Università degli Studi di Napoli Federico II
Ph.D. Program in
Computational and **Q**uantitative **B**iology
XXXVI Cycle

Thesis for the Degree of Doctor of Philosophy

Dissecting DNA methylation patterns at single-molecule level from bulk bisulfite sequencing experiments

PhD student
Antonella Sarnataro

Advisor: Prof. Sergio Cocozza

Co-advisor: Dr. Giovanni Scala

Dipartimento di Medicina Molecolare e Biotecnologie Mediche

Try again. Fail again. Fail better.

Samuel Beckett

Dissecting DNA methylation patterns at single-molecule level from bulk bisulfite sequencing experiments

Ph.D. Thesis presented
for the fulfillment of the Degree of Doctor of
Philosophy in Computational and Quantitative Biology

by

Antonella Sarnataro

December 2023

Approved as to style and content by

_____

Prof. Sergio Cocozza

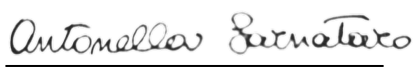
_____

Dr. Giovanni Scala

Candidate's declaration

I hereby declare that this thesis submitted to obtain the academic degree of Philosophiæ Doctor (Ph.D.) in Computational and Quantitative Biology is my own unaided work, that I have not used other than the sources indicated, and that all direct and indirect sources are acknowledged as references.
Parts of this dissertation have been published in international journals and/or conference articles (see list of the author's publications at the end of the thesis).

Napoli,



Antonella Sarnataro

Index

Abbreviations

CCA Canonical Correlation Analysis

CGIs CpG Islands

CTCF CCCTC-binding Factor

DHR Differentially Heterogeneous Regions

DMP Differentially Methylated Positions

DNA Deoxyribonucleic Acid

DNMTs DNA Methyltransferases

FDRP Fraction of Discordant Read Pairs

GEO Gene Expression Omnibus

HD Huntington disease

NGS Next Generation Sequencing

ONT Oxford Nanopore Technology

PCA Principal Component Analysis

TETs Ten-Eleven Translocation enzymes

WGBS Whole Genome Bisulfite Sequencing

RRBS Reduced Representation Bisulfite Sequencing

RTS Read Transition Score

SE Shannon Entropy

SRA Sequence Read Archive

TSS Transcription start site

Abstract

DNA methylation is one of the most well-studied epigenetic modifications and plays a central role in important biological processes, such as gene expression regulation, genome imprinting, and genome stability.

In the mammalian genome, this modification occurs predominantly at CpG dinucleotides, with non-CpG methylation described only in specific cell lineages, such as neurons and embryonic stem cells.

DNA methylation is a highly reversible and dynamic process, so there can be high variability in DNA methylation patterns between distinct cells. This variability in cell-to-cell DNA methylation has been depicted as a major driver of cellular plasticity, which in turn contributes to several pathophysiological processes.

Over the last decade, several computational tools have been developed to extract DNA methylation heterogeneity information from bulk bisulfite sequencing, avoiding the experimental and practical limitations of single-cells assays.

However, most of these tools lack additional functionalities for a comprehensive analysis of DNA methylation heterogeneity. This work contributes to the previous research in this field by developing a novel computational workflow designed for the analysis of both CpG and non-CpG-based DNA methylation patterns. Offering customizable user-driven analyses, strand-specific heterogeneity assessments, locus annotation, and gene set enrichment analyses, EpiStatProfiler adds versatility to the exploration of DNA methylation patterns.

Finally, it is also shown that extending the workflow for extracting and analyzing DNA methylation patterns to third-generation sequencing experiments can be valuable to

provide novel insights into the foundation of epigenetic heterogeneity in both normal and pathological conditions.

Summary in lingua italiana

La metilazione del DNA è una delle modificazioni epigenetiche maggiormente studiate, in quanto implicata in processi biologici chiave, come la regolazione dell'espressione genica, l'imprinting genomico e la stabilità del genoma.

Nei mammiferi, questa modificazione avviene prevalentemente nei dinucleotidi CpG, con la metilazione non-CpG descritta invece solo in linee cellulari specifiche, come i neuroni e le cellule staminali embrionali. La metilazione del DNA è altamente reversibile, quindi può esserci un'elevata variabilità nei livelli di metilazione del DNA tra single cellule.

Questa variabilità nella metilazione del DNA da cellula a cellula è stata descritta come uno dei principali fattori di plasticità cellulare, che a sua volta contribuisce a diversi processi fisiopatologici. Nell'ultimo decennio sono stati sviluppati diversi strumenti computazionali per estrarre informazioni sull'eterogeneità della metilazione del DNA dal sequenziamento del bisolfito, superando i limiti dei saggi su singole cellule.

Tuttavia, la maggior parte di questi strumenti non permette un'analisi completa dell'eterogeneità della metilazione del DNA. Questo lavoro contribuisce alla precedente ricerca in questo campo attraverso lo sviluppo di un nuovo strumento computazionale progettato per l'analisi di modelli di metilazione del DNA sia basati su CpG che non basati su CpG. EpiStatProfiler aggiunge versatilità all'esplorazione dell'eterogeneità di metilazione del DNA.

Infine, si dimostra che estendere il flusso di lavoro per l'estrazione e l'analisi dei modelli di metilazione del DNA agli esperimenti di sequenziamento di terza generazione può essere rilevante per fornire nuove

informazioni sulla regolazione dell'eterogeneità epigenetica sia in condizioni normali che patologiche.

List of figures

1. Summary of chapters content

In *1942* the developmental biologist Conrad Waddington first introduced the term *epigenetics (Noble 2015)*, indicating a phenomenon through which alterations in the gene phenotype could occur in absence of DNA sequence changes.

Over the last decades, epigenetics became a well-established field of research which embraces the study of all those stable modifications which are able to regulate gene expression during development or in response to environmental stimuli (Gibney and Nolan 2010). To do so, such modifications are characterized by fine-tuned dynamics orchestrated by a series of enzymes classified based on their function either as writers, readers or erasers ("Writers, Readers, and Erasers of Epigenetic Marks" 2015).

In this context, two major molecular mechanisms have been largely characterized for their contribution to epigenetic phenomena: *DNA methylation* and *histone modifications ("DNA Methylation: A Historical Perspective" 2022; Millán-Zambrano et al. 2022).*

DNA methylation is the most extensively studied epigenetic modification, exerting a major role in key molecular processes, such as gene expression regulation, genome imprinting, X-chromosome inactivation and genome stability(Smith and Meissner 2013; Greenberg and Bourc'his 2019). In mammalian genomes, it mainly consists in the addition of a methyl group to the fifth carbon of a cytosine base in those DNA sequence contexts where the cytosine is immediately followed by a guanine, i.e. the CpG dinucleotides (Jones 2012).

Recently, DNA methylation at other genomic contexts, such as CpA dinucleotides, has been described for exerting a biological function in specific tissues, such as the brain (de Mendoza et al. 2021).

There is an entire machinery involved in the regulation of such modifications, and dysregulation of this machinery has been extensively described in various disorders, such as imprinting diseases, STR expansion disorder, and foremost cancer (Greenberg and Bourc'his 2019; Jin and Liu 2018; Mishra and Bishnupuri 2016; Thiagalingam 2015).

In Chapter 2, an extensive overview of DNA methylation, along with its regulation across the genome, is provided. Major writers of DNA methylation are described, and the landscape of methylated cytosines across the genome is portrayed. How this work is intended to contribute to deepen the knowledge of DNA methylation patterns maintenance is described through the chapter, along with a description of previous related research in the field.

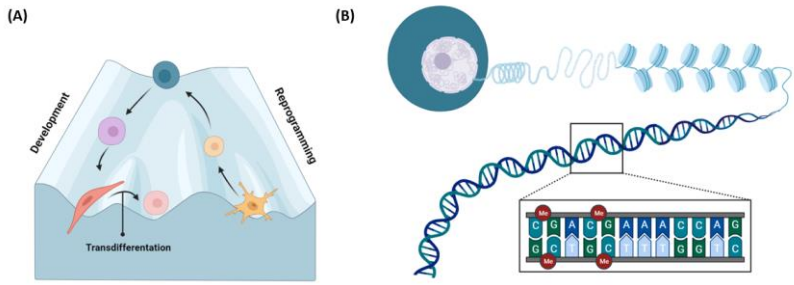


Figure 1.1 Epigenetic landscape and DNA methylation. *Illustration of the epigenetic landscape concept proposed by Waddington* ***(A)****. He envisioned the landscape as a series of valleys and ridges a cell can take during its development and differentiation. Among the different epigenetic modifications which can change a cell phenotype, DNA methylation is one of the most well-studied* ***(B)****. As highlighted in the figure, it consists of the addition of a methyl group to the 5th carbon of the cytosine base, which in mammals mainly occurs in the context of CpG dinucleotides.*

In Chapter 3, the general aim of the project is summarized, followed by a detailed description of the specific goal of this PhD thesis. To summarize, the major goal of the project was the development of a novel

computational framework for the analysis of DNA methylation heterogeneity from bisulfite sequencing samples. Single objectives to reach the goal are then described.

In Chapter 4 the methodology of the work is described. Here, the implementation of the workflow is described in detail, together with the methods employed to perform the benchmarking of the package and additional downstream analysis. Data manipulation and implementation of the novel algorithm are reported as well.

In Chapter 5 the main results of the research thesis are portrayed. The main functions provided with the novel package are documented. In this chapter, application of the workflow to detect relevant loci associated with specific biological conditions is described.

In Chapter 6 the major results of the work are discussed.

2. Background and related work

2.1 *Overview of DNA methylation*

Epigenetics is a rapidly evolving field in biology which explores mechanisms by which genes are regulated and expressed in an organism. The term *epigenetics* literally means *beyond* or *above* genetics, referring to those changes in gene expression or cellular phenotypes that can be passed on from one generation to the next, without alterations to the underlying DNA sequence (Goldberg, Allis, and Bernstein 2007; Waddington 2012).

In this context, the ***epigenome*** refers to all the epigenetic marks on DNA that can be found in a single cell. The epigenome represents a valuable layer to comprehend how genes are regulated and how gene expression gets dysregulated in diseases such as cancer, developmental disorders and various other conditions (Jaenisch and Bird 2003).

Among the other epigenetic modifications, DNA methylation is one of the most well-studied. It consists of the addition of a methyl group on the 5th carbon of a cytosine residue on the DNA backbone, which in mammals mainly occurs in CpG dinucleotides contexts (Moore, Le, and Fan 2013). DNA methylation is also described in other DNA sequence contexts, but less is known about the functional roles attributed to non-CpG methylation. Beyond the others, CpA methylation has been emerged as an important epigenetic marker, relatively abundant in specific cell types, such as the pluripotent stem cells and mature neurons (de Mendoza et al. 2021; Tillotson et al. 2021; J.-H. Lee, Park, and Nakai 2017).

To better comprehend the roles of DNA methylation, it is important to examine how methylation is distributed throughout the genome. As a result of DNA methylation deposited on the cytosine bases, human genome is found to

be CpG depleted, since 5mC can be converted to thymine by spontaneous or enzymatic deamination, thus being eliminated through the cell cycle (Smith and Meissner 2013; Jones 2012). In this context, it is noteworthy to mention that genome-wide methylation assays allowed to deepen the global understanding of methylation patterns through the genome and in the different cell-types (Meissner et al. 2005; Bock et al. 2010; Yong, Hsu, and Chen 2016).

Our current knowledge of DNA methylation is based on the observation that this modification follows a bimodal distribution. Of the roughly 28 million CpG sites found in the human genome, 60-80 % are generally methylated. The remaining CpG sites are found clustered in relatively short genomic domains, named CpG islands (CGI), which are instead resistant to DNA methylation and found near genes (~70% of all gene promoters), even though distal (or orphan, oCGIs) CGIs are also found through the rest of the genome (Smith and Meissner 2013).

The evidence that promoter CGI island methylation state influences the transcription levels of the associated gene was one of the first relevant proofs to establish a main functional role of DNA methylation as epigenetic regulator at specific genomic contexts (Jaenisch and Bird 2003; Holliday and Pugh 1975; Hallgrímsson and Hall 2011).

DNA methylation is now well described as having a pivotal role in major biological mechanisms. As mentioned above, it is well established that there exists a direct association between high levels of DNA methylation at gene promoters and their silencing. Nonetheless, current research is revealing that this relationship can be very nuanced, and additional details to these mechanisms are being elucidated (Jones 2012).

Besides its role within promoter CGI, DNA methylation is shown to exert a key role in many additional genomic contexts.

For example, classified among one of its most conserved functions across species, there is the role of DNA methylation at pericentromeric repeats, where the lack of DNA methylation may cause the transcriptional activation of latent elements which may alter chromosome alignment, segregation and integrity (Chen, Tsujimoto, and Li 2004).

As well, the maintenance of DNA methylation at transposable elements is essential to prevent their activation (Leung et al. 2011). The fine-tuned regulation of parents-of-origin allele-specific methylation is well studied in the context of genomic imprinting (Reik and Walter 2001).

Finally, DNA methylation plays a key role in the gene dosage regulation of X chromosome elements in females, inactivating one copy of the X chromosome (Wutz and Valencia 2015; Loda, Collombet, and Heard 2022).

2.2 *Regulators of DNA methylation*

In mammals, the foundation and the propagation of DNA methylation patterns are ensured by a family of enzymes known as DNA methyltransferases (DNMTs). Three members of the family are responsible for the catalysis of methyl transfer to the DNA cytosines in main mammalian species: DNMT1, DNMT3A and DNMT3B (Gujar, Weisenberger, and Liang 2019; Edwards et al. 2017). Other members, such as DNMT3L, do not have a catalytic activity but work by stimulating the activity of the other DNMT3 enzymes. These writers of DNA methylation have been always described as part of a well-established model in which each member had a specific role.

In this long-established model, DNA methylation marks are erased in the germ cells and in pre-implantation embryos in two distinct waves to allow the establishment of cell type specific DNA methylation patterns during mammalian development. DNMT3A and DNMT3B are the

writers orchestrating the foundation of DNA methylation signatures across the genome, showing no specificity at DNA sequence contexts (Andrews et al. 2023; Okano et al. 1999). In the developing embryo, DNMT1 is the main actor involved in maintenance of DNA methylation, ensuring the heritability of DNA methylation marks during development, showing greater affinity for hemi-methylated DNA molecules (Hermann, Goyal, and Jeltsch 2004; Nishiyama et al. 2020).

Nonetheless, there is now some evidence which supports a not so strict distinction of roles between the two subtypes of enzymes. In particular, observations coming from the selective knock-out of one of these two classes of enzymes show they both can exert a role which is not expected considering the traditional model (Jeltsch and Jurkowska 2014; Jones and Liang 2009; Dahlet et al. 2020).

Despite being highly stable, methylated cytosines can be erased through passive or active demethylation. Passive demethylation is a consequence of the failure of the DNA maintenance machinery upon DNA replication, resulting in a gradual loss of methylated residues (Parry, Rulands, and Reik 2020).

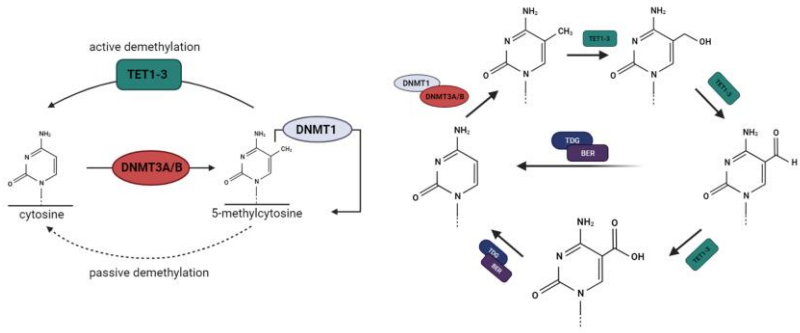


Figure 2.1 DNA methylation process. *Overview of DNA methylation dynamics. DNMT3A/B and DNMT1 regulate de novo and maintenance DNA methylation deposition, respectively. Methylated residues can be then erased through either passive or active DNA methylation, depending on whether the*

process occurs through enzymatic activity exerted by the Ten-To-Eleven translocation (TET) enzymes.

Active demethylation is instead exerted by another class of enzymes, belonging to the Ten-to-Eleven Translocation (TET) protein family. These enzymes catalyze a reaction which oxidize 5mC into novel compounds, which are 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC). 5hmC is of particular interest, being also classified as an epigenetic mark itself, since it has been shown to exert a role in regulatory processes (Parry, Rulands, and Reik 2020).

2.3 *Techniques to measure DNA methylation*

The advent of high-throughput sequencing has provided the opportunity to investigate epigenetic marks at unprecedented resolution across the entire genome. Over the years, various methods have been developed to identify methylated cytosines at high resolution (Meissner et al. 2005; Bock et al. 2010; Masser, Berg, and Freeman 2013; Gu et al. 2011).

Nowadays, the gold standard for calling DNA methylation is based on the bisulfite conversion (Meissner et al. 2005; Gu et al. 2011). The rationale of bisulfite conversion relies on the selective conversion of unmethylated cytosine into uracil bases upon DNA treatment with sodium bisulfite. Uracil bases are then converted into thymines during DNA amplification steps so that it is possible to call methylated cytosines at each sequenced read by performing sequencing comparison at reference CpG sites.

Bisulfite sequencing analysis thus enables comprehensive profiling of DNA methylation patterns at single base resolution, either at genome-wide scale or at targeted genomic regions (Bock et al. 2010; Masser, Stanford, and Freeman 2015).

Whole genome bisulfite sequencing (WGBS) allows the assessment of DNA methylation levels across the entire genome, thus requiring high sequencing throughputs for each sequenced sample (Bock et al. 2010).

On the other hand, assays such as the reduced representation bisulfite sequencing (RRBS) allows the quantification of DNA methylation upon enrichment of high-density CpGs areas, losing information about other regions of the genome, but acquiring information power at the coverage level (Meissner et al. 2005).

Other strategies to profile only specific genomic regions (such as candidate genes for specific disorders) are available, such as targeted PCR amplicon-based deep bisulfite sequencing (Masser, Stanford, and Freeman 2015). This last method relies on the use of specifically designed oligo primers targeting the region of interest during the sample preparation. This allows for high resolution profiling in terms of sequencing depth and experimental costs.

2.4 *Methylation calling*

The computational pipelines employed to profile DNA methylation usually follow shared steps to be accomplished in order to get methylation calls at sufficiently covered regions (Figure 2.2). Raw sequenced reads are first quality checked and sequencing adapters together with poor-quality bases are trimmed to ensure reliability of the data for downstream analysis. Reads are then aligned to the reference genome using bisulfite-aware alignments tools which create in-silico sequence intermediates to ensure the mapping of the lower complexity reads (Krueger and Andrews 2011).

Methylation calling is then performed at the reference CpG positions by indicating as methylated (1) the positions displaying a C on the analyzed read, whereas

calling as unmethylated (0) the cytosines for which a T is displayed at the given position on the analyzed read.

2.5 *Quantitative DNA methylation assessment*

Most of the computational tools used for the assessment of DNA methylation levels rely on the calculation of the average DNA methylation at single-base resolution by estimating the proportion of mapped reads displaying a C over the total number of reads (methylated + unmethylated) mapped at the same position (Bock et al. 2010; Akalin et al. 2012; Assenov et al. 2014; Müller et al. 2019).

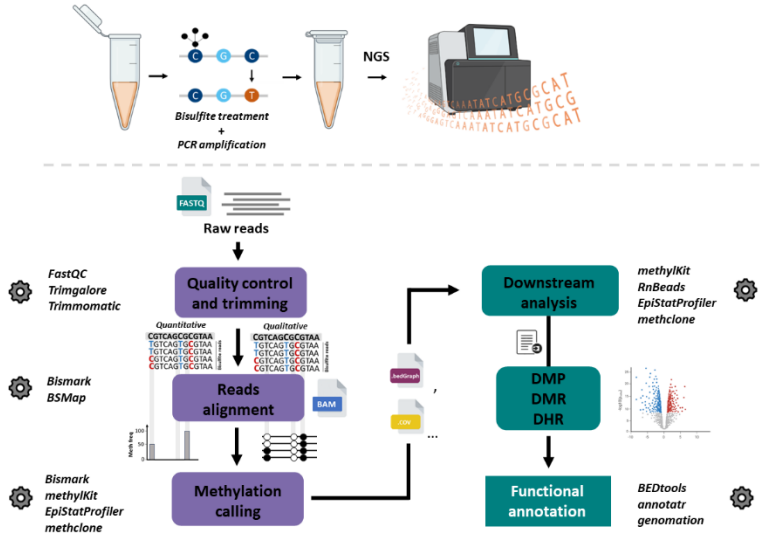


Figure 2.2 DNA methylation analysis workflow. *DNA methylation analysis workflow. The gold standard method, bisulfite sequencing, relies on bisulfite conversion to selectively identify methylated cytosines. The computational pipeline for methylation calling involves quality checking, alignment to the reference genome, and determination of methylated and unmethylated positions at reference CpG sites. Quantitative assessment relies on calculating the average DNA methylation at single-base resolution by estimating the proportion of mapped reads displaying methylated cytosines.*

A plethora of computational tools have been published for the quantification of the average DNA

methylation levels at the genome-wide scale and the comparison of the DNA methylation states between different conditions (Akalin et al. 2012; Assenov et al. 2014; Müller et al. 2019).

These tools include several functions to summarize average DNA methylation levels across genomic features, to perform sample clustering, sample quality visualization and last to perform differential analysis.

Application of these methods has allowed gain of knowledge in the epigenetic field and allowed the discovery of relevant epigenetic signatures associated with specific disease disorders.

2.6 *Qualitative DNA methylation assessment*

Bisulfite sequencing data holds additional information which goes beyond the individual CpG methylation states. This layer of information is represented by the phased methylation states that can be estimated at one single read sequencing (known as epialleles) (Landan et al. 2012). Considering for example a genomic region harboring 4 CpG sites (Figure 2.3.B), there are 16 different possible combinations of methylation states displayed at CpG sites.

The estimation of the composition in terms of these DNA methylation patterns at one genomic location is referred to as DNA methylation heterogeneity.

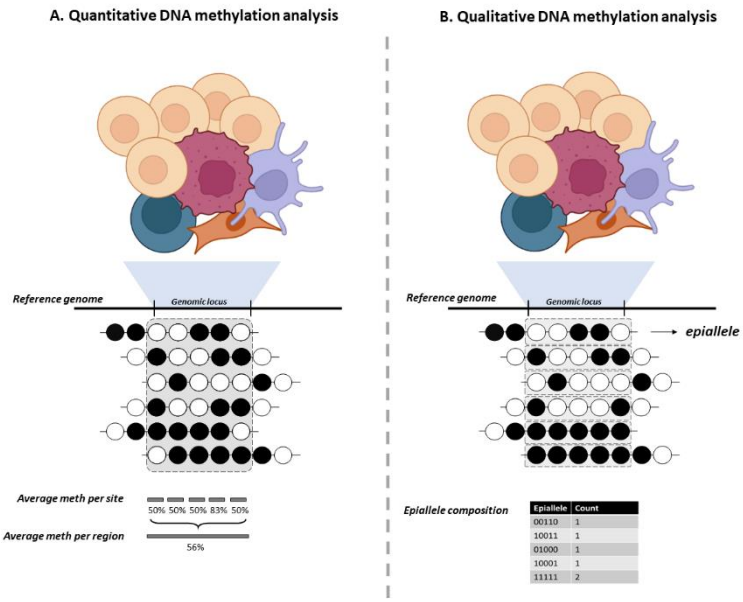


Figure 2.3 Quantitative and qualitative approach to study DNA methylation. *The figure illustrates the distinction between quantitative and qualitative approaches in studying DNA methylation. The quantitative method (A) involves computational tools that calculate average DNA methylation levels across the genome, enabling comparisons between conditions and the identification of epigenetic signatures associated with specific diseases. In contrast, the qualitative approach explores additional layers of information from bisulfite sequencing data, focusing on DNA methylation heterogeneity. This heterogeneity, as depicted in the Figure* ***(B)****, considers the various combinations of methylation states at individual CpG sites within a genomic region.*

Biological sources (Figure 2.4) of DNA methylation heterogeneity can be various (Scherer et al. 2020). A primary source is represented by cell-type composition. High DNA methylation heterogeneity can be displayed from a variety of cell types in a bulk population (Scherer et al. 2020; Sun et al. 2010; Reinius et al. 2012).

On the other hand, high DNA methylation heterogeneity has also been described within homogenous cellular populations (Singer et al. 2014). Allele- and strand-specific DNA methylation may also contribute to the extent

of heterogeneity displayed at one genomic locus (Abante et al. 2020).

Finally, DNA methylation erosion, which consists of the loss of DNA methylation upon DNA replication, may also generate disordered DNA methylation patterns at given loci.

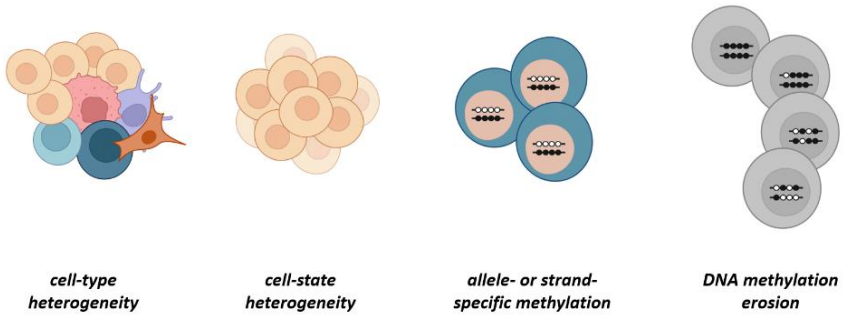


Figure 2.4 Biological sources of DNA methylation. *In the picture, the different biological sources of DNA methylation heterogeneity are depicted. Cell-type composition is a major source of high levels of heterogeneity displayed at bulk tissue. On the other hand, distinct DNA methylation patterns can be found within homogenous cellular populations. Allele- or strand-specific DNA methylation can also impact on DNA methylation heterogeneity. Finally, erosion of DNA methylation upon cell division can result in disordered DNA methylation patterns.*

Several bioinformatic tools have been developed also to explore this additional layer of information from bisulfite sequencing data (Scala et al. 2016; Li et al. 2014; D. Lee et al. 2023; Hetzel et al. 2021), but they either lack the possibility to perform a comprehensive analysis of DNA methylation heterogeneity across multiple genomic loci or to perform downstream differential analysis between different biological conditions.

More in depth, tools leveraging targeted DNA methylation (Scala et al. 2016) enable the examination of methylation profiles through deep targeted bisulfite sequencing. These methods rely on the precision derived from targeted bisulfite experiments, offering a robust characterization of epiallele species within one or more

specific genomic regions. Additionally, they provide the opportunity to depict and analyze the composition of epialleles using population genetics-based approaches.

However, a notable limitation of these tools is their limited capability to analyze only a specific type of input data (e.g., FASTA files). Moreover, acquiring raw data from deep-targeted sequencing in public databases is challenging, limiting the reusability of public methylation datasets.

Conversely, tools that address these limitations by extracting information from genome-wide experiments, such as *methclone (Li et al. 2014)*, lack statistical functionalities for comparing the epiallele composition across different sample groups. This becomes pivotal when the analysis extends beyond the mere characterization of epiallele families in one or more regions across various samples. Indeed, such data can be instrumental in comparing epiallele distributions among sample groups associated with distinct biological conditions. It facilitates the identification of genomic regions with varying epiallelic compositions between conditions, potentially uncovering novel biomarkers that may not be detected using conventional quantitative approaches.

Several metrics have been employed to measure the level of DNA methylation heterogeneity from bulk samples. These metrics can be classified into two major categories, depending on whether they are employed to assess the methylation heterogeneity within (-intra) or between (-inter) the single molecules (Figure 2.5).

Intra-molecule metrics are adopted to quantify the correlation between the methylation states of neighboring CpG sites at the level of the single read. For example, the PDR (Proportion of Discordant Reads) score is employed as a measure of locally disordered methylation. When computing this measure, a bisulfite sequenced read is

considered concordant or not depending on the agreement of the methylation states of the CpG sites mapped by that read. Similarly, the read transition score (RTS) measures the probability that two consecutive CpG sites share the same methylation state, ranging from 0 (when all the CpG sites on the read show the same methylation state) to 1 (when there are no neighboring CpG sites sharing the same methylation state).

On the other hand, to measure the diversity of epialleles populations, inter-molecule scores have been employed. These metrics evaluate the variance of DNA methylation patterns displayed between the molecules in a given genomic locus. Among the other scores, the Shannon Entropy (SE) is the most common employed measure to assess this diversity of epialleles species at one analyzed genomic region. These scores generally are at their maximum values when no two reads display the same DNA methylation pattern (when all the possible epialleles combinations are found in one genomic region across the sequenced reads).

Figure 2.5 Intra- and inter-molecule heterogeneity scores. *The figure shows the different significance of intra- and inter-molecule scores used to measure DNA methylation heterogeneity between neighbor CpG sites on the same molecule, and between the different molecules at a given genomic locus. Several different scores are grouped in two major classes according to this distinction.*

2.7 Third-generation sequencing

Nevertheless, the deterioration of DNA induced by bisulfite conversion and the prevailing utilization of Illumina sequencing technology impose constraints on the

single-molecule examination of bisulfite sequencing data to abbreviated segments spanning only a few hundred base pairs. The restricted count of CpGs captured in these sequences imposes limitations on the exploration of intra-molecular DNA methylation heterogeneity (Kerr et al. 2023).

In contrast, third-generation sequencing such as nanopore sequencing facilitates the production of reads exceeding a megabase in length, a magnitude four times greater than those scrutinized by routine bisulfite sequencing experiments (Liu et al. 2021). Furthermore, these reads enable the direct detection of DNA modifications, including 5-methylcytosine, without resorting to chemical conversion (Liu et al. 2021).

This capability potentially opens avenues for analyzing molecular heterogeneity in DNA methylation patterns on a significantly broader scale than has been previously achievable (Figure 2.6).

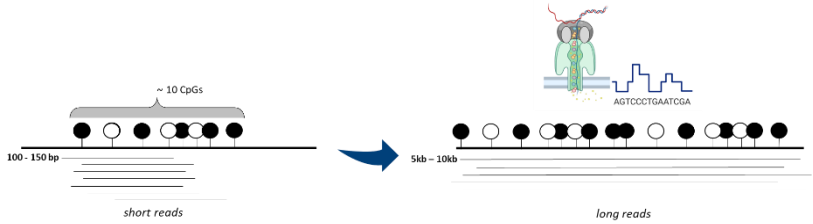


Figure 2.6 Short- vs Long-reads sequencing in the context of DNA methylation heterogeneity analysis. *Comparison of Bisulfite Sequencing and Nanopore Sequencing for DNA Methylation Analysis. Bisulfite conversion and Illumina sequencing technology, while widely used, limit the examination of DNA methylation to short segments due to DNA deterioration. This results in restricted CpG counts, hindering the exploration of intra-molecular DNA methylation heterogeneity. In contrast, third-generation sequencing, exemplified by nanopore sequencing, produces longer reads, exceeding a megabase, and allows direct detection of DNA modifications, including 5-methylcytosine, without chemical conversion. This capability potentially offers a broader scale for scrutinizing molecular heterogeneity in DNA methylation patterns compared to traditional bisulfite sequencing.*

3. Aims of the thesis

The goal of this PhD project was to develop a novel computational tool for the analysis of DNA methylation heterogeneity from bulk sequencing data provided with additional functionalities for a comprehensive analysis of biological datasets in order to detect differential heterogeneous loci.

Specific objectives:

1. Develop a R package provided with customizable functions for the extraction of DNA methylation heterogeneity information and statistical analysis from genomic intervals of choice.

2. Benchmarking the developed computational tool on two public bisulfite sequencing datasets for the detection of significant differentially heterogeneous regions.

3. Extend the workflow to third-generation sequencing methods for the assessment of DNA methylation heterogeneity at the large-scale.

4. Materials and Methods

4.1 Algorithm implementation

To increase the computational efficiency of the analysis, a specific function to identify genomic regions with an adequate coverage of reads in the alignment files has been implemented.

First, the coverage is computed across the entire genome at a single-base resolution. Subsequently, contiguous regions that meet a user-defined coverage threshold are merged, resulting in a collection of genomic ranges with varying lengths.

The first step in the workflow involves the partition of the covered areas into genomic intervals designed to satisfy user-defined criteria, as explained below.

Two distinct approaches have been employed for this step. In the first approach, genomic intervals are determined using sliding windows of flexible lengths containing a user-defined number of CpG/CpH sites. The window steps up by 1 CpG/CpH site at a time.

Alternatively, target regions are generated using a sliding window with a fixed length and step sizes defined by the user, encompassing a user-defined minimum and maximum number of CpG/CpH sites.

As the identified methylation patterns originate from single fragments, the maximum length of the regions is contingent on the read length of the sequencing experiment. For each region slated for profiling, the composition of epialleles is then extracted, as described below.

Initially, all sequenced reads mapped to the corresponding locus are selected. These reads are processed separately based on the strand to which they are aligned. The sequence of each read is then compared with the

reference genome at the positions of the CpG/CpH sites. A table is compiled row-wise, consisting of *n* cytosines (Cs) columns (where *n* is the number of Cs in the CpG/CpH context at one locus) and *r* rows (where *r* is the number of reads spanning the entire region considered). For each read and each CpG/CpH position, if a T (or an A on the reverse strand) is found corresponding to the reference C position, the corresponding position in the table is marked as 0 (unmethylated); otherwise, it is marked as 1 (methylated).

Depending on the user's criteria, the epiallele composition from different strands can be analyzed separately, allowing for a stranded epiallele estimation (required in case of CpH methylation). Alternatively, it can be evaluated independently of the originating strand.

4.2 Bisulfite conversion quality check

To ensure the accurate assessment of cytosine methylation status in the analyzed regions, the possibility to estimate bisulfite efficiency was included in the algorithm. In those experiments where non-CpG methylation is a minor phenomenon, it can be assumed that all cytosines besides the CG context are converted to thymines upon bisulfite conversion.

Consequently, bisulfite efficiency is computed for each sequenced molecule, representing the percentage of CHH cytosines (excluding those in the CG context) that have undergone conversion among all CHH cytosines within the covered genomic interval.

The algorithm was implemented in such a way that reads exhibiting low conversion efficiency can be optionally excluded from the epialleles count computed during the analysis. Subsequently, the coverage for each interval is recalculated, and only those regions meeting the user-defined criteria are retained for further analysis.

The same rationale applies when the pattern under examination through the epialleles analysis differs from CG. In this case, cytosines in the CG context have not been considered when calculating the conversion efficiency ratio of cytosines.

4.3 Handling reads ambiguity and polymorphisms

The ability to detect and remove from the analysis ambiguous reads has been added. When filling the binary epialleles matrix, if nor a T (or A) or a C is found in correspondence of a CpG/CpH site, the corresponding position in the matrix is filled with a 2. The rows (reads) of the matrix containing at least one 2 are subsequently removed from the count.

4.4 Output generation

From the binary epiallele matrix obtained from each analyzed locus, two distinct outputs are generated.

The **first output** is a compressed epiallele table. This results from the conversion of each row of the binary epiallele matrix into a string representing an individual epiallele species. The compressed table is then formed by recording the count for each unique epiallele species identified at the analyzed locus.

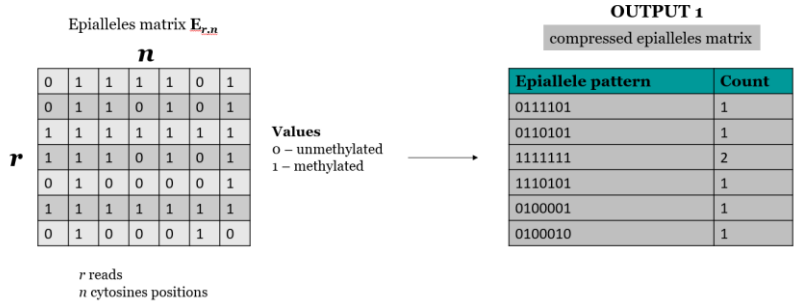


Figure 4.1 Scheme of output 1 generation. *The figure shows the generation of the first output implemented in the algorithm. This consists of a summarized epiallele matrix, which contains the counts of the epiallele species*

observed at a given locus. It is saved in the local disk as a text file containing two columns. The first column represents the precise epiallele pattern, and the count value returns the number of observations of that precise string.

The **second output** is derived by applying a customizable set of functions to the binary matrix. Each function computes a specific summary statistic, and the results are compiled into a data frame. Users have the flexibility to adjust both the number and type of functions applied to compute summary statistics over the epiallele binary matrix. Additionally, users can extend this functionality by incorporating custom-defined functions according to their specific requirements.

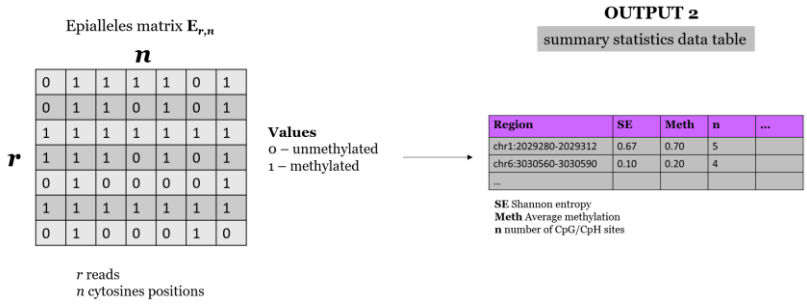


Figure 4.2 Scheme of output 2 generation. *The figure shows a graphical representation of the second output. It consists of a text file displaying as observations the genomic intervals used for the extraction of the epialleles information. For each genomic interval, several summary statistics are reported in the output. Summary statistics to be calculated are indicated by the user during the workflow (i.e., Shannon Entropy value).*

4.5 Summary statistics

The second output is generated by calculating summary statistics operating on the epialleles binary matrix. Different metrics have been developed to estimate DNA methylation heterogeneity within one bulk sample (see Chapter 1). We implemented functions for the calculation of some of these metrics. However, we implemented the generation of the second output so that the user can indicate his own formulas to calculate metrics other than the ones that can be calculated through the functions provided with

the package. These formulas can be provided as new functions in a list as parameter.

The functions to calculate the following metrics have been implemented as basic in the algorithm:

4.5.1 Number of CpG/CpH sites

The number of CpG/CpH (*c*) defines the set of the reference positions used to perform the extraction of the epialleles information at one analyzed region. The metrics can be useful for downstream analysis, especially when the sliding window approach based on a fixed length size is employed.

4.5.2 Number of reads

The number of reads (*r*) indicates the count of reads entirely spanning the genomic regions considered for the extraction of the epialleles composition.

4.5.3 Epiallele species count

The epialleles count value indicates the number of unique epialleles patterns observed in the analyzed regions.

4.5.4 Singleton

The singleton value returns the number of epiallele species found in one single copy among the observed epialleles computed in one analyzed interval.

4.5.5 Highest frequency species

The value indicates the exact string representing the epiallele patterns most represented among the epialleles computed at one given locus.

4.5.6 Mean distance between CpG/CpH sites

The mean distance between CpG/CpH sites displayed in each analyzed genomic intervals indicates the extent of

cytosines aggregation. It is computed using the following formula:

$$d = \frac{\sum_{i=1}^{n}(p_i - p_i + 1)}{n}$$

where

n is number of CpG/CpH – 1 sites, and

p is position (in bp) of the CpG/CpH site in the analyzed region.

4.5.7 Mean DNA methylation

The mean DNA methylation is computed from the binary epialleles matrix by calculating the proportion of methylated cytosines over the total number of cytosines detected at a given region by using the following formula:

$$\mu = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}}{m \times n}$$

where

m is number of rows in the epialleles matrix,

n is number of columns in the epialleles matrix, and

X_{ij} is the value at the intersection of the i-th row and the j-th column (either 0 or 1, indicating unmethylated or methylated, respectively).

4.5.8 Shannon entropy

The Shannon entropy serves as a metric for probabilistic uncertainty, providing an estimation of the heterogeneity within the population of epiallele species at a given locus. The values of Shannon entropy range from 0, indicating uniformity of epialleles patterns across the analyzed reads, to the number of CpG/CpH sites analyzed in

the given region, reflecting maximum diversity when all possible epiallele states are observed with equal frequency.

For a region with n CpG/CpH sites, the Shannon Entropy is computed using the following formula:

$$h = -\sum_{k=1}^{e} p_k \times \log_2 p_k$$

where

e is defined as the maximum number of distinct epiallele species that can be observed in one locus ($2^n, where\ n\ is\ the\ number\ of\ cytosine\ in\ the\ locus$), and

p_k defines the frequency of the k-th epiallele.

4.6 Statistical testing

The comparison of epiallelic compositions across various groups at multiple genomic regions employs a permutational analysis of variance (PERMANOVA). Utilizing the epiallele composition matrix for each sample at a specific locus, first the computation of distances between every pair of samples in the dataset is performed. The test then discerns whether distances among samples from different groups surpass those observed within the same group. Significance is determined by assessing the likelihood that the observed difference within sample data could occur through random allocation of samples to different groups across a series of permutations (n).

Subsequently, each analyzed genomic region is assigned with a *p-value*, with adjusted *p-values* calculated using the false discovery rate (FDR) method. Implementation of this test utilizes the *adonis2* function from the *vegan* R package, renowned for its tools in descriptive community ecology. Post-hoc analyses, comparing epiallelic composition matrices among more than two groups, are conducted using the *pairwise.adonis* R package. Data visualization incorporates dimensionality

reduction through Principal Component Analysis (PCA) and Canonical Correspondence Analysis (CCA).

For the comparison of user-defined summary statistical distributions, a non-parametric test assesses the statistical significance of differences among groups. The Wilcoxon test is applied when only two groups are available for comparing genomic regions; otherwise, the Kruskal-Wallis test is adopted.

Lastly, to compare mean summary statistic values while considering a covariate (e.g., time), an analysis of covariance (ANCOVA) is implemented.

4.7 RRBS raw data processing

RRBS data used for the benchmarking were first downloaded as raw sequenced data from SRA using the GEO accession GSE147156 (considering only the samples collected from the striatal tissue) and GSE48975.

Fastq files were processed using an in-house built pipeline. First, raw reads were quality checked using FastQC v0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Adapters and low quality bases were removed using TrimGalore v0.6.6 using the --rrbs option (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/).

Filtered reads were then aligned to the mm10 reference genome by using Bismark (Krueger and Andrews 2011) v0.23.0 (https://github.com/FelixKrueger/Bismark) using default parameters. Finally, BAM files were sorted and indexed using SAMtoolsKit (https://www.htslib.org/).

4.8 Long reads (ONT) raw data processing

Raw fast5 files were downloaded from Oxford Nanopore amazon bucket (s3://ont-open-data/rrms_2022.07). Reads were first basecalled using the dna_r9.4.1_450bps_hac.cfg configuration files from Guppy v.5.0.11. Basecalled reads were then aligned to the hg38 reference human genome using minimap2 v.2.22 using map-ont parameters. BAM files were sorted and indexed using SAMtoolsKit (https://www.htslib.org/). Methylation calling was performed using nanopolish (https://github.com/jts/nanopolish). Nanopolish index command was used to prepare input data linking reads id with the signal-level data in the fast5 files. Nanopolish call-methylation command was then used to call 5mC. The calling of methylation states to individual CpG sites was determined using the log-likelihood ratio (LLR) calculated by Nanopolish (Simpson et al. 2017). If the LLR was ≤ -2, CpG sites were designated as unmethylated (0), while LLR ≥ 2 indicated a methylated status (1). CpG sites with |LLR| < 2 were considered unassignable (NA).

4.9 Heterogeneity metrics for long reads data

To analyze DNA methylation heterogeneity at the large-scale, a different set of metrics was employed.

The read transition score (RTS) was employed to assess DNA methylation heterogeneity at the single-molecule level (intra-molecule score). RTS indicates the probability that two neighbors CpG sites share the same methylation state, and was calculated as follows:

$$\mathrm{RTS}_r = \frac{\sum_{i=1}^{N-1} |v_{r,i+1} - v_{r,i}|}{N-1}$$

To assess inter-molecule DNA methylation heterogeneity, the quantitative fraction of discordant reads pair (qFDRP) was employed. Once fixed the genomic

window to analyze, and filtered out not informative reads, the score was calculated using the following formula:

$$\mathrm{qFDRP_c} = \frac{\sum_{r_s \in R_c} \sum_{r_{t,t>s}} \frac{\sum_{i \in \{r_s \cap r_t\}} I(x_{i,r_s} \neq x_{i,r_t})}{|\{r_s \cap r_t\}|}}{\binom{|R_c|}{2}}$$

where

R_c is the set of all reads r covering c

r_s, r_t are the sets of CpG positions (reads), and

$x_{i,r}$ is the methylation state of CpG i in read r.

4.10 Association with genes

Significant differentially heterogeneous regions were associated with closest genes using the *regsToPathway* from EpiStatProfiler. The code was implemented from the *annotatePeak* function from the ChIPseeker R package. Each region was annotated defining as promoters the regions ± 1000 kb flanking the transcription start site (TSS). Each annotated regions was labeled as one of the following: "Distal intergenic", "Promoter", "Exonic", "Intronic", "3' UTR" and "5' UTR".

4.11 Differentially heterogeneous regions (DHR) detection

Differentially heterogeneous regions (DHR) were detected using first *diffStat* from EpiStatProfiler to detect significant loci which differ for their Shannon Entropy values between the two conditions (HD vs wild-type mice and Ctcf -/+ vs wild-type mice), and then the *epiStat* function from EpiStatProfiler to detect significant loci which differed for their epialleles compositions.

4.12 Differentially methylated positions (DMP) detection

Quantitative differential analysis was performed using the *methylKit* R package. Methylation calls were performed starting from the aligned BAM files using the *processBismarkAln* function, by considering only CpG sites covered by at least 10 reads and detected in at least 3 samples per group.

4.13 Enrichment analysis for genomic regions

To evaluate the enrichment of CTCF binding sites within the set of significant genomic regions obtained from the analysis of dataset n.2 (GSE48975), CTCF ChIP-seq peaks of mouse lung tissues were retrieved from ENCODE (accession numbers ENCFF491RJK and ENCFF605YVN). A consensus track was generated from the two tracks and finally used to test the overlap with the bed file of the significant regions using *bedtools fisher*(https://bedtools.readthedocs.io/en/latest/index.html).

5. Results

5.1 EpiStatProfiler: a novel R package for the qualitative analysis of DNA methylation

As a result of this PhD project, an innovative workflow for the qualitative analysis of DNA methylation from bulk bisulfite sequencing data, named EpiStatProfiler (Figure 5.1), is presented (Sarnataro et al. 2022). The toolkit comprises a suite of dedicated functions for extracting and subsequently for the statistical analysis of epialleles compositions from bulk samples at multiple genomic regions. EpiStatProfiler stands as an open-source R package, accessible on GitHub (https://github.com/BioinfoUninaScala/epistats).

The package is provided with functions which can be classified in distinct categories (Figure):

a. Input loading

The first module comes with functions designed to load input data and to filter regions which meet user's depth requirements for downstream analysis. Paths to file location are accepted as parameters to load BAM files and a FASTA reference genome into the R environment. The user can then use an additional function to select only the desired chromosome from the alignment files.

b. Design of genomic intervals

The second module consists of functions used to design genomic intervals to extract the epialleles composition from. Two main functions have been implemented to design loci adopting one of two different approaches as described in the Methods section.

The first approach consists of using a sliding window of variable length, containing a different set of CpG sites.

The second approach, instead, relies on the usage of a sliding window characterized by a fixed length and step size set by the user.

Irrespective of the selected system, the user is required to indicate the pattern intended for exploration within the genome to construct the analysis window set. In the case of CG methylation, this parameter is set as 'CG'; alternatively, it can be designated as 'CA', 'CC', or 'CT'. Notably, the functions responsible for generating these regions have been developed to provide the genomic coordinates corresponding to the analyzed cytosines.

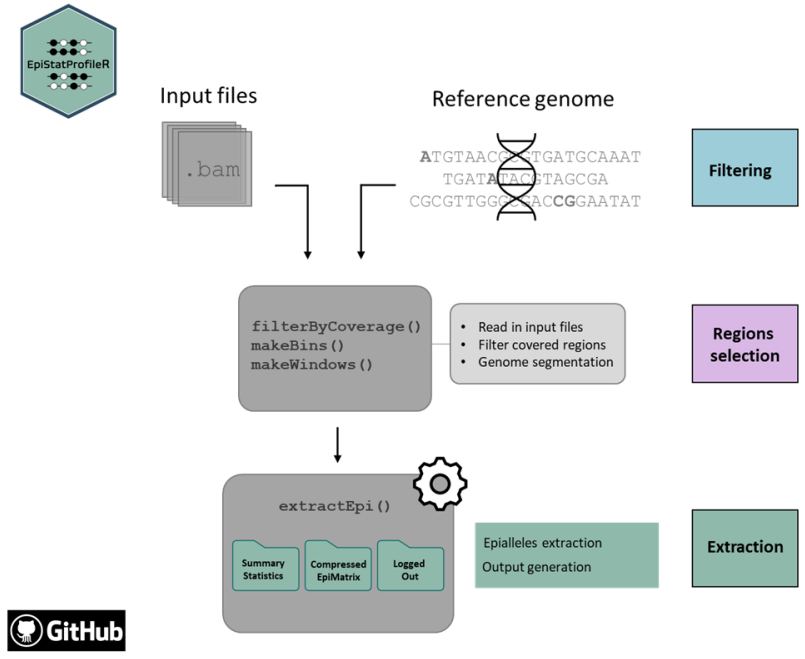


Figure 5.1 EpiStatProfiler workflow. *The main steps of the novel EpiStatProfiler workflow are described. The input data needed to run the analysis are the bam files containing the aligned reads and a reference genome to be provided in a FASTA format. The first functions to be employed are filtering functions. First, only the covered regions are used for the downstream analysis. Then, the user can select specific regions of interest (or chromosomes) to run the subsequent functions. After the filtering step, regions design is allowed through two different approaches: makeBins bins the covered regions into intervals of different lengths but containing the same number of CpG sites*

to be analysed. Otherwise, makeWindows can be employed to bin the regions in intervasl of equal size, containing a different set of CpG sites. Next, the epialleles information, together with the summary metrics, is extracted from each sample provided at the beginning. Once the epialleles information is extracted from all the samples, downstream statistical tests are provided to perform a differential analysis in order to detect relevant loci which differ in their epigenetic heterogeneity between biological conditions.

c. Epialleles extraction

The third module consists of the core functions aimed at the extraction of epiallele composition from each region obtained in the preceding step. Several parameters are retained as customizable, so that the user can adjust them according to the specific goal of the analysis. This includes for example the option to eliminate from the epiallele composition those reads that fail to meet a specified threshold of bisulfite conversion.

Extraction functions return two distinct outputs for each scrutinized region. The first output consists of a compressed epiallele matrix, which summarizes the epiallele composition by computing the frequency of each epiallele within that region. The second output is instead a data table, which reports the numerous summary statistics calculated on the raw binary epiallele matrix used as input. The user can provide functions to calculate metrics of interest with a list of one or more statistics selected from those available in EpiStatProfiler. Additionally, the user can add custom functions that implement other statistics of particular interest.

Optionally, a third output can be produced, encompassing the regions excluded from the analysis. Each output is systematically saved as a text file on the local disk.

d. Statistical analysis

The last module is provided with statistical functions for downstream differential analysis. These functions were implemented considering two different statistical tests to be

performed taking as input the two distinct outputs generated from the previous module's function.

The first set of functions take as input the compressed epiallele matrix. The test is employed in a way that the epialleles observed in one region are considered as a population of individuals (represented by the set of reads covering that region) belonging to different species (represented by the different epiallelic conformations). The *epiStat* function performs the PERMANOVA, as described in the Methods section. The input data needed to run these functions are the list of the compressed epialleles matrices from all the samples and a table containing sample metadata. The user must then indicate which columns are the ones containing the group and the sample information. Additional parameters were provided to perform the analysis, such as the minimum number of samples in each group required to carry out the analysis. The function returns as output a data frame containing all the regions analyzed as rows, along with the results of the statistical test.

Additional functions are provided to perform tests on single regions data. The *pairtest* function can be used to perform the pairwise comparison (see Methods) when more than two groups are present, in order to detect which group is mostly contributing to the dissimilarity observed in the tested locus. Finally, a function is provided to identify the epiallele species driving the dissimilarities between the groups in each genomic region. Both these functions require as one parameter the ID of the region to be analyzed.

The second set of functions take as input the second output instead. Two functions were implemented for comparing summary statistics values among groups: *diffStat* and *diffModel*. Both functions take a table with summary statistics for each analyzed interval as input.

The *diffStat* function identifies regions differing among provided groups for a specified statistic. Users can choose the statistic for the test, and the output is a table with region IDs, test results, and (adjusted) p-values. Additional output includes the test name (see Methods section) and median values of the input statistic for each group.

On the other hand, the *diffModel* function conducts an analysis of covariance (ANCOVA) as outlined in the Methods section. This is useful for evaluating group differences for one statistic across different levels of another variable (e.g., different time points). Users provide the dependent variable and covariate. Output is a table with region IDs, statistical test results, p-values for single variables, interaction term p-values, and coefficients of the fitted model. Slopes of the regression lines for each group are also included.

Results from these functions offer different and complementary biological insights. Functions handling epiallele composition compare general differences, while summary statistics functions compare specific aspects such as epiallele clonality, richness, or the presence of rare epialleles.

e. Dimensionality reduction

An additional module is composed of functions performing dimensionality reduction. The are two main functions implemented in the package for the visualization of the samples. The user can select region of interest to generate ordination plots. The two available functions, *runPCA* and *runCCA*, perform the two distinct ordination analyses described in the Methods section, the PCA and CCA respectively. Both functions accept as input the compressed epiallele matrix along with the sample metadata.

f. Annotation

Finally, to make a biological sense of the significant regions obtained from the previous steps, the *regsToPathway* function is provided to associate significant regions with genes. The function returns a table containing several annotation labels related to each region-gene association, such as the distance to the TSS, the annotation type (Exonic, Intronic, Intergenic, …) and the gene symbol. The output can then be used to perform additional downstream analysis, such as an GO terms enrichment analysis.

5.1.1 EpiStatProfileR allows the identification of relevant genomic differentially heterogeneous loci associated with distinct phenotypes

As proof of concept, we benchmarked the *EpiStatProfiler* R package on two public datasets, consisting of a RRBS and an enhanced RRBS sequencing experiment.

The first datasets consisted of heterozygous Htt knock-in (Q175) and wild-type (Q20) mice, collected for an experimental setting aimed at discerning the potential methylation signatures in mouse models carrying the causative mutation of Huntington's disease.

The size of the dataset consisted of 8 samples from each group, obtained from the striatal tissue of 3-weeks-old mice. First, we selected only those genomic regions covered by a minimum of 30 reads from each input BAM file. Subsequently, using the *makeWindows* function, we designed genomic intervals of a fixed 70 bp (based on the library read length) length size. We choose both CG and CA patterns parameters to compute the epialleles composition at these methylation contexts, separately.

To identify genomic regions exhibiting differential epiallele composition between the Huntington's and wild-type samples, we employed the *diffStat* function. In

particular, we used the Shannon entropy metrics as variables to perform the Wilcoxon non-parametric test. Statistically significant loci were determined by filtering regions with a *p-values* <= 0.05 and an absolute median Shannon entropy delta >= 0.10 between the two groups. Overall, 370 regions (Figure 5.2.a) were found to be significantly different between HD and wild-type mice considering the CG methylation context, while 183 (Figure 5.2.a) regions were found to be significant when looking at the CA methylation (belonging to the + strand).

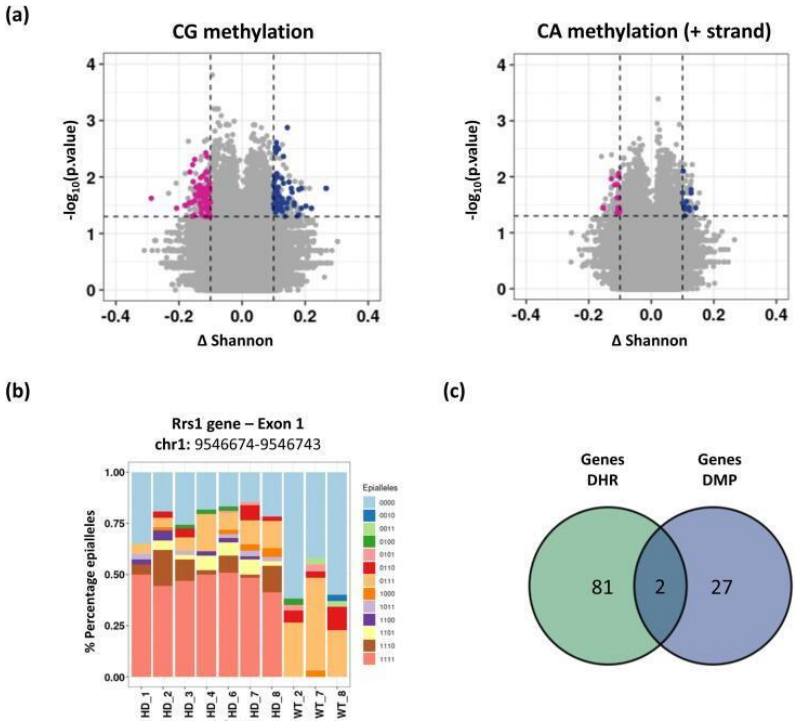


Figure 5.2 Differentially heterogeneous regions between HD and wild-type mice. *(a) Relevant loci of HD mice are shown using volcano plot. Loss and gain in clonality composition are marked in violet-red and blue respectively. The relevant loci were obtained by filtering the regions having a p.value ≤ 0.05 and being characterized by an absolute median Shannon entropy difference ≥0.10 between the two groups and further selected for being characterized by a significantly different epialleles composition among the two groups (PERMANOVA p.value ≤ 0.05). (b) A significant region obtained from the filtering processes used as proof of concept. Regions coordinates and functional annotation are reported on the top of the barplots. (c) Overlap of the genes associated with significant regions obtained performing the qualitative*

(DHR = differentially heterogeneous regions) and the quantitative (DMP = differentially methylated positions) analyses, respectively, within the classical CG context.

Significant regions were further filtered by overlap with the regions which were found to be significantly different as result of the other statistical test performed using the *epiStat* function, which directly compares the epialleles composition between the groups. We identified 135 regions (Figure 5.2.a, colored dots) which differed in their epialleles patterns heterogeneity between the HD and wild-type mice, and 30 regions in the CA methylation context (Figure 5.2.a, colored dots). As proof of concept, we reported an example of one significant region showing increased heterogeneity levels in HD mice compared to wild-type mice (Figure 5.2.b). Of note, this region was found to be associated with the Rrs1 gene, whose altered expression is commonly described in knock-in mice HD models. Next, we associated significant regions with genes (coding and non-coding) and retained only those (n = 102 for CG context and n = 11 fort CA context) annotated in functional domains (Promoter = 45, Intron = 25, Exon = 26, UTRs = 6 in CG context and Promoter = 1, Intron = 8, Exon = 2 in CA context). Genes found to be associated with our significant regions in both CG and CA contexts (CG genes = 83, CA genes = 11) were then used to perform a differential gene set enrichment analysis using the *g:Profiler* R package to identify specific biological processes. We found significantly enriched GO biological terms implicated in the disease pathogenesis looking at both CpG and CpA methylation signatures (see Table 1).

	Source	Term name	Term ID	Adjusted p.value
CG methylation	GO:BP	embryonic morphogenesis	GO:0048598	2.100×10^{-2}
CG methylation	GO:BP	cell differentiation	GO:0030154	3.949×10^{-2}
CG methylation	GO:BP	cellular developmental process	GO:0048869	4.500×10^{-2}
CG methylation	GO:BP	regulation of cellular metabolic process	GO:0031323	4.661×10^{-2}
CG methylation	GO:BP	cerebellar molecular layer development	GO:0021679	4.980×10^{-2}
CA methylation	KEGG	Glutathione metabolism	KEGG:00480	3.058×10^{-2}
CA methylation	WP	Oxidative stress and redox pathway	WP:WP4466	1.880×10^{-2}

Table 1. Results of the gene set enrichment analysis.

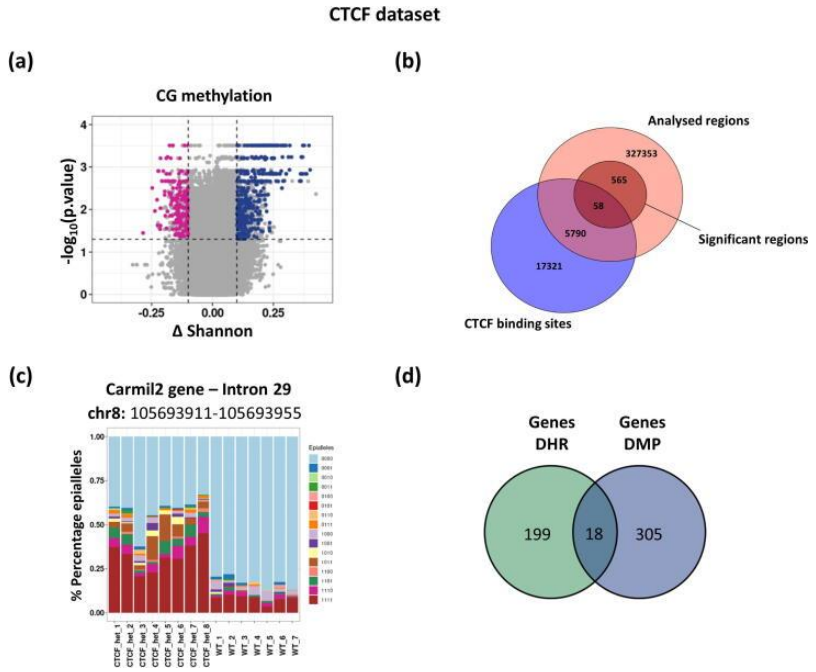


Figure 5.3 Differentially heterogeneous regions between Ctcf -/+ heterozygous and wild-type mice. *(a) Differentially heterogeneous epigenetic regions between wild-type and Ctcf +/– mice are highlighted using a volcano plot. Loss and gain of epigenetic heterogeneity are marked in violet-red and blue respectively. The relevant loci were obtained by filtering the regions having a p.value ≤ 0.05 and being characterized by an absolute median Shannon entropy difference ≥0.10 between the two groups and further selected for being characterized by a significantly different epialleles composition among the two groups (PERMANOVA p.value ≤ 0.05). (b) Scaled Venn diagram showing the enrichment of CTCF binding sites in regions showing statistically significant epiallele heterogeneity between the wild-type and Ctcf +/– samples. (c) Barplots showing the proportion of observed epialleles (different colors) in all the analyzed samples in a selected significant region (coordinates and annotation shown on the top of the barplots). (d) Venn diagram showing the overlap between the genes associated with significant regions from the qualitative (DHR = differentially heterogeneous regions, green) and the quantitative (DMP = differentially methylated positions, blue) analyses within the CpG context.*

The same workflow was then applied to second dataset consisting of whole lung tissue samples from *Ctcf* homozygous knockout (Ctcf +/-, n = 8) and wild-type (n = 7) mice.

To identify genomic regions undergoing epialleles composition heterogeneity differences between Ctcf +/- and

the wild-type mice we used both the *diffStat* and the *epiStat* functions, as described above. Here, we could identify 623 significant regions which differed for their Shannon entropy and their epialleles composition between the two conditions in the CpG context, with the majority (n = 435/623) showing an increase in Shannon entropy (SE) levels in Ctcf depleted samples. We then associated the above reported significant regions with genes and retained only those (n = 549 for CpG context) annotated in functional domains (Promoter = 97, Intron = 88, Exon = 227, UTRs = 137). We then used the genes found to be associated with our significant regions (CG genes = 217) to perform a differential gene set enrichment analysis using *g:Profiler*. Of note, we could identify an enrichment of genes described as CTCF targets. To better explore this observation, we assessed whether the obtained significant genomic regions were enriched for CTCF binding sites. In particular, we tested their overlap with CTCF-binding regions derived from mouse whole lung tissues from ENCODE and found a statistically significant enrichment (Fisher's exact test p.value < 2.5078e-76).

Finally, to compare our results with those obtained from a classical quantitative analysis, a differential methylation analysis was performed for both datasets (see Methods, *Differentially methylated positions*). In particular, we analyzed cytosines in the CpG context covered by at least 10 reads whose methylation status could be assessed in at least 3 samples per group. We then selected differentially methylated positions by filtering the sites showing a *qvalue* ≤ 0.05 and a methylation difference percentage > 25 % among the two groups (WT and HD, WT and Ctcf +/−). Referring to the canonical CpG context, we evaluated the overlap of the genes associated with the significant regions identified through the two different analyses (Genes DHR vs Genes DMP). We found that the two approaches detected significant loci that had some overlaps, showing that the qualitative analysis can capture

additional methylation changes events that could be missed by standard DNA methylation analyses (Figure 5.3.d).

5.2 Extension of the DNA methylation heterogeneity analysis to long-reads sequencing

To analyze DNA methylation heterogeneity at wider genomic contexts, we estimated DNA methylation heterogeneity at the genomic features mappable through the available data. As genomic features to be considered, we elected gene promoters, CG islands, CG shores and CG shelves. We further selected only those features which were mapped by at least 10 long-reads and retaining only those reads which displayed the 30% of NA called sites at the most (see Methods). We then used both intra-molecule and inter-molecule scores to evaluate DNA methylation heterogeneity among CG sites on the same molecule and the heterogeneity of the methylation patterns among the molecules at the previously defined genomic contexts.

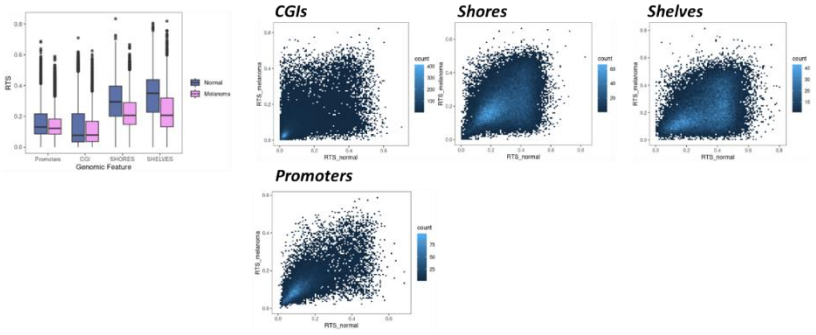


Figure 5.4 Comparison of the RTS score between normal e melanoma cell line samples. *Boxplot of RTS scores in normal e melanoma cell line pair (A). Scatterplot showing the correlation between the two measures in distinct genomic contexts (CG islands, CG shores, CG shelves, and promoter regions)*

We retrieved heterogeneity information as described above from two COLO829BL normal and one melanoma COLO829 cell line samples. By correlating heterogeneity scores between the normal and the melanoma samples, we could observe large differences between the two conditions,

to various extent between the different genomic contexts (Figure 5.4). To confirm that the changes in the heterogeneity were not associated with noise in the data, we evaluated the correlation between the heterogeneity scores between the two normal samples and showed that the methylation heterogeneity values were highly correlated, showing high stability of the large patterns within samples (Figure 5.5).

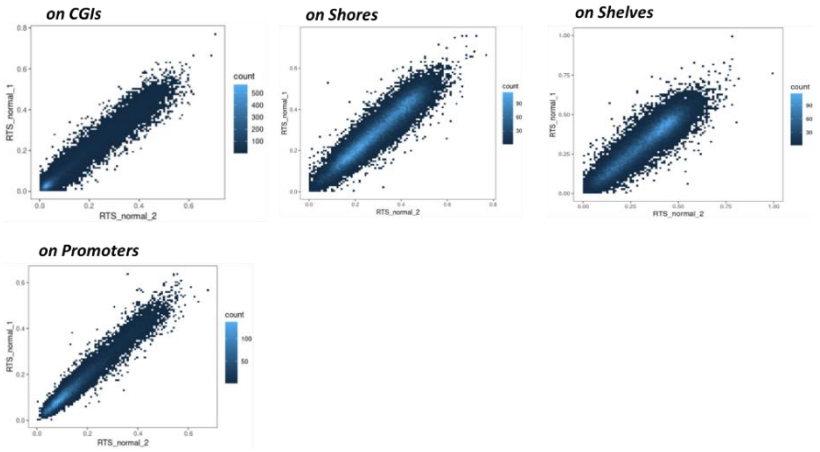


Figure 5.5. Comparison of RTS-score between two normal samples. *Scatterplot showing the correlation between the RTS values of two normal cell line samples.*

Of interest, as shown by one CGI promoter region 4kb long, we could demonstrate the power of long reads sequencing to better examine the alterations of methylation heterogeneity at large domains and in the context of genomic context. Here, we could describe hypermethylation of the CGI promoter regions, with high methylation heterogeneity levels within the boundaries of the CGI islands, whereas the flanking regions undergo hypermethylation at a different dynamic, showing high ordered patterns between the molecules (Figure 5.6).

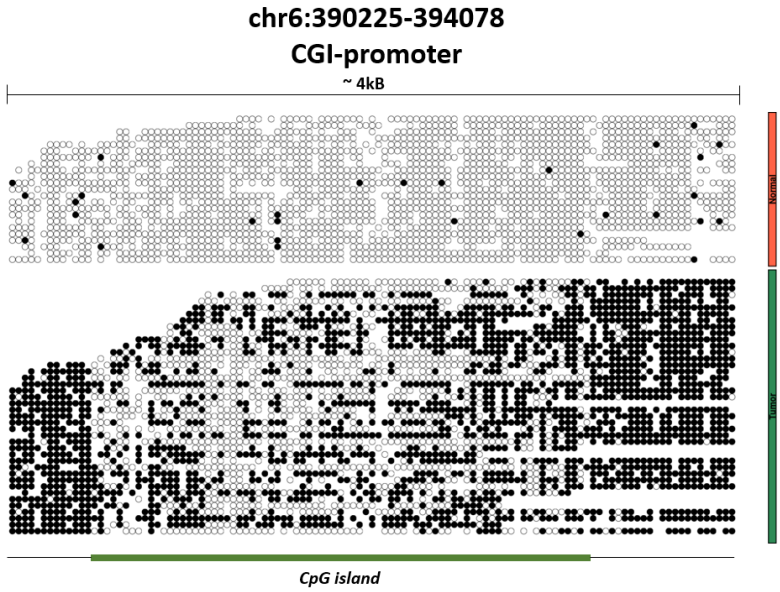


Figure 5.6 Example of one region showing differential DNA methylation heterogeneity between normal and melanoma cell line samples. *Lolliplot shows a 4kb promoter region displaying differential heterogeneous patterns between the CG island and its flanking regions in the tumor sample.*

6. Discussion

Epigenomes are characterized by a dynamic and reversible nature (Carter and Zhao 2020). This high plasticity has a pivotal role in orchestrating main biological processes, including cell differentiation and senescence programs. Variability of epigenetic modifications patterns is the main source cells adopt to acquire new diverse functions (Carter and Zhao 2020). Alterations of epigenetic variability have been widely described as hallmarks of cancer cells, being able to better adapt and survive common treatments (Landau et al. 2014; Landan et al. 2012).

Cell-to-cell epigenetic variability is the most investigated expression of the epigenetic plasticity. Single-cell assays have now been developed to analyze this phenomenon at high resolution (Huan et al. 2018; Karemaker and Vermeulen 2018). Nonetheless, these assays remain difficult to be employed for large scale studies, mostly due to their high costs (Teschendorff et al. 2020).

Still, epigenetic variability information can be retrieved from bulk bisulfite sequencing data, by looking at the specific configuration of methylation states of the cytosines at each sequenced read, using this measurement as a proxy of single cell epigenetic analysis (Li et al. 2014; D. Lee et al. 2023; De Riso et al. 2022). Each of these configurations is usually referred to as *epiallele*, and the diversity of these patterns displayed at one genomic locus can be used as a measure of epigenetic heterogeneity.

In this context, other studies have demonstrated that assessment of DNA methylation heterogeneity can be particularly relevant in distinguishing pathophysiological conditions (Scherer et al. 2020). For example, the extent of DNA methylation heterogeneity at promoter regions in leukemic cancer cells has been proven as a valuable

biomarker of bad prognosis in CLL patients (Landau et al. 2014).

Several computational tools have been developed for the computation of DNA methylation heterogeneity from bulk bisulfite sequencing data (Li et al. 2014; Hetzel et al. 2021; D. Lee et al. 2023).

The major goal of this PhD thesis project was the development of a novel computational tool, designed for the comprehensive analysis of DNA methylation patterns heterogeneity from bulk sequencing experiments.

We developed EpiStatProfileR (Sarnataro et al. 2022), a novel R package for a comprehensive analysis of DNA methylation heterogeneity from bulk sequencing experiments, provided with additional statistical functions for the detection of putative loci relevant to distinguish biological conditions.

EpiStatProfiler enables the estimation and comparison of the distribution of the distinct methylation patterns (epialleles) at a given genomic locus. This qualitative perspective provides valuable insights into cell-to-cell epigenetic heterogeneity, a dimension often overlooked in standard quantitative analyses.

One notable strength of EpiStatProfiler is its versatility, allowing extraction of epialleles from various bisulfite sequencing data types. In extension, EpiStatProfiler offers dedicated functions to select, filter, and analyze genomic regions, making it particularly useful for wide low-coverage assays such as WGBS and RRBS.

Comparing EpiStatProfiler with existing tools (see Figure 6.1), the package allows epiallele extraction and it offers a range of statistical tests to identify significant loci for further investigation as potential epigenetic biomarkers. This is a notable improvement, as many existing tools lack comprehensive analyses and dedicated statistical tests for

comparing epiallele compositions across different experimental conditions.

	Language	Input data	Experiment	Stranded analysis	Non-CpG methylation	Statistics	Bisulfite efficiency QC
EpiStatProfiler	R	BAM files	Genome-wide, Targeted	✓	✓	✓	✓
mHapTools	C, Command line	BAM files	Genome-wide	✗	✗	✗	✗
methclone	C++	BAM files	Genome-wide	✗	✗	✗	✓

Figure 6.1 Comparison of EpiStatProfiler with other computational tools for the analysis of DNA methylation heterogeneity. *The main functionalities of EpiStatProfiler are highlighted in the figure. Compared with other tools, EpiStatProfiler offers the possibility to perform a comprehensive analysis of DNA methylation patterns in any CpX methylation context, providing the user with additional statistical functions which aim to detect significant loci which differ for their heterogeneity between distinct biological conditions.*

Moreover, EpiStatProfiler introduces novel functionalities which are not present in other tools. For instance, it allows the analysis of epiallele composition based on non-CpG sites, accommodating the study of DNA modifications like CA methylation, particularly relevant in certain biological contexts such as brain tissues. Additionally, EpiStatProfiler accommodates diverse experimental designs, providing statistical tools for identifying regions with differential epiallele compositions in two or more sample groups or time points.

By applying EpiStatProfiler to RRBS-based dataset, we demonstrated its effectiveness in identifying epigenetic signatures associated with disease pathogenesis in Huntington mice models. The enrichment analysis revealed distinct sets of GO biological terms associated with the HD condition in both CG and CA methylation contexts.

Furthermore, our tool proved valuable in identifying genomic regions with significant changes in epiallele heterogeneity in the lung tissue of mice with genetic disruption of one Ctcf allele. The results suggest a potential

role of CTCF in maintaining cellular expression stability by stabilizing promoter-enhancer interactions.

Therefore, EpiStatProfiler stands as a comprehensive and versatile tool for characterizing epiallele composition in various biological systems. It provides researchers with a complementary perspective to explore epigenetic information. EpiStatProfiler is publicly available at https://github.com/BioinfoUninaScala/epistats.

Finally, by extending the EpiStatProfiler workflow to long reads sequencing experiments, we analyzed DNA methylation heterogeneity at large genomic contexts in a normal-tumor cell line pair. As preliminary results, it allowed the identification of defined genomic regions undergoing major changes in methylation heterogeneity levels within the borders of established genomic features.

Overall, this work contributed to the existing literature by providing additional tools for dissecting DNA methylation patterns at genome wide level and for the identification of putative relevant loci undergoing major epigenetic alterations in biological conditions.

Acknowledgements

I would like to thank my supervisor prof. Sergio Cocozza for mentoring me, especially during the first part of PhD. Thanks to him, I got to know more about statistics, evolutionism, classic music and Neapolitan culture.

I thank my colleague Giulia, who was also my mentor at the beginning of this journey. I am thankful for the time we spent together in the lab dealing with bioinformatics and life issues.

I would also like to thank my co-supervisor Giovanni Scala, for his valuable contribution to my training as computational biologist, and for his patience in this process. It is basically his fault that I have now an identity crisis.

I thank my host supervisor Pavlo Lutsik and the students from his group in Heidelberg for their help during the last year.

I thank Francesca, which is still supporting me as it was since I started my master's internship. I am glad I can call you my friend now and always.

Beside the people who scientifically contributed to this project and to my scientific education, I would like to thank also the ones that made me keep on moving during these years.
At this point I don't really know what I am, but I am grateful for the people I have around me that are helping me trying to figure it out.

List of publications

a. 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 remodeling of Fxyd1 promoters in developing heart and brain tissues. Sci Rep 12, 6471 (2022). https://doi.org/10.1038/s41598-022-10365-y

b. Sarnataro A., De Riso G., Cocozza S., Pezone A., Majello B., Amente S., and Scala G., A novel workflow for the qualitative analysis of DNA methylation data, Comput Struct Biotechnol J. 2022; 20: 5925–5934. doi: 10.1016/j.csbj.2022.10.027.

c. 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

Bibliography

Abante, J., Y. Fang, A. P. Feinberg, and J. Goutsias. 2020. “Detection of Haplotype-Dependent Allele-Specific DNA Methylation in WGBS Data.” *Nature Communications* 11 (1): 5238.

Akalin, Altuna, Matthias Kormaksson, Sheng Li, Francine E. Garrett-Bakelman, Maria E. Figueroa, Ari Melnick, and Christopher E. Mason. 2012. “methylKit: A Comprehensive R Package for the Analysis of Genome-Wide DNA Methylation Profiles.” *Genome Biology* 13 (10): R87.

Andrews, Simon, Christel Krueger, Maravillas Mellado-Lopez, Myriam Hemberger, Wendy Dean, Vicente Perez-Garcia, and Courtney W. Hanna. 2023. “Mechanisms and Function of de Novo DNA Methylation in Placental Development Reveals an Essential Role for DNMT3B.” *Nature Communications* 14 (1): 1–12.

Assenov, Yassen, Fabian Müller, Pavlo Lutsik, Jörn Walter, Thomas Lengauer, and Christoph Bock. 2014. “Comprehensive Analysis of DNA Methylation Data with RnBeads.” *Nature Methods* 11 (11): 1138–40.

Bock, Christoph, Eleni M. Tomazou, Arie B. Brinkman, Fabian Müller, Femke Simmer, Hongcang Gu, Natalie Jäger, Andreas Gnirke, Hendrik G. Stunnenberg, and Alexander Meissner. 2010. “Quantitative Comparison of Genome-Wide DNA Methylation Mapping Technologies.” *Nature Biotechnology* 28 (10): 1106–14.

Carter, Benjamin, and Keji Zhao. 2020. “The Epigenetic Basis of Cellular Heterogeneity.” *Nature Reviews. Genetics* 22 (4): 235–50.

Chen, Taiping, Naomi Tsujimoto, and En Li. 2004. “The PWWP Domain of Dnmt3a and Dnmt3b Is Required for Directing DNA Methylation to the Major Satellite Repeats at Pericentric Heterochromatin.” *Molecular and Cellular Biology*, October. https://doi.org/10.1128/MCB.24.20.9048-9058.2004.

Dahlet, Thomas, Andrea Argüeso Lleida, Hala Al Adhami, Michael Dumas, Ambre Bender, Richard P. Ngondo, Manon Tanguy, et al. 2020. “Genome-Wide Analysis in the Mouse Embryo Reveals the Importance of DNA Methylation for Transcription Integrity.” *Nature Communications* 11 (1): 3153.

De Riso, Giulia, Antonella Sarnataro, Giovanni Scala, Mariella Cuomo, Rosa Della Monica, Stefano Amente, Lorenzo Chiariotti, Gennaro Miele, and Sergio Cocozza. 2022. “MC Profiling: A Novel Approach to Analyze DNA Methylation Heterogeneity in Genome-Wide Bisulfite Sequencing Data.” *NAR Genomics and Bioinformatics* 4 (4): lqac096.

“DNA Methylation: A Historical Perspective.” 2022. *Trends in Genetics: TIG* 38 (7): 676–707.

Edwards, John R., Olya Yarychkivska, Mathieu Boulard, and Timothy H. Bestor. 2017. “DNA Methylation and DNA Methyltransferases.”

Epigenetics & Chromatin 10 (May): 23.
Gibney, E. R., and C. M. Nolan. 2010. “Epigenetics and Gene Expression.” *Heredity* 105 (1): 4–13.
Goldberg, Aaron D., C. David Allis, and Emily Bernstein. 2007. “Epigenetics: A Landscape Takes Shape.” *Cell* 128 (4): 635–38.
Greenberg, Maxim V. C., and Deborah Bourc’his. 2019. “The Diverse Roles of DNA Methylation in Mammalian Development and Disease.” *Nature Reviews. Molecular Cell Biology* 20 (10): 590–607.
Gu, Hongcang, Zachary D. Smith, Christoph Bock, Patrick Boyle, Andreas Gnirke, and Alexander Meissner. 2011. “Preparation of Reduced Representation Bisulfite Sequencing Libraries for Genome-Scale DNA Methylation Profiling.” *Nature Protocols* 6 (4): 468–81.
Gujar, Hemant, Daniel J. Weisenberger, and Gangning Liang. 2019. “The Roles of Human DNA Methyltransferases and Their Isoforms in Shaping the Epigenome.” *Genes* 10 (2). https://doi.org/10.3390/genes10020172.
Hallgrímsson, Benedikt, and Brian K. Hall. 2011. *Epigenetics: Linking Genotype and Phenotype in Development and Evolution*. Univ of California Press.
Hermann, Andrea, Rachna Goyal, and Albert Jeltsch. 2004. “The Dnmt1 DNA-(cytosine-C5)-Methyltransferase Methylates DNA Processively with High Preference for Hemimethylated Target Sites.” *The Journal of Biological Chemistry* 279 (46): 48350–59.
Hetzel, Sara, Pay Giesselmann, Knut Reinert, Alexander Meissner, and Helene Kretzmer. 2021. “RLM: Fast and Simplified Extraction of Read-Level Methylation Metrics from Bisulfite Sequencing Data.” *Bioinformatics* 37 (21): 3934–35.
Holliday, R., and J. E. Pugh. 1975. “DNA Modification Mechanisms and Gene Activity during Development.” *Science* 187 (4173): 226–32.
Huan, Qing, Yuliang Zhang, Shaohuan Wu, and Wenfeng Qian. 2018. “HeteroMeth: A Database of Cell-to-Cell Heterogeneity in DNA Methylation.” *Genomics, Proteomics & Bioinformatics* 16 (4): 234–43.
Jaenisch, Rudolf, and Adrian Bird. 2003. “Epigenetic Regulation of Gene Expression: How the Genome Integrates Intrinsic and Environmental Signals.” *Nature Genetics* 33 (3): 245–54.
Jeltsch, Albert, and Renata Z. Jurkowska. 2014. “New Concepts in DNA Methylation.” *Trends in Biochemical Sciences* 39 (7): 310–18.
Jin, Zelin, and Yun Liu. 2018. “DNA Methylation in Human Diseases.” *Genes & Diseases* 5 (1): 1–8.
Jones, Peter A. 2012. “Functions of DNA Methylation: Islands, Start Sites, Gene Bodies and beyond.” *Nature Reviews. Genetics* 13 (7): 484–92.
Jones, Peter A., and Gangning Liang. 2009. “Rethinking How DNA Methylation Patterns Are Maintained.” *Nature Reviews. Genetics*

10 (11): 805–11.
Karemaker, Ino D., and Michiel Vermeulen. 2018. “Single-Cell DNA Methylation Profiling: Technologies and Biological Applications.” *Trends in Biotechnology* 36 (9): 952–65.
Kerr, Lyndsay, Ioannis Kafetzopoulos, Ramon Grima, and Duncan Sproul. 2023. “Genome-Wide Single-Molecule Analysis of Long-Read DNA Methylation Reveals Heterogeneous Patterns at Heterochromatin That Reflect Nucleosome Organisation.” *PLoS Genetics* 19 (10): e1010958.
Krueger, Felix, and Simon R. Andrews. 2011. “Bismark: A Flexible Aligner and Methylation Caller for Bisulfite-Seq Applications.” *Bioinformatics* 27 (11): 1571–72.
Landan, Gilad, Netta Mendelson Cohen, Zohar Mukamel, Amir Bar, Alina Molchadsky, Ran Brosh, Shirley Horn-Saban, et al. 2012. “Epigenetic Polymorphism and the Stochastic Formation of Differentially Methylated Regions in Normal and Cancerous Tissues.” *Nature Genetics* 44 (11): 1207–14.
Landau, Dan A., Kendell Clement, Michael J. Ziller, Patrick Boyle, Jean Fan, Hongcang Gu, Kristen Stevenson, et al. 2014. “Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia.” *Cancer Cell* 26 (6): 813–25.
Lee, Dohoon, Bonil Koo, Jeewon Yang, and Sun Kim. 2023. “Metheor: Ultrafast DNA Methylation Heterogeneity Calculation from Bisulfite Read Alignments.” *PLoS Computational Biology* 19 (3): e1010946.
Lee, Jong-Hun, Sung-Joon Park, and Kenta Nakai. 2017. “Differential Landscape of Non-CpG Methylation in Embryonic Stem Cells and Neurons Caused by DNMT3s.” *Scientific Reports* 7 (1): 1–11.
Leung, D. C., K. B. Dong, I. A. Maksakova, P. Goyal, R. Appanah, S. Lee, M. Tachibana, et al. 2011. “Lysine Methyltransferase G9a Is Required for de Novo DNA Methylation and the Establishment, but Not the Maintenance, of Proviral Silencing.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (14). https://doi.org/10.1073/pnas.1014660108.
Li, Sheng, Francine Garrett-Bakelman, Alexander E. Perl, Selina M. Luger, Chao Zhang, Bik L. To, Ian D. Lewis, et al. 2014. “Dynamic Evolution of Clonal Epialleles Revealed by Methclone.” *Genome Biology* 15 (9): 472.
Liu, Yang, Wojciech Rosikiewicz, Ziwei Pan, Nathaniel Jillette, Ping Wang, Aziz Taghbalout, Jonathan Foox, et al. 2021. “DNA Methylation-Calling Tools for Oxford Nanopore Sequencing: A Survey and Human Epigenome-Wide Evaluation.” *Genome Biology* 22 (1): 1–33.
Loda, Agnese, Samuel Collombet, and Edith Heard. 2022. “Gene Regulation in Time and Space during X-Chromosome Inactivation.” *Nature Reviews. Molecular Cell Biology* 23 (4): 231–49.

Masser, Dustin R., Arthur S. Berg, and Willard M. Freeman. 2013. "Focused, High Accuracy 5-Methylcytosine Quantitation with Base Resolution by Benchtop next-Generation Sequencing." *Epigenetics & Chromatin* 6 (1): 33.
Masser, Dustin R., David R. Stanford, and Willard M. Freeman. 2015. "Targeted DNA Methylation Analysis by next-Generation Sequencing." *Journal of Visualized Experiments: JoVE*, no. 96 (February). https://doi.org/10.3791/52488.
Meissner, Alexander, Andreas Gnirke, George W. Bell, Bernard Ramsahoye, Eric S. Lander, and Rudolf Jaenisch. 2005. "Reduced Representation Bisulfite Sequencing for Comparative High-Resolution DNA Methylation Analysis." *Nucleic Acids Research* 33 (18): 5868–77.
Mendoza, Alex de, Daniel Poppe, Sam Buckberry, Jahnvi Pflueger, Caroline B. Albertin, Tasman Daish, Stephanie Bertrand, et al. 2021. "The Emergence of the Brain Non-CpG Methylation System in Vertebrates." *Nature Ecology & Evolution* 5 (3): 369–78.
Millán-Zambrano, Gonzalo, Adam Burton, Andrew J. Bannister, and Robert Schneider. 2022. "Histone Post-Translational Modifications — Cause and Consequence of Genome Function." *Nature Reviews. Genetics* 23 (9): 563–80.
Mishra, Manoj K., and Kumar S. Bishnupuri. 2016. *Epigenetic Advancements in Cancer*. Springer.
Moore, Lisa D., Thuc Le, and Guoping Fan. 2013. "DNA Methylation and Its Basic Function." *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology* 38 (1): 23–38.
Müller, Fabian, Michael Scherer, Yassen Assenov, Pavlo Lutsik, Jörn Walter, Thomas Lengauer, and Christoph Bock. 2019. "RnBeads 2.0: Comprehensive Analysis of DNA Methylation Data." *Genome Biology* 20 (1): 55.
Nishiyama, Atsuya, Christopher B. Mulholland, Sebastian Bultmann, Satomi Kori, Akinori Endo, Yasushi Saeki, Weihua Qin, et al. 2020. "Two Distinct Modes of DNMT1 Recruitment Ensure Stable Maintenance DNA Methylation." *Nature Communications* 11 (1): 1–17.
Noble, Denis. 2015. "Conrad Waddington and the Origin of Epigenetics." *The Journal of Experimental Biology* 218 (Pt 6): 816–18.
Okano, M., D. W. Bell, D. A. Haber, and E. Li. 1999. "DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for de Novo Methylation and Mammalian Development." *Cell* 99 (3): 247–57.
Parry, Aled, Steffen Rulands, and Wolf Reik. 2020. "Active Turnover of DNA Methylation during Cell Fate Decisions." *Nature Reviews. Genetics* 22 (1): 59–66.
Reik, W., and J. Walter. 2001. "Genomic Imprinting: Parental Influence on the Genome." *Nature Reviews. Genetics* 2 (1): 21–32.

Reinius, Lovisa E., Nathalie Acevedo, Maaike Joerink, Göran Pershagen, Sven-Erik Dahlén, Dario Greco, Cilla Söderhäll, Annika Scheynius, and Juha Kere. 2012. “Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility.” *PloS One* 7 (7): e41361.

Sarnataro, Antonella, Giulia De Riso, Sergio Cocozza, Antonio Pezone, Barbara Majello, Stefano Amente, and Giovanni Scala. 2022. “A Novel Workflow for the Qualitative Analysis of DNA Methylation Data.” *Computational and Structural Biotechnology Journal* 20 (October): 5925–34.

Scala, Giovanni, Ornella Affinito, Domenico Palumbo, Ermanno Florio, Antonella Monticelli, Gennaro Miele, Lorenzo Chiariotti, and Sergio Cocozza. 2016. “ampliMethProfiler: A Pipeline for the Analysis of CpG Methylation Profiles of Targeted Deep Bisulfite Sequenced Amplicons.” *BMC Bioinformatics* 17 (1): 484.

Scherer, Michael, Almut Nebel, Andre Franke, Jörn Walter, Thomas Lengauer, Christoph Bock, Fabian Müller, and Markus List. 2020. “Quantitative Comparison of within-Sample Heterogeneity Scores for DNA Methylation Data.” *Nucleic Acids Research* 48 (8): e46.

Simpson, Jared T., Rachael E. Workman, P. C. Zuzarte, Matei David, L. J. Dursi, and Winston Timp. 2017. “Detecting DNA Cytosine Methylation Using Nanopore Sequencing.” *Nature Methods* 14 (4): 407–10.

Singer, Zakary S., John Yong, Julia Tischler, Jamie A. Hackett, Alphan Altinok, M. Azim Surani, Long Cai, and Michael B. Elowitz. 2014. “Dynamic Heterogeneity and DNA Methylation in Embryonic Stem Cells.” *Molecular Cell* 55 (2): 319–31.

Smith, Zachary D., and Alexander Meissner. 2013. “DNA Methylation: Roles in Mammalian Development.” *Nature Reviews. Genetics* 14 (3): 204–20.

Sun, Yan V., Stephen T. Turner, Jennifer A. Smith, Pamela I. Hammond, Alicia Lazarus, Jodie L. Van De Rostyne, Julie M. Cunningham, and Sharon L. R. Kardia. 2010. “Comparison of the DNA Methylation Profiles of Human Peripheral Blood Cells and Transformed B-Lymphocytes.” *Human Genetics* 127 (6): 651–58.

Teschendorff, Andrew E., Tianyu Zhu, Charles E. Breeze, and Stephan Beck. 2020. “EPISCORE: Cell Type Deconvolution of Bulk Tissue DNA Methylomes from Single-Cell RNA-Seq Data.” *Genome Biology* 21 (1): 221.

Thiagalingam, Sam. 2015. *Systems Biology of Cancer*. Cambridge University Press.

Tillotson, Rebekah, Justyna Cholewa-Waclaw, Kashyap Chhatbar, John C. Connelly, Sophie A. Kirschner, Shaun Webb, Martha V. Koerner, et al. 2021. “Neuronal Non-CG Methylation Is an Essential Target for MeCP2 Function.” *Molecular Cell* 81 (6): 1260–75.e12.

Waddington, C. H. 2012. “The Epigenotype. 1942.” *International*

Journal of Epidemiology 41 (1): 10–13.
"Writers, Readers, and Erasers of Epigenetic Marks." 2015. In *Epigenetic Cancer Therapy*, 31–66. Academic Press.
Wutz, Anton, and Karmele Valencia. 2015. "Recent Insights into the Regulation of X-Chromosome Inactivation." *Advances in Genomics and Genetics*, May, 227.
Yong, Wai-Shin, Fei-Man Hsu, and Pao-Yang Chen. 2016. "Profiling Genome-Wide DNA Methylation." *Epigenetics & Chromatin* 9 (June): 26.