UNIVERSITA' DEGLI STUDI DI NAPOLI
"FEDERICO II"

RESEARCH DOCTORATE IN

COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

XXVIII Cycle

A novel qualitative approach to analyse DNA methylation data from Deep Bisulfite Amplicon Sequencing (Deep- Bis)

Tutor

Prof. Sergio Cocozza

Candidate

Ornella Affinito

Co- Tutor

Prof. Gennaro Miele

March 2016

Table of contents

Chapter 1

Introduction

1.1 DNA methylation: a general overview

Epigenetics is referred to changes in gene expression, which are heritable through multiple cell division cycles. These changes result from a set of reversible modifications that occur without alterations in the DNA sequence and include nucleic acid modification, chromatin remodelling and histone modifications [1]. Regarding nucleic acid modification, DNA methylation is a common epigenetic mark in many eukaryotes, involved in the regulation of gene expression [2,3]. DNA methylation has been recognized to play an important role in many biological processes, such as cellular differentiation [4], development [5], disease [6], aging [7], X-inactivation [8], imprinting [9], silencing of repetitive DNA (i.e. transposons) [10] and chromosomal stability [11].

In mammals, the predominant form of DNA methylation is the covalent attachment of a methyl group to the 5′ position of the cytosine residues in CpG context, although there is evidence that cytosine methylation is not limited to those in CpG sequences. Methylation on cytosine in other sequence context (CHH e CHG, with H=A, T or G) is widespread in plants [12] and some fungi [13] and has recently been reported in mammals [14,15]. Cytosine methylation results in transcriptional repression.

Methylation patterns of adult somatic cells are mostly determined during embryogenesis and then it is stably inherited through mitotic divisions, passing over to differentiating cell and tissues [16-18]. There are two types of DNA methylation: i) a stable and invariant form that represents the basis for imprinting and is sex-specific and identical in individuals and cells [9]; ii) a metastable somatic type that changes with age and differs among individuals and cells [19].

DNA methylation is mediated by DNA methyltransferases (DNMTs), which are essential for development and viability [20-21] and are responsible to establish methylation pattern in early development and maintain it during cell division [22]. They are classified into two groups: maintenance and *de novo* methylases. Maintenance methylases DNMT1 shows a preference for the hemi-methylated DNA in order to propagate the existing methylation from the old strand of the mother cells to the newly synthesised strand of the daughter cells during DNA replication. DNMT3A and DNMT3B are *de novo* methyltransferases, responsible for the establishment of DNA methylation patterns early in mammalian development and in germ cells. There is another member of the family, DNMT3L, which is catalytically inactive but interacts with DNMT3A and DNMT3B, facilitating their enzymatic activity. DNMT3L is best known for its role in imprint methylation maintenance during gametogenesis. In the absence of DNMT3L, genomic imprinting is lost [23-24]. The categorisation of the above-mentioned methyltransferases as either maintenance or *de novo* methyltransferases is an oversimplification as DNMT1 can have *de novo* activity as well [25].

On the other side, DNA demethylation can takes place through two mechanisms: passive and active demethylation. In the passive demethylation, 5-methylcytosine (5mC) can be lost or erased during cell division, when proteins that can copy it (DNMTs) are absent or non-functional. In contrast, active demethylation can happen without DNA replication and cell division. In this case, the 5- methylcytosine (5mC) is oxidised by the TET (Ten Eleven Translocation) proteins, generating 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) [26].

In the human genome, both the CpG sites and their degrees of methylation are unevenly distributed in the genome [27-28]. Two fractions with distinct properties are distinguishable: a major fraction (~98%), in which CpGs are relatively infrequent (on average 1 per 100 bp) but highly methylated (approximately 70%–80% of all CpG sites), and a minor fraction (<2%) that comprises short stretches of DNA (approximately 1 kb in length and longer than 200bp) in which CpG sites are frequent (~1 per 10 bp; CpG- rich regions), G+C base content is high (above 50% G+C content) and the observed-to-expected CpG ratio is greater than 60 %. The latter are known as CpG islands (CGIs) and they are usually methylation- free. They are found within the promoters of ~60-70% of human genes [29-30]. Due to their unmethylated status, CGI- promoters are characterised by a transcriptionally permissive chromatin state and are indeed generally associated with constitutively expressed genes in all cell type (housekeeping genes), but a subset of them

may be subject to tissue- specific gain of methylation [31] or during the development [32], resulting in a stable transcriptional repression. However, CGI hypermethylation is required for the long-term silencing of genes located on the inactive X chromosome [8] or associated with imprinted loci [9], germline-specific genes [33] and pluripotency-associated genes [34].

Aberrant DNA methylation is associated with many human diseases and is a hallmark of cancer [6]. These epigenetic changes impact the biological activity of cells through their modification of transcriptional states and regulatory machinery. Hypermethylation of CGI promoters may contribute to carcinogenesis by inactivating tumor suppressor genes or DNA repair genes, while hypomethylation contributes to carcinogenesis by activating oncogene or promoting genomic instability.

1.2 Heterogeneous DNA methylation

Tissues exhibit cellular heterogeneity, arising from genetic, epigenetic or cell- specific transcriptional mechanisms, which confer cellular identity [35-36]. In a mixed cell population (such as a tissue), cells may demonstrate similar phenotypes but with distinct methylation patterns at a specific genomic locus, or different phenotypes. Moreover, the heterogeneity in cellular composition was recognized as an important confounding factor that could compromise the resulting interpretations for methylation studies [37-38]. These findings emphasize the importance of examining the methylation pattern heterogeneity within a cell population or between different cell types.

Mammalian genome contains about 29 million of CpG sites, which are non-randomly distributed along the genome [27-28]. Because each of CpG sites may exist in a methylated or unmethylated state, the number of possible combinations is huge ($10^{8,700,000}$) and may therefore enormously increase the potential information content of genomic DNA, without considering the further increase provided by other cytosine modifications such as hydroxymethylation and non- CpG methylation. Although at genomic level it is practically impossible, or at least a very hard task, to verify all the possible ^{m}CpG combinations present in the cells of a tissue, in principle each cell may bear a specific combination of methylated CpGs at specific loci (epiallele) that may reflect the origin of the cell and/or the functional state of a given gene in that cell. These epialleles differ in their pattern of methylated and unmethylated CpG positions. These molecules (epialleles) can be grouped

in methylation classes, defined as number of methylated cytosines per molecule, independently of their position.

This introduces the concept of epipolymorphism, by means that cells of a tissue may be considered a population of epigenetically heterogeneous cells in which each combination of ^{m}CpG at a given locus represent a specific epiallele. All of these epialleles show variable frequencies. It has been reported that, unlike H1 embryonic stem cells (ESCs) and testicular cells, somatic tissues are highly polymorphic, exhibiting a spectrum of heterogeneous methylation states that range from a complete lack of methylation to full methylation [39].

Closely connected to the epipolymorphism is the concept of methylation heterogeneity. The terms of methylation heterogeneity is used to refer to a mixture of multiple epialleles in varying proportions, which differ in the pattern of methylated and unmethylated CpG sites. In the case of homogeneous methylation pattern, all the cells share the same methylation state, indicating strong constraints in methylation control for the entire cell population. On the opposite side, there are different levels of methylation heterogeneity. At the simplest level, it has been used to refer to a mixture of fully- methylated and unmethylated alleles. Alternatively, a heterogeneous mixture of cells may comprise a spectrum of epialleles with all possible methylation patterns, spanning from a complete lack of methylation to full methylation. Genomic loci with different methylation heterogeneity may share the same average level of methylation.

The occurrence of a uniform pattern of DNA methylation indicates a high fidelity of methylation inheritance. In contrast, DNA methylation heterogeneity can arise in a variety of ways including but not limited to: (i) more than a single population of cells is analysed that differ in DNA methylation at the locus of interest (cell- specific DNA methylation); (ii) the locus of interest is imprinted, i.e. two different epialleles (unmethylated and fully-methylated) are present in each cell (allele- specific DNA methylation); (iii) the locus is inherently heterogeneous in its DNA methylation composition; or iv) a decreased fidelity of methylation inheritance, which can give arise to an asymmetric DNA methylation (hemi-methylated CpG/CpG dyads), indicating a failure in methylation inheritance or stochastic DNA methylation events [40]. Moreover, it is not clear whether this heterogeneity derives from a gradual accumulation of changes in DNA methylation during semiconservative replication of the DNA methylation pattern, or whether there is a continual flux of DNA methylation patterns within the mitotic progeny of a cell. The DNA

methylation level may also vary for some gene loci during aging [41-43] and tumor progression [44].

1.3 Stochastic and deterministic methylation

Genome-wide methylation analysis suggests that different mechanisms may explain the loss or gain of methylation underlying epigenetic diversity in cells: deterministic and stochastic processes [45]. Highly variable methylation patterns reflect stochastic fluctuations in DNA methylation, whereas well-structured methylation patterns imply deterministic methylation events.

Deterministic events may contribute to the gain or loss of methylation at specific loci. Deterministic epigenetic heterogeneity arises during differentiation of cells, in which progressive and predictive changes are accumulated. This kind of events leads to cell type-specific differences, which are readily recapitulated across individuals, and in general to tissue-specific differentiation hierarchies. According to the deterministic model for DNA methylation, locus-specific targeting of DNMT enzymes induces and maintains methylation. The choice of the target is not random, but determined by specific affinities of transcription factors. Eventually, the preference of DNMT1 for hemi-methylated DNA stabilizes and propagates the methylation profiles. Thus, the deterministic DNA methylation patterning must be the consequence of tissue- specific transcription factors and also epigenetic factors that possess stage-specific expression patterns [46]. Allelic-specific methylation, typical of imprinted and X-linked genes, constitutes an example of deterministic methylation changes [47-48]. This deterministic model fails to explain the extreme polymorphism of methylated alleles found with deeper sequence coverage of the genome [39].

Stochastic processes may produce the high rate of methylation polymorphism. Stochastic gain and loss of DNA methylation continuously generate and destroy epialleles (metastable epialleles), leading to increased epipolymorphism. Stochastic variability leads to cell-to-cell variability within differentiation hierarchies such that even individual cells of the same differentiation stage display marked heterogeneity (epigenetic mosaicism). The result of this stochastic variability is that cell-to-cell epigenetic variability cannot be explained by cellular differentiation states, gene expression patterns, or any other characteristics usually related to the deterministic patterns outlined above. It remains unclear whether the

stochastic cell-to-cell variation in DNA methylation patterns arises by errors introduced during DNA replication due to poor fidelity of the DNA-methylating enzymes, or whether it represents a mechanism for introducing an epigenetic buffer state into a cell population. As a consequence, initially organized epigenomic structure may be perturbed over time through a locally stochastic process of methylation aberration. Tumors tend to exhibit an increased genome-wide disorder in DNA methylation patterns [49].

Genome- wide methylation studies have identified numerous regions that on average are differentially methylated between tissues but have provided little evidence of the dynamics of the process generating these differences [14,50,51]. Whether this process changes one methylation pattern to another in a scheduled and regulated fashion integral to the developmental program (deterministic factors) or is instead a noisy and stochastic process occurring independently and in parallel at different sites remains debated.

Nevertheless, although only a low number of molecules/locus (10-40) is usually analysed both in gene-specific and genome-wide studies, almost all previous studies based on bisulfite sequencing revealed that different combinations of ^{m}CpG are indeed present at given loci. These are usually interpreted as the effect of stochastic methylation. However, with a such low number of molecules/locus is difficult to safely say if the origin of the epipolymorphism is a stochastic event or a genetically and/or environmentally driven phenomenon, developmentally regulated, leading to an orchestrated distribution of epialleles among the entire population of cells of a tissue. Moreover, it should be taken in consideration that each of the detected profile corresponds to the configuration of a single allele in a single cell of the complex cell mixture present in an analysed tissue.

Elucidating the stochastic and deterministic elements of methylation is important because of the involvement of epigenetic effects in determining the phenotypic variation among individuals and human disease [52- 56], as well as intra-individual changes over time [57-60].

1.4 Methods for DNA methylation analysis

Various techniques can be used to examine and quantify methylation according to the pre-treatment that the DNA receives and the resolution of the detection (specific-region or genome- wide methylation).

1) Methylation sensitive restriction enzyme digestion. Enzyme restriction digestion treatment is based on the ability of some methylation-sensitive restriction endonucleases (*HpaII*, and *SmaI*) and their corresponding isoenzymes, in particular the isoschizomer *MspI* for *HpaII* and the neoschizomer *XmaI* for *SmaI* (which are not inhibited by CpG methylation) to distinguish methylated from unmethylated cytosines in the CpG sequence context [61]. An important limitation is that all restriction enzyme-based techniques provide methylation data only at the restriction enzyme recognition sites.

2) Affinity enrichment. The affinity enrichment of methylated regions employs antibodies that are specific for methylated cytosine to immunoprecipitate denatured genomic DNA (MeDIP [62-63]) or methyl-CpG binding domain (MBD)- harbouring proteins (MeCP2 [64] or recombinant MBD2 [65-67] or the MBD2b/MBD3L1 complex [68-70]) with affinity for methylated native genomic DNA. An important point regarding affinity-based techniques is that they measure the density of methylation in a specific region. Thus, the genome coverage is limited by the distribution of the potential affinity targets in the genome, e.g. the density of methylated cytosines or CpG sites, which are unevenly distributed in the genome. Therefore, a methylated stretch of DNA where methylation CpG target sites are sparse might be difficult to differentiate from an unmethylated region. Specifically, in mammalian genomes, CG density is generally low and CG-dense sequences are typically unmethylated [62]. The exact methylation state of individual CpG sites cannot be determined using this approach.

3) Sodium Bisulfite conversion of DNA. The method is based on the selective deamination of cytosine but not 5-methylcytosine by treatment with sodium bisulfite [71-72]. Briefly, in the presence of sodium bisulfite, all the unmethylated cytosines are chemically converted to uracil, which is amplified as thymine during PCR. In contrast, the methylated cytosines are not converted, such that in the final sequencing the 5-methylcytosine will be still detected as cytosine. The bisulfite conversion efficiency is critical for the accuracy and the reliability of the results, especially for non-CpG methylation analysis. A poor bisulfite conversion or inappropriate conversion of methylcytosine to thymine will result in an overestimation of the methylation level and will subsequently influence the accuracy of the calculated DNA methylation by increasing the background [73-74].
Bisulfite conversion is the most conventional approach for pre-treatment and has been regarded as the gold standard for determining DNA methylation status because it provides

the most reliable and detailed information on the methylation pattern at single CpG sites resolution [75]. Moreover, for a genome- wide DNA methylation analysis, bisulfite conversion can be apply on the whole genomic DNA and is not limited by the presence of certain restriction enzymes recognition sites or the high CpG density.

The combination of pre-treatment methods with subsequent molecular biology techniques (PCR, mass spectrometry, pyrosequencing, DNA microarrays and high-throughput sequencing) have been developed and applied for DNA methylation analysis on specific loci or genome-wide scale. Indeed, these methods have progressed from small- scale candidate gene analysis to the ability to construct whole- genome methylation profiles.
As regards genome- wide technique, although methylation microarrays are a powerful tool for epigenetic studies, they have an important limitation in that they analyse only a small part of the CpG sites of the genome. The fast development of Next Generation Sequencing (NGS) methods, which can generate millions of reads each corresponding to the sequence of a single DNA molecule in one run without subcloning, has brought new opportunities to the wide usage of the bisulfite sequencing method for genome-wide DNA methylation analysis. Sodium bisulfite conversion followed by massively parallel sequencing (Bisulfite-seq) has become an increasingly used method for performing epigenetic profiles in the human genome [76]. Bisulfite-seq is well suited to the investigation of epigenetic profile from clinical tissue samples [77-78], and can be applied to very small quantities of DNA [79] including formalin-fixed samples [80]. Comprehensive mapping of DNA methylation in relevant clinical cohorts is likely to identify new disease genes and potential drug targets, helps to establish the relevance of epigenetic alterations in disease and provides a rich source of potential biomarkers [81].

1.5 Limitations of genome- wide techniques and quantitative approach

DNA methylation has been extensively analysed from a genomic point of view both in normal and in pathogenic tissues [82-84], but the dynamics that form, maintain and reprogram differentially methylated regions are still not well- defined. Also, little is known about genome-wide variation of DNA methylation patterns.

Genome-wide sequencing of bisulfite-DNA is unbiased relative to the sequence representation, but limited in the coverage/single locus. This limitation is essentially due to the fact that genome- wide methylomes cannot detect epialleles below a certain frequency; approximately variants below 10% escape detection. The reasons are the following. First, the coverage attained is very limited when individual loci are considered and the detection of changes of methylation states, occurring in a fraction of cells, requires sequencing of the locus several hundred times, not attainable with the current protocols. Second, the molecules sequenced represent a statistical collection of methylated cytosines deriving from physically different molecules. Thus, epipolymorphisms linking *in cis* several CpGs in the same DNA molecule cannot be deduced from these sequences. As example, 2 CpGs methylated in 50% of the molecules may derive from 2 molecules, one with a ^{m}CpG and another carrying the other ^{m}CpG; or from 2 molecules, 1 methylated at both sites and the other unmethylated.

As a consequence, the vast majority of studies on DNA methylation, regardless of the techniques employed, used a quantitative approach, namely they took in consideration the average methylation level, summarizing data into average percentage of methylated CpGs in specific genomic regions, or the methylation percentage for single CpG site, or looking at CpGs genome-wide distribution with an only relatively high resolution [85-91]. Such kind of approach is useful when methylation status is uniform in the population of cells under study, but fails to dissect and recognise different DNA methylation patterns when a heterogeneous population is investigated. Indeed, percentage methylation described in most DNA methylation studies hides important pattern and positional information of DNA methylation with potential functional and regulatory relevance [92]. Such information is lost when the average methylation is considered (quantitative approach).

1.6 Aims

In this work, in order to better decode epigenetic data, a new way to analyse DNA methylation, based on qualitative approach, was developed. The qualitative approach allows methylation profiles of cell populations to be studied at the single molecule level, thus providing an added value to the quantitative one. Considering that a single molecule corresponds to the configuration of a single epiallele in a single cell of the complex cell mixture present in an analysed tissue, such qualitative approach could be useful to

recognise different methylation profiles inside an heterogeneous cell population (i.e., tissues).

In order to obtain a detailed overview of a biological sample with an huge repertoire of epialleles and to carry out the analyses with sufficient power, it needs of a large number of sequences for the same genomic region. This can be accomplished through Deep Bisulfite Amplicon Sequencing (Deep- Bis), which allows to obtain a very high coverage (about 200.000-300.000 reads/sample) of selected loci, leading to observations that are not achievable with the low coverage of the genome- wide techniques.

As the number of sequences increases, the ability to analyse this type of data becomes a significant challenge. In the present study, it has been developed AmpliMethProfiler, a python-based pipeline for the extraction at the single molecule level of CpG methylation profiles from Deep- Bis of multiple DNA regions. The output reports the methylation status of each CpG site in a read in binary code (0 if the site is unmethylated, 1 if the site is methylated) and the epiallelic methylation patterns.

To describe epiallelic composition of a sample and to assess differences among samples, diversity measures borrowed from ecology and population genetics were used. From this point of view, each biological sample is not a single organism, but a micro- environment consisting of thousands of species, represented by the specific epialleles.

The qualitative approach is highly versatile and can be easily adaptable to different contexts and biological systems. In this work, it is applied on two experimental models: mouse development and AML progression before and after the demethylating therapy, in order to describe the methylation and demethylation dynamics, to investigate the epialleles distribution, to follow their evolution and to gain insight on epigenetic heterogeneity degree at specific loci.

Chapter 2

Qualitative DNA Methylation Analysis

2.1 A new qualitative approach to analyse DNA methylation data – Single molecules methylation analysis

DNA methylation patterns within a population represent outcomes of markedly different epigenetic mechanisms but may result in identical average methylation profiles.

In order to better decode epigenetic data, a new way to analyse DNA methylation, based on qualitative approach, has been developed. The qualitative approach, specifically looking at the individual methylation conformation of single molecules, could provide an added value to the quantitative one. Considering that a single molecule corresponds to the configuration of a single epiallele in a single cell of the complex cell mixture present in an analysed tissue, such qualitative approach could be useful to recognise different methylation profiles inside an heterogeneous cell population (i.e., tissues). Furthermore, this approach could help to better understand the mechanisms underlying the changes of methylation state of these cells during methylation and demethylation processes and to evaluate the stochastic and /or deterministic components of these phenomena. Moreover, this approach could be useful to assess the variability of DNA methylation pattern that might be observed for a given genomic locus in a cell population.

Conceptual differences between quantitative and qualitative methylation analysis are shown in the Fig. 2.1.

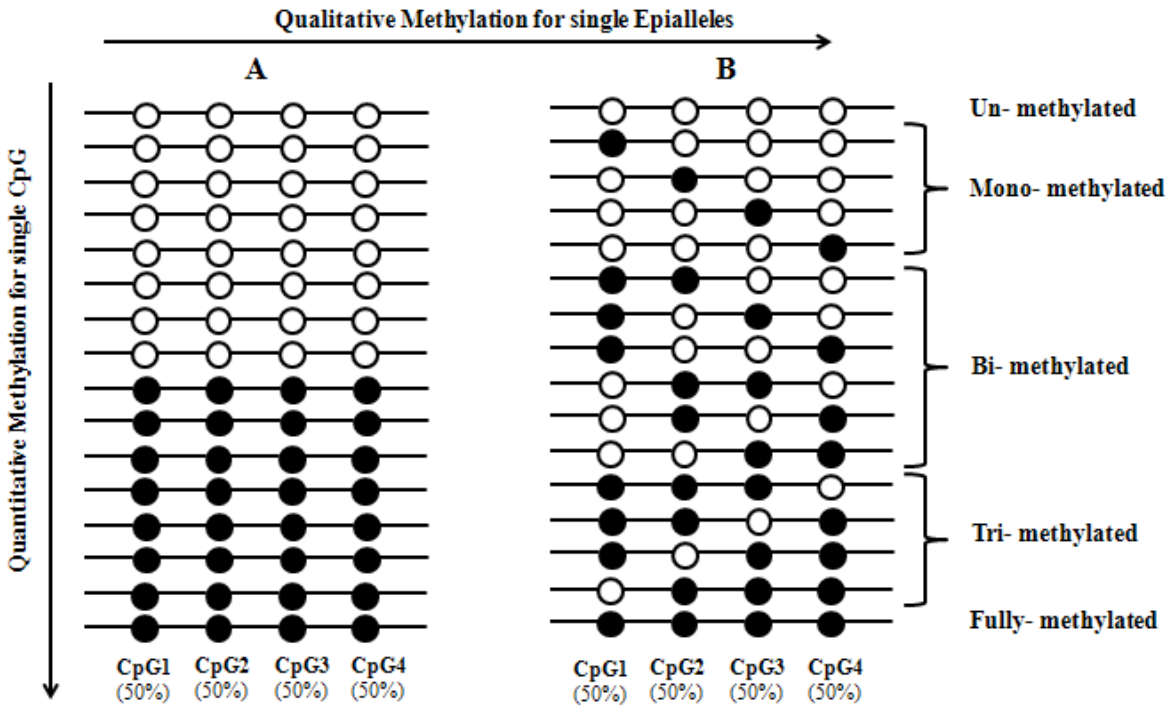

Figure 2.1. Scheme of the distributions of methylation profiles (epialleles) in two cell populations (A and B) and the corresponding methylation classes to which they belong. Empty and filled circles represent unmethylated and methylated CpGs, respectively. Each line represents a single molecule (epiallele). It should be noted that it is impossible to distinguish the DNA methylation scenarios of cell populations A and B by methodologies (e.g., pyrosequencing) that can quantify methylation at individual CpG sites.

Let's consider a locus composed by 4 CpG sites. By a quantitative point of view, both cell populations (A and B) have the same methylation value for each CpG site. However, from a qualitative point of view, DNA methylation scenarios in A and B are very different: in A two main epiallelic forms are present, while in B there is a great epiallelic diversity. In particular, as shown in B, these 4 CpGs can give arise to 16 possible methylation states or epialleles (2^{nCpGs}; 2^4) belonging to different methylation classes, including 1 unmethylated, 4 mono- methylated, 6 bi- methylated, 4 tri- methylated, 1 fully (tetra) – methylated. Thus, the methylation profiles inside the cell population B are very different and each one of the possible epialleles can have different frequency.

Qualitative approach allows to analyse DNA methylation at two different levels: i) methylation classes, defined as the number of methylated cytosine per molecule, independently of their position; 2) epialleles, defined as different combination of methylated CpG sites (^{m}CpGs) at a given locus.

This kind of approach nicely accounts for the high polymorphism of methylation profiles in cell populations derived from individual somatic tissues [39] and can provide new insights on the epipolymorphism degree inside an heterogeneous cell population and if its genesis has a stochastic or deterministic nature.

2.2 Deep Bisulfite Amplicon Sequencing (Deep- Bis)

Most current methylation data sets derive from tissue samples with heterogeneous cell population and each one with its own repertoire of epialleles for a given genomic locus.

An hurdle for most DNA methylation detection methodologies is the limitation when heterogeneous methylation patterns, rather than homogenous ones, are present. Heterogeneous DNA methylation patterns cannot be fully characterized without a method that: 1) provides a comprehensive overview of a sample with an huge repertoire of epialleles, 2) allows the direct visualization of individual epiallele and 3) guarantees an effective statistical representation. Only in such cases the entire population of heterogeneously methylated epialleles can be quantified. Otherwise, the information obtained therefore is compromised by stochastic effects.

The sensitivity in the recognised different methylation patterns increases as depth of coverage increases and the need to achieve high analytical sensitivity for heterogeneously methylated loci is met by massively parallel sequencing. During the past decade, DNA methylation analysis has undergone a major technological revolution. The recently developed Next Generation Sequencing (NGS) methods, in particular when coupled to bisulfite conversion techniques, allow to conduct DNA methylation analysis at single base resolution with high speed and high throughput. However, it is only using clonal sequencing approaches with allelic outputs that it is possible investigate heterogeneous DNA methylation and the extent of epiallelic methylation patterns that exist within a single sample. Importantly, the number of methylated alleles can be substantially underestimated unless clonal approaches are used [93].

The key feature of the qualitative approach is to perform an in-depth methylation analysis and to obtain a very high coverage (about 200.000-300.000 reads/sample) of selected loci by means of the Deep Bisulfite Amplicon Sequencing (Deep- Bis). This technique leads to observations that are not achievable with the low coverage of the genome- wide techniques. Indeed, it is based on the clonal amplification of single bisulfite- converted

DNA molecules (amplicons) and allows to generate multiple sequence reads (high sequencing depth) for each genomic locus/sample. Each sequence read represents a methylation pattern. Each CpG position of a single molecule provides a binary answer: methylated or unmethylated. This high sequencing depth allows to gain a true representation of different epialleles in a sample. This is important to not only derive very precise average methylation levels for each CpG sites, but to also infer cell population characteristics accurately, to determine the extent of heterogeneous DNA methylation and to track the epipolymorphisms. Moreover, it enables the sequencing of different DNA templates from multiple regions simultaneously, providing a true representation of the diversity and extent of heterogeneous DNA methylation patterns derived from a given sample. Also, it allows for detailed comparisons of methylation patterns from different biological samples.

In this way, observed changes in average methylation levels can then be interpreted according to epiallelic diversity, discerning, for example, a regulated increase in the frequency of a specific epiallele from multiple stochastic changes in the frequencies of many epialleles [94].

2.3 AmpliMethProfiler: a pipeline for determining CpG methylation profiles of amplicons from Deep- Bis

2.3.1 Introduction

Recent DNA sequencing technology, the so-called Next Generation Sequencing (NGS) technology, enables researchers to obtain many thousands of sequence reads in a single sequencing run and at a cost that is several orders of magnitude smaller than the previous generation DNA sequencing technologies. This high coverage allows, not only for quantitative, but also for qualitative locus-specific methylation analysis to be performed. Qualitative analysis allows methylation profiles (epialleles) of a given locus inside cell populations to be studied at the single molecule level. Analysis of epiallele diversity is an innovative approach that is now applied in epigenetics, in fields as diverse as carcinogenesis [39], developmental biology [95] and ecology [96]. However, as the number of sequences increases, the ability to analyse this type of data becomes a significant challenge. Moreover, currently available tools for methylation analysis lack output formats that explicitly report CpG methylation profiles at the single molecule level.
In the present study, it has been developed AmpliMethProfiler, a python-based pipeline for the extraction at the single molecule level of CpG methylation profiles of amplicons from Deep- Bis of multiple DNA regions. The output reports the methylation status of each CpG site in a read in binary code (0 if the site is unmethylated, 1 if the site is methylated) and summarises DNA methylation according to epiallelic methylation patterns and can be readily used for the downstream quantitative and qualitative analysis. This software has been used to analyse multiplex bisulfite amplicon PCR, coupled to massively parallel deep sequencing, from different tissue samples and different loci (see Chapters 3 and 4).

2.3.2 AmpliMethProfiler: a general overview

AmpliMethProfiler requires a local installation of BLASTn [97] and uses the Biopython >= 1.65 [98] python library (python version >= 2.7). This pipeline can be used on any python-enabled operating system. The software is designed to work with single end reads. Output from paired-end sequencing should be first converted to single end format, provided that the two paired ends overlaps.

As input data, requires three input files:

1. `readsFile`: a file containing the reads from the sequencer in FASTA format;
2. `primFile`: a comma-separated file containing information on the sequenced regions;
3. `refFile`: and a FASTA file containing the reference sequences of the analysed regions.

Fig. 2.2 depicts the flowchart of AmpliMethProfiler modules. The pipeline is composed of two python scripts and its execution involves two main steps:

1. `preprocessFasta.py` module, for the preprocessing the input FASTA file**,** in terms of filtering and demultiplexing of the sequencer output in single end FASTA format;
2. `methProfiles.py` module, for the generation of the CpG methylation profiles and the computation of summary statistics on methylation status and quality assessments (bisulfite efficiency).

A third python module (`methylUtils.py`) contains necessary functions to run the two scripts above.

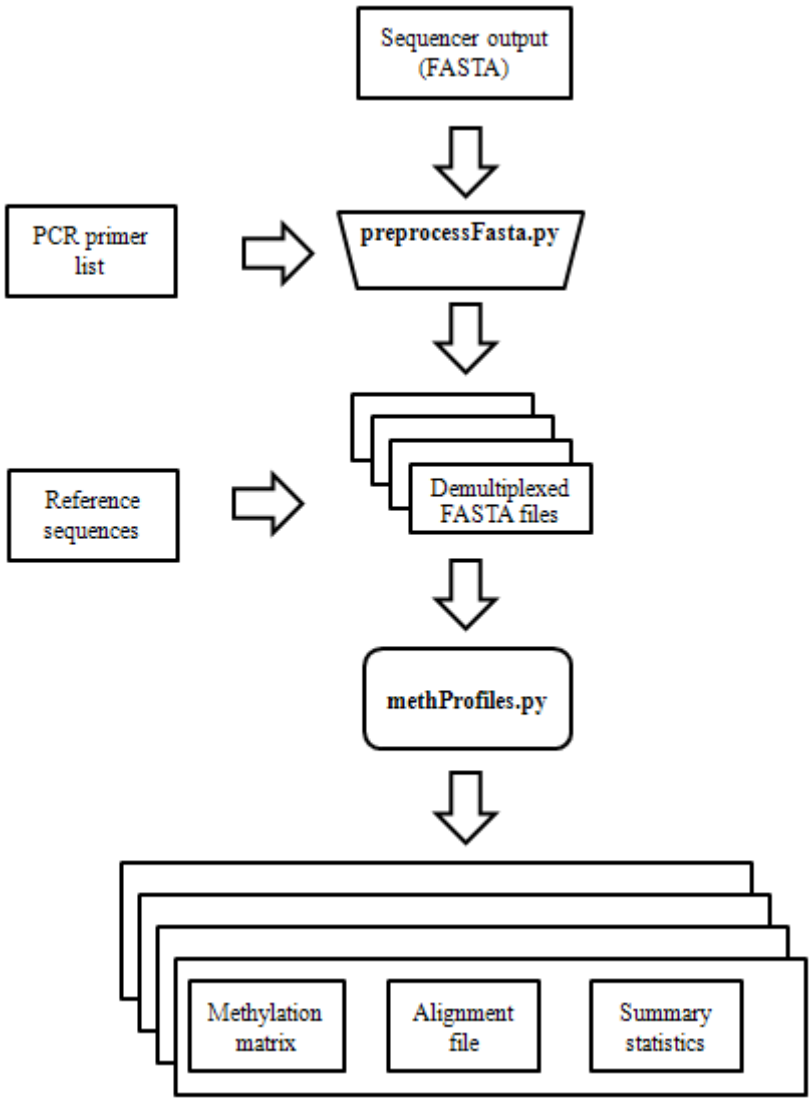

Figure 2.2. AmpliMethProfiler pipeline workflow.

2.3.2.1 Demultiplexing and filtering

The module `preprocessFasta.py` works on three types of inputs:

- Sequence files:
  - ✓ `readsFile`: A file containing reads from the sequencer in FASTA format. Reads are expected to be represented as single end reads.
  - ✓ `primFile`: A comma separated file containing info about the sequenced regions. For each analysed sequence the file should contain two rows (one for the 5' and another one for the 3' primer sequences) with the following

fields:

- a sequence ID
- the PCR primer sequence
- the length of the target sequence

- Filtering parameters:
  - ✓ `threshLen` (optional; 0.5 is the default value): A real between 0 and 1 indicating the minimum percent of length similarity between the sequence read and its reference sequence. Namely, if *refLen* is the length of the reference sequence and *readLen* is the length of the read, the module will retain all reads for which the following relation holds:

    $$refLen - (refLen)*threshLen <= readLen <= refLen + (refLen)*threshLen$$

  - ✓ `primTresh` (optional; 0.8 is the default value): A real between 0 and 1 indicating the minimum percentage of sequence similarity (1 indicates identical sequences, 0 express completely different sequences) between the 5' or 3' end of a read sequence and the corresponding PCR primer sequences contained in the file specified via `primFile`. If no matches between the 5' and the 3' end of the read are found the read will be discarded from the de-multiplexed output.

- Experiment ID:
  - ✓ `sample`: The sample ID on which the experiment is carried on.

Reads from Deep- Bis of multiple regions are demultiplexed by `preprocessFasta.py` by comparing their 5′ and 3′ ends with a list of provided PCR primers. The demultiplexing procedure is based on a user-provided percentage of similarity (setting the parameter `primTresh`) between the 5′ or 3′ end of a read sequence and the corresponding PCR primer sequences. It starts searching with threshold 1 (perfect sequence match) and decreases the threshold value until a match is found or the maximum number of allowed mismatches (specified in the parameter `primTresh`) is reached.

Reads are then filtered out if no match is found between at least one of the read ends or if, given a user-provided threshold, their length does not match the expected one (specified in the parameter `threshLen`).

As output, the `preprocessFasta.py` module returns, for each sequenced region, a demultiplexed, filtered FASTA file, containing the passing filter reads belonging to that region. For each read it will be produced a new header.

2.3.2.2 Extraction of methylation profiles

The next module `methProfiles.py` module runs on each of the demultiplexed, filtered FASTA files generated by `preprocessFasta.py` and computes CpG methylation profile matrices and several summary statistics. This module uses three types of inputs:

- Sequence files:
  - ✓ `refFile`: The FASTA file containing the reference sequences of the analysed sequences
  - ✓ `primFile`: A comma separated file containing info about the sequenced regions. For each analysed sequence the file should contain two rows (one for the 5' and another one for the 3') with the following fields:
    - sequence ID;
    - PCR primer sequence;
    - length of the target sequence.

- Filtering parameters:
  - ✓ `bisu_thresh` (optional; 0.98 is the default value): Bisulphite efficiency threshold.
  - ✓ `cgAmbThresh` (optional; 1 is the default value): Maximum percentage of ambiguously aligned CpG sites.
  - ✓ `alignPropThresh` (optional; 0.6 is the default value): Minimum proportion of not clipped bases on the aligned read.
  - ✓ BLASTN aligner parameters.
    - `blastExecPath`: Path to the local BLASTN binary directory.

- `nThreadBlast (optional)`: Number of thread for BLASTN execution.
- `Reward (optional)`: BLASTN reward for a nucleotide match.
- `penalty (optional)`: BLASTN penalty for a nucleotide mismatch.
- `gapopen (optional)`: BLASTN cost to open a gap.
- `gapextend (optional)`: BLASTN cost to extend a gap.
- `word_size (optional)`: BLASTN word size for initial match.
- `window_size (optional)`: BLASTN multiple hits window size.

➢ Experiment ID:

✓ `sample`: The sample ID on which the experiment is carried on.

The `methProfiles.py` module works in the following way.
First, the amplicons from Deep- Bis are aligned to the corresponding bisulfite-converted reference sequence using the locally installed BLASTN program. Then, `methProfiles.py` inspects the C and CpG aligned positions for each input read.

Bisulfite efficiency for each aligned read is computed as the percentage of conversion of non-CpG cytosine (C) residues to thymine (T) residues over the total number of non-CpG C residues. A cytosine not CpG is expected to be always converted in T after bisulphite treatment. If the percentage of non-CpG deaminated C residues (blue Ts in the read of the example below) over the total number of non-CpG C residues (blue Cs in the reference sequence of the example below) is below the given threshold (setting the parameter `bisu_thresh`), the read is discarded. The methylation status (0 for unmethylated e 1 for methylated) of each CpG site inside of each aligned read is determined by evaluating the eventual deamination of CpG sites as a result of the bisulfite treatment.

```
Ref:       TGCGCGGAACTCTGATTCTGGTAATCCGTGTATTAGAGTGTCTATTC
Bisu_Ref:  TGCGCGGAATTTTGATTTTGGTAATTCGTGTATTAGAGTGTTTATTT
Read:      TGTGTGGAATTCTGATTTTGGTAATCTGTGTATTAGAGTGTTTATTT
```

For each CpG position in the aligned reference sequence (green Cs in the bisulfite-converted reference sequence in the example below), the corresponding position in the aligned read sequence is inspected. A CpG cytosine is expected to be C (methylated) or T (unmethylated) in each read. Because of several causes (e.g. misaligned reads, deletions, polymorphisms..) other bases (G or A) or gaps may be present in such sites: in this case, the methylation state of the CpG site is reported as uncertain (marked in red in the example below). If the percentage of such ambiguous sites (red bases in the read of the example below) over the total number of CpG sites in the region exceed the given threshold (setting the parameter `cgAmbThresh`) the read is discarded.

```
Bisu_Ref: TGCGACGGAATTTTGATTTTGGTAATTCGTGTATTAGAGCGTTTATT
Read:     TGTG-GGAATTCTGATTTTGGTAATCAGTGTATTAGAGCGTTTATT
```

Methylation percentages for each CpG site are then computed as the number of non-deaminated bases (methylated CpG sites) mapped on that site over the total number of non-ambiguous CpG positions. The same method is used to compute bisulfite efficiency for all C (non-CpG) sites.

Moreover, it may be the case that some reads will not be completely aligned. In such cases, if the proportion of aligned bases (blue bases in the read of the example below) is under the given threshold (setting the parameter `alignPropThresh`) the read is discarded.

```
Bisu_Ref: TGCGACGGAATTTTGATTTTGGTAATTCGTGTATTAGAGCGTTTATT
Read:     --------AATTCTGATTTTGGTAATCAGTGTATTAGAGCGT-----
```

2.3.3 Output files

For each sequenced region, `methProfile.py` returns the following files (in the following list *idReg* refers to the provided id for the sequenced region and *sample* refers to the provided experiment ID):

- idReg.bisu: The bisulphite converted reference sequence in FASTA format.
- Summary and quality statistics file (idReg_sample.out.stats): this file contains summary and quality statistics for the analyzed region (information about the

number of passing filter reads, the methylation percentage of each C in CpG sites, and the bisulfite efficiency for each C in non-CpG sites). In particular, the file is structured as follows:

- ✓ number of analyzed reads;
- ✓ number of passing filter reads;
- ✓ position of CpG sites in the analyzed sequence;
- ✓ methylation percentage of each C in CpG sites;
- ✓ number of CpG contributing to the computed methylation assessment for each site;
- ✓ position of C (not CpG) in the analyzed sequence;
- ✓ bisulfite efficiency for each C in non-CpG sites (percentage of deaminated C for each C site);
- ✓ number of C contributing to the computed deamination assessment for each C site.

```
114987: Analyzed reads
112283: Passing filter reads

Meth percent
Position                105         138         150         226         293         343
%Meth                   0.197       0.232       0.117       0.207       0.166       0.112
# of valid Cs           108934      111512      111609      111454      106958      110158

Bisulfite efficiency on non-CpG sites
Position                99          111         113         116         120         123     ...
%Efficiency             0.99        0.99        0.99        1           0.99        0.99    ...
# of valid Cs           112072      111799      110839      112242      112250      106601  ...
```

Figure 2.3. Summary and quality statistics file

- Alignment file (idReg_sample.out.align): this file contains BLAST-aligned sequences in clear text format. Each entry of the txt format file reports a passing filter aligned read represented by five rows, structured as follows:

  (1) read ID, read length, experiment ID, region ID;

  (2) bisulfite efficiency, calculated as the percentage of deaminated Cs (non-CpG) over all Cs (non-CpG);

  (3-5) alignment of the read sequence against the bisulfite-converted reference sequence (the third line is the reference sequence; the fourth are alignment bars, the fifth is the read sequence).

```
>M02436:14:000000000-AE7UP:1:1101:8171:3243 length=405 sample=64 targSeqPrimer=DDO1
bisulphite efficiency: 0.9864
TGAATAATGCGGTAGGGGAAACGAGTGTTTGGTTTTAGTTGGGATTTTTTGGAAGTTAGATAGATTATTTATTTATTTGTTATGT...
||||||||| |||||||||| |||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||...
TGAATAATGTGGTAGGGGAAATGAGTGTCTGGTTTTAGTTGGGATTTTTTGGAAGTTAGATAGATTATTTATTTATTTGTTATGT...

>M02436:14:000000000-AE7UP:1:1101:6105:3186 length=405 sample=64 targSeqPrimer=DDO1
bisulphite efficiency: 1.0
TGAATAATGCGGTAGGGGAAACGAGTGTTTGGTTTTAGTTGGGATTTTTTGGAAGTTAGATAGATTATTTATTTATTTGTTATGT...
||||||||| |||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||...
TGAATAATGTGGTAGGGGAAATGAGTGTTTGGTTTTAGTTGGGATTTTTTGGAAGTTAGATAGATTATTTATTTATTTGTTATGT...
```

Figure 2.4. Example of the alignment file

- Methylation profiles file (idReg_sample.out): this file contains the CpG methylation profile matrix in which columns represent CpG sites and rows represent single molecules (reads). The methylation status of each CpG site in a read is coded as 0 if the site is recognized as unmethylated, 1 if the site is recognized as methylated, and 2 if the methylation state could not be assessed (i.e. other bases other than C or T or alignments gaps are found). Row entries are reported in the same order as in the "Alignment file" (idReg_sample.out.align), and column order represents the CpG positions reported in the "Summary and quality statistics file" (idReg_sample.out.stats). Each row of this file can be considered as the CpG methylation profile of a single read and defines an epiallele in subsequent analyses.

```
0 0 0 0 0 0
0 0 0 0 0 0
1 0 1 1 1 0
0 0 0 0 1 0
0 0 0 0 0 0
2 1 0 0 0 0
0 1 0 0 0 0
0 0 1 0 0 0
0 0 0 0 0 0
1 1 1 1 1 1
```

Figure 2.5. Methylation profiles file

2.3.4 Discussion

AmpliMethProfiler is a python- based pipeline composed by two python scripts for the generation of CpG methylation profiles at single molecule level from Deep- Bis on multiple genomic regions.

AmpliMethProfiler can de-multiplex and filter the experiment output by using provided PCR primer sequences and corresponding reference sequences. It can filter out reads based on the expected length or the impossibility to recognize their reference sequence. Bisulfite specific filters can be applied to filter out reads where the bisulfite treatment failed (or, however, had not the desired effects) or reads where the methylation status of a high number of CpG sites could not be ascertained. AmpliMethProfiler uses BLASTN to align in a fast and reliable way input reads to a bisulfite converted reference sequence, and reports the methylation status at each CpG locus in each input read. Classic quantitative methylation assessment, namely the methylation percentage, is reported for all input sequences at each CpG site in each analyzed region.

2.4 Data analysis

Qualitative methylation analysis has been approached by applying tools borrowed from ecology and population genetic. The populations of molecules (epialleles) produced by NGS technology have been handle as populations of haploid organisms and biological samples as the micro- environments in which these organisms are harbored. By this way, it is possible to describe the general methylation landscape of the various samples. Analyses include: i) estimation of methylation classes and epialleles frequencies; ii) estimation of diversity indexes inside each sample; iii) estimation of epigenetic distance between pair of samples. Analysis have been performed using R statistics environment (http://www.R-project.org).

2.4.1 Methylation classes and epiallelic frequencies

For each locus, molecules were subdivided in classes, according the number of methylated CpG sites per molecule, independently of their position, defining in this way different methylation classes (un-, mono-, bi-, tri-, ….. , fully- methylated).

Different combinations of methylated CpG sites constitute the epialleles.

For each sample, the frequency of each methylation class was calculated as a ratio of the number of reads, belonging to each methylation class, over the total number of reads found in each sample.

For each sample, the frequency of each epiallele was calculated as a ratio of the number of reads of a specific epiallele over the total number of reads found in each sample.

2.4.2 Measuring epigenetic diversity of samples

In order to quantify the epiallelic diversity inside each sample, diversity measures have adapted from ecology [99]. When applying an ecological perspective, each sample is not a single organism, but a micro- environment consisting of thousands of species that are represented by the epialleles. Each epiallele represents a single allele of a cell. Thus, each biological sample can be seen as a community, each epiallele can be seen as a specie, while each read can be seen as an individual.

Species diversity (or epiallelic diversity) has two separate components: 1) the number of species (epialleles) present, called as species richness, and 2) their relative abundances,

termed dominance or evenness. As a result, many different measures (or indices) of biodiversity have been developed. A simple measure is the number of epialleles in the sample, expressed as species richness. Ecological measures of diversity typically integrate both number and abundance of epialleles: one of these measures is the Shannon diversity index (H) [100], which has been used for the analysis.

2.4.2.1 Species richness (S)

This is the simplest of all the measures of species diversity. In ecological field, it is the count of the total number of species found in a community. In the epigenetic field, it is the number of epialleles found in a sample. There are two problems associated with this measure. First, the number of species in a sample depend on the size of the sample: larger samples have more species. This is because not all species have the same probability of being in a sample because some of the species are common (high probability) and some are rare (low probability). As a consequence, small samples have common species and few rare species while larger samples pick up more rare species. Second, this measure does not take into account the proportion and distribution of each species within the community. Thus, this measure does not indicate how the diversity of the population is distributed or organized among species.

2.4.2.2 Shannon diversity index (H)

A diversity index is a mathematical measure of species diversity in a given community. Diversity indices provide more information about community composition than simply species richness. Indeed they take into account not only the species richness, but also the relative abundances (dominance or evenness) of different species that are present in the community. The ability to quantify diversity in this way allows to understand community structure. Let's consider two communities, each one composed of 100 individuals and 10 different species. One community has 10 individuals of each species; the other one has one individual of each of nine species, and 91 individuals of the tenth species. The more diverse community is first one is, but both communities have the same species richness. By taking relative abundances into account, a diversity index depends not only on species

richness but also on the evenness, or equitability, with which individuals are distributed among the different species.

The Shannon diversity index (*H*) is a diversity index that is commonly used to characterize species diversity in a community. It is also known as Shannon- Wiener index or Shannon entropy. The Shannon entropy quantifies the uncertainty in predicting the species identity of an individual that is taken at random from the dataset. The Shannon index assumes all species are represented in a community and that they are randomly sampled from an infinitely large population.

The Shannon diversity index (*H*) is defined as:

$$H = -\sum_{i=1}^{S} p_i \ln p_i$$

where:

S=total number of species (epialleles) in the community (richness)

p_i= proportion of individuals (reads) belonging to the *i*-th species in the community (or the proportion of epiallele *i* relative to the total N epialleles). It is estimates as $p_i = n_i/N$, where n_i is the number of individuals in species i and N is the total number of individuals in the community.

The negative out front is needed to offset the obtained negative when the natural log of p_i (since p_i is a fraction, its log is negative) is taken. Since by definition the p_{is} will all be between zero and one, the natural log makes all of the terms of the summation negative, which is why the inverse of the sum is taken.

Shannon index takes into account species richness and proportion (evenness) of each species within a community. Evenness expresses how evenly the individuals in the community are distributed over the different species. In other words, evenness is a measure of how much similar is the abundance of different species. So the H value allows us to know not only the number of species but how the abundance of the species is distributed among all the species in the community.

The Shannon index increases as both the richness and the evenness of the community increase. Its value spans from 0 to, theoretically, + ∞ (H_{max}=logS).

- High values of H indicate a diverse and equally (evenly) distributed species. This means that the maximum H value is obtained when all the species have the same frequency: in this case, the community is characterized by a high diversity.
- Low values of H represent less diverse community. In this case, the relative abundance of the various species are very dissimilar, so that there are some common (dominant species) and some rare species. If all abundance is concentrated to one specie, and the other species are very rare (even if there are many of them), Shannon entropy approaches zero.
- A value of Shannon entropy exactly equals 0 would represent a community with just one species: there is no uncertainty in predicting the type of the next randomly chosen entity.

2.4.3 Epigenetic distance

The degree of epigenetic diversity across individuals was evaluated using the concept of epigenetic distance. [101]. Each of the epialleles is binary coded, with 0 for an unmethylated cytosine and 1 for a methylated cytosine. Each epiallele is, therefore, represented by a row vector of *n* 0 and 1, where *n* is the number of cytosines in the tested region.

The degree of epigenetic dissimilarity was measured by Euclidean distance, by use of the following equation:

$$d_{12} = \sqrt{\sum_{i=1}^{n} (m_{1i} - m_{2i})^2}$$

where:

- m_{1i} is the average methylation of sample 1 at site *i*;
- m_{2i} is the average methylation of sample 2 at site *i;*
- d_{12} is the Euclidean DNA methylation distance between samples 1 and 2.

The larger the distance, the more dissimilar the two samples' methylation profiles are to each other.

With this metric, the distances between all possible pairs of samples for each promoter locus of p14 and p15 has been calculated.

2.4.4 Principal Coordinate Analysis (PCoA)

When a dataset is composed by a multidimensional data, it can be hard to find patterns between samples. To better understand the structure of large data sets with a large number of variables, it is useful to use the multivariate techniques. Multivariate techniques are very useful to summarize many variables into a smaller number of variables (i.e., reduce the number of dimensions) to simplify the analysis of a dataset and to obtain the full picture from a large amount of data.

Different methods are used to identify patterns and to highlight similarities and differences between samples or groups. One of these is the Principal Coordinate Analysis (PCoA), also known as metric multidimensional scaling (MDS), a tool for multivariate analysis. PCoA is a commonly used method to compare groups of samples based on phylogenetic distance metrics and to explore and to visualize similarities or dissimilarities of data.

The input of PCoA is a distance matrix (a similarity or dissimilarity matrix) between a set of individuals and assigns for each item a location in a low-dimensional space, e.g. as a 2D or 3D graphics. The aim of the PCoA is to map the samples present in the distance matrix to a new set of orthogonal axes (low-dimensional graphical plot) that explain the maximum amount of variance. Samples are mapped on this low-dimensional graphical plot in such a way that distances between points (samples) in the plot are close to original dissimilarities. In other words, it tries to find an arrangement of samples such that that the distances between the samples in the graph match as closely as possible those in the distance matrix.

PCoA is a scaling or ordination technique, which computes a linear transformation of the variables in order to reduce multidimensional datasets into a lower dimensional space, retaining the maximal amount of information. In other words, the algorithm attempts to explain most of the variance in the original data set. This technique helps to extract and visualize a few highly informative components of variation from complex, multidimensional data.

To summarise the variability in the data set, PCoA produces a set of uncorrelated (orthogonal) axes, called as principal coordinate (PC) axes (columns), for each sample (rows). Each axis has an eigenvalue whose value indicates how much variance is shown on (captured in) that axis, or, in other words, the amount of variation explained for each PC. The proportion of a given eigenvalue to the sum of all eigenvalues reveals the relative 'importance' of each axis. Eigenvalues are provided in a sequence of largest to the smallest value. This means that most variance will be shown by making an ordination graph with

the first two axes. The first axis accounts for as much of the variability in the data as possible and each succeeding axes accounts for as much of the remaining variability as possible. A successful PCoA will generate a few (2-3) axes with relatively large eigenvalues, capturing above 50% of the variation in the input data, with all other axes having small eigenvalues. Objects (samples) are represented as points in the ordination space. Each object (sample) has a score along each axis. The object scores provide the object coordinates in the ordination plot. Objects (samples) ordinated closer to one another are more similar than those ordinated further away.

The principal coordinates can be plotted in two or three dimensions to provide an intuitive visualization of the data structure to look at differences between the samples, and look for similarities by sample category (i.e., affected or healths categories, before and after a treatment samples categories). In this way, it is intuitive to identify if samples from a category (i.e. affected) cluster together, compared to samples belonging to another category (i.e. healthy).

Chapter 3

Tracking the evolution of DDO gene methylation profiles during mouse development

3.1 Introduction

DNA methylation landscape is dynamically remodelled during the mammalian life cycle through distinct phases of reprogramming and *de novo* methylation [16]. This remodelling occurs through an establishment of a globally demethylated state during early embryogenesis. Then, at late stages of embryonic development and early post- natally, a lineage-specific methylome that maintains cellular identity is shaped [34,102-104]. Once such reconfiguration comes to end, it may be maintained throughout the life. Such phenomenon, crucial to complete embryonic development, is the result of a dynamic interplay between DNA methylation and demethylation events [105-107] assisted by different DNA methyltransferases (DNMT1, DNMT3a and DNMT3b), which may convert cytosine to methylcytosine [22], or by Ten- eleven translocation (TET) enzymes which oxidize 5mC into 5-hydroxymethylcytosine, 5- formylcytosine and 5- carboxylcytosine, thus promoting 5mC demethylation [26]. DNA methylation plays a critical role in the cell differentiation and embryonic development and its alteration can lead to loss of cell identity, cellular transformation or disease, and is generally incompatible with normal development [20,21].

The epigenomic landscape reflects the cellular and tissue diversity and changes in DNA methylation are a major feature of development. Genome- wide DNA methylation studies

have shown that there are numerous DNA methylation differences between adult tissues (T- DMRs) and within the same tissue at different developmental stages (DS- DMRs) [107-109]. These differences are due to the distinct cell types forming a tissue and to changes in the proportion of specific cell populations during development. Furthermore, because of tissues are made up of many cell types, each cell type acquires a very distinctive epigenomics signature. This means that, despite sharing similar global mCpG content, they gain a cell-type specific genomic DNA methylation landscape that gives an identity card for different cells and governs and stabilizes an elected gene expression program. In summary, DNA methylation profiles are unique to individual cells or tissue types [110-112]. However, if on one side this kind of study have identified numerous region that on average are differentially methylated between tissues, on the other side they have provided little evidence about the dynamics of the process generating these differences. This is because these studies take in consideration the genome- wide CpG sites distribution with an only relatively high resolution.

In this chapter, using the Deep- Bis it has been performed an in-depth methylation analysis (coverage 200.000-300.000 reads/sample) of DDO promoter region in different tissues at different developmental stages. This has allowed to analysing the methylation status at single molecule resolution with a high level of resolution. D-aspartate oxidase (DDO) is FAD-containing enzyme that selectively deaminates bicarboxylic D-amino acids, such as D-aspartate (D-Asp) and D-glutamate [113-115]. In fact, it has been shown here a peculiar feature of DDO gene, to undergo physiological methylation and demethylation processes in different tissues and in different stages of development. Molecules are grouped in methylation classes, defined as the number of methylated cytosine per molecule, independently of their position. This allowed to describing the methylation and demethylation dynamics. Then, it has been investigated the epialleles distribution in different mice and different tissues during the developmental, in order to test the hypothesis that the origin of epipolymorphism in a tissue is not a stochastic event, rather a genetically and/or environmentally driven phenomenon, developmentally regulated, leading to an orchestrated distribution of epialleles among the entire population of cells.

3.2 Results

3.2.1 Choice of biological system and quantitative methylation analysis

An ideal model to study the dynamics of CpG methylation and demethylation processes should include the following features: 1) being a physiological model involving tissues and not cell lines; 2) containing a genomic region containing a limited number of CpG sites; 3) these CpG sites should undergo opposite changes in methylation state in different tissues possibly reaching a similar methylation levels at the end of a dynamic process. Because it was recently found that DDO promoter region contains a limited CpG sites and that these undergo demethylation in brain during development [116] it has been checked here whether the same region could have an opposite behaviour in other tissues.

DNA methylation was assessed through a strategy based on the locus- specific amplification of bisulfite- treated genomic DNA. An in-depth analysis to investigate the methylation state of the DDO promoter region at different pre- and post- natal developmental stages (E15, P0, P7 for the brain; from P0, P15 and P60 for the lung; from P0, P15 and P90 for the gut) was performed. The region upstream the transcription start site (TSS), spanning nucleotides -444 to -88 and containing 6 CpG sites (positions -363, -330, -318, -242, -175, -125) was analysed (Fig. 3.1).

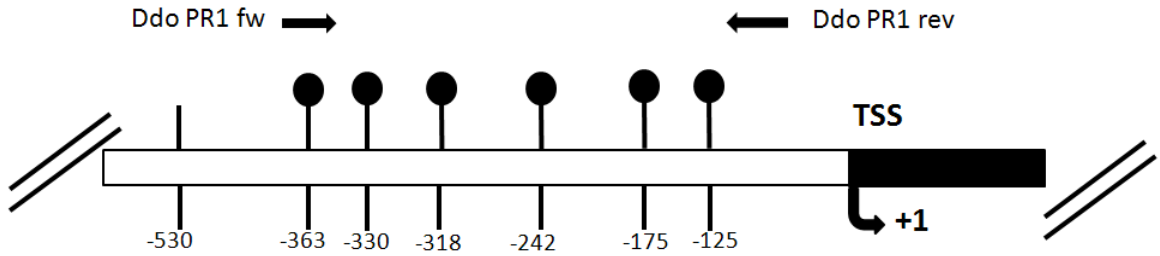

Figure 3.1: Structure of the putative mouse DDO gene promoter. The transcriptional start site (TSS, +1) is indicated by an arrow. The putative regulatory upstream region (white box), exon 1 (black) are indicated. Position of CpG sites is indicated as relative to TSS. Position of the primers used for bisulfite analysis is indicated by arrows at the top of the map (DDO PR1 fw and DDO PR1 rev).

By a quantitative analysis, the average methylation of the analysed region had an opposite behaviour in gut and lung compared to brain (Fig. 3.2). In particular, a significant increase of average methylation degree (one-way ANOVA; p-value=0.001) was observed in gut from P0 (about 20% methylation) to P90 (about 40%) and in lung from P0 (about 20%) to P60 (about 40%), while a significant decrease (one-way ANOVA; p-value=0.001) in brain from E15 (60%) to P7 (40%) was observed (Fig. 3.2). Note that at the final time points for each tissue a similar methylation degree (about 40%) was observed. These observations prompted to perform a bioinformatics methylation classes analysis (qualitative analysis) aimed to gain insights into the dynamics of methylation and demethylation processes.

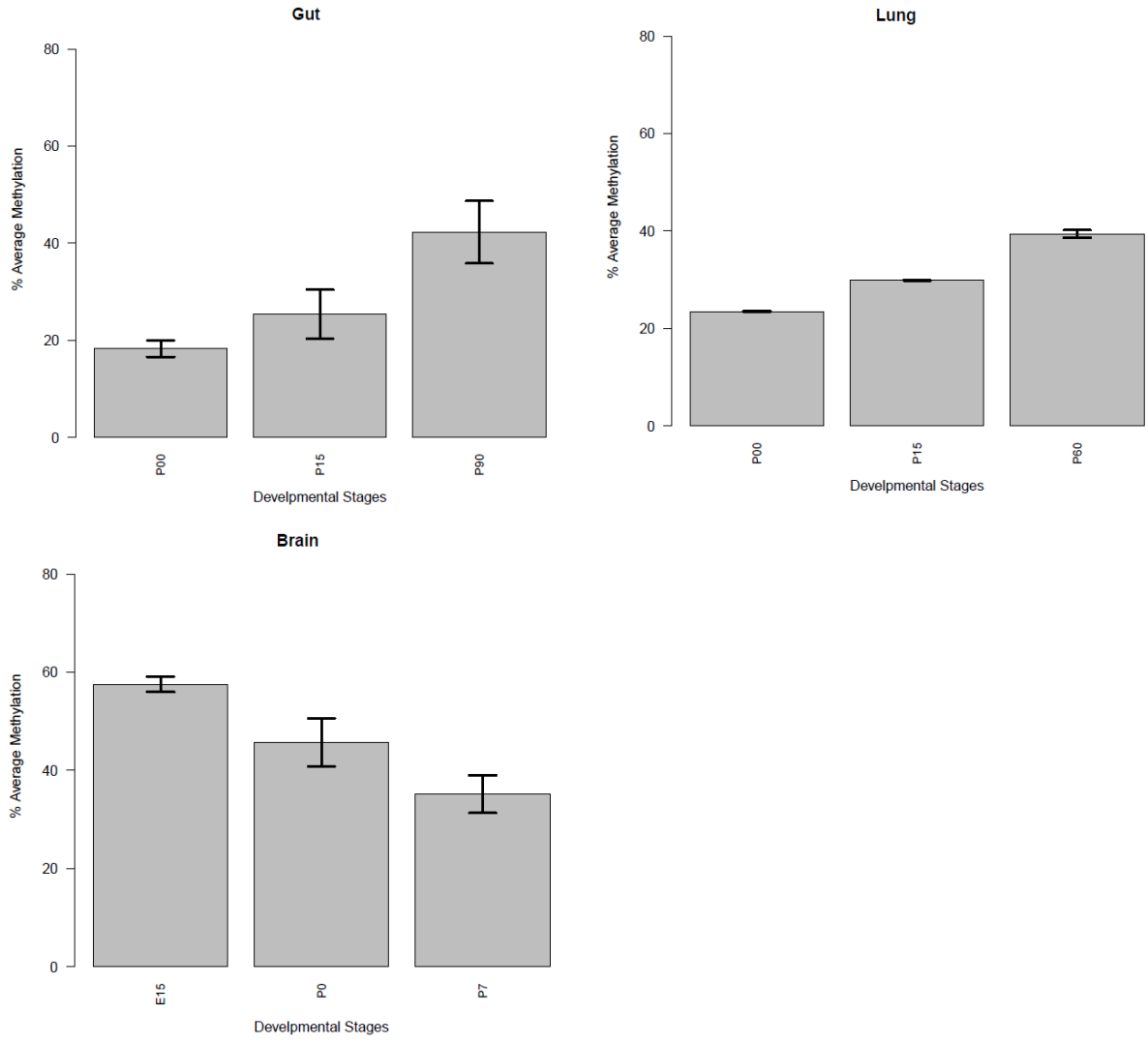

Figure 3.2. Average DNA methylation degree of the DDO gene during mouse ontogenesis for gut, lung and brain (c). The graphs show the global average methylation degree for each developmental stage. Each bar represents the mean percentage of DNA methylation, along with its corresponding 95% CI.

3.2.2 Methylation classes analysis during CpG methylation process

The methylation phenomenon, observed in gut and lung, was analysed in terms of methylation classes distribution (Fig. 3.3). Based on the number of observed methylated CpGs, amplicons-reads were classified into seven methylation classes: unmethylated, mono- methylated, bi- methylated, tri- methylated, tetra- methylated, penta- methylated, hexa- methylated. A methylation class is defined as the number of methylated CpGs independently by their position. Fig. 3.3 shows how the frequencies of these classes vary in lung and gut along with CpG methylation increase. Moving from early to late developmental stages, a decrease of the un-methylated molecules (~2-fold for gut and ~1.6–fold for lung) and mono- methylated molecules (~2-fold for gut and ~1.4–fold for lung) frequency was observed. The frequency of bi- methylated molecules seems to not vary during the global methylation process related to ontogenesis while an increase of tri-, tetra-, penta- and hexa- methylated was observed.

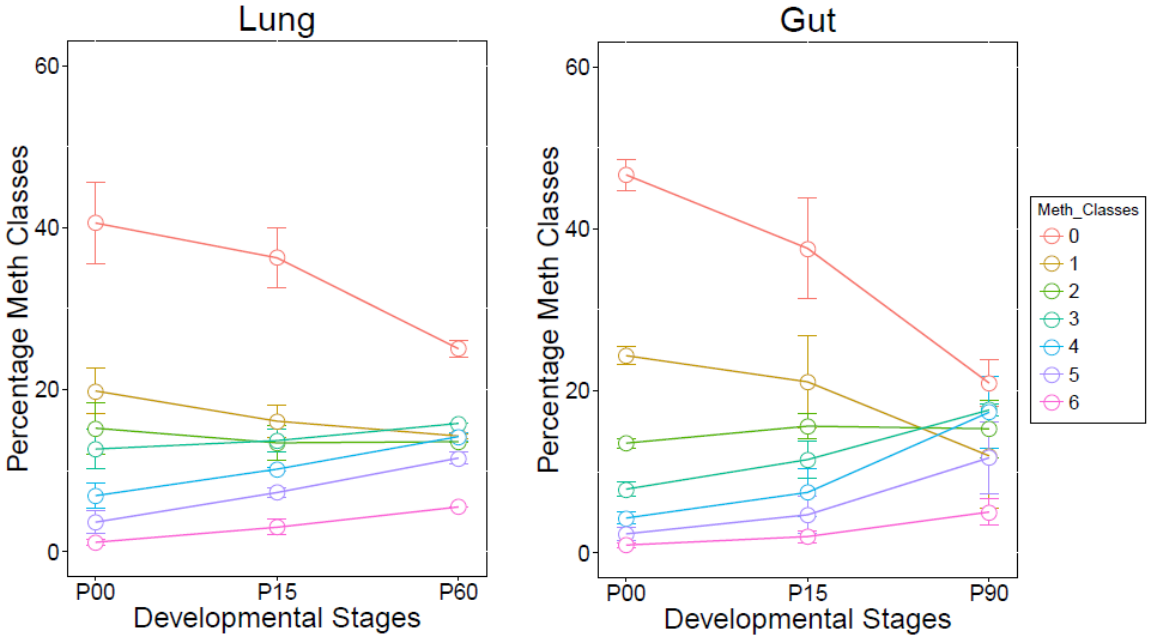

Figure 3.3. Frequency of each methylation class during mouse ontogenesis for lung and gut. Each methylation class is represented with a specific colour. All the values are expressed as the mean and corresponding 95% CI.

3.2.3 Methylation classes analysis during CpG demethylation process

When the demethylation phenomenon was analysed in terms of changes in methylation classes distribution, it has been found that, also in this case, different classes are differently

affected by the phenomenon (Fig. 3.4). In particular, a strong increase of un-methylated molecules frequency (~3.6-fold) and a mild increase (~1.6-fold) of mono-methylated ones was observed. The frequency of higher methylation classes (from bi - to hexa- methylated) generally decreased.

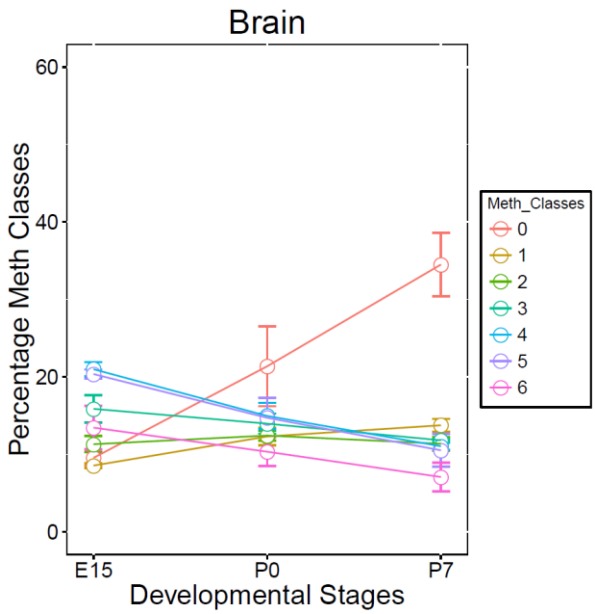

Figure 3.4. Frequency of each methylation class during mouse brain ontogenesis. Each methylation class is represented with a specific colour. All the values are expressed as the mean and the corresponding 95% CI.

In summary, at the final time points, all the tissues analysed showed a similar amount (around 40%) of average methylation (Fig. 3.2): statistical test failed to show any significant difference among them (one- way ANOVA, p-value > 0.1974). Therefore, by quantitative point of view, these tissues were not distinguishable. Conversely, qualitative analysis, based on the distribution of methylation classes, showed statistically relevant differences among the three tissues (Chi- squared test, p-value < 0.05) and, in particular, the profiles distributions in tissues undergoing methylation increase (lung and intestine) (Fig. 3.3) appeared very similar each with other and were substantially different with respect to the brain, which undergo demethylation (Fig. 3.4).

3.2.4 Epialleles frequency distribution at DDO promoter in brain is conserved in different mice

The DDO region under investigation contains 6 CpG sites, potentially giving origin to 64 epialleles (2^6) (Table 1). These ranges from unmethylated to fully methylated epialleles, including all the possible configurations of mono-methylated (n=6), di-methylated (n=15), tri-methylated (n=20), tetra-methylated (n=15), and penta-methylated (n=6).

Table 1 shows the epialleles frequency obtained from the analysis of brains (n=3) at E15 stage. In Fig. 3.5a a graphical representation of the 64 epialleles in brain from different mice is reported. Several interesting aspects were worth of note. At E15, where the average methylated level was evaluated at about 56% (Fig.3.2a), the fully methylated molecules represented about 13%, the unmethylated molecules were about 10% while the remaining 77% of cells bear one of the 62 intermediate epialleles (spanning from mono- to penta-methylated) which, although with slightly different frequency, were almost all represented in the brain cell mixture. Thus, all the 62 epialleles give a variable contribution to the global methylation (about 56%) observed at this developmental stage.

Surprisingly, to be noted that the frequency of each of the 64 possible epialleles is extremely conserved in the brain from different mice at E15 stage. Indeed, analysing the differences between epiallele frequencies of the three mice, it was found that the 90% of the methylation profiles shows an absolute difference in the percentage of occurrence lower than 0.5% and a relative difference lower than 0.33 (see Methods).

Taken together, these results suggest that the profile (frequency distribution) of all the epialleles present in the brain cell mixture (glia, glia subtypes, neurons) is constant in all mice analysed, providing an important insight: the generation of methylation profiles is deterministic and not stochastic.

3.2.5 Epialleles frequency distribution at DDO promoter is conserved in different mice also in other tissues

The eventuality that the frequency distribution epialleles was conserved in different mice also in other tissues was investigated. This hypothesis was tested in gut (Table 1 and Fig. 3.5b) and in lung (Table 1 and 3.5c).

Table 1 shows the epialleles frequency obtained from the analysis of gut (n=3) and lung (n=2) at P0 stage. In Fig. 3.5b-c a graphical representation of the 64 epialleles in gut and

lung from different mice is reported. In these tissues, at P0 stage fully methylated molecules represented about 1% for gut and for lung, the unmethylated molecules were about 45% for gut and about 40% for lung, while the remaining percentage (54% for gut and 59% for lung) of cells bear one of the 62 intermediate epialleles with slightly different frequency. So, as observed in brain, also in gut and lung there is a variable contribution of the 62 intermediate epialleles to the global methylation (about 20%) observed at this developmental stage.

Surprisingly, it was then confirmed that the frequency of each of the 64 possible epialleles is extremely conserved in these tissues from different mice at P0. Indeed, analysing the differences between epiallele frequencies of the three mice, it was found that the 90% of the methylation profiles shows an absolute difference in the percentage of occurrence lower than 0.7% and a relative difference lower than 0.33 (see Methods).

Taken together, these results suggest that the profile (frequency distribution) of all the epialleles present in the cell mixture of these tissues, is highly conserved in different mice.

Table 1. Epialleles frequency distribution at DDO promoter in different mice and different tissues (whole brain, lung and gut). A and B) 64 epialleles obtained with 6 CpG sites. These ranges from unmethylated to fully-methylated epialleles including all the possible configurations of mono-methylated (n=6), di-methylated (n=15), tri-methylated (n=20), tetra-methylated (n=15), and penta-methylated (n=6). White circles represent un-methylated CpG, while black circle represent methylated CpG. C) Epialleles frequency in different mice (n=3) in whole brain. D) Epialleles frequency in different mice (n=3) in gut. E) Epialleles frequency in different mice (n=2) in lung.

A	B			C			D	
	E15 BRAIN			P0 GUT			P0 LUNG	
Epialleles	M1	M2	M3	M1	M2	M3	M1	M2
UNMETH	9.9	10.2	8.6	47.6	47.6	44.7	43.1	38.0
FULLYMETH	10.9	15.7	13.7	1.0	0.7	1.2	1.4	1.0
-125	0.9	0.5	0.7	2.9	2.8	3.2	4.4	3.9
-175	1.3	0.8	0.8	5.0	3.4	4.5	2.8	2.8
-242	1.2	1.3	1.2	5.4	5.7	4.9	4.2	6.8
-318	0.3	0.4	0.3	1.5	1.1	0.8	0.4	0.5
-330	3.3	3.7	3.2	5.4	5.5	5.2	2.9	3.7
-363	1.8	1.6	2.2	5.2	5.8	4.8	3.7	3.5
-175-125	0.4	0.5	0.4	0.9	0.9	1.5	2.2	2.4
-242-125	0.4	0.3	0.3	0.8	0.7	0.9	1.6	2.3
-242-175	0.8	0.6	0.6	1.4	1.2	1.8	2.2	2.7
-318-125	0.1	0.0	0.1	0.2	0.1	0.1	0.1	0.0
-318-175	0.2	0.1	0.1	0.2	0.1	0.2	0.1	0.0
-318-242	0.3	0.2	0.3	0.4	0.6	0.3	0.3	0.5
-330-125	0.6	0.3	0.5	0.7	0.3	0.5	0.4	0.5
-330-175	0.8	0.8	0.8	1.1	0.8	1.2	0.3	0.2
-330-242	1.6	1.4	1.9	1.8	1.5	1.5	1.3	1.0
-330-318	1.7	1.3	1.6	1.0	1.5	0.9	1.9	2.8
-363-125	0.2	0.2	0.2	0.5	0.3	0.4	0.4	0.4
-363-175	0.5	0.4	0.4	0.9	0.8	0.9	0.2	0.3
-363-242	0.5	0.4	0.7	1.2	1.1	1.1	0.6	0.7
-363-318	0.3	0.4	0.4	0.4	0.3	0.3	0.3	0.3
-363-330	3.0	3.4	3.8	2.0	3.0	2.6	2.0	2.8
-242-175-125	0.6	0.4	0.7	0.5	0.7	0.8	2.0	2.8
-318-175-125	0.1	0.1	0.0	0.0	0.1	0.0	0.0	0.1
-318-242-125	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.2
-318-242-175	0.2	0.1	0.2	0.2	0.1	0.1	0.2	0.2
-330-175-125	0.4	0.4	0.3	0.2	0.2	0.3	0.2	0.2
-330-242-125	0.9	0.5	0.9	0.3	0.2	0.2	0.5	0.5
-330-242-175	1.1	1.0	1.0	0.7	0.3	0.5	0.7	0.5
-330-318-125	0.4	0.3	0.3	0.1	0.1	0.1	0.3	0.4
-330-318-175	0.7	0.6	0.6	0.3	0.3	0.2	0.3	0.3
-330-318-242	2.0	1.7	2.0	0.7	0.9	0.8	1.5	1.7
-363-175-125	0.2	0.1	0.2	0.2	0.2	0.3	0.1	0.3
-363-242-125	0.2	0.2	0.2	0.2	0.1	0.2	0.2	0.3
-363-242-175	0.5	0.3	0.4	0.4	0.3	0.5	0.2	0.3
-363-318-125	0.1	0.1	0.1	0.1	0.0	0.1	0.0	0.0
-363-318-175	0.1	0.2	0.1	0.1	0.1	0.1	0.0	0.0
-363-318-242	0.4	0.4	0.5	0.2	0.2	0.2	0.2	0.1
-363-330-125	0.6	0.7	0.7	0.3	0.2	0.3	0.2	0.4
-363-330-175	1.2	0.9	0.9	0.6	0.5	0.7	0.2	0.2
-363-330-242	2.6	2.7	2.9	0.9	1.6	1.3	1.2	1.1
-363-330-318	4.4	3.3	4.4	1.1	1.8	1.6	3.1	4.4
-318-242-175-125	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1
-330-242-175-125	1.3	1.1	1.3	0.3	0.1	0.3	0.7	0.5
-330-318-175-125	0.6	0.6	0.6	0.1	0.1	0.1	0.2	0.2
-330-318-242-125	1.3	1.2	1.3	0.1	0.3	0.2	0.8	0.5
-330-318-242-175	1.9	1.6	1.8	0.3	0.4	0.4	1.0	0.5
-363-242-175-125	0.3	0.4	0.5	0.1	0.1	0.2	0.2	0.2
-363-318-175-125	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0
-363-318-242-125	0.2	0.3	0.2	0.1	0.0	0.0	0.1	0.1
-363-318-242-175	0.5	0.3	0.3	0.1	0.0	0.1	0.1	0.0
-363-330-175-125	0.8	0.8	0.7	0.1	0.2	0.2	0.2	0.2
-363-330-242-125	1.3	1.4	1.3	0.2	0.3	0.3	0.4	0.5
-363-330-242-175	2.3	2.5	2.2	0.4	0.4	0.7	0.6	0.3
-363-330-318-125	1.1	1.2	1.0	0.1	0.2	0.2	0.4	0.4
-363-330-318-175	2.4	1.8	1.9	0.3	0.3	0.4	0.4	0.3
-363-330-318-242	7.7	7.1	7.3	1.3	1.7	1.7	2.4	2.3
-330-318-242-175-125	2.3	2.3	2.6	0.3	0.3	0.5	1.0	0.5
-363-318-242-175-125	0.6	0.4	0.5	0.1	0	0.1	0.1	0.1
-363-330-242-175-125	2.8	3.4	2.7	0.3	0.3	0.3	0.6	0.5
-363-330-318-175-125	1.7	2.1	1.8	0.1	0.1	0.2	0.3	0.3
-363-330-318-242-125	5.1	4.8	4.8	0.4	0.4	0.7	1.0	0.7
-363-330-318-242-175	7.7	8.0	7.6	0.8	0.8	1.3	1.4	0.9

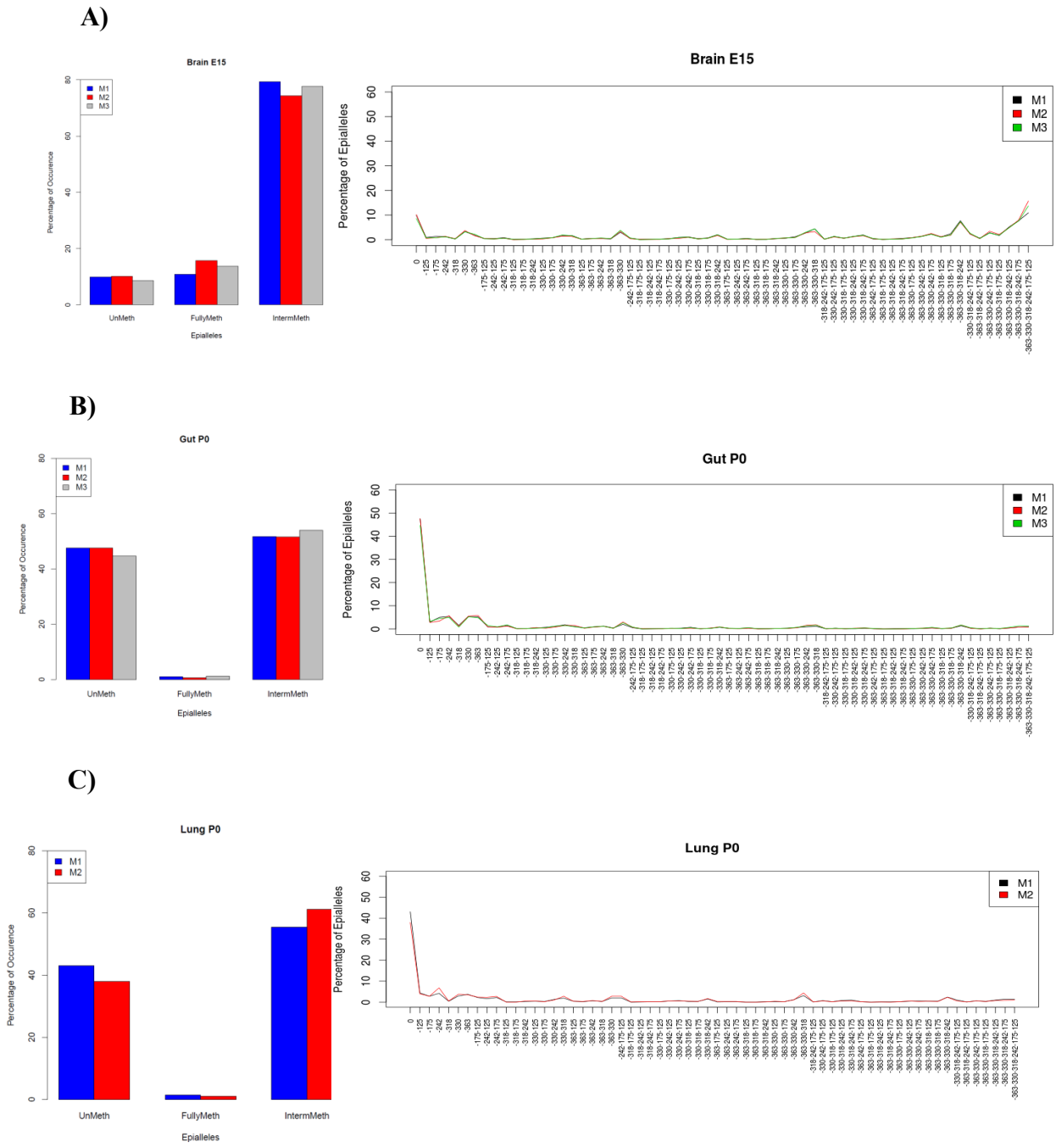

Figure 3.5. Graphic representation of epialleles frequency distribution at DDO promoter in different mice and different tissues: A) brain, B) gut, C) lung. On the left, the percentage of unmethylated, fully-methylated and intermediate methylation epialleles is reported for each mouse. On the right, on x-axis all the possible ^{m}CpGs combinations (epialleles; unmethylated, mono-, bi-, tri-, tetra-, penta-, fully- methylated) are reported, while on y- axis the epialleles frequency is reported.

3.2.6 Epialleles frequency distribution at DDO promoter is conserved in different mice also in different developmental stages

It has been then evaluated if the conservation of epialleles frequency distribution at DDO promoter in different mice was retained also in different developmental stages.

To be tested if the correlation was affected by the extreme epialleles (unmethylated and fully- methylated) frequency values, it was calculated both including and removing them. As an example, in Fig. 3.6 a scatterplot with and without the extreme epialleles is reported only for a pair of mice (Mouse 1 and Mouse 2) at adult stage (P7) in brain.

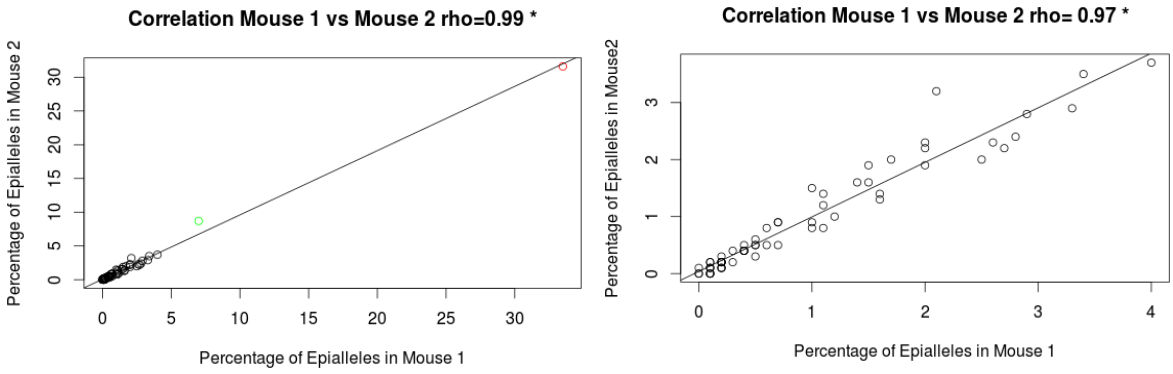

Figure 3.6. Pearson correlations between frequency of epialleles in mouse 1 and mouse 2 at adult stage (P7) in brain. On the left, the scatter plot includes unmethylated (red circle) and fully- methylated (green circle) epialleles. On the right, the same scatter plot without including unmethylated and fully- methylated epialleles. * indicates statistically significant correlations.

To be note that in both cases, there is a high correlation between the two mice as far as the epialleles frequency (Pearson correlation coefficient > 0.9, p-value < 2.2e-16).

Although the analysed region undergoes significant methylation changes during development, this high conservation was strikingly retained in all analysed tissues at each of the successive analysed developmental stages (P0 and P7 for brain, P15 and P90 for gut, P15 and P60 for lung). By excluding the unmethylated and fully- methylated epialleles, the observed distribution of epialleles frequency is highly correlated between mice pairs (Pearson correlation coefficient > 0.9, p-value<0.05; Table 2).

Taken together, these results suggest that each epiallele contributes with a different frequency to the global methylation, but its relative frequency is always conserved among different mice for each tissue under investigation (brain, lung and gut) at a given

developmental stage. Thus, the methylation profiles trend in different mice, during the same developmental stage, is highly conserved, indicating that probably the CpG sites methylation is due to a deterministic event rather than a stochastic one. This demonstrates again the deterministic nature of the genesis of the methylation profiles and of their distribution inside cell population.

Table 2. Pearson correlation values between epialleles frequency distribution in different mice in different developmental stages and in different tissues. The unmethylated and fully methylated epialleles were removed. All the correlations are statistically significant (p-value < 0.05)

Tissue	Stage	M1 *vs* M2	M1 *vs* M3	M2 *vs* M3
HB	E15	0.98	0.99	0.98
HB	P0	0.96	0.94	0.88
HB	P7	0.97	0.92	0.96
Gut	P0	0.97	0.98	0.97
Gut	P15	0.93	0.88	0.97
Gut	P90	0.75	0.80	0.96
Lung	P0	0.95	-	-
Lung	P15	0.99	-	-
Lung	P60	0.99	-	-

3.2.7 Stable intermediate epialleles among different tissues at adult stage

In order to evaluate if there is a correlation between tissues as regard the intermediate epialleles distribution, for each tissue, the epialleles frequencies have been averaged among the three mice. The analysis was carried out at the adult stage (Brain=P7; Gut=P90; Lung=P60), because at this stage methylation and demethylation processes reach about a same global level (40%; Fig. 3.2). Unmethylated and fully- methylated epialleles were excluded from this analysis, because their frequencies are highly different from the intermediate epialleles ones. (Fig. 3.7a). If the 62 intermediate epialleles, ranging from mono- to penta- methylated, are considered, a strong positive correlation among the three tissues was found, and in particular in the case of:

- gut against brain, rho=0.85, p-value $<2.2*10^{-16}$ (Fig. 3.7b);
- gut against lung, rho=0.63, p-value = $1.3*10^{-9}$ (Fig. 3.7c);
- brain against lung, rho=0.87, p-value $<2.2*10^{-16}$ (Fig. 3.7d)

Taken together, these results suggest that the intermediate epialleles frequencies are highly conserved among different tissues at adult stage.

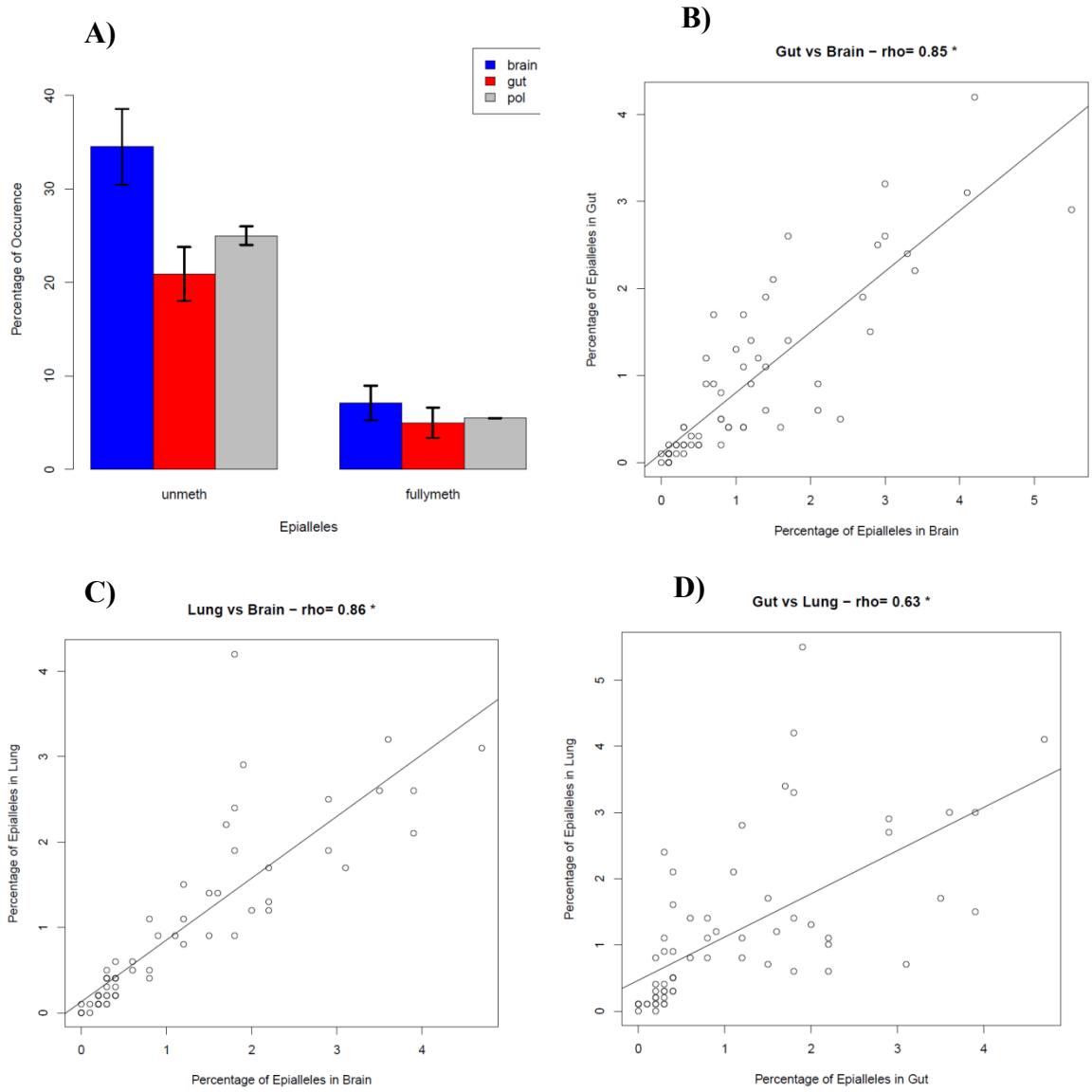

Fig. 3.7 Pearson correlations between tissues at adult stage (Brain=P7; Gut=P90; Lung=P60). A) Each bar represents the mean percentage of unmethylated and fully-methylated epialleles in brain (blue), gut (red) and lung (grey), along with its corresponding 95% C.I. B) Correlation between the percentage of epialleles in gut and in brain; C) Correlation between the percentage of epialleles in lung and in brain; D) Correlation between the percentage of epialleles in gut and in lung. * indicates statistically significant correlations.

3.2.8 Dynamic change of epialleles during brain development

Because DDO region undergoes to significant demethylation, along with DDO gene activation, during early post-natal stages (from 57% (in E15) to 35% (in P7) on average; Fig. 3.3a), it has been also decided to follow the dynamic change of each of the 64 epialleles during this period (Fig. 3.8).

Analysis of the results showed that the frequency of the unmethylated epialleles increases from 10% (in E15) to 30% (in P7) where fully methylated molecules decreases from 14% (in E15) to 5% (in P7). To identify those epialleles whose frequency changes in a statistically significant manner from the early to late developmental stage, it has been defined a coefficient of epialleles frequency variation (see Methods). Only those epialleles, whose 4-standard error confidence interval of the coefficient of variation does not include the value zero, were considered to be varying in a statistically significant manner among stages. As shown in Fig. 3.9, with exception of few penta- (-363-330-318-242-175; -363-330-318-242-125) and one tetra-methylated epiallele (-363-330-318-242) that showed a consistent decrease (≈4%), most of the intermediate epialleles underwent only slight changes with a prevalence of increased frequency of some mono and di-methylated and decreased frequency of some tri, tetra- and penta-methylated molecules.

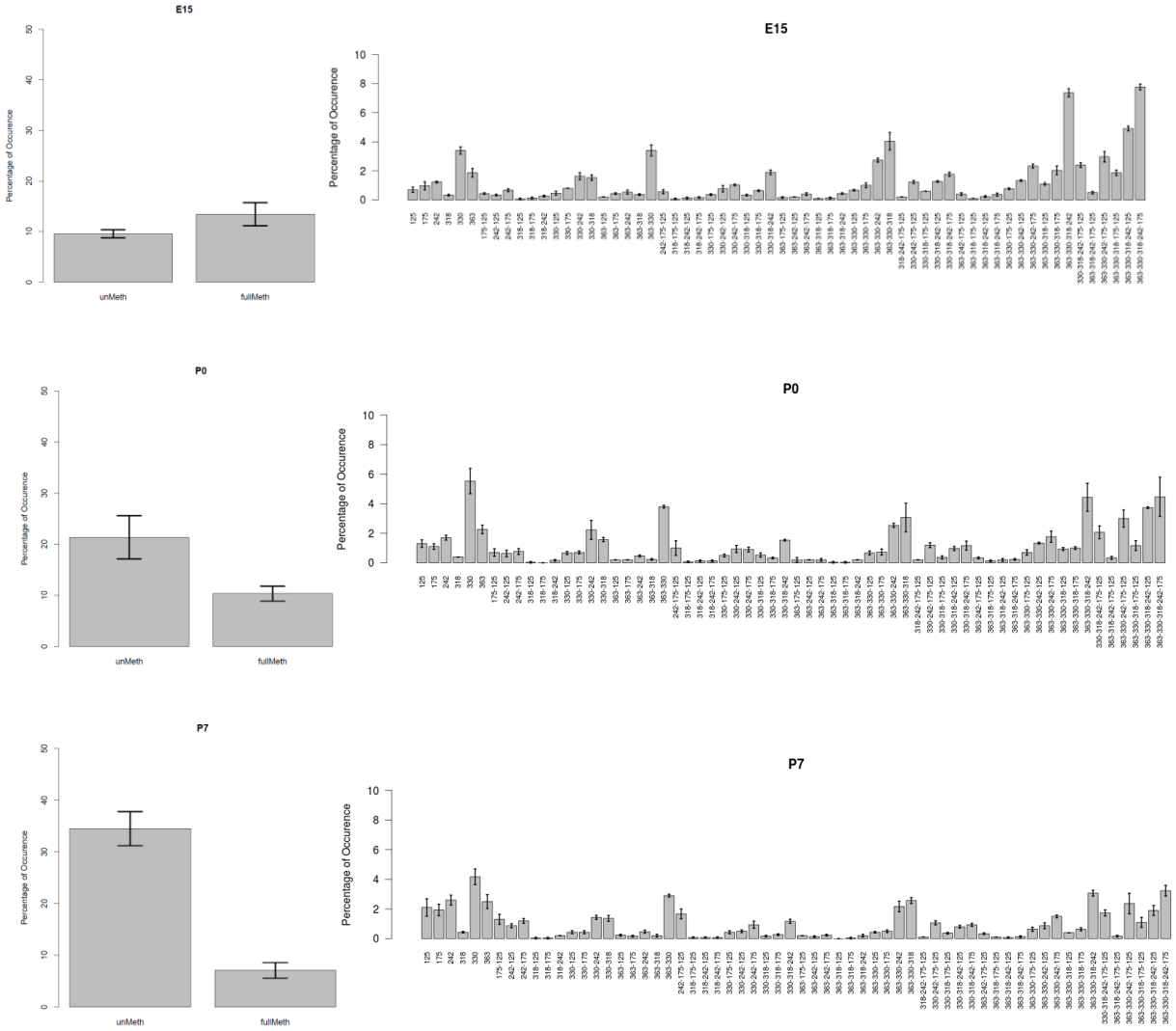

Figure 3.8. Methylation profiles (epialleles) analysis of the DDO promoter region during early brain developmental stages (from E15 to P7). On the left the frequencies of un- and full- methylated epialleles are reported. On the right the methylation profiles trend of all the intermediate epialleles (from mono- to penta- methylated), generated by the combination of the 6 CpG sites present in this genomic region, is shown.

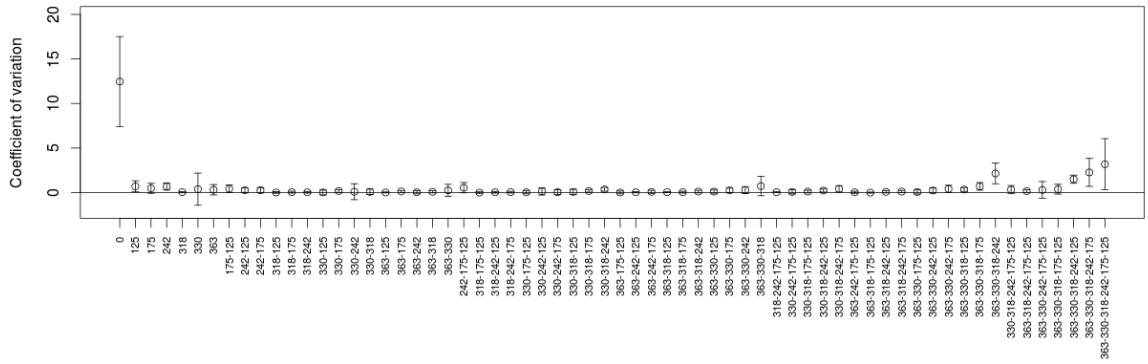

Figure 3.9. Coefficient of epialleles frequency variation from early to late developmental stage. For each epiallele the coefficient of epialleles frequency variation among developmental stages is reported with the corresponding 4 standard errors C.I. Coefficient variation values, which are different from zero, indicate statistically significant variations among stages.

3.3 Discussion

Methylation and demethylation phenomena have been addressed by several different techniques and have been traditionally investigated through quantitative approaches, namely by studying CpG average methylation degree in a given genomic region [85-91]. Such kind of approach gives important information on the general relationships between methylation and expression but might be not able to describe the complex epigenetic structure and dynamics within a cell population. Indeed, quantitative approaches assume that methylation status is quite uniform in the population of cells under study.

In this work, it has been tested a different approach capable of looking at the individual methylation conformation, in terms of methylation classes and epialleles, of single molecules. This qualitative approach is potentially able to describe methylation dynamics with a higher level of definition than quantitative one. Quantitative and qualitative approaches were compared in describing methylation and demethylation dynamics occurring through developmental stages of three different mouse tissues. DDO was used as model gene because, during development, this gene undergoes methylation in lung and gut (as described in this work) and demethylation in the brain [116]. Moreover, the chosen system allowed to study these processes in a natural context (tissues rather that cell lines), and to follow methylation changes occurring during development, thus avoiding any artefacts due to the action of drugs. Finally, the analysed DDO methylation/demethylation processes are relevant for gene function, and thus possibly a consequence of a

deterministic phenomenon, since these are accompanied by gene activation (brain) or gene repression (gut and lung) [116].

Overall, the results show that the qualitative approach has a greater informative content than the quantitative one. First, the results notice that at the end of the methylation/demethylation events occurring in the analysed time window, equal global methylation degree in gut, lung and brain, corresponded to a different distribution of methylation classes. In particular, the distribution of methylation classes are very similar in lung and gut and clearly differ from that observed in brain. The similarity between tissues undergoing methylation and their difference from a tissue undergoing demethylation suggests an effect of the processes (methylation *versus* demethylation) that produced the observed molecules on their class distribution. Thus, qualitative approach allowed to describe in detail the molecules created during the methylation and demethylation processes, respectively.

Then, to evaluate if qualitative approach allows to gain further information, the evolution of methylation profiles (epialleles) during mouse development in somatic tissues was evaluated.

When the epialleles level analysis has been performed, it has been found that:

1. the frequency of each of the 64 possible epialleles and their distribution is extremely conserved in different mice, also analysing different tissues (whole brain, lung and gut).
2. This high conservation is strikingly retained at each of the successive analysed developmental stages (E15, P0 and P7 for brain, P0, P15 and P90 for gut, P0, P15 and P60 for lung).
3. During development, there is a considerable change of extreme epialleles (unmethylated and fully- methylated), while most of the intermediate epialleles undergoes only slight changes.
4. These intermediate epialleles frequencies are highly conserved among different tissues at adult stage.

These results suggest that each epiallele contributes with a different frequency to the global methylation, but its relative frequency is always conserved among different mice for each tissue under investigation (brain, lung and gut) at a given developmental stage. Each of the observed combination of ^{m}CpG positions represents a specific single molecule methylation profile occurring in a given percentage of cells of the analysed cell populations. Thus, the

observed phenomenon appear to be a consequence of a precise mechanism governing the formation and maintenance of predetermined methylation profiles in each of the different cell types forming a tissue rather than a consequence of a random action of DNMTs at the analysed region's CpGs. In other words, these results led to the conclusion that in the somatic tissues, epialleles are generated in a perfectly conserved fashion and the frequency of each epiallele is determined in a well- orchestrated fashion. Indeed, this highly conserved methylation profiles trend indicates that probably the CpG sites methylation (and thus, the generation of epialleles) is not a stochastic event, rather a deterministic one, developmentally regulated, leading to an orchestrated distribution of epialleles among the entire population of cells. This deterministically regulated distribution of different epialleles evokes the possible existence of a novel combinatorial code of CpG methylation. These results might encourage the practice of the novel qualitative approach to study DNA methylation, because it accounts for the high polymorphism of methylation profiles in cells populations derived from individual somatic tissues [39] and may potentially detect cell origin, functional state or structure state. Indeed, once again it is clear that the simple evaluation of averaged degree of methylation does not give a complete picture of methylation status inside a cell population. On the contrary, the qualitative analysis can provide an adding value to the traditional methylation analysis, because it allows a more detailed and faithful tracking of epialleles inside a cell population. This is useful above all in tissues, which are composed by a mosaic of epigenetically different cells.

3.4 Materials and Methods

3.4.1 Deep Bisulfite Amplicon Sequencing (Deep- Bis)

Whole brain, lung and gut tissues were collected from three mice at different developmental stages including the following time points:

- Brain: embryonal stage 15 days (E15), at birth (P0), 7 post-natal days (P7)
- Gut: at birth (P0); 15 post-natal days (P15); 90 post-natal days (P90)
- Lung: at birth (P0); 15 post-natal days (P15); 60 post-natal days (P60)

Methylation status was assessed through a strategy based on the locus- specific amplification of bisulfite- treated genomic DNA, followed by Illumina MiSeq sequencing.

The following bisulfite-specific primers were used to obtain tiled amplicons: DDO PR1 fw 5'-gtgtgtttTtgaggaggtgaTaTtTa- 3' (nt position from -468 to -444) - and DDO PR1 rev 5'-aActtaccctccattAAtccatAcc-3' (nt -88 to -63) (amplicon size 405 bp). The capital letters in the primers sequences indicate the original C or G, respectively.

Reads in FASTQ format obtained from sequencing were processed with Paired-End reAd mergeR (PEAR; http://sco.h-its.org/exelixis/web/software/pear/) for paired end assembling and initial quality filtering. Only those reads with the following features were retained:

- a mean quality score (Phred) greater than 30;
- a read length between 400 and 500 nucleotides;
- an overlapping region within paired-end reads of at least 40 nucleotides.

Resulting reads were then converted in FASTA format using PReprocessing and INformation of SEQuence (Prinseq; http://prinseq.sourceforge.net/).

Reads were then aligned to the corresponding bisulfite converted reference sequence using AmpliMethProfiler (see Chapter 2).
By applying a series of quality filters on the read length and the alignment quality, it has been retained only those reads characterized by:

- length ±50% compared to the reference length;
- primer of the corresponding gene identified with at least 80% of similarity;
- at least 98% of bisulfite efficiency, calculate as percentage of conversion of non-CpG cytosines into thymines over total number of C in a context not CpG;
- to be aligned for at least 60% of their bases with the reference sequence
- C status at the all CpG positions recognised as methylated (1) or unmethylated (0). Aligned reads containing ambiguous calls (presence of gaps or A or G) at the CpG dinucleotide were removed.

Methylation state was estimated by observing base calls (T/C) at CpG sites in the mapped reads.

On average, for each sample and for each developmental stage, we obtained an average of 152,572 amplicon reads for the brain, 49,976 for the lung and 74,400 amplicons for the gut.

3.4.2 Quantitative methylation analysis

The methylation percent of each developmental stage was calculated by averaging CpGs methylation percentages of all CpG sites in the target region and then over all samples belonging to the same developmental stage.

Let $T=\{t_1, \ldots, t_n\}$ be the set of analysed mice.

Let $S=\{s_1, \ldots, s_k\}$ be the set of analysed developmental stage.

Let $n\_reads\ (t_i, s_j)$ be the number of reads obtained for the mouse t_i at the developmental stage s_j.

Let $m\ (t_i, s_j, r,k) \in 0,1$, the methylation status of the k-th CpG site in the read r in the mouse t_i at developmental stage s_j.

For each mouse t_i and each developmental stage s_j, the methylation percentage of the k-th CpG site was computed as:

$$m\ (t_i, s_j, k) = \frac{\sum_{r=1}^{n\_reads(ti,sj)} m(ti, sj, r, k)}{n\_reads(t_i, s_j)}$$

For each mouse t_i and each developmental stage s_j, the average methylation percentage of the whole region was computed as:

$$\mathrm{m}\ (t_i, s_j) = \frac{\sum_{k=1}^{\mathrm{nCpG}} \mathrm{m}\ (t_i, s_j, k)}{\mathrm{nCpG}}$$

where *nCpG* is the number of the CpG sites.

Finally, for each developmental stage j, the average methylation percentage was computed as:

$$\mathrm{m}\ (s_j) = \frac{\sum_{i=1}^{\mathrm{n}} m\ (t_i, s_j)}{n_{sl}(j)}$$

where $n_{sl}\ (j)$ is the total number of samples belonging to the developmental stage j.

3.4.3 Methylation classes frequency

The frequency of each methylation class in each developmental stage was calculated as the ratio of the number of reads belonging to a given methylation class over the total number

of reads obtained for that developmental stage. The obtained value was then averaged on all samples belonging to the same developmental stage.

Denoting with $(t_i, s_j) \in SxT$ the sample t_i at the developmental stage s_j and with $n_{sl}(j)$ the number of samples belonging to the developmental stage s_j.

For each methylation class j in {1,...,k}, the frequency of the methylation class j in each stage s_l was computed as:

$$classMeth\ (j, s_l) = \frac{\sum_{i=1}^{n_{sl}} \frac{n\_reads\ (t_i, s_l, j)}{n\_reads\ (t_i, s_l)}}{n_{sl}(j)}$$

where $n\_reads(t_i, s_l, j)$ is the number of passing filter reads for the mouse t_i at the developmental stage s_l falling into the methylation class j.

3.4.4 Epialleles frequency

For each sample, the frequency of each epiallele in each developmental stage was calculated as the ratio of the number of reads of a given epiallele over the total number of reads found in each developmental stage and in each sample. The obtained value was then averaged on all samples belonging to the same developmental stage.

Denoting with $(t_i, s_j) \in SxT$ the sample t_i at the developmental stage s_j and with $nreads(ti, sj)$ the total number of reads belonging to the sample t_i at the developmental stage s_j.

For each epiallele e in $\{1,\ldots,2^{nCpG}\}$, the frequency of epiallele e in the stage s_j in the sample t_i was computed as:

$$EpiFreq\ (t_i, s_j, e) = \frac{n\_reads\ (t_i, s_j, e)}{n\_reads\ (t_i, s_j)}$$

where $n\_reads(t_i, s_l, e)$ is the number of passing filter reads for the mouse t_i at the developmental stage s_l for the epiallele e.

For each epiallele e in $\{1,\ldots,2^{nCpG}\}$, its frequency in the stage s_j was computed as:

$$EpiFreq\ (s_j, e) = \frac{\sum_{i=1}^{n_{sl}(s_j)} EpiFreq\ (t_i, s_j, e)}{n_{sl}\ (s_j)}$$

where $n_{sl}(s_j)$ is the total number of samples belonging to the developmental stage j.

3.4.5 Epialleles similarity

The epialleles similarity among mice belonging to the same developmental stage was assessed by analysing the distribution of absolute and relative difference between their observed epiallele frequencies.

For each epiallele *j*, let $EpiFreq\left(t_i, s_j, e\right)$ be the frequency of epiallele *e* in the mouse t_i at the developmental stage s_j, computed as described above.

For a pair of mice, t_i and t_l, the absolute and relative differences of epiallele *e* frequency at the developmental stage s_j were respectively calculated as:

$$\text{Abs\_Diff} = EpiFreq\left(t_i, s_j, e\right) - EpiFreq\left(t_l, s_j, e\right)$$

$$\text{Rel\_Diff} = \frac{\text{Abs\_Diff}}{EpiFreq\left(t_i, s_j, e\right) + EpiFreq\left(t_l, s_j, e\right)}$$

Then, the distribution of relative (absolute) differences for each developmental stage was defined, by considering the 2^{nCpG} relative (absolute) differences between the frequency of the same epiallele in each pair of mice belonging to that developmental stage.

Finally, the epiallele similarity for each developmental stage was evaluated by considering the 90th percentile of its relative (absolute) difference distribution.

3.4.6 Coefficient of epialleles variation from the early to late developmental stage

To identify those epialleles whose frequency changes in a statistically significant manner from the early to late developmental stage, the following approach has been used.

For each epiallele, it has been modelled the observed frequency of the epiallele in each mouse as linear function of the developmental stage. In this model, a zero slope means no variation, in terms of epiallele frequency, from the early to late stage. Hence, by means of a linear regression the intercept and the slope of the above model were estimated, along with their corresponding standard errors, for each epiallele. Only those epialleles, whose 4-standard error confidence interval of the slope (named as coefficient of epialleles variation) does not include the value zero, were considered to be varying in a statistically significant manner among stages.

3.4.7 Statistical analysis

An alpha level of 0.05 was used for all statistical tests.

All statistical analyses were performed using R statistical package ver. 3.2.1 (http://www.R-project.org).

Chapter 4

Tracking the evolution of CDKN2A and CDKN2B genes methylation profiles during acute myeloid leukemia progression

4.1 Introduction

Aberrant DNA methylation is extensively investigated for its role in promoting tumor evolution and chemo-resistance [117,118]. These epigenetic changes impact the biological activity of cells through their modification of transcriptional states and regulatory machinery.

Recent models of tumor origin and progression have compared carcinogenesis to evolutionary processes [119-121], where tissue provides the context for cancer cell evolution. According to these models, every round of cell division is driven by the acquisition of new mutations, arising multiple, genetically diverse subclonal populations of cells inside the tumor [122]. During tumor progression, cancer cells undergo clonal selection (selective pressure), leading to emergence, from the population of cancer cells, of clones best adapted to conditions within the tissue ecosystems, which provide the determinants of fitness selection, i.e. the adaptive landscape [123]. Clonal evolution involves the interplay of selectively advantageous or driver lesions, selectively neutral or passenger lesions and changes to the microenvironment [124] that modify the fitness of tumor cells carrying those lesions.

Besides genetic diversity, there is increasing recognition that epigenetic changes, such as DNA methylation and histone deacetylation, can occur throughout tumorigenesis [125]. Tumor cells carrying specific epigenetic variants may have an adaptive advantage. Since both genetic and epigenetic variations result in abnormal gene expression, it is possible that genetic and epigenetic mechanisms synergize to determine the speed of cancer progression.

Epigenetic alterations are linked to tumour heterogeneity: within an individual tumour, DNA methylation patterns are highly polymorphic. Increasing coverage of methylation at single nucleotide level shows in normal and cloned cells in culture an extreme degree of polymorphism [39]. A possible explanation of this heterogeneity is the relation between DNA methylation and DNA damage and repair [126-129].

Acute myeloid leukemia (AML) is a hematological malignancy characterized by uncontrolled proliferation of clonal neoplastic hematopoietic precursor cells leading to the disruption of normal hematopoiesis and bone marrow failure. Several recent studies point to a major pathogenic role for aberrant epigenetic programming in acute myeloid leukemia [130-133].

Aberrant methylation in AML is associated with hypomethylation in oncogenes, such as H-RAS [134] and hypermethylation in tumor suppressor genes (TSGs) [135,136], which are gene involved in the cell cycle. Inside this TSGs group, of particular note are the cyclin-dependent kinase inhibitors (CKI) CDKN2A (p14ARF) and CDKN2B (p15). These genes control the progression from the G1 to the S phase of the cell cycle [134,137,138].

Despite strong evidence that AMLs, and tumors in general, are composed of mixtures of distinct epigenetic clones rather than being monoclonal, the accurate description of this epigenetic heterogeneity overtime, i.e., at different stages of AML progression and before and after treatment is still lacking. Indeed, currently, most of the studies on methylation in cancer cells, both genome-wide or of specific DNA segments, measure quantitative differences, i.e., percentage of methylation of single CpGs in a given sequence. It is impossible to deduce from these data the configuration of methylated CpGs in the same molecule.

In the present study, deep sequencing of bisulfite treated DNA derived from bone marrow and peripheral blood cells of a patient during different stages of AML was carried out. Combining qualitative and quantitative approaches, it has been analysed methylation of the

promoter regions of CDKN2A (p14ARF) and CDKN2B (p15) suppressor genes during the progression of the disease, the demethylating therapy with 5-azacytidine (5-azaC, Vidaza), during the remission and the final relapse. The analysis was carried out by investigating the methylation classes and the distribution of specific epialleles, alleles differing only by methylation. The clonal composition at epigenetic level was dissected during the different phases of disease progression focusing on few specific loci. The restriction of the epigenomic landscape increases the coverage and allow the identification of rare epialleles, which may evolve and become dominant at the end stage of the disease. Importantly, a primary human cells carrying dominant oncogene mutations enter replicative senescence and does not grow if CDKN2A or 2B are expressed [139]. Methylation and silencing of these suppressors is positively selected during tumor progression. This is the first extensive qualitative methylation analysis of CDKN2A-B during the AML in the same patient.

4.2 Results

4.2.1 Quantitative methylation analysis of p14 and p15 genes

DNA methylation was assessed through a strategy based on the locus- specific amplification of bisulfite- treated genomic DNA. In the case of p14, a region of 431 bp around the transcription start site (TSS), spanning nucleotides -33 to +398 and including 35 CpG sites was analysed. In the case of p15, a region of 365 bp around the transcription start site (TSS), spanning nucleotides -275 to -+90 and including 27 CpG sites was analysed.

Samples are derived from cells of the bone marrow cells of a single patient at time 0 (MDS diagnosis, myelodysplasia, no signs of acute leukemia) to 2 months later (AML, diagnosis acute myeloid leukemia), 6 months after the therapy with a demethylating drug (THER), 8 months later (REL I) and final relapse (REL II).

The quantitative methylation analysis p14 and p15 region is shown in the Fig.4.1.

Quantitative analysis of p14 methylated alleles shows that the stages MDS and AML have a comparable methylation level (about 7%), which decreases in the successive stages THER, REL I and REL II (Fig.4.1, left).

Quantitative analysis of p15 methylated alleles shows that global methylation of the region analysed is comparable in all stages, except in stage THER (demethylating therapy) (Fig.4.1, right). Specifically, again MDS and AML has a comparable global methylation level (about 15%, and precisely 15% for MDS and 17% for AML), which dramatically decreases during therapy (THER), until to reach an about 2%. Then, going towards the last stage (REL II), the global methylation level gradually increases again, until to reach almost the initial level (about 20%, and precisely 14% for REL I and 19% for REL II).

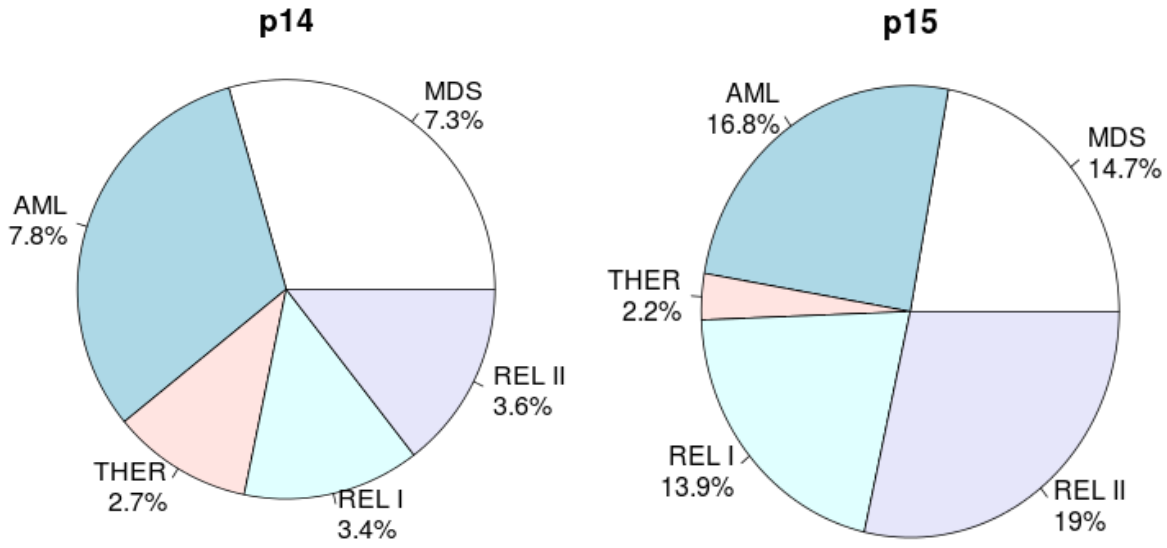

Figure 4.1. DNA methylation degree of p14 and p15 genes during disease progression. The pie charts show the average methylation degree for each stage of the disease.

4.2.2 Distribution of methylation classes of p14 and p15 epialleles

In this section, it has been determined in the same samples indicated in Fig.4.1 the distribution of the various classes of epialleles. Specifically, on the basis of the methylated CpGs, amplicons-reads have been classified into methylation classes, from unmethylated to fully- methylated. A methylation class is defined by the number of methylated CpGs, independently on the location or position in the DNA molecule. Considering that the sequences derive from identical molecules with the same 5' and 3' ends, two sequences represent a single cell.

Analysing how methylation classes are distributed in the different stages of disease for p14 gene (Figure 4.2), it is to be observed that:

- in MDS and AML (onset of disease), the 7% of global methylation (Fig. 4.1a) is mainly represented by mono-, bi- and tri- methylated molecules, which account for more than the 60% of the population (64% for MDS and 68% for AML). 10% of cell population has constituted by unmethylated molecules (11% for MDS and 5% for AML). The remaining percentage (30%) constitute by molecules with higher levels of methylation (from tetra-methylated molecules onwards; 25% for MDS and 27% for AML).

In the successive stages (THER, REL I and REL II), the 5% of global methylation, observed in Fig. 4.1a, is given by molecules with low level of methylation and in particular:

- after therapy (THER), there is a net increase of unmethylated molecules, which constitute almost the 45% of the molecules population; the 52% of the population is composed by mono-, bi- and tri- methylated molecules; only the 3% of the population is composed by molecules with an higher level of methylation (from tetra-methylated molecules onwards).
- In the last two stages of disease (REL I and REL II), about the 35% of the molecules is un- methylated (36% for REL I and 35% for REL II); approximatively, the 60% is composed by mono-, bi- and tri- methylated molecules (59% for REL I and 58% for REL II); about the 5% (5% for REL I and 6% for REL II) of the population is composed by molecules with an higher level of methylation (from tetra-methylated molecules onwards).
- There is a consistent increase of molecules with low methylation level, essentially mono- methylated in REL I, which pass from 31% in THER to 34% in REL I, and bi- methylated in REL II, which pass from 15% in THER to 21% in REL II.

Taken together, these results suggest that p14 epialleles, at the onset of disease are highly heterogeneous, spanning from unmethylated to higher levels of methylation. After one year, induced by the demethylating action of the therapy, molecules with high methylation level (5- AzaC sensitive molecules) disappear, while low methylated molecules increase (5-AzaC resistant molecules). In the last two stages, corresponding to the relapse and death, these 5-AzaC resistant molecules increase their frequency and suggesting a strong positive selection on a specific class of epialleles.

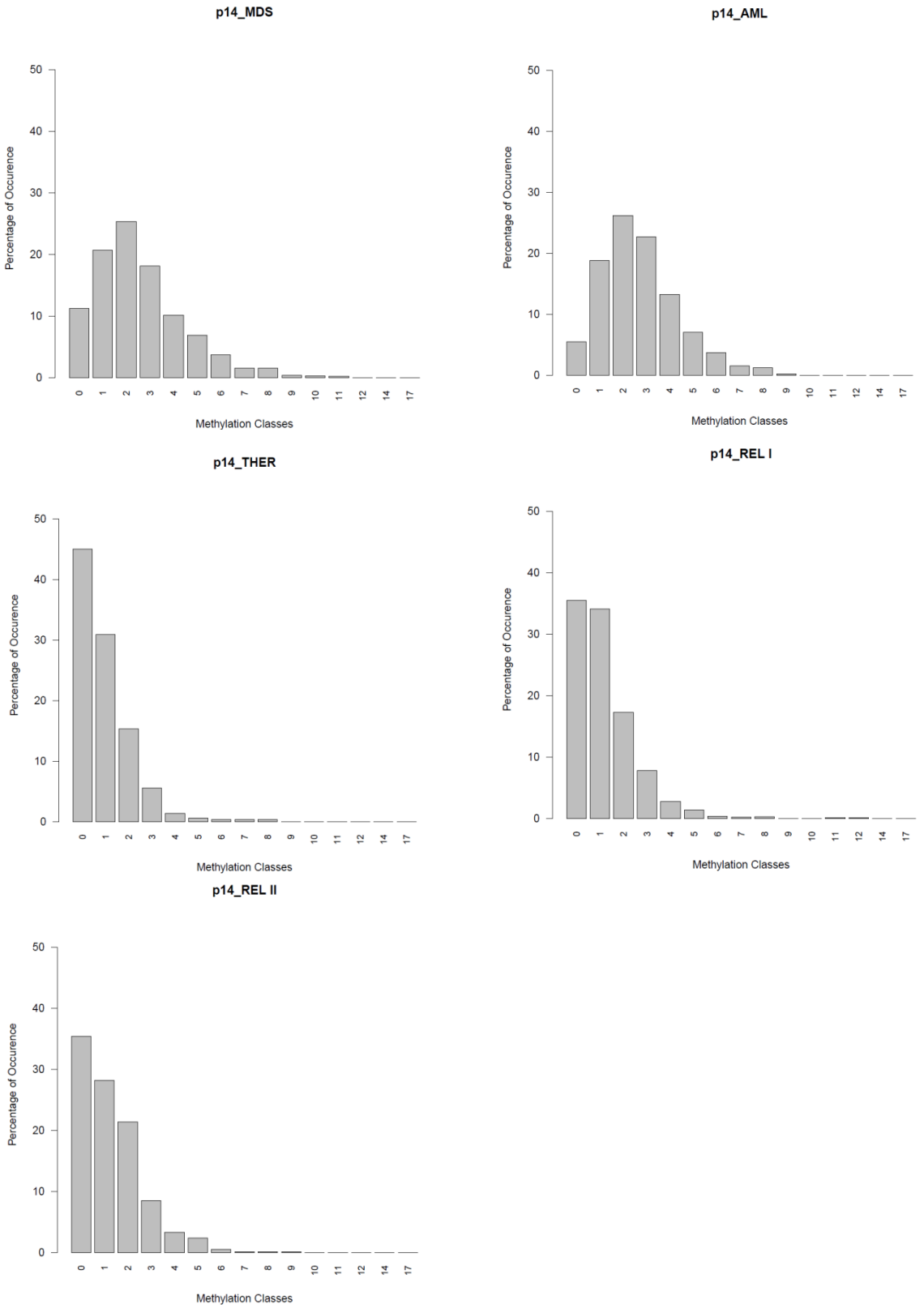

Figure 4.2. Percentage of each methylation class during disease progression for p14 gene. On x-axis the methylation classes are reported, while on the y- axis the frequency of each methylation classes is reported. Each bar represents a methylation class.

Fig.4.3 shows the distribution of the various classes of p15 epialleles during the evolution of the disease:

- in MDS and AML (onset of disease) 15% of methylated molecules (Fig. 4.1b) is represented by highly heterogeneous methylated molecules. Molecules with one to eight methyl groups represent 85% of the population (86.7% for MDS and 88.6% for AML) and each one of them is present with the similar frequency (less than 15%). The unmethylated molecules constitute less than 10% (7% for ToC and 5% for ToD), while the contribution of molecules belonging to higher methylation classes (containing from nine to 17 methylated CpG sites) is minimal (6% for ToC and ToD).
- after one year of therapy (THER), the net decrease of methylation level observed (about 2%, Fig. 4.1b) is balanced by the net increase of unmethylated molecules, which represent the 64% of the whole population, and of the mono- methylated classes, which represent the 24% of the whole population. The remaining 12% is represented by molecules belonging to other methylation classes (from bi-methylated onwards), that, individually, are less than 10%.
- In the last two stages of disease (REL I and REL II), the composition of molecules inside the 20% of global methylation (Fig. 4.1b) highly heterogeneous: in general, these molecules belong to all methylation classes (from mono- methylated onwards). They are present individually with a similar frequency (less than 15%) and globally contributed for about 80% (77% for REL I and 83% for REL II) to the composition of molecules population of these stages. There is a net loss of unmethylated molecules with respect to THER stage: these molecules constitute only the about 20% (23% for REL I and 17% for REL II). To be note that in REL II there is a light increase of molecules containing from 7 to 11 methylated CpG sites with respect to REL I and to the initial stages (MDS and AML).

Taken together, these results suggest that p15 epialleles, at the onset of disease (MDS and AML) are highly heterogeneous, represented by molecules with one to eight methylated CpG sites, all with similar frequency. After one year, in line with the demethylating action of the therapy, molecules with high methylation level (5-AzaC sensitive molecules) disappear, while those ones hypo- and, in particular unmethylated, seems to be resistant (5-AzaC resistant molecules). In the last two stages of the disease, corresponding to the relapse and death, heterogeneous methylated molecules re- appear, and each class is homogeneously distributed. There is a dramatic decrease of unmethylated molecules and

the occurrence of molecules with higher methylation level (containing from 7 to 11 methylated CpG sites), mainly in ToG. Some of them could be generated *de novo* or could be derived by amplification of the few 5-AzadC resistant molecules. So, we have the loss and/or the gain of methylation classes during the progression of the disease.

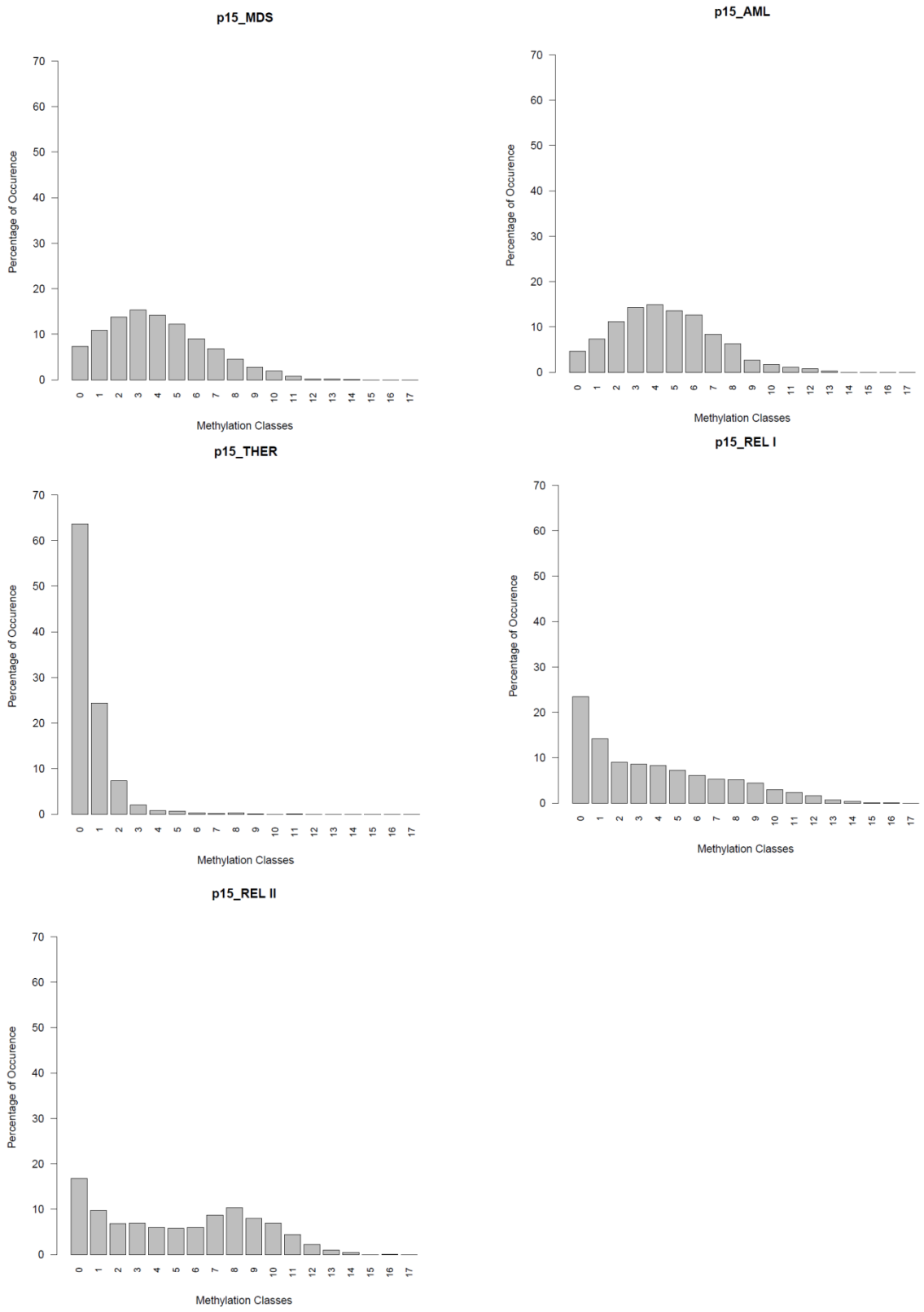

Figure 4.3. Frequency of each methylation class during disease progression for p15 gene. On x-axis the methylation classes are reported, while on the y- axis the frequency of each methylation classes is reported. Each bar represents a methylation class.

4.2.3 Intra-individual epiallelic diversity

This analysis determines the epiallelic configuration of methylated CpGs in single DNA molecules. To quantify the epiallelic diversity (or clonal diversity) inside each sample, it has been adapted diversity measures borrowed from ecology [99]. Indeed, cancers can be viewed from an ecological perspective that focuses on interactions of organisms (in this context, the organisms are the cell clones) with their environment (in this context, the environment is the tissue) and among each other [119,140,141]. When applying an ecological perspective to human cancers, each sample is not a single organism, but a micro- environment consisting of thousands of species (clonal populations of tumor cells), that in this context are represented by the specific epialleles. Each epiallele represents a single allele of a cell.

A simple measure is the number of epialleles (clones) in the sample, expressed as species richness. Ecological measures of diversity typically integrate both number and abundance of epialleles (or clones): one of these measures is the Shannon diversity index (H), which has been used in the following analysis.

The estimated intra- individual epiallelic diversity for p14 gene is reported in the Table1 and Fig. 4.4.

First of all, the number of epialleles (expressed as species richness, S) has been estimated for each sample- stage (Table 1 and Fig. 4.4a). With the same number of reads, the initial stages of disease (MDS and AML) show the higher number of epialleles (856 and 903, respectively), while during therapy (THER) and after (REL I and REL II) the species richness dramatically decreases (377 for THER, 427 for REL I and 368 for REL II).

However, this parameter (the number of epialleles) is not sufficient to describe the intra-individual epiallelic diversity because it does not take into account for the proportion and distribution of each species (epiallele) within the samples. Thus, Shannon diversity index (H) has been used to assess the composition of the epiallelic repertoire: it is high at the onset of disease (MDS and AML), reaching a value of about 5.5 and decreases from the therapy to last stage, reaching a value of about 3. This further confirmed what it has been observed in the methylation classes analysis: before therapy there is a higher level of clonal (or epiallelic) heterogeneity than after therapy. As regard the distribution of these epialleles, high value of H (near to H_{max}) indicates the presence of diverse and equally

distributed species. At the onset of disease (MDS and AML), the Shannon index is closer to its expected value (H_{max}), suggesting that they are equally distributed. With the progression of disease (THER, REL I and REL II) the observed Shannon index (3.4, 3.7 and 3.6, respectively) becomes almost the half of its expected value (H_{max}=about 6), suggesting that the various epialleles are not homogenously distributed, but dominant species (epialleles) are selected.

Taken together, these results suggest that in the case of p14 gene, before therapy there is a higher level of clonal heterogeneity than after therapy (high Shannon index in MDS and AML). Because of the Shannon diversity index incorporates a combination of richness and evenness, the increase of the diversity index cannot be attributed to changes in the distribution itself, but instead to an increase of the total epialleles repertoire itself. The occurrence of different clonal families indicates that at the onset (MDS and AML) the disease is highly polyclonal. On the other side, therapy with 5-AzaC (Vidaza) considerably reduces all the clonal families, by eradicating clones with higher methylation level (5-AzaC sensitive clones). The last stages of disease (REL I and REL II), corresponding to the relapse and depth, show a clonal heterogeneity similar to that found after (THER), suggesting a strong positive selection of few clones resistant or selected by therapy. These clones could be generated *de novo* (from REL I to REL II) or could be derived by amplification of the few 5-AzaC resistant molecules. In conclusion, in the last stages, the disease is epigenetically oligoclonal as far as p14 epiallelic distribution is concerned.

Table 1. Intra- individual diversity summary statistics for p14 gene in different stages of the disease. For each stage the number of reads (N), the species richness (S), the Shannon Diversity Index (H) and the expected Shannon Diversity Index (H_{max}=logS) are reported.

Samples	N	S	H	H_{max}
MDS	2494	856	5.46	6.75
AML	2494	903	5.69	6.80
THER	2494	377	3.40	5.93
REL I	2494	427	3.70	6.06
REL II	2494	368	3.62	5.91

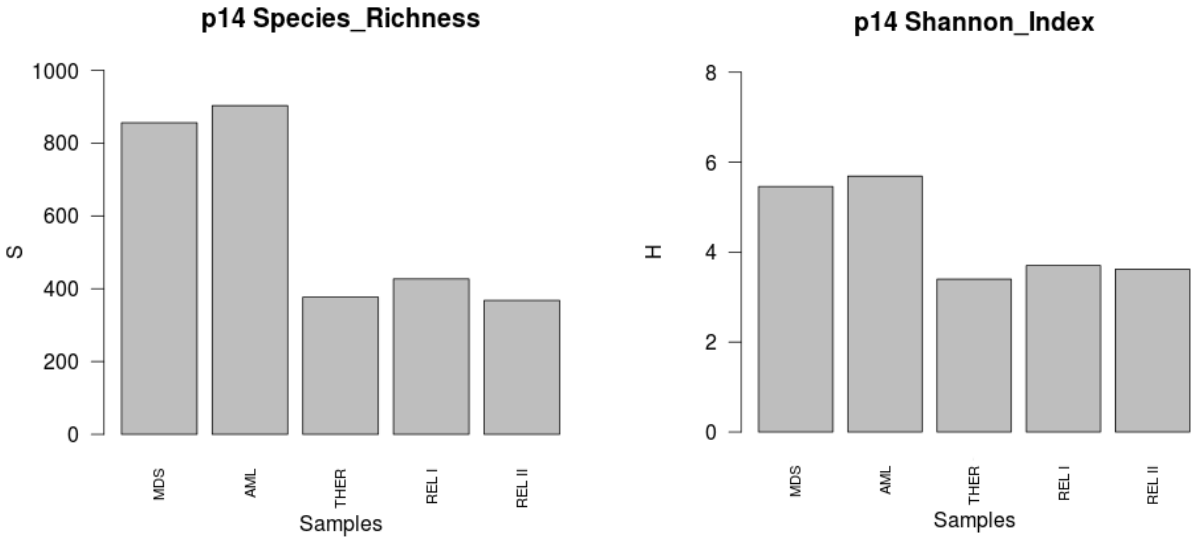

Fig. 4.4. Graphic representation of the summary statistics reported in the Table 1. On the x-axis the different stage of the disease are reported; on the y- axis the species richness, S (a) and the Shannon diversity Index, H (b) is reported.

The estimated intra- individual epiallelic diversity for p15 gene is reported in the Table2 and Fig. 4.5.

Table 2. Intra- individual diversity summary statistics for p15 gene in different stages of the disease. For each stage the number of reads (N), the species richness (S), the Shannon Diversity Index (H) and the expected Shannon Diversity Index (H_{max}=logS) are reported.

Samples	N	S	H	H_{max}
MDS	6392	2496	6.9	7.8
AML	6392	2938	7.3	7.9
THER	6392	338	2.3	5.8
REL I	6392	2077	5.8	7.6
REL II	6392	2024	6.2	7.6

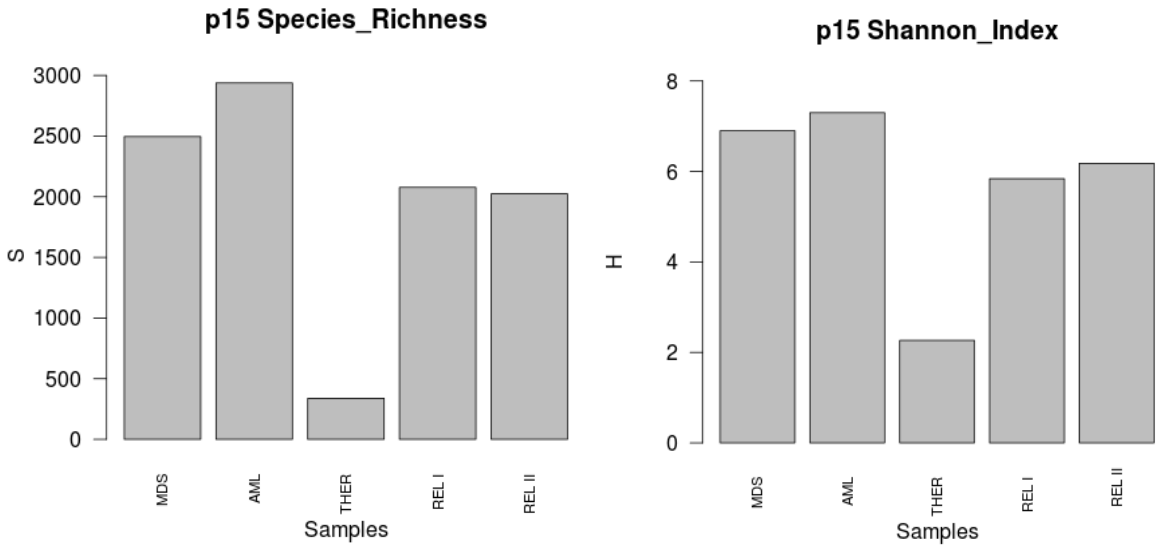

Fig. 4.5. Graphic representation of the summary statistics reported in the Table 2. On the x-axis the different stage of the disease are reported; on the y- axis the species richness, S (a) and the Shannon diversity Index, H (b) is reported.

First of all, the number of epialleles (expressed as species richness, S) has been estimated for each stage (Table 2 and Fig. 4.5a). With the same number of reads, the initial stages of disease (MDS and AML) are associated with the higher number of epialleles (2496 and

2938, respectively). With the therapy (THER) a lot of them disappear and the species richness dramatically decreases (338). In the last stages of disease (REL I and REL II), there is again an increase of the number of epialleles (2077 for REL I and 2024 for REL II).

However, this parameter (the number of epialleles) is not sufficient to describe the intra-individual epiallelic diversity because it does not take into account the proportion and distribution of each species (epiallele) within the samples. Thus, Shannon diversity index (H) has been used to assess the composition of the epiallelic repertoire. The diversity index is higher at the onset of disease (MDS and AML), reaching a value of about 7, decreases with the therapy (THER, 2.27) and then increases again in REL I and REL II (about 6). This further confirms what it has been observed in the methylation classes analysis: at the initial stages of disease there is a higher level of clonal (or epiallelic) heterogeneity, compared to therapy. Therapy reduces considerably the epigenetic heterogeneity of p15 epialleles, which increases again in the last stages (REL I and REL II), reaching levels comparable to the initial stages. As far as the distribution of these epialleles, high value of H (near to H_{max}) indicates the presence of diverse and equally distributed species. At the onset of disease (MDS and AML), the Shannon index is closer to its expected value (H_{max}), suggesting that the epialleles are equally distributed. After one year of therapy (THER), the observed Shannon index (2.2) decreases significantly and is almost the half of its expected value (H_{max}=about 6), suggesting that the various epialleles are not homogenously distributed, but one or more than one dominant species (epialleles) are present. In the last stages of disease (REL I and REL II) the observed Shannon index (5.8 and 6.1, respectively) are almost close to the expected value (H_{max}=about 7.6), suggesting that the various epialleles are equally distributed.

Taken together, these results suggest that in the case of p15 gene, at the initial stages of disease there is a higher level of clonal (or epiallelic) heterogeneity (high Shannon index in MDS and AML). Because of the Shannon diversity index incorporates a combination of richness and evenness, the increases in the diversity index cannot be due to changes in the distribution itself, but instead to an increase of the total epialleles repertoire itself. The occurrence of different clonal families indicates that at the onset (MDS and AML), the disease is highly polyclonal, as far as p15 epiallelic distribution is concerned. On the other side, therapy with 5-AzaC (Vidaza) considerably reduces all the most of the families, by eradicating a lot of clones (5-AzaC sensitive clones). The last stages of the disease (REL I

and REL II), corresponding to the relapse and death, are composed by different clonal families, indicating that the disease at these stages is highly polyclonal, due or to amplication of older variants (5-AzaC resistant clones) or generation of new clones.

4.2.4 Inter- individual epiallelic diversity

The degree of epigenetic diversity across samples was evaluated using the concept of epigenetic distance [101] based on Euclidean distance for p14 (Table 3) and p15 (Table 4) genes. The larger the distance, the more dissimilar the two samples' methylation profiles are to each other.

Euclidean distance of pairwise comparisons among samples has been represented by Principal Coordinate Analysis (PCoA) (Figs. 4.6 and 4.7). Both for p14 and p15 genes, the PCoA plots show a clustering of the tumor samples before therapy on one side (MDS and AML) and after therapy (REL I and REL II) on the other side. This means that inside each one of two cluster, samples have a small Euclidean distance in terms of epialleles composition. Indeed, in the case of p14 gene this distance is estimated 21.7 for MDS/AML and 10.6 for REL I/ REL II, while in the case of p15 it is estimated 41.2 for MDS/AML and 54.4 for REL I/ REL II. For both genes, THER is separated from these two clusters, even if in the case of p14 gene THER stage is closer to the REL I/ REL II cluster than MDS/AML.

Table 3. Euclidean distance matrix (tables) for p14 gene.

	MDS	AML	THER	REL I
AML	21.7			
THER	56.7	68.0		
REL I	40.5	52.5	21.0	
REL II	34.4	47.1	29.5	10.6

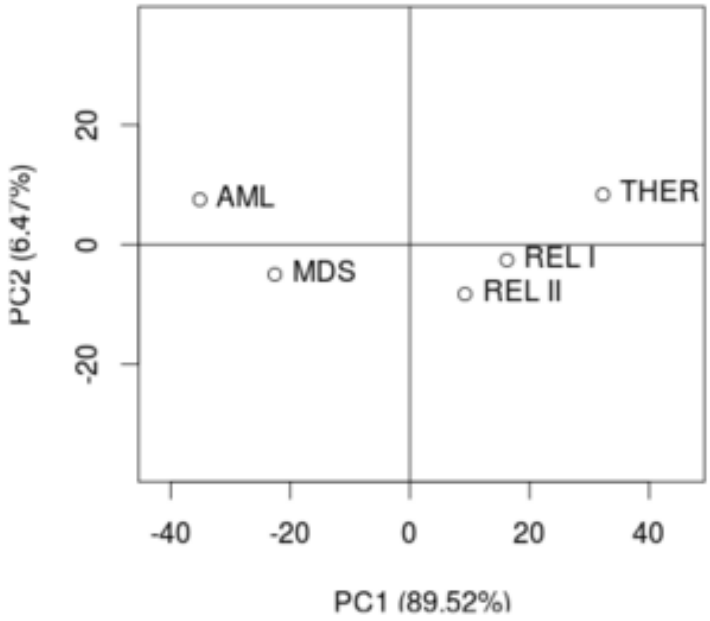

Figure 4.7. Principal coordinate analysis among samples computed by the Euclidean distance matrix of the Table 3 for p14 gene. Samples that are ordinated closer together have smaller dissimilarity values (smaller Euclidean distance in species composition)than those ordinated further apart.

Table 4. Euclidean distance matrix (tables) for p15 gene.

	MDS	AML	THER	REL I
AML	41.3			
THER	84.0	102.3		
REL I	45.1	58.7	78.2	
REL II	84.3	91.8	123.5	54.4

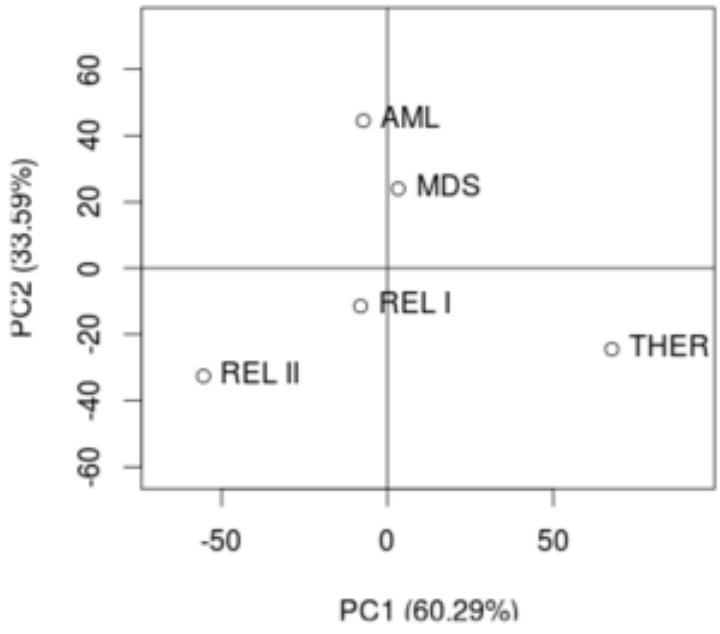

Figure 4.8. Principal coordinate analysis among samples computed by the Euclidean distance matrix of the Table 3 for p15 gene. Samples that are ordinated closer together have smaller dissimilarity values (smaller Euclidean distance in species composition) than those ordinated further apart.

4.3 Discussion

Currently, methylation analysis, both genome-wide or of specific DNA segments, gives information on the percentage of methylation of single CpGs in a given sequence. It is impossible to deduce from these data the methylation status of all the CpGs belonging to the same molecule. Moreover, quantitative analysis is not able to describe the complex epigenetic structure of a cell population, unless the methylation status is quite uniform in the cell population under study. On the contrary, the qualitative one helps to better

understand changes of methylation state inside a heterogeneous cell population, such as a tissue, or even better a tumor tissue.

In the present study, it has been shown the first deep and extensive qualitative analysis of methylated alleles of two suppressor genes CDKN2A (p14ARF) and CDKN2B (p15) during different stages of acute myeloid leukemia (progression, demethylating therapy 5-azaC, remission and relapse) of a patient. Combining quantitative (methylation percentage for sample) with qualitative (configuration of methylated CpGs in single DNA molecules), it has been possible to define the epigenetic landscape of different stages of disease. The qualitative analysis was performed at methylation classes and at single molecule (epialleles) level. The information obtained by these two different levels can be integrated in order to gain insight about methylation heterogeneity degree and distribution and about tumor clonality in different stages of disease. This approach provides a powerful method to track tumor evolution and to analyse the source of heterogeneity of tumor cells. The relevant results are discussed below.

4.3.1 High degree of methylation heterogeneity at the onset of disease for p14 and p15 genes

The high level of heterogeneity before demethylating therapy (5-AzaC, Vidaza) for p14 and p15 genes has given by molecules (epialleles or clones) belonging to different methylation classes. In other words, there is not a dominant methylation classes or a dominant molecules (epialleles or clones), but a lot of molecules, belonging to different methylation classes, equally distributed. The occurrence of different clonal families indicates that the disease at this stage is highly polyclonal. This does not exclude the presence of possible cell clones carrying founder epi-mutations, but maybe the presence of other cell clones carrying the passenger epi- mutations could help to create a micro-environment that favour the growth and the proliferation of cells carrying founder epi-mutations, conferring them a selective advantage.

4.3.2 Low degree of methylation heterogeneity during demethylating therapy (5-AzaC, Vidaza) for p14 and p15 genes

Therapy with 5-AzaC drug eliminates and/or drastically reduces molecules with high methylation level (5-AzaC sensitive molecules), while those ones with low methylation level seems to be resistant (5-AzaC resistant molecules). Thus, therapy with 5-AzaC dramatically reduces the most of the clonal families. This suggests a strong positive selection of few clones carrying alleles with a low number of methylated CpG sites and maybe they have acquired a higher fitness. Besides, selection is context- specific. As a consequence, some of the epi- mutations that are selectively advantageous at certain stages of tumor progression may not be present in the other stages.

Indeed, the landscape of tissue ecosystems of cancer can be radically altered by anti-cancer therapy, in this case by 5-AzaC drug. Therapeutic intervention may decimate cancer clones and erode their habitats, but at the same time it may provide novel selective pressures, as well as new resources and opportunities, for the expansion of those therapy-insensitive cancer cells [142], causing the eventual relapse of the disease [143-145]. In this case, the pre-existence of resistant clones within a tumor can make the difference between tumor extinction (treatment success) and tumor evolutionary adaptation (treatment failure).

4.3.3 Different behaviour after demethylating therapy for p14 and p15 genes

The two onco- suppressors show different behaviour in the last stages of diseases, corresponding to the relapse and death.

In the case of p14 gene, the low clonal heterogeneity, obtained with demethylating therapy, is conserved. This suggests that for this gene at these stages the disease is oligoclonal. Some of them could be generated *de novo* (from REL I to REL II) or could be derived by amplification of the few 5-AzaC resistant molecules. However, PCoA analysis seems to be in favour of the second possibility, given the presence of a cluster composed by THER, REL I and REL II.

In the case of p15 gene, in the last stages of the disease there is a re- occurrence of a high clonal heterogeneity, derived from molecules belonging to different methylation classes, each one of them that is homogeneously distributed. This suggests that at these stages the disease is highly polyclonal composed by different clonal families, derived from either

amplication of older variants (5-AzaC resistant clones) or generation of new variants. However, PCoA analysis seems to be in favour of the second possibility, given that THER is distant from the cluster REL I/ REL II.

The different behaviour of these two genes can find different explanations:

- The tumor microenvironment generated after therapy could exercise an higher selective pressure for cell clones of p14 than those one of p15.
- Stochastic processes may produce the high rate of methylation polymorphism.
- Deterministic events may contribute to the gain or loss of methylation at specific loci

4.4 Materials and Methods

4.4.1 Deep Bisulfite Amplicon Sequencing (Deep- Bis)

The methylation profiles of the region near the TSS of CDKN2A (p14ARF) and CDKN2B (p15) suppressor genes were analysed by deep sequencing based on the locus- specific amplification of bisulfite- treated genomic DNA derived from bone marrow (BM) and peripheral blood cells of a patient (aged 60) during different stages of AML. This patient showed an initial disease progression from myelodysplasia (MDS) to acute myeloid leukemia (AML), with an increase of BM blasts (MDS=12% blasts and AML=35% blasts). He was then treated with demethylating therapy azacitidine (5-AzaC; Vidaza), with haematological remission achievement after 6 months (THER=2% blasts). Unfortunately, after 8 months of remission, the patient underwent to relapse (REL I=22% blasts and REL II=80%blasts) resulted rapidly fatal.

Methylation status was determined by amplicon-based bisulfite sequencing using Illumina MiSeq. The following bisulfite-specific primers were used to obtain tiled amplicons: p14 PR1 fw 5'-ggtgYgtgggtTTTagtTtgTa-3' (nt position from -33 to -12) - and p14 PR1 rev 5'- AaaAcctccaccRAcRAtta-3' (nt +379 to +398) (amplicon size 431 bp) and p15 PR1 fw 5'- attaggagTtgagggTagtgg-3' (nt position from -275 to -296) - and p15 PR1 rev 5'- AacRcAccRaActcaaaAcc-3' (nt +71 to +90) (amplicon size 365 bp). The capital letters in the primers sequences indicate the original C or G, respectively. CpGs including in the primers have been excluded from the further analysis.

Reads in FASTQ format obtained from sequencing were processed with Paired-End reAd mergeR (PEAR; http://sco.h-its.org/exelixis/web/software/pear/) for paired end assembling and initial quality filtering. Only those reads with the following features were retained:

- a mean quality score (Phred) greater than 30;
- a read length between 400 and 500 nucleotides;
- an overlapping region within paired-end reads of at least 40 nucleotides.

Resulting reads were then converted in FASTA format using PReprocessing and INformation of SEQuence (Prinseq; http://prinseq.sourceforge.net/).

Reads were then aligned to the corresponding bisulfite converted reference sequence using AmpliMethProfiler (see Chapter 2).

By applying a series of quality filters on the read length and the alignment quality, it has been retained only those reads characterized by:

- length ±50% compared to the reference length;
- primer of the corresponding gene identified with at least 80% of similarity;
- at least 98% of bisulfite efficiency, calculate as percentage of conversion of non-CpG cytosines into thymines over total number of C in a context not CpG;
- to be aligned for at least 60% of their bases with the reference sequence
- C status at the all CpG positions recognised as methylated (1) or unmethylated (0). Reads with ambiguous calls (presence of gaps or A or G) at the CpG dinucleotide were removed.

Methylation state was estimated by observing base calls (T/C) at CpG sites in the mapped reads.

4.4.2 Rarefaction analysis

For both genes, a highly different number of amplicon reads (depth) for each sample was obtained. The sequence depth is only related to the number of sequences for sample obtained after a next generation sequencing experiment. In order to standardize the data obtained for each sample with different sequencing counts and to avoid bias in the successive analysis due to a different sampling, a rarefaction step was performed. A random subsampling of reads, corresponding to the minimum number of sequences belonging to a sample within the dataset, was taken for each sample. Thus, the successive

analysis have been performed on 2494 amplicon reads for sample for p14 gene and on 6932 amplicon reads for sample for p15 gene.

4.4.3 Quantitative methylation analysis of p14 and p15 genes

For each sample, the global methylation degree was calculated by averaging CpGs methylation percentages of all CpG sites in the sample.

4.4.4 Methylation classes frequency for p14 and p15 genes

For each sample, the frequency of each methylation class was calculated as a ratio of the number of reads, belonging to each methylation class, over the total number of reads found in each sample.

4.4.5 Intra- individual diversity

In order to evaluate the intra- individual diversity of the samples, diversity measures such as richness (number of unique epialleles) and the Shannon diversity index [100] were calculated using "vegan" package of R statistics environment (see Chapter Data Analysis for more details). The larger the Shannon diversity index, the more diverse the distribution of the epialleles.

4.4.6 Inter- individual diversity

The degree of epigenetic similarity was measured by Euclidean distance and visualized through Principal coordinates analysis (PCoA) analysis (see Chapter Data Analysis for more details).

All statistical analyses were performed using R statistical package ver. 3.2.1 (http://www.R-project.org).

Chapter 5

Discussion

Most of the studies on DNA methylation, regardless of the techniques employed, uses the conventional quantitative approach, namely they take in consideration the average methylation level, summarizing data into average percentage of methylated CpGs in specific genomic regions, or the methylation percentage for single CpG site, or looking at CpGs genome-wide distribution with an only relatively high resolution [85-91]. Such kind of approach gives important information on the general relationships between methylation and expression but might be not able to describe the complex epigenetic structure and dynamics within a cell population. Indeed, quantitative approaches assume that methylation status is quite uniform in the population of cells under study. As a consequence, this approach obscures important positional information encoded within the epiallelic DNA methylation patterns.

In the present work, in order to better decode epigenetic data, a new way to analyse DNA methylation, based on qualitative approach, was developed. Qualitative analysis of methylation profiles is an innovative way to look at methylation of a genomic region. The qualitative approach, specifically looking at the individual methylation conformation of single molecules, provides an added value to the quantitative one. This qualitative approach is useful to dissect the clonal composition at epigenetic level and to recognise different methylation profiles inside an heterogeneous cell population (i.e., tissues) for a given genomic locus and to evaluate the stochastic and /or deterministic components.

True representation of multiple methylation patterns can only be fully characterised by clonal analysis. This implies to restrict the genomic space under investigation and to increase the coverage of a specific region by at least 1000 folds in order to obtain an effective statistical representation. Indeed, the qualitative approach takes advantage of the Deep Bisulfite Amplicon Sequencing (Deep- Bis), which allows to reach a very high coverage (about 200.000-300.000 reads/sample) of selected loci, overcoming in this way the limitations linked to the low coverage of the genome- wide technique. Working with DNA molecules with the same 5' and 3' ends and sequencing for many thousands folds the same locus to get a deep coverage of the site, it is possible to evaluate the methylation of hundreds molecules at time at single nucleotide level and derive the configuration of each C in the sequence relative to the other Cs. Deep sequencing provides the ability to investigate clonal methylation patterns with an unprecedented resolution level, enabling the proper characterization of the heterogeneity of methylation patterns, and allow to infer cell population characteristics accurately. In this way, observed changes in average methylation levels can then be interpreted according to epiallelic diversity, discerning, for example, a regulated increase in the frequency of a specific epiallele from multiple stochastic changes in the frequencies of many epialleles [94].

As the number of sequences increases, the ability to analyse this type of data then becomes a significant challenge. Moreover, currently available tools for methylation analysis lack output formats that explicitly report CpG methylation profiles at the single molecule level. In the present study, it has been developed AmpliMethProfiler, a python-based pipeline for the extraction at the single molecule level of CpG methylation profiles of amplicons from Deep- Bis of multiple DNA regions. AmpliMethProfiler processes FASTA files and uses BLASTN to align in a fast and reliable way input reads to a bisulfite converted reference sequence. The output reports the methylation status of each CpG site in a read in binary code (0 if the site is unmethylated, 1 if the site is methylated) and summarises DNA methylation according to epiallelic methylation patterns and can be readily used for the downstream quantitative and qualitative analysis.

The qualitative approach provides a new analytical framework to analyse methylation and allows the use of standard population genetics methods. Indeed, this way of analysing methylation data is accompanied by the possibility to apply notions and techniques derived from the population genetics and ecology fields. Metrics, statistical methods and tools to

analyse population structures in term of species composition, species richness, difference in the population composition and structure among samples can be easily imported and adapted for the analysis of methylation profiles from Deep- Bis. Specifically, the epialleles can are subdivided in several classes, according their content of methylated CpGs (mono, bi and tri- or more-methylated molecules). Then, the mixture of the molecules (epialleles) derived from the sequencing can be treated as a population of haploid organisms, biological samples as the micro- environement in which these organisms are harboured and differences among samples can be treated as differences among populations. Thus, each biological sample can be seen as a community, each epiallele can be seen as a specie, while each read can be seen as an individual. By this way, it is possible to describe the general methylation landscape of the various samples.

The qualitative approach is highly versatile and can be easily adaptable to different contexts and biological systems. In this work, it is applied on two experimental models: mouse development and AML progression before and after the demethylating therapy, in order to describe the methylation and demethylation dynamics, to investigate the epialleles distribution, to follow their evolution and to gain insight on epigenetic heterogeneity degree at specific loci. In both cases, it is clear that the qualitative approach has a greater informative content than the quantitative one, because it allows to dissect the epigenetic complexity of a sample. The most relevant results for these two biological systems are the following.

During mouse development, epialleles are generated in a perfectly conserved fashion and the frequency of each epiallele is determined in a well- orchestrated fashion in the somatic tissues. Indeed, the highly conserved methylation profiles trend indicates that probably the CpG sites methylation (and thus, the generation of epialleles) is not a stochastic event, rather a deterministic one, developmentally regulated, leading to an orchestrated distribution of epialleles among the entire population of cells. This deterministically regulated distribution of different epialleles evokes the possible existence of a novel combinatorial code of CpG methylation. Moreover, the qualitative approach allowed to describe in detail the molecules created during the methylation and demethylation processes, respectively.

Using leukemia samples at different stages of disease, it is possible to gain insights about methylation heterogeneity degree and tumor clonality during the different phases of disease progression. During different stages of AML, the distribution of polymorphic

methylated molecules (epialleles) changes over time. In particular, at the onset of disease there is a high degree of methylation heterogeneity. Then, cells carrying particular epialleles undergo selection during demethylating treatment, leading to a decrease in epigenetic heterogeneity. Finally, the high degree of methylation heterogeneity re- occurs at some genomic sites. Thus, the epiallele dynamics at different stages may indicate important genomic regions involved in these biological processes and may be used as an estimate of the evolutionary distance between stages of diseases or development, thus expanding the knowledge of epigenetic heterogeneity and how epialleles can change within a patient, over time, across the genome.

In general, the qualitative approach could be a novel, rapid mean by which to detect and trace genomic areas with shifts in their cells' epigenetic states and can be used to define epiallelic clonality, tumor evolution, and epigenome dynamics. This approach allows to analyse and to follow the genesis, the variability and the evolution of the epialleles in specific genomic regions and in different biological systems. Furthermore, it could help to better understand the mechanisms underlying the changes of methylation status inside of single cells during methylation and demethylation processes. This approach also allows to evaluate the possible stochastic and/or deterministic components during methylation and demethylation phenomena, because it accounts for the high polymorphism arisen from the mixture of epialleles with variable frequencies in cells populations derived from individual somatic tissues. Indeed, with this approach, coupled with the increase of the coverage of a specific genomic region, observed changes in average methylation levels can be interpreted according to epiallelic diversity (epipolymorphism), discerning, for example, a regulated increase in the frequency of a specific epiallele from multiple stochastic changes in the frequencies of many epialleles. Indeed, the stochastic processes can lead to an methylation degree polymorphism, while the deterministic ones can determine the loss or the gain of methylation at specific loci.

In conclusion, the method developed in this study can provide an added value to the traditional methylation analyses and can greatly extend the capacity to dissect the epigenetic heterogeneity in a cell population. The detection and the tracking of methylation patterns is an important step for unlocking the biological meaning of epigenetic heterogeneity. Moreover, the tracking of the methylation profiles is more faithful to the

epigenetic state of different loci and allows a more detailed overview of the methylation landscape in a tissue, which is composed by a mosaics of epigenetically different cells.

References

1. Allis C, Jenuwein T, Reinberg D. *Epigenetics*. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press; 2007.

2. Suzuki M, Bird A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet*. 2008;9(6):465-476.

3. Law J, Jacobsen S. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet*. 2010;11(3):204-220.

4. Khavari D, Sen G, Rinn J. DNA methylation and epigenetic control of cellular differentiation. *Cell Cycle*. 2010;9(19):3880-3883.

5. Smith Z, Meissner A. DNA methylation: roles in mammalian development. *Nat Rev Genet*. 2013;14(3):204-220.

6. Robertson K. DNA methylation and human disease. *Nat Rev Genet*. 2005;6(8):597-610.

7. Fraga M, Esteller M. Epigenetics and aging: the targets and the marks. *Trends in Genetics*. 2007;23(8):413-418.

8. Sharp A, Stathaki E, Migliavacca E et al. DNA methylation profiles of human active and inactive X chromosomes. *Genome Research*. 2011;21(10):1592-1600.

9. Li E, Beard C, Jaenisch R. Role for DNA methylation in genomic imprinting. *Trends in Genetics*. 1994;10(3):78.

10. Slotkin R, Martienssen R. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet*. 2007;8(4):272-285.

11. Rizwana R, Hahn PJ. CpG methylation reduces genomic instability. *J Cell Sci*. 1999; 112(Pt24): 4513-4519.

12. Cokus S, Feng S, Zhang X et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*. 2008;452(7184):215-219.

13. Rountree M, Selker E. DNA methylation inhibits elongation but not initiation of transcription in Neurospora crassa. *Genes & Development*. 1997;11(18):2383-2395.

14. Lister R, Pelizzola M, Dowen R et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*. 2009;462(7271):315-322.

15. Laurent L, Wong E, Li G et al. Dynamic changes in the human methylome during differentiation. *Genome Research*. 2010;20(3):320-331.

16. Hackett J, Surani M. DNA methylation dynamics during the mammalian life cycle. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2012;368(1609):20110328-20110328.

17. Feng S, Jacobsen S, Reik W. Epigenetic reprogramming in plant and animal development. *Science*. 2010;330(6004):622-627.

18. Fulka H, Mrazek M, Tepla O, Fulka J. DNA methylation pattern in human zygotes and developing embryos. *Reproduction*. 2004;128(6):703-708.

19. Rakyan V, Blewitt M, Druker R, Preis J, Whitelaw E. Metastable epialleles in mammals. *Trends in Genetics*. 2002;18(7):348-351.

20. Li E, Bestor T, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*. 1992;69(6):915-926.

21. Okano M, Bell D, Haber D, Li E. DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. *Cell*. 1999;99(3):247-257.

22. Jurkowska R, Jurkowski T, Jeltsch A. Structure and Function of Mammalian DNA Methyltransferases. *ChemBioChem*. 2010;12(2):206-222.

23. Hata K, Okano M, Lei H, Li E. Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development*. 2002;129(8):1983-1993.

24. Bourc'his D. Dnmt3L and the Establishment of Maternal Genomic Imprints. *Science*. 2001;294(5551):2536-2539.

25. Bestor T. The DNA methyltransferases of mammals. *Human Molecular Genetics*. 2000;9(16):2395-2402.

26. Branco M, Ficz G, Reik W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nat Rev Genet*. 2011; ;13(1):7-13.

27. Bird A, Taggart M, Frommer M, Miller O, Macleod D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*. 1985;40(1):91-99.

28. Ehrlich M, Gama-Sosa M, Huang L et al. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucl Acids Res*. 1982;10(8):2709-2721.

29. Larsen F, Gundersen G, Lopez R, Prydz H. CpG islands as gene markers in the human genome. *Genomics*. 1992;13(4):1095-1107.

30. Saxonov S, Berg P, Brutlag D. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*. 2006;103(5):1412-1417.

31. Song F, Smith J, Kimura M et al. Association of tissue-specific differentially

methylated regions (TDMs) with differential gene expression. *Proceedings of the National Academy of Sciences*. 2005;102(9):3336-3341.

32. Li E. Chromatin modification and epigenetic reprogramming in mammalian development. *Nat Rev Genet*. 2002;3(9):662-673.

33. De Smet C, Lurquin C, Lethe B, Martelange V, Boon T. DNA methylation is the primary silencing mechanism for a set of germ line- and tumor-specific genes with a CpG-rich promoter. *Molecular and Cellular Biology*. 1999;19(11):7327-7335.

34. Mohn F, Weber M, Rebhan M et al. Lineage-specific polycomb targets and de novo dna methylation define restriction and potential of neuronal progenitors. *Molecular Cell*. 2008;30(6):755-766.

35. Fernández L, Torres M, Real F. Somatic mosaicism: on the road to cancer. *Nature Reviews Cancer*. 2015;16(1):43-55.

36. Brena R, Huang T, Plass C. Toward a human epigenome. *Nature Genetics*. 2006;38(12):1359-1360.

37. Liu Y, Aryee M, Padyukov L et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol*. 2013;31(2):142-147.

38. Adalsteinsson B, Gudnason H, Aspelund T et al. Heterogeneity in white blood cells has potential to confound dna methylation measurements. *PLoS ONE*. 2012;7(10):e46705.

39. Landan G, Cohen N, Mukamel Z et al. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nature Genetics*. 2012;44(11):1207-1214.

40. Xie H, Wang M, de Andrade A et al. Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic Acids Research*. 2013;41(14):7184-7184.

41. Ahuja N, Li Q, Mohan AL, Baylin SB, Issa JP. Aging and DNA methylation in colorectal mucosa and cancer. *Cancer Res*. 1998;58(23):5489-5494.

42. Issa J, Ottaviano Y, Celano P, Hamilton S, Davidson N, Baylin S. Methylation of the oestrogen receptor CpG island links ageing and neoplasia in human colon. *Nature Genetics*. 1994;7(4):536-540.

43. Nakagawa H, Nuovo GJ, Zervos EE, Martin EW Jr, Salovaara R, Aaltonen LA, de la Chapelle A. Age-related hypermethylation of the 5' region of MLH1 in normal colonic mucosa is associated with microsatellite-unstable colorectal cancer development. *Cancer Res*. 2001;61(19):6991-6995.

44. Pasquali L, Bedeir A, Ringquist S, Styche A, Bhargava R, Trucco G. Quantification of CpG island methylation in progressive breast lesions from normal to invasive carcinoma. *Cancer Letters*. 2007;257(1):136-144.

45. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer?. *Nature Reviews Cancer*. 2012;12(5):323-334.

46. Ramirez J, Lukin K, Hagman J. From hematopoietic progenitors to B cells: mechanisms of lineage restriction and commitment. *Current Opinion in Immunology*. 2010;22(2):177-184.

47. Reik W, Walter J. Genomic imprinting: parental influence on the genome. *Nat Rev Genet*. 2001;2(1):21-32.

48. John R, Lefebvre L. Developmental regulation of somatic imprints. *Differentiation*. 2011;81(5):270-280.

49. Shaknovich R, De S, Michor F. Epigenetic diversity in hematopoietic neoplasms. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*. 2014;1846(2):477-484.

50. Li Y, Zhu J, Tian G et al. The DNA Methylome of human peripheral blood mononuclear cells. *PLoS Biology*. 2010;8(11):e1000533.

51. Lister R, Pelizzola M, Kida Y et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*. 2011;471(7336):68-73.

52. Horsthemke B. Epimutation in human disease. *Curr. Top. Microbiol. Immunol*. 2006; 310, 45–59.

53. Feinberg A. Phenotypic plasticity and the epigenetics of human disease. *Nature*. 2007;447(7143):433-440.

54. Foley D, Craig J, Morley R et al. Prospects for Epigenetic Epidemiology. *American Journal of Epidemiology*. 2008;169(4):389-400.

55. Flanagan J, Popendikyte V, Pozdniakovaite N et al. Intra- and Interindividual Epigenetic Variation in Human Germ Cells. *The American Journal of Human Genetics*. 2006;79(1):67-84.

56. Schneider E, Pliushch G, El Hajj N et al. Spatial, temporal and interindividual epigenetic variation of functionally important DNA methylation patterns. *Nucleic Acids Research*. 2010;38(12):3880-3890.

57. Siegmund K, Connor C, Campan M et al. DNA Methylation in the human cerebral cortex is dynamically regulated throughout the life span and involves differentiated neurons. *PLoS ONE*. 2007;2(9):e895.

58. Bjornsson H. Intra-individual change over time in dna methylation with familial clustering. *JAMA*. 2008;299(24):2877.

59. Bollati V, Schwartz J, Wright R et al. Decline in genomic DNA methylation through aging in a cohort of elderly subjects. *Mechanisms of Ageing and Development*. 2009;130(4):234-239.

60. Farcas R, Schneider E, Frauenknecht K et al. Differences in dna methylation patterns and expression of the ccrk gene in human and nonhuman primate cortices. *Molecular Biology and Evolution*. 2009;26(6):1379-1389.

61. Zilberman D, Henikoff S. Genome-wide analysis of DNA methylation patterns. *Development*. 2007;134(22):3959-3965.

62. Weber M, Davies J, Wittig D et al. Chromosome-wide and promoter-specific analyses

identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics*. 2005;37(8):853-862.

63. Weber M, Hellmann I, Stadler M et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics*. 2007;39(4):457-466.

64. Cross S, Charlton J, Nan X, Bird A. Purification of CpG islands using a methylated DNA binding column. *Nature Genetics*. 1994;6(3):236-244.

65. Gebhard C. Rapid and sensitive detection of CpG-methylation using methyl-binding (MB)-PCR. *Nucleic Acids Research*. 2006;34(11):e82-e82.

66. Gebhard C. Genome-Wide Profiling of CpG methylation identifies novel targets of aberrant hypermethylation in myeloid leukemia. *Cancer Research*. 2006;66(12):6118-6128.

67. Schmidl C, Klug M, Boeld T et al. Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity. *Genome Research*. 2009;19(7):1165-1174.

68. Rauch T, Pfeifer G. Methylated-CpG island recovery assay: a new technique for the rapid detection of methylated-CpG islands in cancer. *Lab Invest*. 2005;85(9):1172-1180.

69. Rauch TA, Pfeifer GP. The MIRA method for DNA methylation analysis. *Methods Mol Biol*. 2009;507:65-75.

70. Rauch T, Zhong X, Wu X et al. High-resolution mapping of DNA hypermethylation and hypomethylation in lung cancer. *Proceedings of the National Academy of Sciences*. 2007;105(1):252-257.

71. Clark S, Harrison J, Paul C, Frommer M. High sensitivity mapping of methylated cytosines. *Nucleic Acids Res*. 1994; 22(15):2990-2997.

72. Frommer M, McDonald L, Millar D et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences*. 1992;89(5):1827-1831.

73. Grunau C, Clark SJ, Rosenthal A. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Res*. 2001;29(13):E65-5.

74. Genereux D, Johnson W, Burden A, Stoger R, Laird C. Errors in the bisulfite conversion of DNA: modulating inappropriate- and failed-conversion frequencies. *Nucleic Acids Research*. 2009;37(15):5235-5235.

75. Ehrich M, Turner J, Gibbs P et al. Cytosine methylation profiling of cancer cell lines. *Proceedings of the National Academy of Sciences*. 2008;105(12):4844-4849.

76. Laird P. Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet*. 2010;11(3):191.

77. Hansen K, Timp W, Bravo H et al. Increased methylation variation in epigenetic

domains across cancer types. *Nature Genetics*. 2011;43(8):768-775.

78. Berman B, Weisenberger D, Aman J et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nature Genetics*. 2011;44(1):40-46.

79. Adey A, Shendure J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome Research*. 2012;22(6):1139-1143.

80. Gu H, Bock C, Mikkelsen T et al. Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nature Methods*. 2010;7(2):133-136.

81. Bock C. Epigenetic biomarker development. *Epigenomics*. 2009;1(1):99-110.

82. Bergman Y, Cedar H. DNA methylation dynamics in health and disease. *Nat Struct Mol Biol*. 2013;20(10):1236-1236.

83. Shames D, Minna J, Gazdar A. DNA Methylation in health, disease, and cancer. *CMM*. 2007;7(1):85-102.

84. Robertson K, Wolffe A. DNA methylation in health and disease. *Nat Rev Genet*. 2000;1(1):11-19.

85. Ammerpohl O, Haake A, Kolarova J, Siebert R. Quantitative DNA Methylation Profiling in Cancer. *Methods Mol Biol*. 2016;1381:75-92.

86. Hannum G, Guinney J, Zhao L et al. Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Molecular Cell*. 2013;49(2):359-367.

87. Hua D, Hu Y, Wu Y et al. Quantitative methylation analysis of multiple genes using methylation-sensitive restriction enzyme-based quantitative PCR for the detection of hepatocellular carcinoma. *Experimental and Molecular Pathology*. 2011;91(1):455-460.

88. Woodfine K, Huddleston J, Murrell A. Quantitative analysis of DNA methylation at all human imprinted regions reveals preservation of epigenetic stability in adult somatic tissue. *Epigenetics & Chromatin*. 2011;4(1):1.

89. Vaissière T, Hung R, Zaridze D et al. Quantitative analysis of DNA methylation profiles in lung cancer identifies aberrant dna methylation of specific genes and its association with gender and cancer risk factors. *Cancer Research*. 2009;69(1):243-252.

90. Shaw R, Liloglou T, Rogers S et al. Promoter methylation of P16, RARbeta, E-cadherin, cyclin A1 and cytoglobin in oral cancer: quantitative evaluation using pyrosequencing. *Br J Cancer*. 2006;94(4):561-568.

91. Ehrich M, Nelson M, Stanssens P et al. Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. *Proceedings of the National Academy of Sciences*. 2005;102(44):15785-15790.

92. Mikeska T, Candiloro I, Dobrovic A. The implications of heterogeneous DNA methylation for the accurate quantification of methylation. *Epigenomics*. 2010;2(4):561-573.

93. Candiloro I, Mikeska T, Hokland P, Dobrovic A. Rapid analysis of heterogeneously methylated DNA using digital methylation-sensitive high resolution melting: application to the CDKN2B (p15) gene. *Epigenetics & Chromatin*. 2008;1(1):7.

94. Siegmund K, Marjoram P, Tavare S, Shibata D. High DNA methylation pattern intratumoral diversity implies weak selection in many human colorectal cancers. *PLoS ONE*. 2011;6(6):e21657.

95. Dolinoy D, Das R, Weidman J, Jirtle R. Metastable epialleles, imprinting, and the fetal origins of adult diseases. *Pediatr Res*. 2007;61(5 Part 2):30R-37R.

96. Kalisz S, Purugganan MD. Epialleles via DNA methylation: consequences for plant evolution. *Trends Ecol Evol*. 2004;19(6):309-314.

97. Camacho C, Coulouris G, Avagyan V et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10(1):421.

98. Cock PJ, Antao T, Chang JT et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25(11):1422-3.

99. Magurran A. *Measuring Biological Diversity*. Malden, Ma.: Blackwell Pub.; 2004.

100. Shannon CE. *A mathematical theory of communication*. The Bell System Technical Journal; 1948

101. Yatabe Y, Tavare S, Shibata D. Investigating stem cells in human colon by using methylation patterns. *Proceedings of the National Academy of Sciences*. 2001;98(19):10839-10844.

102. Liu H, Liu X, Zhang S et al. Systematic identification and annotation of human methylation marks based on bisulfite sequencing methylomes reveals distinct roles of cell type-specific hypomethylation in the regulation of cell identity genes. *Nucleic Acids Res*. 2015;44(1):75-94.

103. Tanaka S, Nakanishi M, Shiota K. DNA methylation and its role in the trophoblast cell lineage. *Int J Dev Biol*. 2014;58(2-3-4):231-238.

104. Ji H, Ehrlich L, Seita J et al. Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature*. 2010;467(7313):338-342.

105. Brunner A, Johnson D, Kim S et al. Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Research*. 2009;19(6):1044-1056.

106. Meissner A, Mikkelsen T, Gu H et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*. 2008.

107. Song F, Mahmood S, Ghosh S et al. Tissue specific differentially methylated regions (TDMR): Changes in DNA methylation during development. *Genomics*. 2009;93(2):130-139.

108. Liang P, Song F, Ghosh S et al. Genome-wide survey reveals dynamic widespread tissue-specific changes in DNA methylation during development. *BMC Genomics*.

2011;12(1):231.

109. Hirabayashi K, Shiota K, Yagi S. DNA methylation profile dynamics of tissue-dependent and differentially methylated regions during mouse brain development. *BMC Genomics*. 2013;14(1):82.

110. Yagi S, Hirabayashi K, Sato S et al. DNA methylation profile of tissue-dependent and differentially methylated regions (T-DMRs) in mouse promoter regions demonstrating tissue-specific gene expression. *Genome Research*. 2008;18(12):1969-1978.

111. Khulan B, Thompson RF, Ye K et al. Comparative isoschizomer profiling of cytosine methylation: the HELP assay. *Genome Res*. 2006;16(8):1046-55.

112. Rakyan V, Down T, Thorne N et al. An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Research*. 2008;18(9):1518-1529.

113. Still JL, Buell MV, et al. Studies on the cyclophorase system; D-aspartic oxidase. *J Biol Chem*. 1949;179(2):831-837.

114. Van Veldhoven P, Brees C, Mannaerts G. D-Aspartate oxidase, a peroxisomal enzyme in liver of rat and man. *Biochimica et Biophysica Acta (BBA) - General Subjects*. 1991;1073(1):203-208.

115. D'Aniello A, Vetere A, Petrucelli L. Further study on the specificity of d-amino acid oxidase and of d-aspartate oxidase and time course for complete oxidation of D-amino acids. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*. 1993;105(3-4):731-734.

116. Punzo D, Errico F, Cristino L et al. Age-Related Changes in D-Aspartate Oxidase Promoter Methylation Control Extracellular D-Aspartate Levels and Prevent Precocious Cell Death during Brain Aging. *Journal of Neuroscience*. 2016;36(10):3064-3078.

117. Brocks D, Assenov Y, Minner S et al. Intratumor DNA Methylation Heterogeneity Reflects Clonal Evolution in Aggressive Prostate Cancer. *Cell Reports*. 2014;8(3):798-806.

118. Segura-Pacheco B, Perez-Cardenas E, Taja-Chayeb L, Chavez-Blanco A, Revilla-Vazquez A, Benitez-Bribiesca L, Duenas-González A. Global DNA hypermethylation-associated cancer chemotherapy resistance and its reversion with the demethylating agent hydralazine. *J Transl Med*. 2006;4:32.

119. Merlo L, Pepper J, Reid B, Maley C. Cancer as an evolutionary and ecological process. *Nature Reviews Cancer*. 2006;6(12):924-935.

120. Greaves M, Maley C. Clonal evolution in cancer. *Nature*. 2012;481(7381):306-313.

121. Yates L, Campbell P. Evolution of the cancer genome. *Nat Rev Genet*. 2012;13(11):795-806.

122. Visvader JE. Cells of origin in cancer. *Nature.* 2011; 469(7330):314–322.

123. Gatenby R, Gillies R. A microenvironmental model of carcinogenesis. *Nature Reviews Cancer*. 2008;8(1):56-61.

124. Barcellos-Hoff MH, Park C, Wright EG. Radiation and the microenvironment - tumorigenesis and therapy. *Nature Reviews Cancer*. 2005; 5(11):867- 875.

125. Baylin S, Jones P. A decade of exploring the cancer epigenome- biological and translational implications. *Nature Reviews Cancer*. 2011;11(10):726-734.

126. O'Hagan H, Mohammad H, Baylin S. Double strand breaks can initiate gene silencing and sirt1-dependent onset of dna methylation in an exogenous promoter CpG island. *PLoS Genetics*. 2008;4(8):e1000155.

127. Lee G, Kim J, Taylor M, Muller M. DNA methyltransferase 1-associated protein (dmap1) is a co-repressor that stimulates dna methylation globally and locally at sites of double strand break repair. *Journal of Biological Chemistry*. 2010;285(48):37630-37640.

128. Cuozzo C, Porcellini A, Angrisano T et al. DNA damage, homology-directed repair and DNA methylation. *PLoS Genetics*. 2007; 3(7):e110.

129. Morano A, Angrisano T, Russo G et al. Targeted DNA methylation by homology-directed repair in mammalian cells. Transcription reshapes methylation on the repaired gene. *Nucleic Acids Research*. 2013;42(2):804-821.

130. Bullinger L, Ehrich M, Dohner K et al. Quantitative DNA methylation predicts survival in adult acute myeloid leukemia. *Blood*. 2009;115(3):636-642.

131. Figueroa M, Skrabanek L, Li Y et al. MDS and secondary AML display unique patterns and abundance of aberrant DNA methylation. *Blood*. 2009;114(16):3448-3458.

132. Deneberg S, GrÃ¶vdal M, Karimi M et al. Gene-specific and global methylation patterns predict outcome in patients with acute myeloid leukemia. *Leukemia*. 2010;24(5):932-941.

133. Figueroa M, Lugthart S, Li Y et al. DNA Methylation Signatures Identify Biologically Distinct Subtypes in Acute Myeloid Leukemia. *Cancer Cell*. 2010;17(1):13-27.

134. Melki J, Clark S. DNA methylation changes in leukaemia. *Seminars in Cancer Biology*. 2002;12(5):347-357.

135. Boles J, Tirado CA. Epigenesis in acute myeloid leukemia: an update. *J Assoc Genet Technol*. 2013;39(2):61-65.

136. Toyota M, Kopecky KJ, Toyota MO, Jair KW, Willman CL, Issa JP. Methylation profiling in acute myeloid leukemia. *Blood*. 2001; 97(9):2823-2829.

137. Garcia-Manero G, Yang H, Kuang S, O'Brien S, Thomas D, Kantarjian H. Epigenetics of Acute Lymphocytic Leukemia. *Seminars in Hematology*. 2009;46(1):24-32.

138. Matsushita C, Yang Y, Takeuchi S et al. Aberrant methylation in promoter-

associated CpG islands of multiple genes in relapsed childhood acute lymphoblastic leukemia. *Oncology Reports*. 2004.

139. Sharpless N, Sherr C. Forging a signature of in vivo senescence. *Nature Reviews Cancer*. 2015;15(7):397-408.

140. Crespi B, Summers K. Evolutionary biology of cancer. *Trends in Ecology & Evolution*. 2005;20(10):545-552.

141. Pienta K, McGregor N, Axelrod R, Axelrod D. Ecological therapy for cancer: defining tumors using an ecosystem paradigm suggests new opportunities for novel cancer treatments. *Translational Oncology*. 2008;1(4):158-164.

142. Gilbert L, Hemann M. DNA Damage-mediated induction of a chemoresistant niche. *Cell*. 2010;143(3):355-366.

143. Hofmann W. Presence of the BCR-ABL mutation Glu255Lys prior to STI571 (imatinib) treatment in patients with Ph+ acute lymphoblastic leukemia. *Blood*. 2003;102(2):659-661.

144. Roche-Lestienne C, Soenen-Cornu V, Grardel-Duflos N, Laï JL, Philippe N, Facon T, Fenaux P, Preudhomme C. Several types of mutations of the Abl gene can be found in chronic myeloid leukemia patients resistant to STI571, and they can pre-exist to the onset of treatment. *Blood*. 2002;100(3):1014-1018.

145. Shah N, Nicoll J, Nagar B et al. Multiple BCR-ABL kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. *Cancer Cell*. 2002;2(2):117-125.