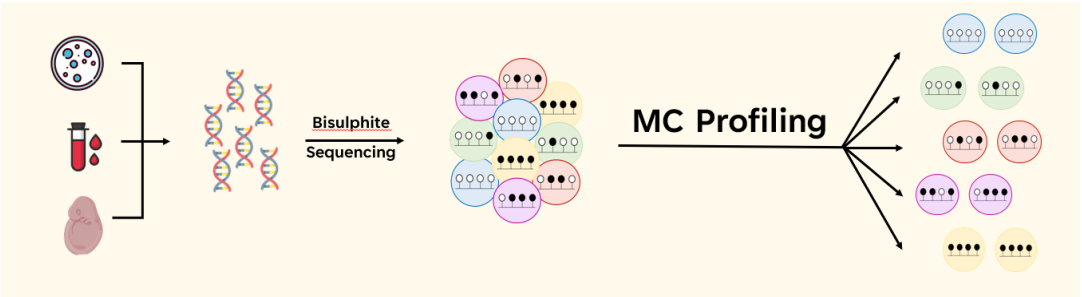
UNIVERSITY OF NAPLES FEDERICO II

DOCTORATE IN
MOLECULAR MEDICINE AND MEDICAL BIOTECHNOLOGY

XXXV CYCLE

Giulia De Riso

MC profiling: a novel approach to dissect DNA methylation heterogeneity in bulk bisulfite sequencing experiments



2019-2023

UNIVERSITY OF NAPLES FEDERICO II

DOCTORATE IN
MOLECULAR MEDICINE AND MEDICAL BIOTECHNOLOGY

XXXV CYCLE

MC profiling: a novel approach to dissect DNA methylation heterogeneity in bulk bisulfite sequencing experiments

Tutor
Prof. Sergio Cocozza

Candidate
Giulia De Riso

2019-2023

Table of Contents

List of Abbreviations used

DNMT = DNA methyl-transferases

TET = Ten-to-Eleven Translocation enzymes

MC = Methylation Classes

MP = Methylation Patterns

RRBS = Reduced representation Bisulfite Sequencing

WGBS = Whole Genome Bisulfite Sequencing

ABS = amplicon Bisulfite Sequencing

5mC = 5-methyl-cytosine

PMF = Probability Mass Function

ASM = Allele-specific methylation

JSD = Jensen-Shannon Distance

Abstract

DNA methylation is an epigenetic mark implicated in crucial biological processes. Most of the knowledge about DNA methylation is based on bulk experiments, in which DNA methylation of genomic regions is reported as average methylation. However, average methylation does not inform on how methylated cytosines are distributed in each single DNA molecule.

Here, we propose Methylation Class (MC) profiling as a genome-wide approach to the study of DNA methylation heterogeneity from bulk bisulfite sequencing experiments. The proposed approach is built on the concept of MCs, groups of DNA molecules sharing the same number of methylated cytosines. The relative abundances of MCs from sequencing reads incorporates the information on the average methylation, and directly informs on the methylation level of each molecule.

By applying our approach to publicly available bisulfite-sequencing datasets, we individuated signatures of loci undergoing imprinting and X-inactivation, and highlighted differences between the two processes. When applying MC profiling to compare different conditions, we identified methylation changes occurring in regions with almost constant average methylation.

Altogether, our results indicate that MC profiling can provide useful insights on the epigenetic status and its evolution at multiple genomic regions.

1. Background

1.1. DNA methylation: an overview

DNA methylation is a heritable epigenetic mark consisting of the enzyme-mediated addition of a methyl-group to deoxyribonucleotides (Jones, 2012; Kim and Costello, 2017; Moore et al., 2013). Methylation of DNA cytosines is the most prevalent form of DNA methylation, although adenine methylation has also been described (Boulias and Greer, 2022). In mammals, DNA methylation mainly occurs in CpG dinucleotides (Jones, 2012; Kim and Costello, 2017; Moore et al., 2013), with CpX methylation being constrained to specific tissues (neurons) and developmental stages (pluripotency) (de Mendoza et al., 2021; Ramsahoye et al., 2000). In the following text, we will focus on CpG methylation.

DNA methylation has been shown to regulate gene expression and genome stability by recruiting proteins involved in gene repression and transposon silencing, or by inhibiting the binding of transcription factor(s) to DNA (Moore et al., 2013). Recently, a more general involvement of DNA methylation in shaping the 3D chromatin conformation is also emerging (Buitrago et al., 2021), although the precise mechanism and causal relationships remain to be elucidated.

Due to its functional impact, DNA methylation has been implicated in crucial biological processes, such as cellular differentiation (Khavari et al., 2010), development (Smith and Meissner, 2013), disease (Robertson, 2005), aging (Fraga and Esteller, 2007), X-inactivation (Cotton et al., 2015), imprinting (Li et al., 1993), silencing of repetitive DNA (i.e. transposons) (Slotkin and Martienssen, 2007) and chromosomal stability (Rizwana and Hahn, 1999).

Dysregulated patterns of DNA methylation have been described in plenty of pathological conditions (Ehrlich, 2019; Robertson, 2005), including infective disease (Konigsberg et al., 2021) and genetic diseases (Villicaña and Bell, 2021), both complex and monogenic (Levy et al., 2022).

Two counteracting processes shape the DNA methylome: DNA methylation and DNA demethylation (Figure 1) (Smith and Meissner, 2013).

DNA methylation is an enzymatic process regulated by the family of DNA methyltransferases, which transfer the methyl group from the donor molecule SAM to the carbon 5 of the target cytosine (Lyko, 2018). Four DNMTs have been described in human (Lyko, 2018):

a) DNMT1 and DNMT3 (present in two subtypes, 3A and 3B) are the catalytically active members of the family

b) DNMT3L has not a catalytic activity itself but it is found in complex with DNMT3s, where it improves the ability of DNMT3s to bind to SAM (the methyl group donor) and stimulates DNMT3s' enzymatic activity

c) DNMT2 is not involved in DNA methylation. Instead, it participates in the methylation of cytosine-38 in the anticodon loop of the tRNA of the aspartic acid.

Classically, two types of DNA methylation are distinguished (Jeltsch and Jurkowska, 2014):

a) maintenance DNA methylation, which restores the pattern of DNA methylation on the newly-synthesized filament during DNA replication

b) de novo DNA methylation, which establishes patterns of DNA methylation in response to environmental stimuli or of developmental programs.

Maintenance methylation is usually attributed to the activity of DNMT1 due to its higher affinity for hemi-methylated cytosines (Hermann et al., 2004, p. 1). In contrast, de novo methylation is attributed to DNMT3s due to their higher affinity to unmethylated DNA as well as to their higher concentration in pluripotent stem cells (Okano et al., 1999). However, it has become clear that the two subfamilies of DNMTs cooperate in both de novo and maintenance DNA methylation. More exclusive roles have been instead described in the establishment and maintenance of imprinted DNA methylation and X chromosome inactivation (Dahlet et al., 2020; LaSalle, 2022).

Despite its stability, mammalian 5mC can be reversed through passive or active demethylation (Kohli and Zhang, 2013). Passive demethylation occurs during DNA replication as a consequence of the failure of the methylation maintenance machinery, which results in progressive dilution of 5mC (Kohli and Zhang, 2013). Active demethylation is an enzymatic process mediated by the family of Ten-to-Eleven Translocation (TET) proteins (Kohli and Zhang, 2013; Wu and Zhang, 2017). These enzymes belong to the family of dioxygenases, and use alpha-ketoglutarate to mediate the iterative oxidation of 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (Kohli and Zhang, 2013; Wu and Zhang, 2017). Replication-dependent dilution of these oxidized products results in DNA demethylation during replication (Kohli and Zhang, 2013; Wu and Zhang, 2017). For 5fC and 5caC, demethylation can also occur through base removal mediated by thymine DNA glycosylase (TDG) followed by the activity of the base excision repair (BER) pathway (Kohli and Zhang, 2013; Wu and Zhang, 2017).

It is worth noting that other than intermediate products of DNA demethylation, 5hmC, 5fC and 5caC are nowadays thought to act as epigenetic marks themselves, even though their exact role in regulatory processes and the interplay with 5mC remain to be elucidated (Caldwell et al., 2021).

Different isoforms have been also individuated for TET enzymes: TET1, TET2, and TET3 (Wu et al., 2018). Unlike DNMTs, different TETs do not exhibit preferential binding in the genome, and all contribute to maintain the unmethylated status of promoters' CpGs near promoter regions (Wu et al., 2018). However, TET enzymes are expressed at different magnitudes in different tissues both in mouse and in humans, supporting a preferential role of different TET subtypes to the regulation of lineage-specific gene expression (Wu et al., 2018).

CpG sites and their degrees of methylation are unevenly distributed in the human genome (Ehrlich et al., 1982). In the largest fraction of the human genome (about 98%) CpG sites are relatively infrequent (on average 1 CpG per 100 bp) but highly methylated (approximately 70%–80% of CpG sites) (Ehrlich et al., 1982). The remaining fraction of the genome (about 2% of the genome) comprises short stretches of DNA (approximately 1 kb in length and longer than 200 bp), known as CpG islands, 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 % (Ehrlich et al., 1982). CpG islands are found within the promoters of ~60-70% of human genes, characterized by an unmethylated status, a transcriptionally permissive chromatin state and generally associated with constitutive expression in all cell types (housekeeping genes) (Saxonov et al., 2006). However, some CpG islands specifically gain methylation in specific tissues or during the development (Li, 2002), resulting in a stable transcriptional repression. Furthermore, CGI hypermethylation is required for the long-term silencing of genes located on the inactive X chromosome (Cotton et al., 2015) or associated with imprinted loci (Li et al., 1993), germline-specific genes (De Smet et al., 1999) and pluripotency-associated genes (Mohn et al., 2008).

Genomic regions lying 2000 base-pairs to each side of a CpG island are named CpG "shores," whereas regions further extending for 2000 base-pairs from shores are named CpG "shelves", with the rest of the genome termed "open sea" (Carmona et al., 2017). Together with CpG islands, these contexts form the CpG "resort", with the concentration of CpG sites decreasing from islands to the open sea (Carmona et al., 2017).

Whereas DNA methylation within promoter CpG islands exhibits patterns established during cellular differentiation, DNA methylation in CpG shores and shelves is more responsive to external factors and can be of interest when trying to determine whether DNA methylation mediates known associations between exposures and diseases (Carmona et al., 2017).

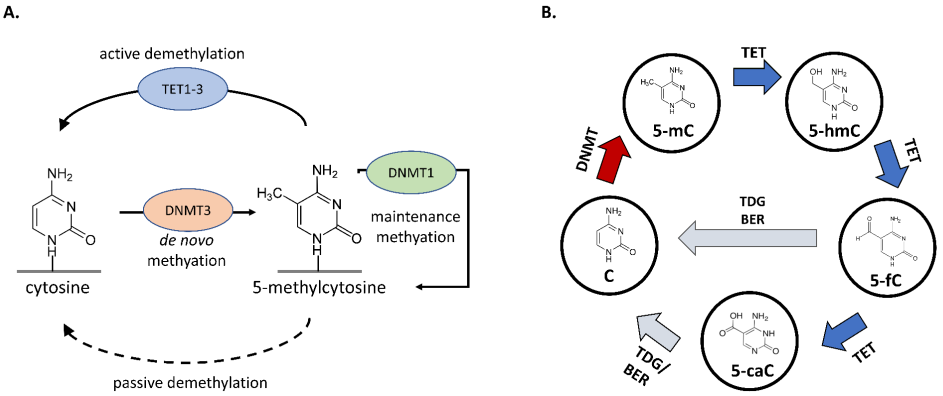


Figure 1. The process of DNA methylation. A. Overview of DNA methylation and demethylation. The two subtypes of DNMTs are shown to regulate the maintenance (DNMT1) and the de-novo (DNMT3A-3B) deposition of DNA methylation. Methylated cytosines can be reverted to their unmethylated status by passive demethylation or by active demethylation mediated by TET enzymes (TET1-3). B. Steps of TET-mediated iterative oxidation of 5-mC. The intermediates 5-hC (5-hydroxy-methyl-cytosine), 5-fC (5-formyl-cytosine) and 5-caC (5-carboxyl-cytosine) are shown. (TET=Ten-to-Eleven Translocation; DNMT= DNA methyl-transferase; TDG=Thymine glycosidase; BER= Base Excision Repair).

1.2 Experimental assays to investigate DNA methylation

Several experimental techniques have been developed to study DNA methylation (Yong et al., 2016). State-of-the-art methods are based on the treatment of genomic DNA with sodium bisulfite, which enable the identification of 5-methylcytosine at single base-pair resolution (Yong et al., 2016). In fact, after bisulfite treatment cytosines are converted into uracil residues through oxidative deamination, whereas 5mCs are immune to this conversion (Yong et al., 2016). Thus, methylated cytosines can be distinguished from unmethylated ones by analyzing the resulting DNA sequence (Figure 2) (Yong et al., 2016).

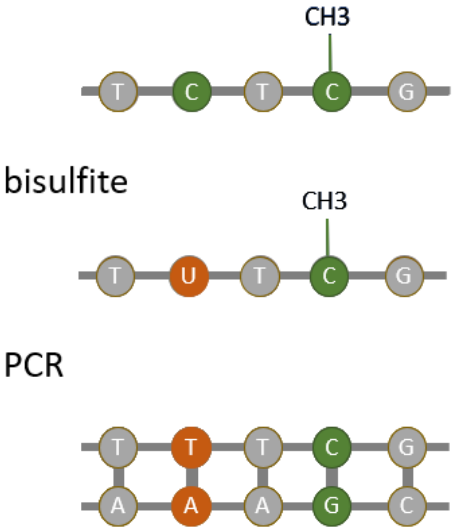


Figure 2: Effect sodium bisulfite treatment on the DNA sequence. Unmethylated cytosines are converted into uracils, and indeed into tymines after PCR amplification. Methylated cytosines are not converted, and indeed not substituted after PCR amplification.

We can distinguish two categories of bisulfite-based assays: hybridization-based assays, like Illumina BeadChips, which uses genotyping probes to discriminate and quantify bisulfite-induced mutations, and sequencing-based assays, which directly detect bisulfite-induced mutations through PCR and sequencing (Yong et al., 2016).

In particular, hybridization assays employ site-specific probes that hybridize onto bisulfite-converted DNA at given CpG loci, resulting in fluorescent signals. With sequencing-based assays, the sequence of a bisulfite-treated DNA molecule is compared to the reference genome at CpG sites. For each CpG, a methylated status (1) is assigned if a C is found in the DNA molecule, and an unmethylated status (0) is assigned if a T is found.

Microarrays platforms allow assessing the methylated status of huge numbers of CpG sites (ranging from about 27.000 to 850.000 for the widely adopted Illumina platforms) distributed along the whole genome (Carmona et al., 2017). Microarrays are therefore a fast and cheap assay to individuate genome-wide DNA methylation alterations at reproducible targets, and are therefore widely adopted, for example, in epigenome-wide association studies where a high number of samples are tested for DNA methylation alterations associated to a given molecular or clinical phenotype (Carmona et al., 2017). However, microarrays are only available for a restricted number of species,and

constrain the observations to pre-defined CpG targets, mostly designed to capture genes and promoter CpG islands (Carmona et al., 2017). Furthermore, they come with some technical issues, like dye-biases and different probe chemistries, and positional effects due to probe cross-reactivity and ambiguous mapping, that must be corrected during data processing and that potentially reduce the number of usable probes (Carmona et al., 2017).

Sequencing-based strategies are instead applicable to whatever species with available reference genome, and can virtually cover all CpG sites, and also CpX sites, in a region of interest, until the whole genome (Carmona et al., 2017). Furthermore, sequencing strategies enable rescuing the information on how methylated residues are phased in a DNA molecule (Landan et al., 2012; Landau et al., 2014; Li et al., 2014). This information, which is lost when using array platforms, is important to explore the inner dynamics of DNA methylation and demethylation, as well as to explore the molecular and cellular heterogeneity of DNA methylation (as will be illustrated in the next section) (Landan et al., 2012; Landau et al., 2014; Li et al., 2014).

Bisulfite sequencing techniques are adopted to assess the methylation status at single base resolution at targeted regions or at genome-wide level (Gu et al., 2011; Masser et al., 2013; Varley and Mitra, 2010; Yong et al., 2016).

A popular targeted sequencing experiment is amplicon bisulfite sequencing (ABS) (Florio et al., 2017; Klobučar et al., 2020). In ABS, the region of interest is selectively amplified, and the amplified products are then sequenced (Florio et al., 2017; Klobučar et al., 2020). ABS is generally achieved through double-step PCR (Florio et al., 2017; Klobučar et al., 2020). In a first step, bisulfite-treated genomic DNA is amplified by using target-specific primers. The 5′ end of these primers contains overhang adaptor sequences that are used in the second step of amplification. This latter step enables the addition of multiplexing indices and sequencing adaptors to the amplified molecules. The obtained sequencing library is then multiplexed and sequenced (Florio et al., 2017; Klobučar et al., 2020). ABS is a useful approach to profile DNA methylation in hypothesis-driven settings, selecting for genes or regulatory regions presumably involved in a biological process (Florio et al., 2017; Klobučar et al., 2020).

An interesting variant of ABS is Deep-ABS, which takes advantage of the throughput of NGS platforms to in deep sequence (10^3 to 10^5 reads per region) amplicon-selected libraries (Florio et al., 2017; Klobučar et al., 2020). Other than enabling an accurate analysis of the amount of methylated molecules per CpG site, Deep-ABS enables a robust analysis of the arrangements of methylated residues in individual DNA molecules (Affinito et al., 2016; Florio et al., 2017).

Sequencing-based approaches for measuring DNA methylation across the human genome have rapidly scaled to the whole genome over the last decade. Whole-genome bisulfite sequencing (WGBS) enables the assessment of the methylation status at almost all CpG sites of the human genome.

In WGBS, DNA is sheared usually by sonication, and then subjected to end repair and to the addition of an adenosine nucleotide at the 3′ end, in a process called end repair and A-tailing (Karemaker and Vermeulen, 2018). The A-tail serves as a binding site for sequencing adapters' ligation. After this, fragments of homogenous size compatible with sequencing requirements are selected and subjected to bisulfite conversion. Bisulfite-converted fragments are amplified by PCR and sequenced (Karemaker and Vermeulen, 2018). At least 500 million reads are needed to provide 1x coverage of the whole genome. Therefore, WGBS requires a large amount of input DNA (1-3 μg) and generates a large amount of data that poses computational costs (Karemaker and Vermeulen, 2018). Furthermore, much of the WGBS data is not informative, due to the invariant DNA methylation status of large genomic regions across conditions and cell-types or to a lack of overlap between samples (Karemaker and Vermeulen, 2018).

To overcome these limitations, enrichment-based assays have been developed to selectively assay more informative regions. Reduced Representation Bisulfite sequencing (RRBS) is a popular bisulfite sequencing assay which enriches for CG-rich parts of the genome, thereby reducing the amount of sequenced genome while capturing the majority of promoters and other relevant regulatory regions (Gu et al., 2011). In RRBS experiments, the purified genomic DNA is digested with the methylation-insensitive restriction enzyme MspI, which recognises the sequence CCCG and cuts between the first and the second cytosine, thus generating CG-ending fragments. The digested DNA is then subjected to end-repair and A-tailing, to enable adapter ligation. The obtained fragments are then size-selected with insertion sizes generally ranging between 40 and 220 base-pairs. The DNA fragments are subjected to bisulfite conversion, PCR amplified and sequenced on a NGS platform (Gu et al., 2011).

RRBS requires less amount of input DNA (10–300 ng) then WGBS, while retaining most information about the DNA methylome (Gu et al., 2011). However, the original version of the RRBS assay poorly covered regions with intermediate CG density, like the CG-shores, that undergo DNA methylation changes upon environmental stimuli, or distal regulatory elements (Gu et al., 2011).

An enhanced version of RRBS (ERRBS) has been therefore developed and is becoming increasingly adopted (Garrett-Bakelman et al., 2015). In ERRBS, the MspI-digested DNA is size-selected to enrich for fragments corresponding to 84–334 base-pairs. This fraction is selected from agarose gel

in two fractions: one ranging between 84 and 184 base-pairs and one ranging between 185 and 334 base-pairs. These fractions are bisulfite treated and PCR-amplified independently, and are t pooled at the same molarity before sequencing, resulting in good representation of the overall 84-334 base-pairs length range (Garrett-Bakelman et al., 2015). ERRBS covers about 10% of genomic CpG sites, and provides a higher coverage of regions outside CpG islands and promoters, including exons, introns, and intergenic regions (Garrett-Bakelman et al., 2015; Kacmarczyk et al., 2018).

It is worth noting that different enrichment strategies are also available to profile DNA methylation at relevant epigenetic regions. In methyl-sequencing assays, target regions are captured from sonicated DNA by hybridization with capture platforms or oligo-covered baits (Kacmarczyk et al., 2018). Several kits are available to target a broad panel of genomic regions known to be epigenetically regulated (Kacmarczyk et al., 2018). Alternatively, target panels have been developed to profile genomic regions relevant for specific biological processes, like genomic imprinting (Ochoa et al., 2022).

Methyl-sequencing assays offer the great advantage to reproducibly and stably profile DNA methylation at same genomic regions in different samples (Kacmarczyk et al., 2018). However, the higher amount of input DNA required (0.25-3 μg) and higher costs have limited their adoption (Kacmarczyk et al., 2018).

Recently, bisulfite sequencing has also been implemented in single-cell settings (Karemaker and Vermeulen, 2018). Studies based on single-cells bisulfite sequencing have highlighted huge cell-to-cell differences even in homogenous conditions (Carter and Zhao, 2021). However, technical issues and elevated costs still limit the diffusion of such assays for DNA methylome (Karemaker and Vermeulen, 2018).

Among these issues, a critical one is the scarce coverage of genomic CpG sites (about 10^6 CpGs covered), dropping from 40% of bulk RRBS to 4% of a single-cell experiment (Karemaker and Vermeulen, 2018). As a consequence, single-cell DNA methylation assays suffer from a restricted overlap in the covered sites between individual cells (Karemaker and Vermeulen, 2018).

Therefore, the dissection of cellular DNA methylation heterogeneity still mostly relies on bulk experiments (Huan et al., 2018; Teschendorff et al., 2020). In this setting, increasing interest is devoted to single-cell guided deconvolution approaches, which enable inferring the proportion of different cell-types in a given sample from bulk DNA methylation data (Scherer et al., 2020a; Teschendorff et al., 2020).

Alternatively, analysis of DNA methylation heterogeneity can be carried out through the analysis of phased chains of methylated cytosines in

individual DNA molecules (Landan et al., 2012; Landau et al., 2014; Li et al., 2014). This latter approach will be discussed in detail in the next section.

1.3 Analytical approaches to DNA methylation

The state of the art of the analysis of DNA methylation is a quantitative approach, which consists in computing the fraction of molecules in which a certain cytosine is methylated out of the total number of analyzed molecules (Bock, 2012).

As such, the methylation status of a cytosine is generally expressed as a proportion, and ranges between zero and one (Bock, 2012). The comparison of methylation proportion across samples enables to identify CpGs whose amount of methylation is associated with environmental exposures (Mitchell et al., 2016), disease conditions (Jin and Liu, 2018) and chronological age (Hannum et al., 2013; Horvath, 2013).

Genome-wide studies of DNA methylomes have highlighted that the proportion of methylation of single cytosines follows a bimodal distribution, with most cytosines peaking near zero or one (Bock et al., 2010). However, most cytosines significantly diverge from the fully unmethylated and the fully methylated values, and about 2% of 26.9 million CpGs in the human genome exhibit intermediate DNA methylation values (between 0.25 and 0.75) in bulk samples (Scherer et al., 2020b). This indicates that most cytosines are not evenly methylated in different DNA molecules of the same sample, highlighting a certain degree of within-sample heterogeneity (Scherer et al., 2020b).

The main biological sources of such heterogeneity include cell-type composition, cell-to-cell differences, allele- and strand-specific DNA methylation (ASM and hemimethylation), and DNA methylation erosion, i.e. the stochastic loss of DNA methylation at a given locus (Scherer et al., 2020b).

Although a single methylated CpG may occasionally be linked to gene expression regulation (Xu et al., 2007) and may affect disease risk (Raval et al., 2007), evidence has been provided that DNA methylation is regulated in larger genomic regions, with sets of neighboring cytosines working as functional units (Haerter et al., 2014; Irizarry et al., 2009; Jaenisch and Bird, 2003; Zhang et al., 2017). DNA methylation analysis has indeed turned to the study of the amount of methylation of DNA regions, which is expressed as average methylation (the fraction of methylated cytosines in a given region), and on the identification of regions with consistently different DNA methylation levels between groups of samples (differentially methylated regions, DMR) (Bock, 2012). Most of the current knowledge on DNA methylation and its implication

in health and disease status is founded on this latter approach (Doi et al., 2009; Hansen et al., 2011; Lokk et al., 2014).

However, the overall average methylation of a region does not inform on how this amount is contributed by the average methylation of single DNA molecules. As an example, an average methylation value of 0.5 for a given locus could result from a homogenous pool of half methylated molecules, or from an heterogeneous, balanced set composed of fully methylated and unmethylated molecules, or even from more heterogeneous pools (Figure 3) (Mikeska et al., 2010).

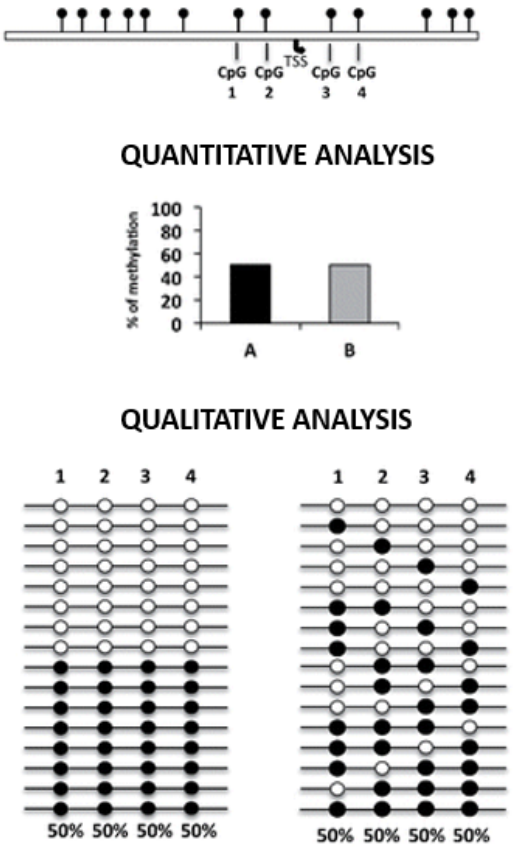


Figure 3. Quantitative versus qualitative analysis of DNA methylation. For a representative locus containing 4 CpG sites, two different configurations are shown. In both cases, average methylation (i.e. the overall fraction of methylated cytosines, here represented as black dots) is equal to 50%. However, in the first condition, two different epialleles with opposite methylation levels (i.e. a fully methylated and a fully unmethylated epiallele) are represented in the pool of DNA molecules. In the second condition, all possible arrangements of methylated cytosines (2^4) are equally represented in the pool of DNA molecules. (Figure adapted from Florio et al., 2016)

Single-cell DNA methylation assays have highlighted extensive cell-to-cell differences in regional DNA methylation (Huan et al., 2018; Karemaker and Vermeulen, 2018), and have demonstrated that cellular heterogeneity can have a functional impact. For example, epigenetic variability at regulatory elements has been linked with gene expression variability (Angermueller et al., 2016; Carter and Zhao, 2021).

Since single-cell DNA methylation assays are still limitedly adopted due to the high cost and large sparsity of produced data (Huan et al., 2018; Teschendorff et al., 2020), alternative approaches have been developed to dissect DNA methylation heterogeneity from bulk experiments.

These approaches are generally based on the analysis of the arrangements of methylated cytosines in individual bisulfite sequencing reads, referred to as epialleles (Huan et al., 2018; Landan et al., 2012; Landau et al., 2014; Li et al., 2014; Scherer et al., 2020b; Xu et al., 2007). Being focused not only on the methylation amount of a given region, but also on how it is distributed across cytosines in individual DNA molecules, these approaches are here referred to as qualitative.

Qualitative analysis of epiallele composition generally relies on two assumptions:

1) each sequenced read comes from and individual DNA molecule
2) each epiallele is representative of the epigenetic status of a given region in an haploid cell

Based on these assumptions, notions and techniques derived from population genetics, ecology and metagenomics have been adopted to directly compare the frequency of epiallele across conditions.

A common approach is to compute an heterogeneity score that summarizes the number of different epialleles and their relative proportion observed in a given region (Scherer et al., 2020b). These scores can be then used to define regions with similar degree of epigenetic variability and to individuate regions under epigenetic drift or clonal selection across conditions (Scherer et al., 2020b).

Some of the most popular scores are the Shannon entropy and the Epipolymorphism.

Shannon entropy of a given locus is a measure of the randomness of DNA methylation patterns in a cell population (Xie et al., 2011).

Epipolymorphism of a given locus represents the probability that two epialleles randomly sampled from the locus differ from each other (Landan et al., 2012).

Both scores range between 0 (minimum degree of epigenetic variability) and 1 (maximum degree of epigenetic variability).

A widely adopted measure to individuate shifts in epiallele composition between pairs of samples is the combinatorial entropy implemented in the tool *Methclone* (Li et al., 2014). Combinatorial entropy ranges between 0 and -144, where 0 corresponds to no change and -144 to the maximum entropy change that can be observed in a locus bearing 4 CpGs (Li et al., 2014).

Beside this approach, other tools such as AmpliMethprofiler and EpiStatProfiler adopt functions from ecology fields to directly compare epiallele composition observed in a given region in different samples from targeted and genome-wide experiments, respectively (Sarnataro et al., 2022; Scala et al., 2016). This latter approach can capture regions in which the same or a similar degree of epigenetic heterogeneity is contributed by different epialleles. These regions would be indeed overlooked when comparing heterogeneity scores (Sarnataro et al., 2022; Scala et al., 2016).

In genome-wide settings, qualitative analysis is limited in the ability to quantify heterogeneity across the genome compared to quantitative analysis. In fact, in order to properly reconstruct the phase of methylated cytosines for a region of interest, each sequencing read has to fully cover the entire region (Scherer et al., 2020b). Furthermore, a minimum number of reads is usually required to have a good representation of the entire range of epigenetic variability of the region (Scherer et al., 2020b).

The choice of target regions is therefore carried out in order to retain the maximum genomic coverage (i.e., the higher number of analyzed targets) while enabling an accurate estimate of epiallele composition of each target. The most adopted setting is to select regions encompassing 4 CpG sites (Landan et al., 2012; Li et al., 2014). Alternatively, regions of fixed size set according to sequencing read length can be selected (Sarnataro et al., 2022). The coverage threshold can be variable depending on the approach, ranging from 16 (the minimum number of reads to observe all possible epialleles in a 4-CpG region) to more stringent values (30-50 reads per target region) (Landan et al., 2012; Li et al., 2014; Sarnataro et al., 2022).

Nonetheless, epiallele analysis adds useful information about the dynamics that form, maintain and reprogram methylated regions in cancer and developing cells. In (Landan et al., 2012), the authors demonstrated increased epigenetic polymorphism in cancer cells, characterized by the co-existence of cell subpopulations gaining DNA methylation and others remaining completely resistant to methylation (Landan et al., 2012). Gain of methylation proceeds by

progressive sensitization of resistant CpG sites to DNA methylation (Landan et al., 2012). Moreover, epiallele analysis was shown to capture DNA methylation shifts in leukemia patients from diagnosis to relapse after treatment, shifts that would be missed when adopting a quantitative approach (Li et al., 2014). In a targeted study, Florio et al. demonstrated that neuronal differentiation proceeds through the emergence of characteristic epialleles from non-organized pools observed in undifferentiated cells (Florio et al., 2017). This pattern of heterogeneity was reproducible among individuals and independent from cell-type composition (Florio et al., 2017).

1.5 Reconciling quantitative and qualitative analysis of DNA methylation

When performing qualitative analysis of DNA methylation, the overall amount of DNA methylation of individual epialleles is usually not taken into account. However, it is expected that the functional properties of individual epialleles is at least influenced, if not determined, by their inherent level of methylation.

In this direction, several scores have been developed to integrate quantitative and qualitative aspects of DNA methylation heterogeneity.

The proportion of discordant reads (PDR) score is one of the first and successful attempts of taking into account the methylation levels to interpret DNA methylation heterogeneity (Landau et al., 2014). This score is based on the analysis of discordant reads, i.e. reads aligned to a certain region and carrying cytosines with different methylation status, independently of their position. PDR is indeed computed as the fraction of discordant reads out of the total number of reads (Landau et al., 2014). Notably, PDR can be computed for each individual CpG site, provided that the reads covering the CpG have a minimum reciprocal overlap (Landau et al., 2014). PDR has enabled to highlight locally disordered DNA methylation as a signature of leukemia cells, which is currently viewed as facilitating tumor evolution through increased epigenetic plasticity (Landau et al., 2014). Moreover, increased PDR resulted in low-levels of gene expression and adverse clinical outcomes of leukemic patients (Landau et al., 2014).

Cell Heterogeneity–Adjusted cLonal Methylation (CHALM) is a promoter-centered methylation quantification method based on the estimate of the fraction of epialleles holding ≥1 methylated cytosines to the total number of epialleles observed for the promoter (Xu et al., 2021).

CHALM was shown to better correlate with gene expression and histone marks better than promoter average methylation (Xu et al., 2021).

Concurrence ratio quantifies the ratio between the percentage of unmethylated CpGs in partially methylated reads to assess the concurrence of DNA methylation and demethylation in a genomic region (Shi et al., 2021).

By using this score, the authors were able to stratify large undermethylated regions into two subgroups with distinct chromatin and gene regulation patterns (Shi et al., 2021). Moreover, the authors demonstrated a strong correlation between high concurrence ratio and the repression of a relevant fraction of tumor suppressor genes (Shi et al., 2021).

As for epiallele analysis, analytical approaches can be adopted to analyze and compare the full distribution of methylation levels exhibited by DNA molecules at a given genomic region.

Mathematical modeling has been indeed applied to infer the distribution of methylation levels from Whole Genome Bisulfite Sequencing (WGBS) data at 150 base-pairs genomic windows (Jenkinson et al., 2017).

The starting point of this approach is a mathematical model incorporating relevant factors known to shape DNA methylation dynamics, including the physical distance among CpG sites and the cooperative effect between neighboring cytosines (Jenkinson et al., 2017). Experimental data, i.e. the epiallele configuration of sequencing reads mapped at each 150 base-pairs windows, are then fitted to the model in order to compute the region-specific parameters (for example, the entity of CpG co-methylation) (Jenkinson et al., 2017). These parameters, integrated in the mathematical model, enable to compute the probability mass function (PMF) of methylation levels that could be observed in a pool of molecules at the inherent genomic region (Jenkinson et al., 2017).

This approach, specifically designed to deal with the low coverage of WGBS experiments, has provided novel insights on DNA methylation heterogeneity and its disposition across the genome, its evolution upon differentiation, aging and cancer, and its relationship with the genetic background (Abante et al., 2020; Jenkinson et al., 2018, 2017).

In previous studies, high-coverage amplicon bisulfite sequencing allowed us to directly estimate the distribution of methylation levels from supporting sequencing reads at targeted regions (Affinito et al., 2016; De Riso et al., 2020). Our approach, here referred to as MC profiling, was based on the concept of Methylation Classes (MCs), i.e. groups of molecules holding the same amount of methylated cytosines (Affinito et al., 2016; De Riso et al., 2020).

The underlying idea of MC profiling is that looking at the distribution of epialleles grouped by their methylation levels adds useful information for the functional interpretation of DNA methylation heterogeneity in a sample.

We already applied the concept of MCs in previous works with the aim to model DNA methylation dynamics at targeted loci assayed through high-coverage bisulfite sequencing (Affinito et al., 2016; De Riso et al., 2020).

The aim of this study is therefore to extend MC profiling to genome-wide bisulfite sequencing data, with the aim to explore DNA methylation heterogeneity of a huge number of genomic regions from the same sample.

2. Aims of your study

The goal of the project was to extend MC profiling to genome-wide bulk bisulfite experiments, which enable to assay the methylation status of thousands of genomic regions in the same sample.

Towards this goal, we planned the following specific aims:

1) assess the feasibility (in terms of accuracy and precision) of MC profiling in low-coverage settings
2) Automate the extraction of MC profiles from genome-wide bisulfite sequencing data
3) develop a computational framework to analyze and compare MC profiles observed in a given region in different conditions (inter-sample analysis)
4) develop a computational framework to analyze and compare MC profiles observed at multiple regions in a given conditions (within-sample analysis)
5) Test the potential of MC profiling on well-known patterns of DNA methylation heterogeneity (both within and between conditions)

3. Materials and Methods

MC profiling

a. MC profile computation

For each epilocus, i.e. a region holding 4 CpGs, we first selected the reads spanning the entire regions. We then computed the set of the relative abundances of the 5 possible methylation classes (MCs), here referred to as MC profile. To this aim, we counted the different configurations of methylated cytosines found supported by the selected reads. We grouped these configurations in 5 MCs according to the number of methylated cytosines. For each MC, we computed the relative abundance as the fraction of sequencing reads supporting the MC out of the total number of reads.

b. Measure of dissimilarity

We adopted the Jensen-Shannon Distance to measure the dissimilarity between two MC profiles. The Jensen Shannon Distance (JSD) quantifies the degree of dissimilarity between discrete distributions P_1 and P_2 (in our case represented by two sets of relative abundances), and is defined as

$$d = \sqrt{\frac{D(P_1,\underline{P})+D(P_2,\underline{P})}{2}}$$ (Lin, 1991)

$\underline{P} = \frac{P_1+P_2}{2}$ is the "average profile", obtained from the mean of the relative abundance of each MC among the two MC profiles.

$D(P_1, \underline{P})$ and $D(P_2, \underline{P})$ represent the Kullback-Leibler (KL) divergence between the average profile $\underline{P}$ and the profile P_1 and P_2, respectively.

The KL divergence is computed as $D(A, B) = A * log2(\frac{A}{B})$ (Cover and Thomas, 2005).

d. Epilocus filtering and multiple samples handling

We limited our analysis to epiloci with a coverage (number of reads spanning the entire epilocus) of at least 50 reads and not higher than a sample-specific cutoff, computed as the 99th percentile of the coverage of all epiloci. To limit our observations to a set of independent epiloci, we applied a sliding selection. Having the set of epiloci ordered by genomic coordinates, we jumped from the first to the nearest non overlapping epilocus, thus retaining only non-overlapping epiloci.

When multiple samples were available for the same condition, we handled MC profiles observed in different samples by computing a consensus MC profile at each epilocus. First, we quantified the inter-individual variability for an epilocus by computing the JSD between the MC profiles observed in the possible sample pairs. After this step, we retained epiloci with low inter-individual variability, i.e. with JSD below 0.26 in all the pairs. For each of these epiloci, we computed the average MC profile by averaging the relative abundance of each MC among all the samples. In this way, we obtained a consensus MC profile representative of all samples in a given condition, that we could directly compare among conditions through the JSD.

e. MC profiles classification

To provide biological interpretation of MC profiles, we adopted a data compression scheme. We defined 5 archetypal profiles, reminiscent of standard discrete distributions and reflecting the reasonable profiles expected given a certain methylation amount. We then assigned each MC profile to one among 5 groups named Methylation Patterns (MPs) according to the most similar archetypal profile.

To assign an MC profile to the nearest MP, we computed the JSD from all the 5 archetypal profiles. We then assigned the MC profile to the MP corresponding to the most similar archetypal profile (i.e. the reference profile with minimum JSD).

We checked the appropriateness of our classification procedure by comparing the JSD of each MC profile from the two nearest MPs, with the lower JSD value representing the distance from the membership pattern centroid (Within Class Distance, WCD) and the higher value representing the distance from the nearest outer pattern centroid (External Class Distance, ECD).

Dataset

We analyzed previously published RRBS data and enhanced RRBS data publicly available in the GEO database (https://www.ncbi.nlm.nih.gov/geo/) with the following accessions: GSE66121, GSE130735, GSE53714, GSE72700. The samples adopted from each dataset are described in Table 1.

Table 1: RRBS datasets adopted in this project

Dataset	GEO accession	Sample accessions	Description
Dataset 1	GSE130735	GSM3752619, GSM3752620, GSM3752621	samples from 3 WT littermate E8.5 embryos
Dataset 2	GSE66121	GSM1614765, GSM1614766, GSM1614767	human CD19+ B-cells isolated from 3 normal controls
Dataset 3	GSE53714	GSM1299332, GSM1299333, GSM1299334	liver samples from 3 F1 mice originating from C57BL/6J and DBA/2J strain cross
Dataset 4	GSE72700	GSM1868584, GSM1868589, GSM1868591	ERRBS data from C57BL6 male mice neurons at different developmental stages (hippocampal precursors:, granule cells:; CA3 neurons:)

Data processing

a. RRBS raw data processing

Fastq files were quality checked by using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Low-quality bases were removed using Trim Galore v0.6.6 with parameters --rrbs and –paired for paired end experiments (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). The obtained fastq were aligned to the reference genomes (hg19 for human samples and mm10 for mouse samples) through Bismark v0.23.0 employing default parameters. The obtained BAM files were sorted and indexed using SAMtoolsKit (http://www.htslib.org/).

b. Deep - Amplicon Bisulfite Sequencing data processing

D-ABS data were processed as previously described (Affinito et al., 2020; Cuomo et al., 2019; Florio et al., 2017). In brief, paired-end reads were merged in a single fastq file through PEAR (minimum overlapping residues equal to 40) (https://cme.h-its.org/exelixis/web/software/pear/doc.html). The fastq file was then converted to fasta through PRINSEQ (http://prinseq.sourceforge.net/).

c. Epiallele counts extraction

For RRBS data, epiallele counts were extracted from BAM files using the utilities provided by the *EpiStatProfiler* R package (Sarnataro et al., 2022). Genomic regions covered by at least 50 reads were individuated through the *filterByCoverage* function. Target regions holding 4 CpGs (the epiloci described in this manuscript) were individuated by using the *makeBins* function (step parameter equals to 1). The maximum length of the target regions was set according to the specific library design (ranging from 70 to 100). Epiloci were then analyzed by using the *epiStatAnalysis* function with default parameters. For each epilocus, this function returns a table with summary statistics (such as average methylation), and a file with epiallele counts. This latter was analyzed through in-house R scripts to compute MC profiles, as described above.

For D-ABS data, epiallele counts were then extracted by adopting AmpliMethProfiler (Scala et al., 2016). The MC profile of the amplicon was then computed following the same procedure of RRBS epiloci.

d. Allele-specific alignment sorting

To perform allele specific MC profiling, we applied the pipeline based on the SNPsplit tool (54) on a dataset of crossed strain mice. First, the positions holding alternative sequences between the strains were extracted from the VCF file downloaded from the Mouse Genomes Project repository (ftp://ftp-mouse.sanger.ac.uk/current_snps/mgp.v5.merged.snps_all.dbSNP142.vcf.gz), and were masked from the reference mm10 genome by using the SNPsplit_genome_preparation function in single strain mode. Fastq files were then aligned to the masked genome by using Bismark 0.23.0 with default parameters. The reads aligned to polymorphic sites were assigned to the respective allele by using the SNPsplit function.

In brief, the reads aligned to variant positions were tagged (SNPsplit-tag internal function), assigned to the reference or to the alternative allele (tag2sort internal function), and written down in separate bam files. We ran the SNPsplit function in --bisulfite mode to automatically discard the reads aligned to C/T or T/C variants on the forward strand and to G/A or A/G variants on the reverse strand, since these variants cannot be distinguished from a methylation status call. The bam files relative to the reference and the alternative allele were processed independently with the EpiStatProfiler tool to obtain the epiallele counts and to compute the MC profile.

At the end, we were able to profile 2749, 460, 314 autosomal epiloci in three mice, with a minimum coverage of 50 reads on both alleles.

e. Epiloci annotation

Epiloci were annotated by using the *annotatr* R package against hg19 and mm10 tracks (CpG islands and coding regions).

Epiloci were associated with the nearest genes by using the seq2pathway R package, setting the *'adjacent'* parameter to assign each epilocus to the closest genes only. For our analysis, we considered the FullResult output which also included non-coding genes.

To test the association between MC profiles and chromatin marks, epiloci of Dataset 2 were annotated using the chromHMM segmentation tracks for the GM12878 lymphoblastoid cell line from the RoadMap Epigenomics project (https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final), whereas epiloci of Dataset 5 were annotated using the segmentation tracks for mouse hindbrain (E10 and P0) obtained from UCSC (van der Velde et al., 2021). For Dataset 4, the E10 hindbrain track was used to annotate epiloci of hippocampal precursors,

whereas the P0 hindbrain track was adopted to annotate epiloci in differentiated neurons (Granule cells and CA neurons). Epiloci overlapping with genomic segments with different labels were annotated based on the label of the genomic segment with the highest overlap.

Association of MPs with expression level

For Dataset 2, we downloaded normalized expression data (FPKM) for 3 samples from GEO with the accession GSE66121. For each gene, we computed the average value among the 3 samples. We then assigned to the highly-expressed group those genes with expression value above the median, and assigned to the lowly-expressed group those genes with expression values below or equal to the median.

Epiloci were assigned to gene promoters, exonic or intronic regions by using the *annotatr* R package against hg19 genes track.

The association between the proportion of epiloci assigned to the different MPs and the expression status was tested through the chi-square test and post-hoc analysis of chi-square residuals (see the Statistical test Section).

Statistical analysis

a. Classification concordance of neighboring epiloci

To test the concordance of neighboring epiloci from Dataset 1 and 2, we binned the genome into 1 kb long regions. We removed the bins harboring less than 3 epiloci, and labeled the remaining ones as concordant if all the epiloci were assigned to the same MP, and discordant otherwise.

We used bootstrapping to test whether the number of concordant bins was higher than the one expected by chance. In brief, we scrambled the epiloci grouped in each bin, such that the overall number of bins together with the number of epiloci for each bin reflected those observed in experimental data, but the epiloci were no longer grouped in a bin based on their proximity. We repeated this procedure 1000 times, and each time we counted the number of bins classified as concordant. We thus obtained the distribution of the number of concordant bins expected by chance, that we used to compute the empirical probability of observing the number of concordant bins found in experimental data.

Statistical test

All the statistical analyses were performed using R software (version 4.0) with an alpha value set for $p < 0.01$.

Association between categorical variables was tested for statistical significance through either Fisher test (when both categorical variables were dichotomous) or chi-square test and post-hoc analysis of chi-square residuals (chi.square.posthoc.test function from the homonymous R package, adopting Bonferroni correction to control for alpha inflation). In particular, we applied Fisher to test whether MC profile changes more probably involved epiloci that also underwent chromatin changes upon differentiation, epiloci located in promoters or epiloci located in CpG Islands. We instead applied chi-square to test whether epiloci exhibiting inter-individual variability were enriched in peculiar genomic contexts (promoters, exons, introns, or intergenic regions), or whether epiloci assigned to different MPs were enriched in particular genomic regions (for example, regions flanking imprinted genes or regions decorated with different histone marks) or more probably changed MC profiles upon differentiation.

Differences in reciprocal distance among epiloci in concordant and discordant bins was tested through the Mann-Whitney test.

Enrichment analysis for 5129 epiloci with significant changes in MC profiles and stable average methylation upon differentiation was performed using GREAT version 4.0.4, using the coordinates of all the analyzed epiloci (115608) as background.

4. Results

4.1 The MC profiling approach

4.1.1 Rationale

The rationale of MC profiling, and the differences with epiallele-based approaches, is depicted in Figure 4.

Epiallele-based approaches are based on the direct analysis of the arrangements of methylated and unmethylated cytosines (epialleles) in sequencing reads mapped to a region of interest (Figure 4A).

Considering each reads coming from a DNA molecule, several scores have been developed to quantify the heterogeneity observed in a bulk sample, and to compare it among different samples (Figure 4B) (Scherer et al., 2020b). This approach has proved to be particularly suitable, for example, to individuate regions undergoing clonal selection and epigenetic drift in tumors (Landan et al., 2012; Landau et al., 2014; Li et al., 2014). In this setting, the composition of individual epialleles is only indirectly accounted for. Similar heterogeneity values could, indeed, come from different epiallele compositions. Of note, the methylation level of epialleles is usually not, or only partially, incorporated in these heterogeneity scores, which makes difficult to interpret the functional impact of heterogeneity shifts (Landan et al., 2012; Landau et al., 2014; Li et al., 2014; Xu et al., 2021; Zhang and Wang, 2022).

The underlying idea of our approach is that looking at the distribution of epialleles grouped by their methylation levels adds useful information for the functional interpretation of DNA methylation heterogeneity in a sample. The proposed approach, MC profiling, is indeed based on the empirical estimate of the distribution of epialleles grouped by their methylation levels (Figure 4C). We already applied the concept of MCs in previous works with the aim to model DNA methylation dynamics at targeted loci assayed through high-coverage bisulfite sequencing (Affinito et al., 2016; De Riso et al., 2020).

We here extended our approach to enrichment-based genome-wide datasets, like the ones from Reduced Representation Bisulfite Sequencing (RRBS) experiments, thus allowing for the simultaneous analysis of thousands of regions from the same sample. In this context, we implemented a new analytical framework to directly compare MC profiles across regions and samples.

Instead of adopting a numerical index (as, for example,the Shannon Index) to summarize the DNA methylation heterogeneity of a given region, we kept as much information as possible and described, for each DNA region, the relative abundance of the possible MCs. In this setting, we adopted the direct

comparison of MC profiles to analyze differential methylation of a given region among conditions, or to examine the differences among regions in the same condition (Figure 4C). It is important to point out that, in this latter setting, direct comparison of epiallele composition would only be possible through MCs, being these sequence independent, and not through the epialleles themselves.

Comparing MC profiles allowed us indeed to compare not only the heterogeneity but also the different methylation levels of DNA molecules.

In summary, adopting MC profiles can provide the following advantages:

- Considering how they are computed, MC profiles directly incorporate the average methylation of a given region, and inform on how it is contributed by single DNA molecules.

- MC profiles retain all information from a pool of molecules, and enable the direct visualization of DNA methylation heterogeneity of a given region

- MC profiles are empirically estimated from sequencing reads, and are independent on a priori parametrization of DNA methylation dynamics.

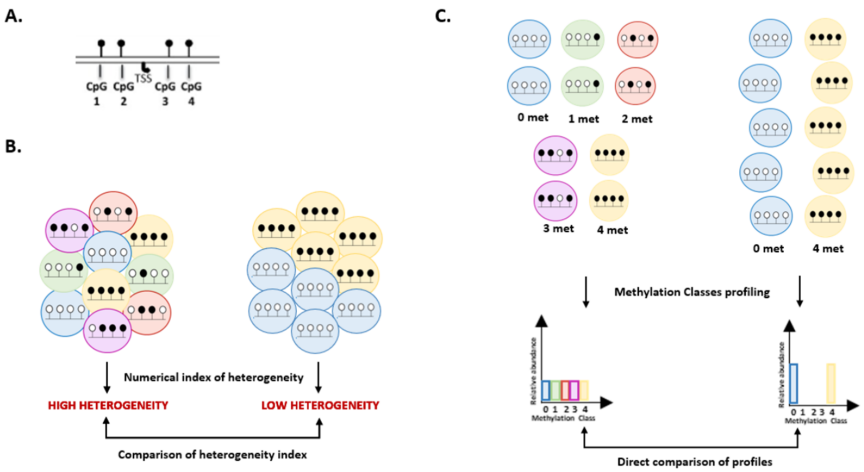


Figure 4: rationale of MC profiling. A: example of a region of interest holding 4 CpGs (TSS=Transcription Starting Site) B: Representation of epiallele-based analysis. DNA methylation heterogeneity for a certain locus is usually quantified through a numerical index (e.g., epipolymorphysm, Shannon entropy, ..). This can be then adopted to compare the heterogeneity of pools of

molecules (for example, to compare the heterogeneity of a certain locus in different samples). C: Representation of MC profiling analysis. The epialleles are first grouped in Methylation Classes (MCs) according to the number of methylated cytosines. The relative abundances of the possible MCs (for a locus holding n CpGs there are n+1 possible MCs), named MC profile, summarize the molecular heterogeneity and the methylation levels of a given region. MC profiles can be directly adopted to perform differential analysis.

4.1.2 Establishment of MC profiling's thresholds

To extend MC profiling to genome-wide datasets, we had to account for two main technical issues:

a) the shortness of reads limiting the number of CpGs that can be assayed in the same DNA molecule

b) the coverage of target regions, limiting the number of DNA molecules that can be analyzed to capture DNA methylation heterogeneity

Regarding point a), a previous study systematically addressed the decrease in the number of informative regions as a function of the number of CpGs to be included in the target region, and highlighted that increasing the number of considered CpGs from 4 to 5 causes a drop of the number of analyzable target regions. Based on these results, we fixed the number of CpGs to 4.

Regarding point b, we didn't find guidelines from previous studies that could fit the specific nature of our analysis. In this context, we reasoned that we could use deep amplicon bisulfite sequencing (D-ABS) data from our previous studies to assess the impact of coverage on MC profiles estimates. We therefore simulated 4-CpG low coverage datasets, starting from an in-house database of D-ABS amplicons. Detailed descriptions of the employed amplicons can be found in Table 2.

Table2: Details about D-ABS data adopted for data simulations

Gene	Organism	Amplicon Coordinates	Genome Assembly
DAO	Human	CHR12:108879926-108880252	GRCh38/hg38
DDOH	Human	CHR6: 110415392-110415789	GRCh38/hg38
SCRN1	Human	CHR7:29990018-29990346	GRCh38/hg38
CDKL5	Mouse	CHRX: 160994844-160994655	GRCm38/mm10
DDO_R3	Mouse	CHR10:40629085-40629513	GRCm38/mm10
DDO_R4	Mouse	CHR10:40629544-40629949	GRCm38/mm10
DDO_R6	Mouse	CHR10:40630278-40630682	GRCm38/mm10
DDO_R7	Mouse	CHR10:40630812-40631211	GRCm38/mm10
DLX6	Mouse	CHR6:6864874-6865260	GRCm38/mm10
TPH1a	Zebrafish	CHR25:8159799-8160116	GRCz11/danRer11

(human_DDO: human D-Aspartate Oxidase, mouse_DDOR4: mouse D-Aspartate Oxidase Region 4, mouse_DDOR6: mouse D-Aspartate Oxidase Region 6, mouse_DDOR7: mouse D-Aspartate Oxidase Region 7)

First, we split each amplicon into non-overlapping regions made up of 4 CpGs, thus obtaining several 4-CpG high-coverage datasets. Among these datasets, we selected those with higher coverage (number of reads > 20000). Since we expect that fully methylated or unmethylated profiles would be better captured at low coverage than intermediately methylated ones, due to the higher number of methylation classes with non-zero abundance, we selected 4-CpG datasets with average methylation levels spanning the entire range from 0 to 1 and enriched for datasets with intermediate average methylation. In this way, we selected 25 datasets, representative of 5 groups according to the average methylation level.

To simulate low coverage datasets, we randomly sampled a fixed number of reads from each 4-CpG dataset.

We simulated low coverage datasets to address the following issues:

a) the minimum coverage to minimize the error between a reference MC profile (i.e., the MC profile computed from the high coverage dataset) and an estimated MC profile (i.e., the MC profile computed from a low coverage dataset)

To address this point, we synthesized 1000 low coverage 4 CpG datasets for coverage values ranging from 20 to 200. From each dataset, we calculated the MC profile, and computed the JSD from the MC profile of the respective 4 CpG high-coverage dataset. As shown in Figure 5A, the JSD values decreased as the coverage increased, as expected. In particular, JSD values dropped between 25 and 50 reads (Figure 5A). A similar gain in accuracy is achieved by triplicating the coverage (i.e. achieving a read number higher than 150). Based on these observations, we considered a region covered by at least 50 reads to be eligible for MC profiling.

b) the minimum value of JSD to consider 2 MC profiles as different.

To address this point, we simulated 1000 pairs of low-coverage datasets with a fixed coverage of 50 reads. For each dataset pair, we computed the JSD among the estimated MC profiles. Ideally, two read groups sampled from the same dataset should exhibit very similar, if not identical, profiles, with a JSD value approaching 0. In practice, however, the estimated profiles differed to a certain extent. As shown in Figure 5B, at a coverage of 50 reads, MC profiles exhibited a JSD lower than 0.26 (min=0.22, max=0.28) for 95% of the experiments. JSD values observed for a wider range of coverage are reported in Table 3.

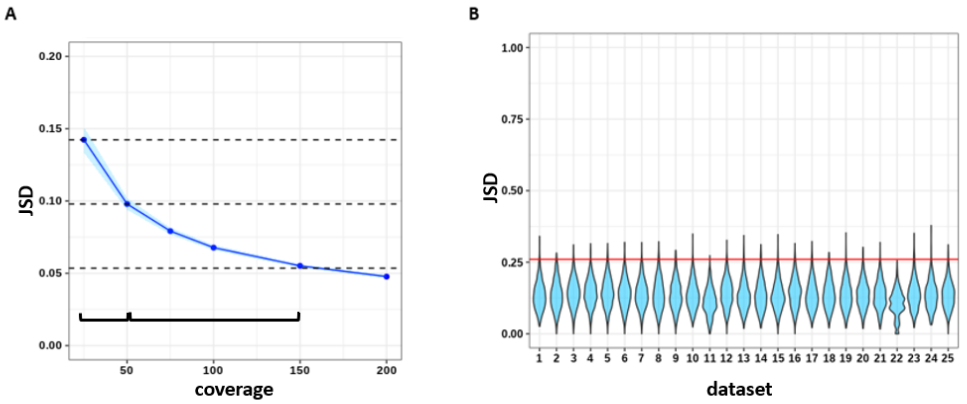


Figure 5: Results from data simulation. A: Accuracy of MC profiles from simulated datasets with increased coverage (y-axis: average JSD value of the MC profiles estimated between 1000 low-coverage 4-CpG datasets and the MC profile computed from the high-coverage 4-CpG dataset. x-axis: number of reads to simulate low coverage datasets. Dashed lines: gain in accuracy when increasing the coverage between 25 and 50 reads. Solid lines: interval of increased coverage to obtain the gain in accuracy observed between 25 and 50 reads). The shaded area indicates the standard deviation of the observed JSD value at a given coverage. B: Precision of MC profiles estimated from simulated 50 reads datasets for each 4-CpG high coverage dataset (y-axis: JSD values between MC profiles estimated from 1000 datasets' pairs; x-axis: 4-CpG datasets).

Table 3: **Jensen-Shannon distance cutoff as a function of coverage**

coverage	cutoff
25	0.37
50	0.26
75	0.21
100	0.18
150	0.15
200	0.13

We concluded that MC profiles having a JSD higher than 0.26 could be defined as different with an error equal or lower than 0.05. Hence, when comparing two MC profiles, we considered them to be different when we observed a JSD above 0.26.

4.1.3 Extraction of MC profiles from genome-wide experiments

The procedure to obtain MC profiles starting from aligned reads produced from bisulfite sequencing experiments is depicted in Figure 6.

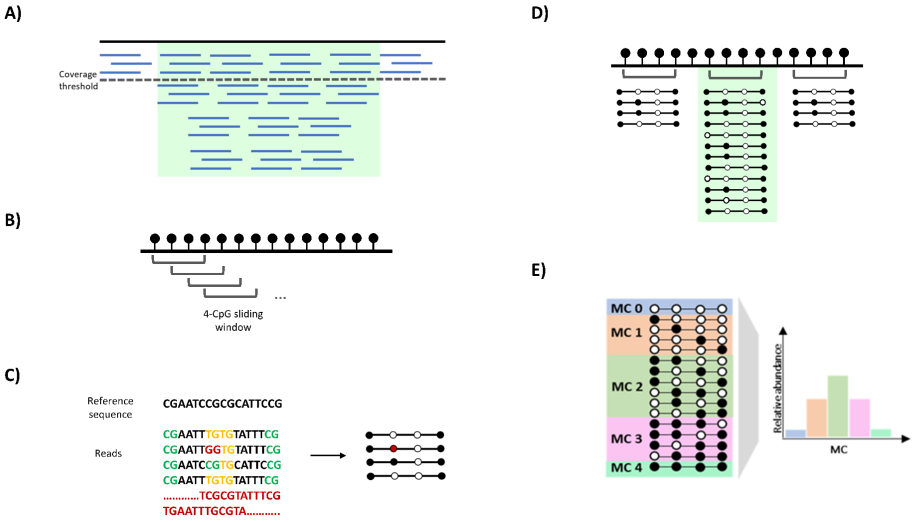


Figure 6. Epiallele extraction and MC profiling. A. Selection of genomic blocks, i.e. contiguous positions above a user-defined coverage threshold B. Selection of target regions for epiallele analysis. In this study, we focus on regions made up of 4 CpG sites, referred to as epiloci C. Extraction of epiallele table. For each epilocus, the sequence of individual reads is compared with the reference genome at the position of the CpG sites, taking into account the strand of the aligned read. If a T is found on the forward strand or an A on the reverse strand in correspondence of the reference C position, then the corresponding position in the matrix is marked as 0 (unmethylated). If instead a C is found on the forward strand or a G on the reverse strand in correspondence of the reference C position, then the corresponding position in the matrix is marked as 1 (methylated). In this way, a nCs (number of cytosines) x nRs (number of reads) table is obtained. Only information from reads completely spanning the 4 CpG sites of an epilocus is retained. Moreover, reads with ambiguous calls in correspondence of CpG positions (i.e., reads for which neither a T or C on the forward strand or an A or G on the reverse strand is found in the correspondence of a CpG site) are filtered out from the epiallele table. D. Selection of epiloci eligible for MC profiling. These are not overlapping epiloci for which the epiallele configuration can be assessed for at least 50 reads E. MC profiling. For a given epilocus, the MC profile is computed as the fraction of reads supporting the possible MCs (i.e., groups of epialleles bearing a given number of methylated cytosines, independently of the position) out of the total number of reads.

An epilocus is defined as a genomic region holding 4 CpGs.

MC profiles are built using the functionalities of *EpistatProfiler*, an R package we recently developed to extract and analyze information on epiallele composition from genome-wide experiments (Sarnataro et al., 2022).

EpiStatProfiler provides a set of functions that allow genome-wide analysis of epialleles composition at thousands of genomic regions that fulfill user-defined criteria.

First, information on aligned reads stored in a bam file is used to compute the genome-wide coverage at single-base resolution. The contiguous regions satisfying a user-defined coverage threshold are then merged to generate a collection of genomic blocks (Figure 6A).

These blocks are subsequently partitioned to build a set of smaller genomic regions constituted by a contiguous set of covered sites meeting user-defined criteria. These can be either the usage of sliding windows of variable length containing a user-defined number of CpG sites or the usage of a sliding window with user-defined fixed length and step sizes, containing a variable number of CpG sites. For the aims of MC profiling, we adopted a sliding window with a fixed number of 4 CpGs for target region selection, that we will refer to as epiloci in the text (Figure 6B).

For each epilocus to be profiled, epialleles composition is extracted by *EpiStatProfiler* in the following manner (Figure 6C).

First, all the sequenced reads mapping to the corresponding locus are selected and their sequence is compared with the reference genome at the position of the CpG sites, taking into account the strand of the aligned read.

Given the number of cytosines in the CpG context at the considered epilocus (n Cs) and the number of reads spanning the entire epilocus (n Rs), a nCs x nRs matrix - composed of n Cs columns and n Reads rows - is compiled.

For each read and each CpG, if a T is found on the forward strand or an A on the reverse strand in correspondence of the reference C position, then the corresponding position in the matrix is marked as 0 (unmethylated). If instead a C is found on the forward strand or a G on the reverse strand in correspondence of the reference C position, then the corresponding position in the matrix is marked as 1 (methylated).

To account for ambiguous methylation calls due to polymorphisms or sequencing errors, if neither a T or C on the forward strand or an A or G on the reverse strand is found in the correspondence of a CpG site, the corresponding cell is filled with the value of 2. Rows harboring these values (i.e. reads with ambiguous calls in correspondence of CpG positions) are filtered out when computing summary metrics and epiallele composition.

At the end, two different outputs are obtained starting from the 0-1 epiallele matrix related to a given epilocus.

The first one is a compressed table of epialleles. To build this output, each row of the binary epiallele matrix is converted to a string that represents one epiallele species. The compressed table is then created by reporting the count for each epiallele species.

The second output is obtained by applying a customizable set of functions that take the binary matrix as input and compute a given summary statistic. A data frame that contains the computed summary statistics is then generated. Available summary statistics include, among the others, the average DNA methylation of the epilocus, computed . The number and the type of functions to compute summary statistics over the epiallele binary matrix can be easily changed and extended with user defined functions.

Building on top of the compressed epiallele table output by EpiStatProfiler, we used in-house R scripts to perform MC profiling. These scripts have been made available in Zenodo (https://doi.org/10.5281/zenodo.7414513).

The procedure to obtain MC profiles for a target epilocus is depicted in Figure 6D-E.

We first selected the epiloci covered by at least 50 reads (Figure 6D). Then, for each epilocus, we grouped the observed epialleles into 5 Methylation Classes (MCs) according to the number of methylated cytosines. For a 4 CpG-locus, 5 MCs can be described.

Finally, we computed the relative abundance of each MC by summing the counts of epialleles belonging to the given MC out of the total number of analyzed DNA molecules, thus obtaining the MC profile of the considered epilocus (Figure 6E).

4.1.4 Analytical framework of MC profiling

The analytical framework of MC profiling is summarized in Figure 7.

Throughout this study, we adopted the Jensen Shannon Distance (JSD) to quantify the dissimilarity between MC profiles (see Methods). Based on the results of simulations performed on high-coverage targeted bisulfite sequencing data, we considered two profiles to be different when we observed a JSD above 0.26 (see Methods).

The JSD can be used to assess the changes of MC profiles at a given epilocus in different conditions. When different samples were available for

each condition, we computed an average MC profile by averaging the relative abundance of each MC among the samples assigned to a given condition. In this way, we obtained a consensus MC profile representative of all samples in a given condition. This enabled us to directly compare the consensus profiles of an epilocus in two conditions through the JSD.

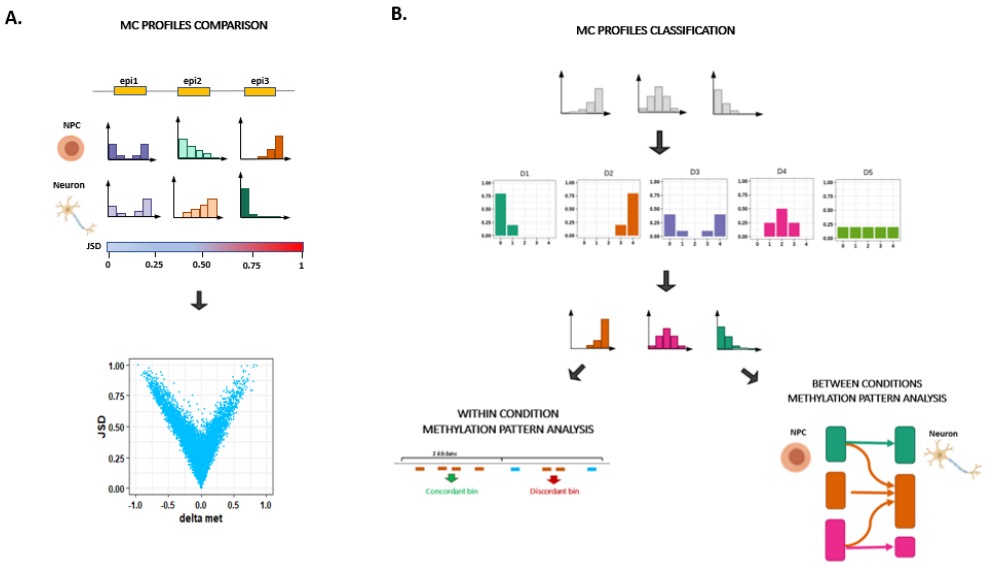


Figure 7: schematic drawing of MC profiling. A. The Jensen-Shannon distance is used to quantify the degree of dissimilarity between MC profiles. Based on the results obtained from simulated data, we considered two MC profiles to be different when observing a JSD above 0.26. The JSD can be used to assess the changes of MC profiles at a given epilocus in different conditions. The JSD can be also compared to other metrics, such as the difference of average methylation (delta met), over the analyzed epiloci. B. MC profiles were assigned to 5 Methylation Patterns (MPs) according to the most similar among 5 archetypal profiles (here indicated in the upper panel, middle row). This data compression procedure provided us with a signature of genome-wide MC profiles composition in a given condition. MPs enabled us to i) directly compare the MP of different epiloci within the same sample/ or condition (within sample analysis) and ii) to compare the MP transitions occurring at a given epilocus in different conditions (between conditions analysis).

The JSD can be also compared to other metrics, such as the difference of average methylation (delta methylation), over the analyzed epiloci.

To improve the interpretability of the data, we adopted a data compression procedure, and assigned each MC profile to a Methylation Pattern

(MP) according to the most similar of 5 archetypal profiles (Figure 7D), hereafter referred to as prototypes (see Methods). The prototypes, which are reminiscent of standard discrete distributions, were chosen because they reflect the reasonable profiles of an epilocus expected at a given methylation amount. In fact, D1 and D2 represent the two symmetric profiles for highly methylated or unmethylated epiloci, in which we expect a prevalence of fully unmethylated and methylated MCs, respectively. D3, D4 and D5, instead, represent the hypothetical profiles of intermediately methylated regions, that can reflect 1) the prevalence of both fully methylated and unmethylated MCs (D3, bimodal profile), 2) the prevalence of intermediately methylated MCs (bell-shaped profile, D4), or 3) the presence of all possible MCs with the same relative abundance (uniform profile, D5).

This data compression procedure provided us with a signature of genome-wide MC profiles composition in a given condition. MPs enabled us to i) directly compare the MP of different epiloci within the same sample or condition (within sample analysis) and ii) to compare the MP transitions occurring at a given epilocus in different conditions (between conditions analysis).

4.2 MC profiles are mostly stable among individuals and across genomic regions

We first applied MC profiling to 2 datasets of samples publicly available in GEO (see Table 1). Dataset1 included samples from 3 wild-type mice embryos, whereas Dataset2 included 3 samples from human CD19+ B-cells isolated from normal controls. Indeed, our datasets came from different species and were representative of different developmental stages, where we expect that DNA methylation heterogeneity probably derives from different dynamics (epigenetic drift in somatic cells vs cell differentiation in mouse embryos). We reasoned that such an experimental plan would have enabled us to generalize the results of our analysis.

For each sample, we profiled about 100000 epiloci in Dataset 1 and 90000 epiloci in Dataset 2.

By examining the average methylation of epiloci belonging to different MPs, we confirmed that the quantitative amount of methylated cytosines of assigned elements was coherent with the expected values for each pattern (Figure 8A).

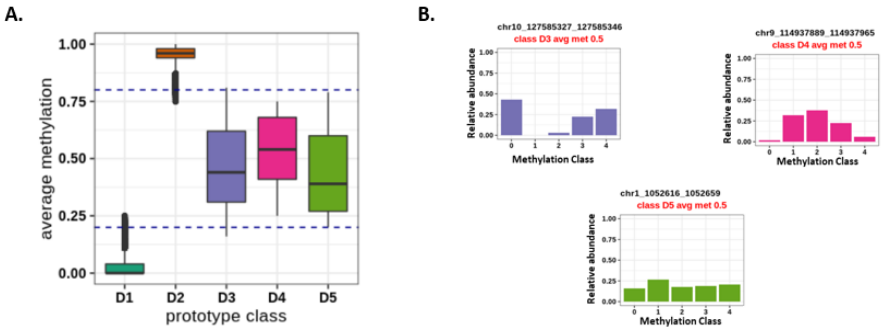


Figure 8. MC profiles versus average DNA methylation. A: Average methylation level of epiloci assigned to each MP in a sample from Dataset 1. B: example of epiloci with same average methylation and different MC profiles.

However, MC profiles add further information depicting the heterogeneity of DNA methylation among DNA molecules. This was particularly evident for the D3, D4, and D5 patterns. In fact, epiloci exhibiting the same average methylation were assigned to different MPs (Figure 8B).

We then investigated the stability of MC profiles across samples. For this aim, we analyzed the epiloci for which the MC profiles were assessed in all the samples in the individual datasets (n=87457 and n=41609) and computed the JSD of MC profiles among sample pairs. We found that 98% of epiloci in Dataset 1 and 96% in Dataset 2 had JSD lower or equal to 0.26 in all sample pairs, meaning that MC profiles at most of the epiloci were very similar between samples (Figure 9A).

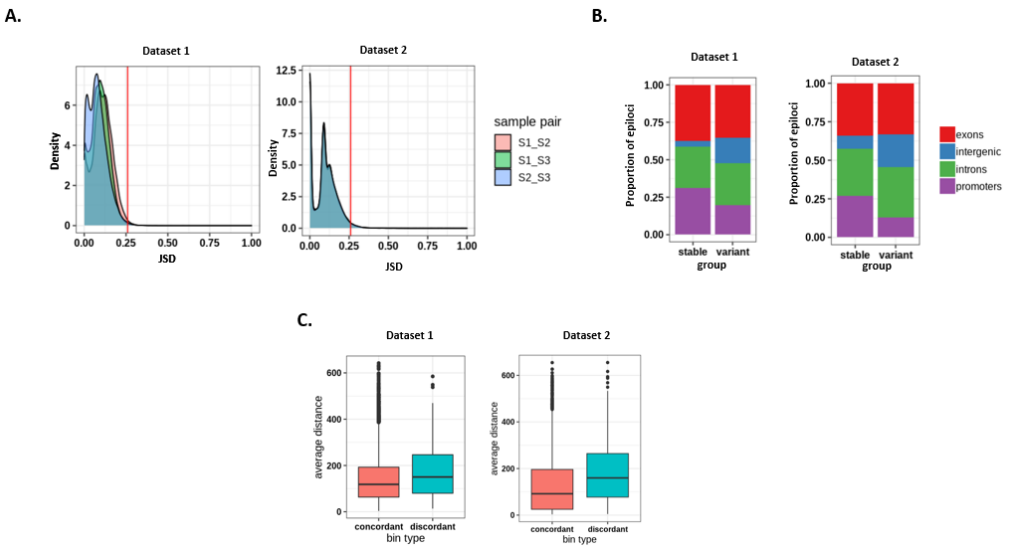


Figure 9. MC profiles are mostly stable among individuals and across genomic regions. A. Density plot of MC profile distance between sample pairs of Dataset 1 and 2. x-axis: JSD values between sample pairs. y-axis:density of epiloci with a given sample-pairs JSD value The red line indicates the cutoff value of JSD. B: Genomic annotation of epiloci with stable or variant MC profiles. C: average distance between epiloci inside concordant and discordant bins in dataset 1 and 2.

In both datasets, we found that stable epiloci, i.e. epiloci with a JSD below the cutoff in all sample pairs (n=86319 and n=39767, in Dataset1 and 2 respectively), were enriched in promoters (chi-square post hoc test p-values < 1e-7) and depleted in intergenic regions (chi-square post hoc test p-values < 1e-7). On the contrary, variant epiloci, i.e. epiloci with JSD above the cutoff in at least one sample pair (n= 1138 and n=1842, in Dataset 1 and 2 respectively), were depleted in promoters (chi-square post hoc test p-values < 1e-7) and were enriched in intergenic regions (chi-square post hoc test p-values < 1e-7). We found no difference in the proportion of stable and variant epiloci located in coding sequences (Figure 9B).

Based on this result, for each dataset we retained for further analysis the stable epiloci, and computed the consensus MC profile by averaging the relative abundances of each MC from the three samples. We then applied the data compression procedure, and assigned the consensus MC profiles to the MPs (see Methods).

Since epigenetic modifications are expected to involve larger DNA regions than individual epiloci, we expected that neighboring epiloci exhibited

concordant MC profiles. To test this hypothesis, we binned the genome in 1 kb regions, and compared the MPs of epiloci located in each bin (see Methods).

Among the bins harboring at least 3 epiloci, 10733 (99%) bore concordant and (1%) 148 bore discordant epiloci in Dataset 1, whereas (95%) 5079 bore concordant and (5%) 249 bore discordant epiloci Dataset 2. We confirmed that the number of bins bearing concordant epiloci significantly differed from the one expected by chance in both datasets (see Methods). This result suggests that MC profiles of neighboring epiloci tend to be similar. This conclusion is further supported by the observation that the reciprocal distance between epiloci in concordant bins tends to be lower than in discordant bins (Mann-Whitney p-value = 0.0005866, Figure 9C).

Overall, MC profiles resulted to be mostly stable among individuals and across genomic regions, thus suggesting that the heterogeneity captured by MC profiles mostly results from controlled DNA methylation dynamics, rather than from stochastic fluctuations of methylation levels.

4.3 MC profiles differentiate functional genomic regions

We reasoned that assigning MC profiles to different MPs could provide us with a signature of genome-wide MC profiles composition in a given dataset. We indeed examined the proportion of epiloci assigned to each MPs.

In accordance with the well-established bimodal distribution of average DNA methylation (Bock, 2012), the most represented prototype classes were D1 (83% and 78% of epiloci in Dataset 1 and 2, respectively) and D2 (about 15% and 16% of epiloci in Dataset 1 and 2, respectively).

The intermediately methylated D3, D4 and D5 classes accounted respectively for 2% of epiloci in Dataset 1 and 5% of epiloci in Dataset 2. Among the intermediately methylated classes, the most represented one was the D5 (90% and 82% of epiloci in Dataset 1 and 2, respectively), followed by the D3 class (9% and 13% of epiloci in Dataset 1 and 2, respectively). The D4 class was strongly underrepresented (1% and 4% of intermediately methylated epiloci in Dataset 1 and 2, respectively) in normal conditions, suggesting that intermediate values of average methylation rarely reflect an intermediate methylation amount on different DNA molecules. Instead, intermediate values of average methylation more often reflected the coexistence of fully unmethylated and fully methylated molecules, in presence (D5) or in absence (D3) of intermediately methylated molecules.

The classification of MC profiles to MPs also enabled us to investigate whether epiloci attributed to the different prototype classes were located in genomic regions with different functional characteristics (Figure 10A).

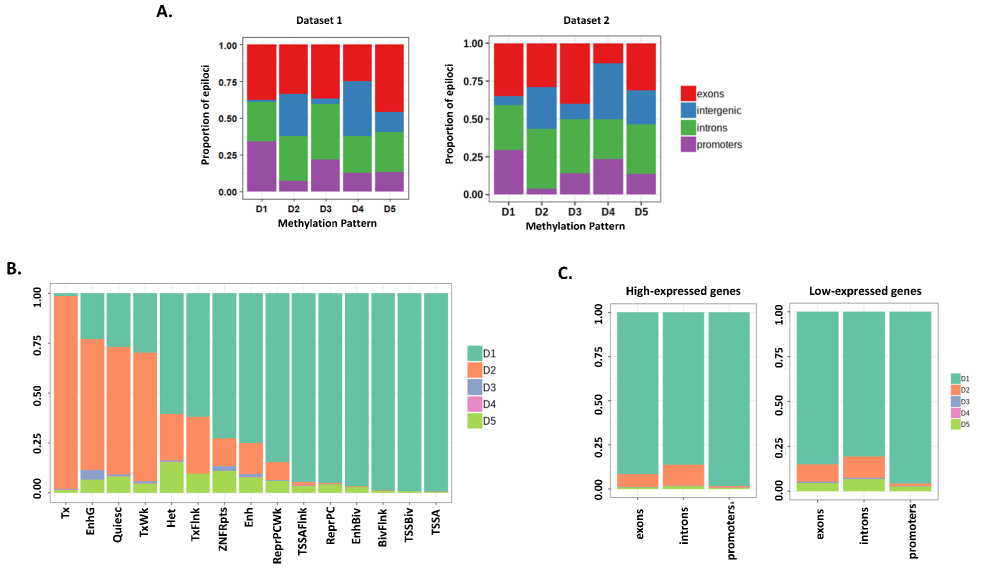


Figure 10. MC profiles differentiate functional genomic regions. A. Genomic annotation of epiloci assigned to the different MPs in Dataset 1 and 2. B. Proportion of MPs in Dataset 2 for regions assigned to different functional categories according to the chromHMM track for GM12878 cell line (Tx= Strong Transcription, EnhG= Genic enhancer, Quies= Quiescent/Low, TxWk= Weak transcription, Het= Heterochromatin, TxFlnk= transcription at gene 5' and 3', ZNFRpts= ZNF genes and repeats, Enh= enhancer, ReprPCWk= Weak Repressed Polycomb, TSSAFlnk= Flanking active TSS, ReprPC=Repressed Polycomb, EnhBiv= Bivalent enhancer, BivFlnk=Flanking bivalent TSS/enhancer, TssBiv= Bivalent/Poised TSS, TssA= Active TSS). C. Proportion of MPs in functional regions (promoters, introns and exons) of highly expressed and lowly expressed genes.

We found that the D1 class was enriched within promoters and exons (chi-square post hoc p-values < 1e-7) and depleted in intergenic regions and introns (chi-square post hoc p-values < 1e-7). On the contrary, the D2 class was mainly located in intergenic regions and introns (chi-square post hoc p-value< 1e-7) and depleted in promoters and exons (chi-square post hoc p-value< 1e-7). Similarly, the D5 class was depleted from promoters and enriched in intergenic regions (chi-square post hoc p-values < 1e-7). We did not find significant differences in the localization of D3 and D4 epiloci.

We found that MPs composition could further distinguish genomic regions decorated with different histone marks in Dataset 2 (Figure 10B). For example, MPs separated constitutive heterochromatin from Polycomb-repressed regions (chi-square p-value < 1e-7), with the former

enriched not only for the methylated D2 but also for the D5 MP (chi-square post hoc p-value< 1e-7), pointing to higher heterogeneity in constitutively inactive genomic regions. Polycomb-repressed regions, on the other hand, were enriched for the D1 MP (chi-square post hoc p-value< 1e-7), pointing to lower levels of DNA methylation in Polycomb-regulated regions.

We also found that MPs composition varied when separately investigating the promoter, exonic and intronic regions of genes with expression levels lower or higher than the median value in Dataset 2 (chi-square p-value < 1e-7, Figure 10C). We found that D1 MP was enriched in promoters, introns and exons of highly-expressed genes (chi-square post-hoc p-value < 1e-7), whereas the D2 MP was enriched in exons (chi-square post-hoc p-value < 1e-7) and slightly enriched in promoters of lowly-expressed genes (chi-square post-hoc p-value < 5e-3). Again, we found an enrichment of the D5 MP in promoters, exons and introns of lowly expressed genes (chi-square post-hoc p-value < 1e-7), suggesting a consistent pattern of increased heterogeneity in low-to-inactive regions.

4.4 MC profiling individuates a signature of imprinted regions and X chromosome inactivation

We tested the capability of MC profiling to discriminate regions undergoing genomic imprinting, a well-known phenomenon of allele specific regulation. In these regions, it is expected that the two alleles differ for their DNA methylation status. Hence, we wondered whether D3 epiloci, in which two pools of molecules exist with opposite DNA methylation status, were enriched at genomic regions flanking imprinted genes.

To test this hypothesis, we assigned each epilocus of Dataset 1 and 2 to its nearest gene (see Methods) and marked the epiloci as associated with imprinted genes if the closest gene was enlisted in Geneimprint (https://www.geneimprint.com/).

As shown in Figure 11A, the five MPs were differentially represented among epiloci flanking imprinted and not imprinted genes (chi-square test p-values < 2.2e-16). Specifically, epiloci assigned to the D3 pattern were strongly overrepresented among epiloci flanking imprinted genes (chi square post-hoc test p-values < 2e-16), thus confirming that D3 epiloci were preferentially, even though not exclusively, associated with allele specific methylation.

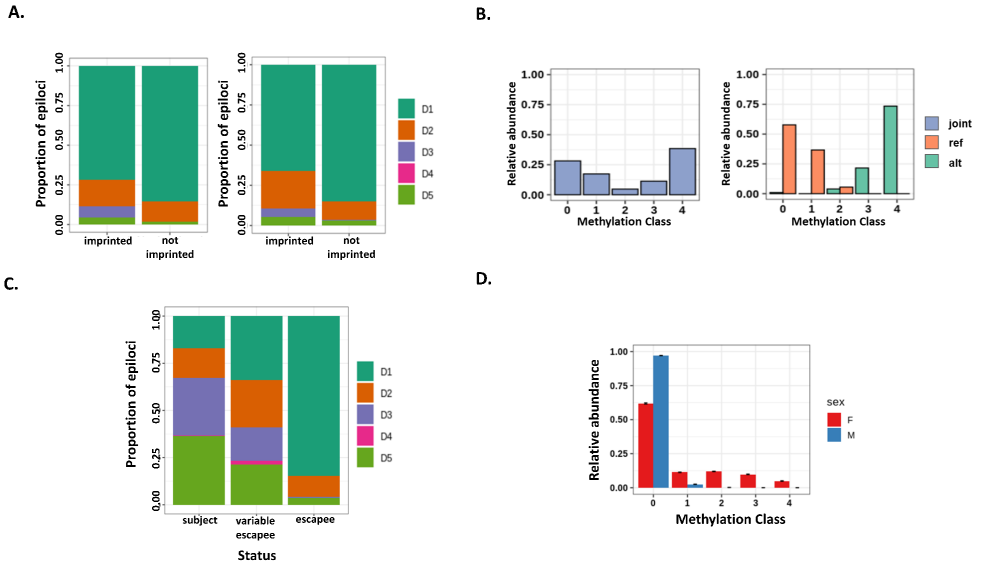


Figure 11: MC profiling of imprinted and X-inactivated regions. A: proportion of epiloci assigned to the different MPs in imprinted and non imprinted genomic regions in Dataset 1 and 2. B: example of bimodal epilocus flanking the Zdbf2 imprinted gene in Dataset 3. The joint MC profile (i.e. the profile obtained without splitting the alleles) is shown in light blue, whereas the profiles of the reference (ref) and the alternative (all) alleles are shown in orange and green respectively. C: classification of epiloci flanking genes undergoing X inactivation (subject), stably escaping X inactivation (escapee), or variably escaping X inactivation (variable escapee); D: average MC profile of epiloci classified as D1 in a male sample from Dataset 1(blue) compared to the average profile of the same epiloci in two female samples from the same dataset (red).

As a confirmatory experiment, we searched for D3 epiloci flanking imprinted genes in a third dataset of DNA methylation data from mice born from two different strains (Dataset 3 in Table 1). Based on known polymorphic sites between the two strains, we were able to attribute each read to the respective allele, and to explore the allele specific MC profile for more than 300 autosomal epiloci in 3 mice (see Methods).

Due to the scarce coverage of imprinted regions also stated in the original study (Orozco et al., 2014), we could find a single epilocus exhibiting a D3 profile and located near imprinted genes. In particular, this epilocus was located on chr1, upstream of the Zdbf2 imprinted gene. In this locus, differentially methylated regions had been previously described (Hiura et al., 2010, p. 1). Figure 11B shows how the bimodal joint MC profile at this

epilocus results from different profiles on the two alleles, with one skewed towards complete demethylation and the other towards complete methylation. It is worth noting that, for both alleles, MC profiling individuated a certain degree of cellular heterogeneity, since intermediate MCs were also represented.

Based on the results obtained from MC profiling of imprinted genes, we decided to investigate whether epiloci located on the X chromosome also exhibited peculiar MC profiles due to the X inactivation process. It is in fact known that, during the inactivation of the X chromosome, most loci are inactivated (subject loci) while others partially or totally escape this inactivation (escapee or variable escapee loci) (Balaton et al., 2015).

We indeed analyzed the MPs to epiloci flanking genes with different inactivation status. First, we assigned X epiloci to the respective MP in two female samples from Dataset 2. Then, we assigned to each epilocus the consensus inactivation status of the nearest gene (Balaton et al., 2015). In this way, we classified 551 epiloci as subject to X chromosome inactivation, 138 as escapee, 56 as variable escapee and 233 as unknown/discordant.

As shown in Figure 11C, MPs were represented in different proportions among subject, escape and variable escape epiloci (chi square post-hoc test p-value < 1e-7). Escape epiloci mostly exhibited unmethylated D1 profiles (chi square post-hoc test p-value < 1e-7), whereas subject epiloci mostly exhibited either bimodal D3 or uniform D5 profiles (chi square test post-hoc p-value < 1e-7). Both groups of MPs (D1 and D3/D5) were represented among variable escape epiloci, none of them significantly enriched.

We hence decided to investigate the MC profile of the inactive X in Dataset 1, for which two female and one male sample was available. We reasoned that we could deduce the profile of the female inactive X by comparing the MC profile of X epiloci in males and females, and that such deduction would have been particularly feasible for epiloci classified as D1 in the male samples. In fact, in this condition, it could be reasonably inferred that methylated molecules in females mostly resemble the methylation status of the inactive X. We indeed compared the average profiles of 1068 epiloci belonging to the D1 MP in males with the respective average profile in female samples (Figure 11D).

The sex difference among the average MC profiles pointed to a heterogeneous DNA methylation status of the inactive X, ranging from being lowly to fully methylated in different cells. Of note, we observed a more gradual methylation status of the inactive X compared to the methylated alleles of imprinted epiloci. This observation is compatible with the previously described discrepancy of average methylation between imprinted and X inactivated genes. In fact, while for imprinted loci one of the alleles is fully methylated, X inactivated genes exhibit partial methylation of the inactive

allele (Balaton et al., 2021). In addition, MC profiles suggest that this partial methylation is due to cell-to-cell differences, and not to a partial methylation in all cells.

4.5 MC profiling individuates DNA methylation changes upon differentiation

We challenged the ability of MC profiling to capture epigenetic changes among conditions. As a model of epigenetic changes, we choose a dataset of neuronal differentiation. To this aim, we analyzed MC profiles changes of 115608 epiloci upon differentiation of hippocampal precursors (HP) to granule cells (GC) (Dataset 4).

For each epilocus, we calculated the difference of average methylation between differentiated cells and neuronal precursors (delta meth), and quantified the MC profiles' change by using the Jensen-Shannon distance (JSD). The relationship between these two measures is shown in Figure 12A. The red lines delineate the difference of average methylation observed in 95% of the considered epiloci (0.14), and the black line indicates the JSD threshold (0.26).

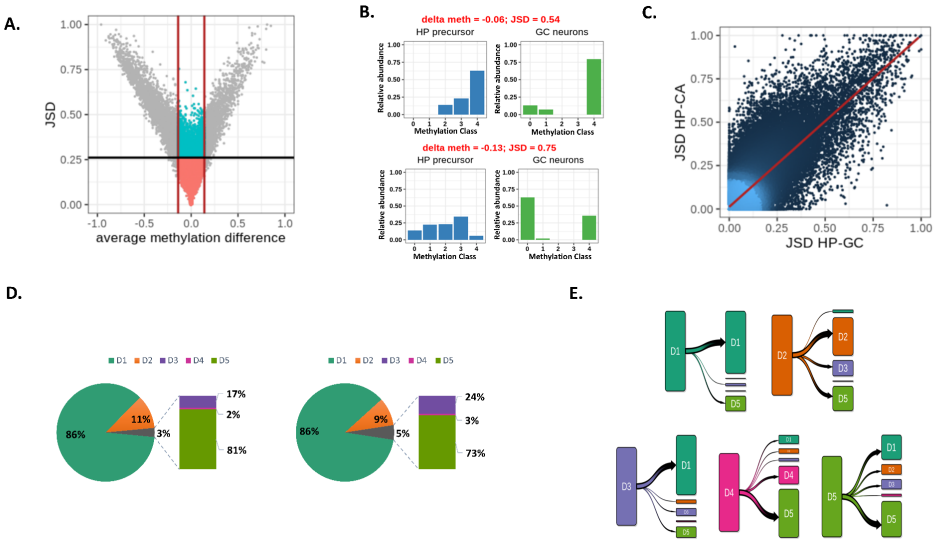


Figure 12: Application of MC profiling on neuronal differentiation. A: DNA methylation changes at 104720 epiloci upon differentiation of hippocampal precursors to granule cells. X-axis: average methylation change (delta met); y-axis: MC profile change (JSD). The black line indicates the JSD cutoff, whereas the red lines indicate the 95th percentile of observed delta values. B: examples of epiloci with low difference in average methylation but high JSD values between HP and GC MC profiles (epiloci coordinates: chr19:57700749-57700788 and chr1:186924297-186924337). C: Comparison of MC profile changes upon differentiation of hippocampal precursors (HP) to granule cells (GC) or CA3 neurons. x-axis: JSD values between MC profiles in HP and GC. y-axis: JSD values between MC profiles in HP and CA. D: MPs composition of hippocampal precursors (on the left) and granule cells (on the right) samples. E: Transition plot of variant epiloci in the HP-GC pair. For the epiloci assigned to the different MPs in HP cells the classification in differentiated GC neurons is shown.

As expected, MC profiles' and average methylation changes were mostly correlated. This relationship strengthened as differences in average methylation approached the maximum, consistent with the fact that huge differences in the amount of methylated cytosines are expected to affect both average methylation and MC profiles. Symmetrically, the relationship between average methylation and MC profiles' changes weakened for lower values of average DNA methylation and was almost lost below 0.14. In this range, despite a large number of epiloci exhibiting stable MC profiles (n= 104720), a group of 5129 epiloci exhibited significant changes in MC profiles upon differentiation (blue dots in Figure 12A). As examples, Figure 12B shows two epiloci with significant changes of MC profiles and little variation of average DNA methylation. This result suggests that, at these epiloci, MC profiles were remodeled without a significant gain or loss of overall DNA methylation.

To test the association between changes in MC profiles and the process of neuronal differentiation, we checked the consistency of the MC profiles changes upon differentiation of the same precursor in a different type of neuron. Notably, we found a high correlation between MC profiles changes for the 97119 epiloci examined upon differentiation of hippocampal precursors to granule cells or CA neurons (Pearson R 0.81, Figure 12C), according to the previously described high similarity among these differentiation processes (Sharma et al., 2016).

To further establish the relationship between the changes in MC profiles and epigenetic remodeling upon cell differentiation, we explored the chromatin landscape, summarized by chromHMM labels, associated with the analyzed epiloci. We observed that MC profile changes more probably involved epiloci located in regions that also underwent chromatin changes upon differentiation

(Fisher test p-value < 2.2 e-16). In fact, only 5% of epiloci located in genomic regions with stable chromatin marks underwent changes in their MC profile, whereas 22% of epiloci undergoing chromatin changes also changed their MC profile, thus suggesting that our approach was probably identifying loci undergoing epigenetic remodeling.

When investigating the genomic localization of developmentally variant epiloci discovered by our approach, we found that they were slightly depleted outside CpG islands and promoters (Fisher test p-values < 2.2 e-16) both in HP-GC and HP-CA transitions.

Being JSD is a symmetric distance, it only quantifies the dissimilarity between two MC profiles, but does not return the information on whether this dissimilarity corresponds to a gain or loss of DNA methylation. Thus, we turned to the analysis of prototype classes to qualitatively interpret MC profile changes upon differentiation. The prototype class composition for HP and GC is shown in Figure 12D.

First, we asked whether changes were occurring at epiloci exhibiting peculiar MC profiles in neural precursors. We found that epiloci classified as D1 remained mostly stable, whereas epiloci assigned to the other classes mostly changed their MC profile upon differentiation (chi-square post-hoc p-values < 1e-7).

We then analyzed the prototype class composition in differentiated neurons, and found a depletion of D2 epiloci and an increased fraction of D5 epiloci (chi-square post-hoc p-values < 1e-7), suggesting that the methylated status in differentiated neurons tends to be more heterogeneous among different cells.

Finally, to better characterize how MC profiles changes were occurring, we analyzed the prototype class transitions upon differentiation. In Figure 12E, for each prototype class in neuronal precursors, we show the final prototype class in differentiated neurons.

We noticed that for a consistent fraction of D1 and D2 epiloci, MC profiles' changes did not correspond to class transitions, meaning that these epiloci were shifting toward a higher or lower DNA methylation heterogeneity. We also noticed that a reduced fraction of epiloci evolved toward the D2 class upon differentiation. Interestingly, most of D3 epiloci evolved to lower methylation upon differentiation, transiting to the D1 class. Thanks to prototype class analysis, we could interpret this demethylation as a negative selection of the fully methylated molecules that were present in neural precursors.

All together, these results indicate that our approach well captures quantitative and qualitative DNA methylation changes upon neuronal

differentiation that might be underestimated or overlooked by an average methylation based approach.

5. Discussion

Each cell is unique. Evidence has accumulated that even in morphologically homogeneous cell populations extensive differences can be highlighted at multiple molecular levels, and that these differences are relevant to biological processes (Carter and Zhao, 2021).

Single-cell DNA methylation assays promise to be the standard technique to study DNA methylation heterogeneity in cell populations (Angermueller et al., 2016; Hui et al., 2018). However, single-cell DNA methylation technologies still generate very sparse data, in a limited number of cells per sample, and at high cost (Huan et al., 2018; Teschendorff et al., 2020). Alternatively, cell-to-cell differences can be deduced by studying the methylation patterns of consecutive cytosines in sequenced reads from bulk experiments, assuming each read coming from a single DNA molecule (Huan et al., 2018; Landan et al., 2012; Li et al., 2014; Scherer et al., 2020; Xu et al., 2021). This approach can stand comparison with single-cell assays when the goal is to obtain a statistical/robust description of DNA methylation of a genomic region in a cell population (Huan et al., 2018).

Building on top of our experience on deep targeted bisulfite sequencing, in this study we propose MC profiling as a genome-wide approach to the study of DNA methylation heterogeneity. Given an epilocus holding 4 CpG sites, we defined its MC profile as the ensemble of the relative abundances of molecules sharing an equal number of methylated cytosines (Methylation Classes, MCs). Such an approach, while incorporating information on the overall average methylation of a region, directly informs on the different methylation levels, and their abundance, observed in a pool of molecules. This information is usually not, or poorly, taken into account by other approaches, which directly quantify the degree of cellular heterogeneity through the analysis of individual arrangements of methylated cytosines in single DNA molecules (epialleles).

In previous studies, several scores have been developed to incorporate methylation level of individual molecules in the estimate of DNA methylation heterogeneity, thus improving prediction of gene expression levels, and correlation with chromatin marks compared to the overall average methylation (Landau et al., 2014; Shi et al., 2021; Xu et al., 2021). MC profiling is in line with this logic, but further enlarges the information on DNA methylation heterogeneity by considering molecules with different methylation levels as separate entities.

A conceptually similar approach to MC profiling has been proposed in (Abante et al., 2020; Jenkinson et al., 2018, 2017). In these studies, DNA methylation is expressed as the probability mass function (PMF) of methylation levels that could be observed in a pool of molecules, which

resemble the concept of our MCs. This approach, specifically designed to deal with the low coverage of WGBS experiments, has provided novel insights on DNA methylation heterogeneity and its disposition across the genome, its evolution upon differentiation, aging and cancer, and its relationship with the genetic background (Abante et al., 2020; Jenkinson et al., 2018, 2017). The biggest difference between this approach and MC profiling is that while the PMF is predicted from a mathematical model applied to DNA methylation data, the frequencies of MCs are empirically estimated from experimental data, thus avoiding time consuming model fitting and releasing the distribution of methylation from a-priori parametrization of DNA methylation dynamics.

To quantify the dissimilarity between MC profiles, we adopted the Jensen-Shannon distance (Lin, 1991). This dissimilarity measure has been applied in bioinformatics and epigenetics (Abante et al., 2020; Guo, 2020; Itzkovitz et al., 2010; Jenkinson et al., 2018, 2017; Kartal et al., 2020).

To set the parameters of our approach, we synthesized low coverage 4 CpG datasets from an in-house database of high-coverage amplicon bisulfite sequencing data (Affinito et al., 2020; Cuomo et al., 2021, 2019; Florio et al., 2017). In this context, we provided a systematic quantification of the impact of coverage on the accuracy of MC profiles, and estimated the expected error associated with MC profiles at a coverage of 50 reads. In our opinion, these results could serve as guidelines to orient qualitative analysis of DNA methylation in low coverage settings.

Here, similarly to previous studies (Jenkinson et al., 2018), we adopted a classification procedure, assigning each MC profile to the most similar among 5 reference profiles. This classification scheme provided us an interpretable representation of each MC profile. Furthermore, it provided us with a qualitative property to be compared across epiloci. Finally, being this a fixed scheme, we could apply it and directly compare the results on different conditions and species.

We demonstrated that MC profiles were stable among different samples and neighboring epiloci. Previous studies illustrated that DNA methylomes exhibit high inter-individual stability, especially in CG dense regions (Bock et al., 2008; Palumbo et al., 2018). Concordant epigenetic marks, including DNA methylation, across genomic blocks have been also described (Ernst and Kellis, 2017; Jenkinson et al., 2017; Zhang et al., 2017). Altogether, our results are in line with previously described patterns of regional and inter-individual stability of DNA methylation, and suggest that MC profiles capture controlled DNA methylation dynamics rather than stochastic fluctuations of methylation levels.

In this paper, we applied MC profiling to gain insights on methylation heterogeneity in various biological contexts.

Firstly, we tested the capability of MC profiling to inspect known examples of mono-allelic regulation, i.e. genomic imprinting and X-inactivation, in which DNA methylation is notably involved.

When we analyzed the MC profiles of epiloci located in proximity of known genomic imprinted regions, we found that bimodal MC profiles were overrepresented. This was expected, considering the known opposite methylation pattern of the two parental alleles at imprinted regions (Edwards and Ferguson-Smith, 2007). For an epilocus located upstream of the Zdbf2 gene, holding a polymorphic site, we were able to clearly show opposite MC profiles on the two alleles.

We then analyzed the MC profiles of epiloci located on the X chromosome in female samples. First, we compared MC profiles of epiloci flanking genes with reported differential inactivation status. Consistent with previous findings (Balaton and Brown, 2021; Cotton et al., 2015; NISC Comparative Sequencing Program et al., 2018), escapee epiloci showed homogeneous DNA methylation on both X copies, being unimodally fully methylated or unmethylated. Subject epiloci, on the contrary, were enriched for more heterogeneous MC profiles (D3 and D5), compatible with different DNA methylation status of the two alleles (Balaton and Brown, 2021; Cotton et al., 2015; NISC Comparative Sequencing Program et al., 2018).

To further inspect the DNA methylation status of the inactive X, we selected the X epiloci with a fully unmethylated profile in male samples, and examined the corresponding MC profiles in female samples to infer the profile of the inactive X. We showed a prevalence of intermediately methylated classes on the inactive X, accompanied by high cellular heterogeneity. Incomplete DNA methylation of the inactive X was described in (Balaton et al., 2021) at single CpG level, thus marking a difference between the X inactivation and the genomic imprinting processes that was well reflected in our analysis. It is worth noting that the prevalence of intermediately methylated MCs that we found with our approach also suggested a difference between the methylated status on the inactive X and at autosomal epiloci, suggestive of peculiar mechanisms intervening in DNA methylation establishment and regulation on the inactive X.

The methylation status of the inactive X appeared also to be highly heterogeneous among different cells. We speculate that this cellular epipolymorphysm could almost in part find its reflection in differences of X inactivation status between equivalent cells described in single-cell RNA-seq studies (Garieri et al., 2018; Keniry and Blewitt, 2018).

Finally, we applied MC profiling to the analysis of DNA methylation changes in different conditions. In particular, we examined profiles' changes upon differentiation, when epigenetic remodeling is expected to occur. We

adopted the Jensen-Shannon distance to capture epiloci with significant differences in MC profiles between neural precursors and differentiated neurons. Being JSD a symmetric distance measure, it didn't return the information on whether MC profiles changes correspond to gain or loss of DNA methylation. Thus, we examined the pattern transitions to gain insights on how profiles' changes were occurring. Combining the analysis of JSD and pattern transitions provided us a comprehensive picture of DNA methylation differences among conditions: in fact, we could distinguish profiles changes associated with unvaried patterns (and thus, with stable reciprocal proportion of DNA molecules with different methylation levels) from profiles changes accompanied with pattern transitions (which indicate a redistribution of the proportions of molecules with different methylation levels).

As expected, we found that MC profiles changes captured by JSD correlated with average DNA methylation gain or loss at most epiloci. However, we described MC profile changes at almost constant DNA methylation for more than 5000 epiloci. Qualitative DNA methylation changes occurring with little to no changes in overall average methylation were also described in (Abante et al., 2020; Jenkinson et al., 2018, 2017), indicating that such an approach can be even more informative than average methylation based approach in the analysis of dynamic systems.

Interestingly, we found that MC profile changes were enriched at CpG islands, which were described to be spared from most epigenetic changes in the original study (Sharma et al., 2016). The association that we found with changes of chromatin marks, as well as the concordance of MC profiles changes upon differentiation in two different neuronal subtypes, pointed to exclude random variations occurring at these epiloci. Instead, considering that most epiloci exhibited stable MPs in precursors and differentiated neurons, it is possible that MC profiling has captured changes in cellular heterogeneity that were overlooked by the average methylation-based approach.

Applying MC profiling to RRBS data can give insights on cellular epigenetic heterogeneity from plenty of already available datasets in public repositories. However, it strongly limits the analysis to CpG islands and immediately proximate regions (Bock et al., 2010; Gu et al., 2011). This limit is further exacerbated when selecting target regions harboring 4 CpGs (the epiloci of this study) shared among multiple samples. The required coverage of 50 reads strongly limits the applicability of the proposed approach outside Whole Genome Bisulfite Sequencing (WGBS) data. However, more unbiased enrichment assays have been developed which combine high throughput sequencing with selection of target regions through PCR or capture-based trapping that are natively less biased toward CG dense regions and could fit the coverage requirements of MC profiling (Kacmarczyk et al., 2018; Klobučar et al., 2020).

Despite these limitations, we here showed that MC profiling could effectively capture cellular differences and changes also in CG dense regions, which are usually reported to be resistant to DNA methylation in normal conditions (Edgar et al., 2014; Mohn et al., 2008). We indeed believe that applying MC profiling to these experiments could further extend our observations outside CG dense regions.

6. Conclusions

We have presented a novel approach named MC profiling aimed at exploring DNA methylation heterogeneity at multiple target regions in bulk bisulfite sequencing experiments. MC profiling is built on the concept of MCs, groups of molecules holding the same number of methylated cytosines. DNA methylation is indeed represented through MC profiles, i.e. the relative abundances of possible MCs for a given region.

MC profiles directly incorporate the average methylation of a given region, and inform on how it is contributed by single DNA molecules. Thus MC profiles offer a functional view of DNA methylation heterogeneity in a sample. MC profiles are empirically estimated from sequencing reads, and are independent on a priori parametrization of DNA methylation dynamics. Moreover, MC profiles retain all information from a pool of molecules, and enable the direct visualization of DNA methylation heterogeneity of a given region.

MC profiling led to the identification of signatures of loci undergoing genomic imprinting and X inactivation, and highlighted differences between the two processes. When applied to a dynamic system, MC profiling identified DNA methylation changes in regions with almost constant average methylation. Altogether, our results indicate that MC profiling can provide useful insights on the epigenetic status and its evolution at multiple genomic regions.

7. Acknowledgements

The greatest thanks go to my lovely family that kindly supported me during these three years.

Thanks to my mum Daniela, my sister Erminia and my brother Pasquale, for patiently understanding my breakdowns and my being-here-but-elsewhere moments. You will always be the root from which I get the energy to grow up and the stability of not falling down.

Thanks to Filippo, who shared my life and embraced my dreams and projects, even at the cost of many denials and absences. You show me each day what love is and how to take care of it.

None of the results I achieved during these three years would have been possible without the precious help of three people that I cannot define with other words than my “scientific family”.

Thanks to Michele Pinelli, with whom I have built a bond of mutual esteem and affection.

Thanks to Antonella Sarnataro, my sister-in-science, for sharing with me this journey, for receiving my confidences with sweetness and discretion, for warming me with her enthusiasm and optimism, for sharing with me some of the places-of-the-heart.

I especially thank Prof. Sergio Cocozza, for thousands of reasons that cannot be enclosed in a few sentences. In my fight against insecurities, you’ve always been the soldier in the first-line. I basically thank you for teaching me how to fill the gap between who I am and who I can and want to be, as a scientist and a human being.

A special thanks to Prof. Lorenzo Chiariotti and all his group for the fruitful collaborations and precious scientific interaction. I especially thank my colleague and friends Mariella Cuomo and Rosa Della Monica.

I also thank Prof. Gennaro Miele, Prof. Stefano Amente and Dr. Giovanni Scala for their significant contribution to this project.

Finally, I thank the students Daniele, Anna and Ermelinda for sharing with confidence a fundamental piece of their traineeship.

Each of these people have played a crucial part in making me who I am today. I hope that this little work will give you back at least a little part of what you did for me.

8. List of your publications

1. Affinito O, Palumbo D, Fierro A, Cuomo M, **De Riso G**, Monticelli A, Miele G, Chiariotti L, Cocozza S. Nucleotide distance influences co-methylation between nearby CpG sites. Genomics 2020. 112;1;144:150.

2. Russo C, Cocozza S, Riccio E, Pontillo G, Petruzzelli LA, Lanzillo R, Spinelli L, Colomba P, Duro G, Imbriaco M, Russo CV, **De Riso G**, Di Risi T, Tedeschi E, Cuocolo A, Brunetti A, Morra VB, Cocozza S, Pisani A. Prevalence of GLA gene mutations and polymorphisms in patients with multiple sclerosis: A cross-sectional study. J Neurol Sci. 2020 May 15;412:116782.

3. **De Riso G**, Cuomo M, Di Risi T, Della Monica R, Buonaiuto M, Costabile D, Pisani A, Cocozza S, Chiariotti L. Ultra-Deep DNA Methylation Analysis of X-Linked Genes: GLA and AR as Model Genes. Genes (Basel). 2020 Jun 4;11(6):620.

4. **De Riso G**, Fiorillo DFG, Fierro A, Cuomo M, Chiariotti L, Miele G, Cocozza S. Modeling DNA Methylation Profiles through a Dynamic Equilibrium between Methylation and Demethylation. Biomolecules. 2020 Sep 3;10(9):1271.

5. Cuomo M, Borrelli L, Della Monica R, Coretti L, **De Riso G**, D'Angelo Lancellotti di Durazzo L, Fioretti A, Lembo F, Dinan TG, Cryan JF, Cocozza S, Chiariotti L. DNA Methylation Profiles of Tph1A and BDNF in Gut and Brain of L. Rhamnosus-Treated Zebrafish. Biomolecules. 2021 Jan 22;11(2):142.

6. **De Riso G**, Cocozza S. Artificial Intelligence for Epigenetics: Towards Personalized Medicine. Curr Med Chem. 2021;28(32):6654-6674.

7. Della Monica R, Cuomo M, Visconti R, di Mauro A, Buonaiuto M, Costabile D, **De Riso G**, Di Risi T, Guadagno E, Tafuto R, Lamia S, Ottaiano A, Cappabianca P, Del Basso de Caro ML, Tatangelo F, Hench J, Frank S, Tafuto S, Chiariotti L. Evaluation of MGMT Gene Methylation in Neuroendocrine Neoplasms. Oncol Res. 2022 Jan 31;28(9):837-845.

8. Cuomo M, Florio E, Della Monica R, Costabile D, Buonaiuto M, Di Risi T, **De Riso G**, Sarnataro A, Cocozza S, Visconti R, Chiariotti L. Epigenetic remodelling of Fxyd1 promoters in developing heart and brain tissues. Sci Rep. 2022 Apr 19;12(1):6471.

9. Roca MS, Moccia T, Iannelli F, Testa C, Vitagliano C, Minopoli M, Camerlingo R, **De Riso G**, De Cecio R, Bruzzese F, Conte M, Altucci L, Di Gennaro E, Avallone A, Leone A, Budillon A. HDAC class I inhibitor domatinostat sensitizes pancreatic cancer to chemotherapy by targeting cancer

stem cell compartment via FOXM1 modulation. J Exp Clin Cancer Res. 2022 Apr 11;41(1):138.

10. Sarnataro A, **De Riso G**, Cocozza S, Pezone A, Majello B, Amente S, Scala G. A novel workflow for the qualitative analysis of DNA methylation data. Comput Struct Biotechnol J. 2022 Oct 23;20:5925-5934.

11. **De Riso G**, Sarnataro A, Scala G, Cuomo M, Della Monica R, Amente S, Chiariotti L, Miele G, Cocozza S. MC profiling: a novel approach to analyze DNA methylation heterogeneity in genome-wide bisulfite sequencing data. NAR Genomics and Bioinformatics, Volume 4, Issue 4, December 2022, lqac096, https://doi.org/10.1093/nargab/lqac096

9. References

Abante J, Fang Y, Feinberg AP, Goutsias J. Detection of haplotype-dependent allele-specific DNA methylation in WGBS data. Nat Commun 2020;11:5238. https://doi.org/10.1038/s41467-020-19077-1.

Affinito O, Palumbo D, Fierro A, Cuomo M, De Riso G, Monticelli A, et al. Nucleotide distance influences co-methylation between nearby CpG sites. Genomics 2020;112:144–50. https://doi.org/10.1016/j.ygeno.2019.05.007.

Affinito O, Scala G, Palumbo D, Florio E, Monticelli A, Miele G, et al. Modeling DNA methylation by analyzing the individual configurations of single molecules. Epigenetics 2016;11:881–8. https://doi.org/10.1080/15592294.2016.1246108.

Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat Methods 2016;13:229–32. https://doi.org/10.1038/nmeth.3728.

Balaton BP, Brown CJ. Contribution of genetic and epigenetic changes to escape from X-chromosome inactivation. Epigenetics Chromatin 2021;14:30. https://doi.org/10.1186/s13072-021-00404-9.

Balaton BP, Cotton AM, Brown CJ. Derivation of consensus inactivation status for X-linked genes from genome-wide studies. Biol Sex Differ 2015;6:35. https://doi.org/10.1186/s13293-015-0053-7.

Balaton BP, Fornes O, Wasserman WW, Brown CJ. Cross-species examination of X-chromosome inactivation highlights domains of escape from silencing. Epigenetics Chromatin 2021;14:12. https://doi.org/10.1186/s13072-021-00386-8.

Bock C. Analysing and interpreting DNA methylation data. Nat Rev Genet 2012;13:705–19. https://doi.org/10.1038/nrg3273.

Bock C, Tomazou EM, Brinkman AB, Müller F, Simmer F, Gu H, et al. Quantitative comparison of genome-wide DNA methylation mapping technologies. Nat Biotechnol 2010;28:1106–14. https://doi.org/10.1038/nbt.1681.

Bock C, Walter J, Paulsen M, Lengauer T. Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. Nucleic Acids Res 2008;36:e55–e55. https://doi.org/10.1093/nar/gkn122.

Boulias K, Greer EL. Means, mechanisms and consequences of adenine methylation in DNA. Nat Rev Genet 2022;23:411–28.

https://doi.org/10.1038/s41576-022-00456-x.

Buitrago D, Labrador M, Arcon JP, Lema R, Flores O, Esteve-Codina A, et al. Impact of DNA methylation on 3D genome structure. Nat Commun 2021;12:3243. https://doi.org/10.1038/s41467-021-23142-8.

Caldwell BA, Liu MY, Prasasya RD, Wang T, DeNizio JE, Leu NA, et al. Functionally distinct roles for TET-oxidized 5-methylcytosine bases in somatic reprogramming to pluripotency. Mol Cell 2021;81:859-869.e8. https://doi.org/10.1016/j.molcel.2020.11.045.

Carmona JJ, Accomando WP, Binder AM, Hutchinson JN, Pantano L, Izzi B, et al. Empirical comparison of reduced representation bisulfite sequencing and Infinium BeadChip reproducibility and coverage of DNA methylation in humans. Npj Genomic Med 2017;2:13. https://doi.org/10.1038/s41525-017-0012-9.

Carter B, Zhao K. The epigenetic basis of cellular heterogeneity. Nat Rev Genet 2021;22:235–50. https://doi.org/10.1038/s41576-020-00300-0.

Cotton AM, Price EM, Jones MJ, Balaton BP, Kobor MS, Brown CJ. Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. Hum Mol Genet 2015;24:1528–39. https://doi.org/10.1093/hmg/ddu564.

Cover TM, Thomas JA. Elements of Information Theory. 1st ed. Wiley; 2005. https://doi.org/10.1002/047174882X.

Cuomo M, Borrelli L, Della Monica R, Coretti L, De Riso G, D'Angelo Lancellotti di Durazzo L, et al. DNA Methylation Profiles of Tph1A and BDNF in Gut and Brain of L. Rhamnosus-Treated Zebrafish. Biomolecules 2021;11:142. https://doi.org/10.3390/biom11020142.

Cuomo M, Keller S, Punzo D, Nuzzo T, Affinito O, Coretti L, et al. Selective demethylation of two CpG sites causes postnatal activation of the Dao gene and consequent removal of d-serine within the mouse cerebellum. Clin Epigenetics 2019;11:149. https://doi.org/10.1186/s13148-019-0732-z.

Dahlet T, Argüeso Lleida A, Al Adhami H, Dumas M, Bender A, Ngondo RP, et al. Genome-wide analysis in the mouse embryo reveals the importance of DNA methylation for transcription integrity. Nat Commun 2020;11:3153. https://doi.org/10.1038/s41467-020-16919-w.

De Riso G, Fiorillo DFG, Fierro A, Cuomo M, Chiariotti L, Miele G, et al. Modeling DNA Methylation Profiles through a Dynamic Equilibrium between Methylation and Demethylation. Biomolecules 2020;10:1271. https://doi.org/10.3390/biom10091271.

De Smet C, Lurquin C, Lethé 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. Mol Cell Biol 1999;19:7327–35. https://doi.org/10.1128/MCB.19.11.7327.

Doi A, Park I-H, Wen B, Murakami P, Aryee MJ, Irizarry R, et al. Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nat Genet 2009;41:1350–3. https://doi.org/10.1038/ng.471.

Edgar R, Tan PPC, Portales-Casamar E, Pavlidis P. Meta-analysis of human methylomes reveals stably methylated sequences surrounding CpG islands associated with high gene expression. Epigenetics Chromatin 2014;7:28. https://doi.org/10.1186/1756-8935-7-28.

Edwards CA, Ferguson-Smith AC. Mechanisms regulating imprinted genes in clusters. Curr Opin Cell Biol 2007;19:281–9. https://doi.org/10.1016/j.ceb.2007.04.013.

Ehrlich M. DNA hypermethylation in disease: mechanisms and clinical relevance. Epigenetics 2019;14:1141–63. https://doi.org/10.1080/15592294.2019.1638701.

Ehrlich M, Gama-Sosa MA, Huang L-H, Midgett RM, Kuo KC, McCune RA, et al. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. Nucleic Acids Res 1982;10:2709–21. https://doi.org/10.1093/nar/10.8.2709.

Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc 2017;12:2478–92. https://doi.org/10.1038/nprot.2017.124.

Florio E, Keller S, Coretti L, Affinito O, Scala G, Errico F, et al. Tracking the evolution of epialleles during neural differentiation and brain development: D-Aspartate oxidase as a model gene. Epigenetics 2017;12:41–54. https://doi.org/10.1080/15592294.2016.1260211.

Fraga MF, Esteller M. Epigenetics and aging: the targets and the marks. Trends Genet 2007;23:413–8. https://doi.org/10.1016/j.tig.2007.05.008.

Garieri M, Stamoulis G, Blanc X, Falconnet E, Ribaux P, Borel C, et al. Extensive cellular heterogeneity of X inactivation revealed by single-cell allele-specific expression in human fibroblasts. Proc Natl Acad Sci 2018;115:13015–20. https://doi.org/10.1073/pnas.1806811115.

Garrett-Bakelman FE, Sheridan CK, Kacmarczyk TJ, Ishii J, Betel D, Alonso A, et al. Enhanced Reduced Representation Bisulfite Sequencing for Assessment of DNA Methylation at Base Pair Resolution. J Vis Exp

2015:52246. https://doi.org/10.3791/52246.

Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc 2011;6:468–81. https://doi.org/10.1038/nprot.2010.190.

Guo X. JS-MA: A Jensen-Shannon Divergence Based Method for Mapping Genome-Wide Associations on Multiple Diseases. Front Genet 2020;11:507038. https://doi.org/10.3389/fgene.2020.507038.

Haerter JO, Lövkvist C, Dodd IB, Sneppen K. Collaboration between CpG sites is needed for stable somatic inheritance of DNA methylation states. Nucleic Acids Res 2014;42:2235–44. https://doi.org/10.1093/nar/gkt1235.

Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. Mol Cell 2013;49:359–67. https://doi.org/10.1016/j.molcel.2012.10.016.

Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, et al. Increased methylation variation in epigenetic domains across cancer types. Nat Genet 2011;43:768–75. https://doi.org/10.1038/ng.865.

Hermann A, Goyal R, Jeltsch A. The Dnmt1 DNA-(cytosine-C5)-methyltransferase Methylates DNA Processively with High Preference for Hemimethylated Target Sites. J Biol Chem 2004;279:48350–9. https://doi.org/10.1074/jbc.M403427200.

Hiura H, Sugawara A, Ogawa H, John RM, Miyauchi N, Miyanari Y, et al. A tripartite paternally methylated region within the Gpr1-Zdbf2 imprinted domain on mouse chromosome 1 identified by meDIP-on-chip. Nucleic Acids Res 2010;38:4929–45. https://doi.org/10.1093/nar/gkq200.

Horvath S. DNA methylation age of human tissues and cell types. Genome Biol 2013;14:R115. https://doi.org/10.1186/gb-2013-14-10-r115.

Huan Q, Zhang Y, Wu S, Qian W. HeteroMeth: A Database of Cell-to-cell Heterogeneity in DNA Methylation. Genomics Proteomics Bioinformatics 2018;16:234–43. https://doi.org/10.1016/j.gpb.2018.07.002.

Hui T, Cao Q, Wegrzyn-Woltosz J, O'Neill K, Hammond CA, Knapp DJHF, et al. High-Resolution Single-Cell DNA Methylation Measurements Reveal Epigenetically Distinct Hematopoietic Stem Cell Subpopulations. Stem Cell Rep 2018;11:578–92. https://doi.org/10.1016/j.stemcr.2018.07.003.

Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at

conserved tissue-specific CpG island shores. Nat Genet 2009;41:178–86. https://doi.org/10.1038/ng.298.

Itzkovitz S, Hodis E, Segal E. Overlapping codes within protein-coding sequences. Genome Res 2010;20:1582–9. https://doi.org/10.1101/gr.105072.110.

Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet 2003;33:245–54. https://doi.org/10.1038/ng1089.

Jeltsch A, Jurkowska RZ. New concepts in DNA methylation. Trends Biochem Sci 2014;39:310–8. https://doi.org/10.1016/j.tibs.2014.05.002.

Jenkinson G, Abante J, Feinberg AP, Goutsias J. An information-theoretic approach to the modeling and analysis of whole-genome bisulfite sequencing data. BMC Bioinformatics 2018;19:87. https://doi.org/10.1186/s12859-018-2086-5.

Jenkinson G, Pujadas E, Goutsias J, Feinberg AP. Potential energy landscapes identify the information-theoretic nature of the epigenome. Nat Genet 2017;49:719–29. https://doi.org/10.1038/ng.3811.

Jin Z, Liu Y. DNA methylation in human diseases. Genes Dis 2018;5:1–8. https://doi.org/10.1016/j.gendis.2018.01.002.

Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 2012;13:484–92. https://doi.org/10.1038/nrg3230.

Kacmarczyk TJ, Fall MP, Zhang X, Xin Y, Li Y, Alonso A, et al. “Same difference”: comprehensive evaluation of four DNA methylation measurement platforms. Epigenetics Chromatin 2018;11:21. https://doi.org/10.1186/s13072-018-0190-4.

Karemaker ID, Vermeulen M. Single-Cell DNA Methylation Profiling: Technologies and Biological Applications. Trends Biotechnol 2018;36:952–65. https://doi.org/10.1016/j.tibtech.2018.04.002.

Kartal Ö, Schmid MW, Grossniklaus U. Cell type-specific genome scans of DNA methylation divergence indicate an important role for transposable elements. Genome Biol 2020;21:172. https://doi.org/10.1186/s13059-020-02068-2.

Keniry A, Blewitt ME. Studying X chromosome inactivation in the single-cell genomic era. Biochem Soc Trans 2018;46:577–86. https://doi.org/10.1042/BST20170346.

Khavari DA, Sen GL, Rinn JL. DNA methylation and epigenetic control of

cellular differentiation. Cell Cycle 2010;9:3880–3. https://doi.org/10.4161/cc.9.19.13385.

Kim M, Costello J. DNA methylation: an epigenetic mark of cellular memory. Exp Mol Med 2017;49:e322–e322. https://doi.org/10.1038/emm.2017.10.

Klobučar T, Kreibich E, Krueger F, Arez M, Pólvora-Brandão D, von Meyenn F, et al. IMPLICON: an ultra-deep sequencing method to uncover DNA methylation at imprinted regions. Nucleic Acids Res 2020;48:e92–e92. https://doi.org/10.1093/nar/gkaa567.

Kohli RM, Zhang Y. TET enzymes, TDG and the dynamics of DNA demethylation. Nature 2013;502:472–9. https://doi.org/10.1038/nature12750.

Konigsberg IR, Barnes B, Campbell M, Davidson E, Zhen Y, Pallisard O, et al. Host methylation predicts SARS-CoV-2 infection and clinical outcome. Commun Med 2021;1:42. https://doi.org/10.1038/s43856-021-00042-y.

Landan G, Cohen NM, Mukamel Z, Bar A, Molchadsky A, Brosh R, et al. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. Nat Genet 2012;44:1207–14. https://doi.org/10.1038/ng.2442.

Landau DA, Clement K, Ziller MJ, Boyle P, Fan J, Gu H, et al. Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia. Cancer Cell 2014;26:813–25. https://doi.org/10.1016/j.ccell.2014.10.012.

LaSalle JM. X Chromosome Inactivation Timing is Not eXACT: Implications for Autism Spectrum Disorders. Front Genet 2022;13:864848. https://doi.org/10.3389/fgene.2022.864848.

Levy MA, McConkey H, Kerkhof J, Barat-Houari M, Bargiacchi S, Biamino E, et al. Novel diagnostic DNA methylation episignatures expand and refine the epigenetic landscapes of Mendelian disorders. Hum Genet Genomics Adv 2022;3:100075. https://doi.org/10.1016/j.xhgg.2021.100075.

Li E. Chromatin modification and epigenetic reprogramming in mammalian development. Nat Rev Genet 2002;3:662–73. https://doi.org/10.1038/nrg887.

Li E, Beard C, Jaenisch R. Role for DNA methylation in genomic imprinting. Nature 1993;366:362–5. https://doi.org/10.1038/366362a0.

Li S, Garrett-Bakelman F, Perl AE, Luger SM, Zhang C, To BL, et al. Dynamic evolution of clonal epialleles revealed by methclone. Genome Biol 2014;15:472. https://doi.org/10.1186/s13059-014-0472-5.

Lin J. Divergence measures based on the Shannon entropy. IEEE Trans Inf

Theory 1991;37:145–51. https://doi.org/10.1109/18.61115.

Lokk K, Modhukur V, Rajashekar B, Märtens K, Mägi R, Kolde R, et al. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. Genome Biol 2014;15:3248. https://doi.org/10.1186/gb-2014-15-4-r54.

Lyko F. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. Nat Rev Genet 2018;19:81–92. https://doi.org/10.1038/nrg.2017.80.

Masser DR, Berg AS, Freeman WM. Focused, high accuracy 5-methylcytosine quantitation with base resolution by benchtop next-generation sequencing. Epigenetics Chromatin 2013;6:33. https://doi.org/10.1186/1756-8935-6-33.

de Mendoza A, Poppe D, Buckberry S, Pflueger J, Albertin CB, Daish T, et al. The emergence of the brain non-CpG methylation system in vertebrates. Nat Ecol Evol 2021;5:369–78. https://doi.org/10.1038/s41559-020-01371-2.

Mikeska T, Candiloro IL, Dobrovic A. The implications of heterogeneous DNA methylation for the accurate quantification of methylation. Epigenomics 2010;2:561–73. https://doi.org/10.2217/epi.10.32.

Mitchell C, Schneper LM, Notterman DA. DNA methylation, early life environment, and health outcomes. Pediatr Res 2016;79:212–9. https://doi.org/10.1038/pr.2015.193.

Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, Stadler MB, et al. Lineage-Specific Polycomb Targets and De Novo DNA Methylation Define Restriction and Potential of Neuronal Progenitors. Mol Cell 2008;30:755–66. https://doi.org/10.1016/j.molcel.2008.05.007.

Moore LD, Le T, Fan G. DNA Methylation and Its Basic Function. Neuropsychopharmacology 2013;38:23–38. https://doi.org/10.1038/npp.2012.112.

NISC Comparative Sequencing Program, Duncan CG, Grimm SA, Morgan DL, Bushel PR, Bennett BD, et al. Dosage compensation and DNA methylation landscape of the X chromosome in mouse liver. Sci Rep 2018;8:10138. https://doi.org/10.1038/s41598-018-28356-3.

Ochoa E, Lee S, Lan-Leung B, Dias RP, Ong KK, Radley JA, et al. ImprintSeq, a novel tool to interrogate DNA methylation at human imprinted regions and diagnose multilocus imprinting disturbance. Genet Med 2022;24:463–74. https://doi.org/10.1016/j.gim.2021.10.011.

Okano M, Bell DW, Haber DA, Li E. DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. Cell 1999;99:247–57.

https://doi.org/10.1016/S0092-8674(00)81656-6.

Orozco LD, Rubbi L, Martin LJ, Fang F, Hormozdiari F, Che N, et al. Intergenerational genomic DNA methylation patterns in mouse hybrid strains. Genome Biol 2014;15:R68. https://doi.org/10.1186/gb-2014-15-5-r68.

Palumbo D, Affinito O, Monticelli A, Cocozza S. DNA Methylation variability among individuals is related to CpGs cluster density and evolutionary signatures. BMC Genomics 2018;19:229. https://doi.org/10.1186/s12864-018-4618-9.

Ramsahoye BH, Biniszkiewicz D, Lyko F, Clark V, Bird AP, Jaenisch R. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proc Natl Acad Sci 2000;97:5237–42. https://doi.org/10.1073/pnas.97.10.5237.

Raval A, Tanner SM, Byrd JC, Angerman EB, Perko JD, Chen S-S, et al. Downregulation of Death-Associated Protein Kinase 1 (DAPK1) in Chronic Lymphocytic Leukemia. Cell 2007;129:879–90. https://doi.org/10.1016/j.cell.2007.03.043.

Rizwana R, Hahn PJ. CpG methylation reduces genomic instability. J Cell Sci 1999;112:4513–9. https://doi.org/10.1242/jcs.112.24.4513.

Robertson KD. DNA methylation and human disease. Nat Rev Genet 2005;6:597–610. https://doi.org/10.1038/nrg1655.

Sarnataro A, De Riso G, Cocozza S, Pezone A, Majello B, Amente S, et al. A novel workflow for the qualitative analysis of DNA methylation data. Comput Struct Biotechnol J 2022;20:5925–34. https://doi.org/10.1016/j.csbj.2022.10.027.

Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci 2006;103:1412–7. https://doi.org/10.1073/pnas.0510310103.

Scala G, Affinito O, Palumbo D, Florio E, Monticelli A, Miele G, et al. ampliMethProfiler: a pipeline for the analysis of CpG methylation profiles of targeted deep bisulfite sequenced amplicons. BMC Bioinformatics 2016;17:484. https://doi.org/10.1186/s12859-016-1380-3.

Scherer M, Nazarov PV, Toth R, Sahay S, Kaoma T, Maurer V, et al. Reference-free deconvolution, visualization and interpretation of complex DNA methylation data using DecompPipeline, MeDeCom and FactorViz. Nat Protoc 2020a;15:3240–63. https://doi.org/10.1038/s41596-020-0369-6.

Scherer M, Nebel A, Franke A, Walter J, Lengauer T, Bock C, et al. Quantitative comparison of within-sample heterogeneity scores for DNA

methylation data. Nucleic Acids Res 2020b;48:e46–e46. https://doi.org/10.1093/nar/gkaa120.

Sharma A, Klein SS, Barboza L, Lohdi N, Toth M. Principles Governing DNA Methylation during Neuronal Lineage and Subtype Specification. J Neurosci Off J Soc Neurosci 2016;36:1711–22. https://doi.org/10.1523/JNEUROSCI.4037-15.2016.

Shi J, Xu J, Chen YE, Li JS, Cui Y, Shen L, et al. The concurrence of DNA methylation and demethylation is associated with transcription regulation. Nat Commun 2021;12:5285. https://doi.org/10.1038/s41467-021-25521-7.

Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 2007;8:272–85. https://doi.org/10.1038/nrg2072.

Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet 2013;14:204–20. https://doi.org/10.1038/nrg3354.

Teschendorff AE, Zhu T, Breeze CE, Beck S. EPISCORE: cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data. Genome Biol 2020;21:221. https://doi.org/10.1186/s13059-020-02126-9.

Varley KE, Mitra RD. Bisulfite Patch PCR enables multiplexed sequencing of promoter methylation across cancer samples. Genome Res 2010;20:1279–87. https://doi.org/10.1101/gr.101212.109.

van der Velde A, Fan K, Tsuji J, Moore JE, Purcaro MJ, Pratt HE, et al. Annotation of chromatin states in 66 complete mouse epigenomes during development. Commun Biol 2021;4:239. https://doi.org/10.1038/s42003-021-01756-4.

Villicaña S, Bell JT. Genetic impacts on DNA methylation: research findings and future perspectives. Genome Biol 2021;22:127. https://doi.org/10.1186/s13059-021-02347-6.

Wu X, Li G, Xie R. Decoding the role of TET family dioxygenases in lineage specification. Epigenetics Chromatin 2018;11:58. https://doi.org/10.1186/s13072-018-0228-7.

Wu X, Zhang Y. TET-mediated active DNA demethylation: mechanism, function and beyond. Nat Rev Genet 2017;18:517–34. https://doi.org/10.1038/nrg.2017.33.

Xie H, Wang M, de Andrade A, Bonaldo M de F, Galat V, Arndt K, et al. Genome-wide quantitative assessment of variation in DNA methylation patterns. Nucleic Acids Res 2011;39:4099–108. https://doi.org/10.1093/nar/gkr017.

Xu J, Pope SD, Jazirehi AR, Attema JL, Papathanasiou P, Watts JA, et al. Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. Proc Natl Acad Sci 2007;104:12377–82. https://doi.org/10.1073/pnas.0704579104.

Xu J, Shi J, Cui X, Cui Y, Li JJ, Goel A, et al. Cellular Heterogeneity–Adjusted cLonal Methylation (CHALM) improves prediction of gene expression. Nat Commun 2021;12:400. https://doi.org/10.1038/s41467-020-20492-7.

Yong W-S, Hsu F-M, Chen P-Y. Profiling genome-wide DNA methylation. Epigenetics Chromatin 2016;9:26. https://doi.org/10.1186/s13072-016-0075-3.

Zhang L, Xie WJ, Liu S, Meng L, Gu C, Gao YQ. DNA Methylation Landscape Reflects the Spatial Organization of Chromatin in Different Cells. Biophys J 2017;113:1395–404. https://doi.org/10.1016/j.bpj.2017.08.019.

Zhang X, Wang X. MeConcord: a new metric to quantitatively characterize DNA methylation heterogeneity across reads and CpG sites. Bioinformatics 2022;38:i307–15. https://doi.org/10.1093/bioinformatics/btac248.