 in Western countries, diagnosed by the presence of a typical ECG pattern with ST-segment elevation and right bundle branch block [66]. BrS is characterized by ventricular instability, which may lead to malignant tachycardia and sudden cardiac arrest in young asymptomatic adults [67] and is estimated to be responsible for 4% of all sudden deaths (SD) and 20% of SD's among patients with structurally normal hearts. The only widely accepted treatment of BrS is the implantation of cardioverter-defibrillator (ICD).

In the past BrS was regarded as "monogenic" disorder [68, 69]. Today 17 genes, mainly ion channel genes [44], are actually associated to BrS. In particular the *SCN5A* mutations account for the vast majority. Despite the number of genes associated has increased, approximately 70% of BrS cases cannot be genetically explained with the current knowledge depicting a complex picture of BrS genetics [44].

Recently Risgaard et al investigated about the relationship between 12 genes previously associated to BrS by using the ESP data [70]. They found a very high prevalence of BrS mutations (1:23) in ESP data compared to the prevalence of BrS in general population [70]. Further investigation [44] consider the entire set of mutations coming from all the 17 genes know to be associated to BrS and they estimated an even higher prevalence of BrS mutations of 1:21. Despite no clinical data are available for the ESP samples, have been demonstrated that none of the patients are affected by channelopathies [71, 72].

On one hand these data really interrogate on the real pathogenic role of the newly BrS associated variants, on the other hand reveals that the use of new population dataset, obtained by NGS technology, is extremely useful

for discriminating the disease-causing instead of common variants in human population.

1.5. Aim of the study

The goals of this work are the following:

1. Develop an automated pipeline to perform the primary bioinformatics analysis of NGS data such as quality control, mapping and variant calling.
2. Develop an annotation and database integration system to select and prioritize mutations probably the pathogenic role.

To validate the application of the bioinformatics pipeline, we analysed two case studies with different NGS strategies:

1. Case 1: Target sequencing on Brugada Syndrome patients.
    a. We designed a custom Agilent capture system in order to analyse all the coding exons of 158 selected genes in our cohort of 91 patients. The goal is to discover novel mutations in genes not previously associated to BrS.
2. Case 2: Undiagnosed arrhythmogenic disease.
    a. We performed whole-exome sequencing of a family trio where the proband is the child. We will focus the analysis on the variation difference between the unaffected parents and the proband. The main purpose is to find causative mutations related to the severe proband phenotype.

2. MATERIALS AND METHODS

2.1. Case 1: Target sequencing of 91 Brugada patients

2.1.1. Patients Clinical Profile

The investigation conformed to principles outlined in the Declaration of Helsinki. Written informed consent for genetic analysis was obtained. 91 consecutive Italian BrS patients (age 50.5± 13.2, 87% males) were selected based on the presence of electrocardiogram (ECG) type I, either spontaneous (n=55) or induced by flecainide or ajmaline infusion (n=36) [73] (section 6.1.1)

Patients were subjected to personal and family history acquisition, evaluation of blood electrolytes, 24 hours Holter-ECG, echocardiography and, whenever possible, cardiac magnetic resonance imaging to exclude the presence of other conditions associated to ST elevation. 35 patients were symptomatic for cardiac arrest, ventricular tachycardia (VT)/ventricular fibrillation (VF), or syncope, while in 56 asymptomatic patients diagnosis of BrS was suspected during routine ECG or family screening. 27 patients showed family history for SCD. Therapeutic options were evaluated according to current guidelines [74] and clinical judgment, leading to ICD implantation in 55 patients. Patients were followed at least yearly by clinical assessment, ECG, echocardiography and ICD check. Mean follow-up was 63.4± 21 months. 8 patients had a major arrhythmic event (SCD, documented VT/VF, appropriate ICD shock). All relevant clinical and genetic data were collected in an electronic database.

2.1.2. Genes selection

For target sequencing we drew a panel of 158 genes that may influence cardiac electric conduction and have a role in the arrhythmogenic mechanism (section 6.1.2). In order to capture the coding region of all the gene in the panel, all the exons of the 158 genes has been selected. Briefly we downloaded all the genomic coordinates of the longest isoform for each gene to capture by using the information stored in UCSC Genome Browser tables. The probe design also included the 5' UTR flanking the first exon for every gene.

We finally performed the design of 2320 coding exons corresponding to 498.094 nucleotides. Target capture was performed with the solution-based kit SureSelectXT Custom 1kb-499kb (Agilent technologies, Santa Clara, CA) according to the manufacturer's version 1.0, compatible with Illumina paired-end sequencing library.

We applied several criteria to choose these candidate genes. Since BrS is characterized by electrical instability, we selected mainly ion channels genes Figure 14.

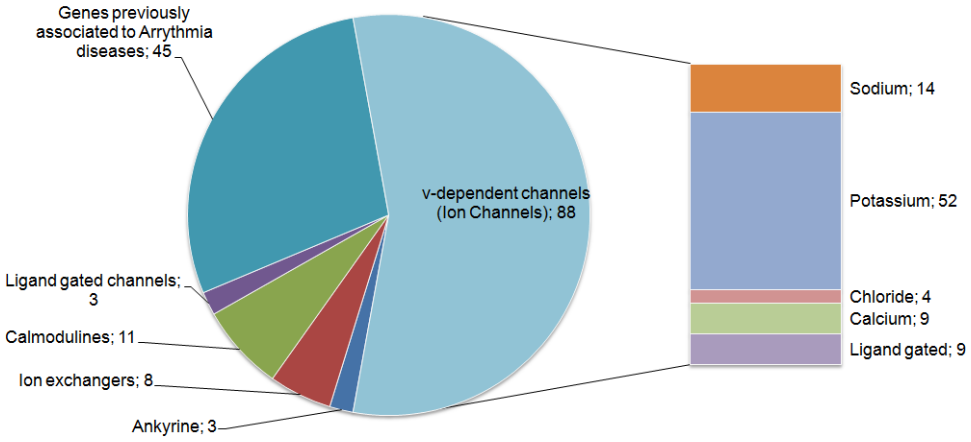


Figure 14 ***Functional categories of the 158 selected genes***

In particular we considered the genes belonging to those 10 categories:

1) *All subtypes of voltage gated Na^+ channels* because of their leading role in the upstroke of cardiac action potential [75–79];
2) *Voltage gated potassium, chloride and calcium channels*, playing a role in cardiac action potential or expressed in cardiac tissue [76, 78, 80–84];
3) *Accessory subunits of sodium and potassium v-dependent channels,* having a known modulator role on the gating of their own alpha subunits [76, 85];
4) *Ligand-gated channels*, such as adrenergic, serotonin, nicotinic and muscarinic acetylcholine receptors, modulating the parasympathetic control of the heart via the vagal nerve activity [86–88];
5) *Ankyrins* [89–91] and subtypes of calmoduline [90];
6) *Transporters and ion exchangers* [76, 92];
7) *Scaffold protein* such as calsequestrins [76], calreticulins [93], all genes associated to arrhythmogenic right ventricular dysplasia (ARVD) and caveolins interacting with sodium channels [92, 94, 95]; anchoring, adapter proteins and modulators of ECG intervals, previously associated with arrhythmia susceptibility [76, 81, 96–100];
8) *Gap junctions* expressed in the heart and involved in other forms of arrhythmias [76, 101];
9) *Structural proteins* involved in cardiomyopathies (such as actin family, myosin, lamin A/C, syntropin and desmin) [76, 92];
10) *Other genes* associated with increased arrhythmic risk in BrS patients [102], angiotensin receptors associated with increased susceptibility to SCD [103] and other genes described in literature in correlation with arrhythmic events [76, 89, 92, 94, 97].

2.1.3. Target sequencing

The wet lab workflow involved several technique for indexing, enrich and sequence the samples into pool of 5 to 7 samples per Illumina GAIIx lane.

Briefly, each DNA sample was targeted with a tag (DNA sequence) in order to recognize the sample after sequencing, afterward each sample DNA tagged were enriched with the Agilent target region probe previously designed.

Genomic DNA (gDNA) of selected patients was extracted from peripheral blood using the automated extractor Maxwell16 (Promega, Milano, Italy); the concentration and high quality of gDNA ($A_{260/280}$ 1.8 to 2.0) was evaluated by Nanodrop Spectrophotometer (Thermo Scientific).

Three micrograms of 91 gDNA patients were fragmented using the Covaris shearing system (Covaris inc., Massachusetts, USA). Selected regions of each patient were then subjected to Illumina protocols for cluster generation and massive sequencing.

Paired-end multiplexed sequencing was performed on the Illumina Genome Analyzer*II*$_x$ platform (Illumina, San Diego, CA), combining 9 patients identified by different index sequences in each lane and performing 86-cycle runs. Image analysis and base calling were performed with CASAVA 1.7 software.

2.2. Case 2: Whole-exome sequencing of Trios

2.2.1. Family pedigree and clinical profile

The patient is a 2-years boy born in Afghanistan. At six-month age he had the first episode ventricular fibrillation that led to a sudden cardiac arrest. After that an ICD (Implantable cardioverter-defibrillator) was implanted. In the family history two siblings (brother and sister) suddenly died at 2- and 3-years old respectively, even if they were completely asymptomatic. He has other two sisters (6- and 7-years old) and 1 brother (1-year old), who are alive with no symptoms until now. Their parents are healthy. The

proband's clinical phenotype is not clear and it was not possible to define his clinical diagnosis because of the presence of an atypical electrocardiogram pattern and a high recurrence of episodes of ventricular fibrillations (*Figure 15*).

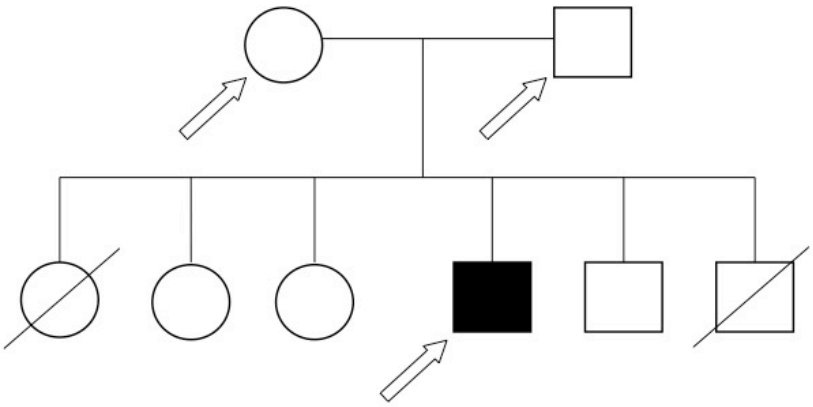


Figure 15 ***Complete family tree.*** *Arrows indicate the samples sequenced by whole-exome sequencing*

2.2.2. Whole-exome sequencing

Whole-exome capture and library preparation was performed using the Agilent SureSelect^{XT} Human All Exon 50Mb kit (Agilent Technologies, Santa Clara, CA, USA), according to the SureSelect^{XT} Target Enrichment System for Illumina Paired-End Sequencing Library protocol (version 1.1.1; Agilent Technologies). Briefly, 3 ug of genomic DNA per sample were sheared using a Covaris S2 AFA instrument (Covaris, Woburn, MA, USA) to a target peak size of 150-200 bp. After fragmentation check by Agilent 2100 BioAnalyzer microcapillary electrophoresis (Agilent Technologies), the exome capture protocol was performed according to manufacturer's instructions and included genomic DNA library preparation and then target enrichment by using Agilent SureSelect exome capture baits. Finally, post-

capture libraries were loaded, one sample per lane, onto the Illumina cBot Cluster Generation System (Illumina, San Diego, CA, USA) at a 8 pM final concentration, and then sequenced by Illumina GAIIx instrument, in a paired-end 85-cycle run. For raw data processing, the Sequencing Control Software (SCS, Illumina) was used to convert raw image data into qseq files and then CASAVA software (Illumina) to generate final fastq files.

2.3. Bioinformatics data analysis

2.3.1. Quality assessment and Mapping

Paired-end reads were mapped against the NCBI human reference genome GRCh37 build using BWA [32] with default parameters. BWA is a tool for mapping raw reads to the reference sequence and its algorithm is based on backward search with Burrows-Wheeler Transform (BWT), to efficiently align reads against a reference sequence, allowing mismatches and gaps.

Briefly, all the reads with more than 5 mismatches or those with mapping quality (MAPQ) less than 15 were filtered out. The MAPQ score is very important in short read aligning because indicates the probability of each mapped read to be misplaced. The score is subsequently transformed in log scale. Afterward, we removed duplicate reads due to clonal amplification during library preparation in order to avoid allele frequency errors and we converted them into standard BAM file, using SAMtools (v. 0.1.12a) [34] obtaining an high-quality alignment (HQ) file for each sample. All the mapping statistics were performed on HQ-mapping files using in-house scripts including calculation of the percentage of read mapping the human genome, calculation of target/exome coverage, calculation of read

mean depth for every sample and calculation of not-covered regions due to technical issues.

2.3.2. Genotype call

Mutations discovery were performed using GATK software [37] that use a Bayesian genotype likelihood model to estimate simultaneously the most likely genotypes and allele frequency in a population of 91 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. This method greatly improves the specificity of the calls and reduces the false positive SNV calls.

Briefly, HQ-mapping files were used for single nucleotide variation and insertion/deletion analysis using GATK software (v.2.1-9), with the following procedure: (a) reads were recalibrated using TableRecalibration walker to bring quality scores closer to their actual probability of mismatching the reference genome; (b) reads were realigned around known indels using IndelRealigner walker and the 1000 Genomes Project data (http://www.1000genomes.org) in order to improve the alignment quality on regions with known INDEL data; (c) the UnifiedGenotyper walker was used to perform SNV/INDEL calling, thus generating a Variant Calling Format (VCF) file containing all the raw variations detected for each sample; (d) groups of three SNVs within 10 base-windows were flagged as “SNPCluster” in the FILTER field of the VCF file.

2.3.3. Mutation filtering

The analysis part of mutation filtering is the key point to ensure the reliability and the strength for mutation data. Applying stringent criteria as heuristic filters on sequencing data like minimum sequencing depth or

strand bias evaluation is the best way to eliminate the maximum number of false positive calls.

To select very confident and high-quality single nucleotide mutations (HQ-SNV) and INDEL mutations (HQ-INDEL), a series of filtering options were applied for further analyses. The filtering criteria were: (i) minimum read depth of 15x; (ii) variation quality score above 150 and (iii) non-reference allele frequency above 20%. Other GATK specific filters, such as strand bias (FS), HaplotypeScore, ReadPosRankSum and Quality by Depth (QD) were applied following the GATK Best Practice v4 (http://bit.ly/13ffAtu). Afterwards, only those variations on the coding regions were selected.

A good way to narrow down the number of possible mutations in exome/target sequencing project is to eliminate those variants already known as polymorphisms in public database. Indeed, to create a subset of novel variations, those not already reported as known polymorphism in 1000 Genome Project (Phase I) [35] or in dbSNP v137, were selected and classified as novel coding HQ-SNVs and novel coding HQ-INDELs. After that, HQ variants were classified and divided into functional category such as silent (synonymous), missense (non-synonymous), nonsense (stop codon) or splice site for HQ-SNVs and frame-shift, codon insertion or codon deletion for HQ-INDELs. Synonymous HQ-SNVs were subsequently ignored creating a subset of novel possibly deleterious HQ-SNVs (NS-SNV) (*Figure 16*).

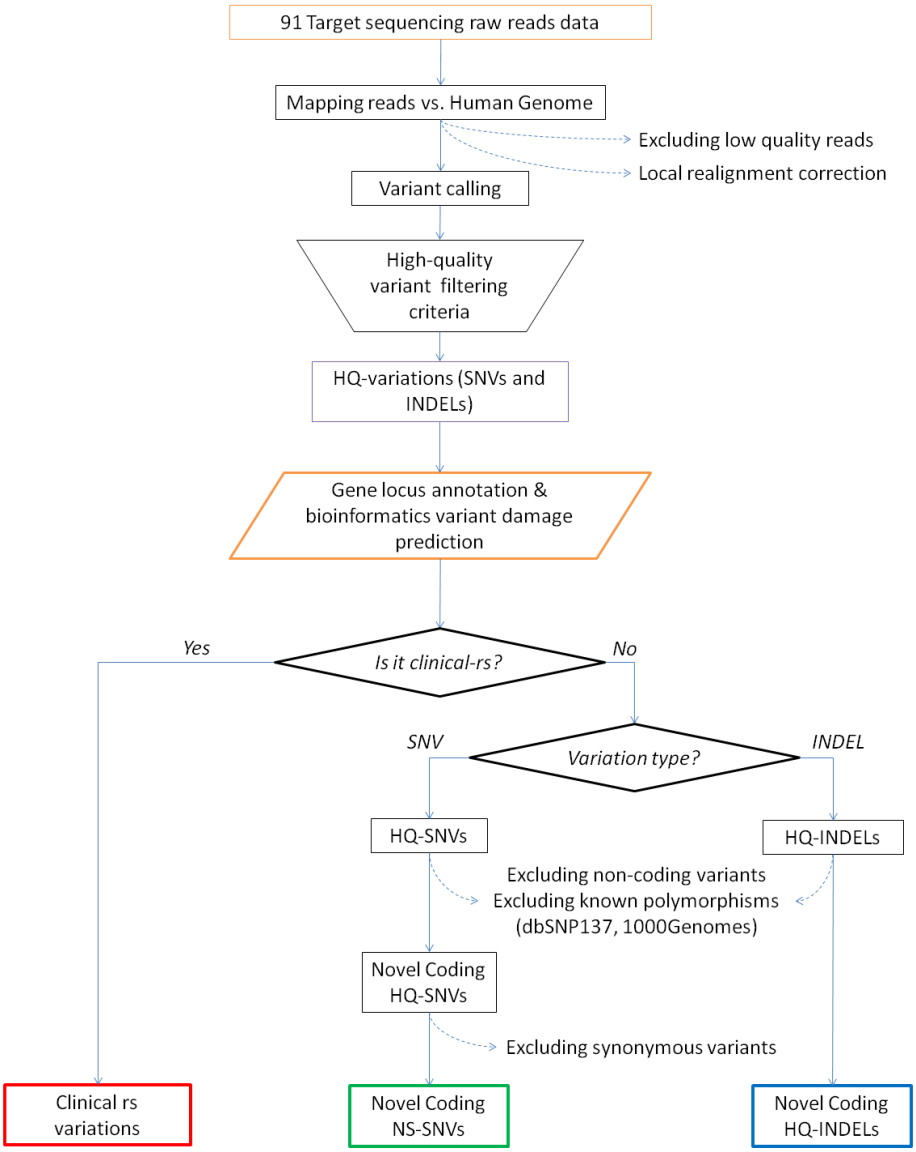


Figure 16 ***Bioinformatics analysis workflow for target sequencing projects.*** *The sequencing reads of the 91 BrS patients have been mapped against human genome reference and the low quality reads were discarded. After the variation call performed by sample, the detected variations have been annotated and filtered by quality. The HQ-variations have been categorized into three category: Clinical rs variations (based on dbSNP annotation), Novel Coding NS-SNV containing the novel deleterious single nucleotide variations and Novel Coding HQ-INDEL including the novel coding insertion or deletions found within the samples.*

2.3.4. Variant annotation and variant quality filtration

Raw variants, known as polymorphisms, were annotated using the VariantFiltration walker in GATK, according to dbSNP v137, 1000genomes data and Exome Sequencing Project (ESP) variations database (http://evs.gs.washington.edu/EVS/). The use of the 1000genomes and ESP databases as resources is very important to evaluate the allele frequency in the human population compared to variants found in our samples. Particularly, the 1000genomes project calculated an estimation of the variations within the healthy human population by generating a value called minor allele frequency (MAF) for every SNV found in their dataset. On the other hand, the ESP database calculated in the same way the MAF value but the 6500 exomes analysed till now, were not collected as healthy samples. That information is fundamental in order to filter-out or annotate a particular variant found in those resources. Variations with known clinical significance were annotated and treated separately.

Gene locus annotation and amino acid variation were generated using SNPEff [40] considering only protein-coding transcripts. Non-synonymous SNVs were subsequently annotated with PolyPhen2 [72], SIFT [58] and Mutation Taster for pathogenicity prediction and with GERP++ [56] for conservation score. Pathogenicity prediction annotation and conservation score were computed using SNPSift [104] in combination with dbNSFP v(2.0b3) [60] as described in http://snpeff.sourceforge.net/SnpSift.html#dbNSFP

2.3.5. Statistical analysis with 1000 Genomes data

Mutation data from 1000 Genomes Project has been downloaded and used for mutation comparison in Case 1 study. Briefly all the chromosomes raw VCF files containing all the integrated call set phase 1 v3 were selected

and downloaded from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/. The integrated call set contained all the variations present in the 1000 Genomes samples produced by genome and exome sequencing. The variations were annotated with VariantAnnotator walker (GATK 2.5-2) by using the annotation file (ALL.wgs.integrated_phase1_release_v3_coding_annotation.20101123.snps_indels.sites.vcf.gz) contained into the 1000 Genomes Project FTP site. Afterward only the 379 European (EUR) samples were selected in order to evaluate the most comparable mutational rate with our cohort. The description of the selected population is available at http://www.1000genomes.org/category/frequently-asked-questions/population. SNV and INDEL variations occurring in target region genes were extracted and selected those variations annotated as coding, non-synonymous narrowing the focus to those variations called by just exome data or confirmed by both genome and exome sequencing data.

To generate a random dataset of the 379 European samples comparable to the Brugada cohort in Case 1, 50 permutations of 91 samples were performed to calculate the number of rare variations (found at least in two individuals) for each subsampling and for every target gene. This dataset were used to draw a distribution of mutation rate of 91 healthy controls and to compare the mutational rate of every gene normalized by gene length with the 91 Brugada patients of Case 1.

3. RESULTS AND DISCUSSION

3.1. Case 1: Target sequencing of 91 Brugada patients

3.1.1. Target Sequencing Statistics

We designed a custom Agilent capture system in order to analyse all the coding exons of 158 selected genes in our cohort of 91 patients. Each sample capture was followed by sequencing on Illumina GAIIx, generating an average 10,833,247 +/- 7,419,828 mapped reads (*Table 1*).

	Mean +/- SD	Median	1st quartile	3rd quartile
Raw Reads	11,188,822 +/- 7,802,174	9,379,338	7,733,175	11,006,395
Mapped Reads	10,833,247 +/- 7,419,828	9,127,572	7,479,202	10,680,006
Depth (fold)	327.22 +/- 137.35	334.20	243.46	411.46
Coverage (%)				
1X	99.16 +/- 1.06	99.44	99.30	99.54
15X	92.94 +/- 9.78	95.18	94.01	96.11

Table 1 ***NGS sequencing statistics of 91 Brugada samples****. Summary of the sequencing performance across the 91 Brugada samples represented by the number of read generated (Raw reads), the number of the reads actually mapped against the Human genome (Mapped Reads), the number of reads, on average, that covered the target region (Depth) and the percentage of target region covered by at least 1 read (Coverage, 1X) and by at least 15 reads (Coverage, 15X).*

An automatic pipeline for data analysis was developed in order to facilitate the analysis steps and to standardize the results for each sample.

After duplicate removal and low-quality reads filtering, we obtained a mean target coverage of 99.16% +/- 1.06 (*Figure 17*) with a mean sequencing depth of 327.22x +/- 137.35 among the samples. Moreover, thanks to the high-sequencing depth obtained, we had a 92.94% +/- 9.78 of the covered regions supported by at least 15 reads, which is the minimum read depth for SNVs.

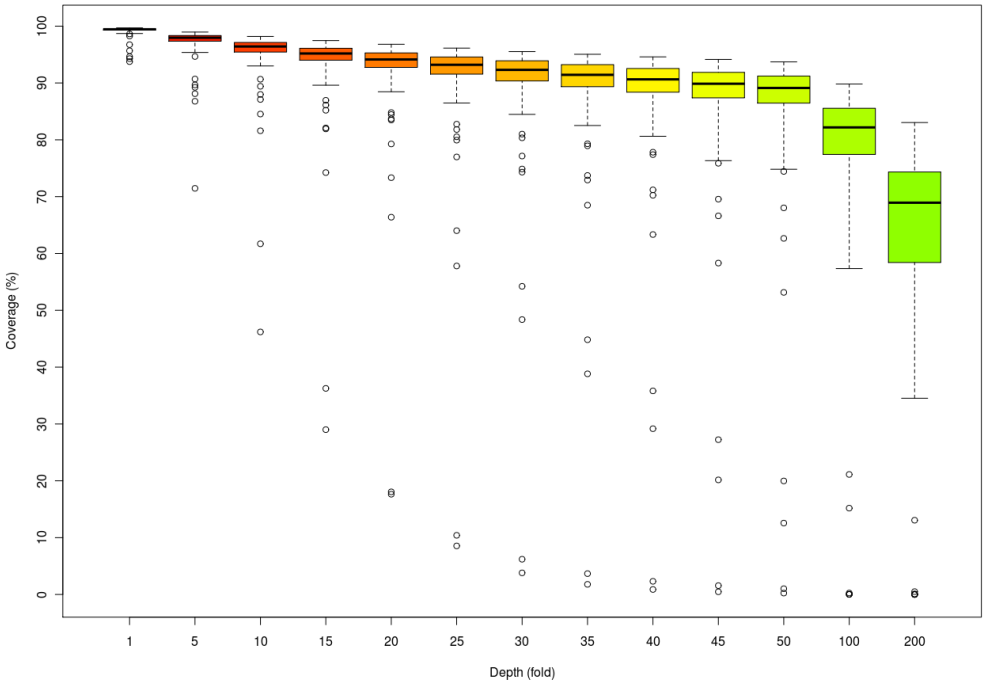


Figure 17 ***Target region coverage by incremental depth****. Boxplot representing the target coverage distribution across the 91 Brugada samples (y-axis) considering different read depth (x-axis).*

Overall, we found 118,388 SNVs and 14,590 INDELs, with an average of 1461 variations per sample and 273 variations per sample considering only those annotated in coding regions (range: 98 to 310). In order to select and prioritize candidate mutations, we filtered out all variations already known as polymorphisms within the human population as well as variations present in non-coding regions, narrowing down to 158 variants with an average 1.7±1.43/patient.

We further classified this filtered variations dataset in three categories: novel NS-SNVs possibly affecting protein function (missense, nonsense and splice-site variations), clinical-rs variants previously reported in dbSNP v137 and novel coding INDELs. In particular, we identified 85 NS-SNVs, 60 clinical-rs variations and 13 novel coding HQ-INDELs (section 6.1.3, 6.1.4, 6.1.5).

3.1.2. Novel variations

82 out of the 85 NS-SNVs were missense substitutions, 2 were predicted to alter canonical donor splice sites, respectively on *KCNJ15* and *SCN10A* genes, and only 1 variant in *NRG1* was reported as nonsense**.** We decided to sequence by Sanger sequencing the lowest covered mutations (with depth < 30X) to exclude the presence of false positive calls introduced by the variant calling system and filtration. All the 9 NS-SNVs below the 30X coverage threshold were confirmed by Sanger sequencing, indicating a robust calling method and quality filtration.

All the NS-SNVs variants were identified in a total of 53 genes and private, with the exception of one missense variant (Ensembl transcript ID ENST00000361308:c.599A>G; p.K117R) in the *LMNA* gene, identified in two BrS patients of our cohort.

In the analysis of the 84 private variants, we found several recurrent genes. In particular, *ANK2* and *ANK3* genes, coding for ankyrin proteins, presented respectively 5 and 4 missense substitutions, and *CACNA1H* gene, encoding the calcium voltage gated channel, harboured 5 NS-SNVs (*Figure 18*)

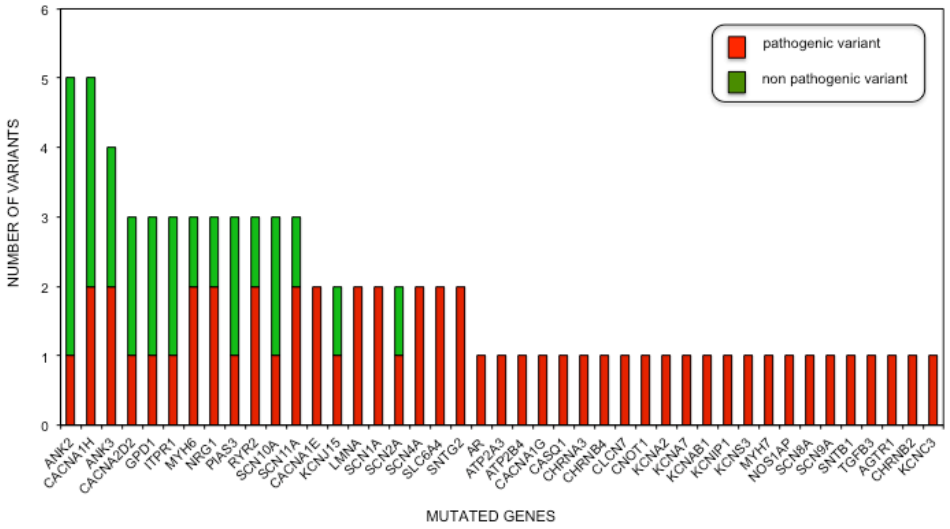


Figure 18 ***Most mutated genes in novel variations category****. Gene mutations count distribution (descending order). Red bar showed the number of mutations predicted to be damaging by at least two out of three bioinformatics tools. Green bar showed the number of mutations with low pathogenicity value (predicted to be neutral)*

In silico pathogenicity analysis revealed that 31 genes carried one variation with a potential pathogenic impact and 12 carried two: voltage-gated sodium channels and their interacting proteins (*SCN1A, SCN4A, SCN11A, ANK3* and *SNTG2*), voltage-gated calcium channels (*CACNA1E* and *CACNA1H*), ligand-gated channels (*NRG1* and *SLC6A4*), genes previously involved in cardiomyopathies (*MYH6* and *LMNA*) and *RYR2* encoding ryanodine receptor. These putative pathogenic variants were present in 38 patients, and in half of them (19/38) these were not associated to other pathogenic variations (either novel or clinical-rs).

Among the 13 novel HQ-INDEL variations identified in 11 patients, 8 were deletions and 5 insertions. The deletion detected in *KCNK17* gene and the insertion in *CLCN2* occurred twice in different patients. Six variations were classified as frameshift with possible dramatic change in protein translation, including two found in the same patient affecting *AGTR1* and *SCL6A4*

genes. The remaining four frameshift variations were private and involved *CHRNB2, MYH6, SCN11A* and *KCNC3* genes.

3.1.3. Clinical-rs variations

We also identified 60 SNVs previously described in dbSNP v137 and, although they were flagged as clinically associated, only a minority was defined as clearly pathogenic. Those particular variants have been previously associated to particular diseases in dbSNP database.

Indeed, 31 out of 60 were missense variations and only 10 of them had been marked as pathogenic: rs72544141 and rs121912706 in *ANK2* and rs74315448 in *KCNC3* had been previously described in cases of Long QT Syndrome (LQTS); rs121913013 in *DSG2* and rs121912998 in *DSP* had been associated with ARVD; rs45454496 in *ANK2* had been linked to a cardiac arrhythmia syndrome; rs121917810 in *AGTR2* and rs121434525 in *ACTN2* had been associated with cardiomyopathy; rs121908919 in *SCN9A* had been described as pathogenic in epilepsy and rs1800888 in *ADRB2* associated with an ischemic heart disease. All these clinical-rs were private in our cohort of patients with the exception of rs45454496 and rs121912998 that were identified in two patients, and rs1800888, present in three different patients.

We analysed pathogenicity of all the other clinical-rs coding variations classified as “untested”, “other” or “probably non pathogenic” in dbSNP v137, by using bioinformatics prediction and conservation tools as described in Methods. According to our criteria three of them, rs121918769 in *SCN1A*, rs141423405 in *KCNE2* and rs71318369 in *CLCN2*, were considered probably damaging. The clinical rs141423405 had been previously reported in a case of LQTS, corroborating the hypothesis that this variant might have a pathogenic role in arrhythmogenic mechanisms.

In addition, five patients were affected by the same missense variant (rs41280102) in the *SCN4A* gene, which is the most recurrent missense variant in our cohort (5.5% frequency). Although this variant was reported as non-pathogenic in dbSNP v137 but untested, it was significantly over-represented in our cohort compared to 1000 Genomes allele frequency ($P<0.05$; Fisher Test).

3.1.4. Permutation analysis with 1000 Genomes data

To verify the robustness of the genes choice, the overall mutations found in the 91 Brugada patients was compared to random permutations of 91 healthy control samples coming from the EUR population of the 1000 Genomes Project. Considering only the NS-SNVs called in our cohort, the overall count (85 mutations) was clearly higher the non-synonymous rare mutations present in 1000 Genomes data (55.94 +/- 6.9) suggesting a real involvement of those genes in BrS development (P-value<0.01) (*Figure 19*).

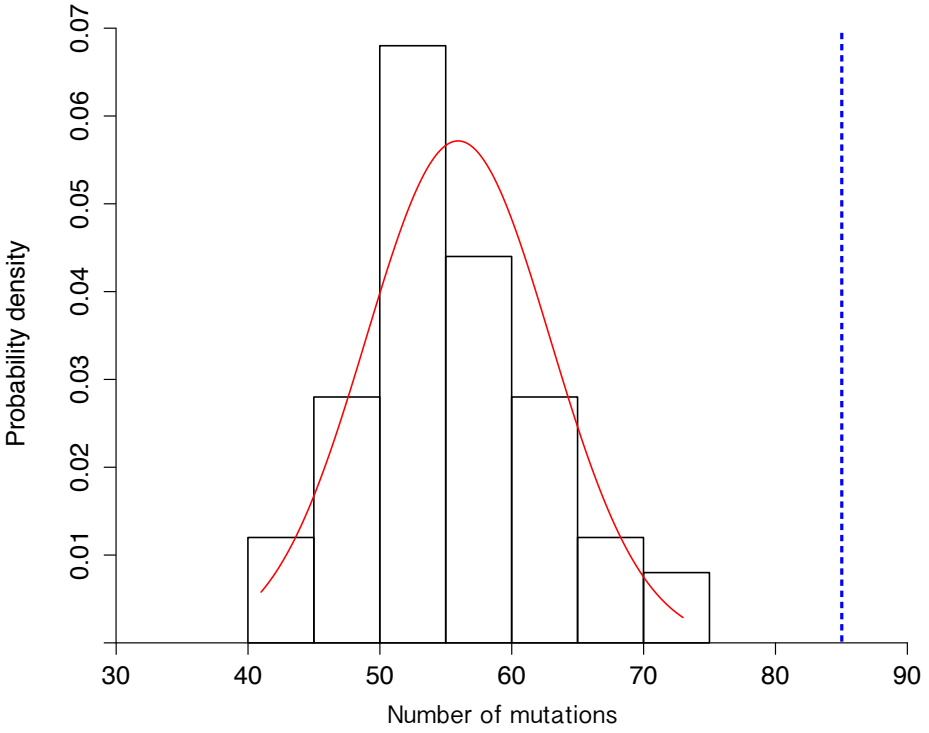


Figure 19 ***1000 Genomes variations distribution compared to Brugada Samples****. Representation of the cumulative mutations, among the 70 genes mutated in the Brugada samples, applying 50 random sampling of 91 individuals from the 1000 Genomes data. The blue dotted vertical line indicates the number of mutations found in our cohort.*

Moreover, to examine in depth the significance of the mutations, to better classify the mutated genes and to prioritize them in relationship with BrS, we compared the mutational rate of the most recurrent mutated genes (two or more mutations found in our cohort) with the mutational rate of those genes in a healthy control dataset as the 1000 Genomes Project data (*Figure 20*).

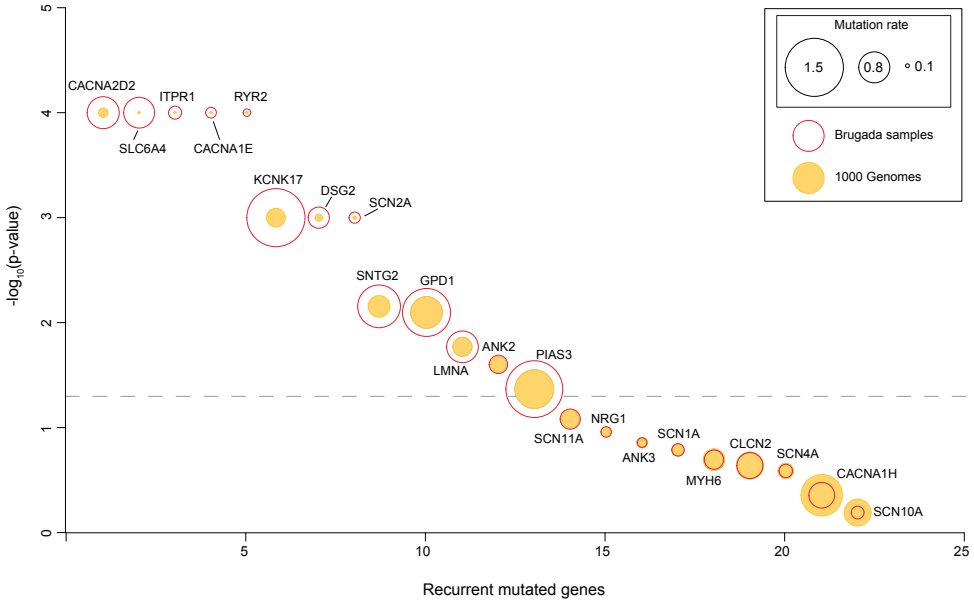


Figure 20 ***Mutation rate comparison of recurrent mutated genes in Brugada samples against 1000 Genomes****. Evaluation of the mutational rate (number of mutation per kilobase) for recurrent genes (at least 2 mutations across all samples) found in our cohort in comparison to variation found in 1000 Genomes data. Yellow circles indicate the average mutational rate in 1000 Genomes. Red circles indicate the mutational rate found in Brugada samples Y-axis indicates the -log$_{10}$ of P-values, grey dotted line represents the 0.05 significance threshold).*

This analysis revealed that a few genes with high number of mutations in our cohort, including *ANK3, CACNA1H, MYH6*, were affected by the same mutational rate (P-value>0.05) as in the 1000 Genomes data actually decreasing the strength and the involvement of those genes in the disease. On the other hand, we found 13 genes were strongly mutated compared with the mutational rate in the 1000 Genomes data (P-value<0.05, -log$_{10}$P-value>1.3). Among all the statistically significant genes the most relevant ones for the biological function were *RYR2*, which was never been associated to BrS, and two genes encoding for calcium channels such as *CACNA2D2* and *CACNA1E*.

3.1.5. Downstream analysis and mutations discussion

Candidate BrS genes were evaluated based on two main criteria: frequency of variations in our cohort and their predicted *in silico* pathogenicity. These results suggested several new candidate genes with a possible involvement in the pathogenesis of BrS. Overall, possibly pathogenic NS-SNVs were detected in 38 patients and in half of them this was the only pathogenic variation identified, suggesting a causative association with the disease.

The most mutated gene in our cohort was *RYR*2, encoding ryanodine receptor and previously associated to ARVD and catecholaminergic polymorphic ventricular tachycardia (CPVT), where it has a principal role in mechanisms underlying triggered arrhythmias [105]. In this gene we identified 7 clinical-rs, classified in dbSNP v137 as probably non-pathogenic, consisting of non-coding and synonymous variants. In addition, we detected 3 private NS-SNVs, including 2 with a high pathogenic effect predicted by bioinformatics tools. *RYR2* had never been associated with BrS until now and we might thus hypothesize a role in this syndrome and in addition it was never seen mutated in all the 50 sampling permutation of 1000 Genomes mutation data.

Our data also confirms the overlap of clinical phenotypes reported in some BrS and ARVD patients [106, 107]. Indeed, besides *RYR2,* our study detected several variants in genes associated with different types of ARVD, such as *DSG2, DSP* and *TGFB3*, suggesting a common genetic background in the arrhythmogenic mechanism in these different pathologies.

The sodium channel family was the most mutated in our study considering clinical-rs and NS-SNVs. Influx of sodium ions through cardiac voltage-gated sodium channels is responsible for the initial fast upstroke of the cardiac action potential, thereby triggering the initiation and propagation of

action potentials throughout the myocardium [108]. Cardiac sodium channels thus play an essential role in excitability of cardiac cells and proper conduction of the electrical impulse within the heart and, although *SCN5A* gene is the principal causative gene in BrS, we may assume that also other types of voltage-gated sodium channels may be involved in arrhythmogenic mechanisms. In particular, we identified possibly pathogenic variants in *SCN1A, SCN2A, SCN4A, SCN8A, SCN9A, SCN10A* and *SCN11A* genes in 15 patients out of 91 (nearly 17%).

Although when BrS was described for the first time in 1992 it was associated to a loss of function of the *SCN5A* gene, it was later demonstrated an important role for other voltage-gated ion channels in BrS pathogenesis. Indeed, to date, several potassium channels have been associated with BrS, such as *KCNE3* [MIM: 604433] (BrS6), *KCNH2* [MIM: 152427] (BrS8), *KCNE2* [MIM: 603796] (BrS9), *KCNJ8* [MIM: 600935] (BrS10), *KCNE5* [MIM: 300328] (BrS13) and *KCND3* [MIM: 605411] (BrS14) [109].

Our results suggest an important pathogenic role for other potassium channels that were recurrently mutated in our cohort; for example missense pathogenic variants were detected in *KCNA2, KCNA7* and *KCNS3* [MIM: 603888] genes, which are expressed in cardiac cells and responsible for the delayed rectifier current and one splice site variants was identified in *KCNJ15*, encoding the inwardly rectifying potassium channel. Notably, 16 patients out of 91 (nearly 17%) harboured mutations in potassium channels.

Mutations localized in cardiac L-type calcium channels had been clearly associated with typical BrS ECG pattern and the development of polymorphic VT, that can lead to SCD [110], but until now the protein Cav1.2, encoded by *CACNA1C* gene, was the only alpha subunit of cardiac L-type calcium channel associated with BrS (BrS3) [109]. Our analysis suggests a possible pathogenic role also for cardiac Cav3.1 and Cav 3.2

alpha subunits of calcium channel encoded by *CACNA1G* and *CACNA1H* genes, in which we detected respectively 1 and 2 probably pathogenic novel missense variants. In addition, the alpha-2/delta subunit of calcium channel is encoded by at least 4 different genes and one of them, *CACNA2D1*, is already annotated in literature as a BrS susceptibility gene [111]. Our study identified 3 novel missense variants in *CACNA2D2*, encoding alpha2/delta2 subunit of calcium channel expressed in heart and one of these NS-SNVs was predicted as pathogenic and it was the gene with the stronger mutational rate difference between our cohort and the 1000 Genomes samples. In our study we identified 7 mutations carriers out of 91 patients in these 3 genes (8%). We can therefore hypothesize the involvement of other types of alpha-2/delta subunit of calcium channel in BrS pathogenesis.

A gene carrying several novel variations in our cohort is *ANK2* encoding ankyrin-B, an adapter protein whose genetic variants have been previously associated with arrhythmic susceptibility. Indeed, in this gene we detected 3 clinical rs (rs45454496, rs72544141 and rs121912706) previously described in literature in clinical cases of arrhythmogenic disease and LQTS [112]. It is known that several genetic variants can be implicated both in LQTS and BrS, so we may suggest a possible implication of these clinical rs in BrS pathogenesis. Moreover, in one patient we detected the clinical rs66785829 in *ANK2*, previously described in a patient with typical ECG type I [113], corroborating the idea of a causative role of *ANK2* in BrS pathogenesis. *ANK2* is also the most mutated gene considering NS-SNVs, leading us to hypothesize that also the 5 novel missense substitutions detected in our cohort might have a causative role in BrS phenotype.

Another interesting candidate gene may be *ANK3*, encoding ankyrin-G isoform, in which our target sequencing detected 4 NS-SNVs in 4 different BrS patients. It has been shown that ankyrin-G is associated with the cardiac sodium channel and that, like Nav 1.5, it is highly expressed in

cardiomyocytes. Moreover, a human mutation in the *SCN5A* gene was described to block ankyrin-G binding, affecting the expression of Nav 1.5 in the cardiomyocytes membrane and resulting in BrS phenotype [114], suggesting the pathogenic role of ankyrins in BrS with 14 out of 91 (15%) patients harbouring mutations.

NRG1 is another gene that may be associated with BrS phenotype, carrying 3 novel variants, 2 of them predicted as damaging by our *in silico* analysis. Neuregulins are important for maintenance of acetylcholine receptor-inducing activity of nicotine receptors in neurons, skeletal muscle and in the heart where the parasympathetic activation counterbalances β-adrenergic activation [115]. Indeed, it is known that the magnitude of ST-segment elevation can be reduced by adrenergic agonists, whereas it is increased by parasympathetic agonists or adrenergic antagonists [116, 117].

The putative role of the nervous system in arrhythmia induction is confirmed also by the role of Nav1.8 channel, encoded by *SCN10A* gene, that is highly expressed in neurons of dorsal root ganglia and cranial sensory ganglia and described by several genome-wide association studies as a modulator of cardiac electrophysiology [46]. Nav 1.8 based sodium channels are absent in cardiomyocytes but are present in the intracardiac neurons of murine heart and a pharmacologically blockade of this channel may influence the properties of the neural activity[118]. In particular, *SCN10A* expression contributes to late sodium current in heart and represents a new target for antiarrhythmic intervention [119].

BrS was originally considered an electrical disorder occurring in the absence of structural heart disease, and the presence of myocardial abnormalities in these disorders was excluded by definition. However, another interesting data emerging from our study is the presence of novel variants with a probable pathogenic impact in *MYH6* and *MYH7*, encoding the alpha- and beta-myosin cardiac heavy chain, respectively, and involved

in different cardiomyopathies. Indeed, several studies already showed that sodium channel dysfunction may result in structural abnormalities in the myocardium, even if the pathological mechanisms possibly involved have still to be clarified [92, 120].

Interestingly all the detected novel variations are private with exception of the missense variant p.K117R (Ensembl transcript ID ENST00000361308:c.599A>G; p.K117R) identified in the *LMNA* gene, encoding lamin A/C, in two BrS patients of our cohort. Also *LMNA* gene had never been involved in BrS, however it has been associated with more than 10 different clinical phenotypes, including cardiac abnormalities characterized by atrial fibrillation, conduction-system disturbances, dilated cardiomyopathy, sudden death and heart failure [121, 122]. Carriers of lamin A/C mutations can develop VTs with progressive conduction block and sinus block requiring ICD therapy for efficacious primary prevention [123]. It may be therefore important to investigate the possible causative role of *LMNA* mutations in ventricular instability in BrS patients.

Moreover to better understand genotype-phenotype correlation, we performed an enrichment analysis of the most mutated genes of our panel, classified by functional category. We also divided patients in classes according to three phenotype traits such as ECG type I, Flecainide positive and Sudden death familiarity.

Considering the patients of our cohort with a family history for SD (29%), the most mutated genes encode calmoduline and this functional enrichment is statistically significant (P-value<0.05; Fisher Test) (*Figure 21*).

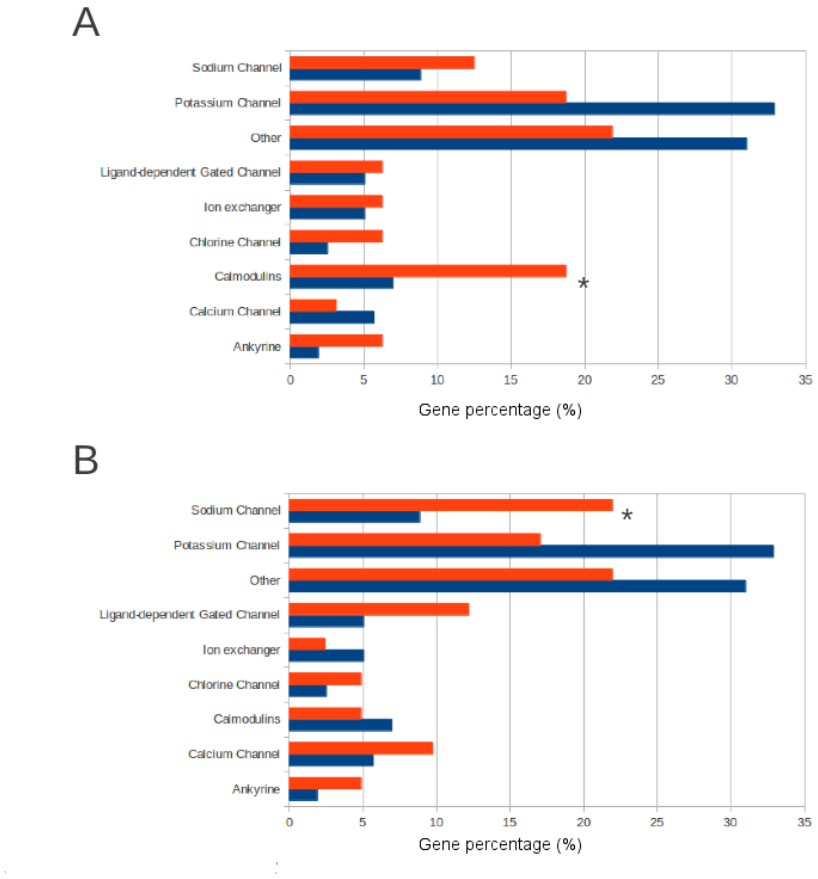


Figure 21 Genes functional enrichment considering to Brugada phenotype: Genes of the target were divided into functional category based on the protein function. Genes mutated in patients (red bars) were compared to the percentage (x-axis) of genes within the target list (blue bars). The Calmoduline gene family resulted statistically enriched in patients with sudden death familiarity (panel A). Whereas the Sodium channel family resulted enriched in patients belonging to type I ECG and Flecainide positive category (panel B). Stars (*) indicate P-value<0.05 for Fisher Exact test.

Interestingly, grouping the samples with spontaneous type I ECG and those patients with a type I ECG induced by Flecainide, the most mutated class of genes is the Sodium channel (P-value<0.05; Fisher Test). This result suggests and confirms the essential role of the sodium channel family in relationship with the BrS onset.

In conclusion, this study identified new several candidate genes associated with BrS ECG. In order to better evaluate their contribution, it will be important to combine genetic studies with molecular and functional experiments that will help identify and prioritize novel genes and pathways potentially relevant for cardiac electrophysiology [124, 125]. In addition, intra-familial segregation studies will be necessary to correlate the role of putative pathogenic variants to the pathological phenotype. Notably, since 20 BrS patients of our cohort were not mutated in any gene of the selected panel, we plan to perform in these cases an exome-sequencing to identify new putative causative genes. These studies may provide new targets for genetic screening and future development of therapeutic strategies [90].
Ultimately, these combined clinical, genetic and translational studies should lead to improved diagnosis, risk stratification, treatment and outcome in patients with diagnosis of BrS.

3.2. Case 2: Whole-exome sequencing of family trio samples

3.2.1. Whole-exome sequencing statistics

We performed a paired-end sequencing for each sample of the family producing an average of ≈ 175 million of reads. To ensure the maximum reliability and robustness for mutation analysis, we performed at least two runs to reach the minimum exome coverage of 94.19% with a minimum mean depth of 87.15X for each sample (*Table 2*).

	112252 (Child)	112315 (Mother)	112316 (Father)
Raw reads	127,090,729	212,210,584	185,347,716
Mapped reads (%)	75,100,902 (59.09)	188,083,331 (88.63)	164,272,934 (88.62)
Mean depth (fold)	87.51	186.05	166.73
Coverage (%)			
1X	94.19	96.63	96.09
15X	85.81	91.67	90.73

Table 2 ***Exome sequencing statistics of trio samples****. Summary of the sequencing performance for all the trio samples indicated by the number of read generated (raw reads), the number of the reads actually mapped against the Human genome with their percentage related to the reads produced (mapped reads), the number of reads, on average, that covered the exome region (depth) and the percentage of exome region covered by at least 1 read (coverage, 1X) and by at least 15 reads (coverage, 15X).*

We obtained that more than the 85% of the entire target was covered by at least 15 reads, which is the minimum read depth for SNV calling, indicating a robust and reliable coverage to evaluate variations in the majority of the target. The statistics for sequencing quality check are essential for the trio analysis because only those positions overlapping the three samples are eligible for the mutation call analysis, the non-overlapping ones were not considered for further analysis. After quality control and mutation calling, all the variations found in all the three samples were analysed by just considering SNVs.

3.2.2. Trio analysis

The big advantage of trios analysis is the opportunity to retrace the inherited mutations found in the disease-affected sample back to the parents and to clearly identify de novo mutations occurred in the affected sample by analysing Mendelian violations.

We found a mean of 39,147.6 SNVs across the samples considering the exome region captured (*Table 3*).

	112252 (Child)	112315 (Mother)	112316 (Father)
No. sites overall	37,828	40,353	39,262
Quality passed sites	32,668	36,541	35,329
Novel mutations	771	928	883
Coding mutations	395	486	450
Synonymous	173	201	173
Non-synonymous	213	271	267
Nonsense	6	7	7
Splice site sites	3	7	3

Table 3 ***Mutations annotations and filtration****. After SNV calling, variations were reported and annotated to evaluate their strength in disease onset. The raw calls (No. sites overall) were filtered by quality parameters described in Materials and Methods section. High-quality variations (quality passed sites) were further filtered by eliminating those variations known as polymorphisms within the population according to 1000 Genomes and dbSNP v137 data. The resulting novel variations (novel variations) were annotated by considering the coding region and divided into 4 categories (synonymous, non-synonymous, nonsense, splice-site).*

For better evaluate those variations with a possible involvement to the disease, we filtered out the variation occurred in non-coding region as well as those annotated as synonymous SNVs which does not change the protein sequence. Moreover all the SNVs present into human variation databases such as 1000 Genomes and dbSNP v137, were considered as polymorphisms within the healthy human population and therefore were excluded from further analysis.

In this particular case, considering the rarity of the disease, the hypothesis was to find homozygous mutations in the child, the affected sample, which could explain the severe arrhythmogenic disease not present in the parents. Indeed on one side, the homozygous mutations in the child were evaluated in order to select candidate mutations related to the disease, secondly the *de novo* mutations were extracted by subtracting the variations found in the parent sample.

Considering the 222 novel possible deleterious mutations (non-synonymous, nonsense and splice-site categories) found in the child, only six occurred as homozygous and therefore were investigated as potential candidate mutations (*Table 4*). Although all the mutations were predicted by at least one bioinformatics tool to be dangerous for the related protein, two of those occurred in genes that were already associated to arrhythmogenic disease.

Chr	Pos	Ref	Alt	AA change	Mutation type	Gene name	MutationTaster	Polyphen2	SIFT
6	123892133	A	G	L56P	MISSENSE	TRDN	-	D	0
11	118531346	G	A	S212L	MISSENSE	TREH	N	B	0.16
11	125453602	G	A	G48R	MISSENSE	EI24	-	-	0
15	91491962	G	A	D591N	MISSENSE	UNC45A	D	D	0.49
16	68011241	A	G	I342T	MISSENSE	DPEP3	D	B	0
22	39032562	T	C	Y132C	MISSENSE	RP1-199H16.5	N	P	0.02

Table 4 ***Six novel candidate homozygous mutations****. Novel possible deleterious mutations were reported. All the missense mutations were also evaluated by three different bioinformatics tools (MutationTaster, Polyphen2, SIFT). Mutattion Taster and Polyphen2 score description: D=Damaging, B=Benign, N=Neutral, P=Probably damaging. SIFT score description: score<0.05 indicates damaging prediction otherwise benign prediction.*

In particular *UNC45A* encodes for a chaperone protein that is involved in cell proliferation and myoblast fusion by binding progesterone receptor and acts as regulator within the progesterone receptor chaperoning pathway [126, 127]. The mutation found was predicted by 2 out of three bioinformatics tools a damaging mutation that could influence the functionality of the protein. Knockdown of *UNC45A* in *Drosophila melanogaster* results in a reduced beat rate as well as reduced cardiac contractility suggesting a key role in heart activity [128].

Another interesting mutation was the p.L56P occurred in the *TRDN* gene. This gene, encoding for an integral membrane protein named triadin, is strictly involved in excitation-contraction coupling as part of calcium release complex in association with ryanodine receptor (*RYR2*). Mutations in *TRDN* gene have been found in catecholaminergic polymorphic ventricular

tachycardia (CPVT) [129] characterized by bi-directional ventricular tachycardia that may degenerate into cardiac arrest causing sudden death. In particular, p.T59R mutant of triadin protein in mice resulted in instability of the protein leading to its degradation. The homozygous mutation found in our affected sample share the same protein domain and could suggest the same behaviour of the T59R mutant triadin and including this mutation in the candidate list. Obviously, further functional analysis needs to confirm the absence of the triadin protein and therefore its implication in the disease onset.

To classify the strength of the candidate mutations, segregation study has been performed among relatives (*Figure 22)*. Using Sanger sequencing validation we confirmed the six heterozygous mutations in the parents and we found that all the six homozygous mutations were heterozygous in the two unaffected sisters and in the one unaffected brother.

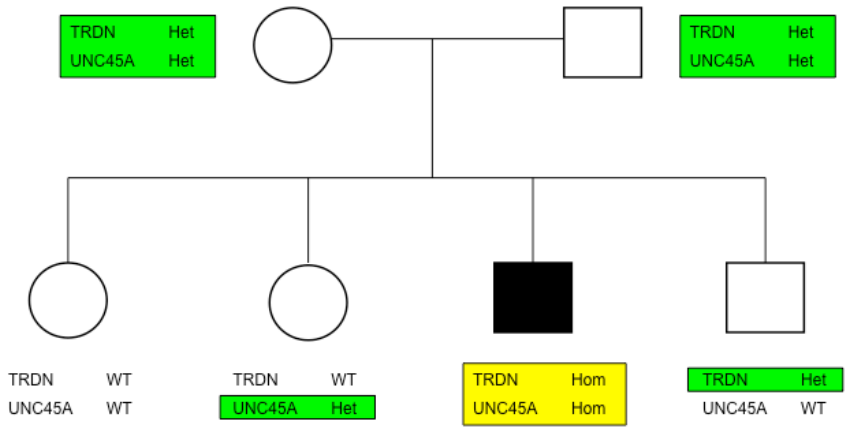


Figure 22 ***TRDN and UNC45A segregation analysis****. Family tree indicating genotype information about the candidate mutations found in the proband. The colours highlight the genotype status of the mutations, green boxes indicate the heterozygous status, yellow boxes indicate the homozygous status and no colour is for the wild type genotype.*

The segregation analysis was not able to identify which of the six mutations could be the driving one for the pathology. Therefore functional studies are needed to clearly correlate the mutations with the disease and to confirm their pathogenic role.

De novo mutations were also evaluated in the patient to expand the mutation analysis. We found that two novel non-synonymous mutations were present all in heterozygous status (*Table 5*).

Chr	Pos	Ref	Alt	AA Change	Mutation type	Gene name	MutationTaster	Polyphen2	SIFT
8	8748608	T	C	N654S	MISSENSE	MFHAS1	D	D	0.7
19	4512964	G	A	T322	SILENT	PLIN4	-	-	-
19	4512979	T	A	T317	SILENT	PLIN4	-	-	-
19	12015991	C	G	S260C	MISSENSE	ZNF69	-	-	0.2

Table 5 ***De novo mutation in affected sample****. Novel coding mutation not present in parents (de novo mutations) were reported. All the missense mutations were also evaluated by three different bioinformatics tools (MutationTaster, Polyphen2, SIFT). D=Damaging, B=Benign, N=Neutral, P=Probably damaging. SIFT: score<0.05 indicates damaging prediction otherwise benign prediction.*

The most relevant variation was the one occurred in the *MFHAS1* gene. The protein structure is characterized by three leucine zipper domains, and a leucine-rich tandem repeat, which are structural or functional elements for interactions among proteins related to the cell cycle.

Although this protein is not correlated to any cardiac disease, it is well known that the overexpression of its transcript is particularly associated to malignant fibrous histiocytoma (MFH) onset [130].

4. CONCLUSION AND FUTURE PERSPECTIVES

The very last part of this PhD dissertation summarises the results obtained in the research projects and suggests future perspectives for further investigations.

For the first project, we performed a target sequencing of 158-selected gene in a cohort of 91 Brugada patients. Since the 30% of BrS cases are associated to *SCN5A* mutations, the remaining 70% are still without a genetic explanation. Therefore we accurately selected a set of 158 genes, mainly belonging to ion channel category, previously associated to arrhythmogenic diseases. We collected and sequenced, for this gene panel set, 91 BrS samples with no disease-causing mutations in the *SCN5A* gene.

Two important results have been achieved from this research project:

1) We tested and developed an automated pipeline for NGS data analysis involving gene panel sequencing project. In particular two main aspects are notable: (1) the use of completely free available tools for each analysis module and (2) the sensitivity of the variant call system which was confirmed by other techniques (Sanger sequencing).
2) We found a set of mutated genes previously not associated to BrS thanks to a statistical validation with 1000 Genomes project data.

In the next future we will implement the pipeline as a free-available package able to perform the complete bioinformatics analysis from raw reads to variant calls and functional annotations. Moreover, we will perform functional analysis of few selected mutations of candidate genes to evaluate the possible damage caused by the mutations. Furthermore we will plan to test novel candidate genes coming from this analysis on a larger dataset in order to better estimate the genetic risk of BrS.

The second study was applied to the whole-exome sequencing of a family trio where the child is affected by an unclear cardiac disease with an atypical electrocardiogram pattern. Because of the unclear diagnosis the use of whole-exome sequencing technology was the optimal and cost-effective solution to widely screening all the coding regions of the genome in a shot. Thanks to the approach of trio sequencing, the candidate variants were filtered by subtracting the germline mutations found in the parents samples. Moreover, because of the severe episodes of cardiac arrest, we focused our attention on the homozygous mutation in the affected sample. We reached three main results from this project:

1) We developed a NGS analysis pipeline for trio family samples. We focused our effort to validate the accuracy of the variant calls and the homogeneity of the callable loci between the samples by comparing the read depth.
2) We found a set of six candidate genes, which presented homozygous mutations in the affected sample coming from the heterozygous genotype of the parents. By looking at the biological process, the *TRDN* and the *UNC45A* were selected for segregation analysis. By segregation analysis with other unaffected individual in the family tree, we were not able to identify the causative variants.
3) We also found *de novo* mutations, which could influence the disease effect. In detail, we found a potential damaging mutation in *MFHAS1* gene that is related to malignant fibrous histiocytoma (MFH) onset and apparently is not related to a cardiac disease. This type of mutation can be catalogued in term of incidental finding result.

Further investigation will need for this project to determine the disease-causing variation by performing functional study on *TRDN* and *UNC45A*

genes. In addition, as for target sequencing pipeline, we will build a package for trio analysis as plug-in in a widely NGS data analysis pipeline.

5. REFERENCES

1. Margulies M, Egholm M, Altman WE, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–80. doi: 10.1038/nature03959

2. Shendure J, Porreca GJ, Reppas NB, et al. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–32. doi: 10.1126/science.1117389

3. Mamanova L, Coffey AJ, Scott CE, et al. (2010) Target-enrichment strategies for next-generation sequencing. Nat Methods 7:111–118. doi: 10.1038/nmeth0610-479c

4. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11:31–46. doi: 10.1038/nrg2626

5. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141. doi: 10.1016/j.tig.2007.12.007

6. Kawashima E, Farinelli L, Mayer P (1998) Method of Nucleic Acid Amplification.

7. Cock PJA, Fields CJ, Goto N, et al. (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38:1767–71. doi: 10.1093/nar/gkp1137

8. Parimoo S, Patanjali SR, Shukla H, et al. (1991) cDNA selection: efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. Proc Natl Acad Sci U S A 88:9623–9627. doi: 10.1073/pnas.88.21.9623

9. Lovett M, Kere J, Hinton LM (1991) Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. Proc Natl Acad Sci U S A 88:9628–9632. doi: 10.1073/pnas.88.21.9628

10. Gnirke A, Melnikov A, Maguire J, et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27:182–189. doi: 10.1038/nbt.1523

11. Gonzaga-Jauregui C, Lupski JR, Gibbs RA (2012) Human genome sequencing in health and disease. Annu Rev Med 63:35–61. doi: 10.1146/annurev-med-051010-162644

12. Schadt EE, Linderman MD, Sorenson J, et al. (2010) Computational solutions to large-scale data management and analysis. Nat Rev Genet 11:647–57. doi: 10.1038/nrg2857

13. Ng PC, Levy S, Huang J, et al. (2008) Genetic variation in an individual human exome. PLoS Genet 4:e1000160. doi: 10.1371/journal.pgen.1000160

14. Ng SB, Turner EH, Robertson PD, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272–6. doi: 10.1038/nature08250

15. Robinson PN, Krawitz P, Mundlos S (2011) Strategies for exome and genome sequence data analysis in disease-gene discovery projects. Clin Genet 80:127–132. doi: 10.1111/j.1399-0004.2011.01713.x

16. Pabinger S, Dander A, Fischer M, et al. (2013) A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. doi: 10.1093/bib/bbs086

17. Coutant S, Cabot C, Lefebvre A, et al. (2012) EVA: Exome Variation Analyzer, an efficient and versatile tool for filtering strategies in medical genomics. BMC Bioinformatics 13 Suppl 1:S9. doi: 10.1186/1471-2105-13-S14-S9

18. Pope BJ, Nguyen-Dumont T, Odefrey F, et al. (2013) FAVR (Filtering and Annotation of Variants that are Rare): methods to facilitate the analysis of rare germline genetic variants from massively parallel sequencing datasets. BMC Bioinformatics 14:65. doi: 10.1186/1471-2105-14-65

19. Asmann YW, Middha S, Hossain A, et al. (2012) TREAT: a bioinformatics tool for variant annotations and visualizations in targeted and exome sequencing data. Bioinformatics 28:277–8. doi: 10.1093/bioinformatics/btr612

20. Planet E, Attolini CS-O, Reina O, et al. (2012) htSeqTools: high-throughput sequencing quality control, processing and visualization in R. Bioinformatics 28:589–90. doi: 10.1093/bioinformatics/btr700

21. Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11:485. doi: 10.1186/1471-2105-11-485

22. Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12:443–451. doi: 10.1038/nrg2986

23. Raney BJ, Cline MS, Rosenbloom KR, et al. (2011) ENCODE whole-genome data in the UCSC genome browser (2011 update). Nucleic Acids Res 39:D871–D875. doi: 10.1093/nar/gkp961

24. Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11:473–83. doi: 10.1093/bib/bbq015

25. Willis J, Adams MD, Yu X, et al. (2012) How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? BioData Min 5:6. doi: 10.1186/1756-0381-5-6

26. Altschul SF, Madden TL, Schäffer AA, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–402.

27. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–8. doi: 10.1101/gr.078212.108

28. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. Bioinformatics 24:713–714. doi: 10.1093/bioinformatics/btn025

29. Jiang H, Wong WH (2008) SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics 24:2395–2396. doi: 10.1093/bioinformatics/btn429

30. Burrows M, Wheeler DJ (1994) A block-sorting lossless data compression algorithm. Syst Res Research R:24. doi: 10.1.1.37.6774

31. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. doi: 10.1186/gb-2009-10-3-r25

32. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–60. doi: 10.1093/bioinformatics/btp324

33. Li R, Yu C, Li Y, et al. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966–1967. doi: 10.1093/bioinformatics/btp336

34. Li H, Handsaker B, Wysoker A, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–9. doi: 10.1093/bioinformatics/btp352

35. Handsaker RE, Korn JM, Nemesh J, McCarroll SA (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet 43:269–276. doi: 10.1038/ng.768

36. Hedges DJ, Burges D, Powell E, et al. (2009) Exome sequencing of a multigenerational human pedigree. PLoS One 4:e8232. doi: 10.1371/annotation/b0fe9dd5-16e1-4b50-b590-263518fbd5eb

37. McKenna A, Hanna M, Banks E, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–303. doi: 10.1101/gr.107524.110

38. Danecek P, Auton A, Abecasis G, et al. (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158. doi: 10.1093/bioinformatics/btr330

39. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38:e164. doi: 10.1093/nar/gkq603

40. Cingolani P, Platts A, Wang LL, et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6:80–92. doi: 10.4161/fly.19695

41. Kitts A, Sherry S (2011) The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation. NCBI Handb.

42. Matullo G, Di Gaetano C, Guarrera S (2013) Next generation sequencing and rare genetic variants: from human population studies

to medical genetics. Environ Mol Mutagen 54:518–32. doi: 10.1002/em.21799

43. Shendure J (2011) Next-generation human genetics. Genome Biol 12:408. doi: 10.1186/gb-2011-12-9-408

44. Nielsen MW, Holst AG, Olesen S-P, Olesen MS (2013) The genetic component of Brugada syndrome. Front Physiol 4:179. doi: 10.3389/fphys.2013.00179

45. Weeke P, Parvez B, Blair M, et al. (2013) Candidate gene approach to identifying rare genetic variants associated with lone atrial fibrillation. Heart Rhythm. doi: 10.1016/j.hrthm.2013.10.025

46. Bezzina CR, Barc J, Mizusawa Y, et al. (2013) Common variants at SCN5A-SCN10A and HEY2 are associated with Brugada syndrome, a rare disease with high risk of sudden cardiac death. Nat Genet. doi: 10.1038/ng.2712

47. Marian AJ, Belmont J (2011) Strategic approaches to unraveling genetic causes of cardiovascular diseases. Circ Res 108:1252–69. doi: 10.1161/CIRCRESAHA.110.236067

48. Thusberg J, Vihinen M (2009) Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat 30:703–14. doi: 10.1002/humu.20938

49. Ng SB, Buckingham KJ, Lee C, et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42:30–35. doi: 10.1038/ng.499

50. Gilissen C, Arts HH, Hoischen A, et al. (2010) Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. Am J Hum Genet 87:418–423. doi: 10.1016/j.ajhg.2010.08.004

51. Wang JL, Yang X, Xia K, et al. (2010) TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. Brain 133:3510–3518. doi: 10.1093/brain/awq323

52. Stitziel NO, Kiezun A, Sunyaev S (2011) Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol 12:227. doi: 10.1186/gb-2011-12-9-227

53. Castellana S, Mazza T (2013) Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools. Br. Bioinform

54. Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11:863–74. doi: 10.1101/gr.176601

55. Miller MP, Kumar S (2001) Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet 10:2319–2328. doi: 10.1093/hmg/10.21.2319

56. Davydov E V, Goode DL, Sirota M, et al. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6:e1001025. doi: 10.1371/journal.pcbi.1001025

57. Felsenstein J, Churchill GA (1996) A Hidden Markov Model approach to variation among sites in rate of evolution. Mol Biol Evol 13:93–104.

58. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–4.

59. Adzhubei IA, Schmidt S, Peshkin L, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–9. doi: 10.1038/nmeth0410-248

60. Liu X, Jian X, Boerwinkle E (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 32:894–9. doi: 10.1002/humu.21517

61. Del Rosario M, Brunner HG, de Ligt J, et al. (2012) Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. N Engl J Med 121003140044006. doi: 10.1056/NEJMoa1206524

62. Lupski JR, Reid JG, Gonzaga-Jauregui C, et al. (2010) Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N Engl J Med 362:1181–1191. doi: 10.1056/NEJMoa0908094

63. Sikkema-Raddatz B, Johansson LF, de Boer EN, et al. (2013) Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. Hum Mutat 34:1035–42. doi: 10.1002/humu.22332

64. Markello TC, Boerkoel CF, Groden C, et al. (2012) The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. Genet Med 14:51–59. doi: 10.1038/gim.0b013e318232a005

65. Coonrod EM, Margraf RL, Voelkerding K V (2012) Translating exome sequencing from research to clinical diagnostics. Clin Chem Lab Med 50:1161–8. doi: 10.1515/cclm-2011-0841

66. Brugada P, Brugada J (1992) Right bundle branch block, persistent ST segment elevation and sudden cardiac death: a distinct clinical and electrocardiographic syndrome. A multicenter report. J Am Coll Cardiol 20:1391–6.

67. Berne P, Brugada J (2012) Brugada syndrome 2012. Circ J 76:1563–71.

68. Chen Q, Kirsch GE, Zhang D, et al. (1998) Genetic basis and molecular mechanism for idiopathic ventricular fibrillation. Nature 392:293–6. doi: 10.1038/32675

69. Antzelevitch C, Pollevick GD, Cordeiro JM, et al. (2007) Loss-of-function mutations in the cardiac calcium channel underlie a new clinical entity characterized by ST-segment elevation, short QT intervals, and sudden cardiac death. Circulation 115:442–9. doi: 10.1161/CIRCULATIONAHA.106.668392

70. Risgaard B, Jabbari R, Refsgaard L, et al. (2013) High prevalence of genetic variants previously associated with Brugada syndrome in new exome data. Clin Genet 84:489–95. doi: 10.1111/cge.12126

71. Nielsen JB, Olesen MS, Haunsø S, et al. (2012) High prevalence of genetic variants previously associated with LQT syndrome in new exome data. Eur J Hum Genet 20:905–908. doi: 10.1038/ejhg.2012.23

72. Andreasen C, Nielsen JB, Refsgaard L, et al. (2013) New population-based exome data are questioning the pathogenicity of previously cardiomyopathy-associated genetic variants. Eur J Hum Genet 21:918–928. doi: 10.1038/ejhg.2012.283

73. Brugada R, Brugada J, Antzelevitch C, et al. (2000) Sodium channel blockers identify risk for sudden death in patients with ST-segment elevation and right bundle branch block but structurally normal hearts. Circulation 101:510–515.

74. Epstein AE, DiMarco JP, Ellenbogen KA, et al. (2008) ACC/AHA/HRS 2008 Guidelines for Device-Based Therapy of Cardiac Rhythm Abnormalities: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Revise the ACC/AHA/NASPE 2002 Guideline. Circulation 117:e350–408. doi: 10.1161/CIRCUALTIONAHA.108.189742

75. Abriel H (2007) Cardiac sodium channel Nav1.5 and its associated proteins. Arch Mal Coeur Vaiss 100:787–93.

76. Gaborit N, Wichter T, Varro A, et al. (2009) Transcriptional profiling of ion channel genes in Brugada syndrome and other right ventricular arrhythmogenic diseases. Eur Heart J 30:487–96. doi: 10.1093/eurheartj/ehn520

77. Chambers JC, Zhao J, Terracciano CMN, et al. (2010) Genetic variation in SCN10A influences cardiac conduction. Nat Genet 42:149–52.

78. Wilde AAM, Bezzina CR (2005) Genetics of cardiac arrhythmias. Heart 91:1352–8.

79. Marangoni S, Di Resta C, Rocchetti M, et al. (2011) A Brugada syndrome mutation (p.S216L) and its modulation by p.H558R polymorphism: standard and dynamic characterization. Cardiovasc Res 91:606–16. doi: 10.1093/cvr/cvr142

80. Ellinor PT, Lunetta KL, Glazer NL, et al. (2010) Common variants in KCNN3 are associated with lone atrial fibrillation. Nat Genet 42:240–4.

81. Summers KM, Bokil NJ, Lu FT, et al. (2010) Mutations at KCNQ1 and an unknown locus cause long QT syndrome in a large Australian family: implications for genetic testing. Am J Med Genet A 152A:613–21.

82. Lundby A, Olesen S-P (2006) KCNE3 is an inhibitory subunit of the Kv4.3 potassium channel. Biochem Biophys Res Commun 346:958–67.

83. Calloe K, Cordeiro JM, Di Diego JM, et al. (2009) A transient outward potassium current activator recapitulates the electrocardiographic manifestations of Brugada syndrome. Cardiovasc Res 81:686–94.

84. Delpón E, Cordeiro JM, Núñez L, et al. (2008) Functional effects of KCNE3 mutation and its role in the development of Brugada syndrome. Circ Arrhythm Electrophysiol 1:209–18.

85. Valdivia CR, Medeiros-Domingo A, Ye B, et al. (2010) Loss-of-function mutation of the SCN3B-encoded sodium channel {beta}3 subunit associated with a case of idiopathic ventricular fibrillation. Cardiovasc Res 86:392–400.

86. Bibevski S, Dunlap ME (2011) Evidence for impaired vagus nerve activity in heart failure. Heart Fail Rev 16:129–35.

87. Herring N, Paterson DJ (2009) Neuromodulators of peripheral cardiac sympatho-vagal balance. Exp Physiol 94:46–53.

88. Gergs U, Baumann M, Böckler A, et al. (2010) Cardiac overexpression of the human 5-HT4 receptor in mice. Am J Physiol Heart Circ Physiol 299:H788–98.

89. Ackerman MJ, Mohler PJ (2010) Defining a new paradigm for human arrhythmia syndromes: phenotypic manifestations of gene mutations in ion channel- and transporter-associated proteins. Circ Res 107:457–65.

90. Abriel H (2010) Cardiac sodium channel Na(v)1.5 and interacting proteins: Physiology and pathophysiology. J Mol Cell Cardiol 48:2–11.

91. Hashemi SM, Hund TJ, Mohler PJ (2009) Cardiac ankyrins in health and disease. J Mol Cell Cardiol 47:203–9.

92. Holm H, Gudbjartsson DF, Arnar DO, et al. (2010) Several common variants modulate heart rate, PR interval and QRS duration. Nat Genet 42:117–22.

93. Rizzi N, Liu N, Napolitano C, et al. (2008) Unexpected structural and functional consequences of the R33Q homozygous mutation in cardiac calsequestrin: a complex arrhythmogenic cascade in a knock in mouse model. Circ Res 103:298–306.

94. Bai C-X, Kurokawa J, Tamagawa M, et al. (2005) Nontranscriptional regulation of cardiac repolarization currents by testosterone. Circulation 112:1701–10.

95. Loussouarn G, Baró I (2010) Neural modulation of ion channels in cardiac arrhythmias: clinical implications and future investigations. Heart Rhythm 7:847–9.

96. Campuzano O, Beltrán-Alvarez P, Iglesias A, et al. (2010) Genetics and cardiac channelopathies. Genet Med 12:260–7.

97. Pfeufer A, Sanna S, Arking DE, et al. (2009) Common variants at ten loci modulate the QT interval duration in the QTSCD Study. Nat Genet 41:407–14.

98. Bezzina CR, Pazoki R, Bardai A, et al. (2010) Genome-wide association study identifies a susceptibility locus at 21q21 for ventricular fibrillation in acute myocardial infarction. Nat Genet 42:688–91.

99. Mazzone A, Strege PR, Tester DJ, et al. (2008) A mutation in telethonin alters Nav1.5 function. J Biol Chem 283:16537–44.

100. Knollmann BC, Roden DM (2008) A genetic framework for improving arrhythmia therapy. Nature 451:929–36.

101. Groenewegen WA, Firouzi M, Bezzina CR, et al. (2003) A cardiac sodium channel mutation cosegregates with a rare connexin40 genotype in familial atrial standstill. Circ Res 92:14–22.

102. Sommariva E, Pappone C, Martinelli Boneschi F, et al. (2013) Genetics can contribute to the prognosis of Brugada syndrome: a pilot model for risk stratification. Eur. J. Hum. Genet.

103. Sotoodehnia N, Li G, Johnson CO, et al. (2009) Genetic variation in angiotensin-converting enzyme-related pathways associated with sudden cardiac arrest risk. Heart Rhythm 6:1306–14.

104. Cingolani P, Patel VM, Coon M, et al. (2012) Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet 3:35. doi: 10.3389/fgene.2012.00035

105. Brunello L, Slabaugh JL, Radwanski PB, et al. (2013) Decreased RyR2 refractoriness determines myocardial synchronization of aberrant Ca2+ release in a genetic model of arrhythmia. Proc Natl Acad Sci U S A 110:10312–7.

106. Kim H, Cho Y, Park Y, et al. (2006) Underlying cardiomyopathy in patients with ST-segment elevation in the right precordial leads. Circ J 70:719–25.

107. Duthoit G, Fressart V, Hidden-Lucet F, et al. (2012) Brugada ECG pattern: a physiopathological prospective study based on clinical, electrophysiological, angiographic, and genetic findings. Front Physiol 3:474.

108. Remme CA (2013) Cardiac sodium channelopathy associated with SCN5A mutations: electrophysiological, molecular and genetic aspects. J. Physiol.

109. Bébarová M (2013) Arrhythmogenesis in Brugada syndrome: Impact and constrains of current concepts. Int J Cardiol. doi: 10.1016/j.ijcard.2012.12.019

110. Antzelevitch C, Pollevick GD, Cordeiro JM, et al. (2007) Loss-of-function mutations in the cardiac calcium channel underlie a new clinical entity characterized by ST-segment elevation, short QT intervals, and sudden cardiac death. Circulation 115:442–449. doi: 10.1161/CIRCULATIONAHA.106.668392

111. Burashnikov E, Pfeiffer R, Barajas-Martinez H, et al. (2010) Mutations in the cardiac L-type calcium channel associated with inherited J-wave syndromes and sudden cardiac death. Heart Rhythm 7:1872–82.

112. Mohler PJ, Splawski I, Napolitano C, et al. (2004) A cardiac arrhythmia syndrome caused by loss of ankyrin-B function. Proc Natl Acad Sci U S A 101:9137–42.

113. Mohler PJ, Le Scouarnec S, Denjoy I, et al. (2007) Defining the cellular phenotype of “ankyrin-B syndrome” variants: human ANK2 variants associated with clinical phenotypes display a spectrum of activities in cardiomyocytes. Circulation 115:432–41. doi: 10.1161/CIRCULATIONAHA.106.656512

114. Mohler PJ, Rivolta I, Napolitano C, et al. (2004) Nav1.5 E1053K mutation causing Brugada syndrome blocks binding to ankyrin-G and expression of Nav1.5 on the surface of cardiomyocytes. Proc Natl Acad Sci U S A 101:17533–8.

115. Okoshi K, Nakayama M, Yan X, et al. (2004) Neuregulins regulate cardiac parasympathetic activity: muscarinic modulation of beta-adrenergic activity in myocytes from mice with neuregulin-1 gene deletion. Circulation 110:713–7.

116. Miyazaki T, Mitamura H, Miyoshi S, et al. (1996) Autonomic and antiarrhythmic drug modulation of ST segment elevation in patients with Brugada syndrome. J Am Coll Cardiol 27:1061–70.

117. Kasanuki H, Ohnishi S, Ohtuka M, et al. (1997) Idiopathic ventricular fibrillation induced with vagal activity in patients without obvious heart disease. Circulation 95:2277–85.

118. Verkerk AO, Remme CA, Schumacher CA, et al. (2012) Functional Nav1.8 channels in intracardiac neurons: the link between SCN10A and cardiac electrophysiology. Circ Res 111:333–43.

119. Yang T, Atack TC, Stroud DM, et al. (2012) Blocking Scn10a channels in heart reduces late sodium current and is antiarrhythmic. Circ Res 111:322–32.

120. Frustaci A, Priori SG, Pieroni M, et al. (2005) Cardiac histological substrate in patients with clinical phenotype of Brugada syndrome. Circulation 112:3680–7.

121. Anselme F, Moubarak G, Savoure A, et al. (2013) Implantable cardioverter-defibrillators in lamin A/C mutation carriers with cardiac conduction disorders. Heart Rhythm null:

122. Carboni N, Mateddu A, Marrosu G, et al. (2013) Genetic and clinical characteristics of skeletal and cardiac muscle in patients with lamin A/C gene mutations. Muscle Nerve

123. Meune C, Van Berlo JH, Anselme F, et al. (2006) Primary prevention of sudden death in patients with lamin A/C gene mutations. N Engl J Med 354:209–10.

124. Kolder ICRM, Tanck MWT, Bezzina CR (2012) Common genetic variation modulating cardiac ECG parameters and susceptibility to sudden cardiac death. J Mol Cell Cardiol 52:620–9.

125. Lodder EM, Scicluna BP, Milano A, et al. (2012) Dissection of a quantitative trait locus for PR interval duration identifies Tnni3k as a novel modulator of cardiac conduction. PLoS Genet 8:e1003113.

126. Price MG, Landsverk ML, Barral JM, Epstein HF (2002) Two mammalian UNC-45 isoforms are related to distinct cytoskeletal and muscle-specific functions. J Cell Sci 115:4013–23.

127. Chadli A, Graham JD, Abel MG, et al. (2006) GCUNC-45 is a novel regulator for the progesterone receptor/hsp90 chaperoning pathway. Mol Cell Biol 26:1722–30. doi: 10.1128/MCB.26.5.1722-1730.2006

128. Melkani GC, Bodmer R, Ocorr K, Bernstein SI (2011) The UNC-45 chaperone is critical for establishing myosin-based myofibrillar organization and cardiac contractility in the Drosophila heart model. PLoS One 6:e22579. doi: 10.1371/journal.pone.0022579

129. Roux-Buisson N, Cacheux M, Fourest-Lieuvin A, et al. (2012) Absence of triadin, a protein of the calcium release complex, is responsible for cardiac arrhythmia with sudden death in human. Hum Mol Genet 21:2759–67. doi: 10.1093/hmg/dds104

130. Sakabe T, Shinomiya T, Mori T, et al. (1999) Identification of a novel gene, MASL1, within an amplicon at 8p23.1 detected in malignant fibrous histiocytomas by comparative genomic hybridization. Cancer Res 59:511–5.

131. Cifola I, Pietrelli A, Consolandi C, et al. (2013) Comprehensive genomic characterization of cutaneous malignant melanoma cell lines derived from metastatic lesions by whole-exome sequencing and SNP array profiling. PLoS One 8:e63597. doi: 10.1371/journal.pone.0063597

6. APPENDIX

6.1. Case 1: Target sequencing of 91 BrS patients

The tables showed below reported supplementary information about the case study 1 such as the patient description, the target genes used for panel construction and the overall mutations found in the mutation analysis classified by the category described in the section 2.3.3

6.1.1. BrS patients description

Description	N (%)
Number of patients	91
Follow-up (months, mean and st.dev.)	63.4 ± 21
Age (years, mean and st.dev.)	50.5 ± 13.2
Males	79 (87)
Family history for SCD	27 (29)
Spontaneous type I ECG	55 (60.4)
Asymptomatic	35 (38.4)
Syncope	25 (27.4)
MAE (SCD, documented VT/VF, appropriate ICD	8 (8.7)
Age at MAE (years, mean and st.dev.)	41 ± 12
ICD implanted	55 (60)

6.1.2. Genes panel for target sequencing

Gene Name	Protein description	Ensembl ID	Cathegory
SCN1A	sodium channel, voltage-gated, type I, alpha subunit	ENSG00000144285	Sodium Channel alpha-subunit
SCN2A	sodium channel, voltage-gated, type II, alpha subunit	ENSG00000136531	Sodium Channel alpha-subunit

Gene Name	Protein description	Ensembl ID	Cathegory
SCN3A	sodium channel, voltage-gated, type III, alpha subunit	ENSG00000153253	Sodium Channel alpha-subunit
SCN4A	sodium channel, voltage-gated, type IV, alpha subunit	ENSG00000007314	Sodium Channel alpha-subunit
SCN5A	sodium channel, voltage-gated, type V, alpha subunit	ENSG00000183873	Sodium Channel alpha-subunit
SCN7A	sodium channel, voltage-gated, type VII, alpha subunit	ENSG00000136546	Sodium Channel alpha-subunit
SCN8A	sodium channel, voltage-gated, type VIII, alpha subunit	ENSG00000196876	Sodium Channel alpha-subunit
SCN9A	sodium channel, voltage-gated, type IX, alpha subunit	ENSG00000169432	Sodium Channel alpha-subunit
SCN10A	sodium channel, voltage-gated, type X, alpha subunit	ENSG00000185313	Sodium Channel alpha-subunit
SCN11A	sodium channel, voltage-gated, type XI, alpha subunit	ENSG00000168356	Sodium Channel alpha-subunit
SCN1B	sodium channel, voltage-gated, type I, beta subunit	ENSG00000105711	Sodium Channel beta-subunit
SCN2B	sodium channel, voltage-gated, type II, beta subunit	ENSG00000149575	Sodium Channel beta-subunit
SCN3B	sodium channel, voltage-gated, type III, beta subunit	ENSG00000166257	Sodium Channel beta-subunit
SCN4B	sodium channel, voltage-gated, type IV, beta subunit	ENSG00000177098	Sodium Channel beta-subunit
HCN1	hyperpolarization activated cyclic nucleotide-gated potassium channel 1	ENSG00000164588	Potassium Channel alpha-subunit
HCN2	hyperpolarization activated cyclic nucleotide-gated potassium channel 2	ENSG00000099822	Potassium Channel alpha-subunit
HCN3	hyperpolarization activated cyclic nucleotide-gated potassium channel 3	ENSG00000143630	Potassium Channel alpha-subunit
HCN4	hyperpolarization activated cyclic nucleotide-gated potassium channel 4	ENSG00000138622	Potassium Channel alpha-subunit
KCND1	potassium voltage-gated channel, Shal-related subfamily, member 1	ENSG00000102057	Potassium Channel alpha-subunit
KCND2	potassium voltage-gated channel, Shal-related subfamily, member 2	ENSG00000184408	Potassium Channel alpha-subunit
KCND3	potassium voltage-gated channel, Shal-related subfamily, member 3	ENSG00000171385	Potassium Channel alpha-subunit
KCNE1	potassium voltage-gated channel, Isk-related family, member 1	ENSG00000180509	Potassium Channel alpha-subunit
KCNE2	potassium voltage-gated channel, Isk-related family, member 2	ENSG00000159197	Potassium Channel alpha-subunit
KCNE3	potassium voltage-gated channel, Isk-related family, member 3	ENSG00000175538	Potassium Channel alpha-subunit
KCNE5	KCNE1-like	ENSG00000176076	Potassium Channel alpha-subunit
KCNQ1	potassium voltage-gated channel, KQT-like subfamily, member 1	ENSG00000053918	Potassium Channel alpha-subunit
KCNQ2	potassium voltage-gated channel, KQT-like subfamily, member 2	ENSG00000075043	Potassium Channel alpha-subunit
KCNH2	potassium voltage-gated channel, subfamily H, member 2	ENSG00000055118	Potassium Channel alpha-subunit
KCNG1	potassium voltage-gated channel, subfamily G, member 1	ENSG00000026559	Potassium Channel alpha-subunit
KCNG2	potassium voltage-gated channel, subfamily G, member 2	ENSG00000178342;	Potassium Channel alpha-subunit
KCHIP2	Kv channel interacting protein 2	ENSG00000120049	Potassium Channel alpha-subunit
KCNIP3	Kv channel interacting protein 3, calsenilin	ENSG00000115041	Potassium Channel alpha-subunit

Gene Name	Protein description	Ensembl ID	Cathegory
KCNIP1	Kv channel interacting protein 1	ENSG00000182132	Potassium Channel alpha-subunit
KCNK5	potassium channel, subfamily K, member 5	ENSG00000164626	Potassium Channel alpha-subunit
KCNK3	potassium channel, subfamily K, member 3	ENSG00000171303	Potassium Channel alpha-subunit
KCNK1	potassium channel, subfamily K, member 1	ENSG00000135750	Potassium Channel alpha-subunit
KCNK6	potassium channel, subfamily K, member 6	ENSG00000099337	Potassium Channel alpha-subunit
KCNK2	potassium channel, subfamily K, member 2	ENSG00000082482	Potassium Channel alpha-subunit
KCNK4	potassium channel, subfamily K, member 4	ENSG00000182450	Potassium Channel alpha-subunit
KCNK12	potassium channel, subfamily K, member 12	ENSG00000184261	Potassium Channel alpha-subunit
KCNK17	potassium channel, subfamily K, member 17	ENSG00000124780	Potassium Channel alpha-subunit
KCNA7	potassium voltage-gated channel, shaker-related subfamily, member 7	ENSG00000104848	Potassium Channel alpha-subunit
KCNA2	potassium voltage-gated channel, shaker-related subfamily, member 2	ENSG00000177301	Potassium Channel alpha-subunit
KCNA4	potassium voltage-gated channel, shaker-related subfamily, member 4	ENSG00000182255	Potassium Channel alpha-subunit
KCNA5	potassium voltage-gated channel, shaker-related subfamily, member 5	ENSG00000130037	Potassium Channel alpha-subunit
KCNA6	potassium voltage-gated channel, shaker-related subfamily, member 6	ENSG00000151079	Potassium Channel alpha-subunit
KCNB1	potassium voltage-gated channel, Shab-related subfamily, member 1	ENSG00000158445	Potassium Channel alpha-subunit
KCNB2	potassium voltage-gated channel, Shab-related subfamily, member 2	ENSG00000182674	Potassium Channel alpha-subunit
KCNC3	potassium voltage-gated channel, Shaw-related subfamily, member 3	ENSG00000131398	Potassium Channel alpha-subunit
KCNF1	potassium voltage-gated channel, subfamily F, member 1	ENSG00000162975	Potassium Channel alpha-subunit
KCNJ2	potassium inwardly-rectifying channel, subfamily J, member 2	ENSG00000123700	Potassium Channel alpha-subunit
KCNJ3	potassium inwardly-rectifying channel, subfamily J, member 3	ENSG00000162989	Potassium Channel alpha-subunit
KCNJ4	potassium inwardly-rectifying channel, subfamily J, member 4	ENSG00000168135	Potassium Channel alpha-subunit
KCNJ5	potassium inwardly-rectifying channel, subfamily J, member 5	ENSG00000120457	Potassium Channel alpha-subunit
KCNJ8	potassium inwardly-rectifying channel, subfamily J, member 8	ENSG00000121361	Potassium Channel alpha-subunit
KCNJ11	potassium inwardly-rectifying channel, subfamily J, member 11	ENSG00000187486	Potassium Channel alpha-subunit
KCNJ6	potassium inwardly-rectifying channel, subfamily J, member 6	ENSG00000157542	Potassium Channel alpha-subunit
KCNJ15	potassium inwardly-rectifying channel, subfamily J,	ENSG00000157551	Potassium Channel alpha-

Gene Name	Protein description	Ensembl ID	Cathegory
	member 15		subunit
KCNJ12	potassium inwardly-rectifying channel, subfamily J, member 12	ENSG00000184185	Potassium Channel alpha-subunit
KCNN3	potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3	ENSG00000143603	Potassium Channel alpha-subunit
KCNS3	potassium voltage-gated channel, delayed-rectifier, subfamily S, member 3	ENSG00000170745	Potassium Channel alpha-subunit
KCNAB1	potassium voltage-gated channel, shaker-related subfamily, beta member 1	ENSG00000169282	Potassium Channel beta-subunit
KCNAB2	potassium voltage-gated channel, shaker-related subfamily, beta member 2	ENSG00000069424	Potassium Channel beta-subunit
KCNAB3	potassium voltage-gated channel, shaker-related subfamily, beta member 3	ENSG00000170049	Potassium Channel beta-subunit
CLCN2	chloride channel, voltage-sensitive 2	ENSG00000114859	Chloride Channel
CLCN3	chloride channel, voltage-sensitive 3	ENSG00000109572	Chloride Channel
CLCN6	chloride channel, voltage-sensitive 6	ENSG00000011021	Chloride Channel
CLCN7	chloride channel, voltage-sensitive 7	ENSG00000103249	Chloride Channel
CACNA1A	calcium channel, voltage-dependent, P/Q type, alpha 1A subunit	ENSG00000141837	Calcium Channel
CACNA1C	calcium channel, voltage-dependent, L type, alpha 1C subunit	ENSG00000151067	Calcium Channel
CACNA1D	calcium channel, voltage-dependent, L type, alpha 1D subunit	ENSG00000157388	Calcium Channel
CACNA1E	calcium channel, voltage-dependent, R type, alpha 1E subunit	ENSG00000198216	Calcium Channel
CACNA1G	calcium channel, voltage-dependent, T type, alpha 1G subunit	ENSG00000006283	Calcium Channel
CACNA1H	calcium channel, voltage-dependent, T type, alpha 1H subunit	ENSG00000196557	Calcium Channel
CACNA2D1	calcium channel, voltage-dependent, alpha 2/delta subunit 1	ENSG00000153956	Calcium Channel
CACNA2D2	calcium channel, voltage-dependent, alpha 2/delta subunit 2	ENSG00000007402	Calcium Channel
CACNB2	calcium channel, voltage-dependent, beta 2 subunit	ENSG00000165995	Calcium Channel
ADRB1	adrenoceptor beta 1	ENSG00000043591	Ligand-dependent Gated Channel
ADRB2	adrenoceptor beta 2	ENSG00000169252	Ligand-dependent Gated Channel
CHRM2	cholinergic receptor, muscarinic 2	ENSG00000181072	Ligand-dependent Gated Channel
CHRNA3	cholinergic receptor, nicotinic, alpha 3	ENSG00000080644	Ligand-dependent Gated Channel
CHRNB2	cholinergic receptor, nicotinic, beta 2	ENSG00000160716	Ligand-dependent Gated Channel
CNP	natriuretic peptide C	ENSG00000163273	Ligand-dependent Gated Channel
NGF	nerve growth factor	ENSG00000134259	Ligand-dependent Gated Channel
NRG1	neuregulin 1	ENSG00000157168	Ligand-dependent Gated Channel

Gene Name	Protein description	Ensembl ID	Cathegory
SLC6A4	solute carrier family 6	ENSG00000108576	Ligand-dependent Gated Channel
ANK1	ankyrin 1	ENSG00000029534	Ankyrins
ANK2	ankyrin 2	ENSG00000145362	Ankyrins
ANK3	ankyrin 3	ENSG00000151150	Ankyrins
ATP1A1	ATPase, Na+/K+ transporting, alpha 1 polypeptide	ENSG00000163399	Ion exchangers
ATP1A3	ATPase, Na+/K+ transporting, alpha 3 polypeptide	ENSG00000105409	Ion exchangers
ATP1B1	ATPase, Na+/K+ transporting, beta 1 polypeptide	ENSG00000143153	Ion exchangers
ATP2A2	ATPase, Ca++ transporting, cardiac muscle, slow twitch 2	ENSG00000174437	Ion exchangers
ATP2A3	ATPase, Ca++ transporting, ubiquitous	ENSG00000074370	Ion exchangers
ATP2B1	ATPase, Ca++ transporting, plasma membrane 1	ENSG00000070961	Ion exchangers
ATP2B4	ATPase, Ca++ transporting, plasma membrane 4	ENSG00000058668	Ion exchangers
SLC8A1 (NCX1)	solute carrier family 8 (sodium/calcium exchanger), member 1	ENSG00000183023	Ion exchangers
CALM1	calmodulin 1	ENSG00000198668	scaffold and calcium binding proteins
CALM3	calmodulin 3	ENSG00000160014	scaffold and calcium binding proteins
CALR	calreticulin	ENSG00000179218	scaffold and calcium binding proteins
CASQ1	calsequestrin 1	ENSG00000143318	scaffold and calcium binding proteins
CASQ2	calsequestrin 2	ENSG00000118729	scaffold and calcium binding proteins
CAV1	caveolin 2	ENSG00000105974	scaffold and calcium binding proteins
CAV2	caveolin 2	ENSG00000105971	scaffold and calcium binding proteins
CAV3	caveolin 3	ENSG00000182533	scaffold and calcium binding proteins
DSC2	desmocollin 2	ENSG00000134755	scaffold and calcium binding proteins
DSG2	desmoglein 2	ENSG00000046604	scaffold and calcium binding proteins
DSP	desmoplakin	ENSG00000096696	scaffold and calcium binding proteins
JUP	junction plakoglobin	ENSG00000173801	scaffold and calcium binding proteins
SNTA1	syntrophin, alpha 1	ENSG00000101400	scaffold and calcium binding proteins
SNTB1	syntrophin, beta 1	ENSG00000172164	scaffold and calcium binding proteins
SNTB2	syntrophin, beta 2	ENSG00000168807	scaffold and calcium binding proteins
SNTG2	syntrophin, gamma 2	ENSG00000172554	scaffold and calcium binding proteins
GJA1	gap junction protein, alpha 1	ENSG00000152661	gap junctions
GJA5	gap junction protein, alpha 5	ENSG00000143140	gap junctions

Gene Name	Protein description	Ensembl ID	Cathegory
GJA7 (GJC1)	gap junction protein, gamma 1	ENSG00000182963	gap junctions
ACTC1	actin, alpha, cardiac muscle 1	ENSG00000159251	structural proteins
ACTN2	actinin, alpha 2	ENSG00000077522	structural proteins
LMNA	lamin A/C	ENSG00000160789	structural proteins
MYH6	myosin, heavy chain 6	ENSG00000197616	structural proteins
MYH7	myosin, heavy chain 7	ENSG00000092054	structural proteins
TELT	titin-cap	ENSG00000173991	structural proteins
DES	desmin	ENSG00000175084	structural proteins
ACE	angiotensin I converting enzyme	ENSG00000159640	Other
AGTR1	angiotensin II receptor, type 1	ENSG00000144891	Other
AGTR2	angiotensin II receptor, type 2	ENSG00000180772	Other
AKAP9	A kinase (PRKA) anchor protein (yotiao) 9	ENSG00000127914	Other
AR	androgen receptor	ENSG00000169083	Other
CNOT1	CCR4-NOT transcription complex, subunit 1	ENSG00000125107	Other
CXADR	coxsackie virus and adenovirus receptor	ENSG00000154639	Other
FGF12	fibroblast growth factor 12	ENSG00000114279	Other
GPD1	glycerol-3-phosphate dehydrogenase 1	ENSG00000167588	Other
GPD1L	glycerol-3-phosphate dehydrogenase 1-like	ENSG00000152642	Other
ITPR1	inositol 1,4,5-trisphosphate receptor, type 1	ENSG00000150995	Other
KNG1	kininogen 1	ENSG00000113889	Other
LITAF	lipopolysaccharide-induced TNF factor	ENSG00000189067	Other
MDR1	ATP-binding cassette, sub-family B (MDR/TAP), member 1	ENSG00000085563	Other
NDRG4	NDRG family member 4	ENSG00000103034	Other
NEDD 4	neural precursor cell expressed, developmentally down-regulated 4, E3 ubiquitin protein ligase	ENSG00000069869	Other
NEDD4L	neural precursor cell expressed, developmentally down-regulated 4-like, E3 ubiquitin protein ligase	ENSG00000049759	Other
NOS1AP	nitric oxide synthase 1 (neuronal) adaptor protein	ENSG00000198929	Other
NPPA	natriuretic peptide A	ENSG00000175206	Other
NPPB	natriuretic peptide B	ENSG00000120937	Other
NRG3	neuregulin 3	ENSG00000185737	Other
PKP2	plakophilin 2	ENSG00000057294	Other
PLN	phospholamban	ENSG00000198523	Other
PPP3CA	protein phosphatase 3, catalytic subunit, alpha isozyme	ENSG00000138814	Other
PSEN1	presenilin 1	ENSG00000080815	Other
PSEN2	presenilin 2	ENSG00000143801	Other
PTPH1	protein tyrosine phosphatase, non-receptor type 3	ENSG00000070159	Other
RANGRF	RAN guanine nucleotide release factor	ENSG00000160789	Other
RYR2	ryanodine receptor 2	ENSG00000198626	Other
PIAS3	protein inhibitor of activated STAT, 3	ENSG00000131788	Potassium Channel alpha-subunit
TBX5	T-box 5	ENSG00000089225	Other
TGFB3	transforming growth factor, beta 3	ENSG00000119699	Other

Gene Name	Protein description	Ensembl ID	Cathegory
TMEM 43	Transmembrane protein 43	ENSG00000170876	Other
YWHAH	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, eta polypeptide	ENSG00000128245	Other

6.1.3. List of NS-SNVs

Sample	Chr	Pos	Ref	Alt	AA Chage	Gene Name
Brugada_73683	10	61819149	C	T	C1682Y	ANK3
Brugada_061826	10	61822964	A	G	V146A	ANK3
Brugada_101728	10	61831718	T	A	N2974I	ANK3
Brugada_84973	10	61835330	G	C	T1770R	ANK3
Brugada_8467	1	111147277	C	T	R43Q	KCNA2
Brugada_061826	1	145584024	C	T	P384S	PIAS3
Brugada_73051	1	145584179	A	G	K409E	PIAS3
Brugada_53353	1	145584276	C	T	P441L	PIAS3
Brugada_061826	1	155257090	C	T	A535V	HCN3
Brugada_74659	1	156085059	A	G	K117R	LMNA
Brugada_8579	1	156085059	A	G	K117R	LMNA
Brugada_09172	1	160163637	T	G	D149E	CASQ1
Brugada_10578	1	162257218	A	G	K88E	NOS1AP
Brugada_93200	1	181689358	C	T	R197W	CACNA1E
Brugada_71746	1	181690939	C	T	R275C	CACNA1E
Brugada_84973	1	203696604	G	A	D1036N	ATP2B4
Brugada_81290	1	237617717	C	A	A424E	RYR2
Brugada_101196	1	237711861	C	T	R1011W	RYR2
Brugada_93200	1	237811912	C	T	T2488M	RYR2
Brugada_93793	12	50501160	T	C	M173T	GPD1
Brugada_74487	12	50501193	G	A	G184E	GPD1
Brugada_091856	12	50501362	G	A	V186M	GPD1
Brugada_92497	12	52100339	G	A	R290H	SCN8A
Brugada_73980	14	23853673	C	T	R1848H	MYH6
Brugada_72060	14	23863351	G	A	R871C	MYH6
Brugada_74200	14	23886163	C	T	G1520R	MYH7
Brugada_103377	14	76425594	A	G	V392A	TGFB3
Brugada_7376	15	56125285	A	C	D1163E	NEDD4
Brugada_73051	15	78893644	C	T	S254N	CHRNA3
Brugada_6582	15	78921890	C	T	V253I	CHRNB4
Brugada_72060	16	1251967	G	A	R506Q	CACNA1H
Brugada_73051	16	1254406	T	C	M800T	CACNA1H
Brugada_63555	16	1259233	C	T	R1189C	CACNA1H
Brugada_5258	16	1260047	C	T	R1253C	CACNA1H
Brugada_84023	16	1260902	C	T	S1385L	CACNA1H
Brugada_062107	16	1510492	A	G	F116S	CLCN7
Brugada_93793	16	58589261	T	C	M358V	CNOT1
Brugada_63620	17	28543101	C	G	G324A	SLC6A4
Brugada_08404	17	3844281	T	C	N695S	ATP2A3
Brugada_74159	17	48703921	C	G	P2109A	CACNA1G
Brugada_072337	17	62026786	C	T	E986K	SCN4A

Sample	Chr	Pos	Ref	Alt	AA Chage	Gene Name
Brugada_74200	17	62029217	G	A	A807V	SCN4A
Brugada_95342	18	28672091	T	C	I109M	DSC2
Brugada_93061	18	29122693	G	A	A738T	DSG2
Brugada_09172	18	29122733	C	T	T751I	DSG2
Brugada_103905	19	13323494	G	A	P2001S	CACNA1A
Brugada_83255	19	49573517	C	T	V392I	KCNA7
Brugada_092178	20	62038356	C	T	A723T	KCNQ2
Brugada_73051	2	1094040	G	A	G90E	SNTG2
Brugada_72939	2	1251167	G	C	K192N	SNTG2
Brugada_71603	21	35742967	A	G	I64V	KCNE2
Brugada_103969	21	39668946	G	A	Splicing	KCNJ15
Brugada_81290	2	166172087	G	A	S497N	SCN2A
Brugada_63620	2	166246047	A	G	I1911V	SCN2A
Brugada_62293	2	166848701	A	G	V1667A	SCN1A
Brugada_61687	2	166892998	C	T	A969T	SCN1A
Brugada_95342	2	167060467	A	G	V1580A	SCN9A
Brugada_73370	2	167298099	T	C	E655G	SCN7A
Brugada_62396	2	18113554	G	T	D427Y	KCNS3
Brugada_95342	2	40404894	T	G	K637T	SLC8A1
Brugada_72939	3	155861029	C	T	S21F	KCNAB1
Brugada_101728	3	38648255	C	T	D349N	SCN5A
Brugada_72939	3	38768095	A	G	Splicing	SCN10A
Brugada_6583	3	38805052	G	A	S212L	SCN10A
Brugada_71603	3	38888249	T	C	Q1733R	SCN11A
Brugada_08404	3	38968422	C	G	E163D	SCN11A
Brugada_72900	3	4741557	G	A	E1466K	ITPR1
Brugada_53353	3	4810254	G	A	A1866T	ITPR1
Brugada_71563	3	4842201	C	T	R2279W	ITPR1
Brugada_8579	3	50402321	T	C	T1030A	CACNA2D2
Brugada_92497	3	50410496	C	T	D669N	CACNA2D2
Brugada_061826	3	50413411	C	T	D517N	CACNA2D2
Brugada_91596	4	114239689	A	T	K147M	ANK2
Brugada_103969	4	114264197	G	A	V1296I	ANK2
Brugada_73980	4	114276621	A	C	K2250Q	ANK2
Brugada_73370	4	114276847	G	A	S2325N	ANK2
Brugada_93061	4	114284568	G	A	D1517N	ANK2
Brugada_103905	4	170639007	G	A	V811I	CLCN3
Brugada_62293	5	170162766	G	A	D175N	KCNIP1
Brugada_101912	7	120385982	G	T	R539L	KCND2
Brugada_62293	7	91726634	G	A	R1300Q	AKAP9
Brugada_061826	8	121554142	T	C	I478V	SNTB1
Brugada_72900	8	32620767	A	T	K434*	NRG1
Brugada_8265	8	32620793	G	T	Q442H	NRG1
Brugada_09172	X	66766051	G	C	E165Q	AR

6.1.4. List of novel coding HQ-INDELs

Sample	Chr	Pos	Ref	Alt	AA change	Variant effect	Gene name
Brugada_072337	1	154541997	TA	T	-	Frame Shift	CHRNB2
Brugada_092178	4	114214678	TCAC	T	VT29V	Codon deletion	ANK2

Sample	Chr	Pos	Ref	Alt	AA change	Variant effect	Gene name
Brugada_103220	1	116243874	ATCG	A	DD324D	Codon deletion	CASQ2
Brugada_103377	19	50819304	C	CG	-	Frame Shift	KCNC3
Brugada_103969	14	23858896	CTCCAGCG TCCGAGA	C	-	Frame Shift	MYH6
Brugada_6582	6	39278694	AAAG	A	FF108F	Codon deletion	KCNK17
Brugada_74659	3	184071131	C	CCGG	R601PG	Codon insertion	CLCN2
Brugada_84023	3	148458931	A	AT	-	Frame Shift	AGTR1
Brugada_84023	17	28548775	TG	T	-	Frame Shift	SLC6A4
Brugada_85255	3	184071131	C	CCGG	R601PG	Codon insertion	CLCN2
Brugada_92497	6	39278694	AAAG	A	FF108F	Codon deletion	KCNK17
Brugada_92497	19	615946	CCCGCCG CCG	C	PPP715-	Codon deletion	HCN2
Brugada_93103	3	38968409	C	CAGTGAAGA	-	Frame Shift	SCN11A

6.1.5. List of clinical rs

Sample	Chr	Pos	dbSNP137_ID	AA change	Variant Effect	Gene Name
Brugada_6582	1	236849999	rs121434525	Q9R	Non-Synonymous	ACTN2
Brugada_101411	5	148206885	rs1800888	T164I	Non-Synonymous	ADRB2
Brugada_53353	5	148206885	rs1800888	T164I	Non-Synonymous	ADRB2
Brugada_73683	5	148206885	rs1800888	T164I	Non-Synonymous	ADRB2
Brugada_10552	X	115303595	rs121917810	G21V	Non-Synonymous	AGTR2
Brugada_72060	X	115303595	rs121917810	G21V	Non-Synonymous	AGTR2
Brugada_101912	4	114294537	rs45454496	E1022K	Non-Synonymous	ANK2
Brugada_61687	4	114294537	rs45454496	E1022K	Non-Synonymous	ANK2
Brugada_6582	4	114286207	rs66785829	V1540D	Non-Synonymous	ANK2
Brugada_73683	4	114269433	rs72544141	E110G	Non-Synonymous	ANK2
Brugada_74159	4	114294462	rs121912706	R1812W	Non-Synonymous	ANK2
Brugada_73051	3	8775526	rs116840771	-	Transcript	C3orf32
Brugada_74487	3	8787514	rs147250678	V139	Synonymous	CAV3
Brugada_101728	3	184075476	rs71318369	R191Q	Non-Synonymous	CLCN2
Brugada_08240	18	29126057	rs34065672	T903I	Non-Synonymous	DSG2
Brugada_08240	18	29125996	rs34417028	S883P	Non-Synonymous	DSG2
Brugada_91857	18	29099850	rs121913013	V56M	Non-Synonymous	DSG2
Brugada_08404	6	7542236	rs121912998	V30M	Non-Synonymous	DSP
Brugada_7269	6	7542253	rs77445784	G35	Synonymous	DSP
Brugada_85255	6	7542236	rs121912998	V30M	Non-Synonymous	DSP
Brugada_08240	17	39923625	rs193922705	-	Downstream	JUP
Brugada_63684	17	39923625	rs193922705	-	Downstream	JUP
Brugada_62293	21	35821904	rs144917638	T10M	Non-Synonymous	KCNE1
Brugada_74159	21	35743006	rs141423405	R77W	Non-Synonymous	KCNE2
Brugada_93238	21	35742947	rs74315448	I57T	Non-Synonymous	KCNE2
Brugada_08404	7	150642586	rs199473032	A1020V	Non-Synonymous	KCNH2
Brugada_103220	7	150648918	rs143011005	I181	Synonymous	KCNH2
Brugada_91596	7	150648918	rs143011005	I181	Synonymous	KCNH2
Brugada_7650	5	170149737	rs147147696	S116	Synonymous	KCNIP1
Brugada_83255	11	2608850	rs12720457	K266N	Non-Synonymous	KCNQ1
Brugada_92497	20	62039895	rs118192239	-	Downstream	KCNQ2
Brugada_91490	1	156104292	rs12117552	L105	Synonymous	LMNA
Brugada_9544	1	156084760	rs11549668	S17	Synonymous	LMNA
Brugada_74659	14	23887607	rs141764279	N1327	Synonymous	MYH7

Sample	Chr	Pos	dbSNP137_ID	AA change	Variant Effect	Gene Name
Brugada_101411	1	237604747	rs193922621	D362	Synonymous	RYR2
Brugada_101728	1	237948286	rs141528541	-	Intronic	RYR2
Brugada_101728	1	237872887	rs74323916	-	Intronic	RYR2
Brugada_103905	1	237948286	rs141528541	-	Intronic	RYR2
Brugada_63216	1	237875068	rs138073811	N3402	Synonymous	RYR2
Brugada_63447	1	237811766	rs72549416	D2439	Synonymous	RYR2
Brugada_7376	1	237875068	rs138073811	N3402	Synonymous	RYR2
Brugada_81290	1	237811889	rs143906555	L2480	Synonymous	RYR2
Brugada_93103	1	237664004	rs147479514	-	Intronic	RYR2
Brugada_95342	1	237948286	rs141528541	-	Intronic	RYR2
Brugada_101196	2	166900411	rs121918769	R604H	Non-Synonymous	SCN1A
Brugada_72939	2	166900411	rs121918769	R604H	Non-Synonymous	SCN1A
Brugada_091856	17	62028920	rs41280102	S906T	Non-Synonymous	SCN4A
Brugada_53609	17	62028920	rs41280102	S906T	Non-Synonymous	SCN4A
Brugada_7975	17	62028920	rs41280102	S906T	Non-Synonymous	SCN4A
Brugada_9346	17	62028920	rs41280102	S906T	Non-Synonymous	SCN4A
Brugada_95342	17	62028920	rs41280102	S906T	Non-Synonymous	SCN4A
Brugada_711	3	38601665	rs41311123	G1352	Synonymous	SCN5A
Brugada_73980	3	38601665	rs41311123	G1352	Synonymous	SCN5A
Brugada_74487	3	38628879	rs41312419	-	Intronic	SCN5A
Brugada_8467	3	38597180	rs45548237	S1449	Synonymous	SCN5A
Brugada_93061	3	38597180	rs45548237	S1449	Synonymous	SCN5A
Brugada_95342	3	38620970	rs199473189	V1081A	Non-Synonymous	SCN5A
Brugada_092178	2	167138296	rs121908919	K520R	Non-Synonymous	SCN9A
Brugada_81713	2	167159672	rs121908916	R142	Synonymous	SCN9A
Brugada_09172	17	28538374	rs28914832	I425V	Non-Synonymous	SLC6A4

ACKNOWLEDGMENTS

This work has been supported by the fellowship of the Fondazione CARIPLO ("From Genome to Antigen: a Multidisciplinary Approach towards the Development of an Effective Vaccine Against *Burkholderia pseudomallei*, the Etiological Agent of Mieloidosis", Grant number: 2009-3577).

-

I would like to thank all the people that contribute to this work: Cristina Battaglia, Gianluca De Bellis, Roberta Bordoni, Chiara Di Resta, Sara Benedetti, Maurizio Ferrari for supervision they provided at all levels of this project.

-

I would like to thank all the guys of the Sequencing group at ITB-CNR for their constant support and help during my Ph.D. experience

PUBLICATIONS

Papers published on international journals

- Cifola, I., Pietrelli, A., Consolandi, C., Severgnini, M., Mangano, E., Russo, V., De Bellis, G., Battaglia, C. (2013).
  Comprehensive genomic characterization of cutaneous malignant melanoma cell lines derived from metastatic lesions by whole-exome sequencing and SNP array profiling. PloS one, 8(5), e63597. doi:10.1371/journal.pone.0063597

- Rumi E, Pietra D, Guglielmelli P, Bordoni R, Casetti I, Milanesi C, Sant'Antonio E, Ferretti V, Pancrazzi A, Rotunno G, Severgnini M, Pietrelli A, Astori C, Fugazza E, Pascutto C, Boveri E, Passamonti F, De Bellis G, Vannucchi A, Cazzola M. (2013)
  Acquired copy-neutral loss of heterozygosity of chromosome 1p as a molecular event associated with marrow fibrosis in MPL-mutated myeloproliferative neoplasms. Blood, 121(21):4388-95. doi: 10.1182/blood-2013-02-486050

- Peano C, Pietrelli A, Consolandi C, Rossi E, Petiti L, Tagliabue L, De Bellis G, Landini P.(2013)
  An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. Microbial informatics and experimentation, 3(1):1. doi: 10.1186/2042-5783-3-1.

- Pietra D, Brisci A, Rumi E, Boggi S, Elena C, Pietrelli A, Bordoni R, Ferrari M, Passamonti F, De Bellis, G, Cremonesi L, Cazzola M.(2011)
  Deep sequencing reveals double mutations in cis of MPL exon 10 in myeloproliferative neoplasms. Haematologica, 96(4):607-11.

- Maciag A, Peano C, Pietrelli A, Egli T, De Bellis G, Landini P.(2011)
  In vitro transcription profiling of the sigma-S subunit of bacterial RNA polymerase: re-definition of the sigma-S regulon and identification of sigma-S-specific promoter sequence elements. Nucleic Acids Res, 39(13):5338-55. doi: 10.1093/nar/gkr129

Publications on proceedings of international conference

- Surfing into NGS ocean: Guidelines for whole-exome sequencing variation analysis and data visualization. (Pietrelli, A., Cifola, I., Severgnini, M., Consolandi, C., Mangano, E., Russo, V., Battaglia, C. and De Bellis, G.) (2013).

Molecular Med Tri-Conference 2013, February 11-15, San Francisco CA, USA

- ESCAPE: A tool for extending assembly contigs without a reference genome (G Corti, M Severgnini, A Pietrelli, F Fuligni) - Next Generation Sequencing Workshop", Università di Bari (6-10 October 2010)

- In vitro transcription profiling of the σS subunit of bacterial RNA polymerase: σS regulon and σS specific promoter sequence elements re-definition (A Maciag, C Peano, A Pietrelli, T Egli, G De Bellis, P Landini) - 4th Congress of the FEMS, Geneve (26 – 30 June 2011)

- Search of new candidate genes in Brugada syndrome using Next-Generation sequencing (Chiara Di Resta, Alessandro Pietrelli, Roberta Bordoni, Simone Sala, Alessia Mongelli, Gianluca De Bellis, Maurizio Ferrari, Sara Benedetti) - Golden Helix Symposium 2012 "Genomic Medicine", Turin (18-21 April 2012)

- A combined method for rRNA removal in Burkholderia thailandensis - Cortona procarioti (4 May 2012) (Congress talk)