Università degli studi di Padova
Scuola di Dottorato di Ricerca in Bioscienze e Biotecnologie
Indirizzo Biotecnologie

Exome resequencing
Arrhythomogenic cardiomyopathy: a case of study

Coordinatore:
Chiar.mo Prof.
GIORGIO VALLE

Dottorando:
ALESSANDRO ALBIERO

Ciclo XXIV

Abstract

Today massively parallel DNA sequencing platforms are become widely available, reducing the costs and the time of DNA sequencing.
Next Generation Sequencers (NGSs) allow to obtain large amount of data and they open new perspectives in fields like genomic and medical research. One of the most promising application in medical research and in diagnostic is the exome sequencing,a specific targeted re-sequencing of the known exons. There are two advantage in sequencing the exome:

- The human exome is the 1% of the total genome (about 30Mbp) and it is so possible to obtain high coverage with low costs.
- Several variations in exome cause diseases.

These two features make the exome sequencing very interesting and increasingly used by scientists. There are several strategies for exome sequencing but, we considered Illumina and SOLiD approaches.
In details, we analyzed 6 patients affected by arrhythomogenic cardiomyopathy. Genetic variations in these patients were already characterized with Sanger technologies so we could compare different variant detections algorithm with SOLiD reads and with Illumina reads.
Results confirmed the key role of coverage in detecting variants.

Abstract - ITALIANO

Attualmente le tecnologie di sequenziamento massivo del DNA sono diventate ampiamente disponibili e hanno ridotto sia i costi che i tempi di sequenziamento.
I sequenziatori di nuova generazione (NGS) permettono di ottenere grosse moli di dati e hanno aperto nuove prospettive nel campo della genomica e della ricerca medica.
Tra le applicazioni più promettenti nel campo della ricerca medica e della diagnostica spicca il sequenziamento dell'esoma definibile come uno specifico targeted resequencing degli esoni noti. Ci sono due vantaggi nel sequenziare l'esoma:

- L'esoma umano è circa l'1% del totale del genoma (circa 30 Mbp) per cui è possibile ottenere alte coperture con costi ridotti.
- Mutazioni a livello esonico sono alla base di molte patologie.

Queste caratteristiche rendono il sequenziamento dell'esoma molto interessante e sempre più utilizzato dagli studiosi. Esistono molte strategie per il sequenziamento dell'esoma, ma in questa tesi verranno considerati gli approci tramite Illumina e SOLiD. Nel dettaglio verranno analizzati 6 pazienti affetti da cardiomiopatia aritmogenica. Le varianti generiche in questi pazienti sono già state caratterizzate con tecnologia Sanger e si vogliono comparare diversi algoritmi di ricerca delle varianti con le sequenze Illumina e SOLiD. I risultati confermano l'importanza del coverage di sequenza.

Contents

Introduction

In this PHD thesis, I take into consideration a specific application of Next Generation Sequencers (NGSs): the human exome resequencing. Sequencing an exome (specifically the human exome) was unthinkable few years ago but, today it is only one of the applications of NGSs.
Untill NGS, in the genomic field the problems were to obtain sufficient data reducing the costs and time: sequencing an eukaryotic genome could take several years and lot of scientist efforts. Currently, the same goal can be obtained in few weeks with a biologist, a bioinformatics, and a Next Generation Sequencer.
In this scenario, it could seem that NGSs solve the major part of problems of the -omics sciences. But this is not true. NGSs solved the problem of "how to obtain the data" but they do not solve the problem of "how to manage and analyse the data".
NGSs changed the role of bioinformatic that is became a fundamental figure in every laboratory which has or have had data from NGSs. The major problems today are computational power, informatics space and capable bioinformatics.
In this thesis the first 2 chapters are general consideration about NGSs and their principal applications. Chapters 3 is a deepening in exome resequencing. Chapter 4 and 5 are bioinformatical deeping in aligning and SNP calling. Chapter 6 is the application of exome resequencing on the arrhythomogenic cardiomyopathy both for diagnostic and research. I considered 6 patients, already characterized with Sanger technology, and I investigated about the different algorithms.
The aim of this PHD thesis is to understand the limits and the capability of exome sequencing to identify SNPs and INDELs. I analyzed different samples with different coverages and in one case with different technologies. In this scenario, I could understand when a variant can be considered reliable or not, that is very important for using the exome sequencing in diagnostic and in research fields.

Chapter 1

Next Generation Sequencing

Contents

1.1 Introduction

With Next Generation Sequencing (NGS), we consider all sequence technologies where:

- Bacterial cloning phase is by-passed.
- Sequencing is performed at the same time over all DNA fragments

These two improvements allowed to reduce time and costs of sequencing and an increasingly number of laboratories has today access to sequencing technologies.
The bottle-neck is still the data analyses[14] bacause the large amount of data produced by NGS is difficult to manage and analyze.

1.2 Roche 454 sequencer

454 was the first next generation system commercialized by Roche. This sequencer is based on pyrosequencing technology[1] that depends on the detection of pyrophosphate released during nucleotide incorporation.

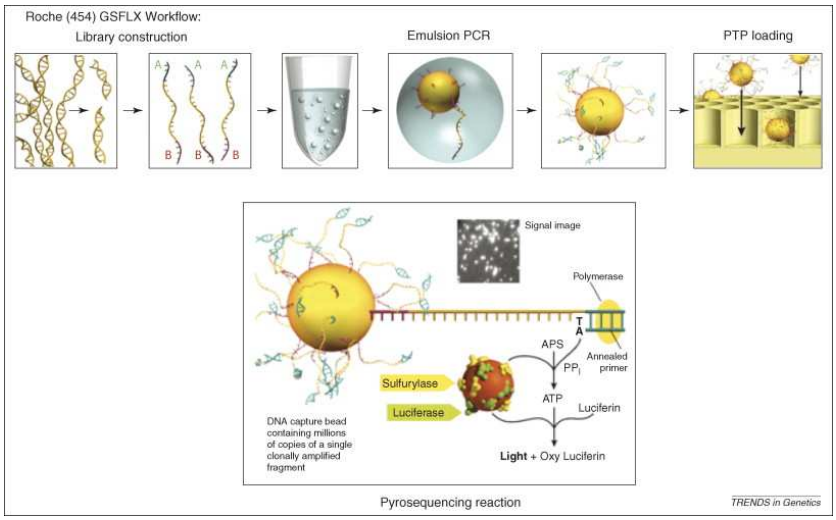

Figure 1.1: 454 Pyrosequencing workflow[44]

The DNA is fragmented with physical methods and specifical adaptors were ligated to the end. The DNA is captured by beads and then it is amplified with an emulsion PCR. Beads are then deposited on a picotiter plate (PTP) with all necessary sequencing enzymes.
Sequencer let flow one of dNTP in a controlled series and the pyrophosphate, released by an incorporation, become substrate of sulfurylase, luciferase and luciferin and there is emission of light[33].
The order of nucleotides allow to know the sequence of the reads, and the light intensity the number of incorporated nucleotides.

The read length of Roche 454 is now around 600/800 bases and the throughput is around 1 Gbp[2] but, 454 throughput is less than the SOLiD or ILLUMINA one, so 454 is not used for exome resequencing. Costs should be too high.
The most outstanding advantage of Roche is its speed and the reads length. One run takes 24 hours and the reads have a length similar to Sanger technology ones.

1.3 Illumina HiSeq sequencer

Illumina sequencers are based on sequencing by synthesis (very similar to Sanger technology). The DNA is broken is small fragment (around 400/600 bases), ligated to specific adaptors and, then placed in a particular flowcell with fixed primers. On the flowcell the DNA is amplified by bridge amplification to create clusters of clonal molecules.
Sequencing is performed by synthesis adding nucleotides containing fluores-

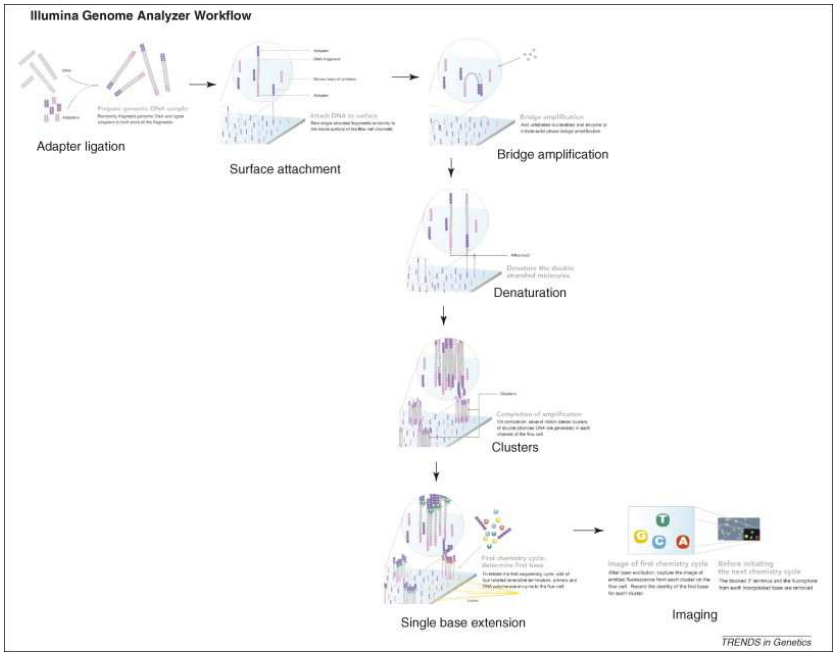

Figure 1.2: Illumina workflow[44]

cent dye; the signal is captured by a CCD camera.
There are several versions of Illumina sequencer, for example HiSeq 1000 produces 300 Gbp per run in 8 days[3]. The reads are from 50 to 150 nucleotides depending on sequencer version and sequencing kit used.

1.4 Applied Biosystem SOLiD sequencer

SOLiD is acronym of Sequencing by Oligo Ligation Detection and its sequencing method is based on ligation. The sequencer adopts the technology of two-base sequencing based on ligation sequencing.
DNA is amplified by emulsion PCR (similar to 454) and then it is placed on flowcell. Sequencing is performed by adding 8 base-probe ligation which contains ligation site (the first base), cleavage site (the fifth base), and 4 different fluorescent dyes (linked to the last base)[44]. Every fluorescent dyes represents 2 bases.
Whit SOLiD technology every base is sequenced two times and the output is in color space format. Color space is different from base space (Illumina, 454 and Sanger output) and it needs of dedicated software.
SOLiD throughput is similar to Illumina one and reads length varies from 35 to 75 base pairs.

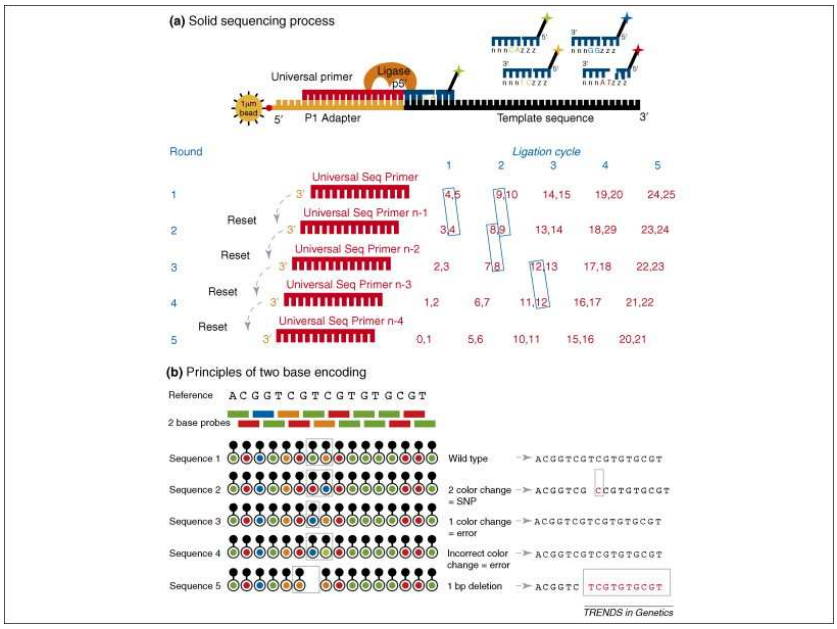

Figure 1.3: SOLiD workflow[44]

1.5 Other Sequencers

Beyond the three sequencers 454, Illumina and SOLiD, there are other next generation sequencers that start to be establish in the NGS market. Among them, Ion Proton and Ion Torrent (both of Applied Biosystem) are the most known. Both sequencers use a sequencing strategy similar to 454, but they do not measure the light intensity of pyrophosphate but the H^+ variation.

1.6 NGS impact on genetic research

New sequencers allowed to obtain large amount of data in very few time. Costs are also reduced (see figure 1.4), so, much more scientists than in the past, have today access to genomic or transcriptomic data. The problem today is not obtain the data but it is the manegment of these data and their processing. One run of SOLiD or ILLUMINA can produce up to 300 Gbp and their processing can double or triple the data.
To manage this data it is necessary to have clusters of hard disk and, analysis can be performed only if it is available big computers, or clusters, with a large number of CPU and lot of RAM.
These problems are often underestimated and scientists have difficult to analyze their data for their researches.
All these troubles can be solved buying hardware or using clouds system such

Sequencer	Read length	throughput	Sequencing method	Output Format
454	up to 800 bp	up to 1Gbp	Pyrosequencing	sff format
Illumina	from 50 to 150 bp	up to 300 Gbp	Sequencing by sinthesis	fastq format
SOLiD	from 35 to 75	up to 200 Gbp	Sequencing by ligation	color space format

Table 1.1: Table of principal NGSs and their output

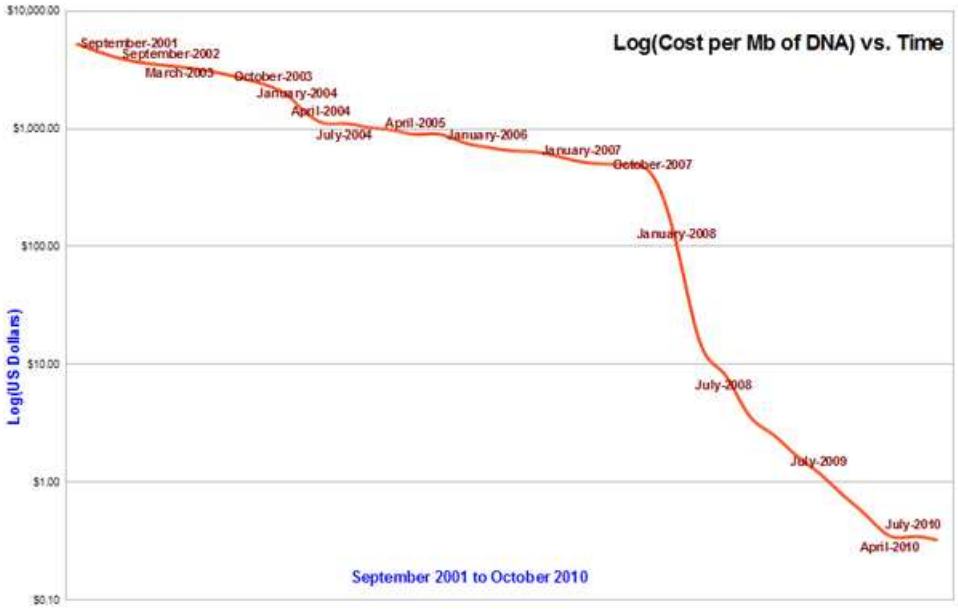

Figure 1.4: Costs of sequencing per base against time

as Amazon (http://aws.amazon.com/ec2/) but this is only a partial solution because these data have to be sent to the remote computer and net transfers can be a real bottle-neck: transferring an Illumina run can take up to 10 or 20 days.
At the same time, having hardware is not always the solution. In fact, bioinformatic capabilities are obligatory to perform the analysis.
In this scenario, the bioinformatic became a figure very important in every laboratories which manage NGS experiments.

Chapter 2

NGS Applications

Contents

2.1 Introduction

NGSs opened new perspective in genomic research. Often the unique limit is the costs. Currently NGSs are used for:

- DeNovo Sequencing
- Resequencing
- RNA-seq and DeNovo transcriptomic sequencing
- Metagenomics

Theorically all the NGSs could be used for these applications but, often the choise is taken considering: costs, bioinformatic analysis, read lengths and read quality.

2.2 DeNovo Sequencing

The term "DeNovo Sequencing" is often confuse with "DeNovo sequence assembly". Even if these two terms seem synonyms, they are very different. NGSs allow today to perform the "DeNovo Sequencing" with low costs (than the past) and with reduced time, but the "DeNovo sequence assembly" remains a challenging tasks. DeNovo Sequencing is the process with which we

obtain a series of read that potentially cover all the genome of an organism. Generally a DeNovo sequencing is measured by coverage:

$$AVG\_Coverage = \left(\frac{Sequenced\_bases}{Genome\_Size}\right)$$

Where AVG_Coverage is the average coverage, Sequenced_bases are the number of bases obtained by sequencing and Genome_Size is the size of sequenced genome in bases.
"De Novo Sequencing assembly" is the process whereby we merge together individual sequence reads to form long contiguos sequences (contig) sharing the same nucleotide sequence reads were derived[43]. De Novo Sequencing assembly is a challenge and, currently there is not a single algorithm or software that perform this tasks. The assembly results are linked to the coverage of the sequencing, the lenght of reads and, the genomic structure of the analyzed organism.[43]. Among the software used for assembly the most know are: Newbler[4], ABYSS[30], CLC[5], SOAPdenovo[23], and Velvet[35].
Generally, for a De Novo sequencing it is requested a coverage from 30X to 50X coverage. Currently these coverages can be obtained with low costs thanks to NGSs. The most used sequencers for this aim are 454 and ILLUMINA.

2.3 Resequencing

Resequencing is very similar to DeNovo Sequencing but, the genome of the analyzed organism is known. The scope of a resequencing is to find variations that can be linked to particular phenotypes. Resequencing can be done over all genome or only in selected regions (amplicons, targeted resequencing and exomes)[44]. In all cases the coverage is the key of the experiments; mutation discovery generally needs a 20X coverage, but studies in amplicons for tumor characterization need very high coverage such us 1000X or 5000X.
In a resequencing project, the first operation to do is to map the reads against the reference admitting mismatches and gaps. Currently there are lots of software to map the reads and the most used are: PASS[18], BOWTIE[41], Newbler[4], Soap[23], BWA[38] and CLC[5].
Output of these programs is an alignment, and the standard output is the SAM/BAM format[6].
These output file are input for SNP calling softwares. Chapter 4 and chapter 5 are a deepening of alignments and SNP callers.

2.4 RNA-Seq and DeNovo transcriptomic sequencing

RNA-Seq is a recently developed approach to transcriptome profiling that uses NGS technologies. Studies using this method have already altered the view of the extent and complexity of eukaryotic transcriptomes[47]. RNA-Seq is generally performed by Illumina or SOLiD and it is requested a reference (genome or transcriptome) where aligning the reads against.
Once reads have been obtained, the first task of data analysis is to map the short reads from RNA-Seq to the reference genome the same software viewed in Resequencing Chapter. The alignment is very important and not trivial, outputs need to be then analyzed with dedicated statistical tools. The major problems of the RNA-Seq alignment are:

- reads that match multiple locations.
- gap openings for spliced alignments.

Despite the problems described above, the advantages of RNA-Seq have enabled to generate an unprecedented global view of the transcriptome and its organization.
454 is generally not used for RNA-Seq but, it is prefered for De Novo transcript assembly. Thanks to the long reads of 454 it is possible to identify transcritps of non model species. The best software to assembly transcriptome is Newbler[4].
Often, for novel organism where the genome sequence is not known, 454 and RNA-Seq are combined to obtain the transcriptional profile and the transcriptional differeces in different condition or tissue of the new organism.

2.5 Metagenomics

The term Metagenomics is very ambiguous because lots of differents experiment can be classified like metagenomics. All cases Metagenomics is tool for studying the diversity and metabolic potential of environmental microbes, whose bulk is as yet non-cultivable[32]. In this scenario we can perform several different experiments focused on the characterisation of bacteria or fungi in a sample, and call them metagenomics.
One of the most used technique for characterizing the bacterial diversity of a sample is the 16S (for fungi ITS) amplicon analysis. In this case, it is used a set of primers for amplifying the variable 16S region. There are several tools to compute this data: CLOTU[7], MOTHUR[8] and QIIME[9].

Chapter 3

Exome Resequencing

Contents

3.1 Introduction

Exome resequencing is a special application of the targeted resequencing and has become a powerful new approach for identifying genes that underlie Mendelian disorders[16][26]. The exome can be defined as the sum of all coding sequencing regions (CDS).

3.2 Why sequencing the human exome?

Despite human exome is only a small part of the entire genome, it contains all the information of the genes and several diseases are related to variations on genes[15].
We can consider three points to give an answer to the question "Why sequencing the human exome":

- Positional cloning studies focused on protein-coding sequences have proved to be highly successful at identifying variants for monogenic diseases[45].

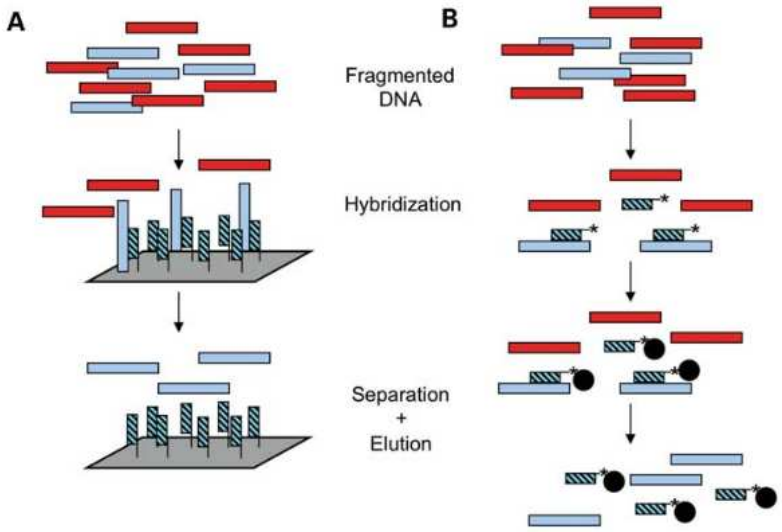

Figure 3.1: Exome Capturing methods: A - Solid-phase. B - Liquid-phase.[20]

- Many Mendelian disorders are caused by disruption of protein-coding sequences[31].
- A large fraction of variant such as missense or nonsense single-base substitutions or small insertion–deletions (indels) in gene coding sequence are predicted to have functional consequences and/or to be deleterious[22].

3.3 Capture Methods

There are 2 principal methods for capturing the exome: Solid-phase hibridization and Liquid-phase hibridization[20].
Solid-phase hibridization utilize probes complementary to sequences of interest fixed to a solid support (microarray or filters). The non-targeted regions are washed out and the regions of interest remains on the support.
Liquid-phase hibridization, at contrary, uses biotinylated probes and the regions of interest are then recovered with magnetic streptavidin beads. Figure 3.1 shows the two principal methods for capturing the exome. Currently the most used method is the Liquid-phase hibridization.

Commercial kits now target, at a minimum, all of the RefSeq collection and an increasingly large number of hypothetical proteins. Nevertheless, all existing targets have limitations. First, the knowledge of all truly protein-

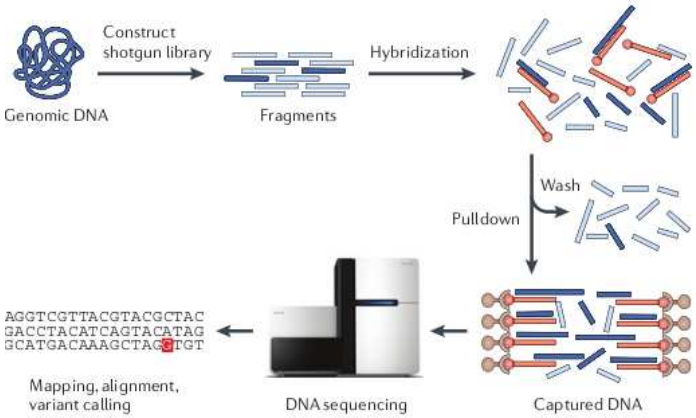

Figure 3.2: Exome Capturing workflow[15]

coding exons in the genome is still incomplete, so current capture probes can only target exons that have been identified so far. Second, the efficiency of capture probes varies considerably, and some sequences fail to be targeted by capture probe design altogether[15]. In this thesis I take into consideration only two commercial kits: the SOLiD and the Illumina kits.
The workflow for the exome enrichment is showed on figure 3.2.

3.3.1 Illumina exome enrichment kit

Illumina exome enrichment kit is called TruSeq Exome Enrichment Kit. It is based on hybrid selection(Fig.3.1),and allows to select 201071 different regions for a total of 62 Mbp and 20846 genes. The probes capture also Untraslated Regions (UTRs).

3.3.2 SOLiD exome enrichment kit

SOLiD exome enrichment kit is called New Target Enrichment Kit. It is based on hybrid selection (Fig.3.1), the kit allows to select 195282 different regions. This kit covers 37 Mbp and 19911 genes.

3.3.3 Comparison of exome enrichment kits

The SOLiD and the ILLUMINA kit are different because of they cover in many cases different regions. More precisely:

- Overlapping regions are 186048 bp.
- 33 Mbp are in common

Sample	% Reads on target	% Reads on target +500	Platform
I2	55,7%	72,1%	ILLUMINA
I4	56,6%	74,3%	ILLUMINA
I5	53,6%	69.4%	ILLUMINA
I6	48,4%	61,8%	ILLUMINA
I7	48,6%	62,6%	ILLUMINA
I12	55,2%	73,1%	ILLUMINA

Table 3.1: Percentage of reads on target in the six patients we analysed exome with Illumina technology.

- ILLUMINA has 29 Mbp exclusive and SOLiD 4Mbp
- The extra regions of ILLUMINA kit are generally UTRs

These differences show also that the definition of exome is not globally accepted.
Another important thing to take into consideration is that these kits capture not only the targeted regions but, often we can find mithocondrial DNA and regions flanking the target:

- Mithocondrial DNA (we found over all samples with high coverage) is very usefull to check the sample before and after sequencing to avoid errors (sample exchange). In fact we can sequence the hypervariable regions (HVR1 and HVR2) before the enrichment and then checking them after the sequencing with NGS. If they are equal we can be sure that there is no sample exchange.

- Flanking regions are also very important. Several mutations can be on these region and they can have damaging effects.

In the table 3.1 are reported the percentage of reads aligned against the target regions and against the target regions plus 500 bp (at 3' and 5') of the six patients we analysed the exome with ILLUMINA technology.

3.4 Application of Exome sequencing

Human Exome sequencing has several application both in diagnostic and in reseach fields. We can find 3 principal applications [20]:

- Medical field

- Human Evolution
- Biological field

3.4.1 Medical Field

In the medical field, human exome sequencing finds a lot of applications. Several disease are associated with DNA variations in exomic regions (mendelian diseases or other well characterized diseases)and exome resequencing can be used for diagnostic purpuses. For example Ng et al[25] sequenced 12 human exome from patients with Freeman-Sheldon disease that is a rare syndrome classified like dominantly inherited rare Mendelian disorder. In the study, researchers were able to find the variations causative of the disease.
Many other studies was performed for mendelian diseases (autosomal recessive ataxia[19], papillorenal syndrome[29]) and, in several cases human exome resequencing allowed to find the causative mutations.
These studies demostrated that exome resequencing can be used for diagnostics purpuses and in this thesis I investigated about the application of exome resequencing for the diagnosis of arrhythmogenic cardiomyopathy.
At the same time it is very important to consider that having the exome means also to have lots of data that can be useful for future studies. In fact whith exome resequencing we have a photo of all the variations of an individual that can be useful for research purpuses. For example if we are investigating about an unknown disease we can analyse all the canditated mutations filtered with common mutations from unrelated patients sequenced for other reasons.

3.4.2 Human Evolution

Like specified in the last subsection, having the exome of an individual means to have a photo of all variants of this individual. These allow to perform comparisons beetween different persons from different populations and extract candidates mutations that can explain the different phenotypes.
A similar study has been performed by Yi et al[34], where it was compared exomes from high-altitude and low-altitude populations to identify possible differences in allele frequencies that can explain different adaptations.
In this study, there were found a discrete number of genes possible candidated for the high altitude adaptation.

3.4.3 Biological Field

Copy number variations (CNVs) and genomic structural variations are large variations that have been considered in the last few years. CNVs are

insertions, deletions or duplications of genes or other regions of the genome while, genomic structural variations are generally inversions or translocations of pieces of genome.
Both variations are in some cases linked to diseases and they can be detected also by exome sequencing[21][42].

Chapter 4

Alignment

Contents

4.1 Introduction

The first step after the sequencing of an exome is the alignment. We are considering the human exome so the alignment have to be performed against the human genome. There are several software to align short reads against a reference but in every case we have to align admitting mismatches and indels. Even if it may be seem a simple task, align short reads is not trivial, there are several software available based on different algorithms.
Mapping results influence the results of SNP Calling software, so it is very important to choose a good aligner with the best parameters.
Mapping the reads against a reference means finding the position of the sequenced piece of genome on the reference taking into considerations sequencing errors and variations.
There are two major problems when we consider the mapping and the NGS output: the first problem is the amount of data and the time necessary to align; the second one is the reads that seem to have multiple solutions[17]. Both problems are very important and they can be connected.
Align billions of reads can be very time consuming and currently algorithms tried to be as faster as possible. The problem of multiple mapping reads is connected to the read length and it is important to consider 2 properties of

a mapped reads:

- The best hit.
- The unique hit.

The best hit is the best position of the reads onto the genome indipendently by the number of mismatches or indels. Generally, every alignment has a score and the best hit is the alignment with the best score. Sometimes, a read can have multiple best hit, so we can map this read in different position and we don't known what is the real position of this read onto the genome. When a reads has only one best hist, it is called unique best hit. Read lenght is stricly correlated to the unique best hit, short reads tend to have several best hit only for statistical questions.
To understand the relation between read lenght and unique hit, we have to consider that a read with length N has N^4 possible combinations (we have 4 nucleotides).
If N is equal to 10, the possible combinations are 10000, so the probability of find our string is 1/10000. The human genome is 3Gbp and we can calculate the number of 10 length strings: it is 3000000000 - 10 + 1, so we expect at least 300000 strings equal to our one. (We consider the human genome like a random string composed by 4 letters).
If N increase, the probabilty of find the same string decrease. This is true if the human genome is a random string, but this not the case of genomes. In addition, genomes have repetitive regions of different lengths that increase the probability to find multiple hits for short reads.
The different algorithms used for mapping short reads can choose 3 different solutions for multiple hit:

- Ignore the multiple hits.
- Consider only a part of all the hits.
- Consider all the hits.

The last solution can increase markedly the processing time for mapping.

4.2 Mapping strategies

Several algorithms had been developped to map reads against a reference; the goal is always find the real position of the reads onto the reference limiting processing time and hardware equipments. In all cases the principal problems are the reads lengths and the number of reads to align.
The most used software are bases on indexing strategies: some software prefers to index the reads, other ones prefer to index the reference. Indexing can take several time and can create large files used then for the alignments.
In this thesis I take into consideration 5 different software:

- Pass.
- Bowtie.
- Bfast.
- CLC.

I don't talk about the alignment algorithms, but, I consider only the principal characterystics of the mapping software.

4.2.1 PASS

Pass[18] is a mapping software developped at CRIBI (University of Padua). PASS can align short reads in bases space (Illumina) and in color space (SOLiD), and it uses a very fast algorithm based on genome indexing. PASS uses short words for placing the reads on the genome and then refines the alignment using a sort of Smith-Watermann algorithm. In my project PASS was used to align SOLiD data.

4.2.2 BOWTIE

Bowtie[41] is based on Burrow-Wheeler transform. Bowtie is very fast but it takes several time to construct the indexes (on the genomes). Another advantage of BOWTIE is the hardware request: BOWTIE can align against the human genome using a laptop, it requires few Giga of RAM. In the thesis BOWTIE had been used to align ILLUMINA reads.

4.2.3 BWA

BWA (Burrows-Wheeler Aligner)[37][36] is an efficient program that aligns short sequences against a long reference sequence such as the human genome. It implements two algorithms, bwa-short and bwa-sw. The former works for query sequences shorter than 200bp and the latter for longer sequences up to around 100kbp. Both algorithms do gapped alignment. BWA needs to index the reference and this operation can take several time. Like BOWTIE it is based on Burrow-Wheeler transform.

4.2.4 CLC

CLC[10] is a commercial suite that offer several tools for genomics and transcriptomics analyses.
CLC mapper is based on a seeding approach. The algorithm iterates over input reads and maps each read individually by applying the following procedure: seeding sequences of 30 nucleotides each are sampled from each third

position of the input read. These seeds are looked up in the index and resulting candidate alignment locations are examined using a banded Smith Waterman.

4.3 Mapper Evaluation

It is very difficult to evaluate the results of a mapper because we can take into consideration different parameters. The best way should be to have a set of reads with known position and with known mismatches.
In my PHD thesis, I take into consideration real data so it is not known the real position of each read. So, the evaluations has been made taking into consideration the number of aligned reads. Results are report in the chapter 6. For all software I used default parameters.

Chapter 5

SNP Caller

Contents

5.1 Introduction

SNP Callers are a series of tools that extract variants from an alignment. The problems, in SNP Callers, are the high error rate of the base calling and the errors in alignments. Under such circumstances, accurate SNP calling are difficult and there is often considerable uncertainty associated with the result[28].
The problem of error rate associated to the NGSs can be by-passed with high coverage; the alignment problems otherwise can be solved only using a good mapper.
In this PHD thesis I take into consideration 2 SNP Caller: CLC Variant Probabilistic caller and GATK[24].

5.1.1 GATK:Genome Analysis ToolKit

GATK is a suite designed to enable rapid development of efficient and robust analysis tools for next-generation DNA sequencers. This is the most used tools and one of the most citated.
GATK includes a series of analysis for variant calling and it accepts a BAM file in input.
There are several workflow for GATK; in this thesis I used the workflow described in figure 5.1.

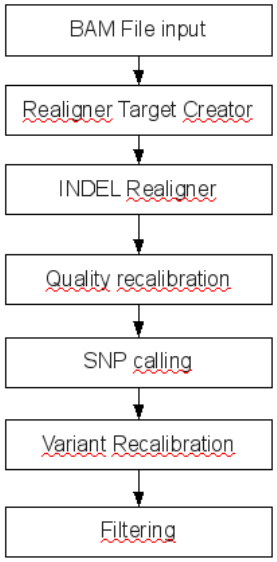

Figure 5.1: GATK Workflow

Realigner Target Creator and Realignment

With this tool, GATK suite performs a realignment of some intervals using Smith-Watermann algorithm[46]. To speed up this operation, GATK in a first phase find the candidate regions analysing the BAM file; then only these regions are realigned using Smith-Waterman.
The idea is to minimize the number of mismatches especially in those regions where there are indels. In general, a large percent of regions requiring local realignment are due to the presence of an insertion or deletion in the individual's genome with respect to the reference genome. Such alignment artifacts result in many bases mismatching the reference near the misalignment, which are easily mistaken as SNPs.

Quality Recalibration

In this phase, GATK performs a correction of the quality score of the reads in the BAM file. To recalibrate the quality score, GATK analyse three parameters:

- The reported quality score.
- The position of the nucleotide in the reads.
- The preceding and current nucleotide.

Using these 3 parameters, GATK is able to correct the quality score of the bases.

SNP Calling

After BAM correction, GATK can perform the SNP calling. GATK is designed also for multiple samples using a Bayesian genotype likelihood model to estimate simultaneously the most likely genotypes and allele frequency in a population of N samples.
SNP calling is performed observing mismatches and indels in the alignment file and taking into consideration the coverage, the frequency of the variations and the strand of the aligned reads.
At the same time GATK gives a score for each variant called (Variant Recalibration and Variant Filtration). Each variant has also a sort of comment to better indentifying problematic result (such as low coverage or strand bias that can create artifacts).

5.1.2 CLC Probabilistic Variant Caller

CLC Probabilistic Variant Caller[11] is a tool of the commercial CLC suite. Probabilistic Variant Caller has been designed for calling variants in haploid (bacteria), diploid (human) and polyploid genomes (cancer or plants). The tool is very simple to use and it take as input a CLC alignments file. The alignment can be performed using CLC or using also other mapper, the result BAM file can be uploaded in CLC Workspace.
The CLC Variant Caller algorithm combines a Maximum Likelyhood approach with a Bayesian model to call the variants and to give to each one a score that represent the probability of the variant.

More precisely it is first calculated a prior probability using only the alignment. The starting parameters are shown in figure 5.2.
These parameters are updated using an Expected Maximization approach.
At the same time it is calculated an error probability taking into consideration also the quality score of the aligned reads, and for each quality score it is calculated a different error probability table.
After the prior and the error probability have been estimated the Variant Caller give in output the most probable allele for each position.

CLC output is a table of variants with several parameters like coverage, forward/reverse reads and probability of the variant.

Site Type	Prior probability
A/A	0.2475
A/C	0.001
A/G	0.001
A/T	0.001
T/C	0.001
T/G	0.001
T/T	0.2475
G/C	0.001
C/C	0.2475
G/G	0.2475
G/-	0.001
A/-	0.001
C/-	0.001
T/-	0.001

Figure 5.2: Initial probability of CLC Variant Caller

Chapter 6

Arrhythomogenic Cardiomyopathy

Contents

6.1 Introduction

Arrhythmogenic right ventricular cardiomyopathy (ARVC) is an inherited myocardial disease associated with significant genotype and phenotype heterogeneity. The structural features of ARVC consist of progressive fibrofatty replacement of myocytes and, clinically, the disease has been associated with ventricular arrhythmias at risk of sudden cardiac death[27].
In my thesys I take into consideration 6 patients with ARCV disease already characterized with Sanger. The exome of the 6 patientshas been enriched and sequenced using Illumina and SOLiD strategy.

6.2 Results

6.2.1 Mapping Results

After sequencing, reads were aligned against the human genome using CLC, PASS, BOWTIE and BWA using default parameters. BOWTIE and BWA required a preliminary indexing of the reference that take several hours.

Sample	Number of reads	Technology
2	17.250.274	ILLUMINA
6S	49.719.032	SOLID
4	81.382.994	ILLUMINA
5	70.603.922	ILLUMINA
6	48.166.720	ILLUMINA
7	52.233.528	ILLUMINA
12	33.786.456	ILLUMINA

Table 6.1: Number of reads sequenced per sample.

Sample	# of Reads	PASS	CLC	BOWTIE	BWA
2	17.250.274	75,97%	87,18%	81,81%	86,89%
6S	49.719.032	67,18%	75,61%	-	-
6	48.166.720	80,64%	86,75%	62,19%	76,90%
12	33.786.456	83,88%	86,75%	86,01%	93,91%

Table 6.2: % of reads aligned with PASS, CLC, BOWTIE and BWA.

CLC and PASS did not required this indexing.
The fastest software was CLC followed by BWA, BOWTIE and PASS. For all the software I take into consideration the number of unique aligned reads admitting 2 mismatches and gaps. For SOLiD reads I used only PASS, CLC and BFAST [12][40][39] (BFASTA uses an algorythm vary similar to PASS). The sequencing of six patients produced different number of reads and for the software evaluation I considered only the 3 patients with the lowest number of reads. This choise has been made for minimize the processing time for the alignments. In the table 6.1 there is reported the number of reads produced by the sequencers.

The samples chosen for the mapper evaluation was the samples 12, 2 and 6 for Illumina and the sample 6S for SOLiD. In the table 6.2 are reported the results obtained with the 4 mappers.

Table 6.2 shows that the mapper with the higher number of unique best hit is CLC; so, we choose CLC like principal software for the alignments. CLC is also the simpliest software to use thanks to its graphical interface.

In table 6.3, there are showed the results of the alignments of all samples and the average coverage of the exome. Like specified in Chapter 3, Illumina

Sample	# of Reads	% aligned (unique)	Avg Exome Coverage
2	17.250.274	87,18%	9,24X
6S	49.719.032	75,61%	40,01X
4	81.382.994	84,79%	51,65X
5	70.603.922	85,32%	41,53X
6	48.166.720	86,75%	25,07X
7	52.233.528	85,91%	27,36X
12	33.786.456	86,75%	20,04X

Table 6.3: % of reads aligned with CLC over all samples.

Sample	Technology	% reads on target	% reference not covered	% reference over 20X
4	ILLUMINA	56,6%	3,96%	81,03%
5	ILLUMINA	53,6%	3,73%	75,08%
7	ILLUMINA	48,62%	4,84%	53,52%
6	ILLUMINA	48,24%	5,76%	48,35%
12	ILLUMINA	55,21%	7,83%	38,65%
2	ILLUMINA	55,76%	22,14%	11,07%
6S	SOLiD	70,11%	9,56%	60,27%

Table 6.4: Table with coverage data of 6 patients. Reads has been aligned with CLC.

and SOLiD have different kits. The average coverage is calculated like:

$$\frac{total\_nucleotide\_aligned\_on\_the\_exome}{total\_nucleotide\_of\_the\_exome}$$

Exome sequencing is a targeted resequencing, and beyond the average coverage there are other parameters that have to be considered for understanding the differences in the sample. These parameters are reported in table 6.4.

Reads on target are all the reads that maps on the exome. With illumina kit, we have the 50% of reads that maps on the target while, with SOLiD kit we had the 70%. If we consider the targeted regions plus 500 bp in 5' and in 3', the percentage of reads on target increase of 20%. These confirm that also these regions are covered and we can consider also the variations calculated for these extra-regions. Like expected the other data in table 6.4 (*% reference not covered and % of reference over 20X*) are stricly related to the average coverage of the samples. I reported also the percentage of reference over 20X coverage because, like explained later, at this coverage we have the best result for variation calling.

Sample	Gene	Chr	Position	Sanger	GATK	CLC	CASAVA
4	DSG2	18	29104698	C/T	C/T	C/T	C/T
4	DSG2	18	29125854	A/G	A/G	A/G	A/G
4	DSG2	18	29126670	T/C	Not found	T/C	Not found
4	DSP	6	7542149	-/A	Not found	-/A	Not found
4	DSP	6	7563983	G	G	G	G
4	DSP	6	7572262	G	G	G	G
4	DSP	6	7572026	A/T	A/T	A/T	Not found
4	DSP	6	7576527	A	A	A	A
4	DSP	6	7584617	C/T	C/T	C/T	C/T
4	DSP	6	7585967	A	A	A	A
5	PKP2	12	32948970	GT/A	Not found	A	Not found
5	PKP2	12	32945721	C/A	C/A	C/A	Not found
5	PKP2	12	32945769	C/G	C/G	C/G	Not found
5	DSG2	18	28666526	+TAA	+TTAA	+TAA	Not found
5	DSP	6	7567970	T	T	T	Not found
5	DSP	6	7572026	A	A	A	Not found
5	DSP	6	7559633	A	A	A	Not found
5	JUP	17	39914070	A/C	A/C	A/C	Not found
5	JUP	17	39913645	A/G	A/G	A/G	Not found

Table 6.5: Table with Variant of Samples 4 and 5. There is reported the Sanger result and the output of CLC, GATK and CASAVA.

6.2.2 SNP Caller Results - Illumina Data

After alignments we performed the SNP Calling. SNP Callers take as input BAM files that are the binary format of SAM, the standard alignment output. We considered GATK and CLC Variant Probabilistic caller.
GATK required lots of step to produce the output and the pipeline took also lots of time (more or less one day per sample).
To evaluate the variant callers I focused my attention on the best samples, the samples 4 and 5 that are the ones with the highest coverage for Illumina sequencing. For SOLiD sequence I used the sample 6S that was the unique avaliable.
Sample 4 and 5 was analyzed using GATK, CLC and CASAVA. CASAVA is the standard suite for Illumina data analyses and performs alignment (with ELAND) and SNP/DIP Calling. Software evaluation was performed considering a series of known variant previously characterized using SANGER Technology. For each variant we checked the SANGER sequence quality and we checked the presence in the variant caller outputs.

Sample	Gene	Chr	Position	Exome position	Coverage
4	DSG2	18	29104698	IN	64
4	DSG2	18	29125854	IN	92
4	DSG2	18	29126670	IN	62
4	DSP	6	7542149	IN	9
4	DSP	6	7563983	IN	67
4	DSP	6	7572262	IN	39
4	DSP	6	7572026	OUT	32
4	DSP	6	7576527	IN	36
4	DSP	6	7584617	IN	74
4	DSP	6	7585967	IN	37
5	PKP2	12	32948970	OUT	8
5	PKP2	12	32945721	OUT	17
5	PKP2	12	32945769	OUT	10
5	DSG2	18	28666526	OUT	18
5	DSP	6	7567970	OUT	14
5	DSP	6	7572026	OUT	14
5	DSP	6	7559633	IN	5
5	JUP	17	39914070	OUT	12
5	JUP	17	39913645	OUT	30

Table 6.6: Positions of variants respect the enriched regions

Table 6.5 shows the results. In this table CASAVA seems to be the worst software but we have to consider that CASAVA extracts variants limited to the enriched regions. Lot of the position reported for samples 4 and 5 are out of the enriched regions (see table 6.6). CASAVA was discarded for its inability to find variants out of enriched regions.

At contrary, GATK and CLC are able to detect variants in all covered regions even if these regions are out of the exome.The performances of GATK and CLC are very similar but observing the table 6.5 we can see that GATK had some difficulties in detect indels (indel -/A in position 7542149 chromosome 6 for the sample 4 and indel -/TAA in position 28666526 chromosome 18 for the sample 5). The unique problem with CLC is the variant in position 32948970 chromosome 12 in the sample 5: here CLC called a variant in omozygosis but SANGER sequences found the same variant in eterozygosis. For better understanding this results I take into consideration also the coverage. Like reported in table 6.6 this variation has a low coverage (8X).
These results suggested that the most reliable software is CLC Variant Probabilistic Detector and I analyzed all the other samples with CLC. In the table 6.7 are reported the results.

Sample	Average Coverage	# of variant from Sanger	# of variant correct from CLC
4	51,65X	10	10
5	41,53X	10	9
6	25,07X	9	7
7	27,36X	10	10
2	9,24X	12	5
12	20,04X	9	6

Table 6.7: Number of variants found by Sanger compared with the CLC output

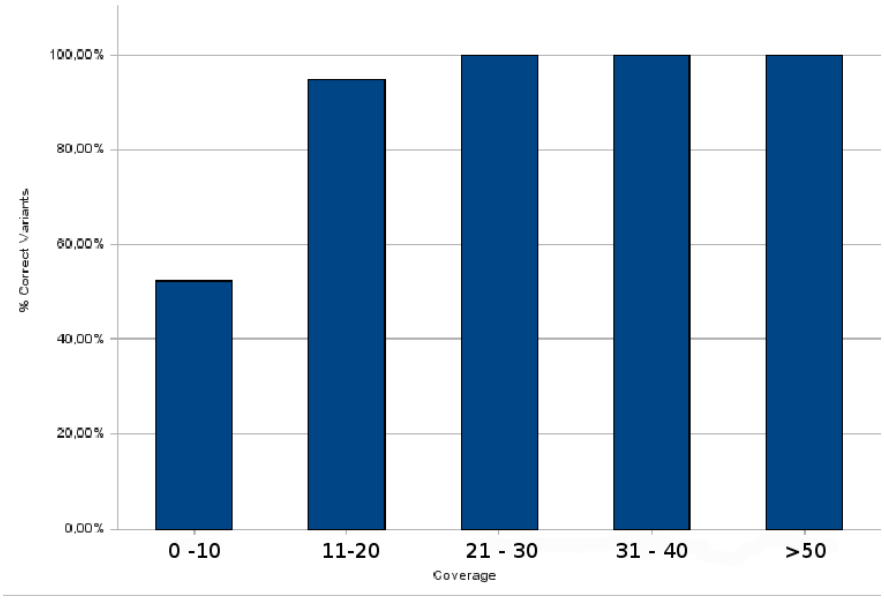

Figure 6.1: % of correct outputs of CLC divided by coverage

Analysing these data, it is clear that there is a stricly relationship among the coverage and the performance of the SNP caller, but the average coverage can be only an approximate parameter. More interesting is the relation among the result of CLC and the coverage of every variant. In the figure 6.1 I considered all the variants indipendently from the sample; I divided the coverage in 5 class and I considered the corrected prediction of CLC against the total of variants. Observing the figure 6.1, it appears that the minimum coverage for having reliable results is 20X. The total results divided per sample are reported in the supplementary materials.

6.2.3 SNP Caller Results - SOLiD Data

For the SOLiD data, we had only one sample (6S) and it is very difficult to extract some statistics having only one sample. For SOLiD data I take

Position	Chr	SANGER	CLC	PASS/GATK	BFAST/GATK
32974422	12	G/-	G/-	NotFound	NotFound
7558318	6	T/C	T/C	T/C	T/C
7578819	6	G	G/A	G	G
7578823	6	A	A/G	A	A
7584617	6	C/T	C/T	C/T	C/T
7585967	6	C	NotFound	NotFound	C
28673760	18	No Coverage	No Coverage	No Coverage	No Coverage
28672067	18	T/C	T/C	T/C	T/C

Table 6.8: Results of SOLiD Data

Algorithm	Total variants found	Variants annotated with dbSNP
CLC	115.681	52.283
BFAST/GATK	48.097	42.213
PASS/GATK	79.478	51.970

Table 6.9: Results of Variant Caller on SOLiD Data

into consideration CLC and GATK.
CLC was used starting from CLC mapping, while GATK was used starting from aligments obtained with PASS and BFAST[12][40][39].
Results are reported in table 6.8. Unlike with Illumina Data, CLC does not perfom very well, probably color space is more complex to align and specialized software like PASS and BFAST perform better.

It is important to consider that 50% of known variants are out of the enriched regions and one has no coverage. The total number of variants called by the three elaborations of sample 6S are reported in table 6.9.

The three algoritms found 34.305 common variants.
These data are very difficult to interpretate. Theoretically within the same sample we should obtain same data. Probably color space is very difficult to trait and the result can vary.

6.2.4 SNP Analyses

The table 6.10 reports the total number of variants called by CLC and the variants that are known according to DBSNP[13]. I considered only the samples sequenced with Illumina.

Sample	Total Number of Variants	Variants with DBSNP code
2	99.600	54.648
4	286.124	165.948
5	320.178	159.853
6	254.541	124.628
7	276.340	134.968
12	195.672	110.996

Table 6.10: Total number of Variant per sample and total number of known variants according to DBSNP

The 50% of called variants are known with a DBSNP code and the number of variant is stricly correlated to the coverage: samples with higher coverage have more variants called; probably the number of false positive increase with the coverage. The six samples share 30.028 variants and 18.374 are known in DBSNP.
Sample 6 and 6S are the same sample sequenced with Illumina(6) and SOLiD(6S). Comparing the variants called with CLC we see that they shared 47.957 variants (Sample 6S has 115.681 variants called using CLC variant caller). Practically all variants found by BFAST and GATK are in common with the PASS/GATK and the CLC ones.

6.2.5 Discussion

Currently, exome sequencing is one of the most challenge approach used to characterize human disease. Results depends on two factor: the mapping and the snp calling algorithms. Moreover results of mapping influence the snp calling results. We saw that changing alignment algorithm, change also the output of snp caller. The most difficult task is to understand the real position of a read on the reference taking into consideration sequencing errors and real differences. On the other hand, SNP caller must to be able to consider different level of coverage and differences in the quality of reads to right assign a variation in a particular coordinate of the reference. At the moment there is not a standard approach to calculate the variants of an exome sequencing, but, in this thesis I observed that CLC suite perform better than the other pipelines using illumina data. With SOLiD data CLC do not perform very well, GATK, using PASS or BFAST alignments, perfomed better.
GATK had problems in deletion/insertion recognization.
CLC performs very well when the coverage is >20X. Observing the table 6.4 we can say that the minumun average coverage for having a reliable snp calling result is at least 70X. At this average coverage we have at least the 80/90% of the exome covered with at least 20 indipendent reads and the results are very robust.

Additionally, CLC is very simple to use and it can be used also by biologists that do not have bioinformatics competences. It is very fast and can be run on laptop computer.
Using the Sanger sequences I tried also to calculate false positive and false negative. I analysed 7.989 nucleotides and I found only 2 false negative results. Observing the coverage I saw that the 2 false negative results is under the 20X coverage (the first is 8X, and the second is 1X), and the rest of nucleotides has very high coverage. These data confim the key role of the coverage in the snp calling results. In the 7.989 nucleotides analized I don't find any false positive results. These don't means that there aren't false positive, I belive that false posivite are present and that these false positive are stricly correlated to the coverage.
Filtering the data by coverage, the number of variants decrease drastically (see table 7.7 in supplementary data); probably also the number of false positive decrease.

Chapter 7

Supplementary Material

Sample	Position	Chr	SANGER	CLC	Coverage
4	29104698	18	C/T	C/T	64
4	29125854	18	A/G	A/G	92
4	29126670	18	C/T	C/T	62
4	7542149	6	-/A	-/A	9
4	7563983	6	G	G	67
4	7572262	6	G	G	39
4	7572026	6	T/A	T/A	32
4	7576527	6	A	A	36
4	7584617	6	C/T	C/T	74
4	7585967	6	C	C	37

Table 7.1: Illumina Sample 4 Results

Sample	Position	Chr	SANGER	CLC	Coverage
5	32948970 – 71	12	T/AC	AC	8
5	32945721	12	C/A	C/A	17
5	32945769	12	G/C	G/C	10
5	28666526	18	-/TAA	-/TAA	15
5	7567970	6	T	T	14
5	7572026	6	A	A	14
5	7559633	6	A	A	5
5	39914070	17	G/T	G/T	12
5	39913645	17	T/C	T/C	30

Table 7.2: Illumina Sample 5 Results

Sample	Position	Chr	SANGER	CLC	Coverage
6	32974422	12	G/-	G/-	18
6	7558318	6	T/C	NotFound	11
6	7578819	6	G	G	10
6	7578823	6	A	A	10
6	7584617	6	C/T	C/T	28
6	7585967	6	C	C	10
6	28673760	18	G/A	NotFound	1
6	28672067	18	T/C	T/C	12

Table 7.3: Illumina Sample 6 Results

Sample	Position	Chr	SANGER	CLC	Coverage
7	7567970	6	C/T	C/T	8
7	7572262	6	A/G	A/G	20
7	7572026	6	T/A	T/A	11
7	7578189	6	G/A	G/A	28
7	7578816	6	G	G	15
7	7578823	6	A	A	12
7	29104714	18	A/G	A/G	37
7	39913645	17	T/C	T/C	11
7	39912145	17	A/T	A/T	10
7	39911771	17	G/A	G/A	34
7	7585967	6	C	C	20

Table 7.4: Illumina Sample 7 Results

Sample	Position	Chr	SANGER	CLC	Coverage
2	7542149	6	A/+A	No Coverage	0
2	7563983	6	G	G	5
2	7565227	6	A/T	No Coverage	0
2	7576527	6	G/A	NotFound	7
2	7584617	6	T/C	T/C	7
2	28649057	18	G	NotFound	2
2	32994007	12	G/-	NotFound	3
2	32977104	12	-/A	-/A	5
2	29104553	18	T/C	T/C	10
2	29104569	18	A/G	A/G	11

Table 7.5: Illumina Sample 2 Results

Sample	Position	Chr	SANGER	CLC	Coverage
12	33030802	12	A/G	NotFound	4
12	30049475	12	G/A	No Coverage	0
12	32949029	12	G	G	11
12	33021819	12	C	NotFound	2
12	28669496	18	C	C	8
12	7563983	6	G	G	25
12	7572262	6	G	G	15
12	7576527	6	A	A	13
12	7584617	6	C/T	C/T	31

Table 7.6: Illumina Sample 12 Results

Sample	Number of variants with coverage $>$ 20X	Total Number of variants
2	22.009	99.600
4	140.088	289.124
5	127.290	320.178
6	78.117	254.541
7	84.979	276.340
12	57.555	195.672

Table 7.7: Variants with a coverage higher than 20X

Bibliography

[1] http://my454.com/products/technology.asp.

[2] http://454.com/products/gs-flx-system/index.asp.

[3] http://www.illumina.com/systems/hiseq_comparison.ilmn.

[4] http://my454.com/products/analysis-software/index.asp.

[5] http://www.clcbio.com.

[6] http://samtools.sourceforge.net/SAM1.pdf.

[7] http://www.bioportal.uio.no/appinfo/show.php?app=CLOTU.

[8] http://www.mothur.org/.

[9] http://qiime.org/.

[10] http://www.clcbio.com/wp-content/uploads/2012/10/whitepaper-on-CLC-read-mapper.pdf.

[11] http://www.clcbio.com/wp-content/uploads/2012/08/whitepaper-probabilistic-variant-caller-1.pdf.

[12]

[13]

[14] Lin Liu at al. Comparison of next-generation sequencing system. *Journal of Biomedicine and Biotechnology*, 2012, 2012.

[15] Michel J. Bamshad at al. Exome sequencing as a tool for mendelian disease gene discovery. *Nature Review*, 12, 2011.

[16] Biesecker. Exome sequencing makes medical genomics a reality. *Nature Genet.*, 2010.

[17] Trapnel C. and Salzberg S.L. How to map billions of short reads onto genome. *Nat.Biotechnol.*, 2009.

[18] Bilardi A. Caniato E. Forcato C. Manavski S. Vitulo N. Campagna D., Albiero A. and Valle G. Pass: a program to align short sequences. *Bioinformatics*, 2009.

[19] Hoischen A. et al. Massevely parallel sequencing of ataxia genes after array-based enrichment. *Hum.Mut.*, 2010.

[20] Kamie K. Teer et al. Exome sequencing: the sweet spot before whole genomes. *Human Molecular Genetics*, 2010.

[21] Krumm N. et al. Copy number variation detection and genotyping from exome sequence data. *Genome Research*, 2012.

[22] Kryukov G.V. et al. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am.J.Hum.Genet.*, 2007.

[23] Li R. et al. De nove assembly of human genome with massively parallel shot read sequencing. *Genome Reserach*, 2010.

[24] McKenna A.H. et al. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data.

[25] Ng S.B. et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 2009.

[26] Ng. S.B. et al. Exome sequencing identifies the cause of a mendelian disorder. *Nature Genet.*, 2010.

[27] Pilichou K. et al. Assessing the significance of pathogenic mutations and autopsy findings in the light of 2010 arrhythmogenic right ventricular cardiomyopathy diagnostic criteria.

[28] Rasmus N. et al. Genotype and snp calling from next-generation sequencing data.

[29] Rava G. et al. Next generation sequencing in research and diagnostics of ocular birth defects. *Maol.Genet.Metab.*, 2010.

[30] Simpson JT et al. Abyss: a parallel assembler for short read sequence data. *Genome Reserach*, 2009.

[31] Stenson P.D. et al. Human gene mutation database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. *Hum. Genomics*, 2009.

[32] Teeling H. et al. Current opportunities and challenges in microbial metagenome analysis–a bioinformatic perspective. *Brief Bioinform.*, 2012.

[33] T.Foehlich et al. High-throughput nucleic acid analysis. *U.S.Patent*, 2010.

[34] Yi X. et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, 2010.

[35] Zerbino D.R. et al. Velvet: algorithms for de novo short assembly using de bruijn graphs. *Genome Reserach*, 2008.

[36] Li H. and Durbin R. Fast and accurate long-read alignment with burrows-wheeler transform.

[37] Li H. and Durbin R. Fast and accurate short read alignment with burrows-wheeler transform.

[38] Li H. and Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 2009.

[39] Nelson SF. Homer N, Merriman B.

[40] Nelson SF. Homer N, Merriman B. Bfast: An alignment tool for large scale genome resequencing. *PLoS ONE*, 2009.

[41] Pop M Salzberg SL. Langmead B, Trapnell C. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology*, 2009.

[42] BJ et al. O'Roak. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 2012.

[43] Konrad Paszkiewicz and David J. Studholme. De novo assembly of short sequence reads. *Briefings in bioinformatics*, 2010.

[44] Elaine R.Mardis. The impact od next-generation sequencing technology on genetics. *Cell*, 2008.

[45] Antonarakis S.E. and Beckmann J.S. Mendelian disorders deserve more attention. *Nature Rev. Genet.*, 2006.

[46] Smith T.F. and Waterman M.S. Identification of common molecular subsequences.

[47] Snyder M. Wang Z, Gerstein M. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.*, 2009.