Alma Mater Studiorum · University of Bologna

School of Science
Department of Physics and Astronomy
Master Degree in Physics

Statistical Analysis of Genetic and Epigenetic Features in Cancer Cells

Supervisor:

Prof. Gastone Castellani

Co-supervisors:
Prof. Luca Morandi

Dott.ssa Alessandra Merlotti

Submitted by:

Francesco Casadei

Academic Year 2020/2021

Abstract

Cancer is one of the leading causes of death in almost every country and, in 2020, 19.3 million new cases and 10 million cancer deaths in the world have been estimated by WHO. The onset of a tumor is often accompanied with a set of genetic and epigenetic alterations, whose understanding can have both diagnostic role and prognostic power for targeted treatments. The spread of NGS platforms, which allow to sequence an entire human genome in a short time and relatively low cost, and the use of statistical methods that help in dealing with such huge amount of data and in finding hidden relationships, have a crucial role in the development of the precision medicine.

The present thesis work consists in two projects. The first one is a study of point mutations and methylation of a cohort of patients diagnosed with Glioblastoma (GBM), involving both Illumina sequencing-by-synthesis platform and Oxford Nanopore Technologies. The second one is an application of Dirichlet Process, a statistical learning method, to a set of Multiple Myeloma (MM) patients characterized by Copy Number Variant (CNV) measures.

The study of GBM patients resulted in a characterization of mutated targeted genes and methylated regions of *MGMT*, which is involved in the cancer evolution. Moreover, this project confirmed that results from ILM data and ONT do agree, giving the opportunity to use ONT for long read sequencing. This approach will reduce misalignment issues when repeats and pseudogenes are present and allows for the identification of point variants far from each other in the same chromosome.

In the second project, the use of two Hierarchical Dirichlet Clustering approaches allowed to identify groups of MM patients with similar CNV evolution between the diagnosis and the post-treatment relapse. The results confirmed the high CNV variability of MM and show that its progression cannot be simply explained by means of clinical parameters about the therapy carried out and patient's response.

Contents

List of Figures

List of Tables

List of Abbreviations

BA = Bland-Altman (plot)
BHDC = Bivariate Hierarchical Dirichlet Clustering
bp = base pair
BS = Bisulfite Sequencing
BWA-MEM = Burrows-Wheeler Aligner Maximal Exact Matches
CCC = Concordance Correlation Coefficient
CN = Copy Number
CND = Copy Number Diagnosis
CNR = Copy Number Relapse
CNV = Copy Number Variation
CpG = Cytosine-phosphate-Guanine (sites)
CR = Complete Remission
CRF = Chinese Restaurant Franchise
CRP = Chinese Restaurant Process
DD = Dirichlet Distribution
dsDNA = double-stranded DNA
GBM = Glioblastoma
GRCh38/hg38 = Genome Reference Consortium Human Build 38
H3-3A = Histone H3.3
HDP = Hierarchical Dirichlet Process
IDH = Isocitrate dehydrogenase
ILM = Illumina
MGMT = O-6-Methylguanine-DNA methyltransferase
MHDC = Multivariate Hierarchical Dirichlet Clustering
MM = Multiple Myeloma
NGS = Next Generation Sequencing
ONT = Oxford Nanopore Technology
PCR = Polymerase Chain Reaction
PN = PolyNucleotide
SNP = Single Nucleotide Polymorphism

ssDNA = single-stranded DNA

TERT = Telomere reverse transcriptase

UHDC = Univariate Hierarchical Dirichlet Clustering

VAF = Variant Allele Frequency

VGPR = Very Good Partial Remission

Introduction

Cancer ranks among the leading causes of death in almost every country in the world. The World Health Organization (WHO) estimated 19.3 million new cases and 10 million cancer deaths all over the world in 2020 [1]. Only in Italy, in 2020, $377,000$ new cases of tumor were diagnosed, of which $195,000$ among men and $182,000$ among women [2].

The onset of a cancer is often accompanied by a set of alterations at a chromosomal, genetic and epigenetic level. These mutations, which take place in somatic cells, can result in the activation or inhibition of a series of biochemical processes, controlled by the mutated genes. As a consequence, the cells can become malignant and induce the onset of a tumor.

The knowledge of the genetic alterations related to the development of a certain cancer, as well as the the discovery and study of new mutations or biomarkers have a crucial impact on tumor diagnosis. Moreover, this kind of information also has a prognostic power, allowing for the design of targeted and effective treatments.

In the last 20 years, the development and the rapid spread of Next Generation Sequencing (NGS) technologies, allowed to sequence an entire human genome in a short time and relatively low cost. This so-called high-throughput and massive-parallel sequencing gave a strong impetus to whole genome studies. On the other hand, this huge amount of data required suitable tools to be processed and analysed. Statistical methods have proved to be powerful tools in fulfilling this task. Statistical methods applied to biological issue led to the development of the bioinformatics, which has greatly increased our knowledge about human genome.

The present thesis work is composed by two different projects, developed in the last year: the first one consists in a study of a cohort of patients diagnosed with Glioblastoma (GBM), a brain tumor, using two different NGS techniques, that are Illumina sequencing-by-synthesis and nanopore sequencing. This analysis will also help in comparing the results from the two techniques, exploiting different statistical methods. The second project is a study of a group of patients affected by Multiple Myeloma (MM), a blood cancer, characterized by sets of Copy Number Variation measurements. This study focuses on the implementation of the Dirichlet Process, that is a statistical learning method with many properties, that can be used to perform clustering to find common evolutionary patterns among patients.

In chapter 1, a biological introduction is presented to the reader. The most basic concepts, such as DNA, RNA and the structure of human genes are provided. Then, a description of the most common genetic and epigenetic modifications is presented. There is an introduction of DNA sequencing, a process that allows to read the order of the nucleotides (A, T, C and G) which form the DNA of an organism. The Illumina sequencing by synthesis technology, considered the gold standard in NGS, is presented, together with its pros, such as the high quality in identifying the nucleotide bases (lower than 1 error each 1000 reads), and cons, including short read length ($150-200$ bp) and the need for chemical treatment to perform methylation analysis. Finally, a revolutionary approach to DNA sequencing is introduced: the nanopore sequencing technique. It basically consists in passing the DNA strand through a biological (or synthetic) pore, across which a ion current is created. The nucleotides that cross the pore induce a current perturbation, that is characteristic of each nitrogen base. So, from the current perturbation it is possible to recognize which base passed through the pore and then, perform DNA sequencing. This method, provided by Oxford Nanopore Technologies (ONT), is cheaper (it does not require chemical agents) and can sequence longer strands. On the other hand, the quality is lower (about 1 error each 100 reads).

In chapter 2, after an introduction to the Bayesian interpretation of probability and a description of some important probability distributions, the theory of Dirichlet Processes is exposed. The Dirichlet Process is a family of stochastic processes whose realizations are probability distribution. It is characterized by a base distribution and a concentration parameter, which plays the same role of the mean and the variance (the inverse of it) of a probability distribution, respectively. The Dirichlet Process draws probability distributions around the base distribution as a probability distribution draws numbers around its mean. An important characteristic of the Dirichlet Process is its clustering property. It also has an extension, called the Hierarchical Dirichlet Process. It basically consists in a Dirichlet Process from which the base distributions of several Dirichlet Processes are sampled. Therefore, it is possible to perform clustering on grouped data and to find clusters that are shared among the groups. Moreover, Dirichlet Process allows for unsupervised clustering: it means that there is no need to set in advance the number of clusters that have to be found because the algorithm will learn it directly from the data.

In chapter 3, the cohort of Glioblastoma patients is presented, together with the Illumina MiSeq System and the ONT MinIon Mk1c used for sequencing. Then, the preparation of the sample is briefly described. Some useful bioinformatics functions and tools are listed and the exploited pipeline is reported. Then, the cohort of Multiple Myeloma patients is introduced. Copy Number Variation measurements are taken at diagnosis and in relapse, and are available for each subject. The *R* language implementation of the Hierarchical Dirichlet Process using a Multivariate mixture model created by Giovanni Sighinolfi, a former student of the University of Bologna, is reported. Finally, some statistical tools exploited for the analyses are described.

In chapter 4, the results of the two projects are shown. A variant analysis on GBM patients is conducted on four target genes, using both Illumina and nanopore, to observe already known point alterations in patients, in order to characterize them. Methylation analysis is performed, studying different regions of *MGMT* gene. Their methylation percentages, obtained from both Illumina and ONT, were compared to assess whether the latter, that has a lower quality score, can be considered reliable.

Two types of Dirichlet Clustering are performed on the cohort of MM patients, in order to find sets of genes which evolve in a common way among them. This kind of information could be helpful in determining how the cancer evolves in patients and in assessing the effect of the treatments these subjects underwent. Moreover, it would be interesting to find common evolution patterns (i. e. trajectories) between patients that could be related to clinical parameters. This could finally help clinicians in defining targeted therapies.

Chapter 1

Genome and Sequencing

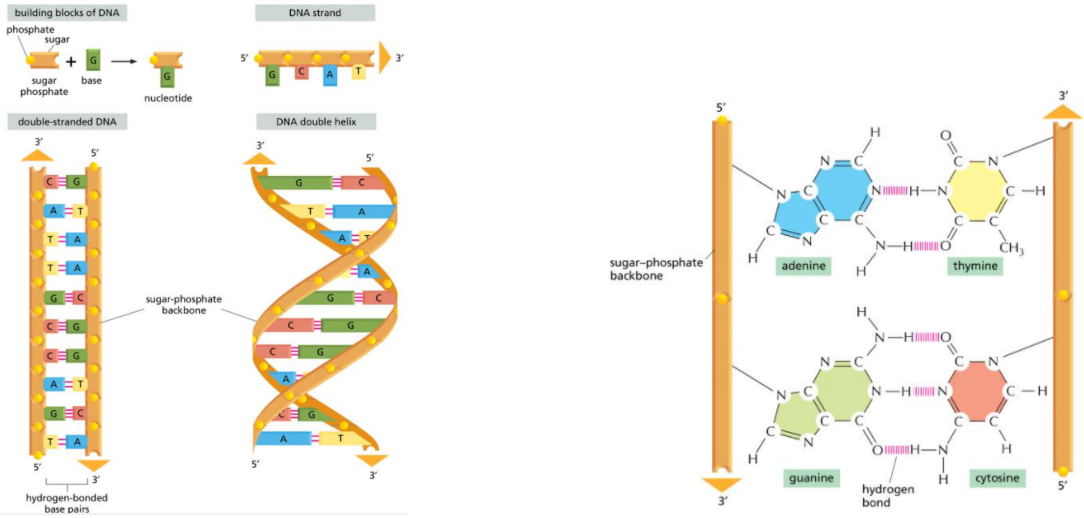
In this chapter, the basic properties and structure of the nucleic acids, DNA and RNA, and their differences, in terms of chemical composition and biological functions are briefly introduced. Then, the important notion of epigenome, the methylation process and its consequences on the genetic expression are described. The concept of DNA variant, with some examples, are also presented. Finally, a brief overview about the 'classical' DNA sequencing platforms is given and the revolutionary nanopore sequencing technology is introduced.

1.1 Nucleic Acids and Genome

1.1.1 DNA

The DeoxyriboNucleic Acid (DNA) is a nucleic acid that contains the genetic information necessary to the proteins biosynthesis, which are fundamental for the proper functioning of many living organisms [3].

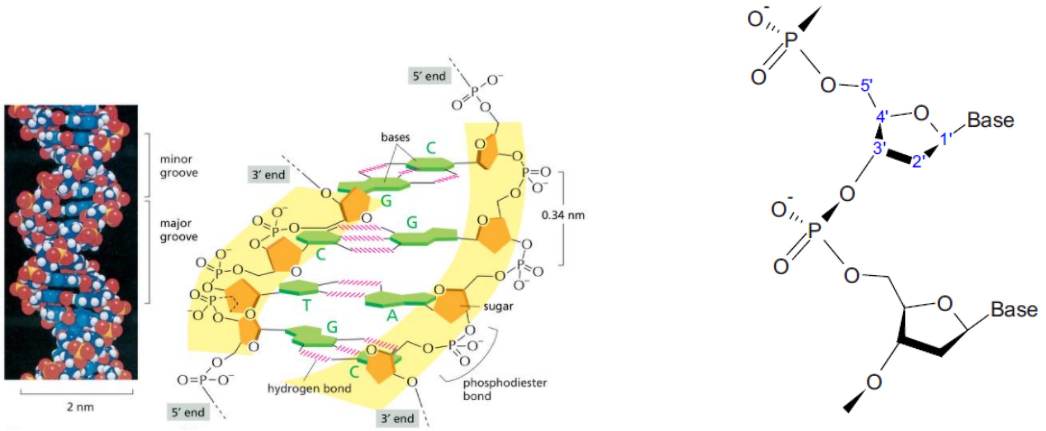
The DNA molecule is made of two long polynucleotidic chains, named DNA chains or DNA strands, each composed by four types of nucleotidic subunits. The two chains are hold together due to hydrogen bonds between the base portions of the nucleotides, as shown in Fig. 1.1a. Each nucleotide, also called monomer, is composed by a base, which contains Nitrogen, and a five Carbon atoms sugar, to which one or more phosphate groups (PO_4) are linked. So, the DNA is a polynucleotide. For the DNA nucleotides, the sugar is the deoxyribose (from which the name DNA), while the base can be an Adenine (A), a Cytosine (C), a Guanine (G) or a Thymine (T), Fig. 1.1b. A and G consist of a six- and a five-membered ring with a common edge, and are called *purine* bases. C and T have just a six-membered ring and they are called *pyrimidine* bases. However, for many physical properties, the distinction into purines and pyrimidines is not helpful, though purines are larger. The nucleotides are covalently linked together in



(a) (b)

Figure 1.1: (a) DNA polynucleotidic strands and hydrogen bonds. (b) Chemical structure of the four DNA nytrogen bases. Figure from [3].

a chain through the sugars and phosphates, which thus form a “backbone” of alternating sugar-phosphate- sugar-phosphate. Each DNA strand has its own chemical polarity, due to nucleotide subunits linking, as shown in Fig. 1.2a. Each sugar has a 5′ phosphate and a 3′ hydroxyl, Fig. 1.2b. Nucleotides are linked together covalently joining 3′-hydroxyl group of one sugar to the 5′-phosphate group of the next one, so that the two ends of the chain will be easily distinguishable. This polarity in a DNA chain is indicated by referring to one end as the 3′ end and the other as the 5′ end. The orientation will be important for the structural dynamics of polynucleotides translocating through a nanopore [4].



(a) (b)

Figure 1.2: (a) A short section of the double helix. Four base pairs are present. The polarities of the two DNA strands are shown. Figure from [3]. (b) Atomistic structure of the nucleic acid backbone. Labels are highlighted. Figure from [4].

The two chains are held together by hydrogen bonds between the bases of the different strands: all the bases are on the inside the double-helix, while sugar-phosphate backbones are on the outside. The formation of the double-stranded (ds) DNA helix occurs by the so-called Watson-Crick base pairing, where A pairs with T and G pairs with C. This complementary base-pairing enables the base-pairs to be packed in the energetically most favourable arrangement, so that each pair is of similar width, holding the backbones an equal distance apart along the DNA molecule. The two backbones wind around each other to form a double helix, with one complete turn every ten base pairs. The members of each pair can fit together only if the two strands are antiparallel (i.e. the polarity of one is oriented opposite to the one of the other strand). As a consequence, each strand contains a sequence of nucleotides that is exactly complementary to the nucleotide sequence of the other strand [3].

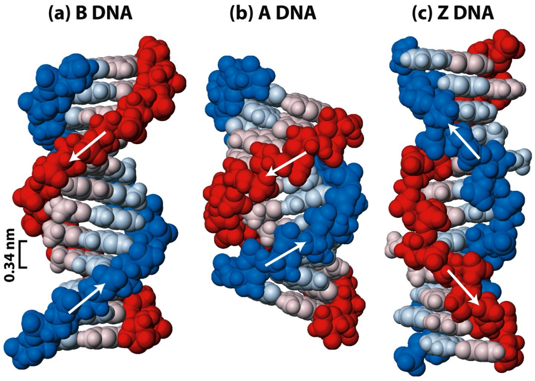


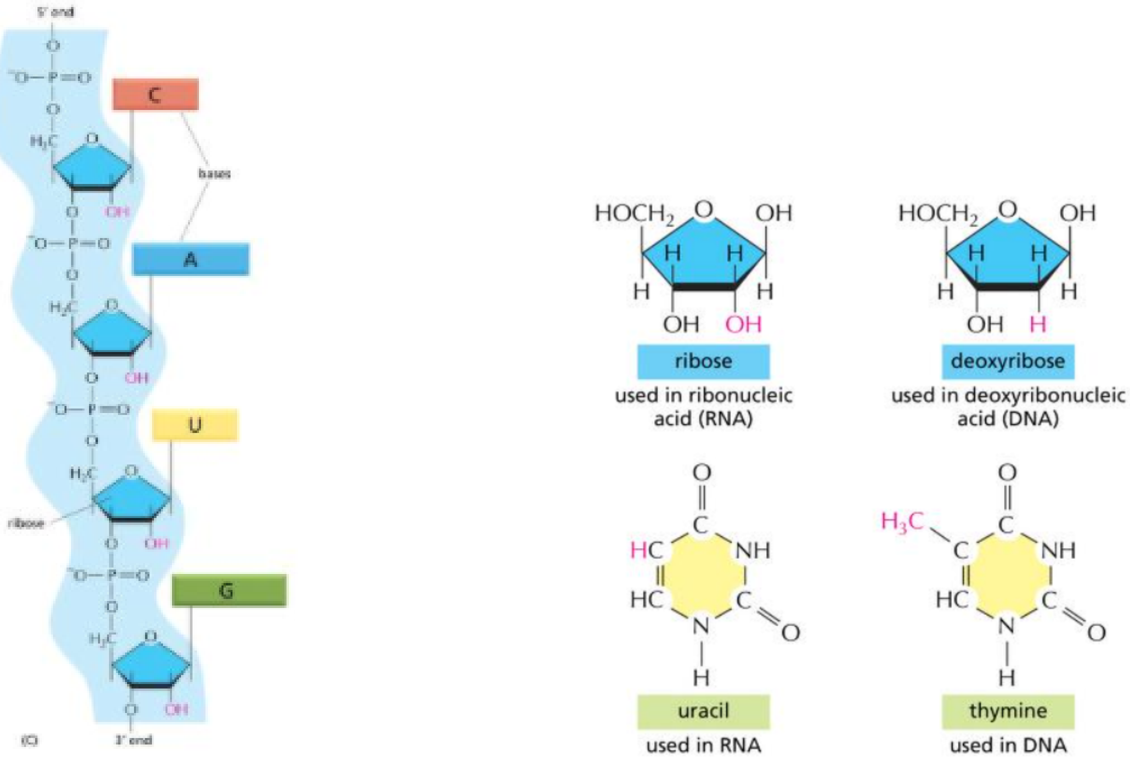
Figure 4-4
Molecular Cell Biology, Sixth Edition
© 2008 W.H. Freeman and Company

Figure 1.3: Secondary structures of DNA. (a) B-DNA; (b) A-DNA; (c) Z-DNA. Figure from [3].

There is also a global property called secondary structure. Several different types of helices do exist and the most common are called A-DNA, B-DNA and Z-DNA, Fig. 1.3. The differences between these two global structures are due to the ionic and water environment. B-DNA is favourite in aqueous environment as water molecules can bind in the channels along the helix. B-DNA has a diameter of $\sim 2\ nm$, a distance between base-pairs of $\sim 0.34\ nm$ and an angle of about 36° between them [4]. A-DNA is favorite in dehydrating conditions. It is right-handed helix as B-DNA, but the structure is more compact, with lower base-pairs distance ($\sim 0.26\ nm$) and angle between them ($\sim 33°$). Z-DNA is quite unusual and it is typical of DNA methylated regions, where the strands wrap themselves in the opposite direction around the axis (left-handed helix). Here, the base-pairs distance is about $0.37\ nm$, the largest. There exists a process called denaturation in which the two strands in the double-stranded (ds) DNA unbind into single-stranded (ss) DNA molecules.

1.1.2 RNA

The RiboNucleic Acid (RNA) is a linear polymer, like DNA, made of four different types of nucleotide subunits linked together by phosphodiester bonds, Fig. 1.4a. From a chemical point of view, it differs from DNA basically in two aspects, depicted in Fig. 1.4b: the nucleotides in RNA are ribonucleotides, i.e. they contain ribose sugar (from which the name RNA), and it contains the base Uracil (U) instead of the Thymine (T), while A, C and G are present. U can base-pair through hydrogen bonds with A and the complementary base property described for DNA does hold for RNA too [3].

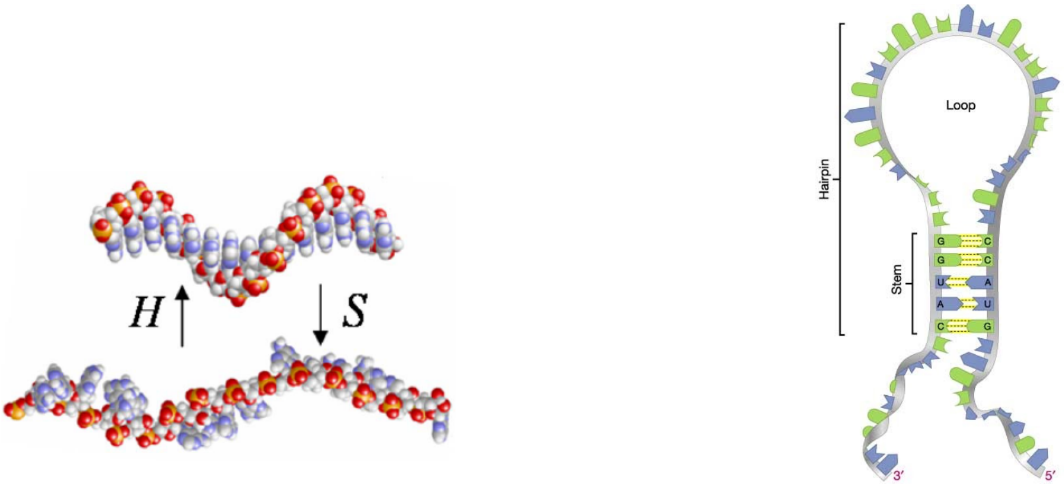


(a) (b)

Figure 1.4: (a) RNA polynucleotidic single strand and hydrogen bonds. (b) Chemical differences between Ribose and Deoxyribose and between Uracil and Thymine. Figure from [3].

Although DNA and RNA are similar from a chemical point of view, they differ in the overall structure and functions. RNA is always single-stranded, while DNA occurs as a double-strand helix. Moreover, RNA can fold up into a particular shape, that gives it precise structural and catalytic functions.

RNA can show secondary structure as well, which mainly depends on ionic environment and temperature. A schematic of the secondary structure of RNA is sketched in Fig. 1.5a. In general, secondary structure can be thought as a result of the competition between entropic and enthalpic factors. The interaction energy of base stacking



(a)

(b)

Figure 1.5: (a) Schematic of the secondary structure of RNA. On the top, single-stranded helix due to enthalpic effects; entropic effects transform it into a random coil polymer, on the bottom. (b) RNA hairpin structure. Figure from [4].

favourites helical secondary structure whereas entropic factors (e. g. temperature) bring the strand to a random coil form [4]. Another type of secondary structure is a hairpin, depicted in Fig. 1.5b. It consists in a single strand (of DNA or RNA) that wraps around to form a double strand. The double-stranded portion is called stem, while there is also a loop of unpaired bases, which causes an unfavourable strain on their formation. Sometimes, a single-stranded portion can occur in the chain as well [4].

Polynucleotides are charged in solution, and this is one of the most significant properties. In several ionic conditions, the PN's backbone is characterized by a single negative charge for each nucleotide unit. On the other hand, in solutions there would be counterions (e.g. Na^+, K^+) that can partially neutralize this charge. However, the fact that the PN's strand is charged is crucial, since it is the basis of nanopore sequencing. In fact, it allows the DNA to pass through a nanopore with the use of an electric field [4].

1.1.3 The structure of a gene

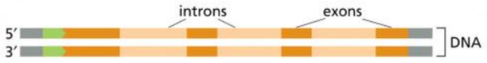


Figure 1.6: Structure of an eukaryotic gene. Exons and introns are highlighted. The green portion is the promoter. Figure from [3].

In the cell nucleus the DNA is packed into chromosomes, which are made of hundreds or even thousands of genes. A gene is a region of DNA which codes for a single RNA or protein. The entire set of genes is called genome. Eukaryotic genes are made of coding sequences (exons), non-coding regions (introns) and a promoter, which controls the gene expression Fig. 1.6. Genes are transcribed into RNA and, through an operation called splicing, they produce the messenger RNA (mRNA), which controls the protein synthesis [5].

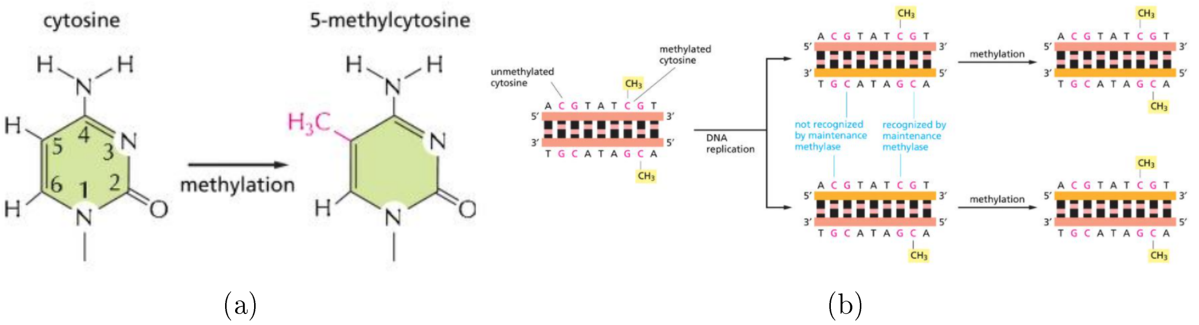
Each gene has its own specific location within a chromosome, called locus. The particular form of the gene is termed allele. Mammalian DNA is characterized by the fact that each gene is present in two allelic forms, that could be identical (homozygous) or may vary (heterozygous). The occurrence of different alleles in the same site of the genome is called polymorphism, [5].

1.2 Epigenome

The epigenetic is defined as the set of the inheritable modifications which cause a genetic expression modulation, but that do not involve a change in the DNA sequence. Epigenetic mechanisms allow the cell to rapidly act in response to environmental changes, determining the genetic expression levels. Unlike the genome, which is mainly static, the epigenome is much more dynamic and can respond to environmental changes, in the diet, to pollution, and vary according to the lifestyle, smoking and alcohol consumption habits, stress, radioactive and chemical exposure [5]. Therefore, epigenetic can be related to pathologic conditions, since epigenetic alterations can influence time and level of expression of some genes. For instance, a gene can be forced to be activated when it is expected to be deactivated, and vice versa. Epigenetic researchers focus on the study of histonic modifications and DNA methylation. The latter will be treated in this chapter.

1.2.1 DNA methylation

The DNA methylation is an epigenetic process that consists in the addition of a methyl group ($-CH_3$) to a nucleotide. The most common forms of methylated nucleotides are 5-methyl-cytosine (5-mC), Fig. 1.7a, by far the most widespread, 5-hydroxy-methyl-cytosine (5-hmC) and 6-methyl-adenine (6-mA). This modification has no effect on base-pairing [3]. This cytosine modification usually deactivates genes, as it attracts proteins that link methylated cytosines and block gene expression. In vertebrates, the DNA methylation is mainly restricted to cytosine nucleotides in the sequence CpG (Cytosine-phosphate-Guanine), which is base-paired to exactly the same sequence on the other strand (opposite orientation). Several studies highlighted that about the $60-80\%$ of all the CpG sites are methylated. CpG sites represent the 1% of the entire genome and, usually, are grouped into the so-called CpG islands, which are genomic regions



(a) (b)

Figure 1.7: (a) Formation of a 5-methyl-cytosine by methylation of a cytosine base in the DNA double helix. (b) Schematic of how DNA methylation patterns are inherited in humans. Figure from [3].

made of about 200 base-pairs (bp), with a CG percentage ≥ 50% [5]. The existing DNA methylated patterns can be inherited by the daughter DNA strands through a mechanism schematized in Fig. 1.7b. The maintenance methyltransferase enzyme acts on those CpG sequences that are paired with an already methylated CpG sequence. In this way, methylation patterns of parental strand is used as a template for methylation in daughter strand [3].

DNA methylation has a role in many cellular and epigenetic processes, including embryonic development, modification of chromatin structure, inactivation of the X-chromosome, genomic imprinting, maintenance of chromosome stability and carcinogenesis [6]. Moreover, DNA methylation patterns are dynamic during vertebrate development and can change in response to developmental or environmental clues. Shortly after fertilization, the great majority of methyl groups are lost from the DNA due to a genome-wide wave of demethylation. Later in development, new methylation patterns are created by several *de novo* DNA methyltransferases that are directed to DNA by sequence-specific DNA-binding proteins where they modify adjacent unmethylated CG nucleotides [3]. Some researchers call 5-mC "a dynamic fifth letter of the DNA code".

So, DNA methylation patterns can provide a significative amount of information due to the numerous processes where it is involved in. However, one of the main problems is that most DNA sequencing technologies cannot directly distinguish between methylated and unmethylated bases in DNA. The most used approaches for identifying 5-mC require bisulfite treatment (explained in par. 1.4.2), that must first be used to convert unmethylated cytosines to uracil. As a consequence, this pre-processing step increases the complexity of library preparation and the potential for artifacts and biases from sources, such as incomplete chemical conversion [6]. Moreover, Illumina-based sequencing techniques suffer from short read lengths, making some regions difficult to map and limiting the study of allele-specific methylation.

1.3 Genetic mutations

The term genetic mutation, or variant, refers to any modification in the nucleotidic sequence of a genome. This modification must be stable and inheritable. Usually, a genetic mutation is due to the effect of an external agent or by chance. Consequently, the genotype of an individual results to be modified, and possibly the phenotype undergoes modifications.

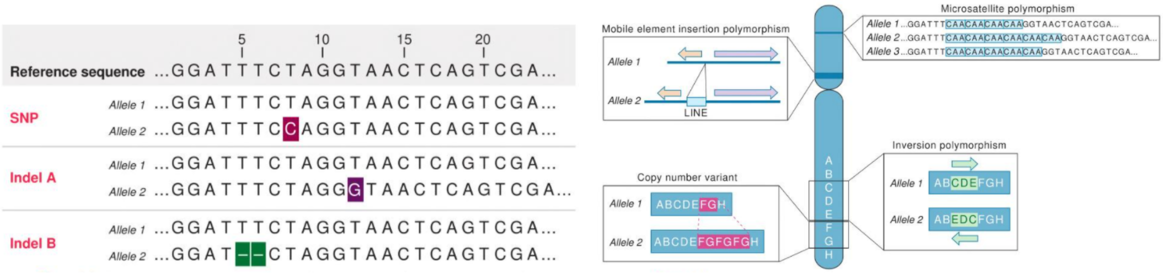
The genome of an individual is characterized by a huge number of modifications and alterations. Here the focus will be on two main groups: Single-Nucleotide Polymorphism (SNP, pronounced 'snip'), which refers to the substitution of a single nucleotide at a specific position, and Copy-Number Variation (CNV), that refers to a relatively large portion of the genome.

1.3.1 Single-Nucleotide Polymorphism

The great majority of the mutations present in the human DNA is made of substitutions of a single base in a precise location within the genome. They are point mutations, i.e. the sequences of two different persons differ from each other for a nucleotidic pair (where one has an A-T pair, the other has C-G, for instance, see Fig. 1.8a). Conventionally, a point mutation must be present in at least 1% of the total population to be considered a SNP. The significance for health of the vast majority of SNPs is unknown and is the subject of ongoing research [7]. Many of these point mutations take place in non-coding regions so that they do not alter the amino acid sequence. It has been estimated that two human genomes randomly chosen within the world population would be different in about 2.5×10^{16} SNP. These polymorphisms are interesting because they may account for the differences between individuals at the level of disease susceptibility, drug metabolism and response to environmental factors [5].

1.3.2 Copy Number Variation

The Copy Number (CN) is a quantity (a number) that basically identifies how many times a certain portion of the DNA is repeated. It consists of variation in the number of copies of larger segments of the genome, ranging in size from 1000 bp to many hundreds of kilobase pairs. A CN refers to a region of the genome, that can vary from a single gene to dozens of them, and thus CNVs are frequently implicated in traits that involve altered gene dosage [7]. If the CN is equal to 2, it means that the mutation is absent (individuals have two copies of DNA portions, each in the chromosome couple). Genetic mutations can be divided into duplication/amplification and deletion. The duplication occurs when a chromosome has two copies of a section of DNA. Typically, a CN larger than 2 indicates the presence of an amplification. The largest CNVs are sometimes found in regions of the genome characterized by repeated blocks of homologous sequences called segmental duplications. The deletion consists in the elision of a portion of genome from a chromosome. A CN lower than 2 means that a deletion of a region of DNA is present. An example of Copy Number Variant is shown in Fig. 1.8b.


5
10
15
20
Reference sequence ...GGATTTCTAGGTAACTCAGTCGA...
SNP
Allele 1 ...GGATTTCTAGGTAACTCAGTCGA...
Allele 2 ...GGATTTCCAGGTAACTCAGTCGA...
Indel A
Allele 1 ...GGATTTCTAGGTAACTCAGTCGA...
Allele 2 ...GGATTTCTAGGGTAACTCAGTCGA...
Indel B
Allele 1 ...GGATTTCTAGGTAACTCAGTCGA...
Allele 2 ...GGAT--CTAGGTAACTCAGTCGA...
Mobile element insertion polymorphism
Allele 1
Allele 2
LINE
Microsatellite polymorphism
Allele 1 ...GGATTTCAACAACAACAAGGTAACTCAGTCGA...
Allele 2 ...GGATTTCAACAACAACAACAACAAGGTAACTCAGTCGA...
Allele 3 ...GGATTTCAACAACAACAACAAGGTAACTCAGTCGA...
Copy number variant
Allele 1 ABCDEFGH
Allele 2 ABCDEFGFGFGH
A B C D E F G H
Inversion polymorphism
Allele 1 ABCDEFGH
Allele 2 ABEDCFGH

(a) (b)

Figure 1.8: (a) Examples of point mutations: Single Nucleotide Polymorphism, insertion and deletion (indels). (b) Example of Copy Number Variation (bottom left). Figures from [7].

Some of the genetic variations are very common, while others are present in only a minority of human individuals. It is known that several structural variants are not associated to a condition of disease. They are called neutral mutations. However, many genome variations are responsible for genetic diseases. It is of crucial importance to be able to recognize those variations which affect human health from the ones that result to be 'harmless'.

1.4 DNA sequencing techniques

Considering that the DNA is a map containing the whole genetic information of an organism, its main property is that this information is encoded into the nucleotide sequence. Being able to read this sequence is of fundamental importance for many applications, from biology to medicine. In the end of $'70s$, the technology needed for DNA sequencing became available and, in the last decades, it has been developed in a fast way, paving the way for the so-called Next Generation Sequencing (NGS) technologies. Compared with the long-established biochemical methodologies, the sequencing and drafting of genomes is constantly evolving as new bioinformatics tools become available. In the present paragraph an overview about the first generation techniques will be given, followed by a description of the main second generation platforms.

1.4.1 First generation techniques

The first DNA sequencing technique was the Sanger method (or dideoxy sequencing), first developed by Frederick Sanger ($1918-2013$) and colleagues in 1977. It was the most widely used sequencing technique for many years, before the spread of the NGS technologies, which allow large-scale and automated genome analyses. However, Sanger method is currently used for smaller-scale applications. It relies on two important techniques: the Polymerase Chain Reaction and the Gel electrophoresis.

Polymerase Chain Reaction

The Polymerase Chain Reaction (PCR) is a technique used to amplify a DNA fragment in a fast way, starting from a complex mixture of starting material, called template DNA. To begin, a pair of DNA oligonucleotides, chosen to flank the desired nucleotide sequence (called target DNA) of the gene, are chemically synthesized. Each oligonucleotide, called primer, is complementary to a stretch of DNA to the $3'$ side of the target DNA (one oligonucleotide for each of the two strands). They are then used to prime DNA synthesis on single strands generated by heating the DNA from the entire genome [3].

The PCR consists of three steps, defined in time and temperature: (i) denaturation, (ii) annealing, (iii) extension, Fig. 1.9. At each cycle, the amount of DNA synthesized in the previous one is doubled. A heat treatment to above 90° is required at each iteration to separate the two template strands (denaturation). The system is then cooled to a temperature between 40° and 60°. In the annealing step, the hybridisation of the two primers, which bind their complementary sites in target DNA, takes place. The extension is carried out by a DNA polymerase. This polymerase is extracted from a thermophilic bacterium, as not to be denaturated during the cycles. The DNA synthesis proceeds from both primers. Finally, the new strands contain a region that is complementary to the other primer, and they can be used as templates [5]. In practice, effective DNA amplification requires $20-30$ reaction cycles, with the products of each cycle serving as the DNA templates for the next, giving rise to an exponential increase (chain reaction). A single cycle requires about 5 minutes.

This technique is now used routinely to clone DNA from genes of interest directly starting either from genomic DNA or from mRNA isolated from cells. One of the main properties of the PCR method is its extreme sensitivity: it can detect a single DNA molecule in a sample.

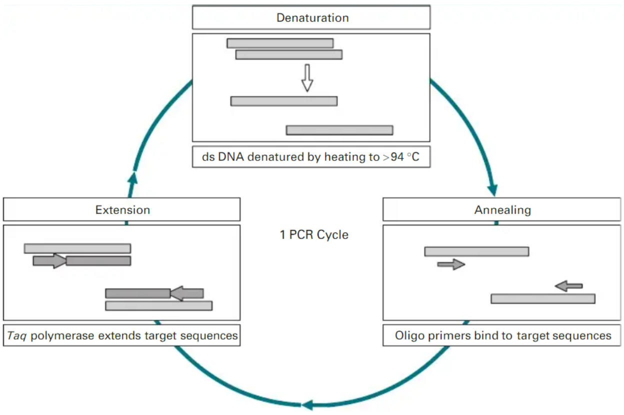

Denaturation
ds DNA denatured by heating to >94 °C
Extension
1 PCR Cycle
Annealing
Taq polymerase extends target sequences
Oligo primers bind to target sequences

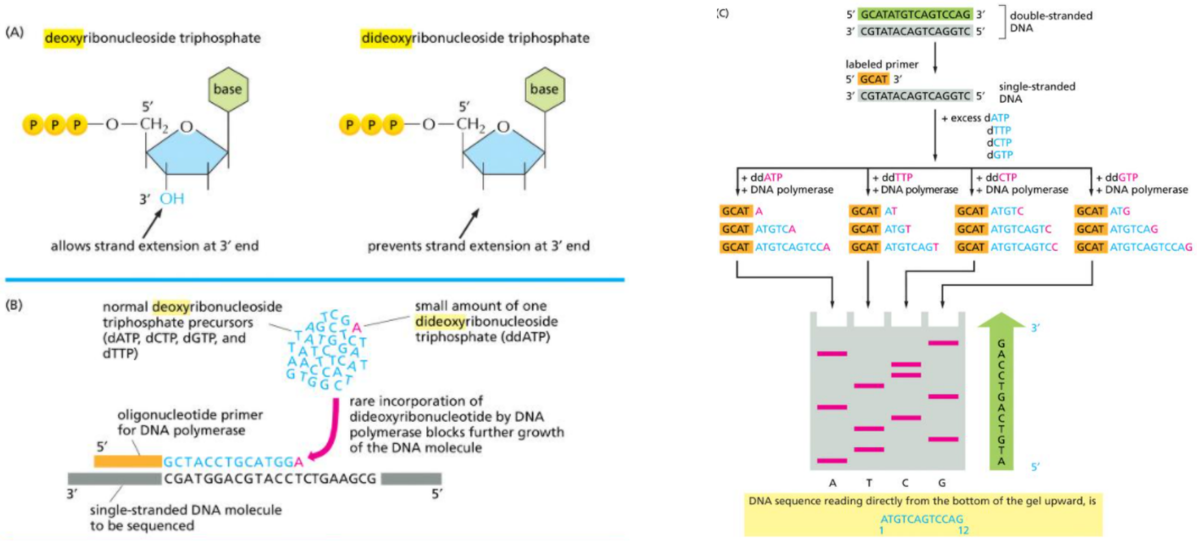
Figure 1.9: A schematic of a PCR reaction cycle, with the three steps. Figure from [5].

Gel electrophoresis

Gel electrophoresis is a method for the separation and analysis of macromolecules, such as DNA and RNA, and their fragments, based on their size and charge [3]. By means of an electric field, molecules (such as DNA) can be made to move through a gel made of polymers (usually agarose or polyacrylamide). When the electric field is applied, the larger molecules move more slowly through the gel, while the smaller ones move faster. Molecules with different size form distinct bands on the gel. If several samples have been loaded into adjacent wells in the gel, they will run parallel in individual lanes. Bands that end up at the same distance from the top in different lanes contain molecules that passed through the gel at the same speed, which usually means they are approximately the same size.

Sanger sequencing

The dideoxy sequencing method requires some agents: a single-stranded DNA template to sequence, a DNA primer, a DNA polymerase, normal deoxynucleotide triphosphates (dNTPs), and modified di-deoxynucleotide triphosphates (ddNTPs, that lack the $3'$ hydroxyl group), Fig. 1.10a, the latter of which terminate DNA strand elongation. The DNA sample is divided into four separate PCRs, all of which contain the four nucleotides (dATP, dGTP, dCTP and dTTP, collectively called dNTPs) and the DNA polymerase. Only one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP) is added to each reaction, while the other added nucleotides are ordinary ones, Fig. 1.10a. The ddNTPs terminate elongation of the DNA strand. Then, the DNA fragments resulting from PCR are thermally denaturated and separated by size, using gel electrophoresis in four different lanes, one for each ddNTP, Fig. 1.10b. The relative positions of the different bands among the four lanes, consecutively from bottom to top, are then used


(A)
deoxyribonucleoside triphosphate
dideoxyribonucleoside triphosphate
base
P P P —O—CH2 O
5′
3′ OH
allows strand extension at 3′ end
prevents strand extension at 3′ end
(B)
normal deoxyribonucleoside triphosphate precursors (dATP, dCTP, dGTP, and dTTP)
small amount of one dideoxyribonucleoside triphosphate (ddATP)
rare incorporation of dideoxyribonucleotide by DNA polymerase blocks further growth of the DNA molecule
oligonucleotide primer for DNA polymerase
5′ GCTACCTGCATGGA
3′ CGATGGACGTACCTCTGAAGCG 5′
single-stranded DNA molecule to be sequenced
(C)
5′ GCATATGTCAGTCCAG 3′
3′ CGTATACAGTCAGGTC 5′
double-stranded DNA
labeled primer
5′ GCAT 3′
3′ CGTATACAGTCAGGTC 5′
single-stranded DNA
+ excess dATP dTTP dCTP dGTP
+ ddATP + DNA polymerase
+ ddTTP + DNA polymerase
+ ddCTP + DNA polymerase
+ ddGTP + DNA polymerase
GCAT A
GCAT ATGTCA
GCAT ATGTCAGTCCA
GCAT AT
GCAT ATGT
GCAT ATGTCAGT
GCAT ATGTC
GCAT ATGTCAGTC
GCAT ATGTCAGTCC
GCAT ATG
GCAT ATGTCAG
GCAT ATGTCAGTCCAG
3′
GACCTGACTGTA
5′
A T C G
DNA sequence reading directly from the bottom of the gel upward, is
ATGTCAGTCCAG
1 12

(a) (b)

Figure 1.10: Schematic of the Sanger sequencing method pipeline. (a) Structure of dNTPs and ddNTPs; (b) PCRs and gel electrophoresis. Figures from [3].

to read the DNA sequence. The introduction of ddNTPs with fluorescence labels allows to perform automated high-throughput sequencing, but gel electrophoresis needed to be replaced by capillary electrophoresis [3].

However, being the Sanger sequencing method slow, laborious and expensive, due to the numerous reactants needed for the sample preparation, NGS, second-generation or massively parallel sequencing technologies, were developed.

1.4.2 Second generation techniques

Many second generation sequencing techniques have been developed in the last years, inspired by Sanger sequencing but exploiting different pipelines. In contrast to sequencing a single DNA fragment, the DNA to be sequenced is typically prepared in the form of random fragments, with defined oligonucleotide sequences at either end that are captured on a flow cell (see Appendix A) by complementary oligonucleotides bonded to the surface. Only a small subset of these techniques will be described in this paragraph. They are called short-read sequencing methods.

Pyrosequencing

It is a technique first described in 1998 that allows parallel sequencing. It is based on the "sequencing by synthesis" principle, in which the sequencing is performed by detecting the nucleotide incorporated by a DNA polymerase. Essentially, the method allows sequencing a single strand of DNA by synthesizing the complementary strand

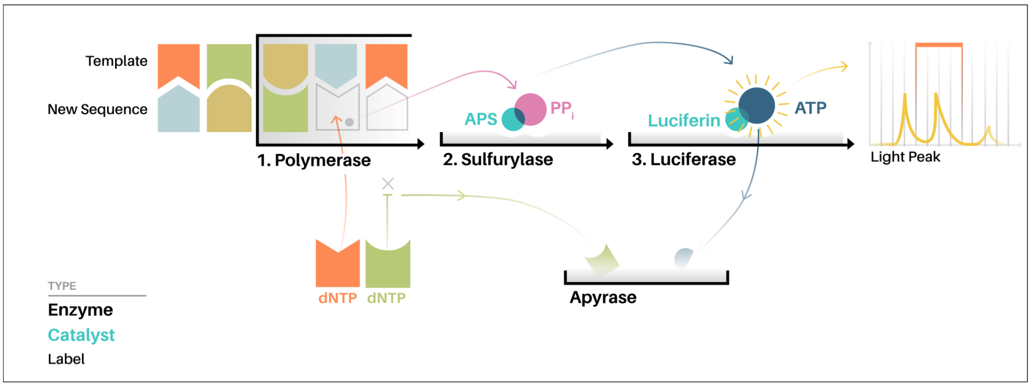


Figure 1.11: Schematic of a pyrosequencing pipeline. Figure from `https://commons.wikimedia.org/wiki/File:How_Pyrosequencing_Works.svg`.

along it, one base pair at a time, and detecting which base was actually added at each step, Fig. 1.11. Several enzymes must be used. The DNA polymerase is responsible for the dNTP incorporation onto the complementary template basis.

For each incorporation, a quantity of pyrophosphate (PPi), proportional to dNTP embedded, is released. Then, PPi is converted into ATP by ATP sulfurylase. ATP and luciferase convert luciferin into oxyluciferin, which generates visible light, proportional to the ATP quantity. So, the light signal is proportional to the number of incorporated nucleotides. Light is detected by a CCD camera and can be viewed as a peak whose heigth is proportional to the number of incorporated nucleotides [8].

Although pyrosequencing is simple to make, it has some limitations. For instance, it can read only sequences with less than 100 bases, which makes it useful for Single Nucleotide Polymorphism (SNP) identification but unfeasible for whole-genome sequencing.

The first commercial NGS platform was Roche 454 sequencing system. Its principle of sequencing was pyrosequencing technology. By incorporating an array-based pyrosequencing technology, it allowed to sequence $400-600$ megabases in a 10 hour period, allowing an entire human genome to be sequenced in ~ 27 days. Its main limitations were in the high cost, if compared to other technologies, and the high error rate in sequencing repeated regions [9].

Illumina dye sequencing

This method basically works in three steps: amplification, sequencing and analysis. Firstly, the DNA is fragmented into small (~ 150 bp) segments. They are bonded with two adapters, one on the left and the other on the right and put into a flow cell. The flow cell is a glass slide with lanes, each one is a channel coated with a lawn composed of two types of oligonucleotides, which serve as anchoring points for the DNA strands. The

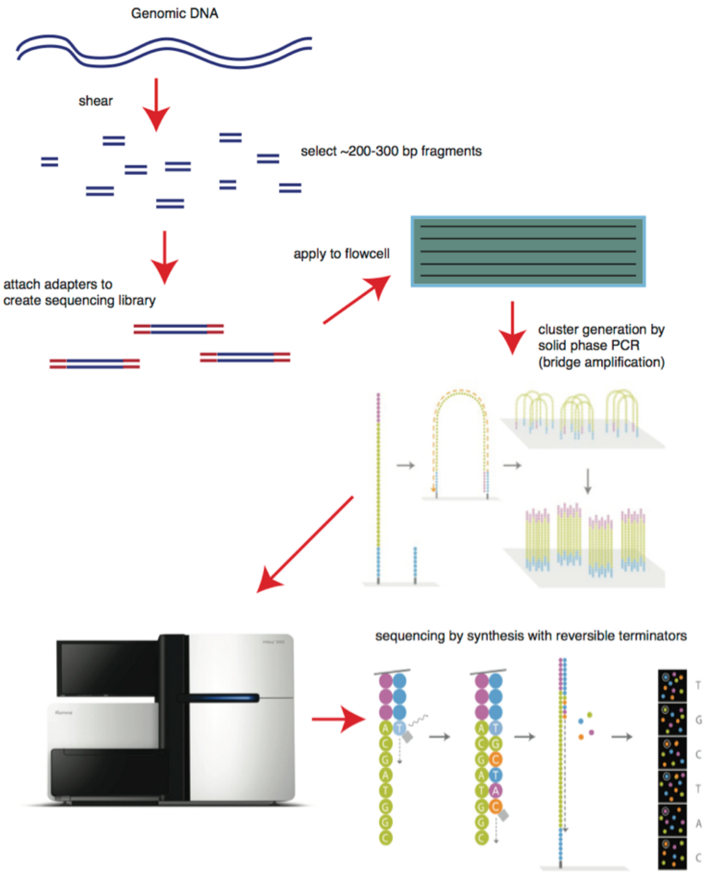


Figure 1.12: Schematic of an Illumina dye sequencing pipeline. Top: sample preparation by shearing DNA in fragments, binding adapters, and applying DNA on the flow cell. Middle: DNA anchoring and bridge amplification. Bottom: Sequencing by synthesis procedure. Figure from https://bitesizebio.com/13546/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology/.

DNA adapter binds complementarily to one in the flow cell and a polymerase creates a complement of the hybridized fragment. The double-stranded molecule is denatured and the original template is washed away.

Then, the strands are amplified through the so-called PCR bridge amplification: the strand folds over and the adapter region hybridizes to the second type of oligos on the flow cell. Polymerases generate the complementary strand. The double-stranded bridge is then denatured, resulting into two single strand copies of the molecule. This process is repeated many times and takes place simultaneously for millions of clusters, Fig. 1.12. The reverse strands are removed. Primers and modified (fluorescent marked) nucleotides are washed onto the flow cell.

Sequencing step begins when primers attach to strands and then, at each cycle, nucleotides compete to be added to the growing chain and only one at a time is incorporated. Clusters are excited by a light source and a fluorescent signal, characteristic for each base, is emitted and recorded by a CCD. This is a sequencing by synthesis process. The procedure is repeated for all the bases, so that millions of clusters are sequenced in a massively parallel process. The read product is washed away, and, with the same procedure, the first index of the DNA strand is read. Then, the template folds over, binds the second type oligo in the flow cell and the second index is read in the same manner. Polymerases extend the second oligo creating a double-stranded bridge, which is linearized, and the original forward strand washed away. The reverse strand read starts, in the same way of the forward strand read.

Finally, in the analysis step, sequences are separated according to the indexes introduced during the sample preparation, reads with similar stretches of base calls are locally clustered and forward and reverse reads are paired. The resulting contiguous sequences are then aligned and analysed.

Illumina dye-sequencing is characterized by a high quality (phred quality score ≥ 30, see Appendix A). One of the main limitations of this technique regards the read length: the DNA must be fragmented into small segments to be read. Moreover, the detection of methylated bases can be done only recurring to bisulfite treatment (see par. 1.4.2).

Bisulfite treatment

It is a method, developed in 1970, that can be used in order to allow the discrimination between methylated and unmethylated cytosines before routine sequencing. Treating DNA with Sodium Bisulfite induces the conversion of all the unmethylated cytosines in the nucleotide sequence into uracils. The method is based on the chemical property of bisulfite to bind to the double bond between 5 Carbon and 6 Carbon of all the unmethylated cytosines. The conversion occurs in three steps [8]:

- sulphonation: bisulfite gives up a sulfur group ($-SO_3H$) to Carbon 6 of cytosine, producing sulphonated cytosine;

- hydrolytic deamination: the sulphonated cytosine is deaminated at Carbon 4, producing sulphonated uracil;

- alcaline desulphonation: in alkaline environment, sulphonated uracil loses the sulfur group on Carbon 6 and converts into uracil.

The final uracils are recognized as thymines and, during PCR amplification on the complementary strand, are paired with adenines.

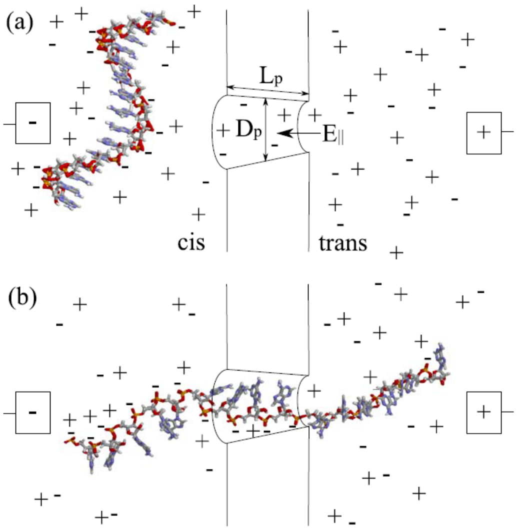


Figure 1.13: Schematic of the nanopore sequencing framework. (a) Electric field and ionic current inside the pore; L_p and D_p are the length and the diameter of the nanopore, respectively. (b) The PN chain crosses the pore, partially blocking the ionic current through the pore. Figure from [4].

1.5 Nanopore sequencing

The full cost of resequencing a haploid human genome with second generation technologies is estimated to be in the region of $\$100,000$ to $\$1,000,000$, that is quite high to be widely accessible by people. In 2004, the US National Institute of Health set a price of \$1000 as a 10-years goal to sequence a human genome [10].

The nanopore sequencing technology approaches this challenging objective. The idea is to pull single-stranded PolyNucleotides (PNs) through a biological (or synthetic) nanopore by the application of an electric potential across the pore that pulls on the charged PN structure [4]. A membrane separates a solution into two halves (*cis* and *trans* chambers) and an electrode is put in each of them. In absence of the polynucleotides chain, an ionic current I_0 is produced by pulling ions through the channel. Phosphate groups are negatively charged and, by introducing the PN chain in the solution, it is pulled to the positively biased half of the solution. The process is shown in Fig. 1.13. The strand is then captured and enters the channel, partially blocking the ion current for the time it needs to cross, called the translocation duration t_d. Each PN is charged but, because of the slow velocity compared to the one of ions, they make a small contribution to the ionic current. The overall effect results in the formation of the so-called blockade current I_b, that accounts for this perturbation, Fig. 1.14. Moreover, from the DNA translocation it is possible to get information about length of the strand and its dynamics and composition. During this passage, the DNA sequencing takes place, according to different approaches that can be adopted: study of the blockade current, study of the transverse current, exonuclease enzyme exploitation and multiple recording sites [4].

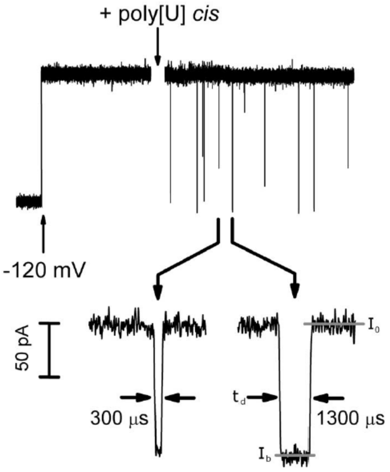


Figure 1.14: Example of an ionic current signal from [11]. Upper: a voltage of -120 mV is applied across the membrane. When Poly[U] enters the *cis* chamber, it causes blockade events. Lower: two examples of blockade events. Ionic current I_0, blockade current I_b and translocation duration t_d are reported.

1.5.1 Types of nanopores

The nanopores that can be employed in this sequencing apparatus can be divided into two main types: biological pores and synthetic ones. The characteristics of the nanopore and the ability to control them are extremely important for sequencing techniques. Methods that exploit nanoscale probes embedded in the pore will require a suitable pore size to maximize the signal difference between the bases beyond the several sources of noise. It means that the pore diameter would have the same width as the single strand of DNA. As a consequence, this maximization of signal difference must be balanced with other effects: DNA-surface interaction is minimal with a large-width pore while DNA capture and translocation cannot occur at small pore sizes. Moreover, the need for inserting the nanoscale probes inside the pore sets a limit on the types of material that can be used for the pore (and its size and shape) [4].

Biological pores

Two types of biological pores are described in this section: α-hemolysin pore and Mycobacterium smegmatis porin A. The α-hemolysin pore (αHL) is shown in Fig. 1.15.

It is composed by 7 sub-units, equal to each others and has axial symmetry about the central channel, whose diameter varies between 14 and 26 Å. It is divided into two parts, both about 50 Å long: the first wider part, called vestibule, with an average diameter of 26 Å, and the second one, called stem, narrower, where the nucleotides read takes place. From Fig. 1.15, it is possible to observe that the two parts are separated by a central constriction of about 1.4 *nm*, which is the smallest restriction: it means that only ss-DNA can cross it (ds-DNA has a diameter of about 2 *nm*). Moreover, the pore is also both small and long enough that the PN must be locally extended, which gives an entropic barrier for transport due to unraveling of the polymer [4].

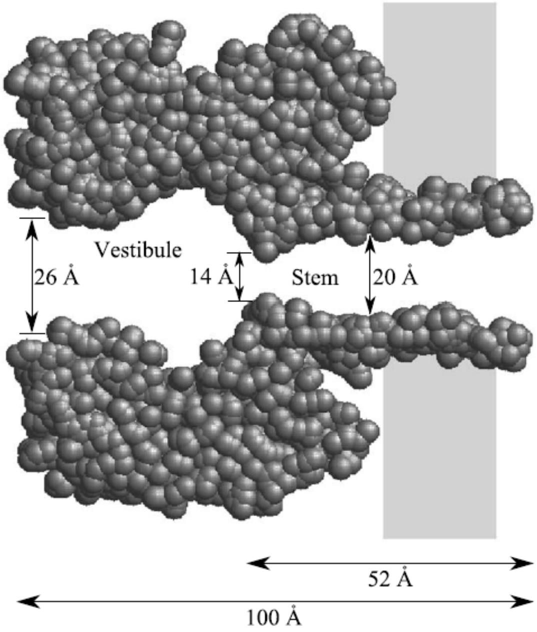


Figure 1.15: Slab view of the α-hemolysin pore. The initial part of the pore is called vestibule and has the wider diameter. Then, there is a central constriction before the 5 *nm* long pore stem. The gray area is the lipid bilayer in which the α-HL assembles to make the pore. Figure from [4].

An alternative to the αHL pore is the Mycobacterium smegmatis porin A (MspA) pore, depicted in Fig. 1.16. MspA it is made of 8 subunits, equal to each others, and has a central channel of variable diameter, which uniformly lowers up to a final short ($\sim$ 0.6 *nm* long) and narrow ($\sim$ 1.2 *nm* in diameter) constriction. MspA has axial symmetry around the central channel. Because of the presence of negative charges near the constriction, that forbid the DNA translocation, a mutant, named M1-MspA, was developed, in which three negatively charged aspartic acids were replaced with neutral asparagines, resulting in a neutral constriction. Its dimensions make it a valid candidate as DNA sequencing nanopore [12].

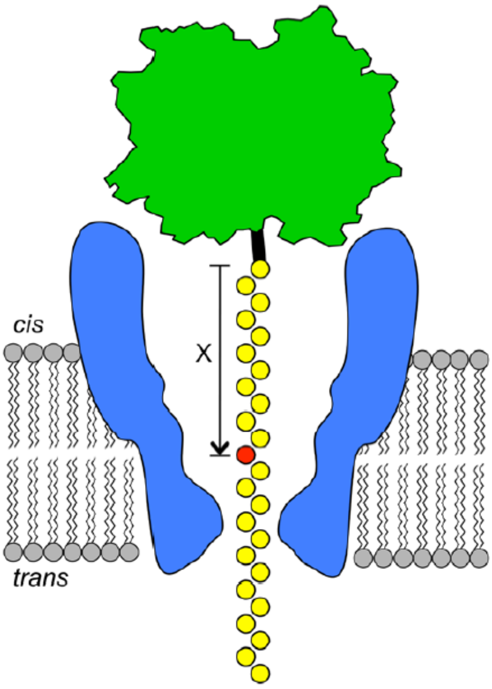


Figure 1.16: Schematic diagram of MspA (blue) set up in a lipid bilayer (grey). Single stranded DNA (ssDNA) was attached to a NeutrAvidin molecule (green) using a biotin linker (black). A specific nucleotide (red) is designated by its position, X, from the biotin-NeutrAvidin 'anchor'. Figure from [12].

Synthetic pores

Synthetic pores are valid alternative for DNA detection and sequencing and offer additional opportunities with respect to biological ones. For example, pore dimensions and properties can be suitably adjusted to meet the needs for a given experiment and it is also possible to integrate them with external sensor and probes (e.g., transverse electrodes). Moreover, they allow to manage different parameters, such as salt concentration, temperature, voltage, and viscosity to obtain optimal detection and sequencing yield, and control translocation velocity and capture rate. Two main approaches to synthetic nanopores fabrication are possible [4].

The first technique is based on the creation of a large-diameter ($\sim 60\ nm$) pore in a solid-state membrane (e.g. Si_3N_4 membrane) using focused ion beam (e.g. Ar^+ beam). When the pore is exposed to the beam, it activates a diffusion process and the pore starts to close, Fig. 1.17. The pore shrinking can be controlled by measuring the ion beam current.

The second technique exploits a high-energy electron beam and SiO_2 membrane. Once the initial pore is made, a Transmission Electron Microscope (TEM) is used to shrink it and, thanks to the imaging capability of this device, it is possible to monitor

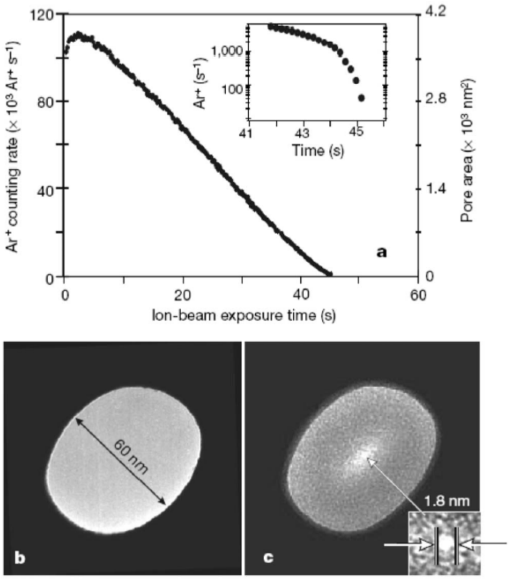


Figure 1.17: Fabrication of a synthetic nanopore. (a) The Ar^+ current decreases with the pore shrinking. (b) A Si_3N_4 membrane with a large pore is first created. (c) The ion beam drives the diffusion process that closes the pore. Figure from [4].

the pore diameter reduction (the shrinking rate is relatively slow, ~ 0.3 nm) and stop the process when the desired size is reached.

However, synthetic pores bring some limitations. For instance, silicon-oxide surfaces can have a negative surface charge density in aqueous solutions, creating additional complications for the understanding of PN transport, and must be neutralized by counterions. On the other hand, biological pores can have unusual potential profiles that may have internal sites with trapped charges. Also graphene nanopore and hybrid pores, resulting from the integration of a protein pore into a solid-state membrane are under exploration.

1.5.2 Properties

The translocation of the DNA through the pore can be divided into two categories, both important: the ones that rely on polymer dynamics (entropic forces, Brownian motion, charges, etc.), which are called universal properties, and the ones which depends on the atomic compositions of the nucleotides (e.g., interaction potentials with the pore surface), called specific properties. The latters will not be treated here. The following information is mainly from [4].

The processes of capture and translocation are the basic properties of the polymer dynamics. The capture is affected by the diffusion of the DNA strand from the bulk to the pore and on local effects around the pore (electric field, interaction between the entrance of the pore and the polymer, etc.). The translocation is not only driven by the applied voltage, but also by polymer-pore interactions, ionic effects, and viscous drags.

It has been shown that the speed of the DNA strand does not depend on its length above about 12 nucleotides (polymer's length equal to pore length). Below this limit, the speed increases with the decreasing length.

Another interesting property is the one which connects the polymer dynamics with the pore size. If the pore diameter is comparable to the size of the polymer width, the interaction between the pore walls and the strand is strong. As a consequence, the resistance to flow increases and the time scale of translocation t_d is controlled by it. Moreover, the translocation duration has been measured to be linearly proportional to the strand length. On the contrary, if the pore is wide, this linear proportionality does not hold anymore.

Considerations about the functional dependence of the time scale of translocation in this case and about balance between drag and driving forces [4] will not be discussed here.

1.5.3 Sequencing and electronic detection

As previously written, ionic currents through the nanopore are fundamental in order to detect polynucleotides, even if there are many things to understand about the translocation. The theory exposed in this subsection is from [4] and [13].

Blockade current

When the polynucleotide enters the pore, it partially blocks the ions crossing the pore, reducing the ionic current according to the amount of pore volume occupied by the DNA strand. It is important to have a clearer understanding of the contribution of the volume exclusion to the blockade current. Since the charge flux depends on both the applied voltage and the diffusion, the steady-state form of the Nernst-Planck equation is reported:

$$J_z = e\mu n E_z - eD\frac{\partial n}{\partial z} \tag{1.1}$$

where J_z is the charge flux along the z direction, e is the electric charge, μ is the ion mobility, n is the density of the charge carriers, D is the diffusion coefficient and E_z is the driving field in the z direction. If the diffusive term can be neglected, the open pore current is simplified in

$$I_0 = e\mu n E_z \tag{1.2}$$

where E_z can be approximated to V/L_p, being L_p the pore length. The quantity $e\mu n = \sigma$ is called the conductivity of the pore. If the change of the current is only due to the volume exclusion

$$n_b = \frac{V_p - V_N}{V_p} n = Fn \tag{1.3}$$

where n_b is the carrier density during a blockade event, V_p is the pore volume for one repeat unit of the polynucleotide and V_N is the volume of a nucleotide. The reduction in current is then given by:

$$\frac{I_0 - I_b}{I_0} = \frac{A_b}{A_0} = 1 - F \tag{1.4}$$

with A_b the portion of area that is blocked while A_0 is the pore area. DNA and RNA nucleotides' volume does not significantly differ: the extra hydroxyl group has a negligible effect on molecular volume and the secondary structure should not cause a variation in the volume occupied per nucleotide. However, experimental data do not fit this theoretical hypothesis, suggesting for additional contributions from other effects. The volume exclusion leads to an increased ions confinement, generating an extra electrostatic barrier which deactivates some charge carriers. Moreover, as previously written, polynucleotides transport a charge, though small, inside the pore.

It is also important to consider the effect of the pore electrostatic environment on the transport. This could be done by extending the Nernst-Planck equation with a simple model of pore electrostatic environment, with dielectric constants and mobilities which depend on the position, and introducing a free-energy potential, whose change would be quite large for small pore diameter, causing a suppression of the ion transport [4]. So, in case of large free-energy change, the majority of the ions do not have enough energy to cross the pore. The number of ions available for the transport is

$$n = n_0 e^{-\Delta F_b/kT} \tag{1.5}$$

where n_0 is the bulk density, k the Boltzmann constant, T the temperature and ΔF_b represents the free-energy barrier, due to both electrostatic and entropic contributions. For large pores, both the contributions can be neglected.

The first encouraging results led to speculate about the possibility of DNA sequencing by measuring the ion blockade. However, experiments with αHL pores demonstrated that the ionic blockade can be exploited only to detect blocks of nucleotides and to extract information about the secondary structure and directionality of polynucleotides. Adding information from translocation duration allowed to distinguish homopolymers constituted by different bases. Moreover, it is possible to obtain some internal information about the strand by analysing the blockade of a polynucleotide with two homogeneous blocks (it has a stair structure).

However, a single-base resolution is needed for DNA sequencing. The different sizes of the bases and the different interactions between pore and bases must be detected. So, in order to perform DNA sequencing, the nanopore must have a length L_p of approximately one nucleotide (< 1 nm) and a diameter between 1 and 2 nm [4]. If the exclusion volume is the only cause of ionic blockade, the differences of the bases in a pore of 1.5 nm diameter and 0.7 nm long would be only a few percent and the noise of the current itself would be much bigger than this, so that it may not be possible to successfully discriminate the different bases. Some possible solutions could be explored. A foreign molecule (an “adapter”) can be inserted in a pore so that it can interact in a peculiar way to each of the four types of nucleotide bases, creating distinguishable ionic currents. Also, exonuclease digestion may be employed for sequencing. It takes a DNA strand and removes one base at a time. As a consequence, the sequencing speed would be considerably slower than the intrinsic speed of a nanopore sequencing experiment.

Transverse current

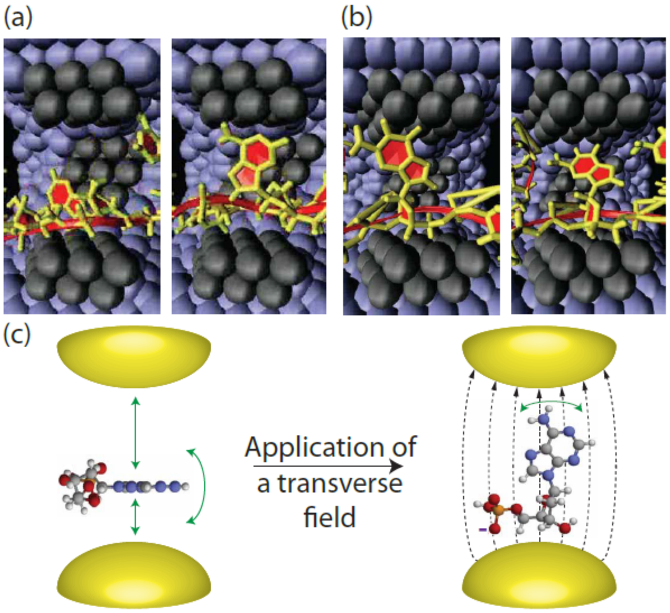


Figure 1.18: Simulation of the molecular dynamics of a PN translocation through a pore. (a) Without a transverse field, base fluctuations are large; (b) with a transverse field, they significantly decrease; (c) schematics of a polynucleotide within electrodes with and without the transverse electric field. Figure from [13].

It is also possible to measure the transverse current, considering that, at the nanoscale, electrons can tunnel. This can be done through the use of electrodes embedded within the pore. Since the difference in both molecular structure and energy level of the nucleotide bases would result in different couplings to electrodes, each base is characterized by its own current signature. This is affected mainly by the couplings of the molecular states to the electrodes, which in turn depends on other factors, such as the driving of the strand through the pore, nucleotide-fluid interaction, polynucleotide movements and structural fluctuations. So, the idea is to discriminate the bases according to their intrinsic electronic characteristics.

By following a Landauer approach, it is possible to compute the electronic current flowing across the atomic constituents of the junction and, at the same time, to account for the structural fluctuations effect. The current is given by:

$$I = \frac{2e}{h} \int_{-\infty}^{\infty} T(E)[f_t(E) - f_b(E)]\, dE \tag{1.6}$$

where e and h are the electron charge and Planck's constant, $T(E)$ is the transmission coefficient, while $f_t(E)$ and $f_b(E)$ are the Fermi-Dirac distributions for the top (t) and bottom (b) electrodes, respectively. The expression of the transmission coefficient is:

$$T(E) = Tr[\Gamma_t G_{DNA} \Gamma_b G_{DNA}^{\dagger}] \tag{1.7}$$

where G_{DNA} is the retarded Green's function:

$$G_{DNA} = \frac{1}{ES_{DNA} - H_{DNA} - \Sigma_t - \Sigma_b - \Sigma_n} \tag{1.8}$$

with S_{DNA} and H_{DNA} being the overlap and Hamiltonian matrices, respectively, of the contents of the junction. The remaining terms are the self-energies of the top electrode, bottom electrode, and external probes (i.e. potential sources of noise). These self-energies represent the influence of the electrodes and external noise on the DNA and other contents of the junction. The other term is:

$$\Gamma_{t(b)} = i\left(\Sigma_{t(b)} - \Sigma_{t(b)}^{\dagger}\right) \tag{1.9}$$

It has been found that base-electrode coupling depends on the base-electrode distance [4]. As a consequence, it is important to control DNA motion in order to avoid the current fluctuations, which are of several orders of magnitude. It has been proposed to exploit the action of an electric transverse field to control the polynucleotide motion. For instance, by placing a capacitor across the nanopore system, it is possible to pull the DNA backbone toward one side of the nanopore, slowing its motion and reducing fluctuations at the same time, as shown in Fig. 1.18. Doing so, the DNA nucleotide bases acquire well-defined electronic current distributions. They approximately follow a log-normal distribution in many circumstances, and only partially overlap, so that they can be discriminated with only a small number of independent measurements of the current [13]. The transport across the junction can be modelled as due to a single electronic state on the DNA, so that the current can be approximately described by the following expression:

$$I \approx \frac{2e^2V}{h}\frac{\gamma^2}{E_0^2+\eta^2} \tag{1.10}$$

where a linear response is assumed. The expression holds for a single electronic level equally coupled to both electrodes (γ coupling strength) and in presence of noise of strength η. The dominant term is the coupling strength. The main feature of the current, that is the Gaussian fluctuations on a log-scale, is due to exponential fluctuations of the coupling of the DNA to the electrodes. The expression above for the current, then, holds for one instant of time. Since there are fluctuations of both DNA and other contents of the junction, the coupling constant will be described by its probability distribution:

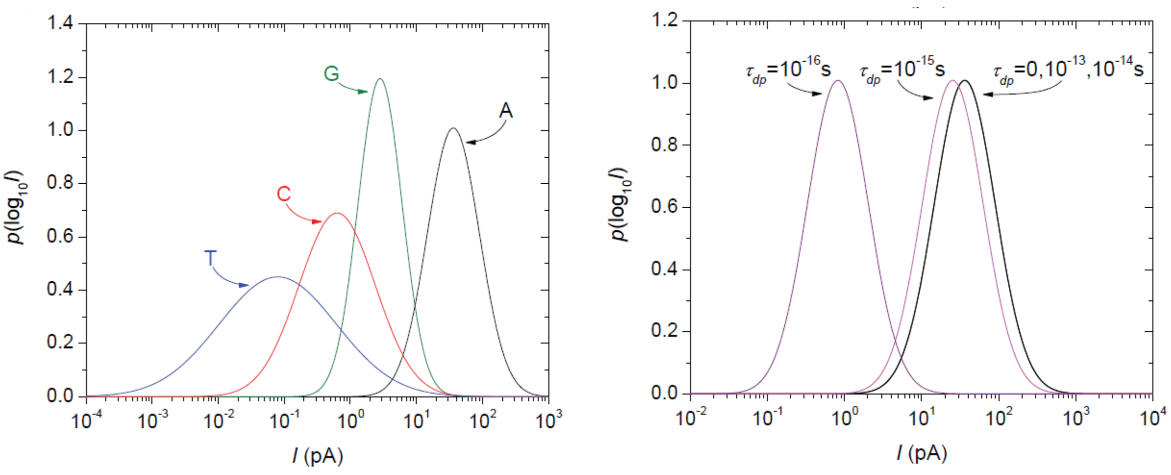
$$p(\ln\gamma/\gamma_m) = \frac{1}{\sqrt{2\pi\sigma_\gamma^2}} exp\left\{-\frac{(\ln\gamma/\gamma_m)^2}{2\sigma_\gamma^2}\right\} \tag{1.11}$$

with standard deviation σ_y and maximum likelihood value γ_m. Only when the DNA motion is controlled by a transverse field the coupling constant takes on the log-normal distribution. These fits to experimental data in terms of current distributions are shown in Fig. 1.19.

As a consequence, since currents are given by continuous probability distributions, it is not possible to discriminate between the bases with just a single measurement, but it is needed to take several measures to obtain a signal which allows single-base detection, as can be inferred from Fig. 1.19a. Also measuring an average current when a single nucleotide is present within electrodes could be a viable alternative. Experimental results demonstrated that the measure is not sensibly affected by ionic and aqueous environment. Moreover, the white noise due to electrodes dephasing should be considered only for intensities greater than the gap between the molecular energy levels and the Fermi level of the electrodes [13], as shown in Fig. 1.19b.

1.5.4 Recognition sites

In order to study to which nucleotides the αHL nanopore is sensitive to, the DNA has to be slowed down while it crosses the pore, so that the translocation time t_d of the single nucleotides will be long enough to perform measurements. It has been shown that the single strand DNA can be ratcheted one base at a time through the pore by the use of a DNA polymerase. Also, the DNA chain can be immobilized by the action of a NeutrAvidin protein, as shown in Fig. 1.20. This approach can improve the resolution of the currents associated with individual nucleotides, because of the prolonged observation time. It has been demonstrated by [14] that the 5 nm-long β barrel of the



(a)

(b)

Figure 1.19: Current distributions for the different nucleotides between electrodes of spacing 1.4 nm and set at a bias of 1 V. (a) Characteristic electronic currents of each base within two golden leads without noise ($\eta = 0$). (b) Variation of the adenine distribution due to white noise from electrodes dephasing $\eta = \hbar/\tau_{dp}$. Figure from [13].

αHL nanopore contains three recognition sites, R_1, R_2 and R_3, each capable to recognize single nucleotide bases of DNA strands: R_1 is located near the internal constriction in the lumen of the pore and recognizes bases at positions $\sim$ 8 to 12 (bases are numbered from the 3′ end of synthetic oligonucleotide probes); R_2 is located near the middle of the β barrel and discriminates bases at positions $\sim$ 12 to 16 while R_3 recognizes bases at positions $\sim$ 17 to 20 and is located near the *trans* entrance of the barrel, Fig. 1.20.

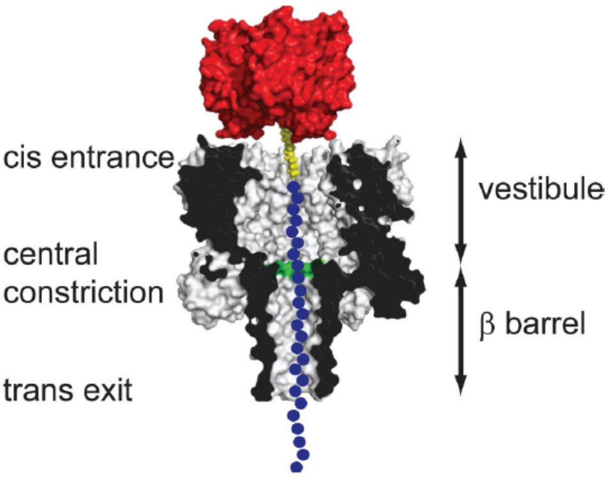
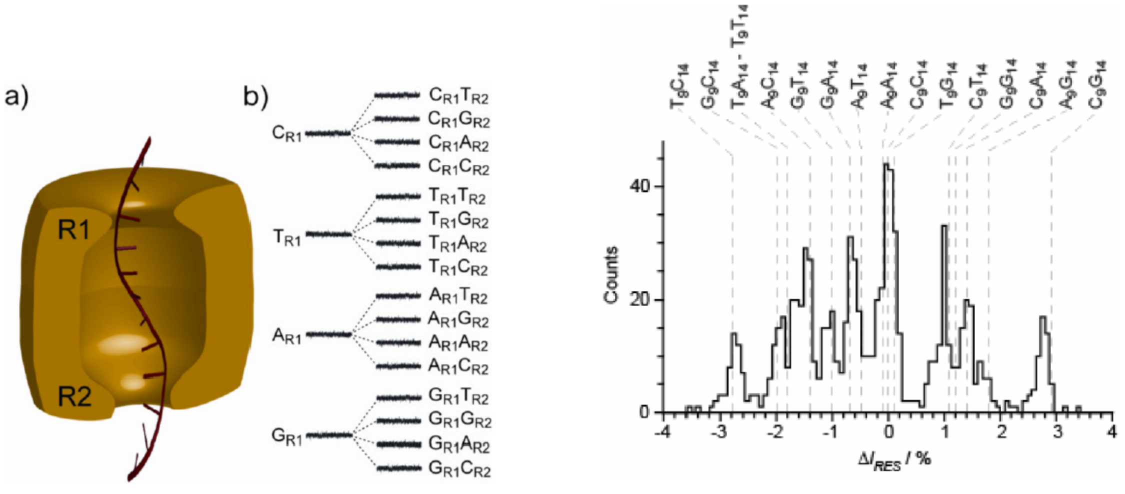


Figure 1.20: Schematic representation of a homopolymeric DNA oligonucleotide (blue circles) immobilized inside an αHL pore (grey) through the use of a biotin (yellow) - streptavidin (red) linkage. The different regions are highlighted. Figure from [15].



(a) (b)

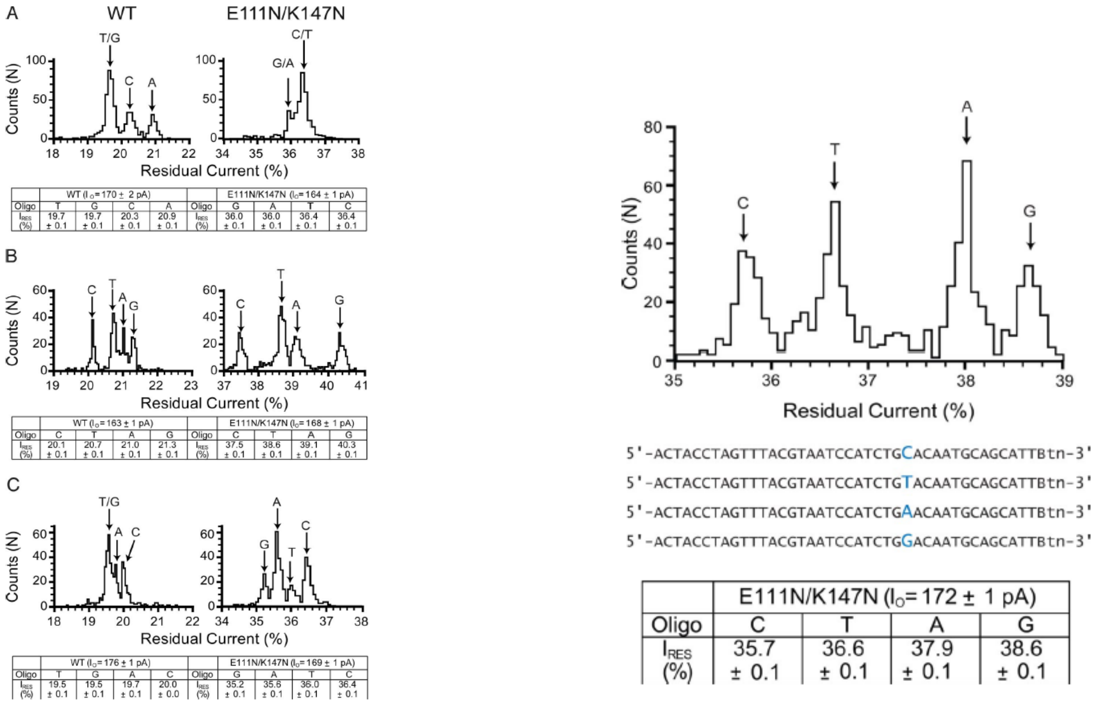
Figure 1.21: (a) Example of a nanopore with two recognition sites R_1 and R_2. with the possible current levels and base combinations. (b) Histogram of the residual currents due to the different nucleotide combinations among the two reading-head system. Figure from [14].

In an αHL pore some amino acids can be removed to reduce the polynucleotide-pore interaction in the desired zones. To determine the pore capability to recognize the four bases, homopolymeric DNA strands with only one substituted base are introduced in the pore. The differences between the peaks of the distributions are of the order of pA.

It is important to understand how to exploit information obtained from recognition sites. It might be advantageous to use two recognition sites within a single pore. In this way, even if they are blunt, sufficient information might be collected to obtain DNA sequence information. Let's consider a nanopore with two recognition sites, called R_1 and R_2, each one capable of recognizing all four bases, Fig. 1.21a. If the first site, R_1 produces a large dispersion of current levels for the four bases and the second site, R_2, produces a lower dispersion, a total of 16 current levels, one for each of the possible base combinations, would be observed as DNA molecules cross the nanopore. Therefore, the current signal would carry information about two positions in the sequence, rather than just one, providing redundant information. In fact, each base is read twice, first at R_1 and secondly at R_2. This mechanism would improve the overall quality of sequencing. Despite the 16 DNA sequences, Fig. 1.21b, did not produce 16 discrete current levels, it is possible to resolve 11. So, an optimal single head system would read the sequence just once, while a perfect two heads system would read the sequence twice, proucing 16 current levels. Therefore, even if the 11-levels system is imperfect, it does produce additional, redundant information about each base, which would provide more solid base identification than a single reading head. An additional third reading head would increase the number of possible base combinations from 16 to 64. Separating 64 levels is a very difficult task, especially considering the electrical noise of the system. Therefore, a two reading-head sensor seems to be optimal [14].

1.5.5 Sequencing results

In the research work of [15], a wild-type αHL pore was compared to a E111N/K147N pore, to test whether they can distinguish between different bases within a DNA chain. All the three recognition sites were tested using homopolymers (poly(dC)) except for a specific position in which a different base was substituted: from Fig. 1.22a, it is clear that site R_2 can distinguish between all the four bases, both for wild-type and engineered pore, while for R_1 and R_3 there is not enough discrimination between the residual currents. Moreover, the ability to discriminate nucleotides within an heteropolymer was tested using the R_2 site of the E111N/K147N pore, which resulted to be the most promising one. The results are shown in Fig. 1.22b. All the four bases were recognized in the same order as in the homopolymeric case. It is possible to notice that the spacing between the four peaks differs from the homopolymeric case. However, it may be advantageous to use 2 recognition sites within a single pore. So, even if they are blunt, they carry sufficient information for DNA sequencing, together with information on the strand length.



	WT (I_O = 170 ± 2 pA)					E111N/K147N (I_O = 164 ± 1 pA)			
Oligo	T	G	C	A	Oligo	G	A	T	C
I_{RES} (%)	19.7 ± 0.1	19.7 ± 0.1	20.3 ± 0.1	20.9 ± 0.1	I_{RES} (%)	36.0 ± 0.1	36.0 ± 0.1	36.4 ± 0.1	36.4 ± 0.1

	WT (I_O = 163 ± 1 pA)					E111N/K147N (I_O = 168 ± 1 pA)			
Oligo	C	T	A	G	Oligo	C	T	A	G
I_{RES} (%)	20.1 ± 0.1	20.7 ± 0.1	21.0 ± 0.1	21.3 ± 0.1	I_{RES} (%)	37.5 ± 0.1	38.6 ± 0.1	39.1 ± 0.1	40.3 ± 0.1

	WT (I_O = 176 ± 1 pA)					E111N/K147N (I_O = 169 ± 1 pA)			
Oligo	T	G	A	C	Oligo	G	A	T	C
I_{RES} (%)	19.5 ± 0.1	19.5 ± 0.1	19.7 ± 0.1	20.0 ± 0.0	I_{RES} (%)	35.2 ± 0.1	35.6 ± 0.1	36.0 ± 0.1	36.4 ± 0.1

(a)

5'-ACTACCTAGTTTACGTAATCCATCTGCACAATGCAGCATTBtn-3'
5'-ACTACCTAGTTTACGTAATCCATCTGTACAATGCAGCATTBtn-3'
5'-ACTACCTAGTTTACGTAATCCATCTGAACAATGCAGCATTBtn-3'
5'-ACTACCTAGTTTACGTAATCCATCTGGACAATGCAGCATTBtn-3'

	E111N/K147N (I_O = 172 ± 1 pA)			
Oligo	C	T	A	G
I_{RES} (%)	35.7 ± 0.1	36.6 ± 0.1	37.9 ± 0.1	38.6 ± 0.1

(b)

Figure 1.22: Histograms and Gaussian fit means of residual current levels. (a) Recognition of the 4 bases by the WT and the E111N/K147N pore. (b) Single nucleotide discrimination in heteropolymers by the E111N/K147N pore. Figure from [15].

A mutation of the αHL pore ($WT-(M113R/N139Q)_6(M113R/N139Q/L135C)_1-am_6amDP_1\beta CD$) was exploited in the experiments performed by [10]. The aim was to demonstrate that it could be possible to use αHL pore to distinguish between methylated and unmethylated cytosines, which could have a crucial role in the epigenetic analysis, without the need to resort to bisulphite treatment. Residual current histograms and Gaussian fits are shown in Fig. 1.23. All the five nucleotides can be distinguished.

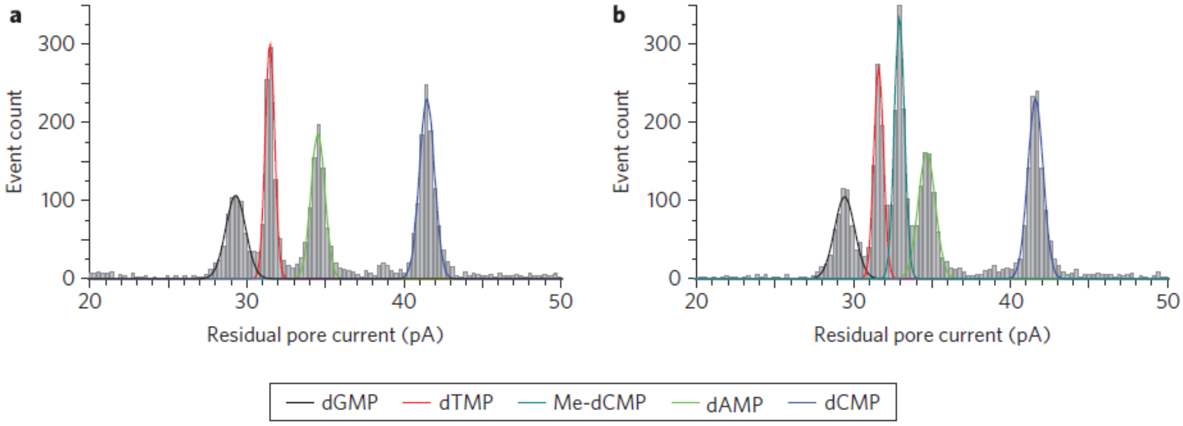
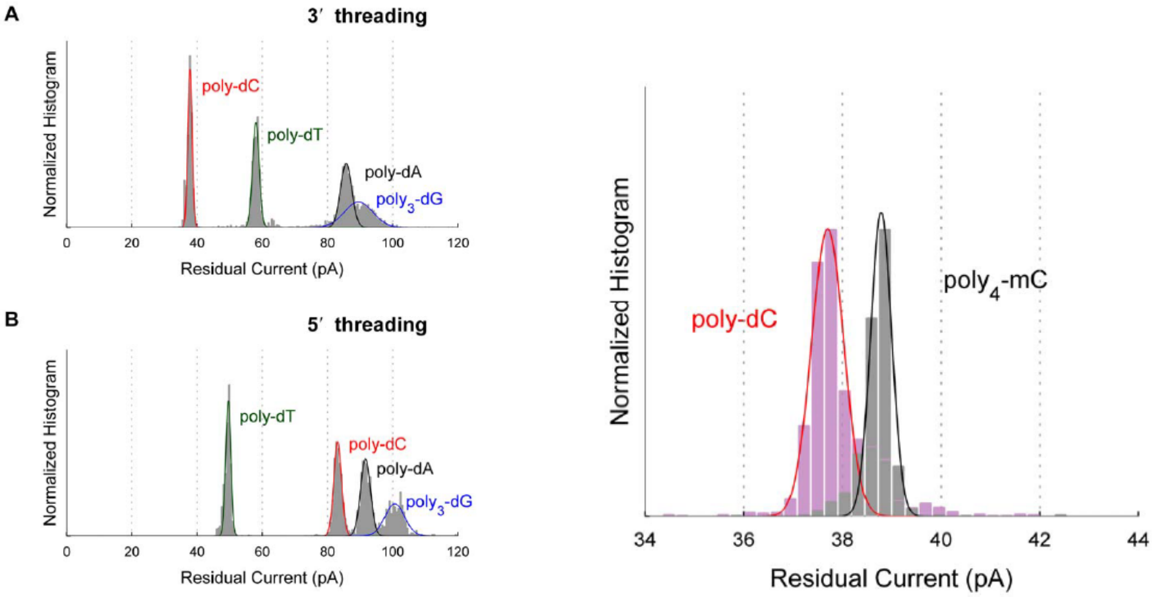


Figure 1.23: Residual current histograms for the ($WT-(M113R/N139Q)_6(M113R/N139Q/L135C)_1-am_6amDP_1\beta CD$) pore in a mixture of (a) dGMP, dTMP, dAMP and dCMP; (b) addition of Me-dCMP. The five nucleotides are clearly distinguishable. Figure from [10].

Considering the work of [12], a Mycobacterium smegmatis porin A (MspA) pore was tested. Residual currents were recorded for homopolymeric strands (poly(dC), poly(dA) and poly(dT)) and (dA)13(dG)3(dA)34. The experiments were performed for both the orientations ($3' \rightarrow 5'$ and $5' \rightarrow 3'$). The histograms of the mean residual currents are shown in Fig. 1.24a. The residual current levels are sensitive to the nature and the orientation of the nucleotides. For poly(dC), poly(dA) and poly(dT) residual current levels are well resolved (at least 8 pA of separation) for both orientations. It is possible to notice an overlap of the Gaussian distributions of poly(dA) and (dA)13(dG)3(dA)34 for $3'$ threading, while it is lower for $5'$ threading. Moreover, in this latter case poly(dT) peak is separated from the other peaks by about 33 pA. Gaussians of (dA)13(dG)3(dA)34 are wider than the others, probably due to the surrounding adenine nucleotides influencing the ionic current. In addition, the ability of discriminating between methylated and unmethylated cytosines was tested. Histograms of residual current levels are shown in Fig. 1.24b. The areas of the two Gaussians overlap of about 2% and the peaks are separated by about 1 pA. A single cytosine was introduced in a poly(dA) strand in different positions to determine which is the most sensitive region in the MspA pore. This



(a) (b)

Figure 1.24: Histograms and fit curves of residual current levels for (a) the four homopolymers in the 3′ → 5′ (up) and 5′ → 3′ (bottom) orientation; (b) the methylated and unmethylated cytosines. The strand-orientation dependence is clear. Figure from [12].

experiment allows to state that there is only a single region sensitive to DNA nucleotides: nucleotides placed in the MspA's constriction (and their neighbours) mostly affect the ionic current. If a SNP in a heteropolymeric strand is studied, the signal is clearly distinguishable when the substituted nucleotide is centred in the constriction (or in its neighbourhood).

From the experiments mentioned above it comes out that the differences in nucleotide residual currents found with MspA are about an order of magnitude larger than those ones found with αHL. Moreover, MspA has only one recognition site. The fact that the short MspA constriction results in a small volume of high current density could be the reason why it has a large specificity to nucleotides in the constriction. On the other hand, differently from αHL, MspA shows a strand-orientation dependent interaction with the pore.

Chapter 2

The Dirichlet Process

In this chapter, the Bayesian interpretation of probability is presented in contrast with the frequentist approach, and the main discrete probability distributions, up to the Dirichlet distribution, are introduced. Then, the Dirichlet Process, a statistical learning method used for many applications, including clustering, is described. The Dirichlet Process is introduced in a formal way and then, practical sampling methods and metaphors are provided. Finally, the Hierarchical Dirichlet Process, a generalization of the latter one to grouped data, is presented.

2.1 Background

In this paragraph, the differences between the two main approaches to statistic, frequentist and Bayesian, are outlined. Moreover, an introduction to the density estimation problem is given. Then, the main discrete probability distribution, important for the introduction of the Dirichlet Process, are presented.

2.1.1 Bayesian interpretation of probability

There are two main approaches to statistics, the frequentist (or classical) approach and the Bayesian approach.

In frequentist statistics, probabilities are associated only with the data, i.e. outcomes of repeatable experiments. Probabilities are specified in terms of random, repeatable events. Probabilities can be only assigned to data, not to statements. The tools of the frequentist philosophy tell us what is expected, under the assumption of certain hypotheses, about hypothetical repeated observations. For example, let's consider a coin toss event. It could be interesting to ask which would be the probability to obtain a 'head'. The frequentist approach will repeat the coin toss experiment many times and count the number of heads over the total number of repetitions, from which it is possible to infer information about this random process. The probability is seen as a limiting frequency.

In Bayesian approach, probabilities are described as a quantification of uncertainty, as a degree of belief. For example, let consider whether the Sun will become too cold to allow life on Earth by the end of the millennium. This is a unique event: there are no events that can be repeated many times in order to define a notion of probability, as in the coin toss example. However, we could have a general idea about how quickly the Sun is cooling. This opinion can be revised if, for instance, we get some diagnostic information from an astronomical observation. It could be interesting to quantify our expression of uncertainty and update the uncertainty in the light of new evidence, as well as to be able to take optimal actions or decisions as a consequence. The Bayesian interpretation allows to deal with such circumstances.

This subjective interpretation of probability is based on the Bayes' theorem. It expresses the conditional probability of an event known that another event happened and can be used to convert a prior probability, which models our opinion, into an updated posterior probability, by incorporating the evidence provided by the observed data. Let's imagine, for instance, to deal with a polynomial curve fitting problem, where we want to make inference about the parameters $\boldsymbol{w}$. Our opinion about $\boldsymbol{w}$, i.e. our assumptions before observing the data, can be expressed through a prior probability distribution $p(\boldsymbol{w})$. The effect of observed data D can be described using the conditional probability $p(D|\boldsymbol{w})$. The Bayes' theorem is given by

$$p(\boldsymbol{w}|D) = \frac{p(D|\boldsymbol{w})p(\boldsymbol{w})}{p(D)} \tag{2.1}$$

where the uncertainty on $\boldsymbol{w}$ after having observed data D is given by the posterior probability $p(\boldsymbol{w}|D)$. The term $p(D|\boldsymbol{w})$ is called likelihood function and it express how probable the observed data set is for different settings of the parameters $\boldsymbol{w}$ (it is a function of the parameters vector $\boldsymbol{w}$). It is possible to notice that the likelihood is not a probability distribution (so, its integral over $\boldsymbol{w}$ is not equal to one). The denominator is a normalization term which ensures that the posterior integrates to one [16]. The Bayes' theorem can also be expressed in words:

$$\textit{posterior} \propto \textit{likelihood} \times \textit{prior} \tag{2.2}$$

The likelihood function is also present in the frequentist approach, but there are some differences to point out. In a frequentist setting, $\boldsymbol{w}$ is considered a fixed vector of parameters and its value can be determined using some estimators by considering the distribution of possible data sets D. On the contrary, in Bayesian philosophy only a single data set D is present, and the uncertainty on the parameters is described through a probability distribution over $\boldsymbol{w}$.

Another important issue is the density estimation problem: the aim is to model the probability distribution $p(\boldsymbol{x})$ of a certain random variable $\boldsymbol{x}$, given a set $\{\boldsymbol{x_1}, ..., \boldsymbol{x_N}\}$ of N observations. Theoretically, infinite distributions can fit this finite dataset (ill-posed problem) [16]. You must distinguish between parametric and non-parametric density estimation. Parametric distributions are the ones governed by a small number of adaptive parameters (e.g., mean and variance for a Gaussian). In a frequentist approach, a specific value for the parameters is obtained by optimizing a likelihood function, for instance. In a Bayesian approach, prior distributions over the parameters must be introduced. Once data have been observed, the posterior is computed using the Bayes' theorem. Non-parametric approaches are based on histograms, nearest-neighbours and kernels. They have some parameters which control the model complexity, instead of the form of the distribution.

2.1.2 Binomial distribution

Let consider a binary random variable, i.e. a random variable that can take on only two values. For example, the variable $x \in \{0, 1\}$ can model the outcome of extracting a card from a deck (with reinsertion), with $x = 1$ representing “clubs card” and $x = 0$ representing “non clubs card”. The probability of $x = 1$ can be denoted by the parameter μ, so that

$$p(x = 1|\mu) = \mu \qquad p(x = 0|\mu) = 1 - \mu \tag{2.3}$$

where $0 \leq \mu \leq 1$, because it is a probability. So, the probability distribution over x can be written in the form

$$Bernoulli(x|\mu) = \mu^x (1 - \mu)^{1-x} \tag{2.4}$$

that is called *Bernoulli distribution.* Mean and variance of the Bernoulli distribution are

$$\mathbb{E}[x] = \mu \tag{2.5}$$

$$var[x] = \mu(1 - \mu) \tag{2.6}$$

Let's suppose to have a dataset $D = \{x_1, ..., x_N\}$ of observed values of x. It is possible to extend the Bernoulli distribution to the case in which the variable is observed N times (e. g. extracting a card N times, with reinsertion). Let suppose m to be the number of observations of $x = 1$ ($N - m$ being the number of observations of $x = 0$). The distribution of the number m is given by

$$Bin(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m} \tag{2.7}$$

where

$$\binom{N}{m} = \frac{N!}{(N-m)!m!} \tag{2.8}$$

represents the number of ways to choose m objects out of a total of N identical objects. Eq. 2.7 is called *Binomial distribution* and an example of it is shown in Fig. 2.1. The mean and the variance of the Binomial distribution are given by

$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} mBin(m|N,\mu) = N\mu \tag{2.9}$$

$$var[m] \equiv \sum_{m=0}^{N} (m-\mathbb{E}[m])^2 Bin(m,|N,\mu) = N\mu(1-\mu) \tag{2.10}$$

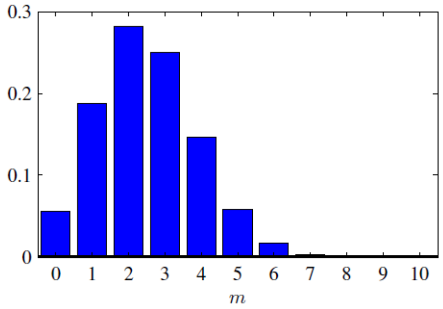


Figure 2.1: Plot of Binomial distribution for $\mu = 0.25$ and $N = 10$. It could represent the “extracting a clubs card” event making 10 trials (with reinsertion). Figure from [16].

2.1.3 Beta distribution

Parameter μ can be estimated by likelihood maximization or in a Bayesian way, through a prior. It could be interesting to find a prior $p(\mu)$ over the parameter μ with the same functional form of the likelihood function, so that the posterior, which is proportional to the product between the prior and the likelihood, will have in turn the same functional form of the prior. This key property is called conjugacy. The *Beta distribution* is a conjugate prior for the Binomial distribution and its expression is given by:

$$Beta(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{2.11}$$

where $\Gamma(z)$ is the gamma function, defined as

$$\Gamma(z) = \int_0^\infty u^{z-1} e^{-u} du \tag{2.12}$$

The parameters a and b are called hyperparameters because they control the distribution of the parameter μ. An example of the Beta distribution for different values of the hyperparameters is plotted in Fig. 2.2.

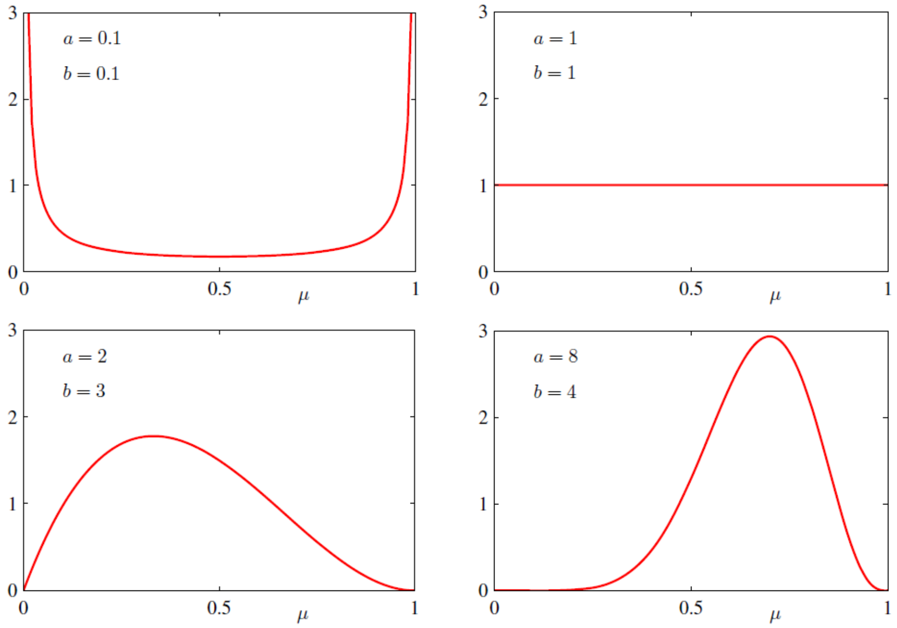


Figure 2.2: Plot of Beta distribution as a function of μ for different values of the hyperparameters a and b. Figure from [16].

The mean and the variance of the Beta distribution are given by

$$\mathbb{E}[\mu] = \frac{a}{a+b} \tag{2.13}$$

$$var[\mu] = \frac{ab}{(a+b)^2(a+b+1)} \tag{2.14}$$

The posterior distribution of μ can be computed through the Bayes' theorem, Eq. 2.2, by multiplying the Beta prior distribution, Eq. 2.11, and the Binomial likelihood, Eq. 2.7:

$$p(\mu|m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)} \mu^{m+a-1}(1-\mu)^{l+b-1} \tag{2.15}$$

where $l = N - m$. Eq. 2.15 shows the same functional form on μ as the prior distribution, demonstrating the conjugacy property. In fact, the posterior distribution is another Beta, for which the hyperparameters a and b can be easily interpreted as an effective number of observations of $x = 1$ and $x = 0$, respectively. In addition, the posterior distribution can act as a prior if we subsequently observe new data: the posterior can be updated by simply multiplying it for the likelihood function of the new data.

2.1.4 Multinomial distribution

It could be possible that discrete variables can assume one of K possible mutually exclusive values. An example is the extraction (with reinsertion) of a card from a deck, where the suit of the card is the outcome, which results in 4 possible values. Such variables can be conveniently represented through the 1 of K scheme, where the variable is described by a K-dimensional vector $\boldsymbol{x}$ in which one of the elements x_i is equal to 1 while all the others are 0. So, as a consequence, $\sum_{i=1}^{K} x_i = 1$ holds. If the probability of $x_k = 1$ is denoted by μ_k, then the distribution of $\boldsymbol{x}$ is given by

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{i=1}^{K} \mu_i^{x_i} \tag{2.16}$$

where $\boldsymbol{\mu} = (\mu_1, ..., \mu_K)^T$, $\mu_i \geq 0$ and $\sum_{i=1}^{K} \mu_i = 1$, because they represent probabilities.

Let's now consider a dataset of N independent observations $\{x_1, ..., x_N\}$. If the number of observations for which $x_i = 1$ is denoted by m_i (i.e. m_1 times outcome 1, m_2 times outcome 2, etc.), it is possible to write down the expression for the joint distribution of the quantities $m_1, ..., m_K$, conditioned on $\boldsymbol{\mu}$ and on the total number of observations N:

$$Mult(m_1, ..., m_K|\boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 ... m_K} \prod_{i=1}^{K} \mu_i^{m_i} \tag{2.17}$$

that is called *Multinomial distribution.* The normalization coefficient is the number of ways of partitioning N objects into K groups of size $m_1, ..., m_K$. Mean and variances for individual m_i are given by:

$$\mathbb{E}[m_i] = N\mu_i \tag{2.18}$$

$$var[m_i] = N\mu_i(1 - \mu_i) \tag{2.19}$$

We can finally notice that variables m_k have to satisfy

$$\sum_{k=1}^{K} m_k = N \tag{2.20}$$

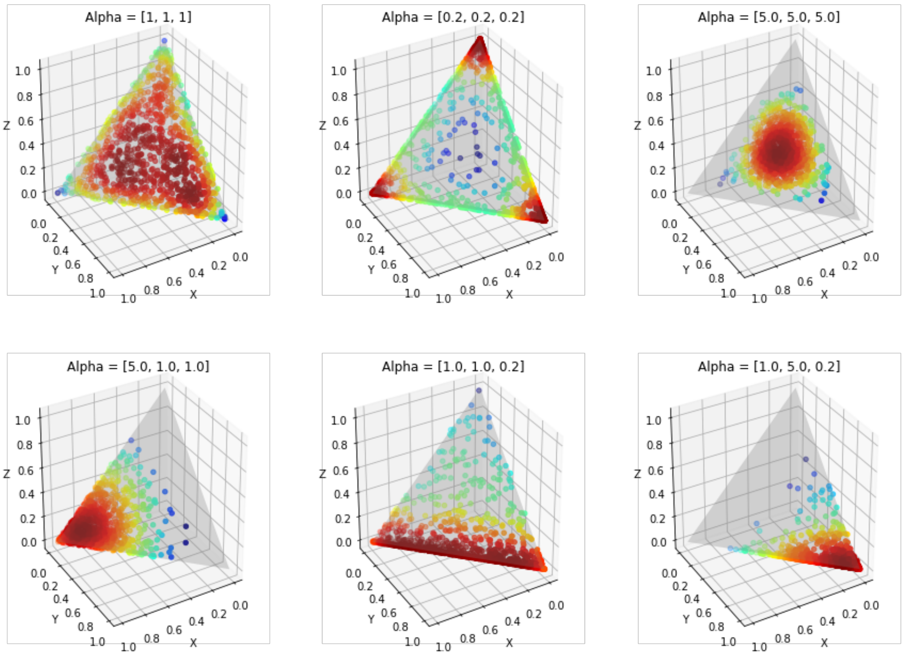


Figure 2.3: Examples of density plot samplings from 3-dimensional Dirichlet Distribution for various settings of the parameters α. Image from `https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet\endline-allocation-437c81220158`.

2.1.5 Dirichlet distribution

It could be interesting to estimate the multinomial parameters $\boldsymbol{\mu}$ in a Bayesian way, through the use of a prior. The *Dirichlet Distribution* (DD) is a family of continuous multivariate probability distributions, parametrized by a vector $\boldsymbol{\alpha}$ of positive real numbers. The DD is a multivariate generalization of the Beta Distribution.

The probability density function of the DD of order $K \geq 2$, with parameters $\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_K\}$ is given by

$$Dir(\boldsymbol{\alpha}) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} \mu_i^{\alpha_i - 1} = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \mu_i^{\alpha_i - 1} \tag{2.21}$$

where $\Gamma(z)$ is the gamma function defined by Eq. 2.12, $\sum_{i=1}^{K} \mu_i = 1$ and $0 \leq \mu_i \leq 1$ for all $i \in \{1, ..., K\}$. An example of sampling from a 3-dimensional Dirichlet Distribution is shown in Fig. 2.3. It is possible to observe that, because of the summation constraint, the distribution over the space of the $\{\mu_i\}$ is confined to a simplex of dimensionality $K - 1$.

It is straightforward to demonstrate that the Dirichlet Distribution is the conjugate prior for the Multinomial distribution. Recalling the Bayes'theorem, Eq. 2.2, by multiplying the prior, Eq. 2.21, for the likelihood function, Eq. 2.17, the posterior distribution for $\{\mu_i\}$ takes the form:

$$p(\boldsymbol{\mu}|D,\boldsymbol{\alpha}) \propto p(D|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{i=1}^{K} \mu_i^{\alpha_i+m_i-1} \tag{2.22}$$

which in turn is a Dirichlet Distribution. It is also possible to determine the normalization coefficient:

$$p(\boldsymbol{\mu}|D,\boldsymbol{\alpha}) = Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}+\boldsymbol{m}) = \frac{\Gamma(\sum_{i=1}^{K}\alpha_i + N)}{\prod_{i=1}^{K}\Gamma(\alpha_i+m_i)} \prod_{i=1}^{K} \mu_i^{\alpha_i+m_i-1} \tag{2.23}$$

As in the case of the Beta distribution, the parameters $\{\alpha_i\}$ can be interpreted as an effective number of observations of $x_i = 1$.

2.2 Dirichlet Process

The Dirichlet Process (DP) is a family of stochastic processes whose realizations are probability distributions. It was formally introduced by Thomas S. Ferguson in 1973 [17]. It is a distribution over distributions, i.e. from a Dirichlet Process it is possible to sample probability distributions as real numbers can be sampled from a probability distribution. The Dirichlet Process is characterized by a base distribution, which plays the same role of the expectation value in probability distribution (i.e. probability distributions are sampled around the base distribution as numbers are sampled around the probability distribution), and a concentration parameter α (or scaling parameter), which is a positive real number that controls the sampling discretization. For $\alpha \to 0$, the realizations are concentrated around a single value, while for $\alpha \to \infty$ they are continuous.

The name is from the Dirichlet Distribution, which is the conjugate prior for the categorical distribution, while the Dirichlet Process is the infinite dimensional extension of the Dirichlet distribution and the conjugate prior for infinite, non-parametric discrete distribution. The theory exposed in this section is mainly from [18] and [19].

2.2.1 Formal definition

The Dirichlet Process can be formally (not constructively) introduced using the language of sets in the following way. Let's consider a random distribution G. For G to be distributed according to a DP, its marginal distributions must be distributed according to a Dirichlet distribution (see par. 2.1.5). Let H be a distribution over Θ (a probability space) and α be a positive real number. So, since G is random, for any finite measurable partition $A_1, ..., A_r$ of Θ, the vector $(G(A_1), ..., G(A_r))$ is random. G is distributed according to a Dirichlet process with base distribution H and concentration parameter α if

$$(G(A_1), ..., G(A_r)) \sim Dir(\alpha H(A_1), ..., \alpha H(A_r)) \tag{2.24}$$

for every finite measurable partition $A_1, ..., A_r$ of Θ.

The roles of H and α can be interpreted as follows: H is basically the expectation value of the DP, while the concentration parameter can be viewed as an inverse variance. The larger α is, the smaller the variance, so that the DP will concentrate more of its mass around the mean. These can be written as

$$\mathbb{E}[G(A)] = H(A) \tag{2.25}$$

$$var[G(A)] = \frac{H(A)(1 - H(A))}{(\alpha + 1)} \tag{2.26}$$

Posterior distribution

Let $G \sim DP(\alpha, H)$. It means that G is distributed according to a Dirichlet Process with base distribution H and concentration parameter α. Since G is a random distribution, it is possible to draw samples from G itself. Let $\theta_1, ..., \theta_n$ be an independent draws sequence from G (since G is a distribution over Θ, the θ_i's assume values in Θ), $\theta_i | G \sim G$. It is possible to compute the posterior distribution of G given observed values of $\theta_1, ..., \theta_n$. Let $n_k = \# \{i : \theta_i \in A_k\}$ be the number of observed values in A_k. We can write

$$(G(A_1), ..., G(A_r)) | \theta_1, ..., \theta_n \sim Dir(\alpha H(A_1) + n_1, ..., \alpha H(A_r) + n_r) \tag{2.27}$$

The posterior distribution over G is a DP as well because the equation above holds for all finite measurable partitions of Θ. Using some little algebra, it turns out that the posterior DP has updated concentration parameter $\alpha + n$ and a base distribution $\frac{\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}}{\alpha + n}$, where δ_{θ_i} is a point mass located at θ_i and $n_k = \sum_{i=1}^{n} \delta_{\theta_i}(A_k)$. The posterior distribution can be rewritten as

$$G | \theta_1, ..., \theta_n \sim DP\left(\alpha + n, \frac{\alpha}{\alpha + n} H + \frac{n}{\alpha + n} \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{n}\right) \tag{2.28}$$

From the above equation, it is possible to state that the posterior base distribution is a weighted average between the prior base distribution H and the empirical distribution $\frac{\sum_{i=1}^{n} \delta_{\theta_i}}{n}$. The prior distribution has a weight proportional to α, while the empirical one has a weight proportional to the number of observations n. So, α can be interpreted as a "strength" associated to the prior. Moreover, Eq. 2.28 gives a key consistency property of the DP, i.e. the posterior DP approaches the underlying 'true' distribution. In fact, taking $\alpha \to 0$, the prior becomes non-informative, so that the predictive distribution is just given by the empirical one. On the other hand, with a large number of observations, $n >> \alpha$, the posterior is dominated by the empirical distribution, that is a close approximation of the true underlying distribution [18].

It is also possible to compute the predictive distribution for θ_{n+1}, conditioned on $\theta_1, ..., \theta_n$ and with G marginalized out. Since $\theta_{n+1}|\theta_1, ..., \theta_n \sim G$, for a measurable $A \subset \Theta$

$$P(\theta_{n+1} \in A|\theta_1, ..., \theta_n) = \mathbb{E}[G(A)|\theta_1, ..., \theta_n] = \frac{1}{\alpha + n}\left(\alpha H(A) + \sum_{i=1}^{n} \delta_{\theta_i}(A)\right) \tag{2.29}$$

where the last step follows from the posterior base distribution of G given the first n observations. By marginalizing out G:

$$\theta_{n+1}|\theta_1, ..., \theta_n \sim \frac{1}{\alpha + n}\left(\alpha H + \sum_{i=1}^{n} \delta_{\theta_i}\right) \tag{2.30}$$

So, as a result, the posterior base distribution given $\theta_1, ..., \theta_n$ is also the predictive distribution of θ_{n+1} [18].

The Dirichlet Process can be equivalently visualized by means of different metaphores, which could facilitate its understanding.

2.2.2 Pólya's (Blackwell-MacQueen) urn scheme

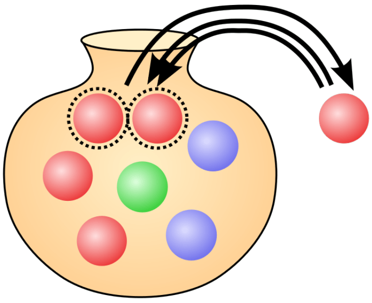


Figure 2.4: Schematic of the Pólya's urn scheme. Figure from http://www.enterprisegarage.io/2017/03/a-mathematical-model-for-innovation/.

The Pólya's urn scheme is a statistical model used as a mental framework to interpret several processes. It is a useful metaphor for the interpretation of Eq. 2.30, as suggested by Blackwell and MacQueen in 1973 [20].

Let's imagine that each value in Θ is a unique color. Draws $\theta \sim G$ are balls with the drawn value being the color of the ball. Let's take an empty urn where the previously seen balls will be placed. Let's imagine to draw a color from H (draw $\theta_1 \sim H$), paint a ball with that color and drop it into the urn. At step $n+1$ you will either pick a new color ($\theta_{n+1} \sim H$), with probability $\frac{\alpha}{\alpha+n}$, paint a ball with that color and drop it into the urn, or reach into the urn and pick a random ball out (draw θ_{n+1} from the empirical distribution) with probability $\frac{n}{\alpha+n}$, paint a new ball with the same color and drop both balls back into the urn, as depicted in Fig. 2.4.

The Blackwell-MacQueen urn scheme can be used to demonstrate the existence of the Dirichlet Process [18]. By starting from the conditional distributions defined in Eq. 2.30, it is possible to construct a distribution over sequences $\theta_1, \theta_2, ...$ by iteratively drawing each θ_i, given $\theta_1, ..., \theta_{i-1}$. For $n \geq 1$ let

$$P(\theta_1, ..., \theta_n) = \prod_{i=1}^{n} P(\theta_i | \theta_1, ..., \theta_{i-1}) \tag{2.31}$$

be the joint distribution over the first n observations, where the conditional distributions are given by Eq. 2.30. This random sequence is clearly infinitely exchangeable. De Finetti's theorem states that for any infinitely exchangeable sequence $\theta_1, \theta_2, ...$ there exists a random distribution G such that the sequence is composed of i.i.d. draws from it:

$$P(\theta_1, ..., \theta_n) = \int \prod_{i=1}^{n} G(\theta_i) dP(G) \tag{2.32}$$

According to the previous setting, the prior over the random distribution $P(G)$ is precisely the Dirichlet Process $DP(\alpha, H)$, so that its existence is established [18].

One of the main property of the predictive distribution 2.30 is that it is a discrete distribution. With positive probability draws from G will take on the same value, regardless of smoothness of H. This implies that the distribution G itself is discrete.

2.2.3 Stick-breaking construction

The formal definition of Dirichlet Process presented in par. 2.2.1 does not provide a mechanism for sampling from Dirichlet Processes. However, as previously mentioned, it is possible to exploit the property that draws from a Dirichlet Process are made of a weighted sum of point masses in order to provide a constructive way of forming G. This method was developed by Sethuraman in 1994 and it is called stick-breaking construction [21]. It is given as follows:

$$\beta_k \sim Beta(1, \alpha) \qquad \theta_k^* \sim H$$

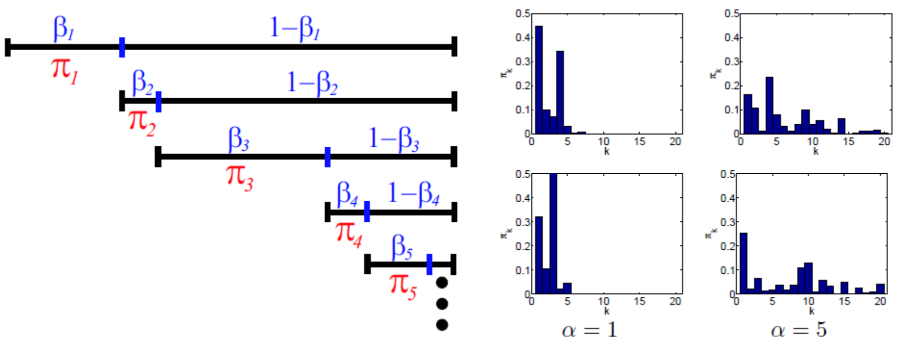


Figure 2.5: Left: Schematic representation of the stick-breaking process. Right: The first 20 weights generated by four random stick–breaking constructions, two with $\alpha = 1$ and two with $\alpha = 5$. Figure from [19].

$$\pi_k = \beta_k \prod_{l=1}^{k-1}(1-\beta_l) \qquad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \tag{2.33}$$

where β_k are called stick-breaking proportions. As a consequence, $G \sim DP(\alpha, H)$. The stick-breaking process is shown in Fig. 2.5. Let's imagine to start with a stick of length 1. Let's break it at β_1, and be π_1 the length of the broken off stick. Then, break the remaining stick at β_2, with π_2 being the length of the broken part. Recursively break the remaining portion, obtaining π_3, π_4 and so on. The stick-breaking distribution over π is called Griffiths-Engen-McCloskey distribution, $\pi \sim GEM(\alpha)$.

This representation of the Dirichlet Process provides another interpretation of the concentration parameter α [19]. From Beta distribution theory, it is known that

$$\mathbb{E}[\beta_k] = \frac{1}{1+\alpha} \tag{2.34}$$

It means that, for small α, the first mixture components are assigned to the majority of the probability mass, while for $\alpha \to \infty$, samples $G \sim DP(\alpha, H)$ approach the base distribution H by assigning small and quite uniform weights to a densely sampled set of discrete parameters $\{\theta_k\}_{k=1}^{\infty}$.

Other stick-breaking processes that sample the proportions β_k from different distributions, such as Poisson-Dirichlet or Pitman-Yor, have been proposed, but they will not be presented in this work.

2.2.4 Chinese Restaurant Process

Eq. 2.30 does not imply only a discreteness property, as already mentioned, but also a clustering property [18]. Usually, DP is used for clustering via the so-called DP mixture models, described in the next section. This clustering effect can be efficiently understood by mean of the metaphor of the Chinese Restaurant Process (CRP), that is going to be introduced.

Let $\theta_1, ..., \theta_n$ be our observations, $\theta_1^*, ..., \theta_m^*$ be the unique values among them and n_k be the number of repeats of θ_k^*. By assigning observations θ_i to distinct values θ_k^*, the Dirichlet Process is implicitly partitioning the data. The predictive distribution can be written as:

$$\theta_{n+1}|\theta_1, ..., \theta_n \sim \frac{1}{\alpha + n}\left(\alpha H + \sum_{k=1}^{m} n_k \delta_{\theta_k^*}\right) \tag{2.35}$$

From the above equation it is possible to notice that the value θ_k^* will be repeated by θ_{n+1} with probability proportional to n_k, i.e. the number of times it has already been observed. The larger n_k, the higher the probability it will increase: this property is called rich-gets-richer phenomenon and it can be deduced also from the Pólya's urn scheme (see par. 2.2.2). Consequently, large clusters grow larger faster.

As previously mentioned, the clustering property of the DP induces partitions. The unique values of $\theta_1, ..., \theta_n$ induce a partitioning of the set $1, ..., n$ into clusters: in cluster k the θ_i's assume the same value θ_k^*. Since $\theta_1, ..., \theta_n$ are random, the partition induced on $1, ..., n$ is in turn random. It condenses all the properties of the DP and it is possible to notice it by inverting the generative process. Starting from the distribution over partitions, it is possible to reconstruct the joint distribution of Eq. 2.31 over $\theta_1, ..., \theta_n$ by first drawing a random partition on $1, ..., n$, then for each cluster k in the partition draw a $\theta_k^* \sim H$, and finally assign $\theta_i = \theta_k^*$ for each i in cluster k [18].

The distribution over partitions is called the Chinese Restaurant Process.Let's imagine to be in a Chinese restaurant with an infinite number of tables, each of which can seat an infinite number of customers. The restaurant's infinite set of tables can be identified with clusters, while customers with observations. Each table is served a different and independently chosen dish θ_k. Let's imagine the first customer entering the restaurant and sitting at the first table. The second customer enters and decides either to sit with the first one, or by herself at a new table. When the $n + 1$st customer enters the restaurant, he can either join an already occupied table k with probability proportional to the number n_k of customers already sitting here, or sits at an unoccupied table with probability proportional to α. So, a new observation can belong to an already existing cluster with probability proportional to the number of observations populating that cluster, or belong to a new cluster without previous observations. After n customers sat down, tables, which correspond to clusters in the metaphor, define a partition of $1, ..., n$ with the distribution over partitions being the same as the one above. Round tables are an important aspect of the CRP: it does not define only a distribution over partitions, but also a distribution over permutations of $1, ..., n$, where each table corresponds to a cycle

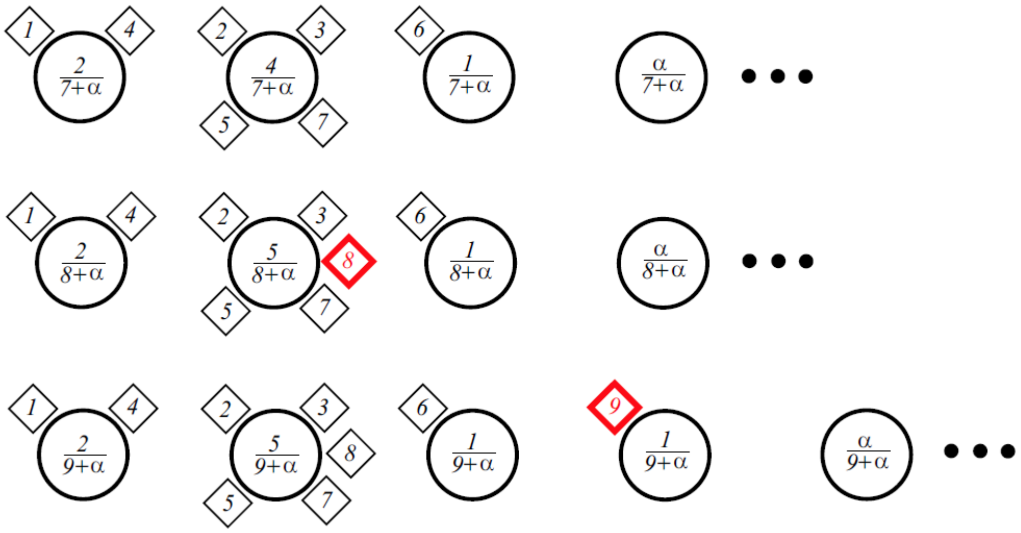


Figure 2.6: Schematic representation of the Chinese Restaurant Process. Circles are tables (clusters) and diamonds are customers (observations). The first line shows occupied tables with probability that next customer (8^{th}) will sit here. Also an unoccupied table is represented. The second line shows the new customer sitting at the most populated table, while third line shows the new customer (9^{th}) occupying a new table. Figure from [19].

of permutation [18].

To conclude, let's consider observation θ_i, with $i \geq 1$. It can assume a new value with probability $\frac{\alpha}{\alpha+i-1}$, independently from the number of clusters among previous θ's. So, the mean and the variance of the number of clusters m are given by:

$$\mathbb{E}[m|n] = \sum_{i=1}^{n} \frac{\alpha}{\alpha+i-1} = \alpha(\psi(\alpha+n) - \psi(\alpha)) \simeq \alpha \log\left(1+\frac{n}{\alpha}\right) \tag{2.36}$$

$$var[m|n] = \alpha(\psi(\alpha+n) - \psi(\alpha)) + \alpha^2(\psi'(\alpha+n) - \psi'(\alpha)) \simeq \alpha \log\left(1+\frac{n}{\alpha}\right) \tag{2.37}$$

where $\psi(\cdot)$ is the digamma function, defined as

$$\psi(\alpha) = \frac{\partial}{\partial\alpha} \log\Gamma(\alpha) \tag{2.38}$$

So, the number of clusters m grows logarithmically with the number of observations n. This is expected because of the rich-gets-richer phenomenon. It is more likely to observe large clusters, so that m has to be smaller than the number of observations n. Moreover, α directly controls the number of clusters: from the above equation a larger α implies a larger number of clusters a priori.

2.2.5 Dirichlet Process mixture models

It is the main application of DP in clustering task. A mixture model is a probabilistic model for representing the presence of subpopulations within an overall population, without requiring that an observed dataset should identify the subpopulation to which an individual observation belongs. They are used to make statistical inferences about the properties of the subpopulations, given only observations on the pooled population. A finite (infinite) mixture model assumes that the data come from a mixture of a finite (infinite) number of distributions.

As shown in par. 2.2.1, DP leads to a posterior distribution with simple and explicit form, Eq. 2.28. Since DP assigns probability one to discrete measures, a DP prior expects multiple observations to take identical values. As a consequence, DP is too restrictive to directly model continuous observations [19]. Mixture models are proposed to address this issue.

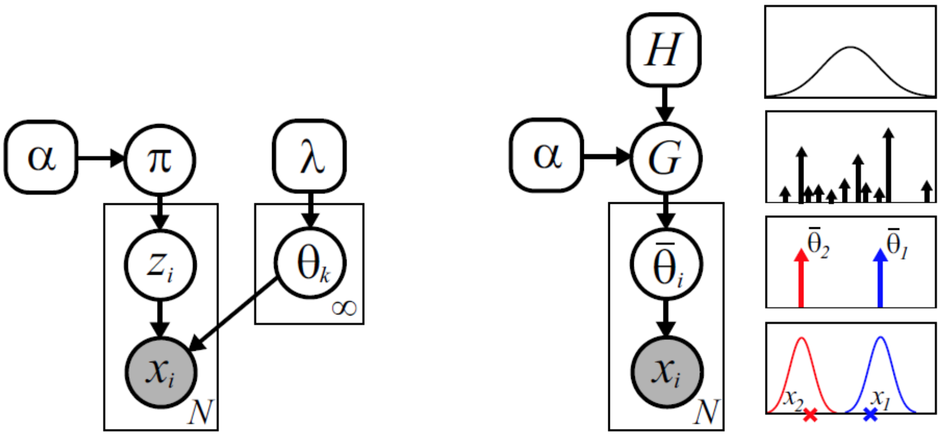


Figure 2.7: Graphical representation of an infinite DP mixture model. Left: representation using indicator variables. Right: distributional form. $\bar{\theta}_i$'s are the parameters of the cluster that generates $x_i \sim F(\bar{\theta}_i)$. Here the example is an infinite Gaussian mixture, with known cluster variances (bottom) and $H(\lambda)$ is a Gaussian prior on cluster means (top). Sampled cluster means $\bar{\theta}_1, \bar{\theta}_2$ and corresponding Gaussians are shown for two observations x_1, x_2. Figure from [19].

Let's model a set of observations $\{x_1, ..., x_n\}$ using a set of latent parameters $\{\theta_1, ..., \theta_n\}$. Each θ_i is drawn independently and identically from G (G is the unknown distribution over parameters modelled using a DP), while each x_i has a distribution $F(\theta_i)$ parametrized by θ_i:

$$x_i|\theta_i \sim F(\theta_i)$$

$$\theta_i|G \sim G$$

$$G|\alpha, H \sim DP(\alpha, H) \tag{2.39}$$

As G is discrete, multiple θ_i's can assume the same value simultaneously, and the model can be seen as a mixture model where the x_i's with the same value of θ_i belong to the same cluster. The stick-breaking construction can be used to represent the mixture perspective. Let z_i be a cluster assignment variable, which can take on value k with probability π_k. So, Eq. 2.39 can be equivalently expressed as:

$$\pi|\alpha \sim GEM(\alpha) \qquad \theta_k^*|H \sim H$$

$$z_i|\pi \sim Mult(\pi) \qquad x_i|z_i, \{\theta_k^*\} \sim F(\theta_{z_i^*}) \tag{2.40}$$

where $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$ and $\theta_i = \theta_{z_i}^*$. π is the mixing proportion, θ_k^* are the cluster parameters, $F(\theta_k^*)$ is the distribution over data in cluster k, and H is the prior over cluster parameters. A schematic of a DP mixture model sampling is shown in Fig. 2.7.

It is important to notice that the DP mixture model is an infinite mixture model, even if only a relatively small number of clusters are used to model the data a priori (due to the fact that the π_k's decrease quickly, Fig. 2.5). Moreover, in the DP mixture model, the number of clusters used to model data is not fixed (as in finite mixture models) and can be automatically inferred from data exploiting the Bayesian posterior inference framework [18]. This is one of the most important property of DP mixture models.

2.3 Hierarchical Dirichlet Process

The Hierarchical Dirichlet Process (HDP) is a generalization of the Dirichlet Process for grouped data. The basic idea is that each individual dataset can be modelled using a separate Dirichlet Process, but the base measure itself is also a Dirichlet Process [19]. It was proposed by Y. W. Teh and M. I. Jordan [22] as a nonparametric Bayesian approach to model grouped data, where each group is associated with a mixture model, and where these mixture models are linked.

Let consider J sets of observations and denote with x_{ji} the i-th observation of j-th group, and with θ_{ji} the associated latent parameter. For each group $j = 1, ..., n$ with data $x_{j1}, ..., x_{jN}$ there is a separate DP, G_j, generating the required parameters θ_{ji}. So, we have an indexed collection of processes $\{G_j\}$, defined over a common probability space. The hierarchical process links these random measures from a probabilistic point of view, allowing them to share the same base distribution. Assuming that base distribution to be random, we have

$$G_0|\gamma, H \sim DP(\gamma, H)$$

$$G_j|\alpha_0, G_0 \sim DP(\alpha_0, G_0)$$

$$\theta_{ji}|G_j \sim G_j$$
$$x_{ji}|\theta_{ji} \sim F(\theta_{ji}) \tag{2.41}$$

where the global measure G_0 is the "mother" of the different G_j and its base distribution H is the prior distribution of the parameters θ_{ji}. Thus, it is possible to share the point masses between the random measures G_j, as each of them inherits the set of atoms from the same "mother" process.

So, the hyperparameters of the HDP consist of the baseline probability measure H and the concentration parameters γ and α_0. The distribution G_0 varies around the prior H, and the amount of variability governed by γ, while the distribution G_j over the parameters in the j-th group deviates from G_0, with the amount of variability governed by α_0. It is possible to use a separate concentration parameter α_j for each group j if the variability is different for each group [22]. An example of a HDP mixture model is represented in Fig. 2.8.

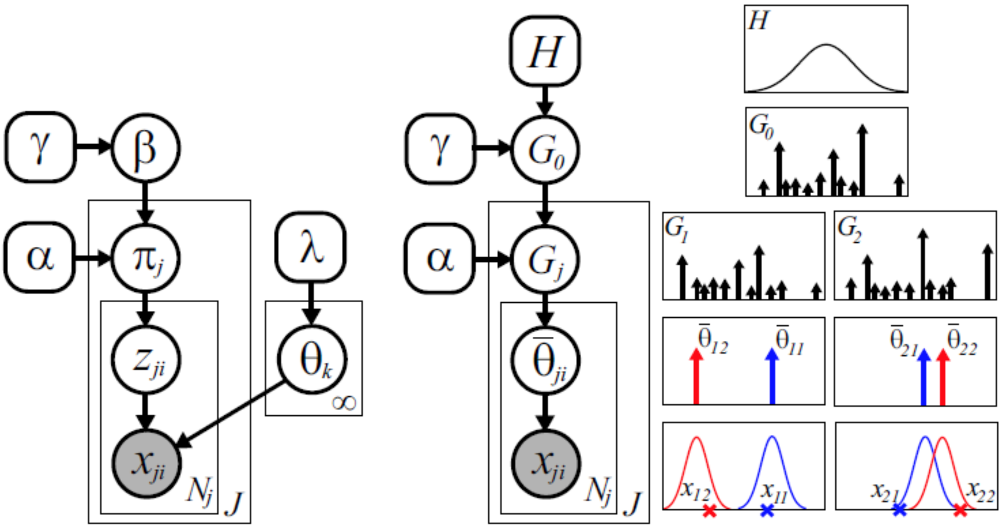


Figure 2.8: Graphical representation of the HDP mixture model. Left: stick-breaking representation. Right: distributional form. The example is a Gaussian mixture. Figure from [19].

2.3.1 Stick-breaking construction

We can recover the stick-breaking construction to represent the HDP mixture model:

$$\boldsymbol{\beta}|\gamma \sim GEM(\gamma) \qquad \theta_k^*|H \sim H(\lambda)$$

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k^*} \tag{2.42}$$

where $\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty}$. The same procedure can be applied to the group-specific mixture distributions G_j, which are independently sampled from a DP with discrete base measure G_0:

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k^*} \tag{2.43}$$

Let's make a parallel with DP. As each θ_{ji} is distributed according to G_j, it takes on the value θ_k^* with probability π_{jk}. If z_{ji} is a variable that indicates $\theta_{ji} = \theta_{z_{ji}}^*$. So, given z_{ji}, it follows that $x_{ji} \sim F(\theta_{z_{ji}}^*)$. The HDP can be represented by:

$$\boldsymbol{\pi_j} | \alpha_0, \boldsymbol{\beta} \sim DP(\alpha_0, \boldsymbol{\beta}) \qquad z_{ji} | \boldsymbol{\pi_j} \sim \boldsymbol{\pi_j}$$

$$\theta_k^* | H \sim H \qquad x_{ji} | z_{ji}, (\theta_k^*)_{k=1}^{\infty} \sim F\theta_{z_{ji}}^* \tag{2.44}$$

where $\boldsymbol{\pi_j} = (\pi_{jk})_{k=1}^{\infty}$. By recalling the stick-breaking construction for DP, the variables β_k are defined as

$$\beta_k^{'} \sim Beta(1, \gamma) \qquad \beta_k = \beta_k^{'} \prod_{l=1}^{k-1} (1 - \beta_l^{'}) \tag{2.45}$$

and the random probability measure $\boldsymbol{\pi_j}$ is given by

$$\pi_{jk}^{'} = Beta\left(\alpha_j \beta_k, \alpha \left(1 - \sum_{l=1}^{k} \beta_l\right)\right) \qquad \pi_{jk} = \pi_{jk}^{'} \prod_{l=1}^{k-1} (1 - \pi_{jl}^{'}) \tag{2.46}$$

that is the stick-breaking construction for HDP [22].

2.3.2 Chinese Restaurant Franchise

By extending the metaphor presented in par. 2.2.4, it is possible to formulate the HDP in terms of a Chinese Restaurant Franchise (CRF), Fig. 2.9, introduced by Y. W. Teh and M. I. Jordan [22]. Let imagine having a franchise to which an infinite number of Chinese restaurants belong. Each separate restaurant j represents a group where customers (observations) x_{ji} can sit at tables (clusters) in the same way as in the CRP, and this happens independently in each restaurant. Each table shares a single dish (parameter) θ_k, that can be ordered from a menu G_0 common to all the restaurants. Variable $\tilde{\theta}_{jt}$ indicates dish served at table t in restaurant j. Note that each x_{ji} is associated with one $\tilde{\theta}_{jt}$, while each $\tilde{\theta}_{jt}$ is associated with one θ_k. Let's introduce indicators to denote these associations and facilitate the understanding: be t_{ji} the index of $\tilde{\theta}_{jt}$ associated with x_{ji}, while k_{jt} being the index of θ_k associated with $\tilde{\theta}_{jt}$. In the CRF, customer i in restaurant j sat at table t_{ji} while table t in restaurant j serves dish k_{jt}. A notation for counts of customers and tables is also needed. Let n_{jtk} denote the number of customers in restaurant j at table t eating dish k, and m_{jk} being the number of tables in restaurant j serving dish k.

In each restaurant, customers sit following the above process, defining a partitioning described by the conditional distribution

$$x_{ji}|x_{j1}, ..., x_{ji-1}, \alpha_0, G_0 \sim \sum_{t=1}^{m_j} \frac{n_{jt}}{\alpha_0 + i - 1}\delta_{\tilde{\theta}_{jt}} + \frac{\alpha_0}{\alpha_0 + i - 1}G_0 \tag{2.47}$$

where m_j is the number of occupied tables in restaurant j and n_{jt} the number of customers in restaurant j at table t. A draw from this mixture can be obtained by drawing from the terms of the right-hand side with probabilities given by the corresponding mixing proportions. If a term in the first summation is chosen, then we set $x_{ji} = \tilde{\theta}_{jt}$ and let $t_{ji} = t$ for the chosen t. If the second term is chosen, then m_j is incremented by one, and we draw $\tilde{\theta}_{jm_j} \sim G_0$ and set $x_{ji} = \tilde{\theta}_{jm_j}$ and $t_{ji} = m_j$. So, an observation can belong to an already existing cluster or can be part of a new cluster, without previous observations.

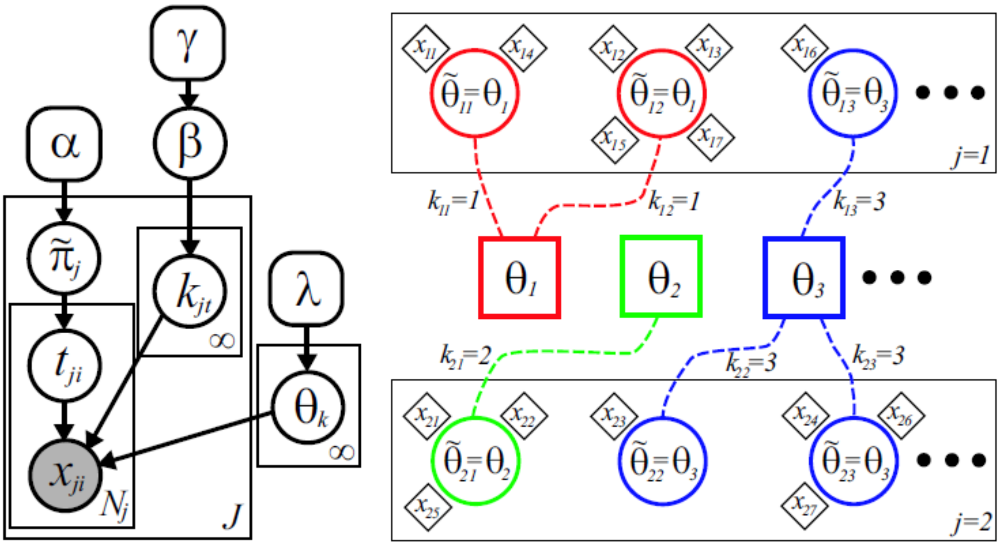


Figure 2.9: Schematic of the Chinese Restaurant Franchise. Left: representation with cluster parameters. Right: example with two restaurants with shared menu with dishes θ_k. There are tables with the same dish in both restaurants. Figure from [19].

The second Dirichlet Process defines how a new dish is chosen once customer in the restaurant j sat down at a new table t^{new}. He either can choose an already ordered dish θ_k by m_k tables of other restaurants with probability proportional to m_k, or he can order a new dish $\theta_k^{new} \sim H$ with probability proportional to γ. So, when a new cluster is generated, it can have same parameters of an already existing cluster of another group or can be different. The second process does complete the description of the partitioning induced by the CRP in each datasets, allowing clusters to be shared, defining the conditional distribution of the variable $\tilde{\theta}_{jt}$:

$$\tilde{\theta}_{jt}|\tilde{\theta}_{11}, ..., \tilde{\theta}_{jt-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_k}{m+\gamma}\delta_{\theta_k} + \frac{\gamma}{m+\gamma}H \tag{2.48}$$

where m_k is the number of tables serving dish k, m is the total number of occupied tables and K the total number of dishes ordered. If $\tilde{\theta}_{jt}$ is drawn by choosing a term in the summation on the right-hand side of the equation, we set $\tilde{\theta}_{jt} = \theta_k$ and let $k_{jt} = k$ for the chosen k. If the second term is chosen, then K is incremented by one and we draw $\theta_k \sim H$ and set $\tilde{\theta}_{jt} = \theta_k$ and $k_{jt} = K$.

This completes the description of the conditional distributions of the x_{ji} variables. To obtain samples of x_{ji}, this procedure must be followed. For each j and i, first sample x_{ji} using Eq. 2.47. If a new sample from G_0 is needed, Eq. 2.48 has to be used to obtain a new sample $\tilde{\theta}_{jt}$ and set $x_{ji} = \tilde{\theta}_{jt}$ [22].

Chapter 3

Materials and methods

In this chapter, the cohort of Glioblastoma patients and the DNA sequencing platforms are described. Then, a series of bioinformatics tools are introduced together with the pipeline used for the analysis. The cohort of Multiple Myeloma patients is presented. Finally, the *R* language implementation of the Hierarchical Dirichlet Process is shown.

3.1 The cohort of Glioblastoma patients

Glioblastoma (GBM) is the most aggressive form of human brain cancer and it carries a poor prognosis. It represents about the 15% of all the brain tumors [23]. Initial symptoms, like headaches, personality changes and nausea, are non-specific [24]. The causes of a wide variety of cases are unknown, with about the 5% of patients that developed it from a low-grade astrocytoma, another brain tumor. Genetic disorders and exposure to ionizing radiation can be considered risk factors [25].

GBM is diagnosed in patients with an average age of 65 and it occurs predominantly in males. The treatment usually consists in surgery, followed by chemotherapy and radiation therapy. Unfortunately, after treatments, the tumor recurs and the median survival after diagnosis is $12 - 15$ months, up to more than five years (about 5%) [25].

In 2016, the World Health Organization (WHO) [26] divided glioblastomas in *IDH* - wildtype, also known as primary glioblastoma, which accounts for about the 90% of cases, predominant in patients with a median age of 62 years, and in *IDH*-mutant, also called secondary glioblastoma, that is diagnosed in patients with a median age of 44. *IDH*-wildtype is characterized by *de novo* development, extensive necrosis and is predominant in males, while *IDH*-mutant often originates from a precursor lesion, such as astrocytoma, has limited necrosis and is balanced between males and females. The median overall survival after a complete treatment is 15 months for primary glioblastoma and 31 for secondary one.

The cohort under study is composed of 24 patients, who has been diagnosed with Glioblastoma. The data were provided by the *IRCCS Istituto delle Scienze Neurologiche di Bologna*, Italy. They are 17 males and 7 females. The average age is 58 and it ranges from 14 to 69 years old. Detailed information is available at the table in Fig. C.1. This study had been approved by the *AUSL Bologna Ethical Committee* and informed consents were obtained from all participants. DNA were purified by salting out protocol from fresh/frozen specimens.

3.2 MiSeq

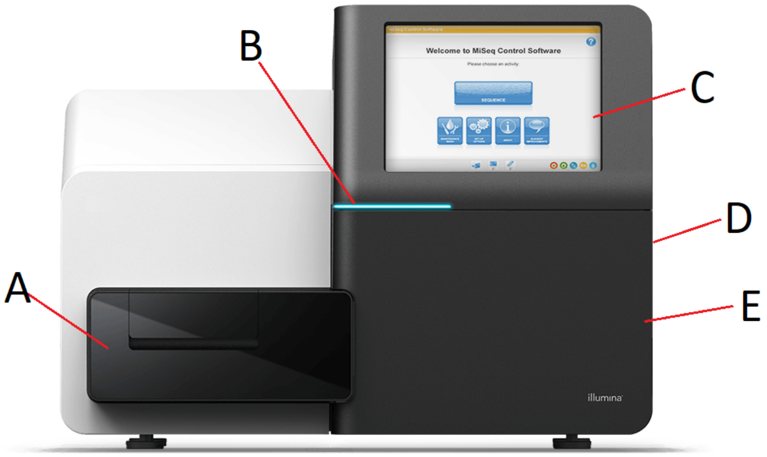


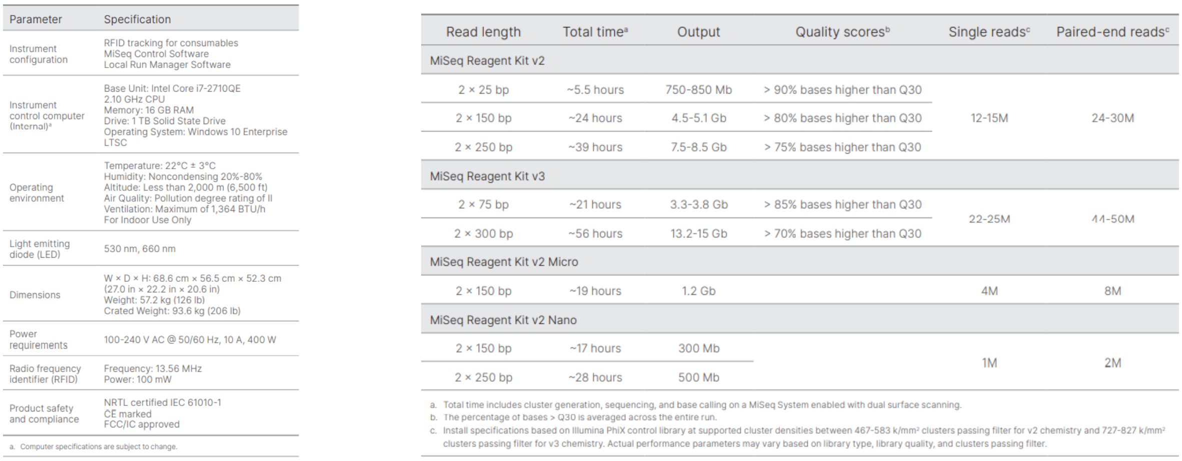
Figure 3.1: Illumina MiSeq System. (A) Flow cell compartment; (B) progress bar of the flow cell; (C) touch screen monitor; (D) power button and USB port; (E) chemical reactants compartment. Figure modified from [27].

MiSeq System [28] is an instrument produced by Illumina, which provides a unique platform that integrates cluster generation, amplification, sequencing and data analysis. In addition, its relatively small size makes it suitable for laboratory environment. It exploits the sequencing by synthesis Illumina NGS (see par. 1.4.2).

The instrument, shown in Fig. 3.1, consists of a touch-screen monitor with an intuitive interface to perform operations, a progress bar which reports the flow cell working status, a USB port to transfer the output files and the compartments that store the flow cell and the chemical reactants. Specifications of the MiSeq System are reported in the table in Fig. 3.2a.

It is interesting to look at the performance parameters, Fig. 3.2b. The quality of the results is exceptional: at least the 70% of the bases have a quality score higher than 30.

It means that for at least the 70% of the called bases, the probability to have made an error is lower than 1 over 1000.



Parameter	Specification
Instrument configuration	RFID tracking for consumables MiSeq Control Software Local Run Manager Software
Instrument control computer (Internal)[a]	Base Unit: Intel Core i7-2710QE 2.10 GHz CPU Memory: 16 GB RAM Drive: 1 TB Solid State Drive Operating System: Windows 10 Enterprise LTSC
Operating environment	Temperature: 22°C ± 3°C Humidity: Noncondensing 20%-80% Altitude: Less than 2,000 m (6,500 ft) Air Quality: Pollution degree rating of II Ventilation: Maximum of 1,364 BTU/h For Indoor Use Only
Light emitting diode (LED)	530 nm, 660 nm
Dimensions	W × D × H: 68.6 cm × 56.5 cm × 52.3 cm (27.0 in × 22.2 in × 20.6 in) Weight: 57.2 kg (126 lb) Crated Weight: 93.6 kg (206 lb)
Power requirements	100-240 V AC @ 50/60 Hz, 10 A, 400 W
Radio frequency identifier (RFID)	Frequency: 13.56 MHz Power: 100 mW
Product safety and compliance	NRTL certified IEC 61010-1 CE marked FCC/IC approved

a. Computer specifications are subject to change.

(a)

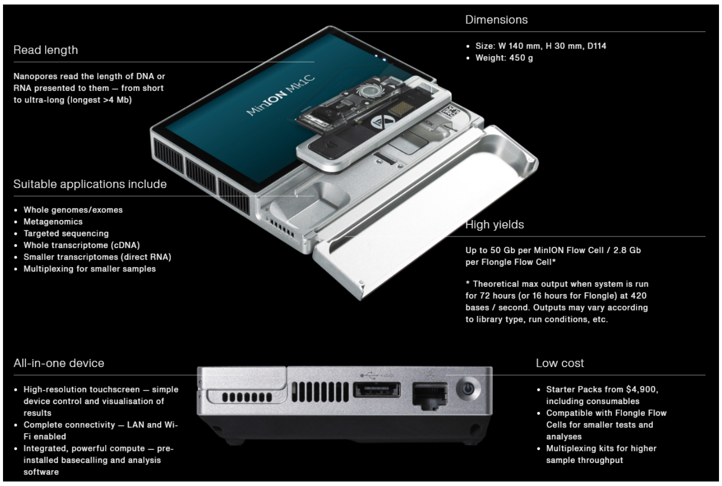
Read length	Total time[a]	Output	Quality scores[b]	Single reads[c]	Paired-end reads[c]
MiSeq Reagent Kit v2					
2 × 25 bp	~5.5 hours	750-850 Mb	> 90% bases higher than Q30	12-15M	24-30M
2 × 150 bp	~24 hours	4.5-5.1 Gb	> 80% bases higher than Q30		
2 × 250 bp	~39 hours	7.5-8.5 Gb	> 75% bases higher than Q30		
MiSeq Reagent Kit v3					
2 × 75 bp	~21 hours	3.3-3.8 Gb	> 85% bases higher than Q30	22-25M	44-50M
2 × 300 bp	~56 hours	13.2-15 Gb	> 70% bases higher than Q30		
MiSeq Reagent Kit v2 Micro					
2 × 150 bp	~19 hours	1.2 Gb		4M	8M
MiSeq Reagent Kit v2 Nano					
2 × 150 bp	~17 hours	300 Mb		1M	2M
2 × 250 bp	~28 hours	500 Mb			

a. Total time includes cluster generation, sequencing, and base calling on a MiSeq System enabled with dual surface scanning.
b. The percentage of bases > Q30 is averaged across the entire run.
c. Install specifications based on Illumina PhiX control library at supported cluster densities between 467-583 k/mm² clusters passing filter for v2 chemistry and 727-827 k/mm² clusters passing filter for v3 chemistry. Actual performance parameters may vary based on library type, library quality, and clusters passing filter.

(b)

Figure 3.2: (a) MiSeq System specifications; (b) MiSeq System performance parameters. Tables from [28].

3.3 Oxford Nanopore Technologies MinIon Mk1c



Read length

Nanopores read the length of DNA or RNA presented to them — from short to ultra-long (longest >4 Mb)

Suitable applications include

- Whole genomes/exomes
- Metagenomics
- Targeted sequencing
- Whole transcriptome (cDNA)
- Smaller transcriptomes (direct RNA)
- Multiplexing for smaller samples

Dimensions

- Size: W 140 mm, H 30 mm, D114
- Weight: 450 g

High yields

Up to 50 Gb per MinION Flow Cell / 2.8 Gb per Flongle Flow Cell*

* Theoretical max output when system is run for 72 hours (or 16 hours for Flongle) at 420 bases / second. Outputs may vary according to library type, run conditions, etc.

All-in-one device

- High-resolution touchscreen — simple device control and visualisation of results
- Complete connectivity — LAN and Wi-Fi enabled
- Integrated, powerful compute — pre-installed basecalling and analysis software

Low cost

- Starter Packs from $4,900, including consumables
- Compatible with Flongle Flow Cells for smaller tests and analyses
- Multiplexing kits for higher sample throughput

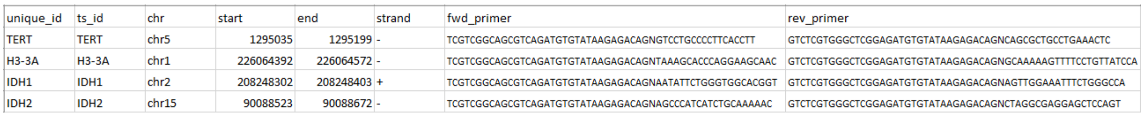
Figure 3.3: Technical information of the MinION Mk1c used for DNA sequencing. Figure from [29].

MinIon Mk1c [30] is a portable instrument produced by Oxford Nanopore Technologies which exploits the nanopore sequencing technique (see par. 1.5). It is characterized by small size and weight and relatively low costs (< 5000$) if compared to MiSeq.

The device is shown in Fig. 3.3. It consists in a high-resolution touchscreen to completely control the instrument and visualize the results, the flow cell compartment, an high-capacity SSD to store the data and LAN and wi-fi connectivity to upload and share data in real-time. It is compatible with flow cells with 126 and 512 active channels. An integrated base calling and analysis software is also present. Differently from Miseq, it allows for long read, high yield and real time sequencing. On the other hand, the base quality is lower than Illumina and it difficultly reaches a phred quality score of 20.

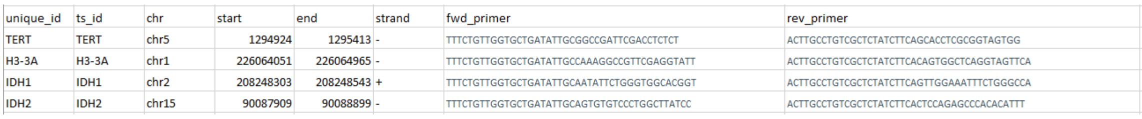
3.4 Sample preparation

Variant analysis was performed using a NGS approach for the following four genes: *IDH1* (exon 4), *IDH2* (exon 4), *H3-3A* (exon 1) and *TERT* (promoter). *IDH1* and *IDH2* are crucial for determining the type of tumor. Tumors without mutations in *IDH1* (usually in R132) often had mutations affecting the analogous amino acid R172 of the *IDH2* gene [31]. Mutations in the *H3-3A*, in G34 and K27 for instance, usually occur in pediatric tumors [32]. *TERT* promoter mutation is commonly associated with *IDH*-wildtype tumors. However, its effect on the prognosis of GBM patients is not well understood [26]. Target regions are reported in the tables of Fig. 3.4 and 3.5 for Illumina and Oxford Nanopore, respectively.



unique_id	ts_id	chr	start	end	strand	fwd_primer	rev_primer
TERT	TERT	chr5	1295035	1295199	-	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNGTCCTGCCCCTTCACCTT	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNCAGCGCTGCCTGAAACTC
H3-3A	H3-3A	chr1	226064392	226064572	-	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNTAAAGCACCCAGGAAGCAAC	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNGCAAAAAGTTTTCCTGTTATCCA
IDH1	IDH1	chr2	208248302	208248403	+	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNAATATTCTGGGTGGCACGGT	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNAGTTGGAAATTTCTGGGCCA
IDH2	IDH2	chr15	90088523	90088672	-	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNAGCCCATCATCTGCAAAAAC	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNCTAGGCGAGGAGCTCCAGT

Figure 3.4: Mutation target considered for the analysis with corresponding loci, forward and reverse primers for Illumina.



unique_id	ts_id	chr	start	end	strand	fwd_primer	rev_primer
TERT	TERT	chr5	1294924	1295413	-	TTTCTGTTGGTGCTGATATTGCGGCCGATTCGACCTCTCT	ACTTGCCTGTCGCTCTATCTTCAGCACCTCGCGGTAGTGG
H3-3A	H3-3A	chr1	226064051	226064965	-	TTTCTGTTGGTGCTGATATTGCCAAAGGCCGTTCGAGGTATT	ACTTGCCTGTCGCTCTATCTTCACAGTGGCTCAGGTAGTTCA
IDH1	IDH1	chr2	208248303	208248543	+	TTTCTGTTGGTGCTGATATTGCAATATTCTGGGTGGCACGGT	ACTTGCCTGTCGCTCTATCTTCAGTTGGAAATTTCTGGGCCA
IDH2	IDH2	chr15	90087909	90088899	-	TTTCTGTTGGTGCTGATATTGCAGTGTGTCCCTGGCTTATCC	ACTTGCCTGTCGCTCTATCTTCACTCCAGAGCCCACACATTT

Figure 3.5: Mutation target considered for the analysis with corresponding loci, forward and reverse primers for Oxford Nanopore.

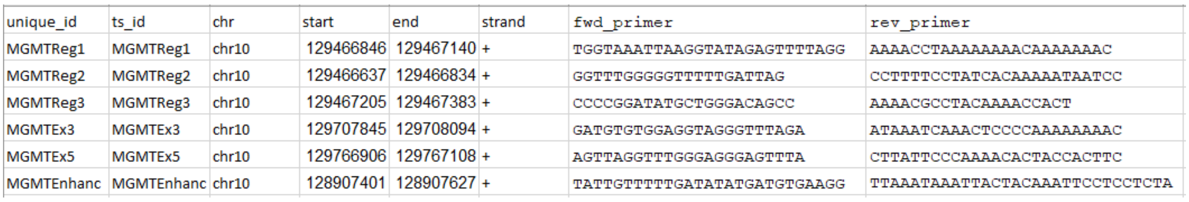
Locus-specific amplicon libraries with tagged primers were generated using overhang adapters based on $Nextera^{TM}$ sequence at the $5'$ for Illumina sequencing; these adapters were recognized by a second round of short PCR step to add Illumina P5/P7 adapters and sample – specific indices to samples. These allow to distinguish DNA belonging to different subjects. This procedure is described in detail in the following section.

Nanopore sequencing followed the same approach, incorporating the adaptors suggested by ONT in the protocol entitled PCR barcoding amplicons (SQK-LSK110).

Amplification products for each sample were mixed together and purified by ampure XT (Agencourt-Beckman Coulter), quantified with the $fluorometerquantus^{TM}$ (Promega) and then employed as template (100 ng) for a second round of PCR (8 cycles) for barcoding. Amplicon products were purified with agencourt Ampure XP beads, quantified with the $fluorometerquantus^{TM}$, pooled and loaded on MiSeq. Each NGS experiment was designed to allocate $\geq 1k$ reads/region, to obtain a depth of coverage $\geq 1000\times$.

With regard to DNA methylation analysis, bisulfite treatment of genomic DNA ($50-500$ ng) was performed using the EZ DNA Methylation-Lightning Kit (Zymo Research Europe, Freiberg, Germany) according to the manufacturer's protocol. DNA methylation was evaluated using targeted bisulfite NGS for *MGMT*.

The epigenetic silencing of the *MGMT* gene by promoter methylation is associated with loss of *MGMT* expression, diminished DNA-repair activity and longer overall survival in patients who receive chemotherapy. In particular, GBM patients received alkylating chemotherapy with temozolomide. Alkylating agents are chemicals that insert alkyl groups in the DNA of rapidly reproducing cells, such as tumor ones, causing their apoptosis. *MGMT* gene encodes a DNA-repair protein which removes alkyl groups from DNA. As a consequence, high levels of *MGMT* activity in cancer cells create a resistant phenotype by blunting the therapeutic effect of alkylating agents and may be a crucial determinant of treatment failure. *MGMT* promoter methylation has a prognostic significance and determination of its status may be an important factor in determining which GBM patients should receive chemoradiotherapy [33]. The enhancer, three regions of the promoter and two exons of *MGMT* have been sequenced in order to evaluate their methylation state, Fig. 3.6.



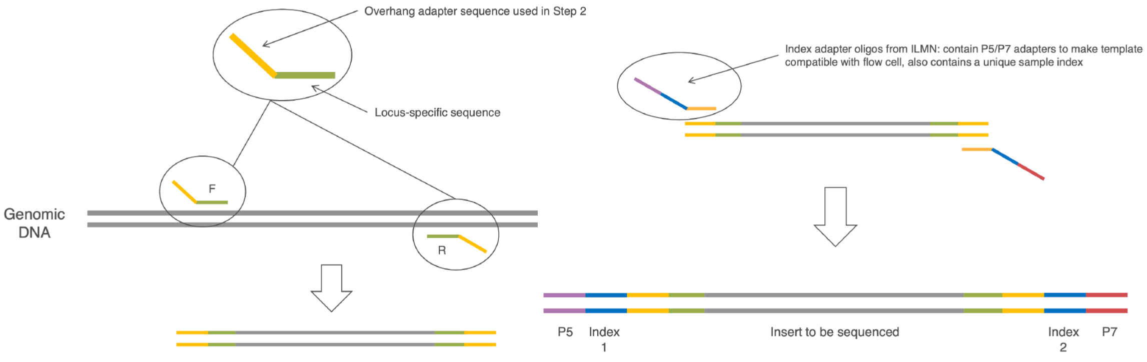
unique_id	ts_id	chr	start	end	strand	fwd_primer	rev_primer
MGMTReg1	MGMTReg1	chr10	129466846	129467140	+	TGGTAAATTAAGGTATAGAGTTTTAGG	AAAACCTAAAAAAAACAAAAAAAC
MGMTReg2	MGMTReg2	chr10	129466637	129466834	+	GGTTTGGGGGTTTTTGATTAG	CCTTTTCCTATCACAAAAATAATCC
MGMTReg3	MGMTReg3	chr10	129467205	129467383	+	CCCCGGATATGCTGGGACAGCC	AAAACGCCTACAAAACCACT
MGMTEx3	MGMTEx3	chr10	129707845	129708094	+	GATGTGTGGAGGTAGGGTTTAGA	ATAAATCAAACTCCCCAAAAAAAAC
MGMTEx5	MGMTEx5	chr10	129766906	129767108	+	AGTTAGGTTTGGGAGGGAGTTTA	CTTATTCCCAAAACACTACCACTTC
MGMTEnhanc	MGMTEnhanc	chr10	128907401	128907627	+	TATTGTTTTTGATATATGATGTGAAGG	TTAAATAAATTACTACAAATTCCTCCTCTA

Figure 3.6: Methylation target considered for the analysis with corresponding loci, forward and reverse primers.

Genomic sequences stored in the Ensembl genome browser (`http://www.ensembl.org/index.html`) were employed as query sequences to identify putative CpG islands in gene promoter regions. MethPrimer (`http://www.urogene.org/cgi-bin/methprimer/methprimer.cgi`) designing was applied to identify CpGs and the best primers of choice. Locus-specific bisulfite amplicon libraries were generated with tagged primers using Phusion U DNA polymerase (ThermoFisher, cod. F555L, Waltham, MA, USA). Libraries were generated using the same approach for mutation analysis with two PCR steps (see par. 3.4.1).

3.4.1 Target enrichment and barcoding

Sequencing the whole genome of each patient is expensive and requires lots of time. Moreover, it is known that there are some regions in the genome which are interesting for the characterization of Glioblastoma, such as $IDH1$ and $IDH2$, for instance. In order to sequence only the regions of interest, a library of amplicons must be created. An amplicon is a fragment of DNA that can be used as a source of an amplification process. To create the library, the following procedure has been performed. It consists in two PCR (see par. 1.4.1) steps. The first PCR is used to amplify the region of interest, the second one to perform the barcoding operation.


Overhang adapter sequence used in Step 2
Locus-specific sequence
F
Genomic DNA
R
Index adapter oligos from ILMN: contain P5/P7 adapters to make template compatible with flow cell, also contains a unique sample index
P5
Index 1
Insert to be sequenced
Index 2
P7

(a)

(b)

Figure 3.7: (a) Target enrichment; (b) barcoding. Figures from [34].

A primer made of about $20 - 25$ bp is designed in order to recognize a specific locus. Two primers are needed: a forward and a reverse primer, Fig. 3.7a. They are also binded to an adapter that would be used in the second step. The primer is used to identify the region and, after that, the amplification through the first PCR begins. It produces $10^6 - 10^9$ copies of the sequence on average. A multiplex PCR is exploited to allow more than one PCR to be carried out at a time to amplify several regions.

After that, sequences from different subjects must be distinguished from each other. This operation is called barcoding and it is depicted in Fig. 3.7b. The used oligonucleotides are made of three parts: a universal primer, a patient-specific barcode (index 1 and 2) and an Illumina barcode ($P5$ and $P7$). The universal primer binds to the locus-specific primer through the adapter. The patient-specific barcode serves to recognize the patient from which the sequence comes. The ILM primer are complementary to the oligonucleotides placed on the flow cell of Illumina so that they can bind. Through a second PCR, with 8 cycles, these composite oligonucleotides are associated to each sample in order to identify it.

3.4.2 Sequencing

Similar procedures have been followed for sequencing with both ILM and ONT.

Before loading the samples on MiSeq, the libraries must be denaturated and diluted. MiSeq micro v2 Reagent Kit – 300 Cycles PE has been used. After the tool maintenance, the sequencing operation is programmed. The type of analysis must be designated, and the samples are named using the patient-specific barcodes. Also, the corresponding ILM barcode are specified. Samples are placed in the cartridge and the flow cell is positioned. The overall analysis lasts about 16 hours.

The paired-end sequencing strategy has been chosen. It allows to sequence both the ends of a fragment. The result is a high-quality and alignable data. Moreover, such a strategy provides more accurate read alignment and the ability to detect indels (insertions-deletions).

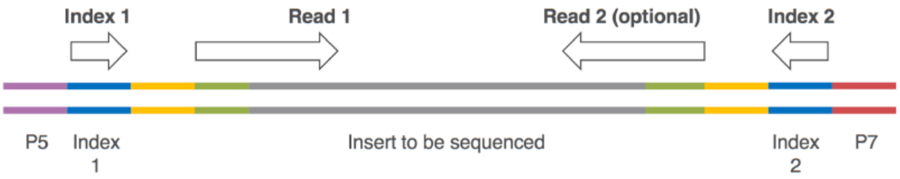


Figure 3.8: Scheme of the sequencing order on MiSeq. Figure from [34].

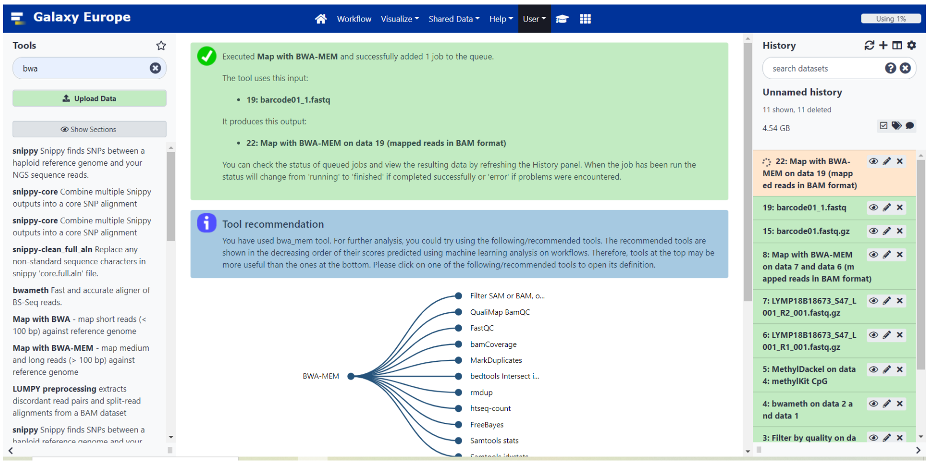
The sequencing order on the MiSeq System, shown in Fig. 3.8, consists in four main steps. Firstly, the amplicon is sequenced in the forward direction (up to 150 bp). Then, the first barcode is read, followed by the second. At this point, the system is able to identify the different samples. Finally, the amplicon is sequenced in the reverse direction (if the chosen strategy is paired-end sequencing).

The procedure followed for ONT sequencing is almost the same. Main differences are in the barcoding, where ONT does not need for $P5$ and $P7$ adapters, and in the sequencing strategy, that in this case was single-end.

3.5 Bioinformatics tools and pipeline

The analysis of FASTQ files (see Appendix A), which contains the final results of the sequencing procedure, requires a bioinformatics pipeline, which is divided in some steps, usually standardized and well-established. Several open-source bioinformatics tools have been used for the analysis of both mutations and methylation in sequence data. They are listed in this paragraph.

3.5.1 Galaxy Europe



snippy Snippy finds SNPs between a haploid reference genome and your NGS sequence reads.

snippy-core Combine multiple Snippy outputs into a core SNP alignment

snippy-core Combine multiple Snippy outputs into a core SNP alignment

snippy-clean_full_aln Replace any non-standard sequence characters in snippy 'core.full.aln' file.

bwameth Fast and accurate aligner of BS-Seq reads.

Map with BWA - map short reads (< 100 bp) against reference genome

Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome

LUMPY preprocessing extracts discordant read pairs and split-read alignments from a BAM dataset

snippy Snippy finds SNPs between a

Executed **Map with BWA-MEM** and successfully added 1 job to the queue.

The tool uses this input:

- **19: barcode01_1.fastq**

It produces this output:

- **22: Map with BWA-MEM on data 19 (mapped reads in BAM format)**

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Tool recommendation

You have used bwa_mem tool. For further analysis, you could try using the following/recommended tools. The recommended tools are shown in the decreasing order of their scores predicted using machine learning analysis on workflows. Therefore, tools at the top may be more useful than the ones at the bottom. Please click on one of the following/recommended tools to open its definition.

Unnamed history

11 shown, 11 deleted

4.54 GB

- 22: Map with BWA-MEM on data 19 (mapped reads in BAM format)
- 19: barcode01_1.fastq
- 15: barcode01.fastq.gz
- 8: Map with BWA-MEM on data 7 and data 6 (mapped reads in BAM format)
- 7: LYMP18B18673_S47_L001_R2_001.fastq.gz
- 6: LYMP18B18673_S47_L001_R1_001.fastq.gz
- 5: MethylDackel on data 4: methylKit CpG
- 4: bwameth on data 2 and data 1
- 3: Filter by quality on da

Figure 3.9: Galaxy Europe interface. On the left, the bioinformatics tools are listed. On centre, the options of the selected tool are reported. On the right, the files uploaded on the Galaxy cloud.

Galaxy Europe [35] is a free-user server, which provides access to several ($\sim$ 2500) scientific tools, covering most of the bioinformatics topics. It provides an intuitive and easy-to-use interface, allowing to search and find the tools one needs. Each function is accompanied with its own description and paper references [36]. The following tools have been used for the fulfilment of this thesis work.

Concatenate datasets tail-to-head (cat) is a tool which concatenates many FASTQ files into a single one. It was used to merge together all the FASTQ files given by MinIon Mk1c. It corresponds to the standard UNIX *cat* command.

FastQC is a tool that perform quality check on the FASTQ files. It has been used before the alignment operation to check whether the sequence data suffered for poor quality score (see Appendix A).

Filter by quality allows to set a lower threshold quality score in order to filter data and remove bad quality reads.

Map with BWA-MEM is a tool which performs mapping and alignment (see Appendix A) of medium and long reads ($>$ 100 bp) against a reference genome (in our case GRCh38/hg38) for mutation data. It is based on the Burrows–Wheeler transform.

Bwameth performs alignment of reads in a bisulfite-sequencing experiment to a reference genome (GRCh38/hg38). It is used for the analysis of methylation data.

Samtools idxstats and *Samtools Depth* require as input the BAM file (see Appendix A) resulting from the alignment operation and returns some basis statistics about it and depth of coverage.

MethylDackel is a tool that processes a coordinate-sorted and indexed BAM file containing BS-seq alignments and extract per-base methylation metrics from them. It requires a reference genome. By default, it only computes metrics for Cytosines in CpG.

3.5.2 Nanogalaxy

Nanopore sequencing requires the development of new bioinformatics approaches to deal with its specific error characteristics. Read mapping and alignment tools are critical building blocks for many such applications. For mapping, reads from nanopore sequencing are particularly challenging due to their higher and non-uniform error profiles [37]. In order to deal with ONT reads, the NanoGalaxy platform [38], an extension of the Galaxy server suitable for nanopore data, was exploited [39]. As a consequence, *Minimap2* has been used to align the input FASTQ files to a reference genome (hg38) and produce the BAM file.

The alignment of methylation data has been performed using *bwameth*. The reason is that available methylation data analysed with ONT were treated by Sodium Bisulfite (like ILM data, even though ONT is able to detect methylated cytosines without the need for BS).

3.5.3 Integrative Genomics Viewer

Integrative Genomics Viewer (IGV) [40] is an open-source visualization tool for the interactive exploration of genomic datasets. It supports flexible integration of all the common types of genomic data and metadata, investigator-generated or publicly available, loaded from local or cloud sources. In Fig. 3.10 is reported a shot from the IGV interface, where methylation data is shown. The BAM and BAI files, which contains the aligned data and coverage information, are loaded. A reference DNA must be chosen (in our case hg38) and it is shown on the bottom of the screen. On the top, the chromosome locus is shown, and a search bar allows to find the gene(s) of interest. The tool detects and highlights where a variant is present. The variant call was set to has at least a 5% of Variant Allele Frequency (VAF).

3.5.4 EPIC TABSAT

EPIC-TABSAT is an easy-to-use tool for the integrated analysis of targeted bisulfite sequencing (TBS) data with array-based studies (e.g. Illumina EPIC). It features an

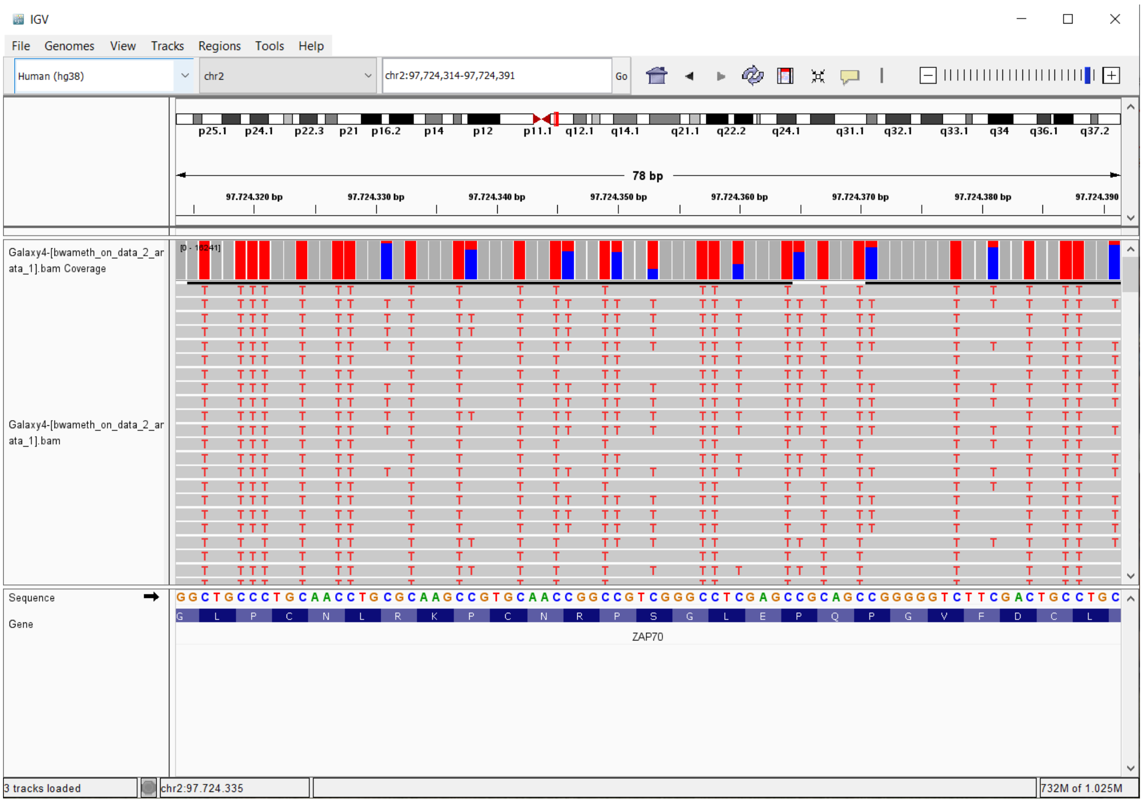

IGV
File Genomes View Tracks Regions Tools Help
Human (hg38)
chr2
chr2:97,724,314-97,724,391
Go
p25.1 p24.1 p22.3 p21 p16.2 p14 p12 p11.1 q12.1 q14.1 q21.1 q22.2 q24.1 q31.1 q32.1 q33.1 q34 q36.1 q37.2
78 bp
97.724.320 bp
97.724.330 bp
97.724.340 bp
97.724.350 bp
97.724.360 bp
97.724.370 bp
97.724.380 bp
97.724.390
Galaxy4-[bwameth_on_data_2_and_data_1].bam Coverage
Galaxy4-[bwameth_on_data_2_and_data_1].bam
Sequence
Gene
GGCTGCCCTGCAACCTGCGCAAGCCGTGCAACCGGCCGTCGGGCCTCGAGCCGCAGCCGGGGGTCTTCGACTGCCTGC
G L P C N L R K P C N R P S G L E P Q P G V F D C L
ZAP70
3 tracks loaded
chr2:97.724.335
732M of 1.025M

Figure 3.10: IGV interface. On the top, the specified chromosome locus is shown. In the middle, aligned data (methylation in this case) is shown together with the coverage. At the bottom, the referece sequence (hg38 in this case) and the amino acid chain are reported.

intuitive user interface, provides a novel approach to study read methylation patterns, and offers an unprecedented way to analyse and interpret TBS data in combination with epigenome wide methylation studies [41]. The EPIC-TABSAT workflow (see Fig. 3.11) covers the complete bisulfite sequence data analysis workflow from quality assessment, methylation calling to interactive result presentation.

This tool has been used to analyse methylation data and produce some statistics. The results from EPIC was compared to the ones obtained through IGV and *MethylDackel* to see whether significant differences are present.

3.5.5 Bioinformatics pipeline

The resulting FASTQ files from the MinION Mk1c device were first concatenated (the device produces files with no more than 400 reads), using *Concatenate datasets tail-to-head (cat)*. Once the unique FASTQ file with the read sequences was available, a quality control was made, exploiting the *FASTQC* tool. Data have been filtered through *Filter by quality*, setting a minimum phred quality score threshold. Then, the alignment

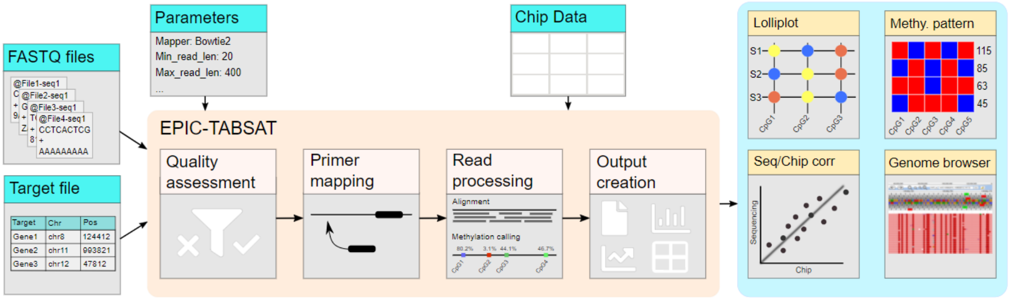


Figure 3.11: Schematic of the workflow of EPIC-TABSAT. Figure from [41].

procedure was performed through the use of the *BAM-MEM* algorithm, which results in the creation of the corresponding BAM file. This file format can be opened in IGV to have an overview about the variants present in some specific regions. Some information about the depth of coverage and statistics can be obtained from the BAM file by using tools such as *Samtools Depth* and *Samtools idxstats*. Finally, variants calling was performed using IGV. By scrolling the screen, highlighted sites have been investigated and annotated. However, variant caller could have been used to fulfil this task. Tools such as *Mutect2 (GATK4)* for ILM and *Medaka* for ONT take a BAM file as input and returns a VCF file, which contains mutation calls (both single nucleotide variants and indels, see Appendix A).

To process and analyse methylation data, a similar pipeline has been used. The differences were basically in the alignment algorithm (here *bwameth* was used) and in the methylation calling tool (*Methyl Dackel* or visually by IGV).

3.6 Agreement analysis

It is interesting to quantify the agreement between the two sequencing methods. Agreement is related to the issue of measuring a variable with two different methods and assessing whether they basically produce the same results. A question such as this must be addressed using appropriate statistical methods. Two possible solutions are proposed in the following subsections.

3.6.1 Bland-Altman plot

The Bland-Altman (BA) plot is a statistical tool, firstly proposed by J. Martin Bland and Douglas G. Altman in 1986 [42], that can be used to compare two different measurement methods, usually the 'gold standard' and a new one, and to assess whether they do agree.

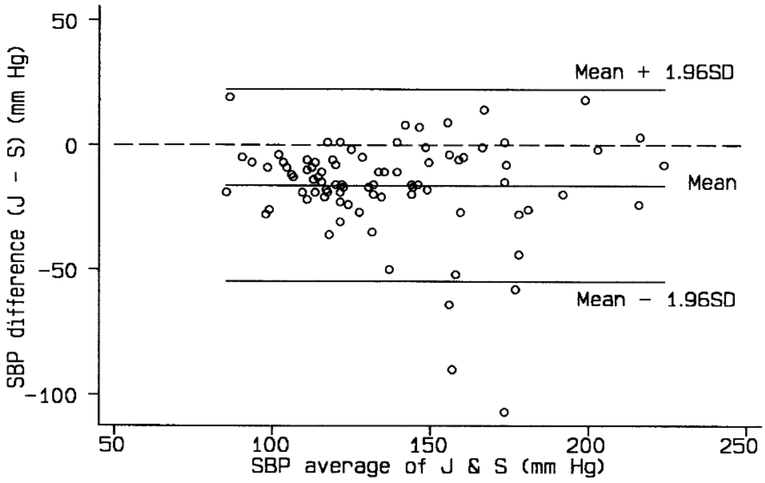


Figure 3.12: Example of BA plot, where the mean difference and the limits of agreement are drawn. Figure from [43].

Let's consider one wants to compare an oxygen saturation monitor and a pulsed oximeter saturation in measuring oxygen saturation, as in the original paper of Bland and Altman. It would be measured in a set of patients, so to have two measurements for each subject, one for each instrument. The BA plot is a graphical representation of the difference between the measurements by the two methods against their mean. An example of BA plot is shown in Fig. 3.12. The mean difference between the two measures and its standard deviation are computed. If the differences are normally distributed, it is expected that 95% of them would lie between ±1.96 standard deviation (they are called limits of agreement). However, if the distribution of the differences is not normal, it could not be a problem as serious as in other contexts [43]. It could happen when the difference and the mean are related, for instance, and in this case some corrective action can be taken.

Here, the BA plot is used to compare and assess the agreement between results from ILM and ONT data, taking the first one as reference.

3.6.2 Concordance Correlation Coefficient

The Concordance Correlation Coefficient (CCC), first proposed by Lawrence Lin in 1989 [44], is a statistical tool aimed to the measurement of the agreement between two variables. It is a sort of 'evolution' of the Pearson's correlation coefficient. The CCC not only evaluates how much the two measurements lie on a straight line, but also how far this line is from the equality line (i. e. $y = x$).

Let consider to have N pairs of measures, taken with two different methods, say

$(x_i,\ y_i),\ i = 1, ..., N$. The CCC is defined as

$$\hat{\rho_c} = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2} = C_b \cdot \rho \tag{3.1}$$

where $\bar{x}$ and $\bar{y}$ are the averages of x_i and y_i, s_x^2 and s_y^2 their variances, s_{xy} the covariance, ρ is the Pearson's correlation coefficient and C_b is the bias correction factor, which evaluates how far the best fit line is far from the equality line, from 1 (no deviation) to 0 (far away). The CCC is a number between -1 (perfect reversed agreement) and 1 (perfect agreement) [44].

3.7 The cohort of MM patients

Multiple myeloma is a blood tumor which mainly affects elderly people. To date, no cures have been found. It is important to study this kind of malignancy, especially from a genetic point of view, to find specific genetic variants and characterize the evolution of this disease from patient to patient.

3.7.1 Multiple Myeloma

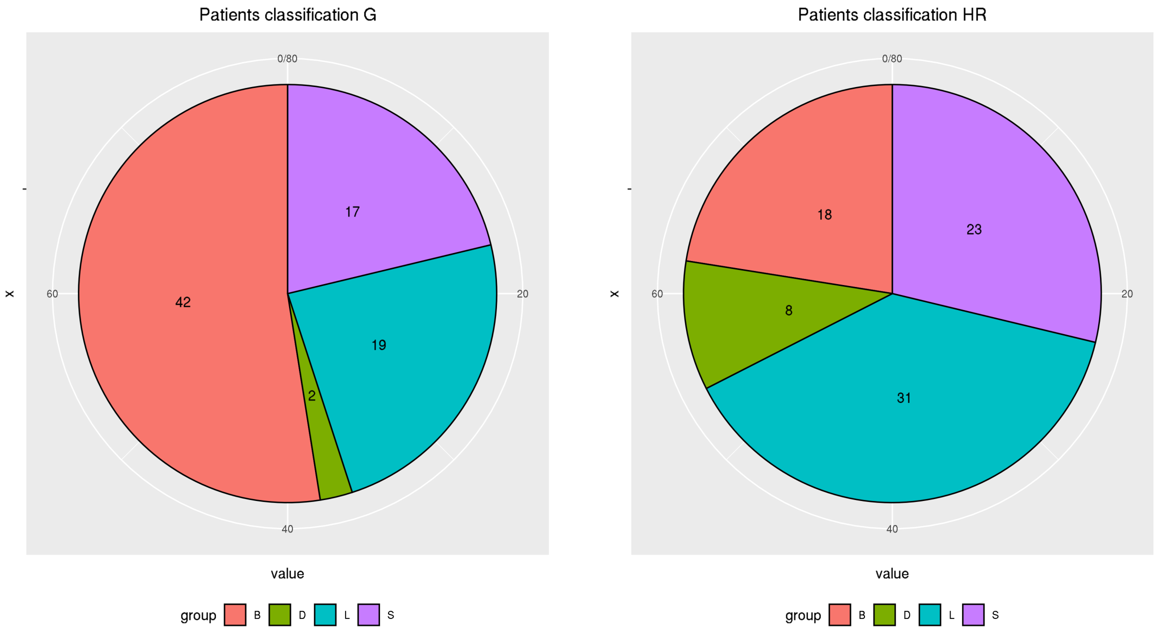
Multiple Myeloma (MM) is a malignancy of plasma cells, which accumulate in bone marrow and overproduce a monoclonal antibody (M proteins) that may thicken the blood and damage the kidneys [45]. The disease is often symptomless for a long time and already advanced at the time of diagnosis. Common symptoms of advanced MM include bone pain, anaemia, frequent infections, and kidney failure. In the past ten years, several novel MM drugs have appeared, which have significantly improved the survival of MM patients. Median overall survival in patients eligible for autologous stem-cell transplantation is estimated to be approximately 10 years, compared with 4-5 years in patients not eligible for transplantation [46]. Despite these recent advances, the vast majority of patients eventually relapse, and the disease remains largely incurable. Mainly affecting elderly individuals (the median age at diagnosis is about 65 [47]), the incidence of MM is expected to increase as a result of the aging of our population. MM is the second most common haematologic malignancy in Europe [48].

The clonal cells suppress normal plasma cell populations, leading to immunosuppression, impaired normal haematopoiesis, lytic bone lesions, and, through a number of different mechanisms, impaired renal function. The clinical outcome of MM is variable, and much of this variability is driven by acquired genetic factors, which immortalize and drive the subsequent progression of the disease. Current knowledge of drivers of disease comes from cytogenetic analyses, which have shown that the genome of MM is diverse and is characterized by structural rearrangements and copy number abnormalities [49]. The cause of MM is still unknown [50]. Risk factors for MM include obesity, chronic inflammation, and exposure to pesticides, organic solvents, or radiation [46].

Multiple myeloma is considered treatable, but generally incurable. Remissions may be brought about with steroids, chemotherapy, targeted therapy, and stem cell transplant. Bisphosphonates and radiation therapy are sometimes used to reduce pain from bone lesions [46].

However, because of its variable clinical outcome, it could be interesting to study the genetic profile of people affected by this disorder and how it changes in time. By measuring the Copy Number of a set of genes, at different time points, it is possible to study the evolutionary pattern of a patient.

3.7.2 Dataset

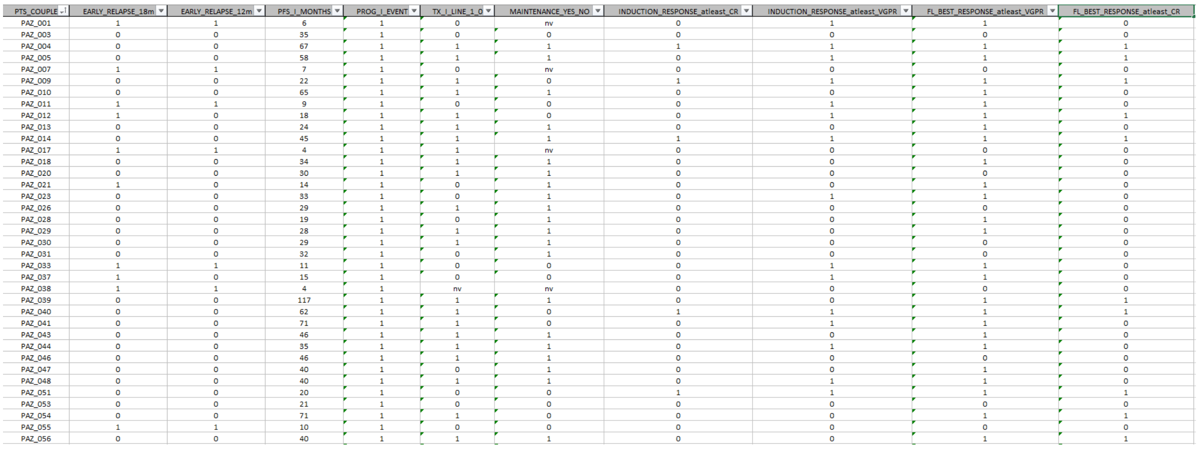


(a) (b)

Figure 3.13: Pie charts of the empirical classifications of the patients. (a) Classification G ; (b) Classification HR.

A cohort of 80 subjects (numbered from 1 to 80) diagnosed with Multiple Myeloma was analysed. For each patient, we have 2409 measures of Copy Number (CN), Fig. 3.13. Each CN refers to one, or more than one, gene (e. g. CN-4 corresponds to *GABRD* and *PRKCZ*, while CN-6 to *SKI*).

The pipeline implemented by the haematologists to obtain Copy Number Variation measurements is described in Appendix B.

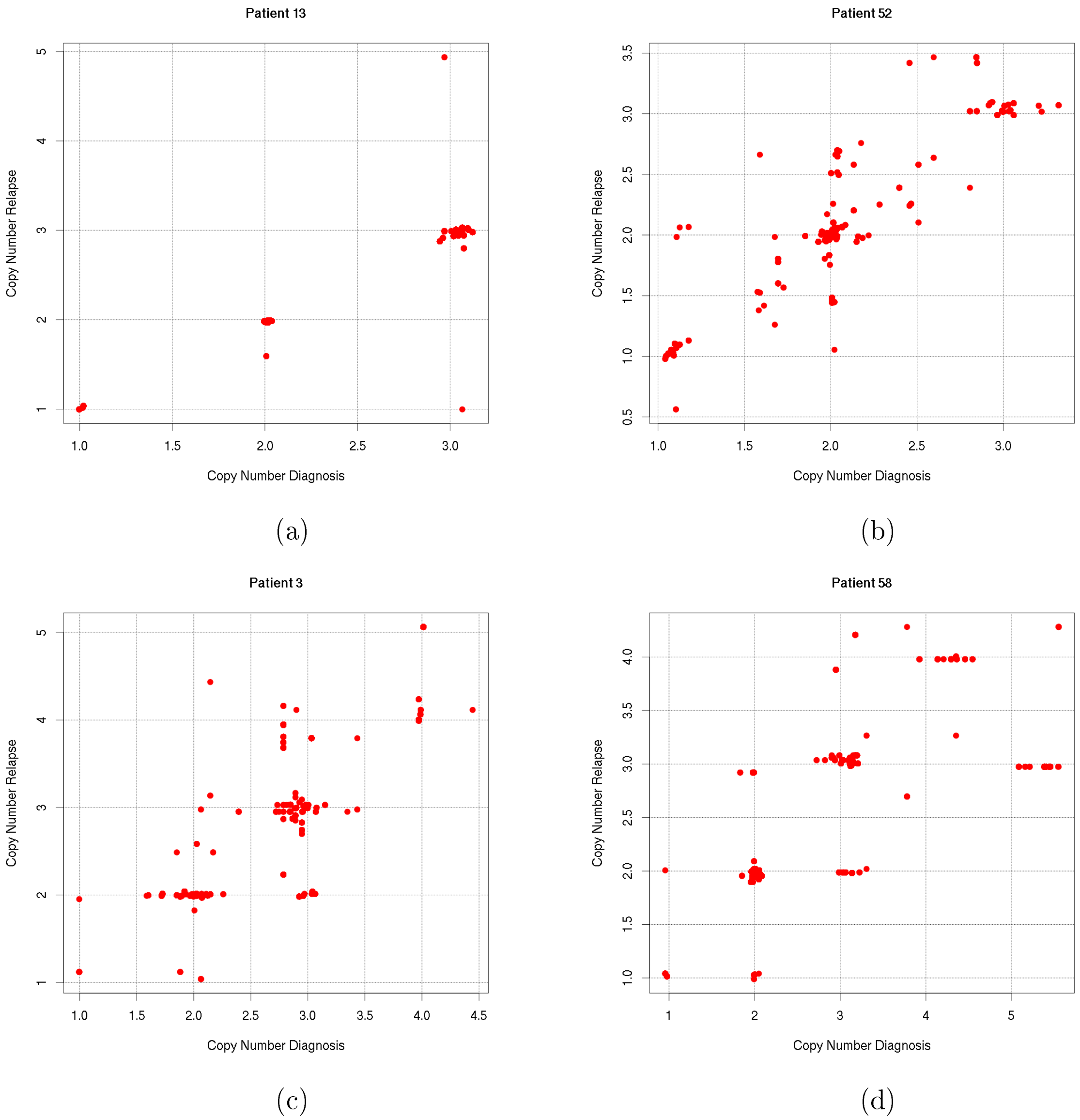


PTS_COUPLE	EARLY_RELAPSE_18m	EARLY_RELAPSE_12m	PFS_I_MONTHS	PROG_I_EVENT	TX_I_LINE_1_0	MAINTENANCE_YES_NO	INDUCTION_RESPONSE_atleast_CR	INDUCTION_RESPONSE_atleast_VGPR	FL_BEST_RESPONSE_atleast_VGPR	FL_BEST_RESPONSE_atleast_CR
PAZ_001	1	1	6	1	0	nv	0	1	1	0
PAZ_003	0	0	35	1	0	0	0	0	0	0
PAZ_004	0	0	67	1	1	1	1	1	1	1
PAZ_005	0	0	58	1	1	1	0	1	1	1
PAZ_007	1	1	7	1	0	nv	0	0	0	0
PAZ_009	0	0	22	1	1	0	1	1	1	1
PAZ_010	0	0	65	1	1	1	0	0	1	0
PAZ_011	1	1	9	1	0	0	0	1	1	0
PAZ_012	1	0	18	1	1	0	0	1	1	1
PAZ_013	0	0	24	1	1	1	0	0	1	0
PAZ_014	0	0	45	1	1	1	1	1	1	1
PAZ_017	1	1	4	1	1	nv	0	0	0	0
PAZ_018	0	0	34	1	1	1	0	0	1	0
PAZ_020	0	0	30	1	1	1	0	0	0	0
PAZ_021	1	0	14	1	0	1	0	0	1	0
PAZ_023	0	0	33	1	0	1	0	1	1	0
PAZ_026	0	0	29	1	1	1	0	0	0	0
PAZ_028	0	0	19	1	0	1	0	0	1	0
PAZ_029	0	0	28	1	1	1	0	0	1	0
PAZ_030	0	0	29	1	1	1	0	0	0	0
PAZ_031	0	0	32	1	0	1	0	0	0	0
PAZ_033	1	1	11	1	0	0	0	1	1	0
PAZ_037	1	0	15	1	0	0	0	1	1	0
PAZ_038	1	1	4	1	nv	nv	0	0	0	0
PAZ_039	0	0	117	1	1	1	0	0	1	1
PAZ_040	0	0	62	1	1	0	1	1	1	1
PAZ_041	0	0	71	1	1	0	0	1	1	0
PAZ_043	0	0	46	1	1	1	0	0	1	0
PAZ_044	0	0	35	1	1	1	0	1	1	0
PAZ_046	0	0	46	1	1	1	0	0	0	0
PAZ_047	0	0	40	1	0	1	0	0	1	0
PAZ_048	0	0	40	1	1	1	0	1	1	0
PAZ_051	0	0	20	1	0	0	1	1	1	1
PAZ_053	0	0	21	1	0	0	0	0	0	0
PAZ_054	0	0	71	1	1	0	0	0	1	1
PAZ_055	1	1	10	1	0	0	0	0	0	0
PAZ_056	0	0	40	1	1	1	0	0	1	1

Figure 3.14: Screenshot from the Copy Number dataset. Each row is a patient, while each column represents a clinical information about the treatment and the response of that patient. From the left: patient, relapse within 18 months, relapse within 12 months, number of months before relapse (progression-free survival), progression of the tumor, transplantation, maintenance, CR to induction, VGPR to induction, VGPR to first line therapy, CR to first line therapy.

Patients underwent different treatments, as shown in Fig. 3.14. The therapy consists in induction (all patients), transplantation, maintenance and consolidation. All of them underwent relapse (progression = 1 for all of them). 45 patients were treated with transplantation while 32 not. Moreover, 37 patients followed a maintenance therapy. All underwent relapse and for each patient the number of months to relapse is provided: 20 subjects are in early relapse, i.e. within 18 months (8 of them within 12 months). Also, the response to the therapy is provided. Induction Response refers to patients' response to induction treatment. The response can be Complete Remission (CR), that is the optimal outcome, or Very Good Partial Remission (VGPR). The criteria and definitions of them were established by the International Myeloma Working Group [51]. Also, response to first line therapy (i. e. the whole therapy one patient followed) is reported. 40 patients resulted into at least Very Good Partial Remission (VGPR) for induction therapy and 12 of them at least Complete Remission (CR), while 58 at least VGPR and 24 at least CR for first line therapy.

For each patient, CNVs were measured at two different time points: at the diagnosis of the disease and after the relapse, and they were named Copy Number Diagnosis (CND) and Copy Number Relapse (CNR), respectively. The plot of CND vs CNR (the evolutionary patterns) are shown in Fig. 3.15. From the figures, it is possible to notice that patients can be grouped according to the evolutionary patterns [52] [53]. There are patients with a stable trajectory (S), as in Fig. 3.15a, characterized by CN pairs concentrated around integer values, while others with linear trajectory (L), Fig. 3.15b,



(a) (b) (c) (d)

Figure 3.15: Examples of patients' evolutionary trajectories. (a) Stable trajectory, patient 13; (b) Linear trajectory, patient 52; (c) Branched trajectory, patient 3; (d) Drifted trajectory, patient 58.

whose CN couples spreads along the main diagonal. Also patients with branched (B) and drifted (D) trajectories are present, Fig. 3.15c and 3.15d, respectively. The firsts present some branches, that is CN pairs with CND ≠ CNR, while the latters have groups of CND with different values that reach the same CNR values. The haematologists, who provided the data, established two different empirical classifications, namely High-Risk (H-R, which considers only a subset of genes known to have a role in tumor onset) and Genomic (G, that takes into account all the genes for which a CN measure is available). A classification consists in assigning a class (S, L, B or D) to each subject, as shown in Tab. 3.1.

Patients	*Class (G)*	*Patients*	*Class (HR)*
1, 3, 4, 5, 6, 7, 8, 10, 11, 16, 17, 18, 19, 20, 21, 22, 25, 27, 28, 32, 35, 37, 38, 42, 47, 48, 49, 50, 53, 55, 56, 58, 59, 64, 65, 67, 70, 71, 73, 74, 75, 79	*B*	1, 3, 4, 10, 17, 22, 25, 27, 37, 53, 54, 55, 58, 64, 65, 73, 75, 79	*B*
2, 9, 15, 26, 29, 30, 33, 34, 36, 41, 44, 45, 51, 52, 54, 66, 77, 78, 80	*L*	5, 6, 11, 15, 16, 18, 19, 20, 21, 23, 28, 32, 33, 36, 38, 41, 44, 45, 47, 48, 49, 51, 52, 56, 66, 67, 68, 71, 74, 78, 80	*L*
12, 13, 14, 23, 24, 31, 39, 40, 43, 46, 60, 61, 62, 63, 68, 69, 72	*S*	2, 8, 9, 12, 13, 14, 24, 26, 29, 31, 34, 35, 40, 43, 46, 60, 61, 62, 63, 69, 70, 72, 77	*S*
57, 76	*D*	7, 30, 39, 42, 50, 57, 59, 76	*D*

Table 3.1: Empirical classification of MM patients provided by the haematologists. The Genomic classification (G) takes into consideration all the analysed genes, while High-Risk classification (HR) considers only some genes which are important for the incoming of the disease.

The aim of this work is to provide a robust and objective classification method, able to reproduce and improve the empirical one, which lacks generalization, by means of using Hierarchical Dirichlet Process. This method would also furnish an approach that could be generalized to larger and different datasets.

3.8 The ‘dirichletprocess’ package in *R*

In order to conduct the analysis and perform Dirichlet clustering, the *R*’s package ‘dirichletprocess’ has been used [54]. *R* is a free software environment for statistical computing and graphics [55]. It is suitable for handling huge amount of data and many statistical techniques are implemented. Moreover, *R* can be easily extended through the use of packages, available online.

The ‘dirichletprocess’ package provides objects and functions that can be used to define Dirichlet Processes and fit them.

3.8.1 Hierarchical multivariate Dirichlet mixture model

The Hierarchical multivariate mixture model is the most widely used non-parametric modelling approach for multivariate data [54]. It is also heavily used in clustering applications. For the unknown parameters, we have

$$\theta = (\boldsymbol{\mu}, \Lambda) \tag{3.2}$$

for d dimensional data, $\boldsymbol{\mu}$ is a column vector of length d and Λ is a $d \times d$ dimensional matrix. The multivariate kernel function is:

$$k(\boldsymbol{y}|\theta) = \frac{|\Lambda|^{1/2}}{2\pi^{-d/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T \Lambda (\boldsymbol{y}-\boldsymbol{\mu})\right) \tag{3.3}$$

For the prior choice, a multivariate normal distribution (N) for $\boldsymbol{\mu}$ and a Wishart distribution (Wi) for Λ are used:

$$G_0(\boldsymbol{\mu}, \Lambda|\boldsymbol{\mu_0}, \kappa_0, \nu_0, T_0) = N(\boldsymbol{\mu}|\boldsymbol{\mu_0}, (\kappa_0\Lambda)^{-1}) Wi_{\nu_0}(\Lambda|T_0) \tag{3.4}$$

where $\boldsymbol{\mu_0}$ is the mean vector of the prior, κ_0, ν_0 are single values and T_0 is a matrix. For clustering of multidimensional data, as in the case of CN measures at diagnosis and relapse, it is common to use a mixture of multivariate normal distributions. Moreover, using mixture of multivariate distributions allows to deal with any continuous data while, for example, mixture of Beta distributions is suitable for data defined on a bounded interval. In addition, hierarchical version can be used to treat grouped data: in this case each patient is a group.

The code implementation of the constructor functions of these objects was done by Giovanni Sighinolfi, a former student of the University of Bologna.

3.8.2 Hierarchical Gaussian Dirichlet mixture model

It is also possible to construct a mixture of Gaussian distributions in order to deal with one dimensional data. For the unknown parameters, here we have

$$\theta = (\mu, \sigma^2) \tag{3.5}$$

The Gaussian kernel function is:

$$k(y_i|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i-\mu)^2}{2\sigma^2}\right) \tag{3.6}$$

The conjugate prior for θ is the Normal-Gamma distribution, with parameters $\gamma = (\mu_0, k_0, \alpha_0, \beta_0)$:

$$G_0(\theta|\gamma) = N\left(\mu|\mu_0, \frac{\sigma^2}{k_0}\right) Inv - Gamma(\sigma^2|\alpha_0, \beta_0) \tag{3.7}$$

where the default setting of the parameters is $\mu_0 = 0$, $\sigma_0^2 = 1$, $\alpha_0 = 1$, $\beta_0 = 1$.

However, hierarchical Gaussian mixture model is not available in the 'dirichletprocess' package, so the functions to construct both the mixing object and the Hierarchical Dirichlet Process had been wrote down.

3.9 Silhouette score

The Silhouette score [56] is a metrics that can be used to assess the goodness of a clustering operation on a certain dataset. Let consider to have N data points that have been clustered through a certain algorithm (e. g. Dirichlet clustering). For any data point i belonging to cluster C_i, it is possible to define

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j) \tag{3.8}$$

where $|C_i|$ is the number of points belonging to cluster C_i and $d(i,j)$ the distance (according to a certain metrics, in this case the simple Euclidean distance) between point i and j both belonging to cluster C_i. The quantity $a(i)$ is a measure of how well i is assigned to its cluster (the smaller the distance from the other points, the better is). We can define also

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j) \tag{3.9}$$

that is a measure of how far is point i belonging to cluster C_i from the other points of different clusters C_k. The Silhouette score of point i can be then defined as

$$s(i) = \begin{cases} \frac{b(i)-a(i)}{max\{a(i),b(i)\}} & if\ |C_i| > 1 \\ 0 & if\ |C_i| = 1 \end{cases} \tag{3.10}$$

As can be easily understood from the above equation, the silhouette score is a number $-1 \leq s(i) \leq 1$. If $s(i)$ is near to 1 it means that the data point has been properly clustered. If $s(i)$ is close to -1, it means that data point i should have been clustered in its neighbouring cluster.

3.10 Logistic regression model

Logistic regression is appropriate for modelling dichotomous variables, which take only two values, representing success and failure. It is used to model the dependency of a binary variable from a set of covariates, which could be either discrete or categorical or continuous [57].

Let Y_i be a binary random variable and y_i its realization. It can assume only two values, 0 and 1. Let π_i be the probability to assume value 1. As a consequence, Y_i follows a Bernoulli distribution (see par. 2.1.2). Now suppose that the units under study could be classified into k groups according to the factors of interest. Let n_i denote the number of observations in group i and y_i to be the number of units which have the attributes of group i. So, y_i can be viewed as a realization of a random variable Y_i which takes the values $0, 1, ..., n_i$. As a consequence, Y_i is binomial with parameters π_i and n_i (see par. 2.1.2).

Then, it would be interesting to find a method that allows to have the probabilities π_i depending on a vector of covariates $\boldsymbol{x_i}$. If the dichotomous variable is described by a linear function of the covariates, the problem is that it would not be guaranteed that the output values will be in the correct range (i. e. [0, 1]). The trick consists in transforming the probability in order to get rid of the range restrictions:

$$odds_i = \frac{\pi_i}{1-\pi_i} \tag{3.11}$$

that is called the odds, the ratio of the probability to its complement (favourable over unfavourable cases). However, odds can take any positive value. So, a second transformation is performed:

$$\eta_i = logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) \tag{3.12}$$

Then, the assumption that the logit of the probability π_i follows a linear model, allows us to introduce the logistic regression model.

Suppose to have k independent observations $y_1, ..., y_k$, each of them being treated as a realization of a random variable Y_i binomial distributed. The logistic regression model is defined as:

$$logit(\pi_i) = \boldsymbol{x_i'\beta} \tag{3.13}$$

where $\boldsymbol{x_i}$ is a vector of covariates and $\boldsymbol{\beta}$ a vector of regression coefficients. The interpretation is as follows: β_j represents the change in the logit of the probability associated with a unit change in the $j-th$ predictor holding all other predictors constant [57]. By exponentiating we get

$$\frac{\pi_i}{1-\pi_i} = exp\{\boldsymbol{x_i'\beta}\} \tag{3.14}$$

where the exponentiated coefficient is called odds ratio. Solving for the probability, it results

$$\pi_i = \frac{exp\{\boldsymbol{x_i'\beta}\}}{1+exp\{\boldsymbol{x_i'\beta}\}} \tag{3.15}$$

The last two operations help in the interpretation of the results in terms of probabilities, rather than logit.

If the output is no more a dichotomous variable, but rather a categorical one, the model must be extended in a multinomial sense. The simplest approach to multinomial data consists in nominating one of the response categories as baseline, calculate the log-odds for all the other categories relative to the baseline and finally let the log-odds be a linear function of the covariates [57].

Let assume the log-odds of each response follow a linear model:

$$\eta_{ij} = log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = \alpha_j + \boldsymbol{x_i'}\boldsymbol{\beta_j} \tag{3.16}$$

where α_j is a constant, J is the number of categories and $\boldsymbol{\beta_j}$ a vector of regression coefficient, $j = 1, ..., J-1$.

The multinomial logistic model is analogous to a logit regression model, where the probability distribution of the response is multinomial (see par. 2.1.4) rather than binomial. Here, moreover, there is a total of $J-1$ equations. Note that it makes no difference which category it is designed as reference (it is always possible to convert from a formulation to another). It is important to be careful in the interpretation of the final results: remember that the coefficients are always computed taking one category as baseline and they must always be referred to it.

The model can also be written in terms of probabilities π_{ij} to have a more 'comfortable' interpretation:

$$\pi_{ij} = \frac{exp\{\eta_{ij}\}}{\sum_{k=1}^{J} exp\{\eta_{ik}\}} \tag{3.17}$$

for $j = 1, ..., J$.

Chapter 4

Results and discussion

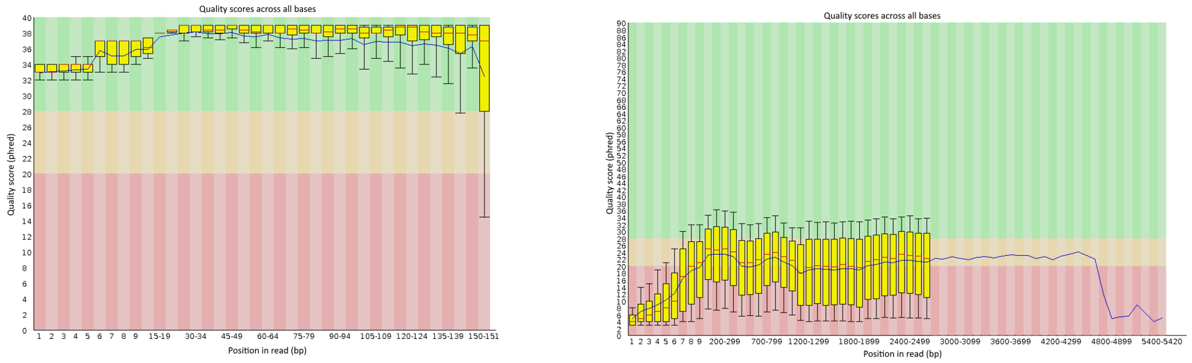
In this chapter, the results obtained from mutation and methylation analyses, both with Illumina and Oxford Nanopore Technologies, are presented, in order to compare these two techniques. Then, the results of the Dirichlet Process clustering are shown. Two different kinds of hierarchical clustering have been made: bivariate and univariate. Finally, these results are discussed in order to find a clinical relevance of them.

4.1 Variant analysis

The data set is composed by 24 patients. Four genes were taken into consideration: *IDH1* ($chr2 : 208.248.303 - 208.248.543$), *IDH2* ($chr15 : 90.087.909 - 90.088.899$), *H3-3A* ($chr1 : 226.064.051 - 226.064.965$) and *TERT* ($chr5 : 1.294.924 - 1.295.413$). For each of them, a region was sequenced both with Illumina and ONT in order to detect variants. It has been possible to sequence longer regions with ONT. The ILM sequence data have been used as validation, given that ILM is considered a gold standard in point mutations. The BAM alignment file have been obtained through *BWA-MEM* for ILM data and *Minimap2* for ONT data.

4.1.1 Quality, alignment and coverage

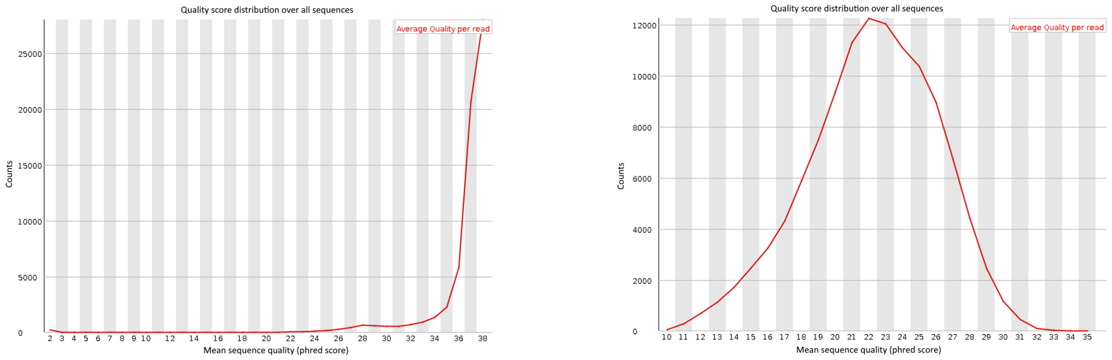
At first, quality results were analysed using *FASTQC*. The quality information for both ILM and ONT is reported in Fig. 4.1. The plots show per-base sequence quality. From the figure, it is clear that ILM results have a higher median phred quality score (> 30) than ONT ones (~ 20). Moreover, by observing the quality score distributions over all sequences in Fig. 4.2, it comes out that many ONT sequences suffer for low quality score, while only few reads from ILM has low quality. However, these results are not surprising: the ONT is expected to have a lower quality with respect to ILM devices, due to the shorter sequencing time and the more susceptibility to errors, because of the overlapping current distributions (see Par. 1.5.5, for instance). The ONT data have been filtered with *Filter by quality*, setting a phred quality score threshold of 10.



(a)

(b)

Figure 4.1: Per-base sequence quality plots for ID2 for (a) ILM and (b) ONT. ILM shows a higher phred quality score on average. The median is always over 30 and sometimes is close to 40, while for ONT in some cases it is below 20.



(a)

(b)

Figure 4.2: Quality score distributions over all sequences for ID2 for (a) ILM and (b) ONT. Many ONT sequences suffer for low quality.

The BAM file created by *BWA-MEM* for ILM data and by *Minimap2* for ONT were visualized by IGV and shown in Fig. 4.3. From the figure it is possible to clearly notice that ONT (top) allowed to sequence a longer region with respect to ILM (bottom). The figure focuses on *H3-3A* region. ONT identifies lots of indels, which are practically absent in ILM results. This could be due to misalignments, which need to be corrected in an additional step. The variant in position $chr1 : 226.064.531$ is detected by both ILM and ONT. The Variant Allele Frequency, that can be obtained by clicking on the rectangular icon corresponding to any position, are similar (100% and 89%, respectively).

Information about the depth of coverage and mapped reads were obtained from the BAM file using *Samtools idxstats* and *Samtools Depth*, and are reported in the heatmaps

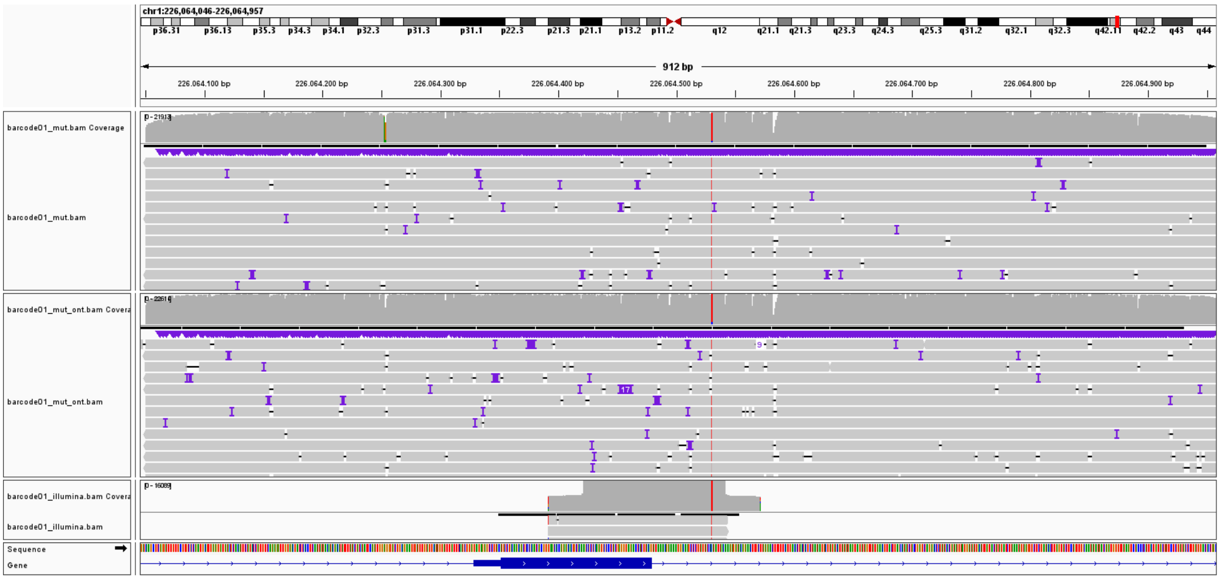
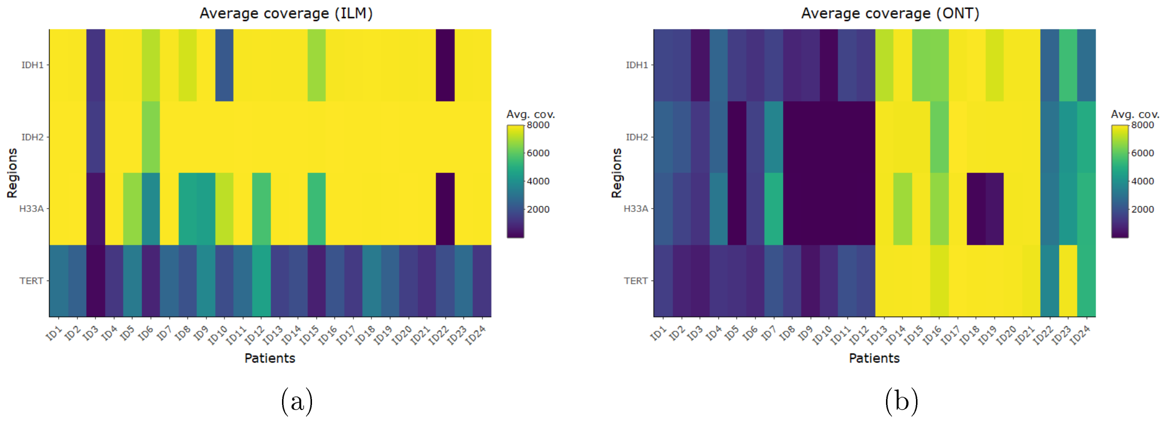


Figure 4.3: BAM files visualized by IGV. Region of *H3-3A* for ID2 is shown for ONT aligned with *BWA-MEM* (top), with *Minimap2* (middle) and ILM (bottom). It can be noticed that ONT allows to sequence a longer region. Moreover, *Minimap2* reduces the amount of false variant calls. Many indels are called by ONT (violet marks on the top) with respect to ILM.

in Fig. 4.4. The coverage is relatively high for both ILM and ONT, considering that ONT data lost about 65% of reads after quality filtering. However, ONT data of 6 patients show a poor coverage for *IDH2* and *H3-3A* ($\sim$ 30 reads). So, variants called in these regions for those 6 patients were not considered reliable.



(a) (b)

Figure 4.4: Heatmap of the mapping performance for mutation data for each target sequenced through (a) Illumina and (b) Oxford Nanopore.

4.1.2 Variant calling

The resulting BAM files have been opened by IGV and analysed by eye, in order to detect point mutations. IGV has been set to highlight only those mutations with a VAF $\geq 5\%$. The point mutations found by ILM are reported in Tab. C.1. Variants detected by ONT sequence data are shown in Tab. C.2 and C.3.

SNP locus	*Reference*	*ILM*	*ONT*	*Annotation*	*Frequency*
$chr2 : 208.248.388$	C	T	T	$p.R132H$	4/24
$chr2 : 208.248.389$	G	A	A	$p.R132C$	1/24
$chr2 : 208.248.468$	G	G	A	$p.G105G$	1/24

Table 4.1: Point mutations detected in the region of *IDH1*. Mutations in *R132* is revealed by both ONT and ILM. From the left: locus of the SNP, base in the reference genome, base called from ILM data, base called from ONT data, variant annotation (according to the standard nomenclature [58]) and frequency of the variant among the dataset.

SNP locus	*Reference*	*ILM*	*ONT*	*Annotation*	*Frequency*
$chr15 : 90.088.015$	G	na	A	$g.90088015G > A$	11/24
$chr15 : 90.088.023$	G	na	A	$g.90088023G > A$	6/24
$chr15 : 90.088.247$	G	na	A	$g.90088247G > A$	11/24
$chr15 : 90.088.345$	C	na	T	$g.90088345C > T$	17/24

Table 4.2: Point mutations detected in the region of *IDH2*. The mutations found by ONT are in regions not available (na) for ILM.

IDH1, Tab. 4.1, is characterized by $p.R132H$ mutation, that occurs in 4 patients. $p.R132C$ is observed both by ILM and ONT, while $p.G105G$, has been found only by ONT. Read lengths are 182 bp and 241 bp for ILM and ONT, respectively.

ONT detected several point mutations in gene *IDH2*, all in non-coding regions. These mutations have been not found by ILM due to the fact that were in regions not covered by ILM. In fact, ILM read length is 182 bp, while ONT has 991 bp. The results are reported in Tab. 4.2. Variants in non-coding region mainly consist in $C > T$ and $G > A$ substitutions. Many indels are called, especially by ONT.

SNP locus	*Reference*	*ILM*	*ONT*	*Annotation*	*Frequency*
$chr1 : 226.064.531$	C	T	T	$g.226064531C > T$	24/24

Table 4.3: Point mutation detected in the region of *H3-3A*. It is a known single nucleotide variant present in most of Caucasian people.

For what concerns *H3-3A*, Tab. 4.3, there is only one variant which is called both by ILM and ONT, that is a $C > T$ substitution in $chr1 : 226.064.531$, always with high

VAF (100% with ILM and 90% with ONT). It has been found in all the patients of the dataset and it is a single nucleotide variant present in most of the Caucasian people. Even here, ILM read length is about 181 bp, while ONT has 915 bp.

$SNP\ locus$	$Reference$	ILM	ONT	$Annotation$	$Frequency$
$chr5 : 1.295.052$	T	T	C	$g.1295052T > C$	24/24
$chr5 : 1.295.113$	G	A	A	$g.1295113G > A$	15/24
$chr5 : 1.295.135$	G	G	A	$g.1295135G > A$	1/24
$chr5 : 1.295.234$	A	na	G	$g.1295234A > G$	16/24
$chr5 : 1.295.338$	C	na	T	$g.1295338C > T$	2/24

Table 4.4: Point mutations detected in the region of *TERT*. Only one mutation has been found by both ILM and ONT.

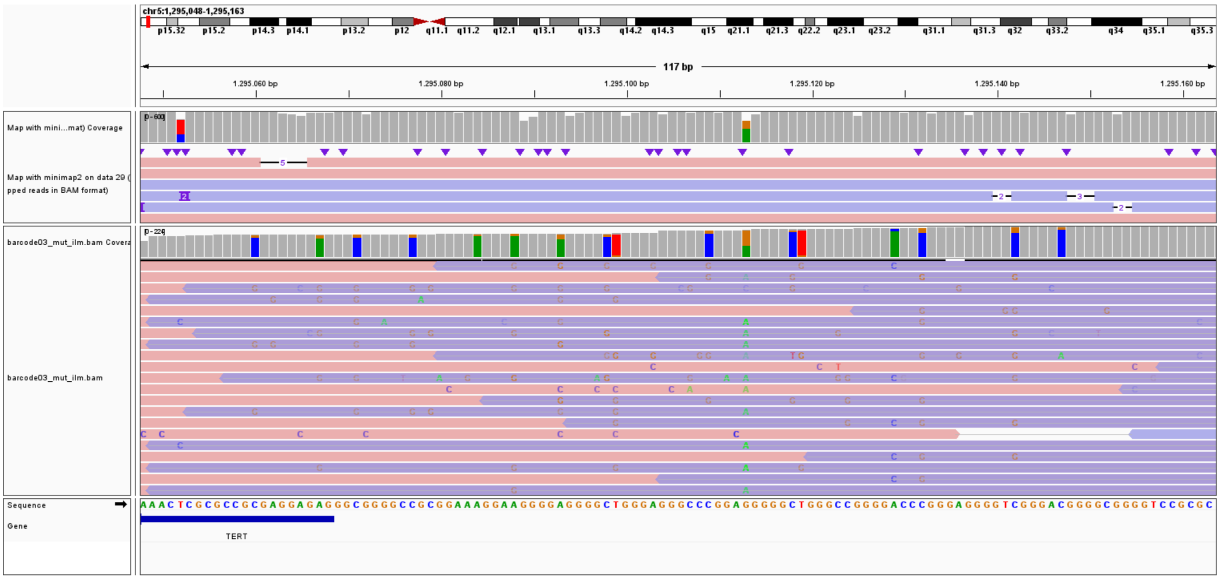


Figure 4.5: Screenshot from IGV. It is possible to see all the (artifact) mutations found in data from ILM (bottom) with respect to ONT (top) for *TERT*. Forward reads are colored in red while reverse ones in blue.

Point mutations found in *TERT* are reported in Tab. 4.4. A recurrent variant, found by both the techniques, is $G > A$ substitution in $chr5 : 1.295.113$. ONT detected also $T > C$ and $A > G$ substitutions. Here, several variants called by ILM are not shown. By looking at the IGV screen, Fig. 4.5, it is possible to observe that ILM mainly calls many variants. In particular, ILM calls lot of $A > G$ substitutions, not detected by ONT. Indeed, for *TERT*, read length is 165 bp for ILM and 490 for ONT.

4.2 Methylation analysis

The methylation of CpG regions was studied for gene *MGMT*. In particular, six regions have been taken into consideration, Fig. 3.6. They have been called Reg1 ($chr10$: 129.466.846 – 129.467.140), Reg2 ($chr10$: 129.466.637 – 129.466.834), Reg3 ($chr10$: 129.467.205 – 129.467.383), Ex3 ($chr10$: 129.707.845 – 129.708.094), Ex5 ($chr10$: 129.766.906 – 129.767.108) and Enhanc ($chr10$: 128.907.401 – 128.907.627). Enhanc is the *MGMT* enhancer; Reg1, Reg2 and Reg3 are regions of the promoter, while Ex3 and Ex5 are exons. For each region, the percentage of methylation of the CpG islands was studied. Sequence results for Reg2, Reg3, Ex3 and Enhanc are available for each patient for both Illumina and ONT. Reg1 is available for both ONT and ILM for only 10 patients, while Ex5 is available only for ONT.

4.2.1 Alignment and coverage

The alignment of the methylation data obtained through Illumina has been made by *bwameth*, an extension of the BWA algorithm, suitable for treating Bisulfite sequences.

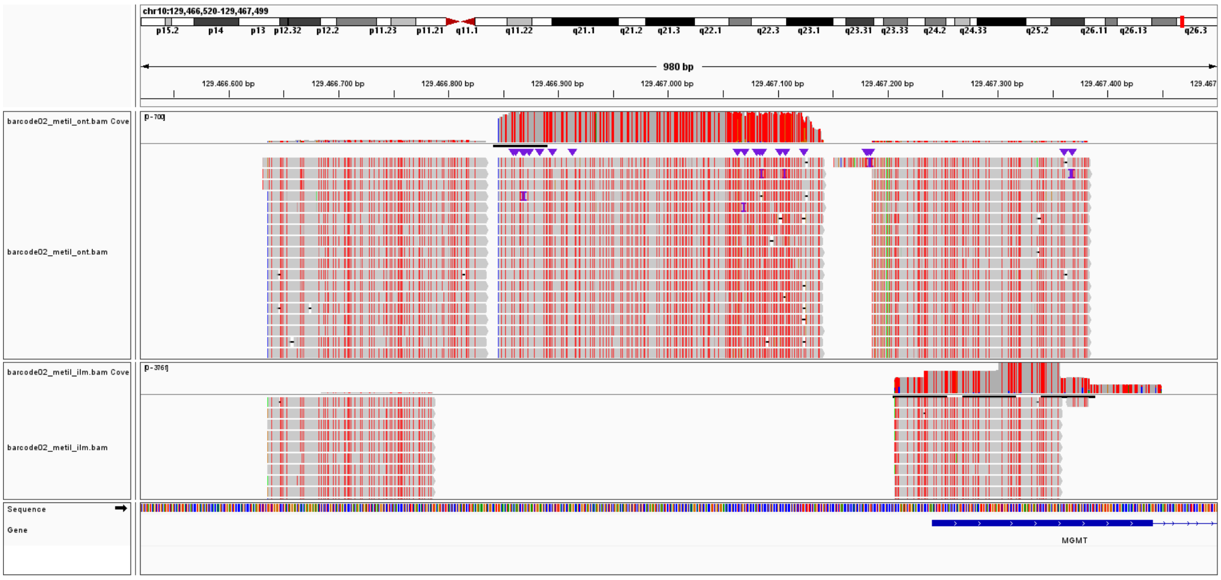
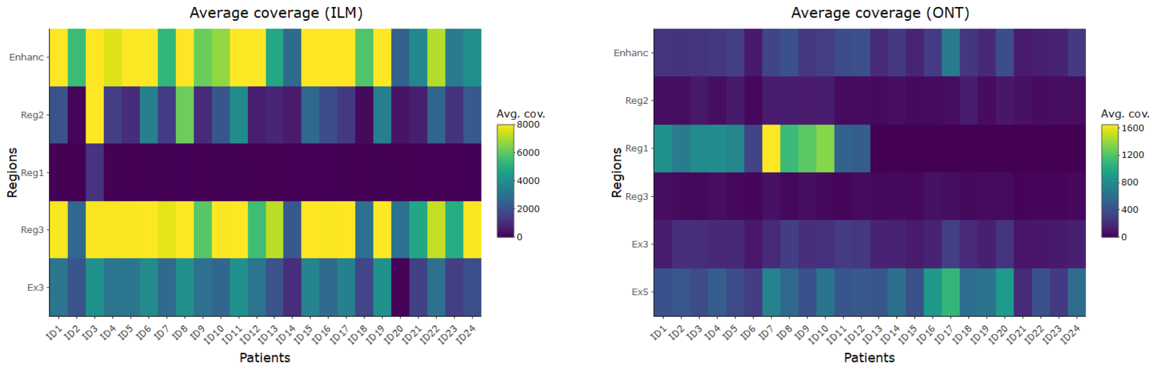


Figure 4.6: BAM files generated by *bwameth* visualized by IGV. Region of *MGMT* for ID2 is shown for both ONT (top) and ILM (bottom). Reg2 (left), Reg1 (center) and Reg3 (right). The red marks indicate thymines (i. e. unmethylated cytosine after Sodium Bisulfite treatment).

The data obtained through ONT has been treated with Sodium Bisulfite too: as a consequence, *bwameth* was used to align also ONT data. However, it is important to highlight that this kind of tool is not suited for nanopore sequencing data. Aligned data can be visualized by IGV, Fig. 4.6. The Sodium Bisulfite converts the unmethylated cytosines into uracils, so that they are read as thymines. Red marks in Fig. 4.6 are the thymines instead of the cytosines, so they indicate an unmethylated CpG region. On the other hand, cytosines (in blue) do indicate methylated bases. From the figure it is also possible to get information about the coverage. The central region, Reg1, available only for ONT (top), is characterized by a high coverage ($\sim$ 700 counts). Reg3, on the right, has $\sim$ 40 counts for ONT and $\sim$ 2700 counts for ILM.

The data were also aligned in EPIC TABSAT (see par. 3.5.4), using *Bowtie* algorithm for ILM and *TMAP* for ONT. Information about the number of mapped reads of the different regions, obtained from *Samtools Depth*, is reported in Fig. 4.7. The mean number of mapped reads for each CpG site is color-coded. Reg1 is the region with the lower coverage (in some cases even not available) for ILM data, Fig. 4.7a, while Reg3 and Enhanc are the ones with the higher depth. Fig. 4.7b shows that, in general, ONT data are characterized by a very low coverage and, in general, the number of mapped reads of ONT data is by far lower with respect to ILM data, except for Reg1. As a consequence, methylation results from ILM sequencing were considered to be more reliable and were taken as reference.



(a) (b)

Figure 4.7: Heatmap of the mapping performance (number of mapped reads divided by CpG sites) for each target sequenced through (a) Illumina and (b) Oxford Nanopore.

4.2.2 Methylation call

Some statistics about the methylation percentage of CpG regions revealed by ILM and ONT is shown in the boxplots in Fig. 4.8. These results are obtained from EPIC TABSAT; however, no significant differences were met between methylation data called by EPIC, Methyl Dackel or visually using IGV.

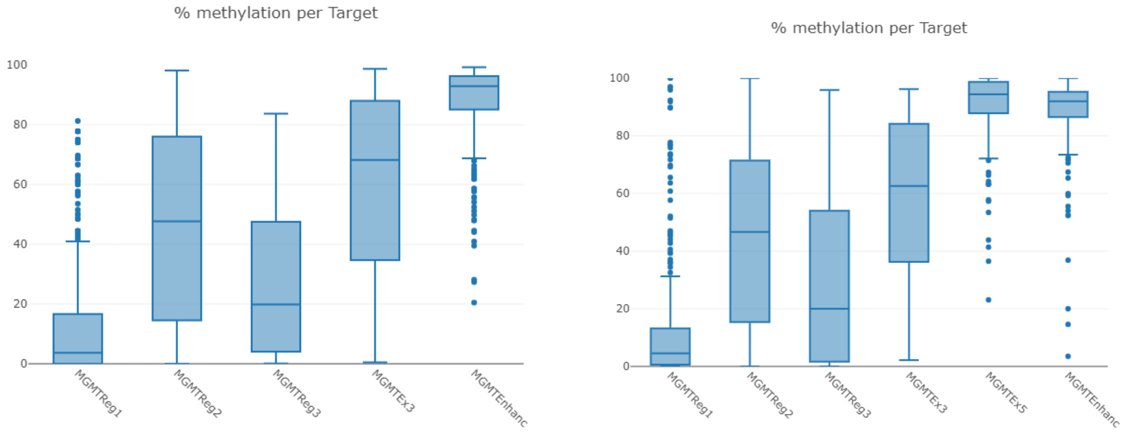
Reg1 is characterized by a low median of percentage for both ILM (3.7%) and ONT (4.5%), but also by many outliers with higher methylation (up to 80% and even 90% for ONT). For some patients, this region has not been analysed, since the number of mapped reads was too small (about 5 mapped reads for ILM).

Reg2, Reg3 and Ex3 have respectively a median of 47.7%, 19.8% and 68.2% for ILM and 46.6%, 19.9% and 62.6% for ONT. They do not present any outliers.

Methylation of Ex5 is available only for nanopore data and it results to have a high percentage, with a median of 94.4%. Also in this case, some outliers can be seen.

The enhancer shows a high methylation percentage, with a median of 92.9% for ILM and 91.9% for ONT, even though several outliers are present.

Mean values of methylation percentage for each region for each patient are reported in Tab. C.4.



(a) (b)

Figure 4.8: Boxplots of the percentages of methylation of CpG sites for the five targets for (a) Illumina and (b) the six targets for Oxford Nanopore. The estimated medians of methylation percentage of the regions seem to agree.

However, it is clear that there are quite big fluctuations across the patients in the methylation percentage of these regions. In Fig.4.9 a lollipop plot of the methylation of the CpG regions in Reg1 for the 24 patients sequenced by ILM is shown. By observing the plot, it is clear that there are a few subjects (e. g. ID7, ID10 and ID17) with a higher percentage of methylation than the others, which in general are characterized by a small fraction of methylated cytosines. They are the outliers that can be seen in Fig. 4.8.

Fluctuations in the methylation percentages are present also in the different CpG regions in the same portion of the genome. Fig. 4.10 shows a lollipop plot of the methylation of the CpG regions in Reg2 for ILM data. Here it is possible to observe that, considering one patient, the percentage of methylated cytosines varies across the region (e. g. ID9 and ID14). This affects the height of the boxplots in Fig. 4.8.

In order to assess the accordance of the medians of methylation percentage of each region estimated by both ILM and ONT, the Wilcoxon test has been computed. It is a

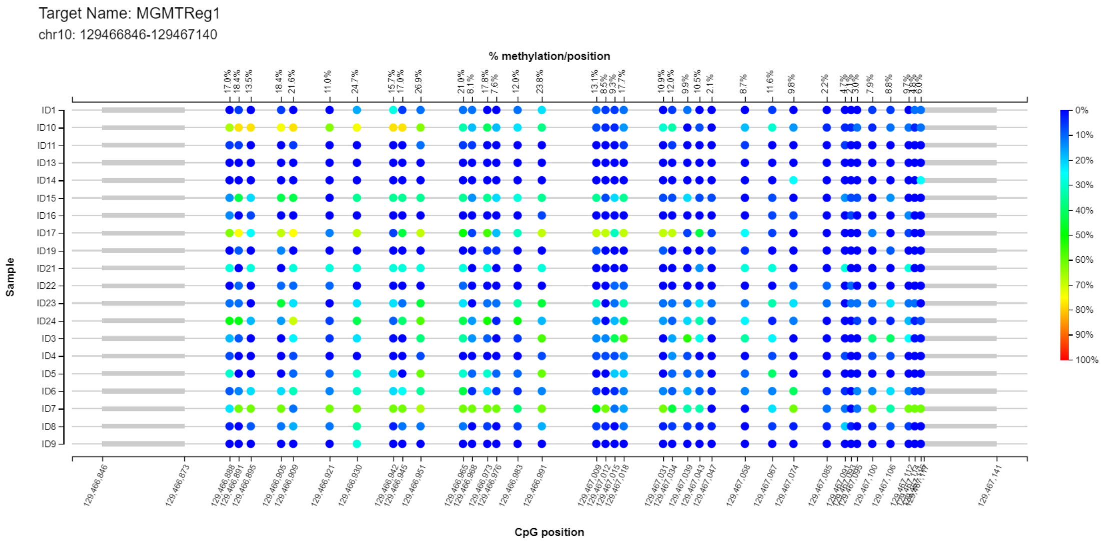


Figure 4.9: Lollipop plot of the percentage of methylated cytosines of Reg1 for the 24 patients sequenced by ILM. The percentage is encoded by a color scale. Here, it is clear that there are some outlier patients, ID7, ID10 and ID17, for instance, with different fraction of methylation than the others. The plot is drawn by EPIC-TABSAT.

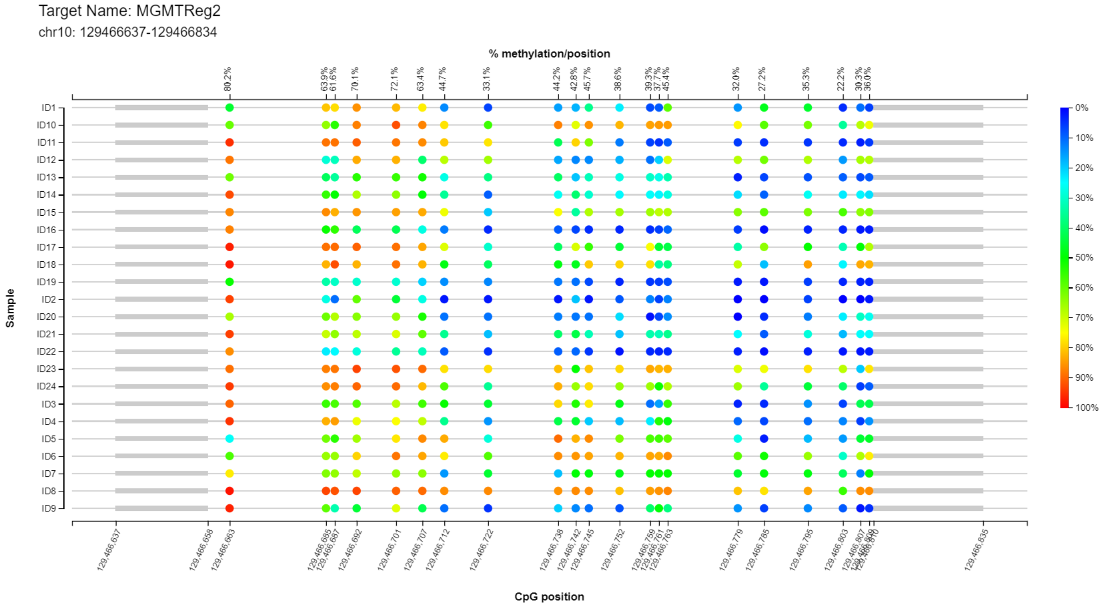
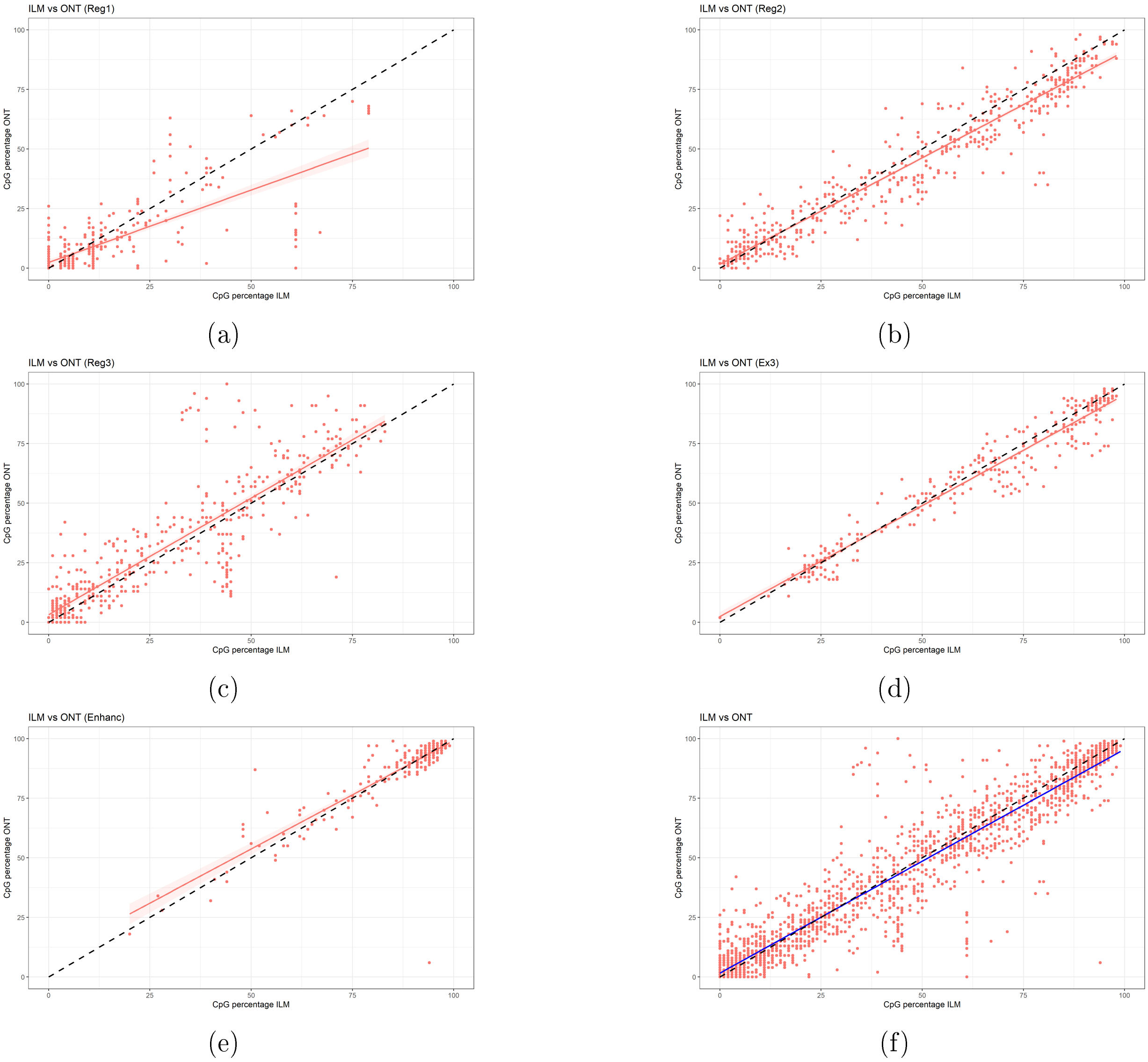


Figure 4.10: Lollipop plot of the percentage of methylated cytosines of Reg2 for the 24 patients sequenced by ILM. The percentage is encoded by a color scale. Here it is clear that there are CpGs with different fraction of methylation across the same region. The plot is drawn by EPIC-TABSAT.

non-parametric test for comparing two statistical samples. The null hypothesis that the two populations are sampled from the same distribution is tested against the alternative one, for which one sample is bigger than the other. The Wilcoxon test assumes the two samples are dependent (they are measurements from the same subjects). It also requires the two distributions to be symmetric, which is not completely satisfied, in particular for Reg1 and Reg3, that have a skewness of about 2 and -2, respectively (so results should be taken with care).

It was computed for Reg1 (p-value $>$ 0.5), Reg2 (p-value $>$ 0.1), Reg3 (p-value $>$ 0.2), Ex3 (p-value $>$ 0.2), Enhanc (p-value $>$ 0.5). It was also calculated for all the regions together, obtaining a p-value $>$ 0.6. So, the null hypothesis cannot be rejected (the significance level can be fixed to $\alpha = 0.05$).



(a)

(b)

(c)

(d)

(e)

(f)

Figure 4.11: Scatter plots of CpG percentage estimated by ILM vs CpG estimated by ONT. Line of equality is drawn (black dotted line). (a) Reg1 ($y = 0.61x + 2.42$), (b) Reg2 ($y = 0.89x + 1.69$), (c) Reg3 ($y = 0.98x + 3.24$), (d) Ex3 ($y = 0.93x + 2.48$), (e) Enhanc ($y = 0.91x + 8.30$), (f) Total ($y = 0.94x + 1.65$); all p-values < 0.05.

Since two different methods were used to estimate the methylation percentage, it could be interesting to study the agreement between them. The scatter plots of the CpG methylation percentage estimated by ILM and ONT for each region, are shown in Fig. 4.11. The equality line is also plotted to be used as reference. Moreover, in the figure caption, the fitted line coefficients are reported. From this kind of plot, it is possible to make some considerations about the agreement between results obtained from ILM and ONT. The more close are the scatter points to the equality line, the better is the agreement.

Considering Reg1, Fig. 4.11a, the points tend to move off from the equality line as the methylation percentage increases. However, two considerations must be made for Reg1: the first is that the majority of CpG sites are unmethylated, so few measurements are available for larger percentages. Secondly, Reg1 is not available for all the patients (only 10) for both ILM and ONT, so the sample is quite small. Moreover, as can be seen in Fig. 4.7a, coverage for ILM data is very low (~ 20 mapped reads for several subjects), so ILM values are not so reliable.

Reg2 and Reg3, Fig. 4.11b and 4.11c respectively, are also affected by the presence of outliers. These large differences between ONT and ILM values can be explained by the fact that in some CpG sites, the depth of ONT was very poor (~ 50 mapped reads), so the found CpG percentages are not completely reliable. The agreement seems to be better for Ex3, Fig. 4.11d. Here, the line of the equality and the fitted line are very close. Also scatter points do not spread out. ONT data has a better coverage, with more than 100 mapped reads for each CpG site.

Looking at Enhanc, Fig. 4.11e, it is possible to see that the majority of CpG sites are methylated and so, again, not so much samples for small percentages are available. The presence of some outliers above the equality line, makes the trend line deviating from the bisector. In particular, there is one point located at $(94, 6)$ immediately captures the attention. The difference between ILM and ONT values are the same also if *MethylDackel* or EPIC are used for the analysis. An hypothesis could be a degradation of the DNA samples due to the bisulfite treatment. It is important to notice that the larger discrepancy between ILM and ONT seems to be in the region of low CpG percentage, where points are more widespread.

The scatter plot for all the regions is shown in Fig. 4.11f. The fit line approaches the line of equality, even if some outliers (both above and below the bisector) are present, mainly in the medium percentage region.

Considering the fit results, Reg1 has the lowest gradient, while the others approach almost 1. Anyway, the regions for which the intercept is relatively small are Reg2 and Ex3, for which points are note much widespread and measures across the whole range of percentages are available. Enhanc intercept, for instance, is heavily affected by the superior outliers, which weigh because few points are present in the medium range. All the regions together show a quite relatively small intercept and a gradient which approaches 1.

The Concordance Correlation Coefficient (CCC) has been computed for all the five regions. It is a quantitative estimate of the distance of the fitted line from the equality line. The computed values together with the confidence intervals are reported in Tab. 4.5. It is important to point out that the CCC behaves well with points sampled from a bivariate distribution. This condition is not met by our data, so the results are only informative and must be taken with care.

The CCC resembles what has been already pointed out: Reg1 measurements are far from the main diagonal, while Reg2 and Ex3 show the best coefficients. They are due to the fact that Reg2 and Ex3 are not affected by outliers, while the points are not so much widespread overall. Also all the regions together have a high CCC, with small confidence intervals. On the other hand, Reg3 and Enhanc show some outliers that are very far from the bisector, which contribute to decrease the CCC.

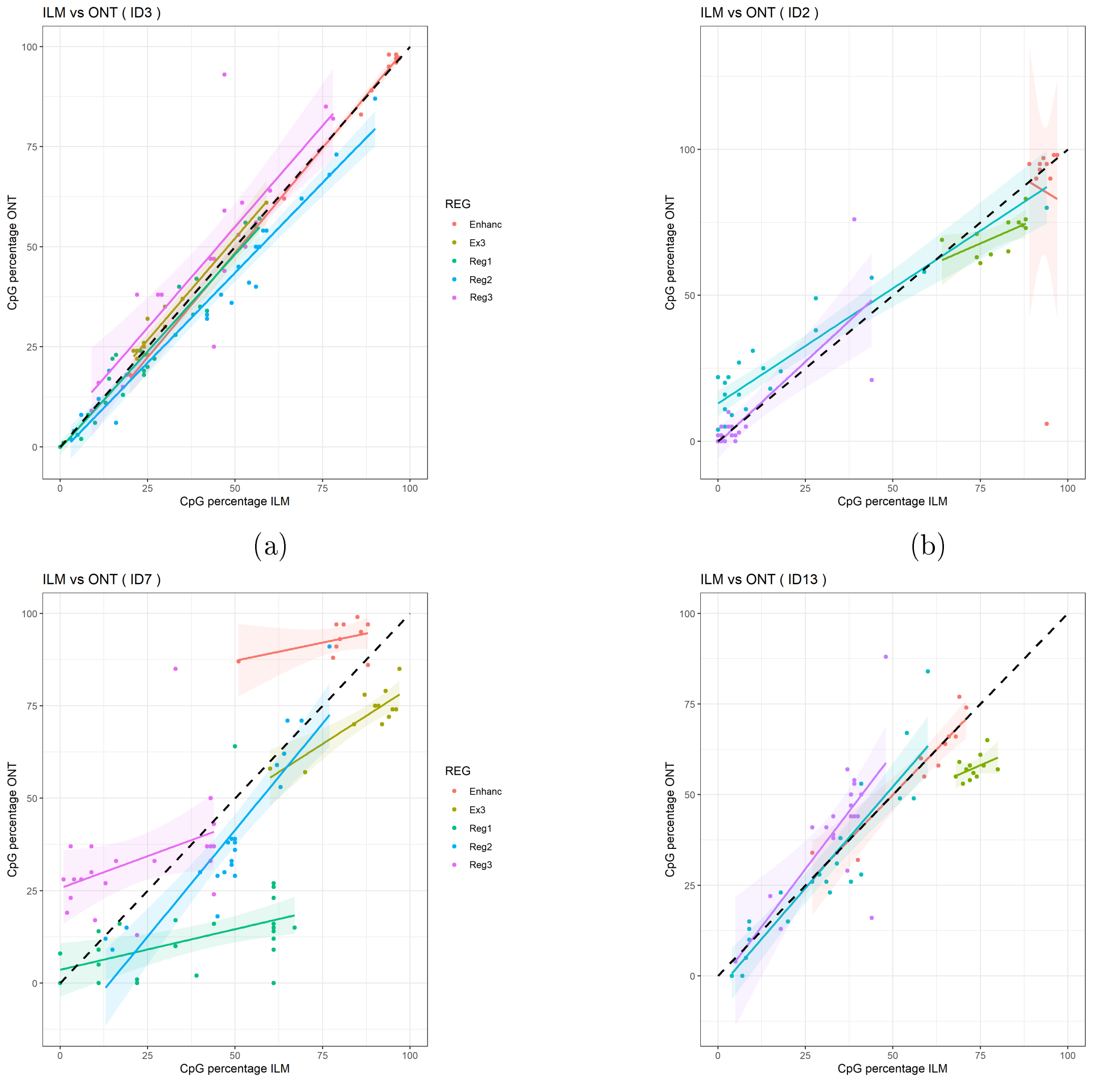
$Region$	CCC	$C.I.$	C_b
$Reg1$	0.736	$[0.689, 0.777]$	0.962
$Reg2$	0.950	$[0.941, 0.957]$	0.992
$Reg3$	0.874	$[0.852, 0.892]$	0.989
$Ex3$	0.964	$[0.955, 0.971]$	0.997
$Enhanc$	0.878	$[0.846, 0.904]$	0.999
$Total$	0.949	$[0.945, 0.954]$	0.999

Table 4.5: Concordance Correlation Coefficient computed for each region, corresponding confidence intervals (C.I.) and bias correction factor C_b.

The same scatter plots were drawn for each patient. However, only four examples are reported, Fig. 4.12, which summarize the different cases that are present. An example of good agreement between ILM and ONT is shown in Fig. 4.12a. Here, all the fit lines for each region approach the bisector. Few outliers can be observed: they are probably caused by regions with lower depth. About 9 patients show this trend.

Examples of bad agreement are in Fig. 4.12b and 4.12c. They show the two outliers of Enhanc, of which it had been already deal with. In Fig. 4.12c almost all the scatter points are far from the main diagonal. They are not even parallel to the line of equality. Probably, by removing these patients from the samples plotted previously, the agreement between ILM and ONT does improve. About 5 patients have this 'bad' trend.

Fig. 4.12d show a patient with good agreement for some regions (in particular Enhanc, that practically coincides with the line of equality) and poor with others, such as Ex3, which is quite far from the bisector. About 11 subjects share this characteristic.



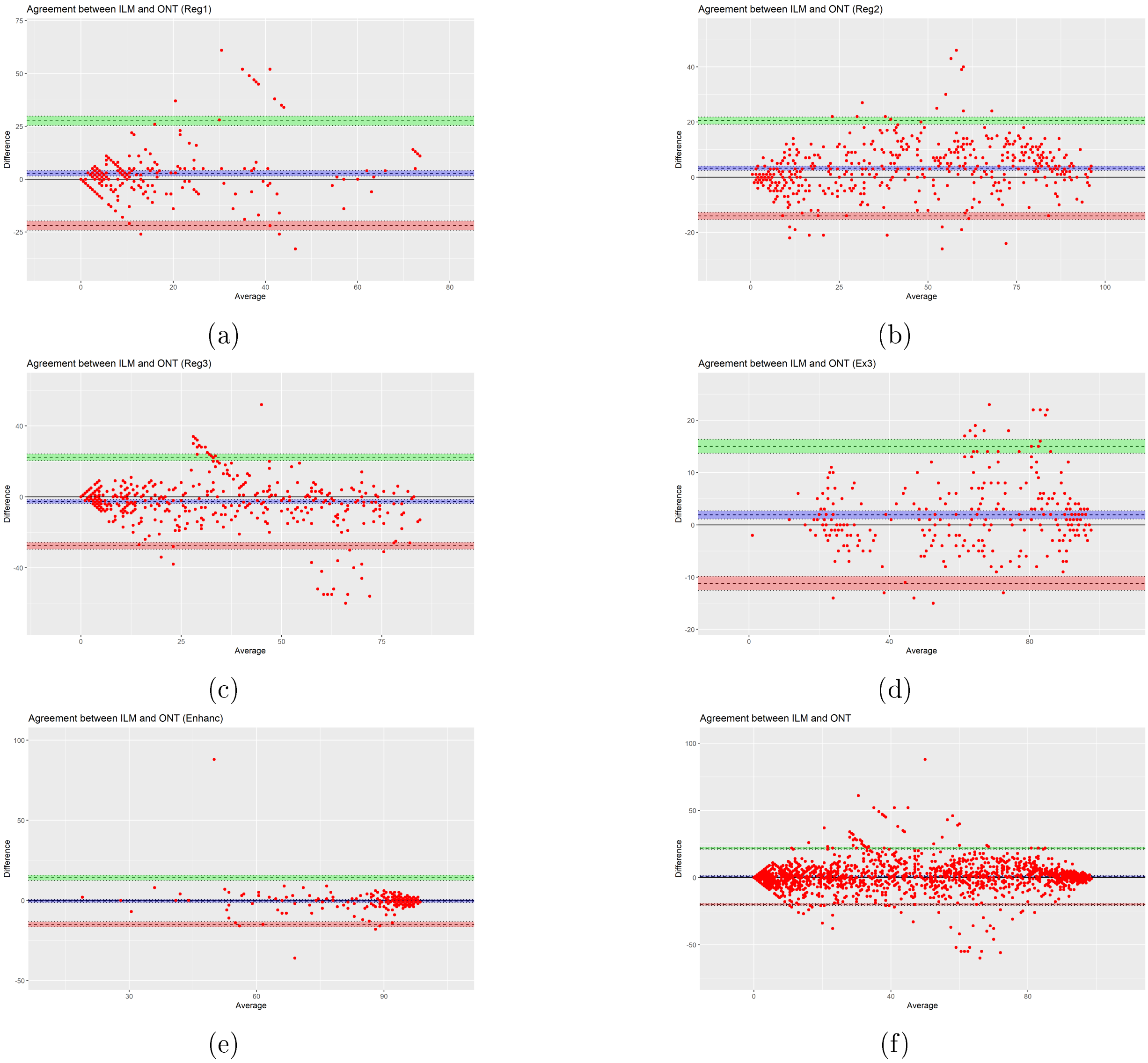
(a) (b)

(c) (d)

Figure 4.12: Scatter plots of CpG percentage estimated by ILM vs CpG estimated by ONT for (a) ID3, (b) ID2, (c) ID7, (d) ID13.

To get more information about the agreement between the two techniques, Bland-Altman plots (see par. 3.6.1) are drawn, Fig. 4.13. The BA plot works well if data are normally distributed. However, if the distribution is not normally distributed, but the 95% of data falls within the mean ± 1.96 standard deviation, then the plots are nevertheless reliable. Again, it is important to invite the reader to take the results with care. The above condition is fully satisfied by the distribution of the differences in Reg2, Enhanc and for all the regions, while does not completely hold for the others.

The BA plot also shows the bias (i. e. the mean difference between the two methods) and the limits of agreement, with their own confidence intervals. The limits of agreement are computed under the hypothesis that the differences are normally distributed, which is not true. However, when two different methods are tested, the biological/clinical



(a)

(b)

(c)

(d)

(e)

(f)

Figure 4.13: Bland-Altman plots of the CpG percentage of methylation, estimated by ILM and ONT. The plot are drawn for each patient for (a) Reg1, (b) Reg2, (c) Reg3, (d) Ex3, (e) Enhanc and (f) all the regions.

problem itself establishes the limits of agreement. For instance, if the problem requires to precisely estimate the percentage of methylation (e. g., to distinguish between 23% and 24%), then the limits of agreement will be very narrow. If the problem needs only to discriminate between methylated and unmethylated sites, wider limits of agreement can be used. So, it is important to be careful in observing and comparing different BA plots.

Reg1, Fig. 4.13a, has a bias of 2.8. Scatter points are concentrated towards the low values, while for larger averages several points fall out from the superior limit of agreement (which are almost at −21 and 27, quite wide). The pattern on the left resembles the vertical pattern of points visible in Fig. 4.11a. In presence of low CpG methylation, ONT and ILM tend to over and underestimate the percentage, due to the difference in mapped reads for both.

Reg2, Fig. 4.13b, show points that are more distributed along the averages interval. The bias is around 3.2. Here, outliers are present beyond both the limits of agreement (-14 and 20) and along all the interval.

The bias of Reg3, Fig. 4.13c, is negative and equal to -2.6. Again, outliers are present above and below the limits of agreement (which are at -28 and 22). Both Reg1 and Reg3 show a weak linear relationship between differences and averages.

A clearly visible characteristic of Ex3, Fig. 4.13d, is that the average difference between ONT and ILM is small than other regions. The outliers are not so distant from the agreement limits (at -11 and 15, relatively narrow). The bias is 1.9.

As expected, Enhanc has the majority of points concentrated around high percentages, Fig. 4.13e. The bias is -0.4 and 0 falls within its confidence interval, so that it is possible to say that no bias is present between ONT and ILM percentages. The great majority of points is within the limits of agreement (at -15 and 14), but the two outliers, about which had been already deal with.

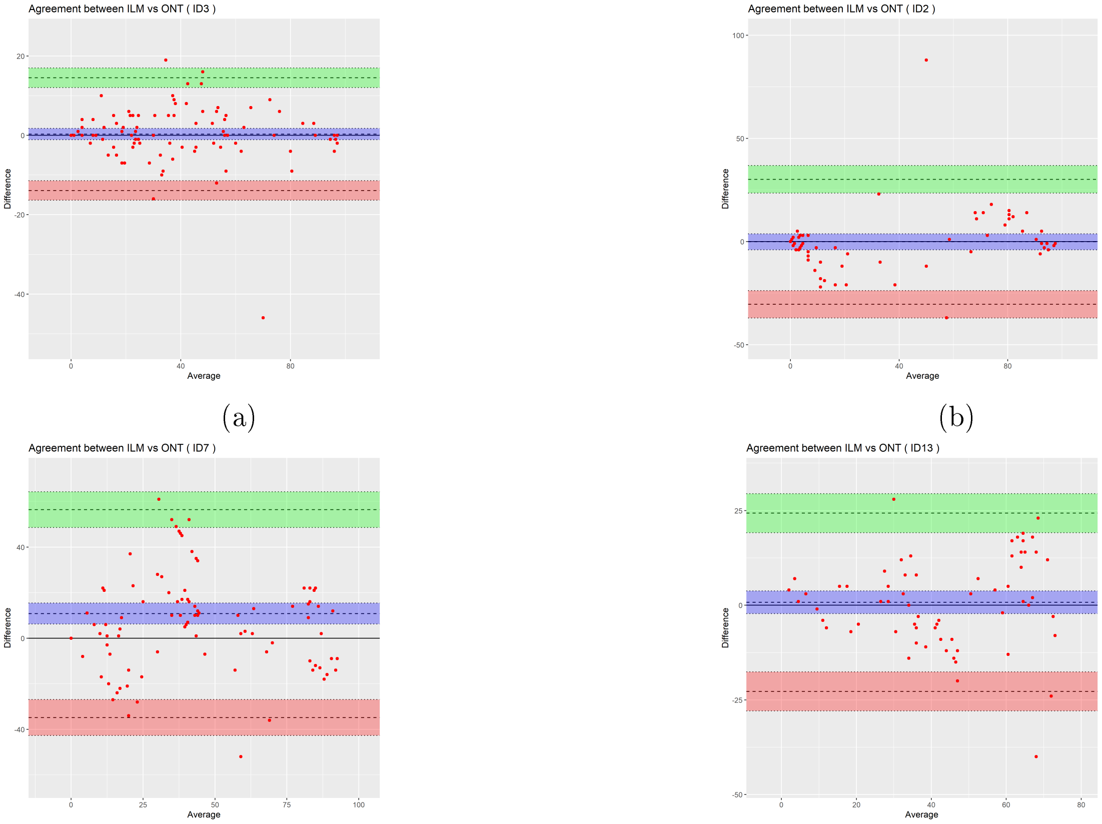
Fig. 4.13f shows the BA plot for all the regions and all the patients together. The estimated bias is 0.9, but it can be neglected since 0 falls within its confidence interval. Here points are well-distributed along all the range interval. Outliers are present mainly in the medium average region, while they are absent towards high values. Limits of agreement are placed at -20 and 22.

Now, the previously discussed good agreement and bad agreement examples are shown using the BA plots, Fig. 4.14. For all of them, the condition under which BA plot behaves well is satisfied. ID3 has almost all the points within the limits of agreement (14 and -14), except for a distant outlier. Moreover, the points are quite spread along all the range of average values.

ID2, Fig. 4.14b, has a few outliers too. However, its points show a sinusoidal-like pattern inside the region of agreement. This pattern could mean that there is a dependence between the average of the CpG methylation and the differences between the two instrument estimates. When the CpG average is small, ONT over-estimates the percentage of methylation, while for high average, ILM seems to over-estimate in turn. When a clear relationship between average and differences is present, it might happen that limits of agreement (at 30 and -30) are over-estimated.

Also, ID7, Fig. 4.14c, seems to be affected by a similar dependency: here the pattern is negative-positive-negative like. Again, there is probably an over-estimation of the limits of agreement, which are at -35 and 56 (quite wide). The bias of 11 is also high.

ID13, Fig. 4.14d, shows a quite oscillating behaviour, with alternating increase and decrease of the difference between ILM and ONT data as the average of the measures increases. However, limits of agreement are narrower, precisely at -22 and 24. The bias is about 0.8 and 0 falls within its confidence interval.



(a)

(b)

(c)

(d)

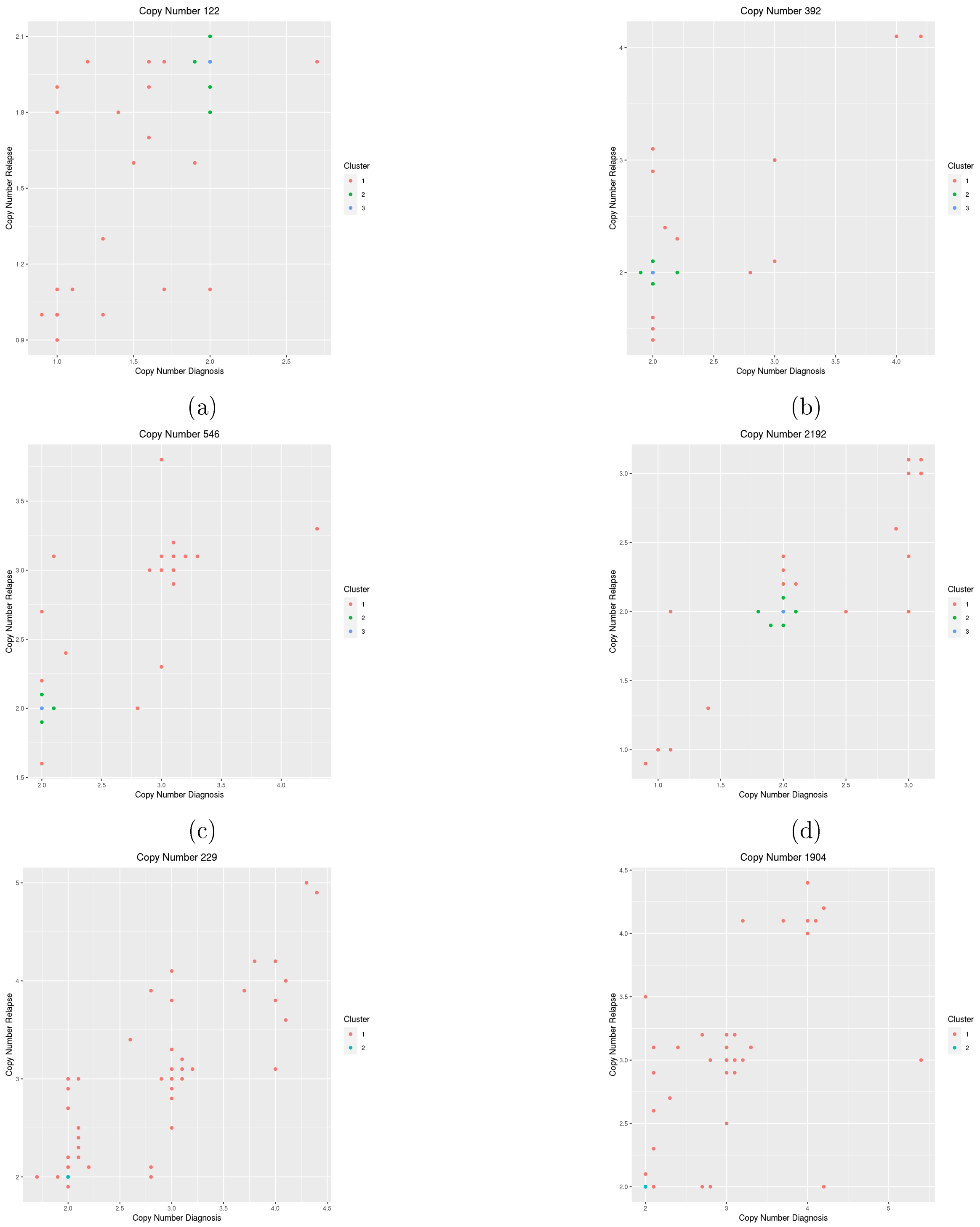
Figure 4.14: Bland-Altman plots of CpG percentage estimated by ILM vs CpG estimated by ONT for (a) ID3, (b) ID2, (c) ID7, (d) ID13.

4.3 MM patients analysis

Multiple Myeloma patients are characterized by a large number of features (i.e. Copy Number values). They can be used to get more information about the patient, the effect of the treatment and the evolution of the malignancy. Dirichlet Process, which is suitable for handling large number of data, is used to try to group patients' features and provide a classification possibly useful for clinical purposes.

4.3.1 Hierarchical Bivariate Dirichlet Clustering

The first idea is to apply the Hierarchical Dirichlet Process, that is an unsupervised method (i.e. it does not require an a priori number of clusters to be specified and does not need to be trained), with a 2 dimensional multivariate (i.e. bivariate) mixture model.



(a) (b)

(c) (d)

(e) (f)

Figure 4.15: CND vs CNR of 6 CNs for each patient and clusters obtained from the 2D MHDC (hierarchy on CN) for 1000 iterations. Cluster 2 and 3 are the ones that were removed. (a) CN-122 (*SNX7*); (b) CN-392 (*RPIA*); (c) CN-546 (*ALCAM*, *CBLB*); (d) CN-2192 (*CDH20*, *RNF152*); (e) CN-229 (*GPR161*, *TIPRL*, *SFT2D2*, *TBX19*, *XCL2*, *XCL1*, *DPT*, *ATP1B1*, *NME7*, *BLZF1*, *CCDC181*); (f) CN-1904 (*VPS13C*, *C2CD4A*, *C2CD4B*, *TLN2*).

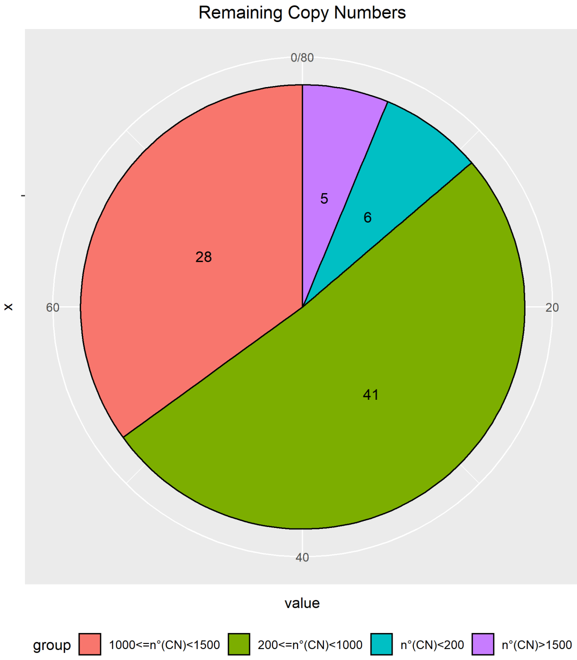
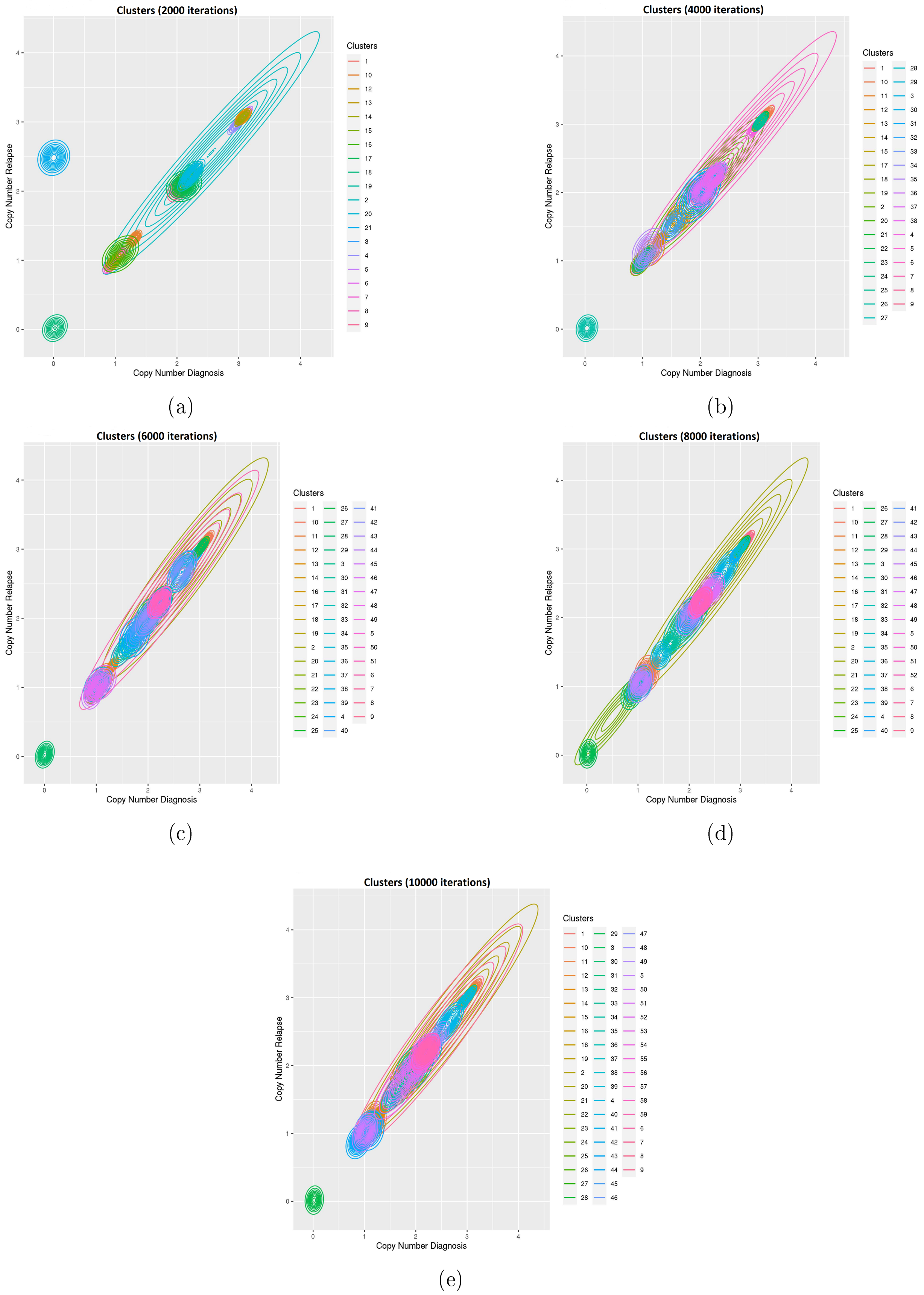


Figure 4.16: Pie chart of the number of remaining Copy Number (n°(CN)) pairs for each patient, after the first Hierarchical Dirichlet Clustering. Blue: $n°(CN) <$ 200, green: $200 \leq n°(CN) <$ 1000, red: $1000 \leq n°(CN) <$ 1500 and violet: $n°(CN) \geq$ 1500.

The aim is to identify clusters that will be shared among the different patients. We chose the bivariate distribution because the data are continuous and not defined on a bounded interval.

However, one of the main issues met was the fact that a huge portion of the Copy Number couples is centered at about $(2, 2)$. This heavily affects the result of the clustering. To deal with this problem, a first clustering was performed using CN as hierarchy: each CN consisted of a group, whose elements were the corresponding 80 patients' CN couples. So, there were 2409 groups made of 80 elements. The CN values were rounded to the first decimal place. The 2D Multivariate Hierarchical Dirichlet Clustering (MHDC), implemented by Giovanni Sighinolfi, was performed for 1000 iterations. Results are shown in Fig. 4.15 (only some plots are reported as examples).

The 2D MHDC found 3 clusters: cluster 3 exactly centered in $(2, 2)$, cluster 2 near the previous and cluster 1 that comprises all the other CN couples. It is possible to notice that, for some groups (i.e. CN), only two clusters have been found (cluster 2 is absent). After that, for each patient, the CN couples assigned to cluster 2 and 3 were removed. By doing so, it is expected an increase of the ability of MHDC to find clusters not centered in $(2, 2)$. This removal operation is justified by the following reasoning: a CN equal to 2 denotes a 'normal' situation (see par. 1.3.2). Consequently, Copy Number pairs centered around $(2, 2)$ represents an initial normal situation (no gain, no loss) that



(a)

(b)

(c)

(d)

(e)

Figure 4.17: Clusters obtained from the 2D MHDC for (a) 2000 iterations (20); (b) 4000 iterations (37); (c) 6000 iterations (50); (d) 8000 iterations (51); (e) 10000 iterations (58).

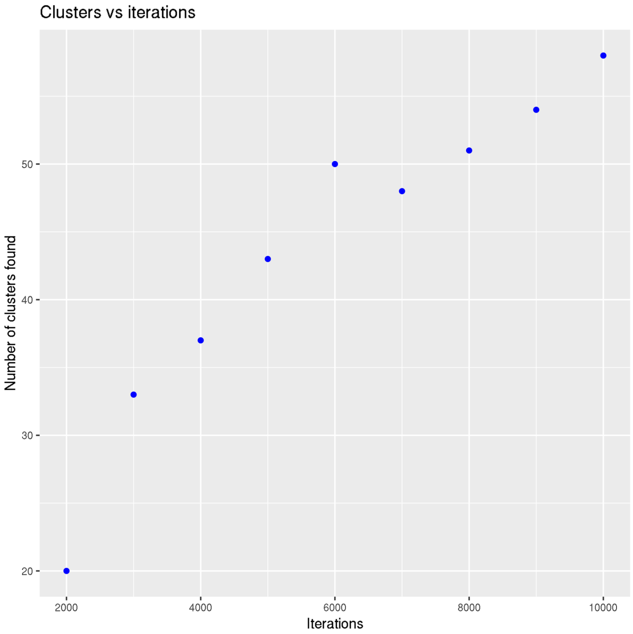


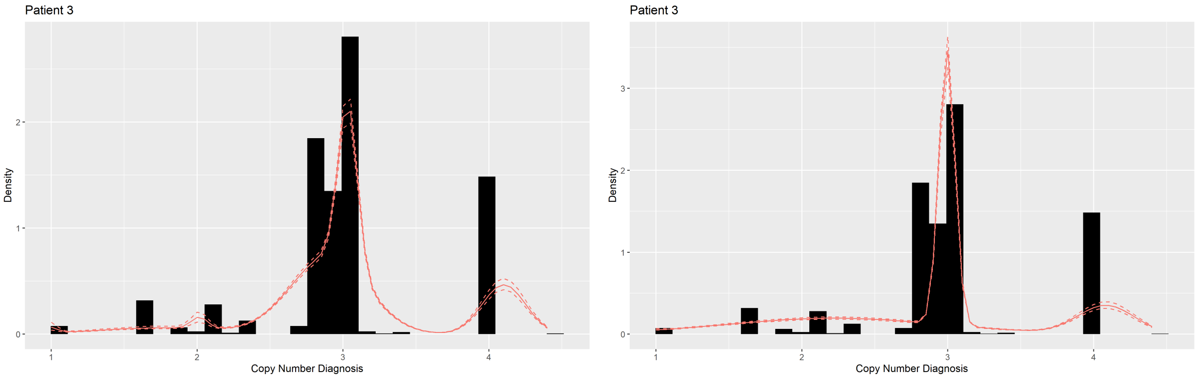
Figure 4.18: Number of clusters found by the MHDC vs iterations. The increasing trend is clearly visible.

remained so after the treatment. This kind of information could be neglected if the aim is to characterize the evolution of the patients' profiles after the therapy.

There are some patients who have lost a huge number of CN pairs after the previous clustering, remaining with less than 100 of features. From Fig. 4.16 it is possible to notice that 6 patients (precisely 2, 23, 34, 62, 63 and 74) have less than 200 of remaining CN pairs. Only five patients (11, 14, 30, 55 and 72) have more than 1500 CN, meaning that the fraction of 'normal' CN values was huge for almost all the patients. Only one CN pair has been removed from patient 14.

Then, the 2D MHDC was performed using patients as hierarchy, by mean of steps of 2000 iterations, until a total of 10000 iterations was reached. Each patient consisted of a group, whose elements were the CN couples (CND, CNR). So, there were 80 groups, each one made of a different number of elements (because of the removal of the CN pairs centered around $(2, 2)$). The results of 2D clustering are shown in Fig. 4.17. Starting from 2000 iterations, Fig. 4.17a, the number of clusters found by the MHDC is increasing, up to reaching 58, Fig. 4.18. However, the clusters are all almost along the main diagonal, except for 2000, Fig. 4.17a, where we observe one cluster out of the main diagonal. This cluster is centered around $(0, 2.5)$. Moreover, the great majority of the clusters is placed around $(0, 0)$, $(1, 1)$, $(2.5, 2.5)$ and $(3, 3)$. Many clusters are almost superposed, so that it is difficult to discriminate between each other. These observations indicate that the final result is heavily affected by the more regular samples (i. e. the ones with CND = CNR, which are genes that started from an abnormal situation and remained so), failing to catch the points out of the main diagonal. This could be due to the fact that the MHDC finds clusters that are shared among the different groups, so that the CN couples present in many patients affect the final results, while more rare couples are not revealed by the algorithm.

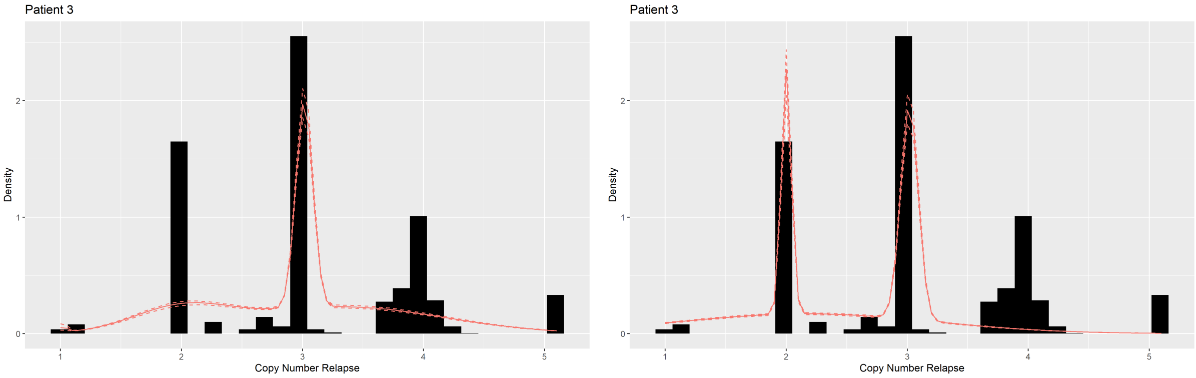
4.3.2 Hierarchical Univariate Dirichlet Clustering



(a)

(b)

Figure 4.19: Examples of the results from two different realizations, with the same parameters, of the Univariate Hierarchical Dirichlet Process applied to CND of patient 3. In (a) there is a component centered in 2 that results to be absent in (b).



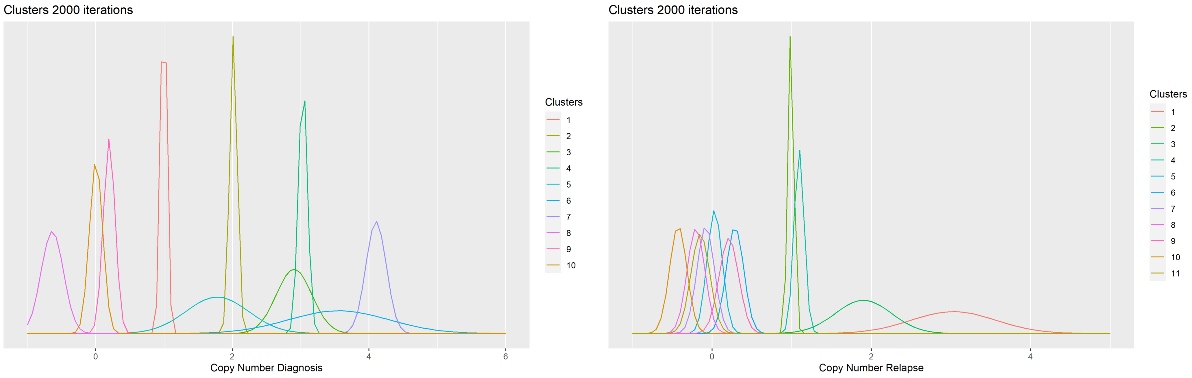
(a)

(b)

Figure 4.20: Examples of the results from two different realizations, with the same parameters, of the Univariate Hierarchical Dirichlet Process applied to CNR of patient 3. In (a) there is a peaked component centered in 2 that is absent in (b).

To better characterize the patients' evolutionary patterns, it is possible to perform a Hierarchical Dirichlet Process separating the diagnosis and the relapse phase (i. e. a 1D HDC). The idea is to keep CND and CNR separated to capture patients' profiles before and after the treatment, and eventually comparing them. By modifying the original code, the 1D version, a Univariate Hierarchical Dirichlet Clustering (UHDC) has been implemented. UHDC was performed by taking into consideration only CND for all the 80 patients. The clustering procedure was repeated 10 times, using the same hyperparameters and prior parameters. At each repeat, the results differ from each other and different components (i.e. Gaussian clusters) were found. The results for 2000 iterations for 2 realizations are shown in Fig. 4.19.

The same procedure has been performed on CNR for all the 80 patients. As before, 10 trials were done, each for 2000 iterations. Results are shown in Fig. 4.20 Again, different components were found.

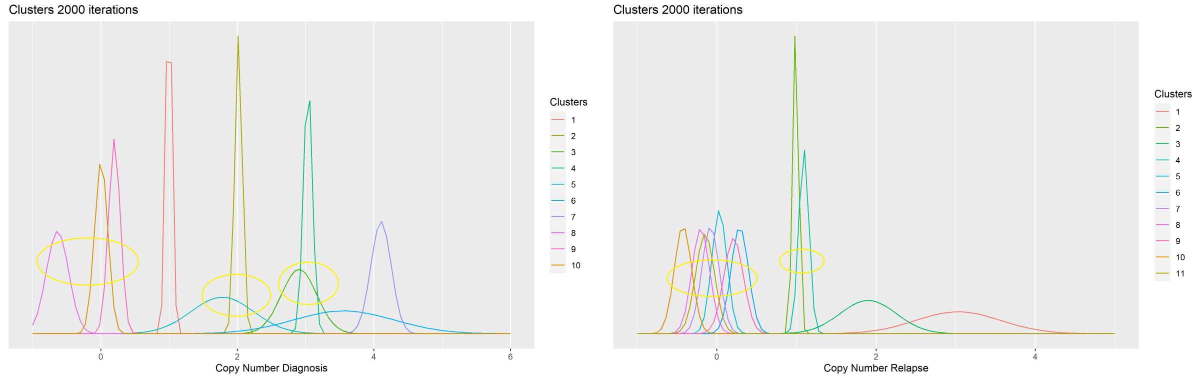


(a) (b)

Figure 4.21: Gaussian clusters (components) found by the UHDC with the lower mean value of misclassified points. The heights of the Gaussians do not correspond to the fraction of CNs in each cluster. (a) 10 clusters found for the diagnosis case; (b) 11 clusters found for the relapse case.

So, the problem is that, at each execution of the Dirichlet Process, the number and the type of the clusters found change. This is due to the fact that the Dirichlet Process has a random component inside the algorithm and, as a consequence, the fit must be performed several times (e.g., 10). A way to determine which is the 'best' clustering is needed. It is possible to assess the 'goodness' of clustering using the Silhouette score (see par. 3.9). The idea is to compute this quantity for patient for each iteration result and, after this step, to evaluate the percentage of misclassified points for each patient for each result. A point is considered to be misclassified if its Silhouette score was ≤ 0.5. Then, we have a vector of percentages for each iteration trial. By computing the mean of this vector, it is possible to choose the result which shows the smallest number, that would be the best clustering option.

Then, all the clusters found for the diagnosis and relapse by the chosen best clustering are plotted, Fig. 4.21. From now on, the chosen clusters will be called components. It is clearly visible that some components are similar to others: they are basically centered around the same value while the variances differ, and the information carried by them is almost the same. So, it is possible to merge them, as to consider them as the same one. An objective way to determine whether two components can be merged is needed: two Gaussians can be treated as same if their mean values differ less than 0.25 and clusters centered around 0 are merged together. From this consideration, components 2 and 5 of diagnosis can be merged to form the 'new' component 2, and the same for components 3 and 4 (they will form component 3) and for components 8, 9 and 10 (new component 6). Component 6 becomes component 4 and component 7 becomes component 5. For the relapse case, components 2 and 4 are merged to have a new component 2, while components from 5 to 11 are substituted by a new component 4. The merging operation is shown in Fig. 4.22.



(a)

(b)

Figure 4.22: Gaussian components after merging together similar components. The height of the Gaussians do not correspond to the fraction of CNs in each cluster. (a) 6 components left for the diagnosis case; (b) 4 components left for the relapse case.

After this merging operation, it is possible to define the patient's profile. The patient's profile in diagnosis (relapse) is defined as the set of components which are present for that patient in the diagnosis (relapse). The profiles of all the patients can be simply represented as a matrix with 80 columns (one for each patient) and a number of rows equal to the total number of components found for diagnosis (6) and relapse (4). Then, if a patient holds a component, the corresponding cell will be colored in yellow, while if the patient is not described by that component, the corresponding cell will be colored in red. The profiles of patients in diagnosis and relapse are shown in Fig. 4.23. By observing the two matrices in the figure, it is clear that there some patients characterized by the same components in diagnosis, Fig. 4.23a and in relapse, Fig. 4.23b.

Now, it is possible to define the concept of patient's evolutionary trajectory: it consists in both the profile in diagnosis and in relapse of a subject. So, patients with the same initial and final profile (i.e. patients with the same initial and final state) will have

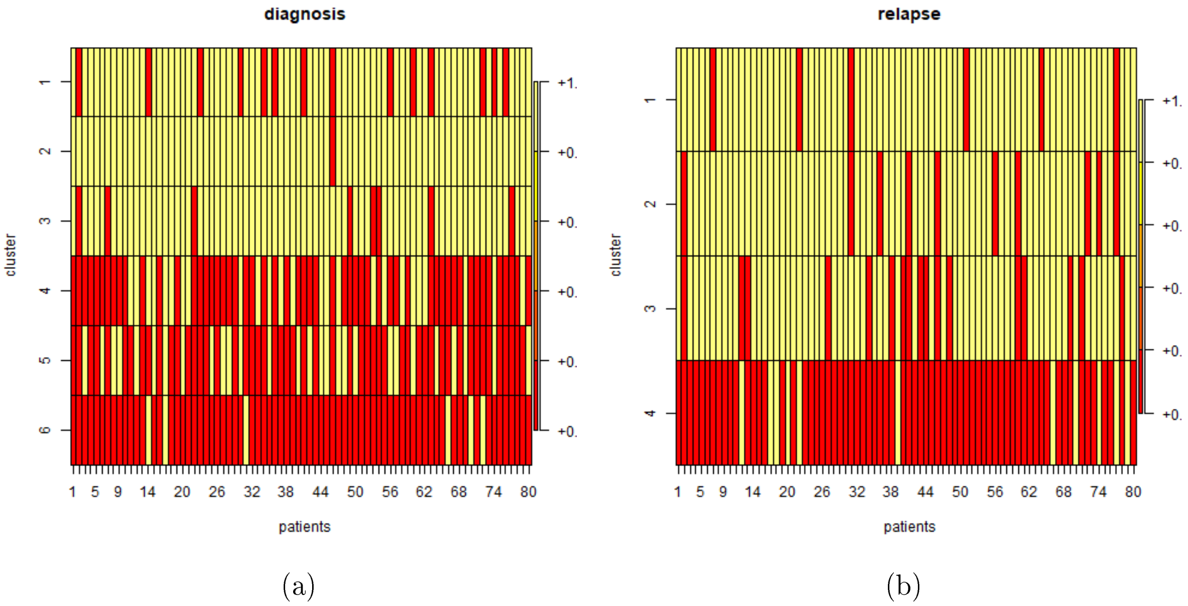


Figure 4.23: Matrix representation of the patients' profiles for 2000 iterations, after component merging. The rows correspond to the components (clusters) while the columns to patients. Present components are highlighted in yellow, while absent ones are coloured in red. (a) Profiles for the diagnosis case; (b) profiles for the relapse case.

the same evolutionary trajectory. In practice, it is quite simple to establish the evolutionary trajectories. Firstly, patients with the same initial profile are grouped together (the same for the relapse profile). Then, subjects which are in the same group both at diagnosis and in relapse (i.e. patients that start from the same initial profile and, after the therapy, reach the same final state) also have the same evolutionary trajectory. Let's make an example: looking at Fig. 4.23, we can consider patient 1 and patient 4. Both are described by components 1, 2 and 3 in diagnosis. So, they share the same initial profile. Looking at the relapse matrix, they are both described by components 1, 2 and 3 at relapse. So, they have final profile in common. Therefore, they can be described by the same evolutionary trajectory: they both start from an initial state characterized by components 1, 2 and 3 and reach a final state described by components 1, 2 and 3.

The new classification obtained using this procedure is shown in Tab. 4.6. There are 11 evolutionary trajectories that are common to at least two subjects. By comparing the new groups and the empirical classification, it comes out that there is not a good agreement. The main reason is that the hierarchical univariate classification has been performed over a part of the entire dataset, after removing the genes with initial and final CN both almost equal to 2, while G classification considered all the features. The evolutionary trajectories are shown in the alluvial plot in Fig. 4.24. The alluvial plot allows to clearly observe the trajectories of each patient, the initial profile and the final one. $D2$ is the most popular initial profile while $R3$ is for the relapse. In particular,

Trajectory	*Patients*
1	1, 4, 5, 10, 16, 19, 24, 26, 29, 32, 52, 59, 65, 68, 73, 75
2	3, 6, 8, 9, 25, 28, 42, 50, 57, 67, 80
3	11, 21, 33, 37, 55, 58, 62
4	13, 38, 43, 71, 78
5	15, 35, 45, 47
6	18, 39, 79
7	27, 40, 48
8	23, 76
9	44, 69
10	49, 54
11	66, 70

Table 4.6: Groups (trajectories) of patients that share the same profile both in diagnosis and relapse, after component merging.

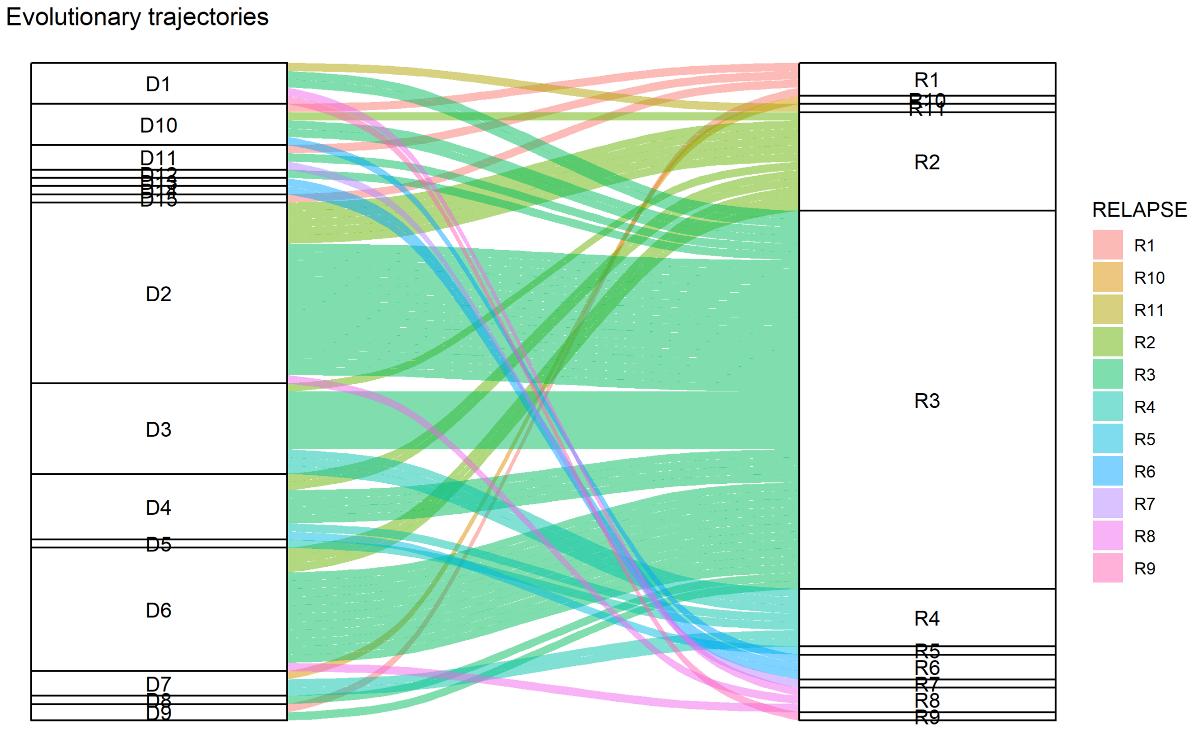
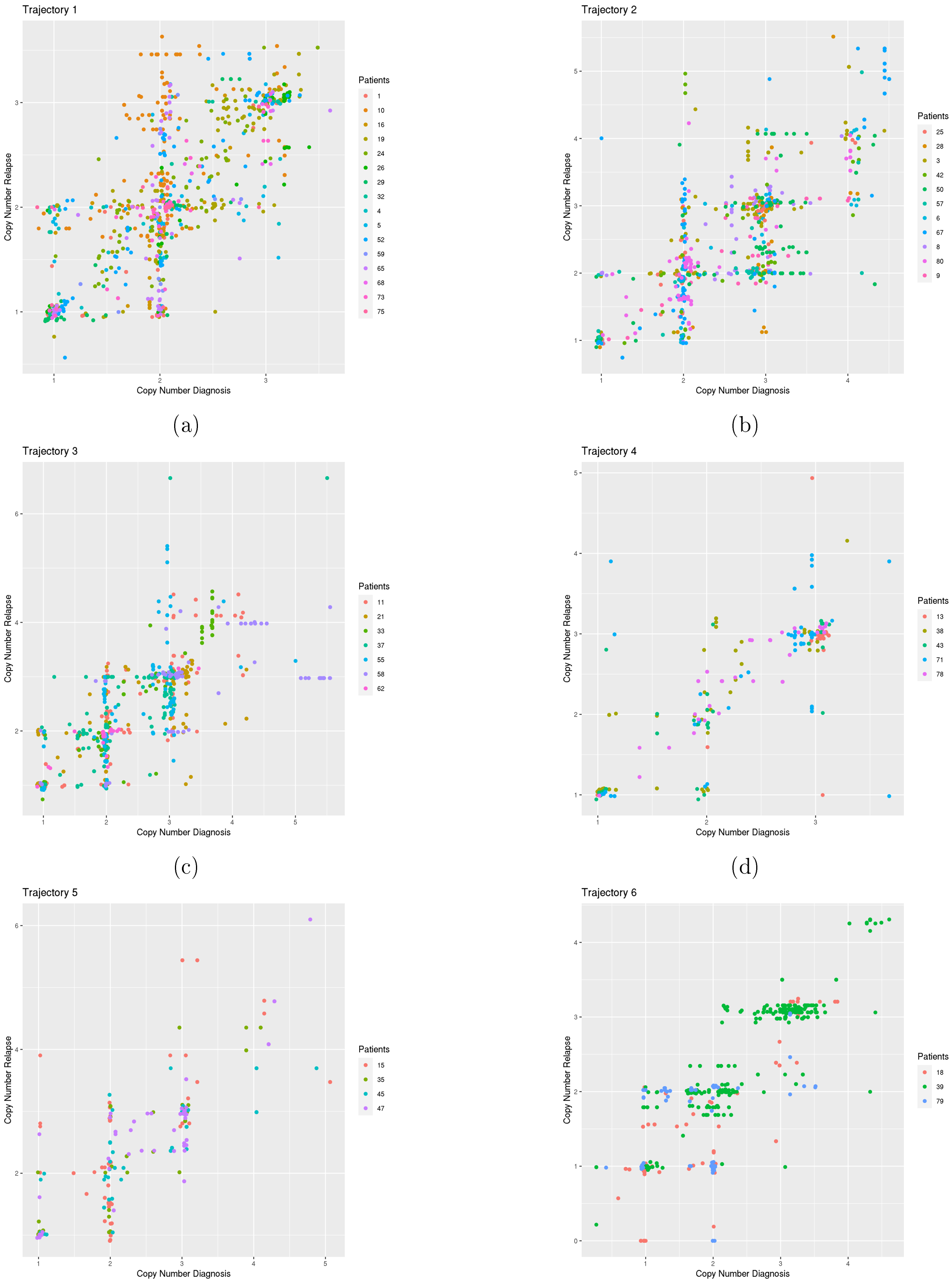


Figure 4.24: Alluvial plot of the evolutionary trajectories of the 80 patients. On the left, the profile in diagnosis and on the right, the profile at the relapse. Colors of the trajectories reflect the final profiles.



(a)

(b)

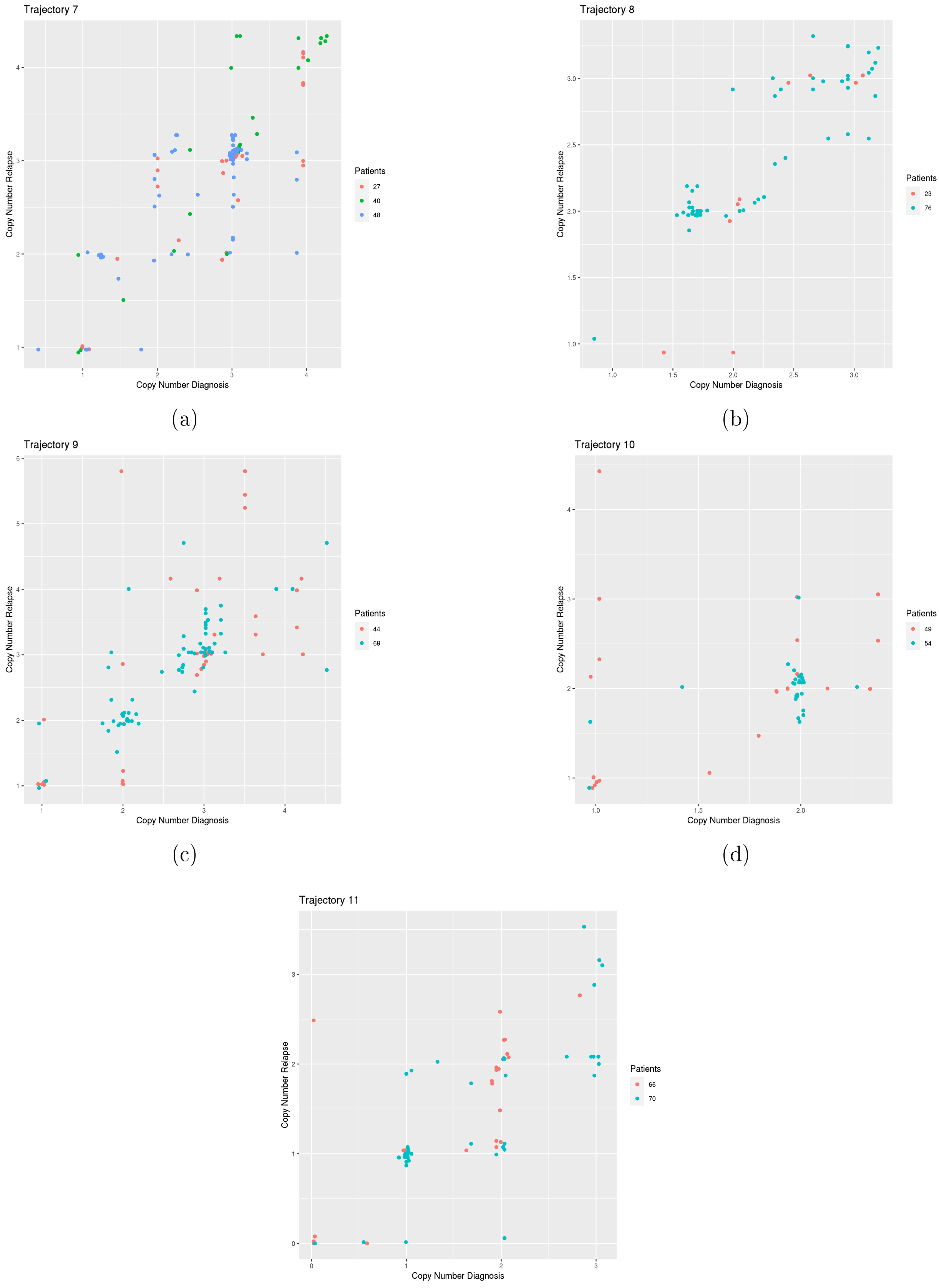
(c)

(d)

(e)

(f)

Figure 4.25: Scatter plots of the first 6 of the 11 evolutionary trajectories found by the UHDC. (a) Trajectory 1, (b) trajectory 2, (c) trajectory 3, (d) trajectory 4, (e) trajectory 5, (f) trajectory 6. (a)-(c) have a branched-like pattern, while (d)-(f) seem to be more drifted.



(a)

(b)

(c)

(d)

(e)

Figure 4.26: Scatter plots of the last 5 of the 11 evolutionary trajectories found by the UHDC. (a) Trajectory 7, (b) trajectory 8, (c) trajectory 9, (d) trajectory 10, (e) trajectory 11. These trajectories seem to be linear with in some cases drifts, in others branches.

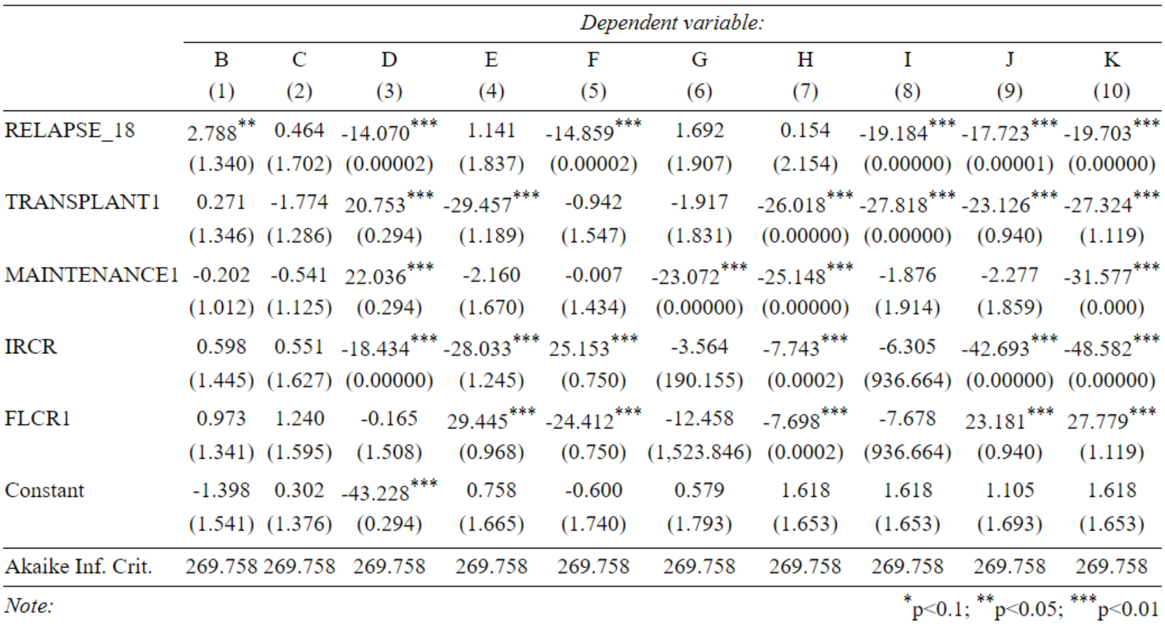
patients with final profile $R3$ came from different profile in diagnosis. This suggests that the treatment that patients underwent results in an effect that is similar for many subjects.

Then, the patients composing each group have been investigated in order to assess if common portions of the genome (i. e. features) belong to the same components. This condition is met by trajectories $4 - 11$, which share features that evolve in the same manner (from a component to another one). Trajectory 1 has portions of genome whose evolution is similar for at maximum 9 out of 16 patients, while trajectory 2 and 3 only for 7 out of 11 and 5 out of 7, respectively.

The scatter plots of the 11 trajectories are shown in Fig. 4.25 and 4.26. For each trajectory, CND vs CNR is plotted for each patient. The first 3 trajectories show a branched-like pattern. However, these groups are quite heterogeneous and some inter-patients differences are present. Trajectory 4, 5 and 6 can be considered drifted. The first two show vertical drifts, that are common for all the patients belonging to the group. Trajectory 6 instead has horizontal drifts. Trajectory 7 can be labelled as linear-branched: the trend is linear and both horizontal and vertical branches are present. Concerning trajectory 8, Fig. 4.26b, it is not easy to assign a label. Patient 23 has a small number of remaining CNs and so he does not show many similarities with patient 76. However, they can be assigned to a linear trend. Trajectory 9, Fig. 4.26c, is characterized by a sort of branches, even if patient 69 has a predominant linear growth. To finish, trajectories 10 and 11 can be assigned to a linear-drifted trend, even if some differences are present.

In order to assess if the obtained classification is consistent with the available clinical parameters, a multinomial logit model has been constructed. The model has the clinical parameters shown in Fig. 3.14 and the output is the trajectory label (here they are letter-coded, A for trajectory 1, B for trajectory 2 and so on). The model has been implemented in R using the function *multinom()* from package *nnet*. The p-value of the estimated model is < 0.05. The estimated output is shown in Fig. 4.27.

Trajectory A is taken as reference, and it does not appear in the final output. As can be seen from the figure, not all the coefficients are significative (i. e. p-value < 0.05). Anyway, some of them are so and a possible, but cautious, interpretation of the model output is the following. It is important to point out that the coefficients ($\boldsymbol{\beta}$) shown in Fig. 4.27 must be exponentiated to have a more intuitive interpretation (odds ratio). If a certain patient underwent relapse after 18 month, its logit of the probability to belong to trajectory B with respect to A will increase, keeping the other variables constant. A transplanted patient would have an increasing logit to stay in trajectory D rather than A, keeping constant the remaining covariates. On the other hand, its logits of belonging to trajectories H-K with respect to A decrease. Also following a maintenance therapy



	Dependent variable:									
	B	C	D	E	F	G	H	I	J	K
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
RELAPSE_18	2.788**	0.464	-14.070***	1.141	-14.859***	1.692	0.154	-19.184***	-17.723***	-19.703***
	(1.340)	(1.702)	(0.00002)	(1.837)	(0.00002)	(1.907)	(2.154)	(0.00000)	(0.00001)	(0.00000)
TRANSPLANT1	0.271	-1.774	20.753***	-29.457***	-0.942	-1.917	-26.018***	-27.818***	-23.126***	-27.324***
	(1.346)	(1.286)	(0.294)	(1.189)	(1.547)	(1.831)	(0.00000)	(0.00000)	(0.940)	(1.119)
MAINTENANCE1	-0.202	-0.541	22.036***	-2.160	-0.007	-23.072***	-25.148***	-1.876	-2.277	-31.577***
	(1.012)	(1.125)	(0.294)	(1.670)	(1.434)	(0.00000)	(0.00000)	(1.914)	(1.859)	(0.000)
IRCR	0.598	0.551	-18.434***	-28.033***	25.153***	-3.564	-7.743***	-6.305	-42.693***	-48.582***
	(1.445)	(1.627)	(0.00000)	(1.245)	(0.750)	(190.155)	(0.0002)	(936.664)	(0.00000)	(0.00000)
FLCR1	0.973	1.240	-0.165	29.445***	-24.412***	-12.458	-7.698***	-7.678	23.181***	27.779***
	(1.341)	(1.595)	(1.508)	(0.968)	(0.750)	(1,523.846)	(0.0002)	(936.664)	(0.940)	(1.119)
Constant	-1.398	0.302	-43.228***	0.758	-0.600	0.579	1.618	1.618	1.105	1.618
	(1.541)	(1.376)	(0.294)	(1.665)	(1.740)	(1.793)	(1.653)	(1.653)	(1.693)	(1.653)
Akaike Inf. Crit.	269.758	269.758	269.758	269.758	269.758	269.758	269.758	269.758	269.758	269.758
Note:								*p<0.1;	**p<0.05;	***p<0.01

Figure 4.27: Coefficients of the multinomial logit model implemented using the clinical parameters. To have a more intuitive interpretation, coefficients must be elevated to exponent.

would increase the logit to belong to trajectory D with respect to A. Patients which both underwent relapse before 18 month and was not transplanted increases its logit to be in trajectory I rather than A, keeping in turn constant the other variables. A patient with CR to induction therapy has a decreasing logit of belonging to trajectories J and K with respect to A. Cautiously, it seems that trajectories H-K, which are the more stable ones, are associated with no transplant, early relapse and poor induction response (no CR). By exponentiating the coefficients, the increase of the logit is transformed into an increase of the odds ratio, which in turn is related to the probability. Unfortunately, the model has an Akaike's Information Criterion quite high, suggesting for an overfitting.

Moreover, this kind of classification has some limitations. The first one is that, at each execution of the Dirichlet process, the number and the type of the clusters found change. This is due to the fact that the Dirichlet Process has a random component inside the algorithm. This can be overcome by selecting the 'best' clustering using the Silhouette score and then computing the percentage of misclassified points. In addition, the final classification includes only a part of the entire dataset (57 patients over 80), while the remaining subjects are classified as singlets.

4.4 Discussion

4.4.1 Variant analysis

Firstly, the variants found in regions available for both ILM and ONT are considered. Mutation $p.R132H$ has been found in four patients, all males and under 60 years old. It is always detected by both ILM and ONT with similar VAFs and large depth. $p.R132C$ and $p.G105G$ were each found in one patient. $R132$ is known to be a recurrent mutation in GBM patients, associated to a better prognosis with respect to *IDH*-wildtype cancer.

ONT detects several variants in *IDH2* but, unfortunately, only one is in a region available also for ILM, which, however, did not reveal it. Mutations in *IDH2* are mainly characterized by $G > A$ substitutions.

H3-3A is characterized only by $C > T$ transition, found in all the patients. This is known to be a single nucleotide variant present in most of the Caucasian people.

Considering *TERT*, from ILM data, many $G > A$ transitions were found, but they have not been revealed by aligned ONT data. So, they were considered as misalignment artifacts. Only one mutation, a $G > A$ transition, was found from both ONT and ILM data in 16 patients. *TERT* is a gene characterized by several repeated regions. Short-read sequencing techniques, such as Illumina, are heavily affected by this issue: short reads are more difficult to be aligned especially when the region of interest is present several times along the whole genome. In some particular circumstances, aspecific amplification products may be generated during the PCR process, and subsequentially mapped in the same region of interest giving false positive calls. On the contrary, with the chance to sequence long reads, as with ONT, a more flexible and specific primer design can be performed avoiding aspecific amplification due to the presence of repetitive sequences.

In general, variant calls on data from ILM and ONT do agree. The VAFs are similar for both methods. ILM is considered the gold standard in variant detection. Its high phred quality score > 30 makes it reliable for variant calling. ONT has a lower quality score, the studied reads have a score not greater than 20, but allows to detect all the mutations in turn found from ILM data.

On the other hand, ILM revealed its limitations in *TERT* analysis, while with ONT is possible to sequence longer reads and detect a higher number of possible mutations. Moreover, longer reads sensibly reduce the problem of misalignment in repeated regions.

4.4.2 Methylation analysis

MGMT is a key gene for the study of Glioblastoma and its methylation have an impact on the survival after chemotherapy and radiotherapy treatment.

Reg1 was found to be unmethylated in 8 patients, considering that reads were available only for 10 patients over a total of 24.

Reg2 is methylated for 20 patients and Reg3 for 15. Four patients have both Reg2 and Reg3 being unmethylated. Moreover, two patients show an average methylation percentage lower than 20% for both Reg1 and Reg3 and lower than 40% for Reg2. Both Ex3 and Enhanc have an average percentage of methylation greater than 20%, even $> 50\%$ for most of patients.

Ex5 has been sequenced only by ONT and resulted to be methylated for all the patients, with mean percentages higher than 90% in 22 subjects.

The methylations of the sequenced regions estimated using both data from ONT and ILM agree. The Wilcoxon test confirmed that there are no significant differences (p-value > 0.05) between the median of the methylation percentages estimated from ILM and ONT data.

The agreement analysis led to interesting results. Reg1 suffers for poor agreement between ILM and ONT results. It is probably due to the poor number of available reads for ILM. Other regions showed a better agreement, in particular Enhanc and Ex3, which are characterized by high methylation percentage, and Reg2.

By analysing patient by patient, it has been found that, for some of them, agreement is very poor. A possible reason could be related to scarse amount of starting DNA material, which was then degraded during Sodium Bisulfite treatment (usually about $80 - 90\%$ of DNA is degraded during this process). Therefore, it is important to collect sufficient DNA to be processed accurately. Low starting DNA means low number of reads and related low accuracy in the quantification process of DNA methylation level. Nanopore sequencing allows instead for methylation detection without need for Bisulfite treatment. This is a crucial point in the comparison between ILM and ONT: the latter avoids DNA degradation issue.

4.4.3 Hierarchical Dirichlet Clustering

The first attempted approach has been the Hierarchical Bivariate Dirichlet Clustering, which tried to identify bivariate-shaped clusters across the entire cohort of patients, considering a subset of the initial features. The algorithm only found clusters along the main diagonal. After 10000 iterations, more than 50 clusters were found, and they result to be redundant. In addition, clusters on the main diagonal describe CN pairs which basically did not change between diagnosis and relapse.

The second approach separated diagnosis from relapse. A Hierarchical Univariate Dirichlet Clustering has been performed on both diagnosis and relapse, separately. The result consists in the identification of evolutionary trajectories. Patients belonging to them are characterized by common portions of genome which evolve in the same way. Smaller groups have portions common to all the subjects, while larger groups have genes that are shared by most of patients. The scatter plots of CND vs CNR for each trajectory show quite clear evolutionary patterns, which can be assigned to a label. However, the found trajectories do not have statistical significance with the available clinical parameters. A simple multinomial logit model confirmed that.

The limits of both the hierarchical approaches can be sought in different causes. The first reason must be sought in the profiles of the patients forming the cohort under study. These subjects show a great heterogeneity in the CND vs CNR scatter plots. Different groups of CNs are present in each patient. This would affect the final result of the hierarchical approach, which cannot identify groups that are shared by a few patients as features common to the whole cohort. Moreover, the great majority of patients have many CND which slightly differ from CNR or do coincide. They heavily affect the final clustering, which in fact found groups only along the main diagonal. It is possible to remove from the analysis those CN pairs which differ less than 0.1, for instance, and try to run the algorithm on the remaining dataset. A biological issue could arise: the genes which refer to those CNs may be important for the characterization of the malignancy. So, it is important to find a compromise between the statistical principle of parsimony and the biological interpretation which needs to consider all the genetic data.

Conclusion

Next Generation Sequencing systems play a crucial role in the analysis of DNA from cancer cells. Illumina and Oxford Nanopore Technologies provide two different sequencing strategies. The first one represents the gold standard in variant detection thanks to its exceptional quality score. Because of the several chemical reactants needed for the analysis, it results to be costly. The latter is a cheaper and faster solution but has a lower phred quality score.

The analysis of a cohort of 24 Glioblastoma patients allowed to compare these two technologies and assess their agreement. Five patients resulted to be *IDH*-mutant for both data from ILM and ONT, all males and under 60 years old. $p.R132H$ and $p.R132C$ mutations have been detected, while $p.G105G$ has been revealed in one patient only from ONT data. The other genes, *IDH2*, *H3-3A* and *TERT*, did not show any variants in coding regions, apart from some point mutations in non-coding regions. Despite its lower phred quality score, all the variants found from ILM data have been found in turn using ONT data. The VAFs associated to each mutation were similar. Moreover, since ONT allows for long read length, it resulted in a better alignment for *TERT*, that is characterized by repeated regions, avoiding many artifacts. Some additional mutations have been called from ONT data, probably because its lower quality score.

Methylation analysis has been conducted with data both from ILM and ONT. Six regions of *MGMT* were evaluated, five of them common to both the methods (only Ex5 has read available only for ONT). The enhancer region resulted to be methylated for all the patients, as well as exons 3 (it has all patients with mean methylation larger than 20%) and 5. However, their role in the onset and evolution of the tumor is not yet fully understood. Promoter regions Reg1, Reg2 and Reg3 have respectively 2, 20 and 15 patients with average methylation percentage larger than 20%. The Wilcoxon test stated the agreement between results from ILM and ONT.

Also, CpG site methylation percentages have been compared. Reg1, which has a low number of reads for ILM, shows some discrepancies between ILM and ONT estimate of methylation. Things became better for Reg2 and Reg3, up to Ex3 and the enhancer, for which the agreement is high. The Bland-Altman plots confirmed these results. The Lin's Concordance Correlation Coefficient estimated the agreement between the two results.

All these considerations led us to affirm that the results from the two sequencing methods do agree. Despite for some additional point mutations, which, however, are not catalogued in COSMIC (Catalogue Of Somatic Mutations In Cancer [59]), variants found by ILM are found in turn from ONT data. Even the methylation analysis, which usually does not require for Bisulfite treatment in case of nanopore sequencing, led to same results.

Considering the problem of the classification of the cohort of 80 Multiple Myeloma patients using Copy Number Variation measurements and Hierarchical Dirichlet clustering, different approaches have been adopted.

Hierarchical Multivariate Dirichlet Process has been used firstly to remove the CN pairs that are in a normal condition and then applied to the remaining set of CNs. As a result, many clusters lying on the main diagonal have been found. This approach failed to capture those CN groups out from the diagonal, which represent changes in the initial state after the therapy. Such a result was not surprising: by having a look at the patients' trajectories, i. e. CND vs CNR plots, almost all the subjects are characterized by lot of CN pairs placed along the main diagonal, while only a small fraction of pairs in different positions (which vary from patient to patient). This fact could lead us to suppose that the therapies followed by patients basically left the situation unchanged for the majority of them, without significantly modify the CN distribution.

A Hierarchical Univariate Clustering approach has been tested. Here, CND and CNR have been clustered separately. The result consisted in a set of clusters, 10 for diagnosis and 11 for the relapse, basically centered around 0, 1, 2, 3 and 4. Components similar to each other have been merged and then each patient was given a profile, i. e. the components which describe it in diagnosis and relapse (separately). Then, subjects which shared initial and final profile were grouped together, forming evolutionary trajectories. These trajectories have been analysed, finding groups of genes which shared the same behaviour among patients. The scatter plots denoted some similarities in the evolution trend and the multinomial logit model found some relationships between the trajectories and some clinical variables. However, this kind of approach suffers for some limitations, starting from the fact that it could identify a trajectory for each patient, but several of them were not shared by more than two patients and in some cases not even by two. This result reflects a peculiar characteristic of Multiple Myeloma, that is its variability: it is difficult to find patients who evolve according to the same trajectory.

So, the fact that both the hierarchical approaches show limitations could be explained, in part, by the great heterogeneity in the cohort of patients. These patients are characterized by a high number of CNs lying on the main diagonal, which were learned by the algorithm as common features between the patients, while only a portion of them standing outside it. The fact that these latters vary among the subjects could probably be the reason why they have not been found by the algorithm. On the other hand, the univariate approach put together patients with branched and drifted trajectories. It could mean that therapies had an effect on a part of genome of these patients, but probably ‘wrong’ targets have been modified.

To conclude, finding clusters shared between patients will be helpful in defining new features that can be used to study the cancer, its heterogeneity, and the treatment response. A possible solution to the problem of Hierarchical Dirichlet Clustering could be to perform the algorithm on those CNs whose genes are known to be crucial for the onset or development of the malignancy. In addition, the dataset must be increased, so to have more samples to compare, which could help in finding common features among patients. A possible future direction could be also the integration of gene expression data, to have a better understanding of which of these genes characterized by a CNV different from 2 actually have an impact on the onset and progression of the tumor.

Appendix A

Biological and Bioinformatics Glossary

Here a non-exhaustive biological and bioinformatics glossary is reported, in order to help in the reading.

- Alignment = it is an operation that arranges the DNA sequences to identify similarities or differences between them.

- Amplicon = a DNA (or RNA) fragment that is used as a source for an amplification process. It usually consists in repetitive sequences. It can be artificially or naturally synthesized.

- BAM file = Binary Alignment Map, is a tab-delimited text file which contains sequence alignment data.

- Coverage = the sequence coverage (or depth) in DNA sequencing is defined as the average number of reads that include a given nucleotide in the final reconstructed sequence. So, if a portion of the genome is sequenced, we could have several sequencing runs. If a given nucleotide is present in many of these runs, the number of times it is present is the coverage. A high coverage provides a better accuracy and reduces the fraction of sequencing errors.

- FASTQ file = it is a format used for text file that contains information about a biological sequence and its corresponding quality score (from the clusters that pass filter on a flow cell). For a single-read run, one Read 1 (R1) FASTQ file is created for each sample per flow cell lane. For a paired-end run, one Read 1 and one Read 2 (R2) FASTQ file is created for each sample for each lane. Each FASTQ file's entry consists on 4 lines: a sequence identifier, the sequence itself, a separator and the quality score.

- Flow cell = it is a channel for adsorbing mobile DNA fragments. It is also a core sequencing reactor vessel: all the sequencing happens here.

- Genome assembly = is the process of putting nucleotide sequences in the right order to represent the original chromosomes. It is required since sequencing products are much shorter than genes.

- GRCh38/hg38 = is the assembly of the human genome released December of 2013.

- Library = a (genomic) library is a collection of fragments of DNA, that has previously been cut, before an amplification reaction, stored in a population of microorganisms.

- Phred Quality Score = it is a measure of the quality of identified bases generated by a sequencing system. Let P be the base-calling error probability. Then, $Q = -10\, log_{10}(P)$. So, if a base has a quality score of 30, it means that the probability of incorrectly call that base is 1 over 1000. The quality score is stored in the FASTQ file.

- Read = a sequence of DNA which results from a sequencing process. The read length is the length of the read (i.e. the number of base pairs sequenced) and it is sequencing platform specific. The NGS systems are divided into short read length, such as Illumina ($\sim 150 - 200$ bp), and long read length as Oxford Nanopore (~ 1000 bp).

- VCF file = Variant Call Format, is a tab-delimited text file which contains the genomic variants. It is made of a header and a body with all the variants, that is divided into 9 columns: chromosome name, base position, variant identifier, reference sequence, alternative sequence, probability of all samples being homozygous, filter, variant information and individual genotype information.

Appendix B

Pre-processing pipeline for Copy Number data

Here, the pipeline followed by haematologists to obtain CNV data is presented.

For each subject, SNPs array raw data (CEL files) were analysed with a bioinformatic pipeline including Rawcopy, ASCAT and GISTIC.

Rawcopy (Rawcopy v1.1 R Package (Mayrhofer, Viklund, Isaksson, 2016)). Rawcopy analysis was used in order to normalize Affymetrix arrays, extract quality metrics and obtain raw logR and BAF signals for each SNP array probe. Quality metrics were also produced by using Chromosome Analysis Suite (ChAS) v3.3 program on a windows 10 machine. We kept only samples that pass all quality thresholds defined as: RawCopy MAPD < 0.23, ChAS MAPD < 0.25 and ChAS QC < 10.00.

The raw logR and BAF tracks of all samples that passed the quality checks were used as the input for ASCAT v2.5.2 (P Van Loo, 2005). This analysis produced a genomic copy number track per patient, adjusted and corrected for its relative computed normal cell contamination level. This step removes the effect of imperfect enrichment of tumor cells, enabling in such a way the detection and quantification of subclonal Copy Number Abnormalities (CNAs) tumor fractions. ASCAT samples with ploidy > 3.5, reflecting an ambiguous possible whole genome duplication event, were refitted to match a diploid state for simplicity of analysis.

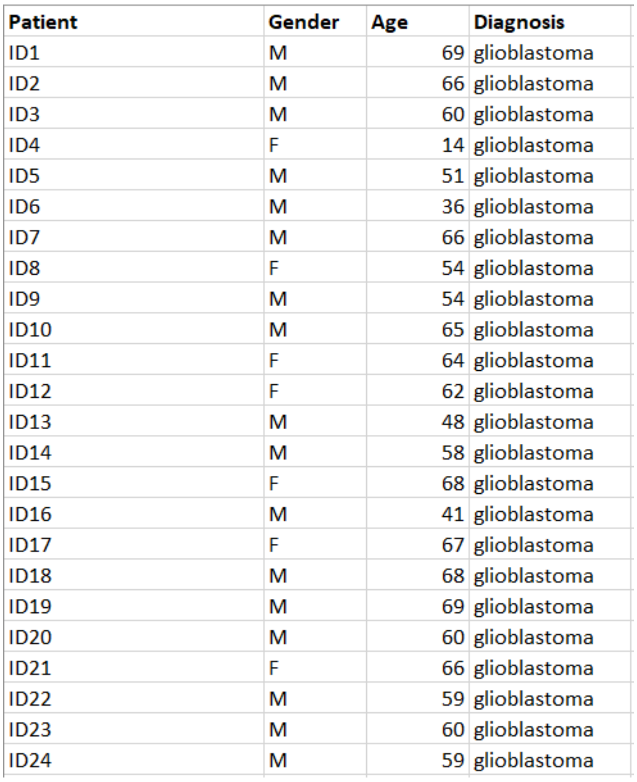
Broad Institute GISTIC v2.0 tool (CH Mermel, 2011) was employed to detect both broad ($> 25\%$ arm length) arm-level CNAs and significative focal CNAs regions. A complete callset for each sample and each chromosome arm was built, keeping in consideration both broad arm-level CNAs calls plus any CNA detected in a focal region. More simply, an arm-level CNAs was called only if it's “broad” ($> 25\%$ arm length) or is located in a genomic position defined as significatively “focal” by GISTIC algorithm.

The entire catalogue of CNAs was drawn up, by describing the distribution of all chromosomal broad CNAs (i.e., covering $> 25\%$ of the chromosomal arm). In addition, GISTIC algorithm (CH Mermel, 2011) was employed to identify a set of focal genomic regions (i.e., covering $< 25\%$ of chromosome arm), with a non-random confluence of highly frequent, small CNAs, covering well-known tumour suppressor genes and oncogenes, widely regarded as relevant in MM biology (e.g. TP53, RB1, MYC, CKS1B).

Appendix C

Tables of Mutations and Methylation

The basic information about the cohort of Glioblastoma patients is shown in the table of Fig. C.1.



Patient	Gender	Age	Diagnosis
ID1	M	69	glioblastoma
ID2	M	66	glioblastoma
ID3	M	60	glioblastoma
ID4	F	14	glioblastoma
ID5	M	51	glioblastoma
ID6	M	36	glioblastoma
ID7	M	66	glioblastoma
ID8	F	54	glioblastoma
ID9	M	54	glioblastoma
ID10	M	65	glioblastoma
ID11	F	64	glioblastoma
ID12	F	62	glioblastoma
ID13	M	48	glioblastoma
ID14	M	58	glioblastoma
ID15	F	68	glioblastoma
ID16	M	41	glioblastoma
ID17	F	67	glioblastoma
ID18	M	68	glioblastoma
ID19	M	69	glioblastoma
ID20	M	60	glioblastoma
ID21	F	66	glioblastoma
ID22	M	59	glioblastoma
ID23	M	60	glioblastoma
ID24	M	59	glioblastoma

Figure C.1: Information about the cohort of GBM patients.

Four regions of genes of interest for the study of glioma have been sequenced: *IDH1* (exon 4), *IDH2* (exon 4), *H3-3A* (exon 1) and *TERT* (promoter). These genes are known to be important for the characterization of the Glioblastoma. Tables with variants annotated using the standard nomenclature [58] are reported here.

Tab. C.1 shows all the point mutations detected by ILM in the four analysed patients.

Those variants that are thought to be artifacts due to misalignments problems (mainly in *TERT*) are not present.

Variants detected from ONT data are shown in Tab. C.2 and C.3.

Patient	$IDH1$	$IDH2$	$H3-3A$	$TERT$
$ID1$			$g.226064531C > T$ (100%)	$g.1295113G > A$ (42%)
$ID2$			$g.226064531C > T$ (100%)	
$ID3$			$g.226064531C > T$ (100%)	$g.1295113G > A$ (43%)
$ID4$			$g.226064531C > T$ (100%)	
$ID5$	$p.R132C$(44%)		$g.226064531C > T$ (100%)	
$ID6$	$p.R132H$(69%)		$g.226064531C > T$ (100%)	
$ID7$			$g.226064531C > T$ (100%)	$g.1295113G > A$ (44%)
$ID8$			$g.226064531C > T$ (100%)	$g.1295113G > A$ (57%)
$ID9$			$g.226064531C > T$ (100%)	
$ID10$	$p.R132H$(34%)		$g.226064531C > T$ (100%)	
$ID11$			$g.226064531C > T$ (100%)	$g.1295113G > A$ (56%)
$ID12$			$g.226064531C > T$ (100%)	$g.1295113G > A$ (49%)
$ID13$	$p.R132H$(25%)		$g.226064531C > T$ (100%)	
$ID14$			$g.226064531C > T$ (100%)	$g.1295113G > A$ (24%)
$ID15$			$g.226064531C > T$ (100%)	$g.1295113G > A$ (55%)
$ID16$			$g.226064531C > T$ (100%)	$g.1295135G > A$ (64%)
$ID17$			$g.226064531C > T$ (100%)	$g.1295113G > A$ (47%)
$ID18$			$g.226064531C > T$ (100%)	$g.1295113G > A$ (44%)
$ID19$			$g.226064531C > T$ (100%)	$g.1295113G > A$ (47%)
$ID20$			$g.226064531C > T$ (100%)	$g.1295113G > A$ (29%)
$ID21$			$g.226064531C > T$ (100%)	$g.1295113G > A$ (5%)
$ID22$			$g.226064531C > T$ (100%)	$g.1295113G > A$ (54%)
$ID23$			$g.226064531C > T$ (100%)	$g.1295113G > A$ (41%)
$ID24$	$p.R132H$(30%)		$g.226064531C > T$ (100%)	

Table C.1: Mutations detected by ILM for the entire cohort of patients. The VAF for each mutation is present. For mutations in non-coding regions, the position is also reported. From the left: patient ID, genes with their variants annotated following the standard nomenclature [58].

Table with the average methylation percentage for each of the five regions (six for ONT) for each patient for both ILM and ONT data is reported in Tab. C.4. The ONT estimates are within brackets. The standard deviation of the mean has been used as uncertainty estimate.

$Patient$	$IDH1$	$IDH2$	$H3-3A$	$TERT$
$ID1$		$g.90088023g > a$ (49%) $g.90088247G > A$ (10%) $g.90088345C > T$ (14%)	$g.226064531C > T$ (92%)	$g.1295052T > C$ (28%) $g.1295113G > A$ (55%) $g.1295234A > G$ (48%)
$ID2$		$g.90088015g > a$ (48%) $g.90088247G > A$ (9%) $g.90088345C > T$ (15%)	$g.226064531C > T$ (93%)	$g.1295052T > C$ (34%)
$ID3$		$g.900088345C > T$ (12%)	$g.226064531C > T$ (91%)	$g.1295052T > C$ (37%) $g.1295113G > A$ (63%) $g.1295234A > G$ (48%)
$ID4$		$g.90088015g > a$ (88%) $g.90088023g > a$ (9%) $g.90088247G > A$ (8%) $g.900088345C > T$ (14%)	$g.226064531C > T$ (93%)	$g.1295052T > C$ (26%) $g.1295234A > G$ (81%)
$ID5$	$p.R132C$(45%)	/	/	$g.1295052T > C$ (28%)
$ID6$	$p.R132H$(73%)	$g.90088023g > a$ (36%) $g.900088345C > T$ (13%)	$g.226064531C > T$ (94%)	$g.1295052T > C$ (34%) $g.1295234A > G$ (88%)
$ID7$		$g.90088247G > A$ (11%) $g.900088345C > T$ (17%)	$g.226064531C > T$ (94%)	$g.1295052T > C$ (34%) $g.1295113G > A$ (47%) $g.1295234A > G$ (94%)
$ID8$		/	/	$g.1295052T > C$ (38%) $g.1295113G > A$ (66%)
$ID9$		/	/	$g.1295052T > C$ (32%) $g.1295113G > A$ (5%)
$ID10$	$p.R132H$(36%)	/	/	$g.1295052T > C$ (28%) $g.1295234A > G$ (47%)
$ID11$		/	/	$g.1295052T > C$ (33%) $g.1295113G > A$ (67%)
$ID12$		/	/	$g.1295052T > C$ (36%) $g.1295113G > A$ (70%) $g.1295234A > G$ (57%)

Table C.2: Mutations detected by ONT for the remaining part of the cohort. The symbol / means that no mutations were annotated because of the poor number of reads. Lowercase letters refer to intronic regions.

$Patient$	$IDH1$	$IDH2$	$H3-3A$	$TERT$
$ID13$	$p.R132H$(24%)	$g.90088015g > a$ (96%) $g.90088345C > T$ (11%)	$g.226064531C > T$ (94%)	$g.1295052T > C$ (31%) $g.1295338C > T$ (11%)
$ID14$		$g.90088023g > a$ (46%) $g.90088247G > A$ (10%) $g.90088345C > T$ (11%)	$g.226064531C > T$ (95%)	$g.1295052T > C$ (22%) $g.1295113G > A$ (32%) $g.1295234A > G$ (6%)
$ID15$		$g.90088015g > a$ (48%) $g.90088023g > a$ (48%) $g.90088345C > T$ (11%)	$g.226064531C > T$ (95%)	$g.1295052T > C$ (33%) $g.1295113G > A$ (59%) $g.1295234A > G$ (95%) $g.1295338C > T$ (12%)
$ID16$		$g.90088015g > a$ (50%) $g.90088023g > a$ (45%) $g.90088345C > T$ (11%)	$g.226064531C > T$ (95%)	$g.1295052T > C$ (27%) $g.1295135G > A$ (49%) $g.1295234A > G$ (32%)
$ID17$		$g.90088015g > a$ (48%) $g.90088247G > A$ (10%) $g.90088345C > T$ (13%)	$g.226064531C > T$ (94%)	$g.1295052T > C$ (29%) $g.1295113G > A$ (58%) $g.1295234A > G$ (38%)
$ID18$		$g.90088015g > a$ (49%) $g.90088247G > A$ (8%) $g.90088345C > T$ (11%)	$g.226064531C > T$ (96%)	$g.1295052T > C$ (31%) $g.1295113G > A$ (62%)
$ID19$		$g.90088015g > a$ (46%) $g.90088247G > A$ (11%) $g.90088345C > T$ (15%)	$g.226064531C > T$ (94%)	$g.1295052T > C$ (31%) $g.1295113G > A$ (63%) $g.1295234A > G$ (35%)
$ID20$			$g.226064531C > T$ (95%)	$g.1295052T > C$ (35%) $g.1295113G > A$ (45%) $g.1295234A > G$ (39%)
$ID21$		$g.90088015g > a$ (96%) $g.90088345C > T$ (11%)	$g.226064531C > T$ (95%)	$g.1295052T > C$ (30%) $g.1295113G > A$ (13%)
$ID22$	$p.G105G$(29%)	$g.90088247G > A$ (10%) $g.90088345C > T$ (14%)	$g.226064531C > T$ (95%)	$g.1295052T > C$ (29%) $g.1295113G > A$ (58%) $g.1295234A > G$ (50%)
$ID23$		$g.90088015g > a$ (95%) $g.90088247G > A$ (11%) $g.90088345C > T$ (16%)	$g.226064531C > T$ (94%)	$g.1295052T > C$ (30%) $g.1295113G > A$ (58%) $g.1295234A > G$ (42%)
$ID24$	$p.R132H$(32%)	$g.90088015g > a$ (49%) $g.90088247G > A$ (9%) $g.90088345C > T$ (12%)	$g.226064531C > T$ (95%)	$g.1295052T > C$ (32%) $g.1295234A > G$ (94%)

Table C.3: Mutations detected from ONT data for the first half of the cohort of patients. Lowercase letters refer to intronic regions.

Patient	*Enhanc* (%)	*Reg1* (%)	*Reg2* (%)	*Reg3* (%)	*Ex3* (%)	*Ex5* (%)
$ID1$	89.4 ± 0.9 (91.6 ± 0.9)	5.6 ± 1.1 (5.4 ± 1.3)	36 ± 6 (35 ± 7)	38 ± 4 (38 ± 5)	59 ± 4 (54 ± 5)	97 ± 4
$ID2$	93.3 ± 0.8 (86 ± 9)	/	16 ± 5 (26 ± 4)	6 ± 3 (7 ± 4)	81 ± 2 (71 ± 2)	99.0 ± 0.5
$ID3$	83 ± 8 (83 ± 8)	17 ± 3 (16 ± 3)	45 ± 6 (39 ± 5)	43 ± 4 (48 ± 5)	28 ± 3 (30 ± 3)	99 ± 3
$ID4$	92.5 ± 0.9 (90.7 ± 1.2)	2.8 ± 0.6 (1.2 ± 0.4)	40 ± 6 (39 ± 7)	5 ± 3 (6 ± 4)	25 ± 2 (19 ± 2)	98.0 ± 0.9
$ID5$	84 ± 4 (84 ± 5)	12 ± 3 (18 ± 3)	53 ± 6 (49 ± 6)	34 ± 5 (37 ± 5)	26 ± 4 (24 ± 4)	98 ± 6
$ID6$	85 ± 4 (84 ± 5)	14 ± 2 (17 ± 2)	72 ± 3 (62 ± 3)	54 ± 4 (53 ± 4)	22.3 ± 1.3 (23.0 ± 1.5)	92.0 ± 5
$ID7$	79 ± 3 (93 ± 2)	42 ± 4 (13 ± 2)	48 ± 4 (40 ± 5)	22 ± 4 (33 ± 3)	87 ± 3 (72 ± 2)	91.0 ± 1.6
$ID8$	95.1 ± 0.8 (94.8 ± 0.8)	6.6 ± 1.1 (5.8 ± 0.9)	85 ± 2 (78 ± 2)	41 ± 4 (36 ± 5)	49 ± 4 (48 ± 4)	98.0 ± 0.7
$ID9$	94.8 ± 2 (94.6 ± 2)	0.8 ± 0.8 (1.0 ± 0.3)	28 ± 5 (28 ± 5)	6 ± 3 (7 ± 4)	90 ± 5 (90 ± 5)	99.0 ± 0.6
$ID10$	85 ± 4 (84 ± 5)	31 ± 5 (25 ± 4)	73 ± 3 (66 ± 3)	65 ± 5 (63 ± 5)	41 ± 6 (42 ± 5)	80 ± 6
$ID11$	94.1 ± 0.8 (93.8 ± 0.9)	1.5 ± 0.4 (1.5 ± 0.2)	45 ± 9 (42 ± 8)	18 ± 5 (18 ± 5)	67 ± 7 (64 ± 7)	98.0 ± 1.2
$ID12$	85 ± 6 (85 ± 5)	/	47 ± 6 (48 ± 7)	24 ± 5 (28 ± 6)	87 ± 4 (82 ± 4)	97 ± 2
$ID13$	59 ± 4 (59 ± 5)	/	29 ± 4 (29 ± 5)	34 ± 2 (41 ± 4)	73.1 ± 1.0 (57.3 ± 0.9)	96.0 ± 1.2
$ID14$	91.6 ± 1.3 (90.5 ± 0.8)	/	35 ± 5 (34 ± 4)	12 ± 3 (19 ± 4)	85 ± 5 (84 ± 4)	99.0 ± 0.7
$ID15$	92.5 ± 0.9 (91.9 ± 1.1)	/	67 ± 4 (59 ± 4)	44 ± 5 (42 ± 4)	85 ± 2 (87 ± 2)	99.0 ± 0.9
$ID16$	86 ± 7 (87 ± 7)	/	18 ± 5 (26 ± 6)	5 ± 3 (10 ± 4)	53.3 ± 1.1 (56.5 ± 1.2)	97.0 ± 0.7
$ID17$	87 ± 4 (89 ± 3)	/	62 ± 5 (56 ± 5)	46 ± 4 (53 ± 5)	55 ± 6 (61 ± 5)	99 ± 3
$ID18$	96.9 ± 0.5 (97.2 ± 0.3)	/	65 ± 5 (49 ± 6)	35 ± 5 (43 ± 6)	83 ± 4 (86 ± 4)	99.0 ± 0.5
$ID19$	77 ± 6 (78 ± 6)	/	14 ± 3 (14 ± 3)	6 ± 3 (10 ± 4)	59.3 ± 1.2 (62.3 ± 1.5)	99.0 ± 0.8
$ID20$	64 ± 4 (72 ± 3)	/	28 ± 5 (25 ± 5)	23 ± 4 (19 ± 5)	73 ± 2 (75 ± 2)	96.0 ± 1.2
$ID21$	93.5 ± 0.9 (94.0 ± 1.3)	/	40 ± 5 (40 ± 5)	18 ± 2 (15 ± 4)	71 ± 2 (68 ± 2)	98.0 ± 0.6
$ID22$	88 ± 4 (85 ± 5)	/	14 ± 4 (14 ± 4)	5 ± 2 (10 ± 4)	87 ± 4 (87 ± 3)	81 ± 3
$ID23$	85 ± 6 (85 ± 5)	/	78 ± 4 (70 ± 3)	32 ± 4 (38 ± 5)	93.7 ± 0.9 (90.9 ± 1.3)	99 ± 9
$ID24$	97.2 ± 0.4 (96.8 ± 0.3)	/	62 ± 6 (55 ± 6)	48 ± 5 (52 ± 6)	26.9 ± 0.6 (29.8 ± 0.6)	98.0 ± 0.3

Table C.4: Mean methylation of the five regions of MGMT sequenced by both ILM and ONT (six regions). The uncertainty has been estimated through the standard deviation of the mean. Nanopore results are within brackets. The symbol / indicates a poor number of reads and thus results are not considered reliable.

Bibliography

[1] Sung H. et al. “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.” In: *CA: A Cancer Journal for Clinicians* 71.3 (2021), pp. 209–249. DOI: 10.3322/caac.21660.

[2] AIRC. *Il cancro in cifre.* URL: https://www.airc.it/cancro/informazioni-tumori/cose-il-cancro/numeri-del-cancro. (accessed: 09.2021).

[3] Alberts B. et al. *Molecular Biology of the Cell.* Fifth edition. Garland Science, June 2008. URL: https://archive.org/details/MolecularBiologyOfTheCell5th_201802/.

[4] Zwolak M. and Di Ventra M. “Colloquium: Physical approaches to DNA sequencing and detection.” In: *Reviews of Modern Physics* 80.1 (2008), pp. 141–166. DOI: 10.1103/RevModPhys.80.141.

[5] Wilson K. and Walker J. *Wilson and Walker’s Principles and Techniques of Biochemistry and Molecular Biology.* Eight edition. Cambridge University Press, June 2018.

[6] Schatz M. C. “Nanopore sequencing meets epigenetics.” In: *Nature methods* 14.4 (April 2017), pp. 347–348. DOI: 10.1038/nmeth.4240.

[7] Nussbaum R. L., McInnes R. R., and Willard H. F. *Thompson and Thompson Genetics in Medicine.* Eigth edition. Elsevier, 2016, pp. 43–56. URL: https://archive.org/details/thompson-thompson-genetics-8th-edition-medicoscompanion.com/.

[8] Maccarone M. *Metodologie biochimiche e biomolecolari.* First edition. Zanichelli, 2019.

[9] Grice E. A. and Hodkinson B. P. “Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research.” In: *Advances in wound care* 4.1 (2015), pp. 50–58. DOI: 10.1089/wound.2014.0542.

[10] Clarke J. et al. “Continuous base identification for single - molecule nanopore DNA sequencing.” In: *Nature nanotechnology* 4 (April 2009), pp. 265–270. DOI: 10.1038/NNANO.2009.12.

[11] Kasianowicz et al. “Characterization of individual polynucleotide molecules using a membrane channel.” In: *Proceedings of the National Academy of Sciences USA* 93.24 (1996), pp. 13770–13773. DOI: 10.1073/pnas.93.24.13770.

[12] Manrao E. et al. “Nucleotide Discrimination with DNA Immobilized in the MspA Nanopore.” In: *PLoS ONE* 6.10 (October 2011). DOI: 10.1371/journal.pone.0025723.

[13] Zwolak M. and Di Ventra M. “DNA Sequencing via Electron Tunneling”.

[14] Stoddart D. et al. “Multiple base-recognition sites in a biological nanopore – two heads are better than one.” In: *Angewandte Chemie International Edition* 49.3 (2010), pp. 556–559. DOI: 10.1002/anie.200905483.

[15] Stoddart D. et al. “Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore.” In: *Proceedings of the National Academy of Sciences USA* 106.19 (May 2009), pp. 7702–7707. DOI: 10.1073/pnas.0901054106.

[16] Bishop C. *Pattern Recognition and Machine Learning.* First edition. Springer, January 2006.

[17] Ferguson T. S. “A Bayesian analysis of some nonparametric problems.” In: *The Annals of Statistics* 1.2 (1973), pp. 209–230. DOI: 10.1214/AOS/1176342360.

[18] Teh Y. W. *Dirichlet Process.* URL: https://www.gatsby.ucl.ac.uk/~ywteh/research/npbayes/Teh2010a.pdf. (accessed: 05.2021).

[19] Sudderth E. B. “Graphical Models for Visual Object Recognition and Tracking.” PhD thesis. MIT, 2006.

[20] Blackwell D. and McQueen J. B. “Ferguson distribution via polya urn schemes.” In: *The Annals of Statistics* 1.2 (1973), pp. 353–355. DOI: 10.1214/AOS/1176342372.

[21] Sethuraman J. “A constructive definition of Dirichlet priors.” In: *Statistica Sinica* 4.2 (1994), pp. 639–650.

[22] Teh Y. W. et al. “Hierarchical Dirichlet Processes.” In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1566–1581. DOI: 10.1198/016214506000000302.

[23] *Glioblastoma.* URL: https://en.wikipedia.org/wiki/Glioblastoma. (accessed: 08.2021).

[24] Young R. M. et al. “Current trends in the surgical management and treatment of adult glioblastoma.” In: *Annals of Translational Medicine* 3.9 (2015). DOI: 10.3978/j.issn.2305-5839.2015.05.10.

[25] Tan A. C. et al. "Management of Glioblastoma: State of the Art and Future Directions." In: *CA: A Cancer Journal for Clinicians* 70.4 (2020), pp. 299–312. DOI: 10.3322/caac.21613.

[26] Louis D. N. et al. "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary." In: *Acta Neuropathol* 131 (2016), pp. 803–820. DOI: 10.1007/s00401-016-1545-1.

[27] Illumina. *MiSeq System.* URL: https://www.illumina.com/systems/sequencing-platforms/miseq.html. (accessed: 08.2021).

[28] *MiSeq™ System.* M-GL-00006 v1.0. Illumina. 2021.

[29] Oxford Nanopore Technologies. *MinION.* URL: https://nanoporetech.com/products/minion. (accessed: 07.2021).

[30] *MinIon.* BR-1002(EN)-V5. Oxford Nanopore Technologies. 2021.

[31] Hai Yan M. D. et al. "IDH1 and IDH2 Mutations in Gliomas." In: *New England Journal of Medicine* 360.8 (2009), pp. 765–773. DOI: 10.1056/NEJMoa0808710.

[32] Bjerke L. et al. "Histone H3.3 Mutations Drive Pediatric Glioblastoma through Upregulation of MYCN." In: *Cancer Discovery* 3 (May 2013), pp. 512–519. DOI: 10.1158/2159-8290.CD-12-0426.

[33] Morandi L. et al. "Promoter methylation analysis of O6-methylguanine-DNA methyltransferase in glioblastoma: detection by locked nucleic acid based quantitative PCR using an imprinted gene (SNURF) as a reference." In: *BMC Cancer* 10.48 (2010), pp. 512–519. DOI: http://www.biomedcentral.com/1471-2407/10/48.

[34] Luca Morandi. *Next Generation Sequencing in the Clinic.* URL: https://virtuale.unibo.it/pluginfile.php/595160/mod_unibores/content/0/Prt3MorandiNGS2020.pdf (visited on 08/11/2021).

[35] *Galaxy Europe Project.* URL: https://galaxyproject.eu/. (accessed: 07.2021).

[36] Afgan E. et al. "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update." In: *Nucleic Acids Research* 44 (2016), W3–W10. DOI: 10.1093/nar/gkw343.

[37] Sovic I. et al. "Fast and sensitive mapping of nanopore sequencing reads with GraphMap." In: *Nature Communications* 7.11307 (2016). DOI: 10.1038/ncomms11307.

[38] De Koning W. et al. "NanoGalaxy: Nanopore long-read sequencing data analysis in Galaxy." In: *GigaScience* 9 (2020), pp. 1–7. DOI: 10.1093/gigascience/giaa105.

[39] *NanoGalaxy.* URL: https://nanopore.usegalaxy.eu/. (accessed: 08.2021).

[40] Robinson J. T. et al. “Integrative Genomics Viewer.” In: *Nature Biotechnology* 29.1 (Jan 2011), pp. 24–26. DOI: 10.1038/nbt.1754.

[41] Krainer J. et al. “EPIC-TABSAT: analysis tool for targeted bisulfite sequencing experiments and array-based methylation studies.” In: *Nucleic Acids Research* 47 (2019), pp. 1–5. DOI: 10.1093/nar/gkz398.

[42] Bland J. M. and Altman D. G. “Statistical methods for assessing agreement between two methods of clinical measurement.” In: *Lancet* 327.8476 (1986), pp. 307–310. DOI: 10.1016/S0140-6736(86)90837-8.

[43] Bland J. M. and Altman D. G. “Measuring agreement in method comparison studies.” In: *Statistical Methods in Medical Research* 8 (1999), pp. 135–160. DOI: 10.1177/096228029900800204.

[44] Lin L. I-K. “A Concordance Correlation Coefficient to Evaluate Reproducibility.” In: *Biometrics* 45 (March 1989), pp. 255–268. DOI: 10.2307/2532051.

[45] Michels T. C. and Petersen K. E. “Multiple Myeloma: Diagnosis and Treatment.” In: *American Family Physician* 95.6 (2017), pp. 373–384.

[46] van de Donk N. W. C. J. et al. “Multiple myeloma.” In: *Lancet* 397 (2021), pp. 410–427.

[47] Rajkumar S. V. “Multiple myeloma: 2020 update on diagnosis, risk-stratificationand management.” In: *American Journal of Hematology* 3.95 (2020), pp. 548–567. DOI: 10.1002/ajh.25791.

[48] HARMONY Alliance. *Multiple Myeloma.* URL: https://www.harmony-alliance.eu/en/focus/multiple-myeloma. (accessed: 06.2021).

[49] Walker B. A. et al. “Identification of novel mutational drivers reveals oncogene dependencies in multiple myeloma.” In: *Blood* 132.6 (2018), pp. 587–597. DOI: 10.1182/blood-2018-03-840132.

[50] *Multiple myeloma.* URL: https://en.wikipedia.org/wiki/Multiple_myeloma. (accessed: 05.2021).

[51] Durie B. G. M. et al. “International uniform response criteria for multiple myeloma.” In: *Leukemia* 20 (2006), pp. 1467–1473. DOI: 10.1038/sj.leu.2404284.

[52] Bolli N. et al. “Heterogeneity of genomic evolution and mutational profiles in multiple myeloma.” In: *Nature Communications* 5.2997 (2014). DOI: 10.1038/ncomms3997.

[53] Jones J. R. et al. "Clonal evolution in myeloma: the impact of maintenance lenalidomide and depth of response of the genetics and sub-clonal structure of relapsed disease in uniformly treated newly diagnosed patients." In: *Haematologica* 104.7 (2019), pp. 1440–1450. DOI: 10.3324/haematol.2018.202200.

[54] Ross G. J. and Markwick D. *dirichletprocess: An R Package for Fitting Complex Bayesian Nonparametric Models.* URL: https://rdrr.io/cran/dirichletprocess/f/inst/doc/dirichletprocess.pdf. (accessed: 05.2021).

[55] *The R Project for Statistical Computing.* URL: https://www.r-project.org/. (accessed: 05.2021).

[56] Rousseeuw Peter J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

[57] German Rodriguez. *Lecture Notes on Generalized Linear Models.* 2007. URL: https://data.princeton.edu/wws509/notes/.

[58] den Dunnen J. T. et al. "HGVS Recommendations for the Description of Sequence Variants: 2016 Update." In: *Human Mutation* 37.6 (2016), pp. 564–569. DOI: 10.1002/humu.22981.

[59] *COSMIC: Catalogue Of Somatic Mutations In Cancer.* URL: https://cancer.sanger.ac.uk/cosmic. (accessed: 09.2021).

Acknowledgements

I would like firsly to thank all my Family, for all the support that always gave me, allowing me to study without worries.

I would like to thank Prof. Gastone Castellani for all the opportunities he gave me in these years and for allowing me to perform my thesis project within his research group.

Then, I would like to thank Prof. Luca Morandi for the time he spent in teaching me a lot of biological, sequencing and bioinformatics concepts and for all the suggestions he gave me.

I would like to thank Dott.ssa Alessandra Merlotti for the help in developing the Dirichlet Process project, for the proactive discussions and all the advices she gave me in the last year.

I would like to thank also Dott. Daniele Dall'Olio for all that he taught me in the last year and the patience in explaining me a lot of concepts.

A special thank to Jacopo, a friend of mine, with a master of science in Statistics, which gave me valuable suggestions for my thesis work during our walks.

Finally, I want to thank all my travel companions during the bachelor degree and the master degree, who shared with me part of this long path, during which we became Friends.