# Acknowledgments

Questa tesi mi ha dato la possibilità di vivere una delle più belle esperienze della mia vita: vivere a Ferrara. Ricorderò sempre gli anni trascorsi in questa splendida città con affetto, con un sorriso e con tantissima nostalgia. Ringrazio chi l'ha vissuta con me, ringrazio di cuore il "boss" Stefano, che mi ha dato questa opportunità di crescere e che mi ha insegnato i segreti della bioinformatica. Grazie ai miei supercompagni Marco e Marco, per aver condiviso con me il lavoro di tutti i giorni. Ringrazio di cuore tutte le altre persone fantastiche e speciali che ho incontrato qui: Ilaria, Sara, Edi.

E tutte coloro che ho incrociato anche solo per brevi istanti. Non vi dimenticherò mai!!!

Grazie al mio fantastico laboratorio dell'Ospedale Riuniti di Bergamo. Grazie alla Dott.ssa Maria Iascone perché sono la SMBI che sono grazie a lei! E ovviamente grazie a tutte le girls,e soprattutto Laura, Dani e Anna Rita, perché mi supportano (e sopportano) ogni giorno!

Infine grazie alla happy family Sana, per aver creduto in me e per essere sempre i miei principali sostenitori!!!

L'ultimo pensiero, che poi è sempre il primo, è per mio marito Sergio. Perché è tutto.

*Alla mia mamma e al mio papà*

*A Sergio*

*“...Credo di poter affermare che nella ricerca scientifica*
*né il grado di intelligenza né la capacità di eseguire e portare a*
*termine il compito intrapreso siano fattori essenziali*
*per la riuscita e per la soddisfazione personale.*
*Nell'uno e nell'altro contano maggiormente la totale dedizione*
*e il chiudere gli occhi davanti alle difficoltà:*
*in tal modo possiamo affrontare i problemi che altri,*
*più critici e più acuti, non affronterebbero...”*

Rita Levi Montalcini

# Contents

**Acknowledgments** ..... i

**List of Tables** ..... vi

**List of Figures** ..... vii

**1 Introduction** ..... 1
1.1 Sanger capillary sequencing ..... 1
1.2 Next-generation sequencing ..... 2
1.2.1 Library preparation ..... 5
1.2.2 DNA amplification ..... 6
1.2.3 Sequencing and imaging ..... 7
1.2.3.1 Roche/454 ..... 8
1.2.3.2 Illumina ..... 9
1.2.3.3 Life Technologies/Applied Biosystems ..... 11
1.2.3.4 Life Technologies/Ion Torrent ..... 11
1.3 Application of next-generation sequencing ..... 12
1.3.1 Analysis of cancer genome by next-generation sequencing ..... 13
1.4 Computational challenges ..... 14
1.4.1 IT infrastructure ..... 15
1.5 Bioinformatic analysis ..... 15
1.5.1 Primary Analysis ..... 16
1.5.2 Secondary Analysis ..... 17
1.5.2.1 Alignment ..... 17
1.5.2.2 Sequence Alignment/Map format ..... 19
1.5.2.3 Variant calling ..... 20
1.5.3 Tertiary Analysis ..... 21
1.5.3.1 Annotation ..... 22
1.5.3.2 Prioritazion and intepretation ..... 23

**2 Genomic analysis in complex diseases** ..... 24
2.1 Ras/Raf/MAPK pathway ..... 24
2.1.1 Colorectal cancer ..... 28
2.1.1.1 CRC treatment ..... 30
2.1.1.2 anti-EGFR monoclonal antibodies ..... 31

2.1.1.3 Kras mutation in colorectal cancer ..... 32
2.1.1.4 Effect of Kras mutations on anti-EGFR therapy ..... 33
2.1.2 RASopathies ..... 34
2.1.2.1 Noonan syndrome ..... 35
2.1.2.2 Leopard syndrome ..... 38
2.1.2.3 Costello syndrome ..... 40

**3 Aims of the present study** ..... 42

**4 Materials and Methods** ..... 43
4.1 Patients ..... 43
4.1.1 Colorectal cancer patients ..... 43
4.1.2 RASopathies patients ..... 43
4.1.2.1 Patient 1 ..... 44
4.1.2.2 Patient 2 ..... 44
4.1.2.3 Patient 3 ..... 44
4.2 Sample sequencing ..... 45
4.2.1 Colorectal cancer sequencing ..... 45
4.2.2. RASopathies sequencing ..... 46
4.3 Bioinformatic analysis ..... 46
4.4 Prioritization and intepretation ..... 47

**5 Results** ..... 48
5.1 Colorectal cancer ..... 50
5.2 RASopathies ..... 50
5.2.1 Patient 1 ..... 52
5.2.2 Patient 2 ..... 54
5.2.3 Patient 3 ..... 57

**6 Conclusion** ..... 57

**References** ..... 59

**Appendix** ..... 67
List of publication ..... 67
Abbreviations ..... 69

# List of Tables

Table 1.1: Comparison of performance among different NGS platforms.....3
Table 1.2: Applications of next-generation sequencing technologies .....12
Table 1.3: Phred quality scores are logarithmically linked to error probabilities .....17

# List of Figures

Figure 1.1: Schematic workflow of Sanger sequencing method ..... 2
Figure 1.2: Reduction of cost per base of DNA sequencing ..... 3
Figure 1.3: Increase of number of entries in dbSNP (2002-2012). ..... 3
Figure 1.4: Cover of Nature and Time journals. ..... 4
Figure 1.5: Paired-end library ..... 5
Figure 1.6: Mate-paired library ..... 5
Figure 1.7: DNA amplification procedures: emulsion PCR. ..... 6
Figure 1.8: DNA amplification procedures:bridge amplification ..... 7
Figure 1.9: Roche/454 sequencing workflow ..... 9
Figure 1.10: Illumina sequencing workflow ..... 10
Figure 1.11: Life Technologies/Applied Biosystems workflow ..... 11
Figure 1.12: Workflow of NGS data processing ..... 15
Figure 1.13: FASTQ format of raw data ..... 16
Figure 1.14: Graphical representation of alignment ..... 19
Figure 1.15: Alignment file in SAM format. ..... 20
Figure 2.1: Ras//MAPK pathway. ..... 25
Figure 2.2: Multiple alignment between human RAS protein ..... 27
Figure 2.3: Key events in the field of Ras research. ..... 28
Figure 2.4: Colorectal cancer. A) biopsy B) histological section ..... 29
Figure 2.5: Mechanism of anti-EGFR monoclonal antibodies ..... 31
Figure 2.6: Ras/MAPK signalling pathway and related genetic syndromes ..... 34
Figure 2.7: Clinical features in Noonan syndrome. ..... 36
Figure 2.8: Phenotypic features in Leopard syndrome. ..... 40
Figure 2.9: Dysmorphic craniofacial features in Costello syndrome ..... 41
Figure 4.1: Pedigree of family of Patient 3 ..... 45
Figure 5.1: Structural modeling of amino acid change Gly12Val in protein ..... 48
Figure 5.2: Visualization of KRAS-mutant SNV using IGV Browser. ..... 50
Figure 5.3: PTPN11 mutation in Patient 1 ..... 52
Figure 5.4: Patient 2 sequencing reads overlapping the missense mutation ..... 53
Figure 5.5: Sanger sequencing validation ..... 54

Figure 5.6: Mutation in MYH7 gene in brother and sister of Patient 1.. .....54
Figure 5.7: Identification of HRAS mutation.....55
Figure 5.8: Pedigrees in which the MYH7 and HRAS mutations segregated.. .....56

# Chapter 1

# Introduction

Since the genetic information is embedded in the order of nucleotides in the DNA and RNA molecules, sequencing technology to determine the order of nucleotides in the DNA or RNA has always been one of the most powerful and fundamental tools in molecular biology. Determining the sequence of DNA can reveal the secrets contained in the genetic code of a person, his susceptibility to diseases and his response to drug treatments.

The sequencing technology has been evolving and improving during the past a few decades, in terms of accuracy and throughput. Recently, several sequencing platforms have been developed, which have very high-throughput and low cost comparing to the traditional Sanger sequencing technology.

## 1.1 Sanger capillary sequencing

Sanger sequencing is a method of DNA sequencing, based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication [Sanger et al, 1977] (Figure 1.1). Developed by Frederick Sanger in 1977, it was the most widely-used sequencing method for approximately 25 years. The sequencing biochemistry takes place in a 'cycle sequencing' reaction, in which cycles of template denaturation, primer annealing and primer extension are performed. The primer is complementary to known sequence immediately flanking the region of interest. Each round of primer extension is stochastically terminated by the incorporation of fluorescently labeled dideoxynucleotides (ddNTPs). In the resulting mixture of end-labeled extension products, the label on the terminating ddNTP of any given fragment corresponds to the nucleotide identity of its terminal position. Sequence is determined by high-resolution electrophoretic separation of the single-stranded, end-labeled extension products in a capillarybased polymer gel. Laser excitation of fluorescent labels as fragments of discreet lengths exit the capillary, coupled to four-color detection of emission spectra, provides the readout that is represented in a Sanger sequencing electropherogram, that has been decoded into DNA

sequence [Smith et al, 1986; Ansorge et al, 1987]. Simultaneous electrophoresis in 96 or 384 independent capillaries provides a limited level of parallelization. After three decades of gradual improvement, the Sanger biochemistry can be applied to achieve read-lengths of up to ~1,000 bp, and per-base 'raw' accuracies as high as 99.999%.

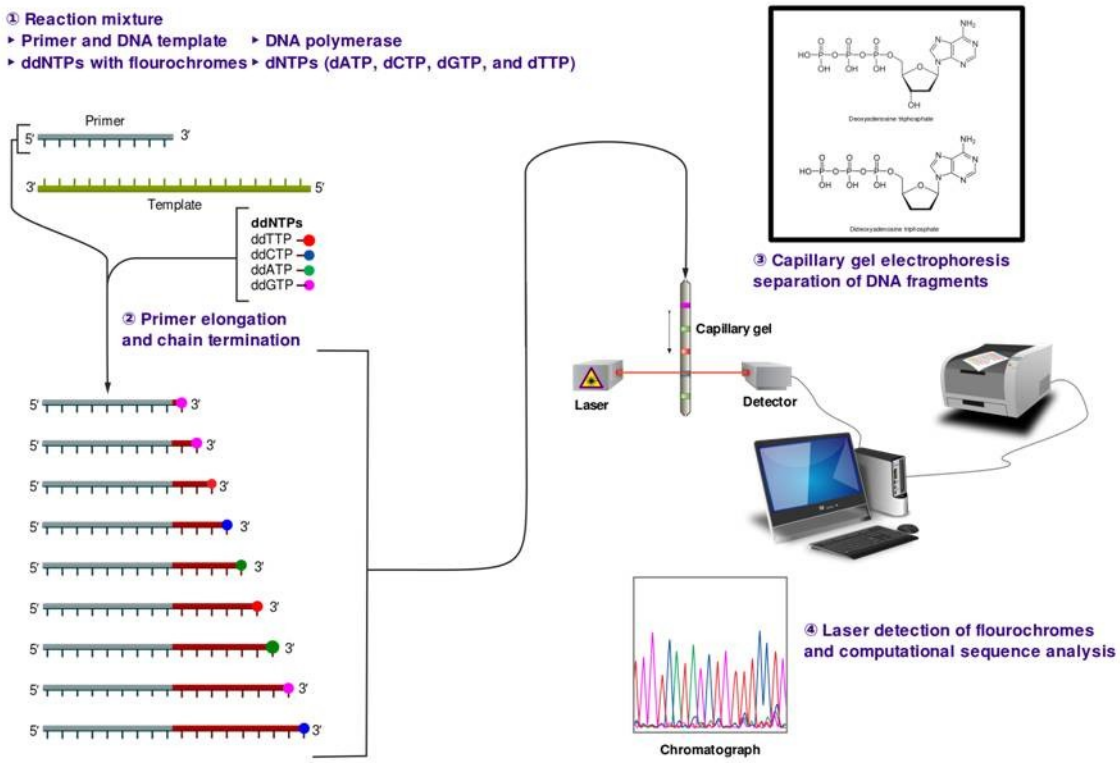


**Figure 1.1:** Schematic workflow of Sanger sequencing method. Source: *http://en.wikipedia.org*.

## 1.2 Next-generation sequencing

After the completion of the first human genome sequence in 2004 [Lander et al, 2001; Venter et al, 2001], the growing need to sequence a large number of individual genomes in a fast, low-cost and accurate way has directed a shift from traditional Sanger sequencing methods towards new high-throughput genomic technologies. In 2005, the first massively parallel DNA sequencing platforms emerged, ushering in a new era of next-generation sequencing (NGS) which allows sequencing at unprecedented speed in combination with low costs per base (Figure 1.2). As a consequence, the number of sequencing related data stored in public available databases has increased significantly (Figure 1.3).

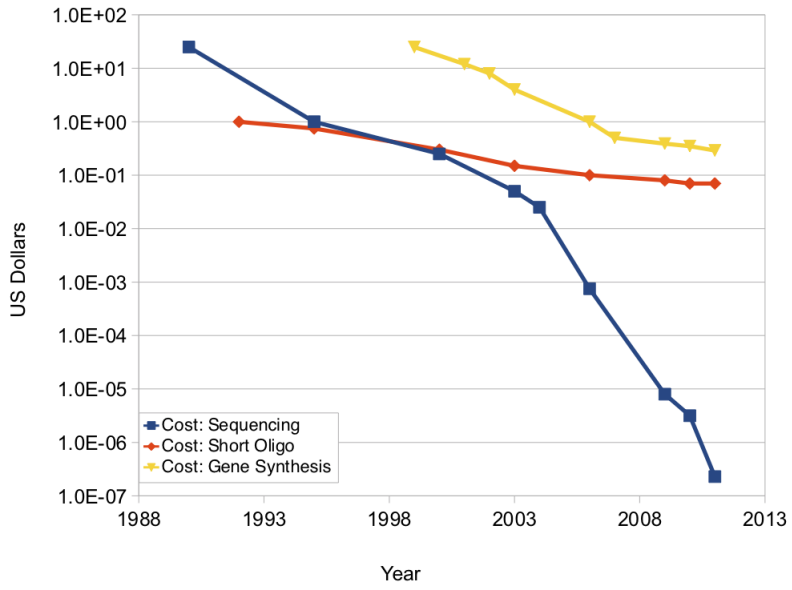


**Figure 1.2:** Reduction of cost per base of DNA sequencing . Source: h*ttp://www.synthesis.cc*.

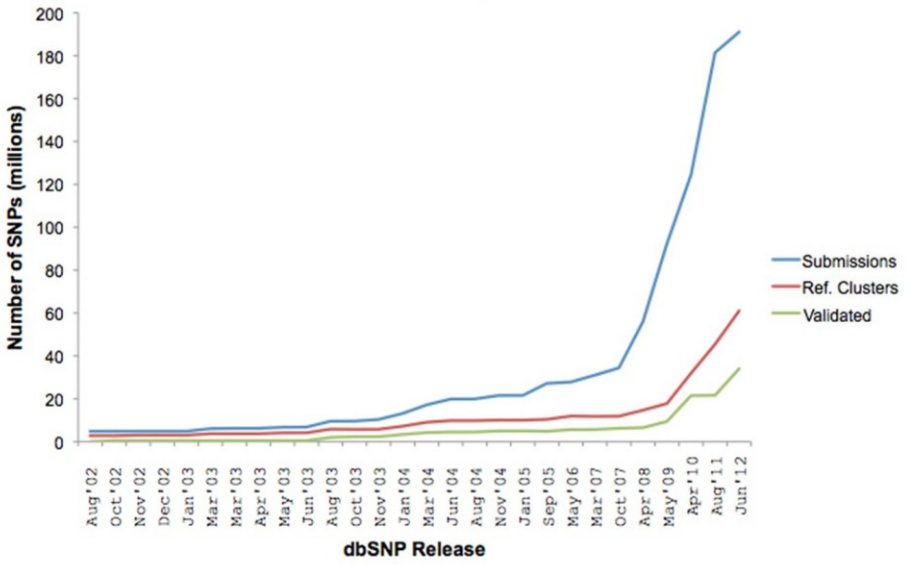


**Figure 1.3:** Increase of number of entries in dbSNP (2002-2012). Source: *http://massgenomics.org*.

By applying genome-wide sequencing with high-throughput platforms the 1000 Genomes project (1KGP) sequenced the complete genome of 185 individuals from four populations, and analyze targeted exons of 697 individuals from seven populations within only two years [Kaiser et al, 2008; 1000 Genomes Project Consortium et al, 2010]. The aim was not only

provide a comprehensive resource on human genetic variation, but also investigate the relationship between genotype and phenotype (Figure 1.4).


**Figure 1.4:** Cover of *Nature* and *Time* journals. The impact of 1KGP on scientific community but also in the politicl and economic world was enormous.

The variety of NGS features makes it likely that multiple platforms will coexist in the marketplace, with some having clear advantages for particular applications over others. In general, each platform embodies a complex interplay of enzymology, chemistry, high-resolution optics, hardware, and software engineering [Shendure et al, 2008; Metzker, 2010]. By different approaches, each technology seeks to amplify single strands of a fragment library and perform sequencing reactions on the amplified strands. The fragment libraries are obtained by annealing platform-specific linkers to blunt-ended fragments generated directly from a genome or DNA source of interest.

The main difference respect to traditional DNA sequencing approach is that molecules can be selectively amplified by PCR because of the presence of adapter sequences.

Because the presence of adapter sequences means that the molecules then can be selectively amplified by PCR, no bacterial cloning step is required to amplify the genomic fragment in a bacterial intermediate as is done in traditional sequencing approaches.

Specifically, the sequencing process could be grouped into library preparation, amplification reaction, sequencing and imaging, and data analysis [Metzker, 2010]. In the following sections, the experimental stages will be discussed, while paragraphs 1.3 and 1.4

will be dedicated to computational challenges introduces by NGS and data analysis procedures.

### 1.2.1 Library preparation

Library preparation is the step of preparation of DNA templates [Roe, 2004]. This process includes the random breaking of DNA sample into smaller fragments, ligating adapter sequences to allow the later use of universal primers, and amplifying the produced DNA templates.

Thre are three different protocols:

- ✓ SINGLE-END library (SE) is created by randomly shearing genomic DNA (gDNA) or complementary DNA (cDNA) into fragments which are less than 1 kb in size;

- ✓ PAIRED-END library (PE) is created as the SE protocol by sequencing primer (SP) sites are ligated at each end. After analyzing the first read with SP1 the templates are regenerated and the second read is sequenced by the use of SP2 (Figure1.5);

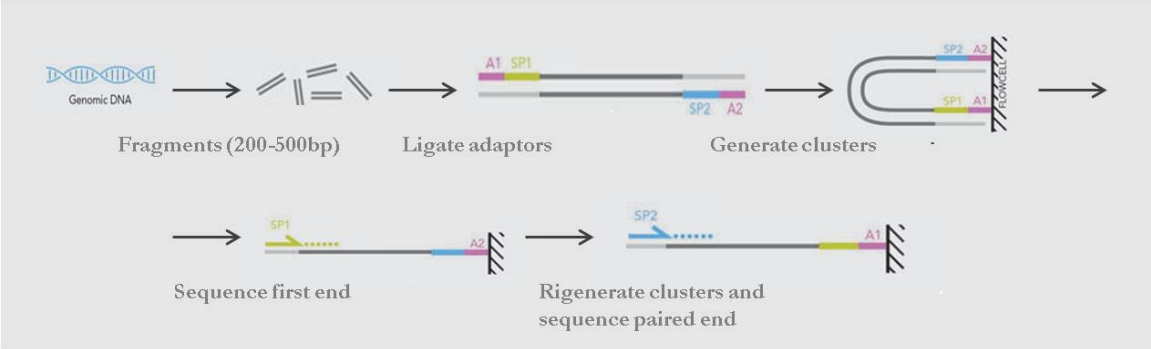


**Figure 1.5:** Paired-end library. Source: *www.illumina.org* (adapted).

- ✓ MATE-PAIRED library (MP) is created by shearing DNA with 2-5 kb in size labeled at the ends, circularized, and again linearized by cutting the cycles. Only fragments containing the label and therefore both ends of the original DNA fragment are selected and sequenced (Figure 1.6).

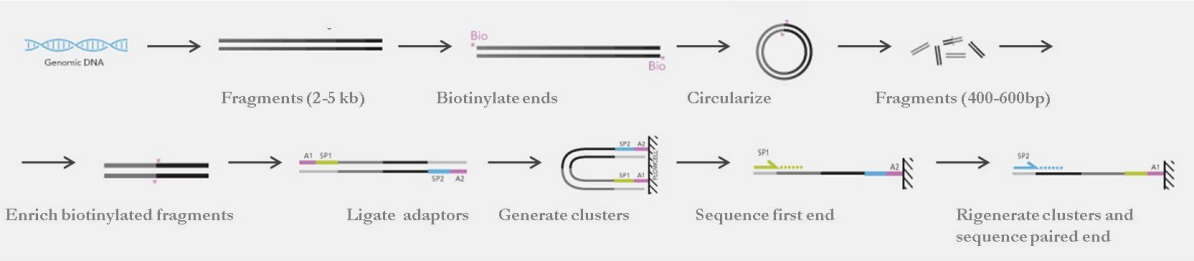


**Figure 1.6:** Mate-paired library. Source: *www.illumina.org* (adapted).

## 1.2.2 DNA amplification

After the library preparation, NGS technologies perform DNA amplification to increase the signal intensity for nucleotide detection. The most commonly used amplification procedures are:

- ✓ Emulsion PCR in which single-stranded DNA (ssDNA) hybridizes onto oligonucleotide bound beads, ideally leaving one DNA template per bead (Figure 1.7). The beads are part of water-inoilmicro-emulsions which additionally contain all necessary components for PCR. After PCR amplification within these micro-reactors, up to thousands of complementary DNA strands are covalently bound to the beads. The original DNA template strands are washed away and the beads are purified and immobilized for later sequencing [Dressman et al, 2003; Metzker, 2010].

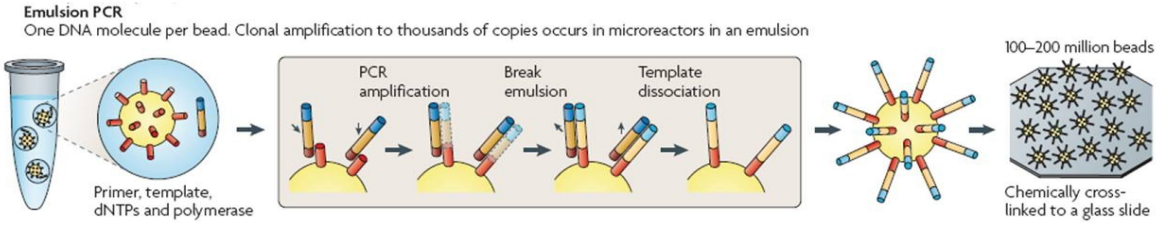


**Figure 1.7:** DNA amplification procedures: emulsion PCR. [Metzker, 2010].

- ✓ Bridge amplification in which the main device is a solid surface which is densely coated with forward and reverse primers. ssDNA templates anneal randomly to the surface and their complementary strands are built [Metzker, 2010]. After denaturation, a washing step removes all original template DNA. The remaining covalently bound complementary strands bind over to nearby reverse primers enabling the creation of yet another replicated strand. The DNA is denaturated

again to yield ssDNA and the next cycle of primer annealing and DNA replication is started. Thereby, clusters of DNA are produced all over the solid surface. As a last step prior to sequencing the template DNA is cleaved and washed away (Figure 1.8).

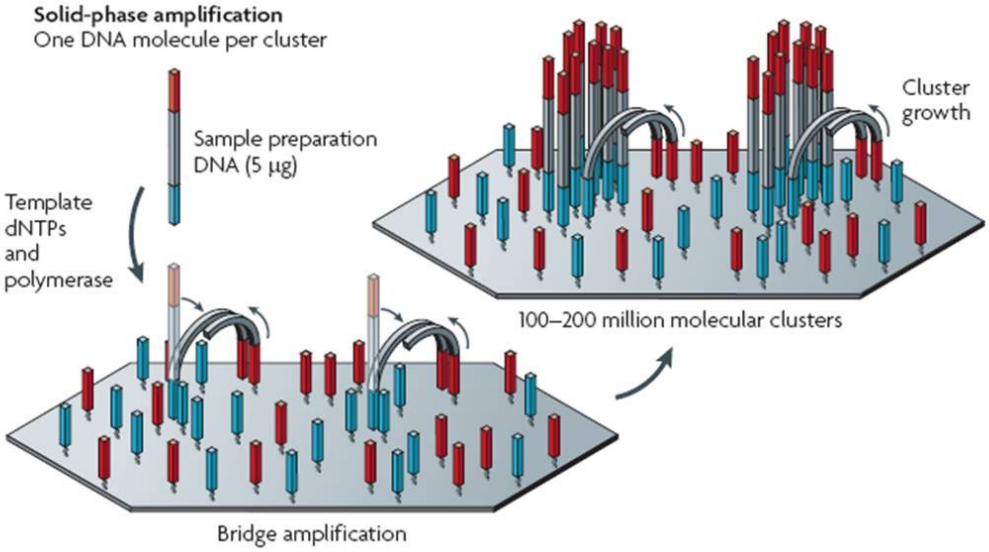


**Figure 1.8:** DNA amplification procedures: bridge amplification [Metzker, 2010].

## 1.2.3 Sequencing and imaging

The platforms for massively parallel DNA sequencing read production widely used up to today are distributed from Roche/454 (Eurofins MWG Operon; Huntsville, AL, USA), Illumina ((San Diego, CA, USA), Life Technologies/Applied Biosystems (Foster City, CA, USA) and Life Technologies/Ion Torrent (South San Francisco, CA, USA) (Table 1.1).

| Company | Platform | Sequencing | Throughput |
|---|---|---|---|
| Roche/454 | GS FLX-T-XL<br>GS Junior | Emulsion PCR/ Pyrosequencing | 700Mb/23h<br>35Mb/10h |
| Illumina | HiSeq 2000<br>MiSEQ | Bridge PCR/Sequencing-by-synthesis | 105-600Gb/ 2-11d<br>540Mb-7Gb/ 4-39h |
| Life Tech/ ABI | SOLID™ 4 | Emulsion PCR/Sequencing-by-ligation | 25-100Gb/ 3.5-16d |

| | | | |
|---|---|---|---|
| Life Tech/Ion Torrent | Ion PGM™ | Emulsion PCR/Ion semiconductor sequencing | 300Mb-1Gb/ 0.9-4.5h |

**Table 1.1:** Comparison of performance among different NGS platforms. Abbreviations: ABI, Applied Biosystems; d, day; h, hours.

All the instruments implement a process in which the synthesis of a complementary DNA strand is used to determine the DNA sequence and the use of a template amplification step to increase signal intensity for nucleotide identification.

Recently, three benchtop high-throughput sequencing instruments are become available, allowing the implementation of high-throughput sequencing in diagnostic settings. These are the 454 GS Junior (Roche/454), MiSeq (Illumina) and Ion Torrent PGM (Life Technologies/Ion Torrent) [Loman et al, 2012].

These platforms are smaller and cheaper that the large instruments and allow the analysis of patients in parallel in short time, since they generate a reduced throughput.

### 1.2.3.1 Roche/454

The Roche/454 was the first commercially available NGS platform, introduced in 2005. The platform analyzes DNA by pyrosequencing technology [Ronaghi et al, 1998; Ronaghi, 2001] in which nucleotides are detected based on the release of pyrophosphate. After the amplification of target DNA with emulsion PCR, beads, pyrosequencing enzymes, and pyrosequencing sulfates are loaded into a pico titer plate device, which places each bead into an addressable position within the plate. After the sequencing primer has been annealed the first sequencing cycle is started. Each cycle, nucleotides of one type (either dATP, dCTP, dGTP, or dTTP) are added to the plate. When incorporated, the nucleotide releases a pyrophosphate thereby triggering a series of downstream reactions. The use of luciferase in these reactions causes emission of a light signal which is proportional to the amount of integrated nucleotides. This signal is detected by a CCD camera and image information is stored for further processing. The remaining nucleotides are washed away and the next type of nucleotides is added to the plate (Figure 1.9) [Ansorge, 2009].

The latest 454 GS FLX platform with Titanium chemistry can produce approximately one million reads with lengths of up to 1000 bp per instrument run. Despite the higher costs, this platform for its long reads is best suitable *for de novo* assembly, metagenomic and characterization and analysis of microbiome.

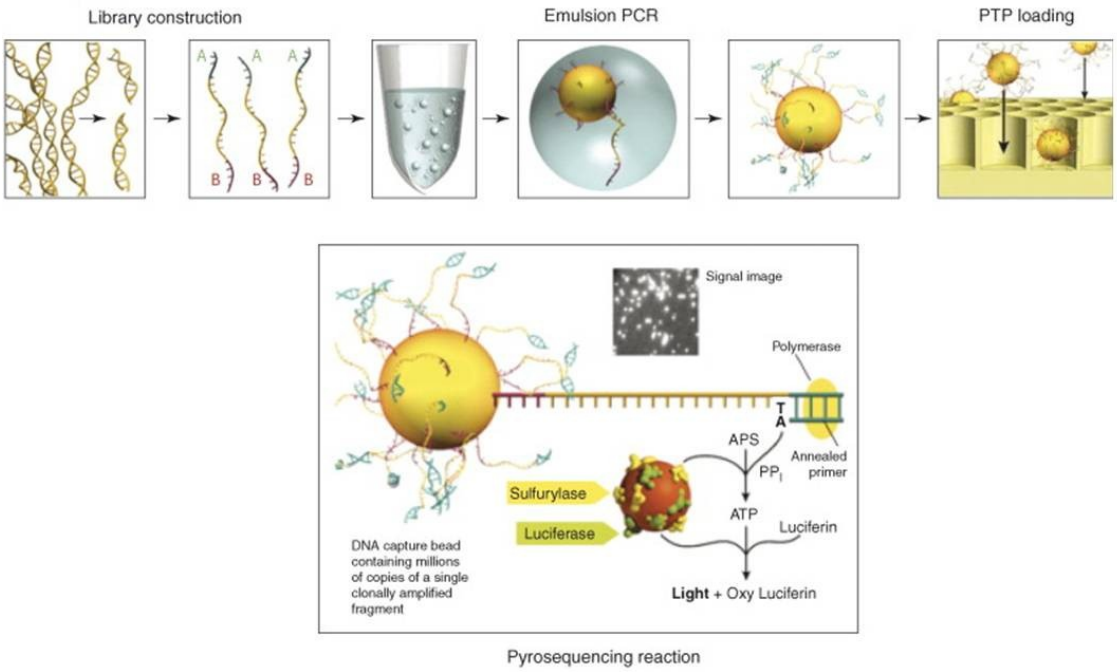


**Figure 1.9:** Roche/454 sequencing workflow [Mardis, 2008].

The benchtop sequencer GS Junior System, available since 2010, is based on the chemistry of 454 Sequencing technology and was developed to allow the long read sequencing in reduced time (10 hours sequencing and 2 hours data processing), with lower set-up and running costs scaled for the needs of diagnostic laboratories.

### 1.2.3.2 Illumina

The Illumina sequencing system employs an array-based DNA sequencing-by-synthesis technology with reversible terminator chemistry. This technology analyzes different DNA samples in parallel by the use of bridge amplification and dye-terminated nucleotides (Figure 1.10).

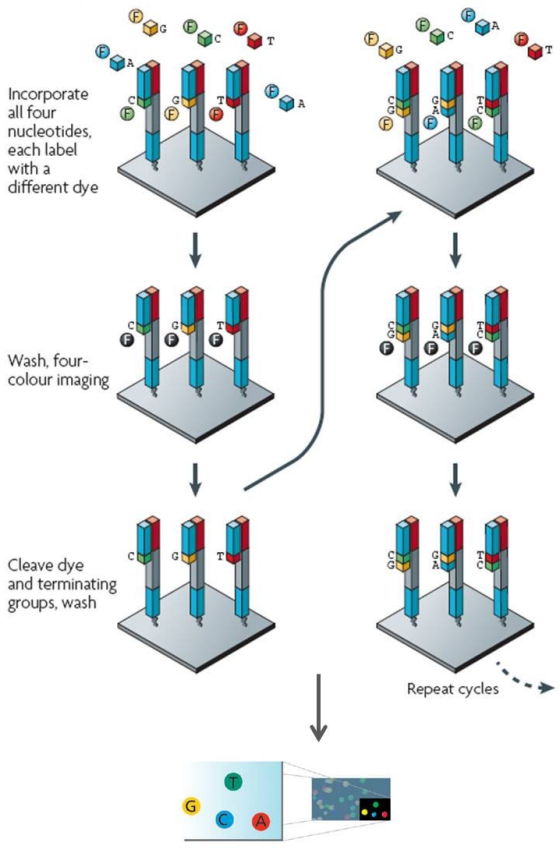


**Figure 1.10:** Illumina sequencing workflow (adapted from Mardis, 2008).

After primer hybridization, nucleotides which are labeled by different fluorescent dyes are added to the slide. In each sequencing cycle, only one nucleotide is incorporated into the complementary strand due to its attached terminator group. A washing step removes all remaining free nucleotides before the newly added base is identified. Finally, the terminator and dye group are cleaved off, and the sequencing cycle is restarted. The first Illumina sequencing platform, the Genome Analyzer (GA), originally produced 35-bp reads and generate more than 1 Gb of high-quality sequence per run in 2–3 days. Latest platforms, GA IIX and HiSeq 2000, allow to increase fragment length up to 150bp and improve the throughput.

Recently, Illumina released a benchtop high-throughput sequencing instrument, MiSeq, developed for clinical settings and diagnostics routine. The MiSeq system is based on the existing Illumina sequencing-by-synthesis chemistry, but with a drastic reduction run times and cost, compared to other platforms.

### 1.2.3.3 Life Technologies/Applied Biosystems

The SOLiD™ system is based on ligating fluorescently labeled dinucleotide probes to the DNA template [Tomkinson et al, 2006; McKernan et al, 2009]. After emulsion PCR, beads are covalently bound to a glass slide and universal sequencing primers, ligases, and a pool of labeled dinucleotide probes are added to the glass slide. The DNA sequence is determined by recording the color code, representing the first two bases of the dinucleotide, in several cycles of DNA ligation and cycles of primer reset. Each nucleotide (color-encoded) in the template is read twice by two fluorescent signals, reducing the error rate (Figure 1.11).

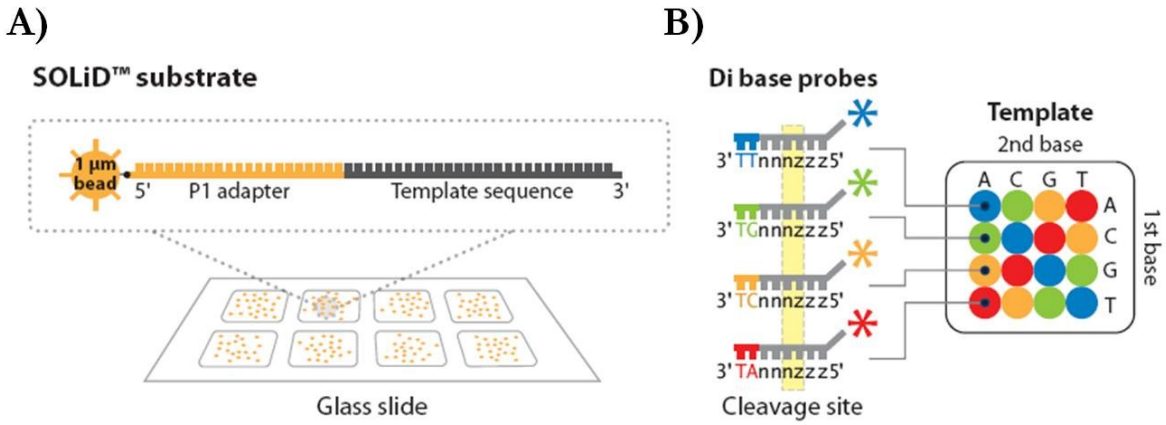


**Figure 1.11:** Life Technologies/Applied Biosystems sequencing workflow. A) The SOLiD™ substrate consists of an emulsion PCR bead, its covalently bound primer site (two sites for MP and PE), and the DNA template to be sequenced. b) SOLiD™ uses '1,2-probes' (a version of a dinucleotide probe) where the first and second nucleotides are analyzed. The remaining six bases consist of either degenerated or universal bases [Metzker, 2010]. Each dye represents 4 of 16 possible dinucleotide sequences. (adapted from Mardis, 2008).

.

### 1.2.3.4 Life Technologies/Ion Torrent

Ion Torrent™ is based on semiconductor sequencing approach, which combines computer software with integrated circuits and complementary metal-oxide semiconductors (CMOS). The technology also adopts an electrochemical detection system, the ion-sensitive field-effect transistors (ISFET), which can detect ions as they are released by DNA polymerase during sequencing by DNA synthesis. All of these electronics are focused on detecting and analyzing the release of a hydrogen ion (or proton) which occurs each time a nucleotide triphosphate is added. The proton release causes a slight pH shift which is detected by a

CMOS sensor. Each chip has at least 1.2 million sensors. These are composed of a well containing the dNTP and an acrylamide bead with a DNA template. Just beneath the well lies the metal oxide sensing layer, which itself lies over a sensor plate and floating metal "gate" that transmits electronic information (the pH changes) to the semiconductor. This technology differs from other sequencing technologies in that no modified nucleotides or optics are used. Ion Torrent Personal Genome Machine (PGM$^{TM}$), available since 2011, is benchtop NGS platform from Ion Torrent, which enables fast, affordable, genome-sequencing with a high throughput (80–100 Mb/h) and with up to 200bp fragment reads.

## 1.3 Application of next-generation sequencing

The combination of different types of sample input and library preparations and the production of low cost reads by next-generation sequencing technologies makes them useful in a variety of areas, such as whole genome sequencing, ChIP-Seq, metagenomics, targeted re-sequencing (e.g. exome sequencing), RNA-Seq, Methyl-Seq, and others [Smith et al, 2008; Wold et al, 2008] (Table 1.2).

| **Application** | **references** |
|---|---|
| De novo sequencing of genomes | Velasco et al., 2007; Bentley, 2006 |
| Targeted resequencing (SNP, indels, CNV and structural variations) | Hodges et al., 2007; Porreca et al., 2007 |
| Epigenome | Johnson et al., 2007; Mikkelsen et al., 2007 |
| DNA Methylation | Cokus et al., 2008 |
| Transcriptome (RNA profiling, gene expression, alternative splicing) | Axtell et al., 2006; Berezikov et al., 2006 |
| Metagenome (human microbiome characterization) | Turnbaugh et al., 2007; Hubert et al., 2007 |

**Table 1.2:** Applications of next-generation sequencing technologies.

Important applications include: (1) *de novo* genome sequencing, whole-genome resequencing or more targeted sequencing for discovery of mutations or polymorphisms; (2) transcriptome analysis for gene expression and RNA profiling; (3) large-scale analysis

of DNA methylation; (4) analysis of DNA-protein interactions by chromatin immunoprecipitation (ChIP-Seq); (5) characterization of human microbiome (epigenomics).

A specific application of NGS is the molecular analysis of cancer genome.

### 1.3.1 Analysis of cancer genome by NGS

Cancers are caused by the accumulation of genomic alterations. Therefore, analyses of cancer genome sequences and structures provide insights for understanding cancer biology, diagnosis and therapy [Ross et al, 2011]. NGS technologies have allowed substantial advances in cancer genomics, since they have facilitated an increase in the efficiency and resolution of detection of each of the principal types of somatic cancer genome alterations, including nucleotide substitutions and small insertions and deletions. Furthermore, these new sequencing methods make it feasible to discover novel intrachromosomal rearrangements, including inversions, tandem duplications and deletions, reciprocal and non-reciprocal interchromosomal rearrangements and microbial infections, and to resolve copy number alterations at very high resolution.

Cancer samples have specific features requiring particular consideration in NGS processing and analysis. The main characteristic is that cancer samples differ in their quantity, quality and purity from the peripheral blood samples, commonly used for germline genome analysis. Surgical resection specimens tend to be large and have been the mainstay of cancer genome analysis [Meyerson et al, 2010]. However, diagnostic biopsies from patients with disseminated disease tend to contain few cells and the quantity of nucleic acids available may be limiting. Extracted nucleic acids from cancer are also often of lower quality than those purified from peripheral blood. This because most cancer biopsy and resection specimens are formalin-fixed and paraffin embedded (FFPE) that makes DNA impure. Moreover, cancer specimens often include substantial fractions of necrotic or apoptotic cells that reduce the average nucleic acid quality histology. Moreover, a cancer specimen contains a mixture of malignant and nonmalignant cells and, therefore, a mixture of cancer and normal genomes (and transcriptomes). For all these reasons, FFPE-derived nucleic acids can require special experimental protocols and computational approach for the processing and analysis.

Furthermore, the cancers themselves may be highly heterogeneous and composed of different clones that have different genomes. Cancer genome analytical models must take these two types of heterogeneity (cancer versus normal heterogeneity and within-cancer

heterogeneity) into account in their prediction of genome alterations. Specifically, cancer genomes vary considerably in their mutation frequency (degree of variation compared to the reference sequence), in global copy number or ploidy, and in genome structure.
Various computational methods have been developed to determine the presence of somatic mutations using NGS data. The detection of somatic mutations in cancer requires mutation calling in both the tumour DNA and the matched normal DNA, coupled with comparison to a reference genome and an assessment of the statistical significance of the number of counts of the mutation in the cancer sequence and its absence in the matched normal sequence. False positives and errors in variant calls can be due to inaccurate detection of mutation in tumor for insufficient coverage or for machine-sequencing biases, incorrect local alignment of individual reads and discordant alignment of pairs.

## 1.4 Computational challenges

Next-generation sequencing has revolutionized the study of human genetics and has immense clinical implications. The possibility to sequence different samples and different genomes in parallel has reduced the cost and increased the throughput of genomic sequencing and these technologies are still evolving [Metzker, 2010]. Using deep sequencing, for example, it is now possible to discover novel disease causing mutations [Ley et al, 2008] and detect traces of pathogenic microorganisms [Isakov et al, 2011]. For the first time, research fields such as personalized medicine for patient treatment are becoming tangible at genomic levels given advances in deep sequencing data integration. At the same time, NGS has introduced in the scientific field a computational challenge. In fact, the amount of data produced by ultra high throughput sequencing run is often tremendous and can reach hundreds of millions of reads in various lengths per experiment [Mardis, 2008]. NGS platforms in production are able to produce data of the order of giga or terabytes per machine day. The storage, processing, querying, parsing, analyzing and interpreting of such an incredible amount of data is a significant task that holds many obstacles and challenges. The emergence of NGS platforms imposes increasing demands on statistical methods and bioinformatic tools for the analysis and the management of the huge amounts of data generated by these technologies. Today a large number of softwares already exist for analyzing NGS data. These tools can be fit into many general categories including alignment of sequence reads to a reference, base-calling and/or polymorphism

detection, *de novo* assembly from paired or unpaired reads, structural variant detection and genome browsing.

### 1.4.1 IT infrastructure

The high amount of DNA data generated by next-generation sequencing techniques is demanding computationally intense hard- and software systems to conquer the computationally expensive tasks of NGS data analysis.
To provide an efficient system for analyzing and visualizing next-generation sequencing data, in this thesis a 64 bit computing cluster, consisting of the following components, was introduced: two Sun FireTMX4600 M2 Servers each with four Quad-Core AMD OpteronTM8356, 2.3 GHz, 80 GB RAM; Serial attached SCSI (SAS) storage of 16 TB (extendable up to 256 TB).

## 1.5 Bioinformatic analysis

There are three typical analysis stages in NGS processing (Figure 1.12):

a. Primary analysis (base calling)
b. Secondary analysis (alignment and variant calling)
c. Tertiary analysis.

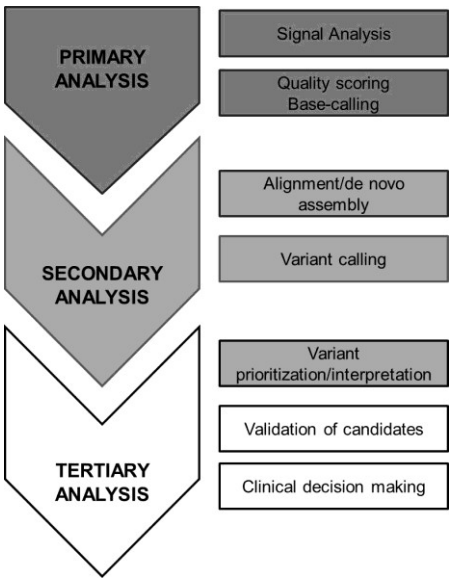


**Figure 1.12:** Workflow of NGS data processing.

### 1.5.1 Primary analysis

Primary analysis can be defined as the machine specific steps needed to call base pairs and compute quality scores for those calls from the image detected during the sequencing. This results in a FASTQ file, the format of raw data, which is just a combination of the sequence data, the so called *read*, as a string of A, C, G and T characters and an associated Phred quality score for each of those bases (Figure 1.13). This output is ready for processing in a secondary analysis pipeline.


**Figure 1.13:** FASTQ format of raw data.

All current commercial next-generation instruments first capture images of many parallel reactions. Analysis software is then applied to locate sequencing reactions and extract information about each reaction. Each platform analysis process is unique and proprietary and may involve several sub-steps. Data filtering may be done subsequent to or during primary analysis. Each platform applies specific rules to eliminate sets of reads and/or images that may have gross artifacts, eliminate individual reads or duplicate sequences. The user may or may not have the ability to view or change these rules.

As quality parameter, the Phred score has become the standard to characterize the quality of DNA sequences and the goodness of sequencing [Ewing et al, 1998; Ewing et al, 1998]. Phred quality score *Q* is defined as logarithmically related to the base-calling error probabilities *P*:

$$Q=-10\log_{10}P \qquad (1)$$

This means that with a Phred of 30 assigned to a base, the chances that this base is called incorrectly are 1 in 1000 (Table 1.3).

| Phred Quality Score | Probability of incorrect base call | Base call accuracy |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1.000 | 99.9% |
| 40 | 1 in 10.000 | 99.99% |
| 50 | 1 in 100.000 | 99.999% |

**Table 1.3:** Phred quality scores are logarithmically linked to error probabilities

## 1.5.2 Secondary analysis

### 1.5.2.1 Alignment

The step conducted after base calling is alignment of the reads to a reference sequence (also called “mapping”) or, if a reference sequence does not exist, for example in the case of microbiology analysis, assembly *de novo* of the sequence reads into contigs and scaffolds.

Alignment of next-generation reads to a reference sequence may be conducted by a variety of different algorithms and may or may not use quality values or intensity values. Mappers such as MAQ [Li et al, 2008] (developed specifically for the Illumina platform) used a hash/seed approach, wherein the reference genome was indexed by k-mers of a given size and the first k-mer of each read was searched for a near-exact match. This approach was well-suited for near-exact matches (i.e. 0-3 single base differences between read and genome) but does not treat small insertions or deletions. It also tends to be both processor and memory intensive. Another algorithms are based on the Burrows-Wheeler transform [Li et al, 2009], for example BWA, which demonstrated significant (>10-fold) improvement in mapping speed. Some mappers are platform-specific features, such as BFAST [Homer et al, 2009] for the Life Technologies/Applied Biosystems which takes advantage of the SOLiD two-base encoding scheme (color base space), and the gsMapper [Shearer et al, 2010] for the Roche/454 platform which uses full flowgram information for alignment. Due to their central role in almost all next-generation applications, mappers remain under active development.

An important feature to be considered in this step is the quality assessment of the data that can improve the mapping performance [Li et al, 2010]. The alignment output contain a Phred based quality score for each of the aligned reads, describing the probability of per-base false alignment. Combination of this quality score together with other alignment

parameters such as mismatches could and should be further assessed using specialized tools [Lassmann et al, 2011] in order to characterize mapped and unmapped reads for potential alignment improvement. These alignment quality scores can be re-assessed and recalibrate taking into account the given base and its quality score, the position within the read and the adjacent nucleotides to account for sequencing chemistry biases. This procedure reduces the effect of sequencing technology derived biases and improve overall variant detection accuracy [DePristo et al, 2011].

Another feature to consider is that most alignment tools allow the user to set the number of allowed mismatches between the read and a reference location and the scoring scale for gap opening and extension. Allowing more mismatches results in a higher portion of mapped reads but at the cost of increased ambiguity and reduced confidence of these alignments. Mismatch allowance should be set while considering the specific experiment at hand. For example, when undergoing microRNA expression profiling, one will want an accurate estimate of the abundance of each microRNA , and should not allow a high mismatch rate if any. On variant calling experiments however, the user should consider the possible expected size range of the variants before setting the allowed mismatch and gap penalty parameters (e.g, if one aims to find a >5nt long deletion, the mismatch limitation should allow it).

To ensure the accuracy of alignment and of the variant detection, during alignment step it is important to considered the multiple mapping. This means that a read could be mapped to multiple loci on the reference genome for sequence homology and repetitiveness. Different alignment tools flag these multiply mapped reads, and provide the user with the option to either randomly assign them to one loci. Discarding multiply mapped reads results in loss of a substantial portion of the data, with potential crucial effects on the following analysis.

An important concept in NGS is depth of coverage. This is a measure of how many reads cover a given locus of the genome (Figure 1.14). After the sequencing process is complete, upper and lower depth thresholds should be applied on the sequencing data before variant calling is performed. Setting a lower coverage limit removes erroneous mismatches caused by sequencing errors and thus supported by very few reads [1000 Genomes Project Consortium et al, 2010; Li, et al, 2009]. Setting a lower limit has been shown to reduce sensitivity without increasing specificity in some tools [Goya et al, 2010] and therefore should be considered in the context of the utilized tool. Setting an upper limit removes

mismatches caused by copy number variations, PCR duplicates introduced by library preparation and reads mapping to paralogous sequences.


**Figure 1.14:** Graphical representation of alignment. Visualization by Integrative Genomics Viewer (IGV) [Robinson et al, 2011; Thorvaldsdóttir et al, 2012].

In secondary analysis, according to the specific application, the choice of appropriate reference sequence in genome re-sequencing is also critical. The University of California Santa Cruz (UCSC) Genome Browser [Kent et al, 2002] provide in FASTA format genome assembly released from Genome Sequence Consortium (the most recent release is GRCh37/hg19 assembly). Projects such as the 1KGP, the Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov/), the 10,000 genomes project (http://www.genome10k.org/), as well as the ongoing NGS of DNA and RNA populations from human tissues may serve to enrich and expand the concept of “reference.”

### 1.5.2.2 Sequence Alignment/Map format

The standard output data file is in SAM (Sequence Alignment/Map) format [Li et al, 2009], which contains several information regarding the the original sequence, or the matching sequence of the reference as well as information about the match (e.g. sample name, SNP calls, or mapping quality value). Most common alignment software generate the alignment output in SAM format, with a multitude of supporting downstream analysis tools.

The SAM format was designed to store nucleotide alignments in a generic way and with the aims to support several sequencing platforms, allowing short and long reads up to 128 Mb, saving SE as well as PE information, and including various types of alignments (Figure 1.15).

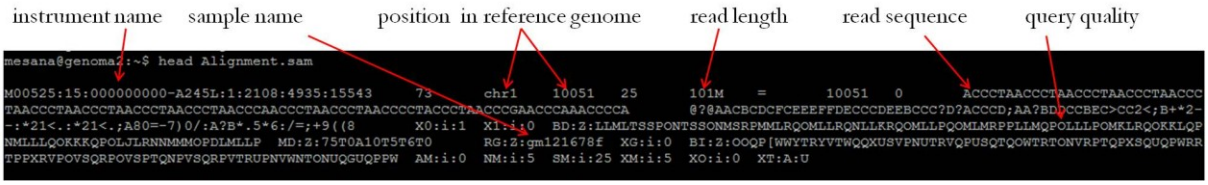


**Figure 1.15:** Alignment file in SAM format.

The tab delimited text format can be divided into a header and an alignment section. Header lines can be identified by the '@' character at the start of each line and contain, among others, information about the reference against the reads have been aligned, and read groups. In the alignment section, a tab delimited line describes a read alignment and contains required information about read name, read flags, reference sequence, alignment position, mapping quality, extended CIGAR, reference sequence of the paired read, position of the paired read, inferred insert size within the read pair, read sequence, and read quality. The extended CIGAR string characterizes the type of alignment operation including clipped, spliced, multi-part, and padded alignment. Allowed operations are match/mismatch (M), insertion (I), deletion (D), skipped base (N), soft clipping (S), hard clipping (H), and padding (P). Additional optional fields allow the documentation of less important or program specific data. Color space read information is also described in the optional fields. To accelerate parsing and ease data processing, the binary file format Binary Alignment/Map (BAM) equivalent to SAM was introduced [Li et al, 2009].

### 1.5.2.3 Variant calling

The secondary analysis includes after alignment the variant calling procedure [DePristo et al, 2011]. Variant calling is the process of accurately determining the variations between a sample and the reference genome. These may be in the form of single nucleotide variants, smaller insertions or deletions (indels), or larger structural variants of categorizations such as transversions, trans-locations, and copy number variants (CNV).

The process is subject to biases for low coverage, sequencing errors, misalignment caused by either low complexity and repeat regions or adjacent variants and library preparation biases. Variant calling depends on an efficient combination between an accurate alignment and sophisticated inference of variance from it.

After aligning deep sequencing reads against a reference genome, SNPs can be inferred from the results by simply denoting each base that is inconsistent between reference and read as a SNP. This straightforward inference of mismatches results in a massive amount of alleged SNPs, many of which suffer from some sort of inaccuracy such as: calling a mismatch in the wrong location, homozygousity and heterozygousity discrepancies and even calling a mismatch in the correct location but with the wrong base. Currently most SNP calling tools [Koboldt et al, 2009; Li et al, 2009; McKenna et al, 2010] apply different probabilistic based considerations and heuristics such as quality assessment and recalibration, SNP filtration, local realignment, coverage assessment, prior probability based on known SNPs, genotype based likelihood and even cancer genomics to elucidate SNPs from alignment results.

Some features, such as local realignment, base quality and transition/transversion ratio have to be considered critically to improve the goodness of variant calling step. Local alignment considers reads that support the presence of an indel in the vicinity of either detected SNPs or known SNP sites retrieved from dbSNP, results in a significant reduction in false positive SNPs [McKenna et al,2010].

The expected ratio between transitions (e.g purine-purine substitutions) and transversions (e.g purine-pyrimidine substitutions) (Ti/Tv ratio) is ~2.3 for whole-genome sequencing and around 3.3 for whole-exome sequencing (coding regions only) [1000 Genomes Project Consortium et al, 2010; DePristo et al, 2011].

After producing a list of detected SNPs, it is highly recommended to compare it against dbSNP, to rescue the number of false positives. The portion of novel SNPs detected in a deep sequencing experiment should range between 1 and 10 percent [DePristo et al., 2011].

### 1.5.3 Tertiary analysis

Variant calling in secondary analysis allows the detection of an extensive catalogues of human genetic variation in the samples. However, pinpointing the few phenotypically

causal variants among the many variants present in human genomes remains a major challenge, particularly for rare and complex traits wherein genetic information alone is often insufficient. Tertiary analysis includes all the approaches to estimate the deleteriousness of single nucleotide variants, which can be used to prioritize disease-causal variants in the context of a specific study [DePristo et al, 2011; paper VI].

Specifically, experimental or computational approaches that provide assessments of variant function can be used to better estimate the prior probability that any given variant is phenotypically important.

#### 1.5.3.1 Annotation

Calling variants using deep sequencing data often results in a multitude of detected variations, even after strict and effective quality filtration as denoted earlier, deep sequencing data reveals thousands to millions of different variations. These ones can result in biological effects through introduction of different amino acids into protein sequences, early termination of coding sequences and alteration of regulatory elements and splice sites. The following process after variant calling is annotating the detected variants and elucidating their effect and biological significance, separating clinically, scientifically and medically relevant variations from neutral, non functional ones. In a large list spanning this many variants, manual annotation of each variant effect is neither feasible or accurate. For annotation, public databases can flexibly use, such as RefSeq, UCSC, ENSEMBL, GENCODE or many other gene definition systems. All these information can be used to determine the position of the mismatch in the gene and whether it is in a coding/non-coding, exon, intron, UTR or intron/exon junction region (intronic regions contiguous to exon starts and exon ends that are important to evaluate splice-site mutations).

To give biological context to NGS data could be useful to query public or commercial databases containing biological/functional annotation. For example OMIM (Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), {February, 2012}. URL: http://omim.org/), which catalogues genetic disorders and traits, Human Gene Mutation Database (HGMD) [Stenson et al, 2009], which reports validate gene mutations, and COSMIC, a large repository of somatic mutations in cancer [Forbes et al, 2009].

Moreover, large projects such as 1KGP, Exome Sequencing Project (ESP) (Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL:

http://evs.gs.washington.edu/EVS/) [February 2012]) allow for studies to be able to determine the relative frequencies of variants in their samples compared to common populations (MAF, minor allele frequency).

A MAF of 1% is used as cutoff to define a variants as rare [Frazer et al, 2009; paper I]. The rarity of a variant or the absence in a healthy control population is classically one of the criteria for evaluating the pathogenicity both in research and in clinical settings. Although there is need of caution in considering this parameter as prediction of deleteriousness. In fact, MAF ≥1% implies that the variant is sufficiently common in the population to be not tremendously deleterious, the opposite is not generally true.

### 1.5.3.2 Prioritization and interpretation

The development of methods to assess the effect of variants has been a major subject of research in the field of bioinformatics during the past decade. Currently, there are several tools available which can predict the disease-causing potential of mismatches for protein structure and function. These programs can predict the effects on the basis of different features, including evolutionary conservation, changes in the physico-chemical characteristics of the aminoacids, the sequence environment of the affected amino acid alteration in structural properties of proteins.

This “tolerance” predictors, which evaluate *in silico* the effect of mutations on the phenotype at protein level include Polymorphism Phenotyping 2 (PolyPhen-2) [Adzhubei et al, 2010], Align-GVGD [Tavtigian et al, 2006], SIFT [Kumar et al, 2009] and MutationTaster [Schwarz et al, 2010].

The Grantham Score, which categorizes the differences of physicochemical properties for codon replacements, , including polarity and molecular volume, can be also retrieved from the original Grantham Score Matrix [Grantham, 1974].

Moreover, the evolutionary nucleotide conservation score in vertebrates (phyloP) or in different species is a reliable method for predicting possible pathogenicity of a missense variant [Flanagan et al., 2010; Siepel et al, 2005].

# Chapter 2

# Genomic analysis in complex disease

The advent of NGS technologies has revolutionized the field of genomics, enabling fast and cost-effective generation of genome-scale sequence data with exquisite resolution and accuracy. To date, NGS technologies have been widely used for many applications, such as rare variant discovery by whole genome resequencing or targeted requencing, transcriptome profiling of cells, tissues and organisms, and identification of epigenetic markers for disease diagnosis. Studies using massive parallel sequencing have yielded surprising and important discoveries regarding the genetic bases of Mendelian and complex diseases, particularly cancer. The identification of molecular mechanism underlying tumorigenesis enables molecular diagnosis and carrier testing in the patient and his or her family. This is of great importance for patient management and family counseling, and serves as a starting point for therapeutic interventions. Furthermore, the identification of Mendelian disease genes contributes to our understanding of gene functions and biological pathways underlying health and disease in general [paper II].

In the following paragraphs, some examples of application of NGS technologies in clinical and diagnostic context will be presented.

## 2.1 Ras/MAPK pathway

The Ras/MAPK (mitogen-activated protein kinases) pathway is one of the most important signaling network inmplicated in growth-factor mediated cell proliferation, differentiation and death [Molina et al, 2006]. It is a chain of proteins in the cell that communicates a signal from a receptor on the surface of the cell to the DNA in the nucleus of the cell. The signal starts when a signaling molecule binds to the receptor on the cell surface and ends when the DNA in the nucleus expresses a protein and produces some change in the cell, such as cell division. The pathway includes many proteins, including MAPK, originally called ERK

(extracellular signal-regulated kinases), which communicate by adding phosphate groups to a neighboring protein, which acts as an "on" or "off" switch (Figure 2.1).

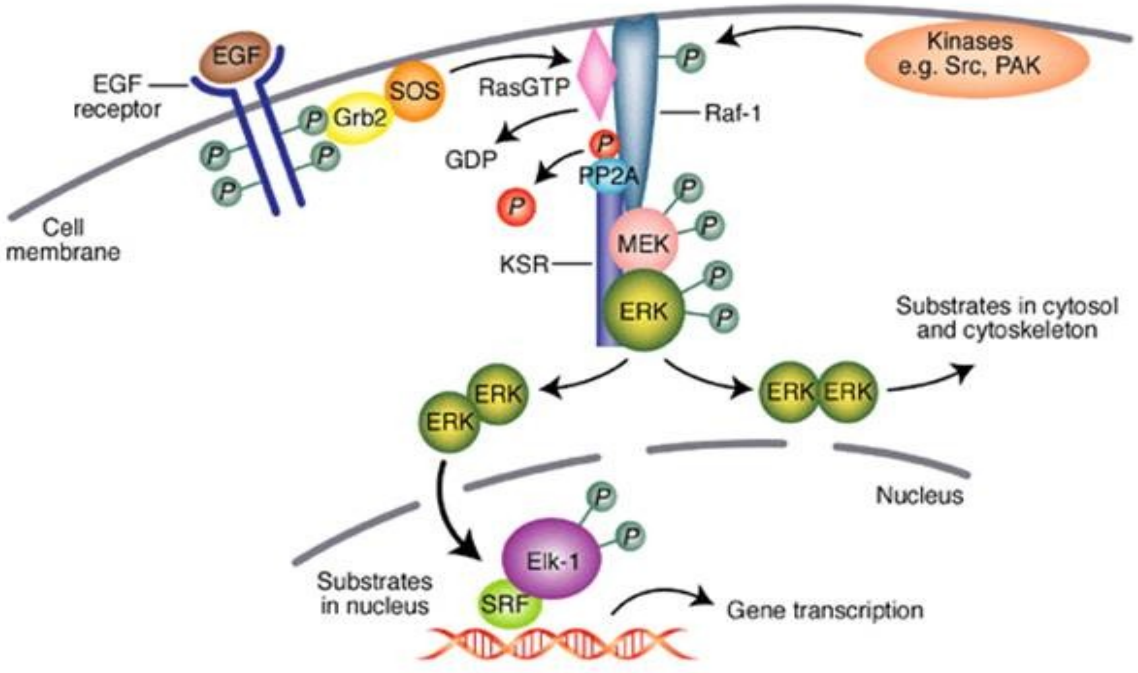


**Figure 2.1:** Ras//MAPK pathway. Source: *http://journals.cambridge.org*

The cascade of activation of this pathway starts when receptor-linked tyrosine kinases such as the epidermal growth factor receptor (EGFR) are activated by extracellular ligands. Binding of epidermal growth factor (EGF) to the EGFR activates the tyrosine kinase activity of the cytoplasmic domain of the receptor. The EGFR becomes phosphorylated on tyrosine residues. Docking proteins such as GRB2 contain an SH2 domain that binds to the phosphotyrosine residues of the activated receptor. GRB2 binds to the guanine nucleotide exchange factor SOS (son of sevenless) to the plasma membrane by way of the two SH3 domains of GRB2. When the GRB2-SOS complex docks to phosphorylated EGFR, SOS becomes activated. Activated SOS displaces GDP from a member of the Ras subfamily (i.e. HRas, KRas), subsequently allowing the binding between Ras and GTP [Molina et al, 2006]. Ras are membrane-associated guanine nucleotide-binding proteins which affect many cellular functions, including cell proliferation, apoptosis, migration, fate specification, and differentiation. In the GTP-bound form, Ras interacts specifically with so-called effector proteins, thereby initiating cascades of protein-protein interactions that may finally lead to

cell proliferation and activates a number of signaling pathways, including the Raf/MEK/ERK pathway. The activity of Ras is limited by the hydrolysis of GTP back to GDP by GTPase activating proteins (GAP) in an enzymatic process. Reactivation of Ras requires the removal of GDP by SOS. Activated Ras activates the protein kinase activity of RAF kinase. RAF kinase phosphorylates and activates MEK (MEK1 and MEK2). MEK phosphorylates and activates a mitogen-activated protein kinase (MAPK).Raf, and MAPK are both serine/threonine-selective protein kinases. MEK (also known as MAPKK) is a tyrosine/threonine kinase. In the technical sense, Raf, MEK, and MAPK are all mitogen-activated kinases.

The Ras/MAPK pathway is probably the best characterized signal transduction pathway in cell biology. The function of this pathway is to transduce signals from the extracellular milieu to the cell nucleus where specific genes are activated for cell growth, division and differentiation. MAPK regulates the activities of several transcription factors. MAPK can phosphorylate C-myc. MAPK phosphorylates and activates MNK, which, in turn, phosphorylates CREB. MAPK also regulates the transcription of the C-Fos gene. By altering the levels and activities of transcription factors, MAPK leads to altered transcription of genes that are important for the cell cycle. The Ras/MAPK pathway is also involved in cell cycle regulation, wound healing and tissue repair, integrin signaling and cell migration. Finally, the Ras/MAPK pathway is able to stimulate angiogenesis through changes in expression of genes directly involved in the formation of new blood vessels. Thus, signaling through the Ras/Raf/MAPK regulates a variety of cellular functions that are important for tumorigenesis. Dysregulation of this pathway is a common event in cancer as Ras proteins are the most frequently mutated oncogenes in human cancer [Malumbres et al, 2003]. Mutations in the Kras oncogene have been localized in codons 12, 13, 59 and 61 with those at codons 12 and 61 occurring most frequently. Kras mutations are present in 15-50% of lung cancers and in 72-90% of pancreatic cancers

Historically, the rat sarcoma (RAS) virus homologue was the first oncogene to be described in human cancer (Der et al 1982). In cancer, the most commonly mutated members of the RAS superfamily include HRAS, KRAS and NRAS, highly conserved proteins . The RAS proteins are small (21 kilodalton) G-proteins that are active with bound GTP and inactive with bound GDP. Although the GTP can self-hydrolyze, there is a class of enzymes termed GAPs (GTPase activating proteins) that facilitate this hydrolysis and terminate RAS activity.

The name 'Ras' is an abbreviation of 'Rat sarcoma', reflecting the way the first members of the protein family were discovered. When Ras is 'switched on' by incoming signals, it subsequently switches on other proteins, which ultimately turn on genes involved incell growth, differentiation and survival. As a result, mutations in ras genes can lead to the production of permanently activated Ras proteins. This can cause unintended and overactive signalling inside the cell, even in the absence of incoming signals. Because these signals result in cell growth and division, overactive Ras signaling can ultimately lead to cancer.Ras is the most common oncogene in human cancer - mutations that permanently activate Ras are found in 20-25% of all human tumors and up to 90% in certain types of cancer (e.g. pancreatic cancer).[2] For this reason, Ras inhibitors are being studied as a treatment for cancer, and other diseases with Ras overexpression [Fernández-Medarde et al, 2011].

All Ras protein family members belong to a class of protein called small GTPase, and are involved in transmitting signals within cells (cellular signal transduction). Ras is the prototypical member of the Ras superfamily of proteins, which are all related in 3D structure and regulate diverse cell behaviours (Figure 2.2) [Malumbres et al, 2003; Karnoub et al, 2008].

```
KRAS_HUMAN      MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG 60
NRAS_HUMAN      MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG 60
HRAS_HUMAN      MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG 60
                ************************************************************

KRAS_HUMAN      QEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDL 120
NRAS_HUMAN      QEEYSAMRDQYMRTGEGFLCVFAINNSKSFADINLYREQIKRVKDSDDVPMVLVGNKCDL 120
HRAS_HUMAN      QEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHQYREQIKRVKDSDDVPMVLVGNKCDL 120
                ****************************:*** **: ***********:*************

KRAS_HUMAN      PSRTVDTKQAQDLARSYGIPFIETSAKTRQRVEDAFYTLVREIRQYRLKKISK-EEKTPG 179
NRAS_HUMAN      PTRTVDTKQAHELAKSYGIPFIETSAKTRQGVEDAFYTLVREIRQYRMKKLNSSDDGTQG 180
HRAS_HUMAN      AARTVESRQAQDLARSYGIPYIETSAKTRQGVEDAFYTLVREIRQHKLRKLNPPDESGPG 180
                .:***:::**::**:*****:********* **************::::*:.  ::   *

KRAS_HUMAN      CVKIKKCIIM 189
NRAS_HUMAN      CMGLP-CVVM 189
HRAS_HUMAN      CMSCK-CVLS 189
                *:    *::
```

**Figure 2.2:** Multiple alignment between human RAS proteins. The amino acid sequences are highly conserved. The alignment was performed using Clustal algorithm [Larkin et al, 2007]

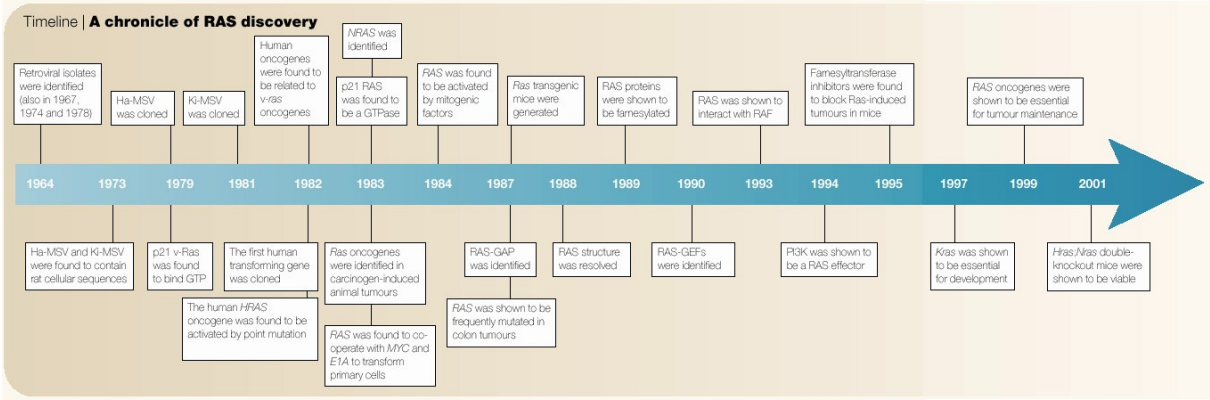


**Figure 2.3:** Key events in the field of Ras research [Malumbres et al, 2003].

.

Oncogenic lesions introduce changes in the primary sequence of RAS so that the protein is constitutively active. RAS pathway still remains one of the most investigated pathways in human cancer (Solit et al 2006), including melanoma, and our current understanding suggests that several possible mutations along this cascade lead to tumor-promoting physiology.

Although RAS proteins are frequently mutated in cancer, there is preferential targeting of specific family members in different tumor types.

## 2.1.1 Colorectal cancer

Colorectal cancer (CRC) is the second most common cancer in the world, accounting for approximately one million new cases each year [Markowitz et al, 2009] (Figure 2.4). The overall mortality is approximately 50%. The surgical stage at diagnosis is the most important factor for predicting patient outcome, with five year survival rates of more than 90% for stage I disease and less than 10% for stage IV disease.

**Figura 2.4:** Colorectal cancer. A) biopsy B) histological section

The surgical stage represents a classification system based on the extent and depth of tumor growth. The system most commonly used in describing CRC is the Tumor, Nodes, Metastasis (TNM) system of the American Joint Committee on Cancer (AJCC) provided in 2002.

In Stage I CRC shows invasive grown into the anatomical layers of the colon, but the tumor is not spread outside the colon wall or into regional lymph nodes. During Stage II there is tumor penetration through the bowel wall involving the serosa; however, there is no involvement of regional lymph nodes or distant metastases. In Stage III CRC have spread to nearby lymph nodes but not yet metastasized to distant sites in the body. Finally, in Stage IV the tumor has spread to distant organs such as the liver, lungs, or other sites. CRC is considered to develop through a multi-step process, originating from normal colon epithelium that develops into precursor lesions termed adenomas. Adenomas can subsequently further progress into invasive CRC with metastatic potential. Spread of CRC occurs by direct growth through the bowel wall and through invasion of lymphatic and venous channels. The most common sites for metastases are regional lymph nodes and the number of lymph node metastases influences prognosis. The liver is the most common distant site for CRC metastases. The vast majority of CRC are adenocarcinomas, with less than 10% of the cancers being distinguished by an abundant secretion of mucin. The tumors are classified according to the degree of morphological differentiation into well, moderately and poorly differentiated. About 80% are well or moderately differentiated with a growth pattern consisting of tumor cells that form irregular glandular structures present at different layers of the bowel wall. Poorly differentiated CRC show no, or only hinted, glandular formation. Overall poor differentiation with a diffuse infiltrative growth pattern is

associated with poor prognosis, although stringent classification systems based on morphological features are lacking

#### 2.1.1.1 CRC treatment

Treatment decisions are thus based on the surgical stage and morphological characteristics.

At stage 0, in which the cancer has not grown beyond the inner lining of the rectum, the strategy consist in removing or destroying the cancer by polypectomy, local excision, or transanal resection. For stage I, during which cancer has grown through the first layer of the rectum into deeper layers but has not spread outside the wall of the rectum itself, surgery is usually the main treatment: a low anterior resection, colo-anal anastomosis, or an abdominoperineal resection may be done, depending on exactly where the cancer is found within the rectum

In some cases, adjuvant chemoradiation (treatment with radiation and chemo together) is advised for patients having such surgery. 5-Fluorouracil (5-FU) is the chemo drug most often used.

Stage II in which the cancer have grown through the wall of the rectum and may extend into nearby tissues is treated by low anterior resection, colo-anal anastomosis, or abdominoperineal resection (depending on where the cancer is in the rectum), along with both chemotherapy and radiation therapy. Most doctors now favor giving the radiation therapy along with the chemo drug 5-FU before surgery (*neoadjuvant treatment*), and then giving adjuvant chemo after surgery, usually for a total of 6 months of treatment (including the time getting chemo and radiation together). The chemo after surgery may be the FOLFOX regimen (oxaliplatin, 5-FU, and leucovorin), 5-FU and leucovorin, CapeOx (capecitabine plus oxaliplatin) or capecitabine alone, based on what's best suited to your health needs. Treatment for stage III cancers that have spread to nearby lymph nodes includes radiation therapy, given along with 5-FU chemo before surgery (called *chemoradiation*). This may shrink the cancer, often making surgery more effective for larger tumors. It also lowers the chance that the cancer will come back in the pelvis. Giving radiation before surgery also tends to lead to fewer problems than giving it after surgery. The rectal tumor and nearby lymph nodes are then removed, usually by low anterior resection, colo-anal anastomosis, or abdominoperineal resection, depending on where the

cancer is in the rectum. After surgery, chemo is given, usually for about 6 months. The most common regimens include FOLFOX, 5-FU and leucovorin, or capecitabine alone. Finally, in stage IV cancer has spread to distant organs and tissues such as the liver or lungs. and the Treatment options disease depend to the extension of metastasis. Treatment options include surgery to remove the rectal lesion and distant tumors, followed by chemo (and radiation therapy in some cases) with combination of chemo and radiation therapy before and after the surgery.

### 2.1.1.2 anti-EGFR monoclonal antibodies

Recently new targeted drugs have been implemented into the treatment of patients with advanced colorectal cancer. Epidermal growth factor receptor is commonly expressed in colorectal tumors and monoclonal antibodies (mAbs) inhibiting EGFR demonstrate clinical efficacy in patients with mCRC. This anti-EGFR monoclonal antibodies, cetuximab and panitumumab, were designed as effective inhibitors of the EGFR (Figure 2.5).

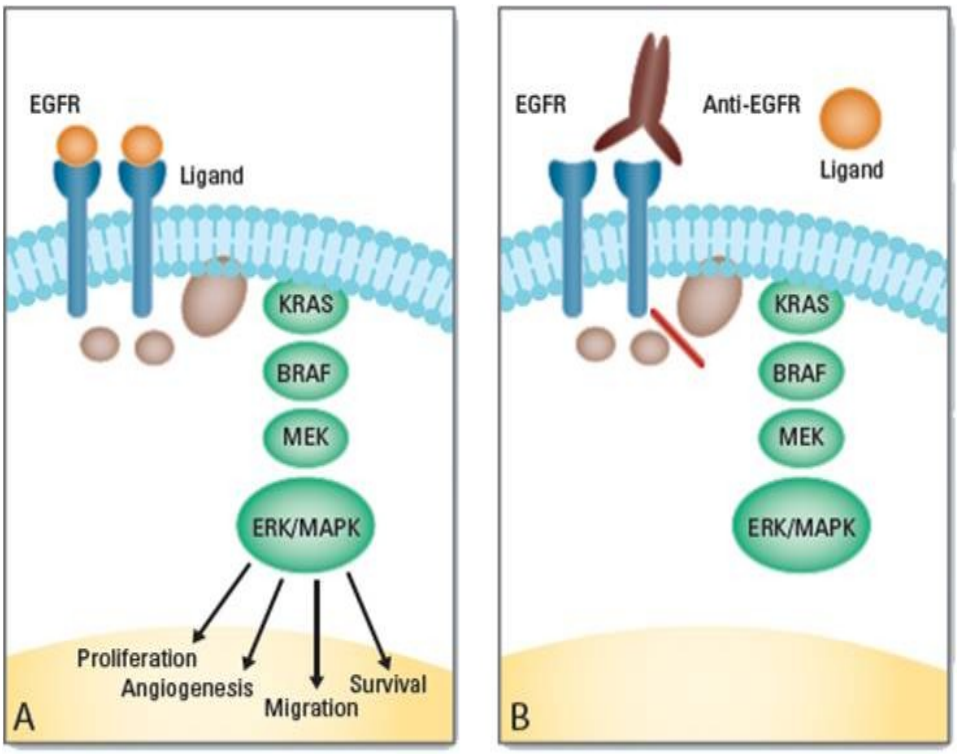


**Figure 2.5:** Mechanism of anti-EGFR monoclonal antibodies. A) EGFR binding to ligand activates the Ras/MAPK cascade. B) monoclonal antibodies anti-EGFR inhibits the EGFR activity.

Cetuximab (Erbitux®, Merck KgaA, Darmstadt, Germany) is a chimeric mouse/human antibody targeted against the extracellular domain of the EGFR. Binding of cetuximab to the receptor prevents ligand binding, induces receptor internalization and causes a direct inhibition of the receptor tyrosine kinase activity.This in turn blocks downstream signal transduction via the PI3K/Akt and Ras/MAPK pathways inducing pro-apoptotic mechanisms and inhibiting cellular proliferation, angiogenesis and metastasis. Cetuximab is indicated for the treatment of patients with epidermal growth factor receptor (EGFR)-expressing, Kras wild-type metastatic colorectal cancer, in combination with chemotherapy, and as a single agent in patients who have failed oxaliplatin- and irinotecan-based therapy and who are intolerant to irinotecan. As an IgG1 antibody cetuximab may also induce antibody-dependent cell-mediated cytotoxicity (ADCC). The side effects include acne-like rash, and dermathological reactions (urticaria, pruritis), hypotension, bronchospasm, dyspnea, wheezing, angioedema, dizziness, anaphylaxis, and cardiac arrest. Photosensitivity and hypomagnesemia have been also observed in some patients. This drug is given by intravenous therapy and costs up to $30,000 for eight weeks of treatment per patient.

Panitumumab (Vectibix®, Amgen Thousand Oaks, CA, USA), by contrast, is a fully human antibody which is also directed against the EGFR but being an IgG2 MoAb lacks ADCC activity.

#### 2.1.1.3 Kras mutation in colorectal cancer

Oncogenic mutations of the Kras genes are observed in about 40% (20–50%) of sporadic colorectal cancers [Lievre et al, 2006; Heinemann et al, 2009]. These are point mutations and are generally observed as somatic mutations. Up to 90% of activating mutations of the Ras gene are detected in codons 12 and 13, but less frequently also in codons 61 and 63. The most frequent types of KRAS mutations in colorectal cancers are G>A transitions and G>T transversions. The codons 12 and 13 code for two adjacent glycine residues located in the proximity of the catalytic site of RAS. In particular codon 12 mutations of the Kras gene were associated with a mucinous phenotype of colorectal cancer, while codon 13 mutations were rather non-mucinous, but were characterized as more aggressive tumors with a greater metastatic potential.

Different KRAS mutations result in an exchange of different amino acids at these catalytic sites, and therefore, may be responsible for the different levels of intrinsic GTPase activity reduction. As a consequence, variable RAS mutations may imply variable effects on the biology of disease.

For this reason, several studies support the importance of mutational activation of KRAS in the progression of CRC [Amado et al, 2008; Souliers et al, 2010].

For the detection of KRAS mutations several methods are known. The hotspots and thus the most frequent mutations g.34G>C (p.Gly12Arg), g.35G>C (p.Gly12Cys), g.34G>A (p.Gly12Ser), g.35G>A (p.Gly12Asp), g.35G>C (p.Gly12Ala), g.35G>T (p.Gly12Val), g.38G>A (p.Gly13 Asp) and rarely g.183G>T (p.Gln13His). Moreover, as the detection of KRAS mutations has become a routine diagnostic method, the protocols in use are selected on the basis of velocity, robustness and easiness. The main problems in detection of KRAS mutations involve the low quality of the DNA that often causes error in specificity and sensitivity of the analysis.

#### 2.1.1.4 Effect of Kras mutations on response to anti-EGFR therapy

The reports on the prognostic relevance of KRAS mutations are inconsistent. While some studies have indicated a negative impact of KRAS mutations on survival (ref) others did not [Lievre et al, 2006; Souliers et al, 2010].

The KRAS mutation status has been identified as a strong predictor of response to anti-EGFR antibodies. In fact, mutations of the KRAS gene may activate downstream signal transduction and confer resistance to upstream inhibition of the EGFR by monoclonal antibodies. For this reason, EGFR directed therapy is not only not effective in KRAS mutant mCRC patients. Moreover, it may also induce unnecessary toxicity and has been associated with an inferior outcome in some clinical trials. In fact, determination of the KRAS mutation status is required in all mCRC patients who may receive anti-EGFR directed antibody therapy. Studies using cetuximab plus chemotherapy for first-line treatment of mCRC could detect an improvement of treatment efficacy only in KRAS wild-type patients. By contrast, KRAS mutant patients either had no benefit from the addition of cetuximab or even showed a worse outcome than their comparators. At present, it is not clear if additional toxicity impairs treatment intensity in KRAS mutant patients or if negative interactions between cetuximab and chemotherapy take place.

## 2.1.2 RASopathies

The Ras/MAPK pathway is critical to normal development and essential in the regulation of the cell cycle, differentiation, growth and cell senescence (Figure 2.6). A class of developmental disorders, the so called RASopathies, is caused by germline mutations in genes that encode protein components of the Ras/MAPK pathway resulting in increased signal transduction down the Ras/MAPK pathway [Tidyman et al, 2008]. The term RASopathies includes different diseases, such as capillary malformation-AV malformation syndrome, autoimmune lymphoproliferative syndrome, Costello syndrome (CS), Leopard syndrome (LS), Noonan syndrome, (NS), neurofibromatosis type 1 etc.. Each syndrome exhibits unique phenotypic features, however, since they all cause dysregulation of the Ras/MAPK pathway, there are numerous overlapping phenotypic characteristics between the syndromes, including craniofacial dysmorphisms, cardiac malformations and cutaneous, musculoskeletal and ocular abnormalities, varying degrees of neurocognitive impairment and an increased risk of developing cancer.

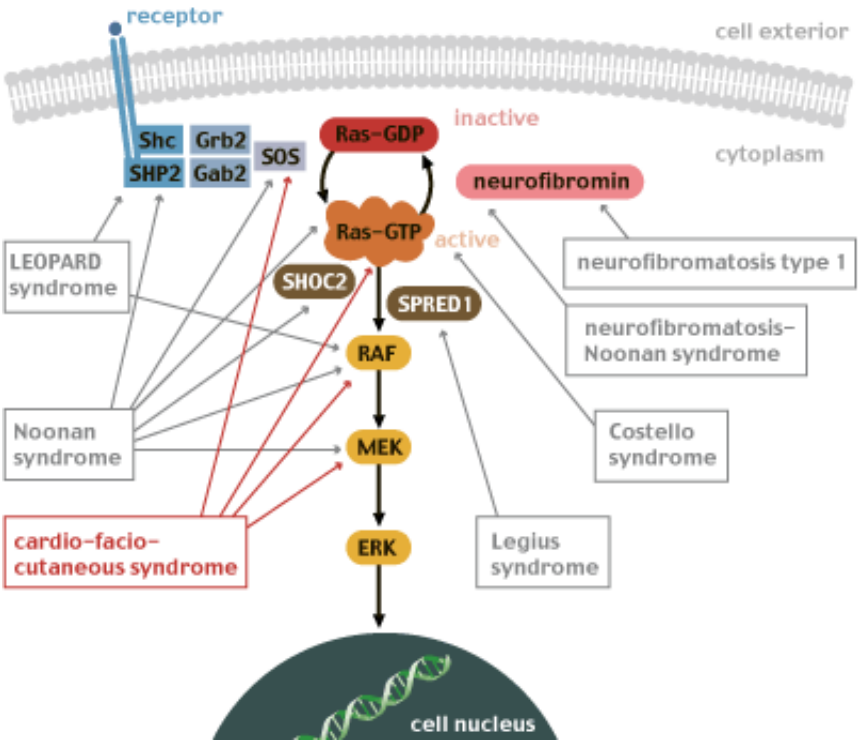


**Figure 2.6:** Ras/MAKP signalling pathway and related genetic syndromes. RASopathies are caused by mutations in the genes which control the production of certain signal proteins.Source: *http://www.socialstyrelsen.se/*

The severity of symptoms will depend on which part of the signalling pathway is impaired. Several types of cardiac anomalies are associated with RASopathies, for example a defective pulmonary valve that obstructs the blood flow from the lungs (pulmonary valve stenosis), or atrial septal defect resulting in abnormal opening in the wall separating the left and right upper heart chambers of the heart, or a thickening of the muscular wall of the right ventricle and of the diving muscle between the right and left ventricles (hypertrophic cardiomyopathy, HCM) which may be progressive. The syndromes are also associated with characteristic facial features. The head is relatively large in relation to the body, and the forehead is wide, narrowing at the temples. The jaws are often small and the palate may be high and narrow. Down-slanting eyes with increased space between them (hypertelorism) and drooping eyelids (ptosis) are common. Impaired vision, i.e myopia, or eyes problem (strabismus, nystagmus) also occur. The ears are low set and some of these children have impaired hearing.

Some patients may have skin disorders, eczema or ichthyosis or hyperkeratosis (abnormal dry skin). Some children have lymphoedema, a tissue swelling caused by a compromised lymphatic system and accumulation of lymphatic fluid in the pleural cavity. Hypotonus can cause delayed motor development. Many have intellectual disabilities with varying degrees of severityThey may also have difficulties organising new information, adapting to new situations, and learning new skills. Language development is delayed in approximately half of these children. Some children have epilepsy and hydrocephalus may also occur.

#### 2.1.2.1 Noonan syndrome

Noonan syndrome is an autosomal dominant disorder that affects approximately 1/1000 to 1/2500 newborns [van der Burgt, 2007, Tidyman, 2008]. Diagnosis of Noonan syndrome is primarily dependent on clinical features, including short stature, typical face dysmorphology, and congenital heart defects. Diagnosis can be difficult due to the wide spectrum of clinical features, not all of which need to be present for diagnosis [Razzaque et al, 2007].

Phenotypic features of Noonan syndrome include hypertelorism, ptosis, and low-set, posteriorly rotated ears with a thickened helix. Cardiac abnormalities most commonly associated with Noonan syndrome include pulmonary stenosis and hypertrophic cardiomyopathy [Schubbert et al, 2006]. Other associated features include the presence of

a webbed neck, chest deformity, mild intellectual deficit, cryptorchidism, poor feeding in infancy, bleeding tendencies, and lymphatic issues (Figure 2.6).

Short stature is one of the main characteristics of the condition, affecting more than 80% of children diagnosed.[6] There could be also a variable degree neurocognitive delay [van der Burgt, 2007; Roberts et al, 2007].

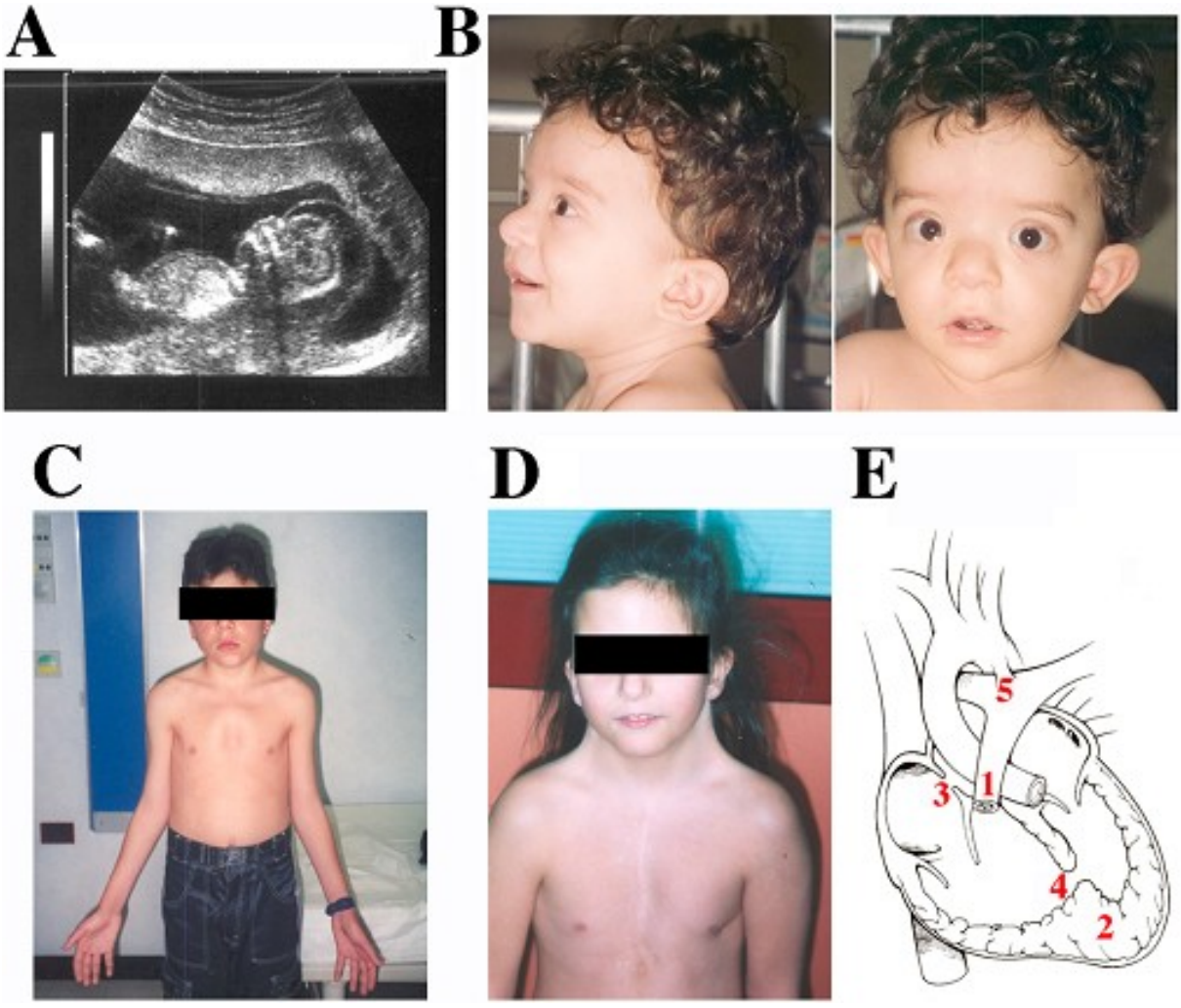


**Figure 2.7:** Clinical features in Noonan syndrome. (A) nuchal cystic hygroma in echography; (B) facial dysmorphisms; (C) pectus deformities; (D) webbed neck; (E) schematic representation of major cardiac defects: 1, pulmonic stenosis; 2, hypertrophic cardiomyopathy; 3, atrial septal defects; 4, ventricular septal defects; 5, patent ductus arteriosus. Source: *http://atlasgeneticsoncology.org.*

Individuals with NS have an increased risk of cancer. Children with Noonan syndrome are predisposed to malignancies, juvenile myelomonocytic leukemia (JMML) most commonly.

JMML, formerly termed juvenile chronic myeloid leukemia or chronic myelomonocytic leukemia, is a myeloproliferative/myelodysplastic disorder of childhood characterized by excessive proliferation of immature and mature myelomonocytic cells that originate from a pluripotent stem cell. In childhood, JMML accounts for approximately 30% of cases of myelodysplastic and myeloproliferative syndromes and 2% of leukemias. It typically presents in infancy and early childhood, and is often lethal. Recent studies have provided strong evidence that hypersensitivity to granulocyte-macrophage colony-stimulating factor (GM-CSF), due to a selective inability to down-regulate the Ras/MAPK cascade, plays a central role in the clonal cell growth characteristic of JMML. In approximately 15-30% of JMML cases, the pathological activation of the Ras/MAPK cascade results from oncogenic NRAS or KRAS2 mutations that specifically affect GTP hydrolysis, leading to the accumulation of RAS in the GTP-bound active conformation .

It is known that four genes, PTPN11, KRAS, SOS1 and RAF1 cause NS with all genes encoding various components of the Ras/MAPK pathway. The most common gene associated with NS is PTPN11 which accounts for approximately half of all cases. SHP2, the protein product of PTPN11, is a non-receptor protein tyrosine phosphatase composed of N-terminal and C-terminal SH2 domains and a catalytic protein tyrosine phosphatase (PTP) domain. Shp2 plays an important role in mediating multiple downstream biological responses, such as proliferation and/or survival, adhesion, and migration.

The catalytic function of the protein is auto-inhibited through a blocking interaction between the N-SH2 and PTP domains. The majority of NS-causing missense mutations in PTPN11 cluster in residues involved with the interaction between the N-SH2 and PTP domains. Mutations in this region interfere with the stability of the catalytically inactive form of SHP2 resulting in impairment of the protein's ability to switch from the active to the inactive protein conformation causing increased signaling down the Ras/MAPK pathway. SOS1 missense mutations are the second most common cause of NS and account for approximately 13% of cases. SOS1 encodes the Ras-GEF protein, SOS1, which is responsible for stimulating the conversion of Ras from the inactive GDP-bound form to the active GTP-bound form. The majority of SOS1 missense mutations are located in codons encoding residues that are responsible for stabilizing the protein in an inhibited conformation [Razzaque et al, 2007].

These mutations disrupt the auto-inhibition of SOS1 Ras-GEF activity resulting in a gain-of-function of SOS1 and a subsequent increase in the active form of Ras and increased Ras/MAPK pathway signaling.

KRAS mutations are associated with a small percentage (<2%) of individuals with NS and result in increasing signaling down the Ras/MAPK pathway through two distinct mechanisms: either by mutations that reduce the intrinsic and GAP stimulated GTPase activity or by mutations that interfere with the binding of KRAS and guanine nucleotides. Both result in a net increase in the active GTP-bound form of KRAS. Mutations in Raf1 also causes NS. Raf1 encodes the protein RAF1, a serine/ threonine kinase that is one of the direct downstream effectors of Ras.

### 2.1.2.2 Leopard syndrome

Leopard syndrome is a rare autosomal dominant disorder characterized by Lentigines, Electrocardiogram abnormalities, Ocular hypertelorism, Pulmonic valvular stenosis, Abnormalities of genitalia, Retardation of growth, and Deafness (Figure 2.8).

The disease is a complex of features, mostly involving the skin, skeletal and cardiovascular systems, which may or may not be present in all patients.

Cafè-au-lait spots occurr in a high number (10,000+) over a large portion of the skin, at times higher than 80% coverage. These can even appear inside the mouth, or on the surface of the eye. These have irregular borders and range in size from 1 mm in diameter to several centimetres in diameter. Also, some areas of vitiligo-like hypopigmentation may be observed. Electrocardiographic conduction abnormalities: Ocular defects are frequent, like hypertelorism. Facial abnormalities are the second highest occurring symptom after the lentigines and include broad nasal root, prognathism or low-set, possibly rotated, ears.

Cardiac abnormalities may be present, including aortic stenosis, or mitral valve prolapsed or hypertrophic cardiomyopathy.

Abnormal genitalia: (usually cryptorchidism (retention of testicles in body) or monorchism (single testicle). In female patients, this presents as missing or single ovaries. Most newborns with this syndrome are of normal birth weight and length, but will often slow within the first year.

The presence of all of these hallmarks is not needed for a diagnosis. A clinical diagnosis is considered made when, with lentigines present there are 2 other symptoms observed, such as ECG abnormalities and ocular hypertelorism, or without lentigines, 3 of the above conditions are present, with a first-degree relative (i.e. parent, child, sibling) with a clinical diagnosis.

Mild mental retardation is observed in about 30% of those affected with the syndrome
Nystagmus (involuntary eye movements), seizures, or hyposmia (reduced ability to smell) has been documented in a few patients
In 2006, a patient affected by Leopard syndrome was reported with acute myelogenous leukemia. The nature of how the mutation causes each of the condition's symptoms is not well known; however, research is ongoing. As Noonan syndrome, LS is caused by germ line missense mutations in PTPN11, encoding the protein-tyrosine phosphatase Shp2. The most common LS associated PTPN11 mutations affect amino acids in the catalytic PTP domain which result in reduced SHP2 catalytic activity. In contrast, PTPN11 mutations associated with NS all produce SHP2 gain-of-function. It has recently been proposed that the residual catalytic activity in the LS mutant SHP2 protein is sufficient to produce a gain-of function-like phenotype due to dysregulation of the protein causing continuous MAPK pathway activity during development . LS is associated with increased risk of malignancy, in particular to acute myelogenous leukemia and neuroblastoma.

**Figure 2.8:** Phenotypic features in Leopard syndrome.

In itself, LEOPARD syndrome is not a life threatening diagnosis, most people diagnosed with the condition live normal lives. Obstructive cardiomyopathy and other pathologic findings involving the cardiovascular system may be a cause of death in those whose cardiac deformities are profound. Mild mental retardation is observed in about 30% of those affected with the syndrome

### 2.1.2.3 Costello syndrome

Costello syndrome is a rare developmental disorder with multiple anomalies, including characteristic dysmorphic craniofacial features (Figure 2.9), failure to thrive, cardiac, musculoskeletal and ectodermal abnormalities and neurocognitive delay (for review see [29]). Individuals with CS are at increased risk of developing neoplasms, both benign and malignant. Heterozygous germline mutations in *H-ras* cause CS. The distribution frequency of mutations reveals that more than 80% of individuals have a Gly12Ser substitution, followed by the second most common, Gly12Ala. These substitutions disrupt guanine nucleotide binding and cause a reduction in intrinsic and GAP induced GTPase activity resulting in Ras remaining in the active state. In addition, less frequently observed mutations, such as K117R and A146T may result in an atypical phenotype [33,34]. Biochemical investigation of the novel HRAS mutant protein, K117R, has demonstrated normal intrinsic GTP hydrolysis and responsiveness to GTPase-activating proteins; however, the nucleotide dissociation rate is increased. The increase in the guanine nucleotide exchange rate results in a net increase in the active GTP bound form of Ras due to the higher concentration of GTP in the cell. It is interesting that amino acid positions 12 and 13, the two most common positions mutated in CS, are also the most frequently mutated positions in oncogenic Ras in approximately 20% of all tumors.


**Figure 2.9:** Dysmorphic craniofacial features in Costello syndrome.

There is no cure for cardiofaciocutaneous, but a great deal can be done to relieve symptoms and prevent complications. In particular, being able to diagnose precocemente congenital heart defects may allow to treat, depending on the type and the severity, the heart defect. Severe narrowing of the pulmonary artery owing to a defective pulmonary valve (pulmonary valve stenosis) may require immediate intervention, either using balloon dilation or surgery. When balloon dilation is used, a plastic catheter with a deflated miniature balloon is inserted in a major blood vessel in the groin and is guided through the bloodstream into the narrowed pulmonary valve. When the balloon is inflated, the narrowed pulmonary valve widens. In less acute cases, surgery or other treatment may be carried out some time during the first few years of the child's life. When the muscular wall of the ventricle is thickened (hypertrophic cardiomyopathy), medical treatment may improve cardiac function, although surgical intervention is sometimes required.

# Chapter 3

# Aims of the present study

The aim of this thesis has been to develop and evaluate pipeline for the analysis of next-generation sequencing data. This for the detection of SNPs and indel within DNA sequences obtained by targeted re-sequencing of a genome's entire set of protein coding regions.

In this manuscript, I have provided an insight into sequencing technologies, allowing the comprehension of their fundamental principles and highlighting their impact on biomedical field, from genomic research to cancer medicine.

I have presented several examples reporting how massive parallel sequencing could be translated into clinical diagnostics for Mendelian and complex diseases, particularly cancer, enabling fast and cost-effective generation of genome-scale sequence data with exquisite resolution and accuracy.

I have been investigate by NGS the role of Ras/MAPK pathway mutations in colorectal cancer and in genetic syndromes, to underline the role of this cascade for morphological changes,gene induction and growth control.

Furthermore, I have discussed about the deep sequencing data analysis and computational and bioinformatic challenges. This overview about NGS technologies has the objective to underlined the progress and advantages in diagnostics and therapeutic management, considering at the same time the limitations. These are represented from the difficulty to find genotype-phenotype correlation and the computational demand for managing the data or interpretation.

# Chapter 4

# Materials and Methods

## 4.1 Patients

### 4.1.1 Colorectal cancer patients

Three patients (mean age: 67years) underwent surgery due to aggressive colorectal carcinoma upon informed consent. A total of eighty six specimens of colorectal carcinoma tissues, containing at least 70% of tumor cells, were obtained from Department of Pathology of Hospital Papa Giovanni XXIII (Bergamo, Italy) as FFPE blocks. Tumors were staged and graded according to criteria of system classification. DNA sample from blood was also collected for comparison. Patient's clinical information about age at diagnosis, gender and family history of colorectal cancer (at baseline) was retrieved. Information about tumour sub-localization, tumor stage and differentiation of the tumour was also retrieved.

### 4.1.2 RASopathies patients

The Cardiovascular Department of Hospital Papa Giovanni XXIII is the leading Italian centre for the surgical treatment of hypertrophic cardiomyopathy (HCM) in severely symptomatic patients. More than 200 patients have been successfully treated by miectomy and the number of patients is increasing each year. After surgery, patients and family are followed in a dedicated outpatient clinic. The Molecular Genetics Laboratory is dedicated to cardiac disease and more than 300 HCM patients/families have been investigated.
The diagnosis of HCM was based on criteria recommended by the criteria proposed by McKenna et al. [McKenna et al, 1997].

A clinical diagnosis of HCM was confirmed when 1) maximal wall thickness is two standard deviations above the mean for age and body surface area; 2) when septal thickness is > 15mm or posterior wall thickness is > 13 mm or 3) electrocardiographic abnormalities and a septal thickness > 14 mm or posterior thickness > 12 mm are present.
All the patients' relatives signed a written informed consent. Available family members were enrolled and evaluated by recording medical history and performing transthoracic echocardiography.

#### 4.1.2.1 Patient 1

Patient 1 was a newborn male, referred at 10 months to our hospital because of respiratory distress. The suspected diagnosis was obstructive hypertrophic cardiomyopathy. He presented an hypertrophy of left ventricle mainly at level of interventricular septum (15mm). Familial anamnesis was negative for cardiovascular diseases.

#### 4.1.2.2 Patient 2

A 7 months male was the first child of healthy, nonconsanguineous parents. Family history was unremarkable. Weight, length and occipital-frontal circumference were within normal limits and no facial dysmorphisms were observed. A few café-au-lait spots were noted. The echocardiogram was consistent with massive left ventricular hypertrophy. No sensorineural hearing loss, ophthalmological alterations or delayed psychomotor development were observed.

#### 4.1.2.3 Patient 3

Patient 3 was a 57 years old man, affected by a severely obstructive form of hypertrophic cardiomyopathy underwent septal miectomy. He presented left ventricular outflow tract gradient 105 mmHg, an interventricular septum of 28 mm and according the NYHA classification, he was declared as IV stage of HCM. Family history was positive with cardiovascular disease (Figure 4.1).

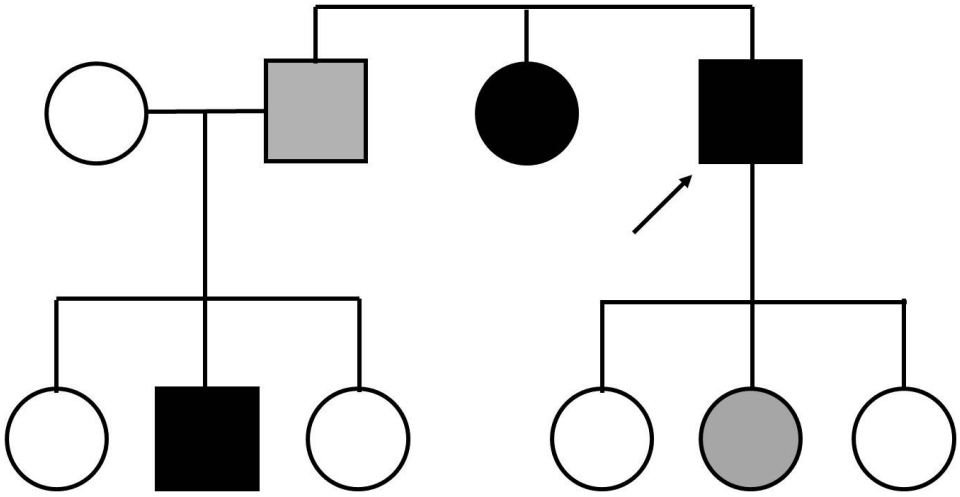

**Figure 4.1:** Pedigree of family of Patient 3. Square male, circle female, solid symbols HCM affected and grey symbols HCM mild affected. The arrow indicates Patient 3.

As shown in family pedigree, the sister of Patient 3 presented clinical findings of severe HCM, while the brother was mild affected. The brother's son, living in France, at cardiological examination presented obstructive HCM.

Interestingly, the daughter of Patient 3 presented a mild hypertrophic interventricular septum and left ventricle.

# 4.2 Samples sequencing

## 4.2.1 Colorectal cancer sequencing

Genomic DNA was extracted from seven manually microdissected 5μm sections of FFPE tissue. Two sections of the frozen specimens were prepared and stained with hematoxylin and eosin for pathological analysis and exact localization of the tumors. Deparaffination of the sections was performed by xylene.

The DNA concentration and purity was measured at 260 and 280 nm. Exon 2 of the Kras gene was amplified from isolated genomic DNA. The mRNA sequence of reference is NM_004985 with the codon 1 which corresponds to the start codon ATG and the nucleotide 1 that corresponds to the first A (van Krieken et al, 2008). The following primers were used for KRAS amplification:

5′-TGTAAAACGACGGCCAGTGGTACTGGTGGAGTATTTGATAGTG-3′ (forward) and 5′-CAGGAAACAGCTATGACCTGGATCATATTCGTCCACAAAA-3′ (reverse).
The annealing temperature was 58°C. PCR products were checked for purity and size by electrophoresis on a 2% agarose gel and subsequently used for direct sequencing. Nextera XT DNA Sample Preparation kit (Illumina) was used for the preparation of sequencing-ready libraries for small PCR amplicons that was sequenced on MiSeq platform by 2x100bp read length.

### 4.2.2 RASopathies sequencing

The genes included in targeted resequencing panel have been selected using "cardiomyopathy" as keyword in OMIM database and HGMD™ Professional database (Release 2012.2). Genomic coordinates of exons belonging to all Reference Sequence (RefSeq) isoforms of the 159 genes were collected using the UCSC table browser assembly hg19 (February 2009). An additional 30 bp of flanking intronic sequence were added to each region of interest (ROI). Genomic intervals were merged using custom script to have each region represented only once. ROI included 3,015 regions encompassing 1,037,636 bp. These regions were submitted to eArray tool (Agilent Technologies, Santa Clara, CA, USA) with Repeat Masker function activated. After a visual review of designed baits, the panel was submitted to Agilent (Agilent Technologies, Santa Clara, CA, USA) for manifacturing. After removing repeat regions, the total target was 939,470 bp.
The genomic DNA was extracted from peripheral blood cells using standard procedures. In solution hybridization capture was carried out using SureSelect kit (Agilent Technologies, Santa Clara, CA, USA) according to manufacturers' protocols. Enriched fragment library were sequenced by 2x100bp sequencing protocol on Illumina HiSeq 2000 (San Diego, CA, USA), following manufacturer procedure instructions.

## 4.3 Bioinformatic analysis

Image acquisition, image processing and signal processing have been performed using the Illumina pipeline with default parameters. Sequences were aligned to the human genome reference (GRCh37/hg19 assembly) sequence using the Burrows-Wheeler Alignment tool (BWA 0.6.2) with default parameters. PCR duplicates were removed using the Samtools

package, and base quality score recalibration, local realignment, and variant calling were performed with the Genome Analysis Toolkit (GATK v2.1.8). Variants with coverage less than 8X, with variant allele frequencies of less than 0.25, or with overall quality scores of less than 20 were discarded. NGSrich tool were used to assess the target enrichment performance, including coverage and quality of sequencing on each region of interest (Frommolt et al.,2012).

The variant calling and annotation of identified mismatches were performed by GAMES (paper VI). For copy number variants, a statistical model developed in our Laboratory (paper in preparation) was used to detect large deletions and duplications on the basis of read depth, using as control the pool of population reference samples [paper III].

## 4.5 Prioritization and interpretation

Several publicly available computational tools were used to evaluate the functional significance of variants and to assess the effect of amino acid substitution on the structure and function of the protein protein. The Grantham Score, which categorizes the differences of physicochemical properties for codon replacements, was retrieved from the original Grantham Score Matrix. The evolutionary nucleotide conservation PhyloP score in 46 vertebrates was retrieved from UCSC Genome Browser.

*In silico* predictive algorithms were applied: PolyPhen-2, Align-GVGD, SIFT and MutationTaster. Using dbSNP Release 137, ESP ( 1000 Genomes Project (1KGP) and HGMDTM Professional (Release 2012.3) , the identified variants were prioritized, according also to minor allele frequency (MAF), when available. Variants were classified as: ‘novel’, if only a single allele was present in a parent and none were seen in dbSNPv137, ESP or 1KGP; ‘rare’, if they did not meet the criteria for novel and were present in ,1% of controls; and ‘common’, if they were present in >1% of controls.

# Chapter 5

# Results

## 5.1 Colorectal cancer

Using targeted NGS sequencing, we obtained a total of 1,981,308 high-quality reads encompassing exon 2 of Kras gene. After mapping to the reference human genome (GRCh37/hg19), 83.74% of the yielded clean reads could be uniquely matched to target regions, and 99.54% of the targeted region was covered in at least a 30-fold depth in each sample. The average coverage depth for exon was 250X. Thus, the coverage should have been adequate to reliably detect DNA variants within the ROI.

In exon 2 of the Kras oncogene, a mutation was observed in codon 12. It was a G>A transition in position chr12:g.25398284C>T (NM_033360.2:c.35G>A) that lead at protein level an aminoacid change Gly12Asp (Figure 5.1 and 5.2). KRAS proteins in which the wild-type glycine residue is replaced by an aspartate represent approximately 50% of tumors of colon.

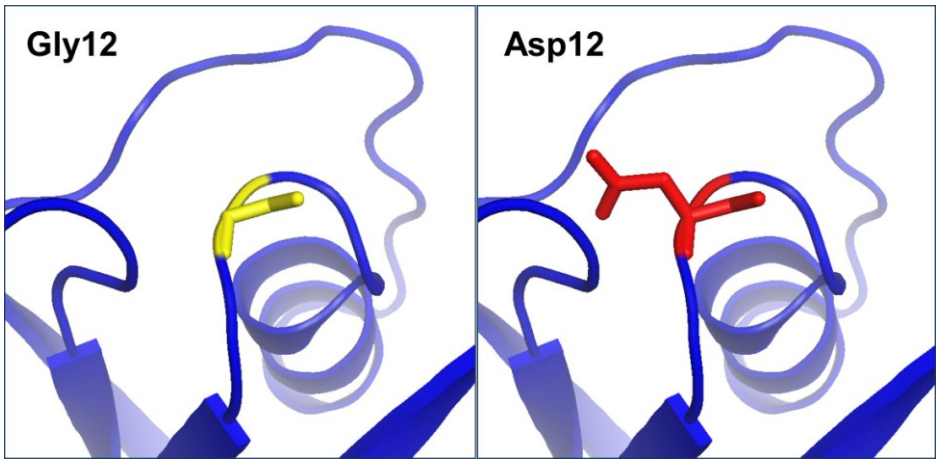


**Figure 5.1:** Structural modeling of amino acid change Gly12Val in KRAS protein.

It is known that the amino acid change position plays a fundamental role in definition of aggressiveness in CRC. For example, CRC carrying Kras codon 12 mutation may increase

aggressiveness not by altering proliferative pathways, but by the differential regulation of Kras downstream pathways that lead to inhibition of apoptosis, enhanced loss of contact inhibition, and increased predisposition to anchorage-independent growth. These results offer a molecular explanation for the increased aggressiveness of the tumors with Kras codon 12 mutations observed in the clinical setting.

For what concerns the type of amino acid change, KRAS Gly12Asp activates phosphatidylinositol 3-kinase and MEK signaling, whereas those with mutant Gly12Cys or mutant Gly12Val had activated Raf signaling and decreased growth factor–dependent Akt activation and are associated were associated with decreased progression-free survival. Molecular modeling studies showed that different conformations imposed by mutant KRas may lead to altered association with downstream signaling transducers.Since this missense mutation abolishes GTPase activity resulting in constitutively activated Ras signaling, patients who carry this amino acid change are unlikely to benefit from the anti-EGFR treatments because their tumors express a protein that signals cell proliferation without the activation of EGFR.

This implicates that therapeutic interventions may need to take into account the specific mutant KRas protein expressed by the tumor

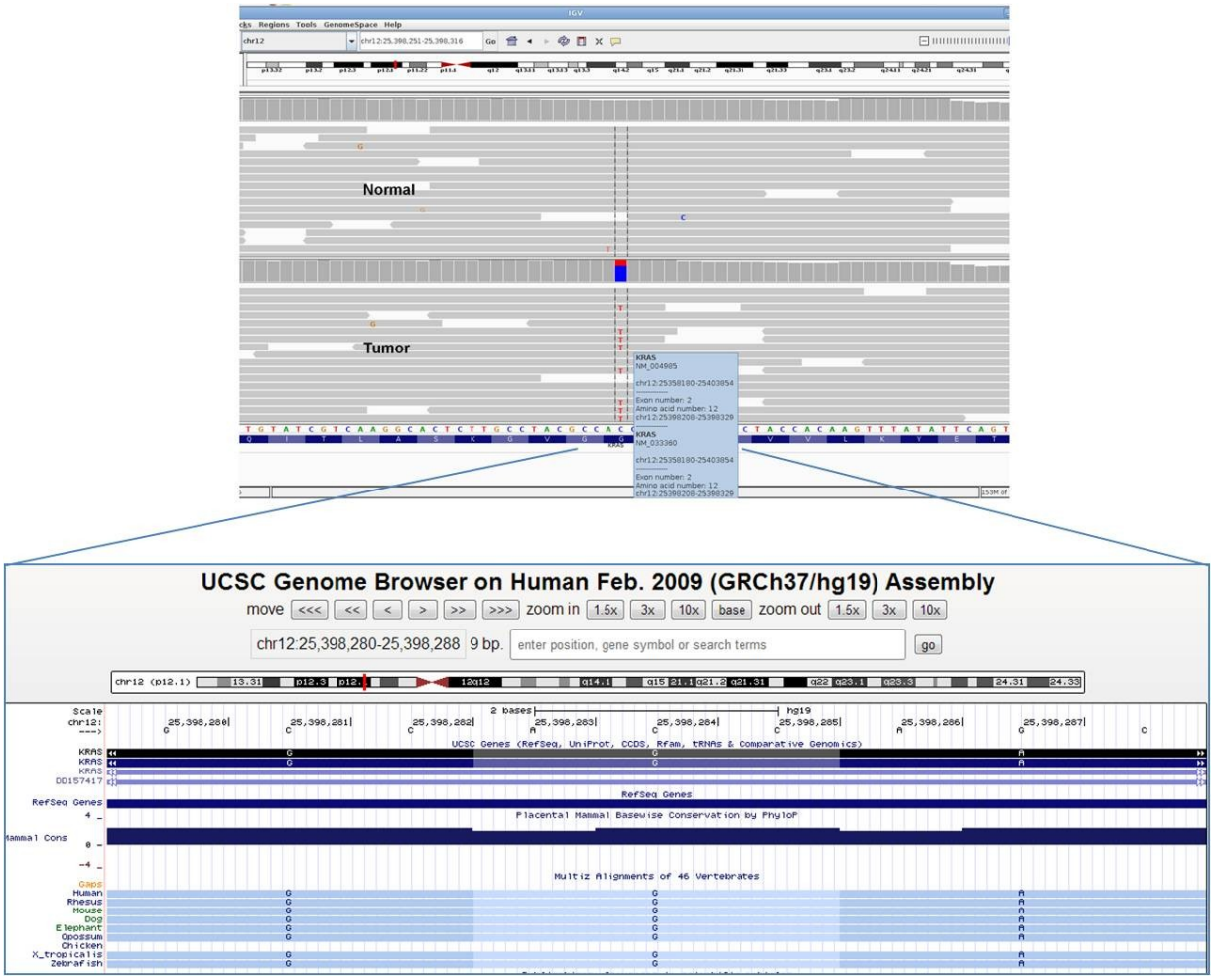


**Figure 5.2:** Visualization of KRAS-mutant SNV using IGV Browser. Coverage file and bam files are shown in the top part of the IGV Browser output. The bottom part of the picture consist of Refseq gene with nucleotides visualized in UCSC Genome Browser. It is shown also the conservation of aminoacid Gly12 in different species.

# 5.2 RASopathies

## 5.2.1 Patient 1

The sequencing run produced 7,871,442 100-bp long sequence reads. The number of mappable reads to the region of interest was 5,002,724. Mean depth of coverage was 419X with 95.52% of target region covered >30X, demonstrating the high performance of the sequence capture approach. All successfully mapped sequence reads were analyzed

to detect sequence variants, including insertions and deletions. Variants were called automatically, generating a list of sequence variants compared with the reference sequence (hg19), which we had annotated with NCBI dbSNP build 135, HGMD, ESP and an internal database comprising disease mutations derived from HCM molecular characterization databases (http://genepath.med.harvard.edu/_seidman/cg3/index.html and http://angis.org.au/Databases/Heart/).

Additionally, all novel nonsynonymous variants were filtered against the 6500 exomes released from the Exome Sequencing project. All together, 3,004 known or novel variants were detected in Patient 1. Of these variants, 387 (19.3%) are in coding regions or exon-intron junction, 265 (71,1%) are predicted to be noncoding or synonymous, whereas the remaining are nonsynonymous, leading to the exchange of 1 or more amino acids.

Using our mutation detection workflow, we identified in Patient 1 a heterozygous missense mutation, in position chr12(GRCh37/hg19):g.112915494 (NM_002834.3:c.893A>G) resulting in an aminoacid change p.Asn298Ser within exon 8 of *PTPN11* gene (Figure 5.3). The substituted nucleotide was moderately conserved (phyloP=3.19 [-14.1;6.4]). The aminoacid change caused small physicochemical difference (Grantham dist:46 [0-215]). *In silico* disease-causing potential prediction considered the missense mutation as deleterious. The mutation is not reported in dbSNP

We did not identified mutations in sarcomeric genes. The results were validated independently by Sanger sequencing.

The identification of this mutation supplied the indication for diagnosis of Noonan syndrome. It is known in literature that this complex genetic disease is characterized by cardiovascular defects as phenotypic features. Patient 1 was affected by hypertrophic cardiomyopathy, but the genetic findings by NGS underlined that this defects is secondary to Noonan syndrome.

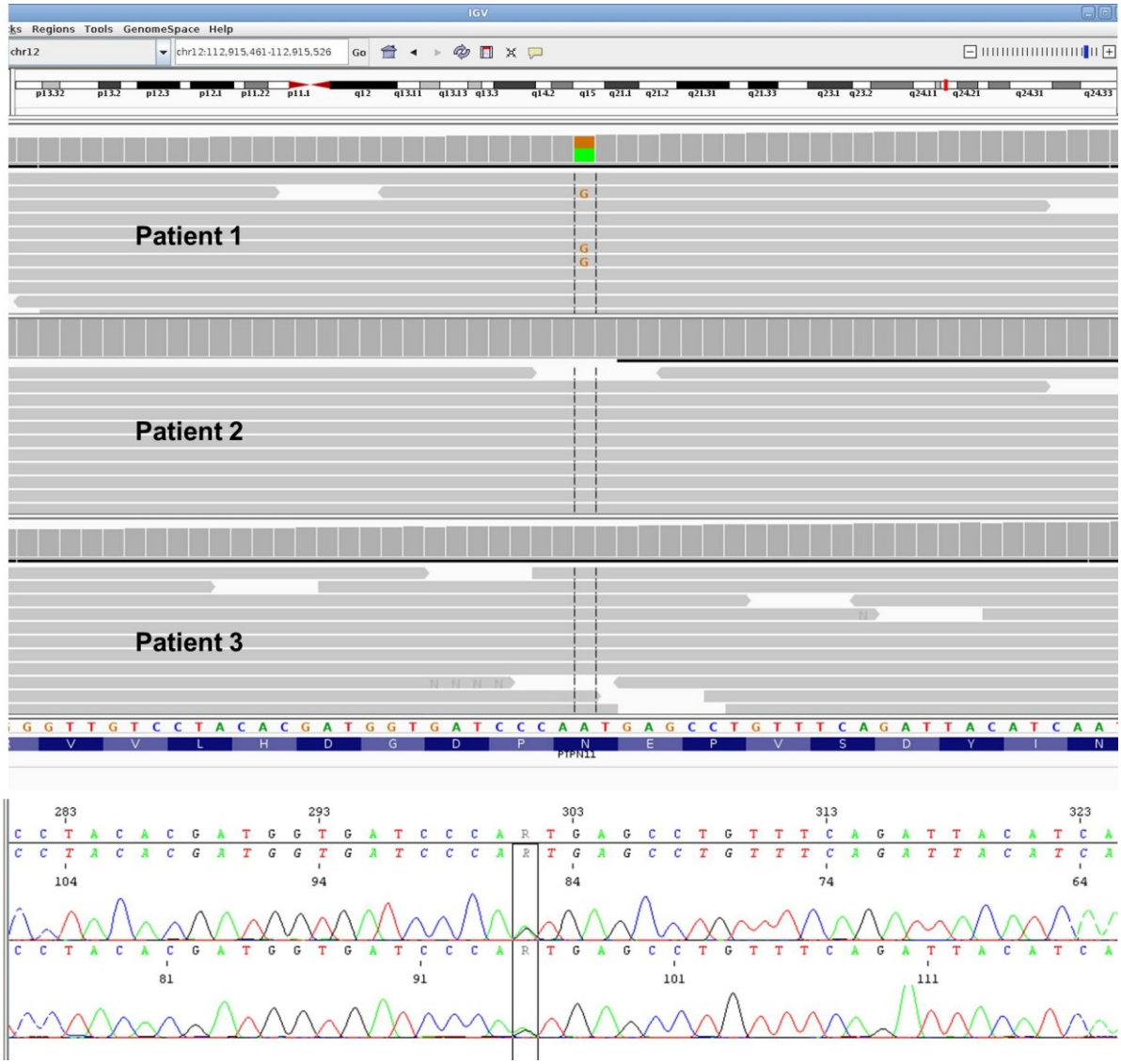


**Figure 5.3:** PTPN11 mutation in Patient 1. Identification of (NM_002834.3:c.893A>G) by NGS, as visualized in IGV in Patient 1. In the bottom part of the figure the Sanger sequencing validation has been reported.

## 5.2.2 Patient 2

We obtained 2,547,492 mapped reads, with 396X as mean coverage. The 93.26% of the target regions was covered more than 30X. By GAMES (Sana et al., 2011) we have identified 2617 variants, of which 287 falling in coding regions and intron-exon junctions. Using dbSNP Release 135, ESP and HGMD™ Professional (Release 2011.4) , we have

filtered out those variants with minor allele frequency (MAF) >0.05%. To evaluate the pathogenic potential of detected variants, we have applied *in silico* predictive algorithms.

By this approach, we detected a homozygous mutation, in position chr12(GRCh37/hg19):g.112910827 (NM_002834.3:c.836A>G) leading to the exchange of tyrosine with cysteine at position 279 in PTPN11 protein (Figure 5.4). Both nucleotide and amino acid resulted as highly conserved. Grantham score for amino acid change was high, revealing high phycochemical difference (Grantham dist: 194). The mutation was reported in dbSNP v135.

Sanger sequencing confirmed the identified mutation (Figure 5.5).

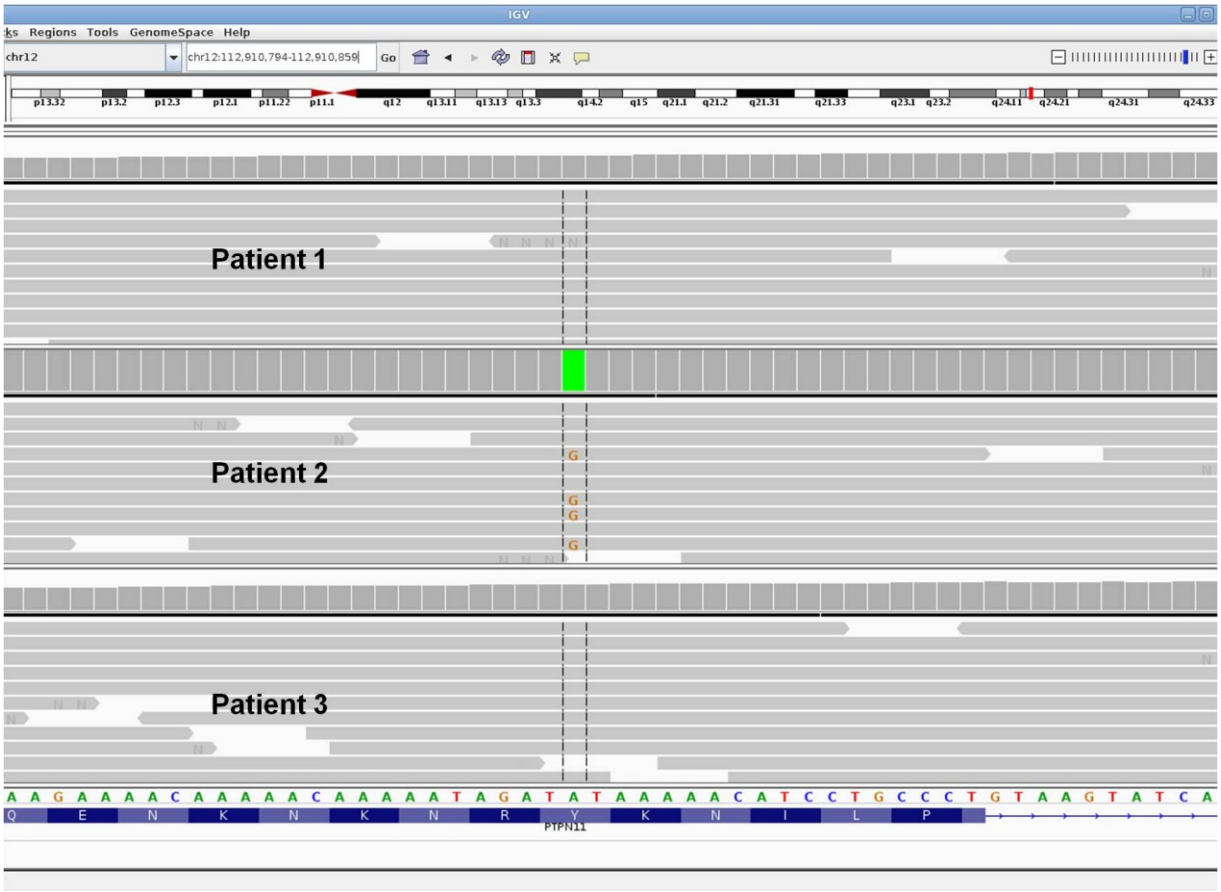


**Figure 5.4:** Patient 2 sequencing reads overlapping the missense mutation. It is display the homozygous mutation c.836A>G.

The identified mutation was previously associated to Leopard syndrome. The diagnosis agrees with the clinical features of patient 2, characterized by hypertrophic cardiomyopathy and café-au-lait spots.

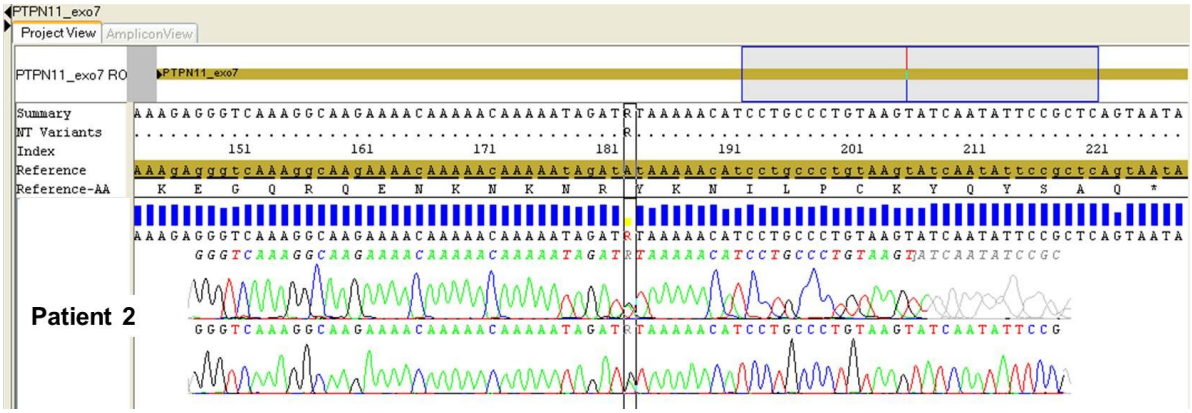


**Figure 5.5:** Sanger sequencing validation.

## 5.2.3 Patient 3

Sequencing of Patient 3 and relative produced on average 15,582,644reads, 69.6% in target regions. Mean coverage was 745X.

Brother and sister of Patient 3 carry a mutation p.Met932Lys in exon 23 of MYH7 gene, known to be associated with HCM (Figure 5.6). Retrospective genetic analysis of nephew detected the same missense mutation. Interestingly, MYH7 was wild-type in Patient 3 and in his daughter (Figure 5.6).

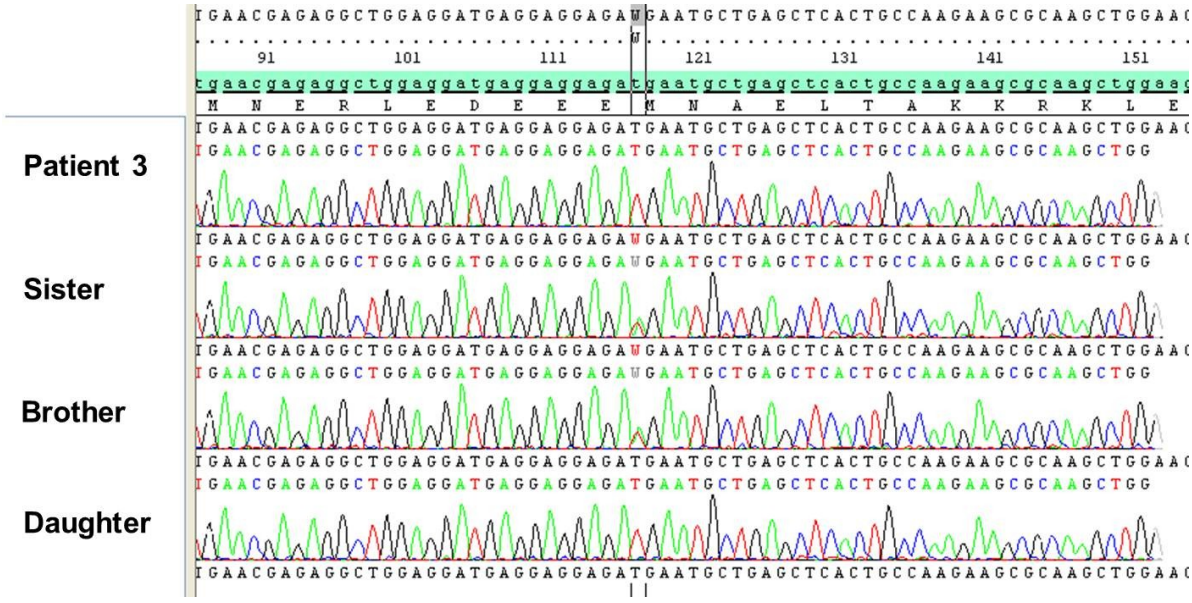


**Figure 5.6:** Mutation in MYH7 gene in brother and sister of Patient 1.

However, the processing of data detected another heterozygous mutation in Patient 3 and his daughter within exon 3 of HRAS gene (GRCh37/hg19, chr11:533854; NM_005343.2:c-202C>T) (Figure 5.7). This nucleotide substitution leads to aminoacid change (p.Arg68Trp), predicted to be deleterious by SIFT, AGVGD and Mutation Taster. It's known that is mutation is associated to Costello syndrome.

The phenotypic features of this disease include hypertrophic cardiomyopathy and mental retardation. A more accurate clinical evaluation of daughter of Patient 3 confirmed the diagnosis of Costello syndrome (Figure 5.8).

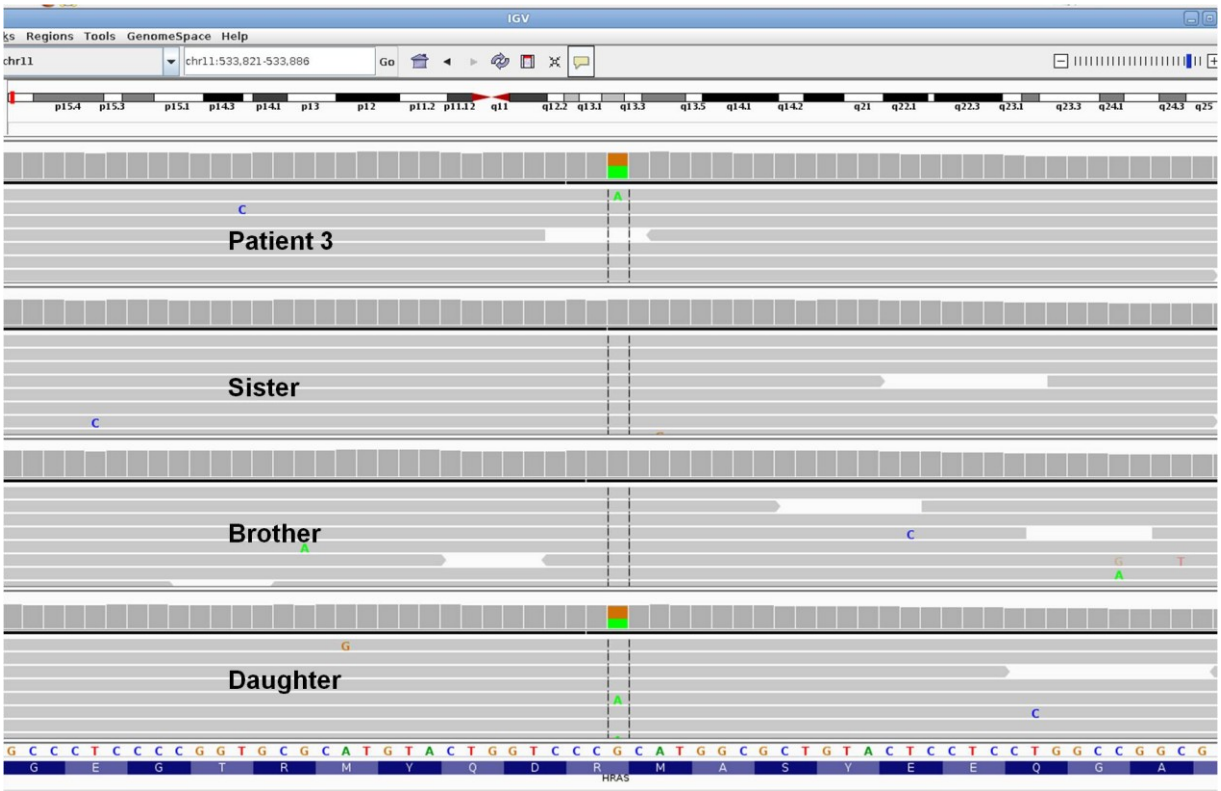


**Figure 5.7:** Identification of HRAS mutation mutation (NM_002834.3:c.893A>G) by NGS, as visualized in IGV in Patient 1.

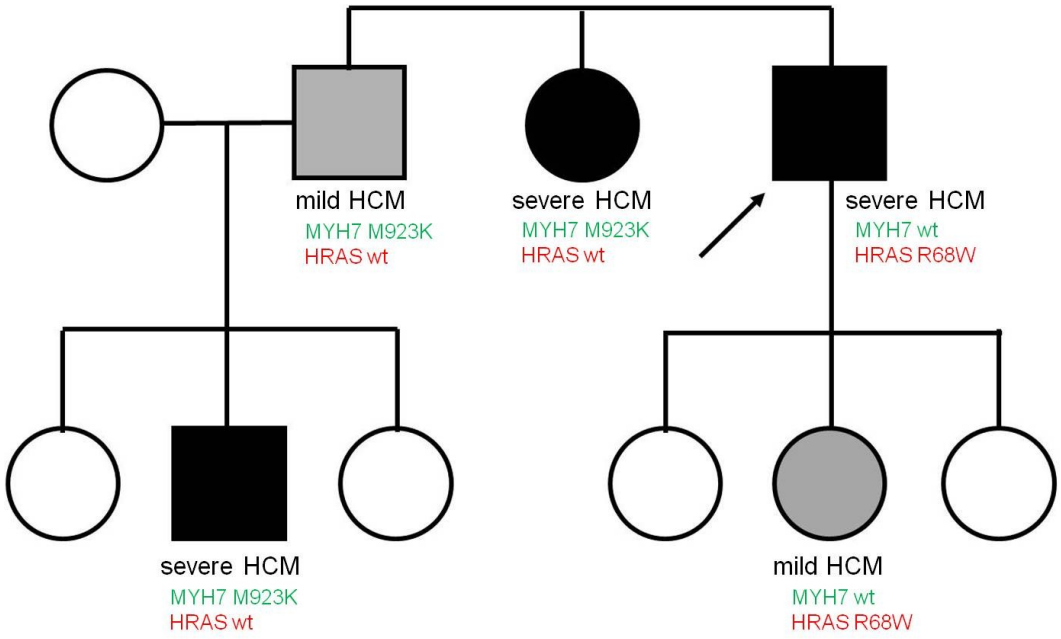


**Figure 5.8:** Pedigrees in which the MYH7 (green) and HRAS (red) mutations segregated. The genotypic and phenotypic status of subjects are also indicated.

# Chapter 6

# Conclusions

Ultra high-throughput sequencing is revolutionizing the study of human genetics and has immense clinical implications. It has reduced the cost and increased the throughput of genomic sequencing. NGS instruments make possible to rapidly generate large amounts of sequence data at substantially lower cost respect to the traditional sequencing approaches. These high-throughput sequencing technologies (e.g. Roche/454, Life Technology SOLiD, Ion Torrent and Illumina) make whole genome sequencing and resequencing, transcript sequencing as well as gene expression, DNA-protein interactions and DNA methylation feasible at an unanticipated scale. Furthermore, NGS enable scientist for the first time to analyze millions of DNA sequences in a single run.

At the same time, massive parallel sequencing is a growing field with many computational challenges. A normal deep sequencing run outputs a massive amount of data which require complex computational processing and interpretation. The overflow of available bioinformatic tools and software for each of the optional analysis steps presents a challenge for the researcher aiming to evaluate and interpret deep sequencing data. This the reason for which currently, new efficient and well designed bioinformatics tools are emerging, which are addressing different tasks in the downstream analysis of NGS data.

During the course of my PhD, I have analyzed and improved current protocols and algorithms for next generation sequencing data, taking into account the specific characteristics of these new sequencing technologies. Combining these tools into an integrated analysis pipeline greatly facilitates the interpretation of NGS results and could help the life science investigator in comprehension of molecular mechanisms.

The presented approaches and algorithms were applied in different projects and are widely used within the diagnostics in Laboratory of Molecular Genetics at Hospital Papa Giovanni XXIII in Bergamo.

Since two years, NGS is becoming a powerful method in clinical setting. With the release of benchtop sequencers and improvements in other platforms, the turnaround time for sequencing is getting quicker. Although, today there are a number of steps in the process -

sample prep, time on the sequencer, Sanger confirmation, interpretation, alignment, and variant calling - that make NGS not yet compatible with diagnostics. The speed of computational analyses and algorithms will get faster, and the instruments are already getting faster. The need to develop confirmation assays will diminish, so the interpretive process will get quicker, and there are a lot of pieces of the process that are all on their own path to improved efficiency. Furthermore, there is considerable interest in the use of next-generation sequencing to help diagnose unidentified genetic conditions, but it is difficult to predict the success rate in a clinical setting that includes patients with a broad range of phenotypic presentations. However, recent studies provide evidence that next-generation sequencing can have high success rates in a clinical setting, for the identification of potential therapeutic targets, for insight into the heterogeneity of cancer and for differential diagnosis in complex diseases.

In conclusion my PhD thesis highlights both the challenges and opportunities in the application of NGS to clinical diagnosis in patients with complex diseases, such as tumors, and heterogeneous genetic syndromes. This study focuses also on the computational challenges introduced by these new technologies.

# References

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. (2010) A map of human genome variation from population-scale sequencing. Nature 467,1061–1073

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

Amado RG, Wolf M, Peeters M, Van Cutsem E, Siena S, Freeman DJ, Juan T, Sikorski R, Suggs S, Radinsky R, Patterson SD, Chang DD. (2008) Wild-Type KRAS Is Required for Panitumumab Efficacy in Patients With Metastatic Colorectal Cancer. J Clin Oncol. 26:1626-34.

American Joint Committee on Cancer. (2002) AJCC Cancer Staging Manual, 6th ed. New York: Springer-Verlag;

Andreyev HJ, Norman AR, Cunningham D, et al . (1998) Kirsten ras mutations in patients with colorectal cancer: a multicenter “RASCAL” study. J Nat Cancer Inst. 90:675-684.

Ansorge W, Sproat B, Stegemann J, Schwager C, Zenke M. (1987) Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. Nucleic Acids Res, 15:4593–4602.

Ansorge WJ. (2009) Next-generation DNA sequencing techniques. N Biotechnol. 25:195-203.

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

Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. (2003) Transforming single DNA molecules into fluorescent magnetic particles fr detection and enumeration of genetic variations. Proc Natl Acad Sci U S A. 100:8817-22.

Ewing B, Green P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res., 8, 186–194.

Ewing B, Hillier L, Wendl MC, Green P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res., 8, 175–185.

Fernández-Medarde A, Santos E. (2011) Ras in Cancer and Developmental Diseases. Genes & Cancer 2: 344

Flanagan, S. E., Patch, A.-M., & Ellard, S. (2010). Using SIFT and PolyPhen to predict loss-offunction and gain-of-function mutations. Genetic Testing and Molecular Biomarkers,14, 533-537.

Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR, Futreal PA. (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. 39(Database issue):D945-50.

Frazer KA, Murray SS, Schork NJ, Topol EJ. (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet. 10:241–251.

Frommolt P, Abdallah AT, Altmüller J, Motameny S, Thiele H, Becker C, Stemshorn K, Fischer M, Freilinger T, Nürnberg P. (2012) Assessing the enrichment performance in targeted resequencing experiments. Hum. Mutat. 33, 635–641.

Goya R, Sun MG, Morin RD, Leung G, Ha G, Wiegand KC, Senz J, Crisan A, Marra MA, Hirst M, Huntsman D, Murphy KP, Aparicio S, Shah SP.(2010) SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics. 26:730-6.

Grantham R. (1974) Amino acid difference formula to help explain protein evolution. Science 185, 862–864.

Gripp KW, Lin AE. (2011) Costello syndrome: A Ras/mitogen activated protein kinase pathway syndrome (rasopathy) resulting from HRAS germline mutations. Genet Med. [Epub ahead of print]

Guerrero S, Casanova I, Farré L, Mazo A, Capellà G, Mangues R.(2000) K-ras codon 12 mutation induces higher level of resistance to apoptosis and predisposition to anchorage-independent growth than codon 13 mutation or proto-oncogene overexpression. Cancer Res. 60:6750-6.

Haigis KM, Kendall KR, Wang Y, Cheung A, Haigis MC, Glickman JN, Niwa-Kawakita M, Sweet-Cordero A, Sebolt-Leopold J, Shannon KM, Settleman J, Giovannini M, Jacks T. (2009) Differential effects of oncogenic K-Ras and N-Ras on proliferation, differentiation and tumor progression in the colon. Nat Genet. 40:600-8.

Heinemann V, Stintzing S, Kirchner T, Boeck S, Jung A. (2009) Clinical relevance of EGFR- and KRAS-status in colorectal cancer patients treated with monoclonal antibodies directed against the EGFR. Cancer Treat Rev. 35:262-71.

Hilger RA, Scheulen ME, Strumberg D.(2002) The Ras-Raf-MEK-ERK Pathway in the Treatment of Cancer. Onkologie 25:511–518

Homer N, Merriman B, Nelson SF. (2009) BFAST: an alignment tool for large scale genome resequencing. PLoS One 4:e7767

Isakov O, Modai S, Shomron N. (2011) Pathogen detection using short-RNA deep sequencing subtraction and assembly. Bioinformatics. 27:2027-30.

Kaiser J. (2008) DNA sequencing. A plan to capture human diversity in 1000 genomes. Science, 319, 395.

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

Karnoub AE, Weinberg RA. (2008) Ras oncogenes: split personalities. Nat Rev Mol Cell Biol. 9:517-31.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. (2002) The human genome browser at UCSC. Genome Res., 12, 996–1006.

Kerr B, Delrue MA, Sigaudy S, Perveen R, Marche M, Burgelin I, Stef M, Tang B, Eden OB, O'Sullivan J, De Sandre-Giovannoli A, Reardon W, Brewer C, Bennett C, Quarell O, M'Cann E, Donnai D, Stewart F, Hennekam R, Cavé H, Verloes A, Philip N, Lacombe D, Levy N, Arveiler B, Black G. (2006) Genotype-phenotype correlation in Costello syndrome: HRAS mutation analysis in 43 cases J Med Genet 43:401–405.

Kranenburg O. (2005) The KRas oncogene: past, present and future. Biochim Biophys Acta. 1756:81-82.

Kumar P, Henikoff S, Ng PC. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc., 4, 1073–1081.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C,

Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ; International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. (2007) Clustal W and Clustal X version 2.0. Bioinformatics, 23, 2947-2948.

Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, Cook L, Abbott R, Larson DE, Koboldt DC, Pohl C, Smith S, Hawkins A, Abbott S, Locke D, Hillier LW, Miner T, Fulton L, Magrini V, Wylie T, Glasscock J, Conyers J, Sander N, Shi X, Osborne JR, Minx P, Gordon D, Chinwalla A, Zhao Y, Ries RE, Payton JE, Westervelt P, Tomasson MH, Watson M, Baty J, Ivanovich J, Heath S, Shannon WD, Nagarajan R, Walter MJ, Link DC, Graubert TA, DiPersio JF, Wilson RK. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature. 456:66-72. d

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

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

Li H, Ruan J, Durbin R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res., 18, 1851–1858.

Lièvre A, Bachet JB, Le Corre D, Boige V, Landi B, Emile JF, Côté JF, Tomasic G, Penna C, Ducreux M, Rougier P, Penault-Llorca F, Laurent-Puig P. (2006) KRAS Mutation Status Is Predictive of Response to Cetuximab Therapy in Colorectal Cancer. Cancer Res. 66:3992-5.

Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ. (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol. 30:434-9.

Malumbres M, Barbacid M. (2003) RAS oncogenes: the first 30 years. Nat Rev Cancer. 3:459-65.

Mardis ER. (2008) The impact of next-generation sequencing technology on genetics. Trends Genet. 24:133-41.

Markowitz SD, Bertagnolli MM. (2009) Molecular origins of cancer: Molecular basis of colorectal cancer. N Engl J Med. 361:2449-60.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297-303.

McKenna WJ, Spirito P, Desnos M, Dubourg O, Komajda M (1997) Experience from clinical genetics in hypertrophic cardiomyopathy:proposal for new diagnostic criteria in adult members of affected families. Heart 77:130–132

McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang L, Sheridan A, Fu H, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM, Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR, Beaudoin RE, Coleman BE, Laptewicz MW, Sannicandro AE, Rhodes MD, Gottimukkala RK, Yang S, Bafna V, Bashir A, MacBride A, Alkan C, Kidd JM, Eichler EE, Reese MG, De La Vega FM, Blanchard AP. (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligationsequencing using two-base encoding. Genome Res. 19:1527-41.

Metzker ML. (2010) Sequencing technologies - the next generation. Nat. Rev. Genet., 11, 31–46.

Meyerson M, Gabriel S, Getz G.(2010) Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet. 11:685-96.

Molina JR, Adjei AA. (2006) The Ras/Raf/MAPK Pathway. J Thorac Oncol. 1:7-9.

Rauen KA. (2007) HRAS and the Costello syndrome. Clin Genet. 71: 101–108

Razzaque MA, Nishizawa T, Komoike Y, Yagi H, Furutani M, Amo R, Kamisago M, Momma K, Katayama H, Nakagawa M, Fujiwara Y, Matsushima M, Mizuno K, Tokuyama M, Hirota H, Muneuchi J, Higashinakagawa T, Matsuoka R. (2007) Germline gain-of-function mutations in RAF1 cause Noonan syndrome. Nat Genet. 39:1013-7.

Roberts AE, Araki T, Swanson KD, Montgomery KT, Schiripo TA, Joshi VA, Li L, Yassin Y, Tamburino AM, Neel BG, Kucherlapati RS. Germline gain-of-function mutations in SOS1 cause Noonan syndrome (2007). Nat Genet. 39:70-4.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. (2011) Integrative genomics viewer. Nat Biotechnol. 29:24-6.

Roe BA. (2004) Shotgun Library Construction for DNA Sequencing. Methods Mol Biol. 255:171-87.

Ronaghi M, Uhlén M, Nyrén P. (1998) A sequencing method based on real-time pyrophosphate. Science. 17;281:363, 365.

Ronaghi M. (2001) Pyrosequencing sheds light on DNA sequencing. Genome Res. 2001 Jan;11(1):3-11.

Ross JS, Cronin M. (2011) Whole cancer genome sequencing by next-generation methods. Am J Clin Pathol. 136:527-39.

Sanger F, Nicklen S, Coulson AR. (1977) DNA sequencing with chain-terminating inhibitors. Proc.Natl. Acad. Sci. U. S. A. 74, 5463–5467.

Schubbert S, Zenker M, Rowe SL, Böll S, Klein C, Bollag G, van der Burgt I, Musante L, Kalscheuer V, Wehner LE, Nguyen H, West B, Zhang KY, Sistermans E, Rauch A, Niemeyer CM, Shannon K, Kratz CP. (2006) Germline KRAS mutations cause Noonan syndrome. Nature Genetics. 38:331-336.

Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. (2010) MutationTaster evaluates disease-causing potential of sequence alterations. Nat. Methods 7, 575–576.

Shearer AE, DeLuca AP, Hildebrand MS, Taylor KR, Gurrola J 2nd, Scherer S, Scheetz TE, Smith RJ.(2010) Comprehensive genetic testing for hereditary hearing loss using massively parallel sequencing. Proc Natl Acad Sci U S A. 107:21104-9

Shendure J, Ji H. (2008) Next-generation DNA sequencing. Nat. Biotechnol., 26, 1135–1145.

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

Smigielski EM, Sirotkin K, Ward M, Sherry ST. (2000) dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res., 28, 352–355.

Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, Woolf B, Shen L, Donahue WF, Tusneem N, Stromberg MP, Stewart DA, Zhang L, Ranade SS, Warner JB, Lee CC, Coleman BE, Zhang Z, McLaughlin SF, Malek JA, Sorenson JM, Blanchard AP, Chapman J, Hillman D, Chen F, Rokhsar DS, McKernan KJ, Jeffries TW, Marth GT, Richardson PM. (2008) Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res. 18:1638-42.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell C, Heiner C, Kent SB, Hood LE. (1986) Fluorescence detection in automated DNA sequence analysis. Nature, 321:674–679.

Soulières D, Greer W, Magliocco AM, Huntsman D, Young S, Tsao MS, Kamel-Reid S. (2010) KRAS mutation testing in the treatment of metastatic colorectal cancer with anti-EGFR therapies. Curr Oncol. 17 Suppl 1:S31-40.

Stenson PD, Ball EV, Howells K, Phillips AD, Mort M, Cooper DN. (2009) The human gene mutation database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics.Hum Genomics. 4:69-72.

Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, de Silva D, Zharkikh A, Thomas A. (2006) Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J Med Genet. 43:295-305.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. (2012) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. [Epub ahead of print]

Tidyman WE, Rauen KA. (2008) Noonan, Costello and cardio-facio-cutaneous syndromes: dysregulation of the Ras-MAPK pathway. Expert Rev Mol Med.10:e37.

Tidyman WE, Rauen KA. (2009) The RASopathies: developmental syndromes of Ras/MAPK pathway dysregulation. Curr Opin Genet Dev. 19:230-6.

Tomkinson AE, Vijayakumar S, Pascal JM, Ellenberger T.(2006) DNA ligases: structure, reaction mechanism, and function. Chem Rev. ;106:687-99.

van der Burgt I. (2007) Noonan syndrome. Orphanet J Rare Dis. 2:1-6.

van Krieken JH, Jung A, Kirchner T, Carneiro F, Seruca R, Bosman FT, Quirke P, Fléjou JF, Plato Hansen T, de Hertogh G, Jares P, Langner C, Hoefler G, Ligtenberg M, Tiniakos D, Tejpar S, Bevilacqua G, Ensari A. (2008) KRAS mutation testing for predicting response to anti-EGFR therapy for colorectal carcinoma: proposal for an European quality assurance program. Virchows Arch. 453:417-31.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X.(2001) The sequence of the human genome. Science 291:1304–51.

Wold B, Myers RM. (2008) Sequence census methods for functional genomics. Nat Methods, 5:19–21.

# Appendix

# List of Publications

This thesis is based on the following articles, which are referred to in the text by their Roman numerals:

I. Iascone M, Sana ME, Ferrazzi P. Letter by iascone et Al regarding article, "population-based variation in cardiomyopathy genes". Circ Cardiovasc Genet. 2012 Dec 1;5:e57.

II. Iascone M, Sana ME, Pezzoli L, Bianchi P, Marchetti D, Fasolini G, Sadou Y, Locatelli A, Fabiani F, Mangili G, Ferrazzi P. Extensive Arterial Tortuosity and Severe Aortic Dilation in a Newborn With an EFEMP2 Mutation. Circulation. 2012 Dec 4;126:2764-8.

III. Pezzoli L, Sana ME, Ferrazzi P, Iascone M. A new mutational mechanism for hypertrophic cardiomyopathy. Gene. 2012 Oct 10;507(2):165-9.

IV. Volinia S, Galasso M, Sana ME, Wise TF, Palatini J, Huebner K, Croce CM. Breast cancer signatures for invasiveness and prognosis defined by deep sequencing of microRNA. Proc Natl Acad Sci USA. 2012 Feb 21;109:3024-9.

V. Zauli G, Voltan R, di Iasio MG, Bosco R, Melloni E, Sana ME, Secchiero P. miR-34a induces the downregulation of both E2F1 and B-Myb oncogenes in leukemic cells. Clin Cancer Res. 2011 May 1;17(9):2712-24.

VI. Sana ME, Iascone M, Marchetti D, Palatini J, Galasso M, Volinia S. GAMES identifies and annotates mutations in next-generation sequencing projects. Bioinformatics. 2011 Jan 1;27(1):9-13.

VII. Volinia S, Galasso M, Costinean S, Tagliavini L, Gamberoni G, Drusco A, Marchesini J, Mascellani N, Sana ME, Abu Jarour R, Desponts C, Teitell M, Baffa R, Aqeilan R, Iorio MV, Taccioli C, Garzon R, Di Leva G, Fabbri M, Catozzi M, Previati M, Ambs S, Palumbo T, Garofalo M, Veronese A, Bottoni A, Gasparini P, Harris CC, Visone R, Pekarsky Y, de la Chapelle A, Bloomston M, Dillhoff M, Rassenti LZ, Kipps TJ, Huebner K, Pichiorri F, Lenze D, Cairo S, Buendia MA, Pineau P, Dejean A, Zanesi N, Rossi S, Calin GA, Liu CG, Palatini J, Negrini M,

Vecchione A, Rosenberg A, Croce CM. Reprogramming of miRNA networks in cancer and leukemia.Genome Res. 2010 May;20(5):589-99.

VIII. Galasso M, Sana ME, Volinia S. Non-coding RNAs: a key to future personalized molecular therapy? Genome Med. 2010 Feb 18;2(2):12.

IX. Galasso M, Sana ME, Volinia S. Deep sequencing of non coding RNA: a model for genomics in personalized medicine. Accademia delle Scienze di Ferrara, Atti, volume 86, Anno Accademico 186, 2008-2009.

# List of Abbreviations

1KGP 1000 genomes project

AGVGD align GVGD

BAM binary alignment mapping

bp base-pairs

BRAF v-raf murine sarcoma viral oncogene homolog B1

BWA Burrows-Wheeler alignment

CM cardiomyopathy

CMOS complementary metal-oxide semiconductors

CNV copy number variant

CRC colorectal cancer

CS Costello syndrome

ddNTP 2',3' dideoxynucleotide

DNA deoxyribonucleic acid

EGF epidermal growth factor

EGFR epidermal growth factor receptor

ERK extracellular signal-regulated kinase

ESP exome sequencing project

FFPE formalin-fixed and paraffin

GA Genome Analyzer

GAP GTPase activating protein

GDP guanosine diphosphate

GEF guanine-nucleotide-exchange factor

GRB2 growth-factor-receptor-binding protein 2

GTP guanosine triphosphate

HCM hypertrophic cardiomyopathy

HGMD human gene mutation database

HOCM obstructive hypertrophic cardiomyopathy

Hras Harvey rat sarcoma viral oncogene homolog

IGV integrative genomcis viewer

INDEL insertion/deletion

ISFET ion-sensitive field-effect transistors
ISV interventricular septum
KEGG Kyoto Encyclopedia of Genes and Genomes
Kras Kirsten rat sarcoma viral oncogene homolog
LS Leopard syndrome
LV left ventricle
LVOT left ventricular outflow tract
mAb monoclonal antibody
MAF minor allele frequency
MAPK Ras/mitogen activated protein kinase
MEK MAPK/ERK kinase
MYH7 myosin, heavy chain 7, cardiac muscle, beta
NGS next-generation sequencing
Nras neuroblastoma RAS viral oncogene homolog
NS Noonan syndrome
NYHA New York Heart Association
PCR polymerase chain reaction
PE paired-end
PGM personal genome machine
PI3K phosphatidylinositide 3-kinases
PolyPhen-2 polymorphism phenotyping v2
PTPN11 tyrosine-protein phosphatase non-receptor type 11
ROI region of interest
SAM sequencing alignment mapping
SIFT sorting intolerant from tolerant
SNP single nucleotide polymorphism
SNV single nucleotide variant
SOS son of sevenless
ssDNA single strand DNA
TCGA cancer genome atlas
UCSC University of California Santa Cruz
wt wild type