SAARLAND UNIVERSITY

CENTER FOR BIOINFORMATICS

GENETICS/EPIGENETICS

MASTER'S PROGRAM IN BIOINFORMATICS

MASTER'S THESIS

---

methylTFR - a computational method to identify DNA methylation signatures in transcription factor binding sites

---

submitted by

Sarath Kumar Murugan

January, 2023

Supervisor

Prof. Dr. Fabian Müller

Reviewers

Prof. Dr. Fabian Müller

Prof. Dr. Jörn Walter

Sarath Kumar Murugan

methylTFR - a computational method to identify DNA methylation signatures in transcription factor binding sites

Master's Thesis in Bioinformatics,

Saarland University
Saarbrücken, Germany
January, 2023

Declaration

I, Sarath Kumar Murugan, hereby confirm that this thesis titled, "methylTFR - a computational method to identify DNA methylation signatures in transcription factor binding sites" is my own work and that I have documented all sources used.

Saarbrücken, January, 2023

Sarath Kumar Murugan

Acknowledgements

I would like to express my deepest gratitude to my thesis advisor, Prof. Dr Fabian Müller, for his unwavering support and guidance throughout the course of my master's programme. His knowledge and expertise in the field of epigenetics were invaluable in helping me to shape my research and develop my ideas. I am deeply grateful for the time he has taken to read and provide feedback on my work and for the encouragement he has given me to pursue my academic goals.

I would also like to thank Prof. Dr Jörn Walter for agreeing to be a second reviewer for my master's thesis.

I am truly grateful for all the unwavering support, guidance, and mentorship Venkatesh Chellappa has provided me, I couldn't have made it this far without his unconditional belief in me. Many thanks to Johan Lindberg and Rebecka Bergström for supporting me in following my ambition. I want to convey my special thanks to Karthick Maniram for many insightful discussions, personal advice and for feeding me with some really delicious food while I worked on my thesis.

I am also grateful to all my labmates, Midhuna Maran, Irem B. Gündüz, Nihit Agarwal, and Ahmed Osman, for their constant help and productive discussions, without which I could not have completed my thesis in a timely and graceful manner.

I want to thank my friends and family for their support throughout this journey. Their love and encouragement kept me motivated and helped me to persevere through the challenges I faced. I would specially thank my mother, Uma, and my partner, Yaazhini, for their unflinching belief in me and their constant support and understanding. I would also like to thank Anish, Geo James, Bhuvaneswar, Nitish, and Nobel for creating a harmonious and personable environment in Germany.

Finally, I would like to thank all the staff and faculty members at the University of Saarland for their support and assistance during my time in the program. I am proud to have been a part of this institution and am grateful for the opportunities it has given me to grow as a student and a researcher.

Thank you all for your support and guidance.

Abstract

DNA methylation is a well-known epigenetic mark and is a process of adding methyl group (-CH3) to the cytosine base in the context of a CpG dinucleotide. This process can significantly impact the regulation of gene expression by affecting the binding of transcription factors to their binding sites (TFBS) in the genome. In most cases, the methylation of CpG sites within TFBS can reduce transcription factors' ability to bind and suppress gene expression. Conversely, demethylation of CpG sites within TFBS can activate gene expression by increasing the TFs binding affinity. Therefore, understanding the relationship between DNA methylation and TFBS is crucial for comprehending the regulation of gene expression and its impact on cell development and disease.

This research project provides a novel computational approach to analyse the methylation around TF motif sequences in order to provide meaningful insights. The method, called methylTFR, is based on TF footprinting which is used to identify specific binding sites for transcription factors (TFs) in DNA using chromatin accessibility. It is adapted for use with bulk bisulfite sequencing data. We observed that there is a drop in methylation level around the motif center in TF footprinting by DNA methylation, which indicates possible transcription factor binding. We captured that signal by calculating a deviation score, which represents the difference between methylation signals at the motif center and in the background. Higher deviations reveal a higher likelihood of TF binding to the motif sequence. The approach was validated using human blood cell samples from blueprint epigenome data and was effective in identifying outliers and clustering samples. Furthermore, by using the distal enhancer regions from the Ensembl regulatory build, we could observe the distinct divergence between different cell types and understand the transcription factors associated with methylation patterns. We wanted to show that motif-based analysis can provide information beyond typical region-based analyses by a cluster comparison study. The result shows that deviation score-based clustering is faster and more efficient than tiling-based analysis. Additionally, deviation scores can be used for downstream analyses such as differential testing and dimensionality reduction. (MethylTFR available at GitHub: https://github.com/EpigenomeInformatics/methylTFR)

Contents

List of Figures

List of Abbreviations

ADP	Adinosine Di Phosphate
BED	Browser Extensible Data
BI	Binomial distribution
BSC	Bisulfite Crick
BSCR	Bisulfite Crick Reverse complement
BSW	Bisulfite Watson
BSWR	Bisulfite Watson Reverse complement
DEEP	Das Dutsche Epigenom Programm
DNA	Deoxyribonucleic Acid
EM	Expectation Maximization
FDR	False Discovery Rate
FMR(s)	Fully Methylated Region(s)
GLM	Generalized Linear Model
HMM	Hidden Markov Model
HTML	Hyper Text Markup Language
IGV	Integrative Genomics Viewer
LMR(s)	Low Methylated Region(s)
NBI	Negative Binomial distribution
MeDIP	Methylated DNA Immunoprecipitation
NGS	Next Generation Sequencing
PCR	Polymerase Chain Reaction
PMD(s)	Partially Methylated Domain(s)
RRBS	Reduced Representation Bisulfite Sequencing
RNA	Ribonucleic acid
RNAPII	RNA polymerase II
SNP	Single Nucleotide Polymorphism
TF	Transcription Factor
TFBS	Transcription Factor Binding Sites
UMR(s)	Unmethylated Region(s)
WGBS	Whole Genome Bisulfite Sequencing
ZINBI	Zero Inflated Negative Binomial distribution

Chapter 1

Introduction

1.1 Biological Background

The central dogma of life is a basic rule that governs the passage of genetic information inside living organisms. This central dogma was originally put forth in 1953 [1]. According to this rule, genetic information flows from DNA to RNA to proteins. Initially, it was believed that this flow was unidirectional. However, it has been established that this flow happens in both directions as well as at different levels. The information flow could happen from RNA to DNA, and proteins have also been reported to affect DNA structure and function.

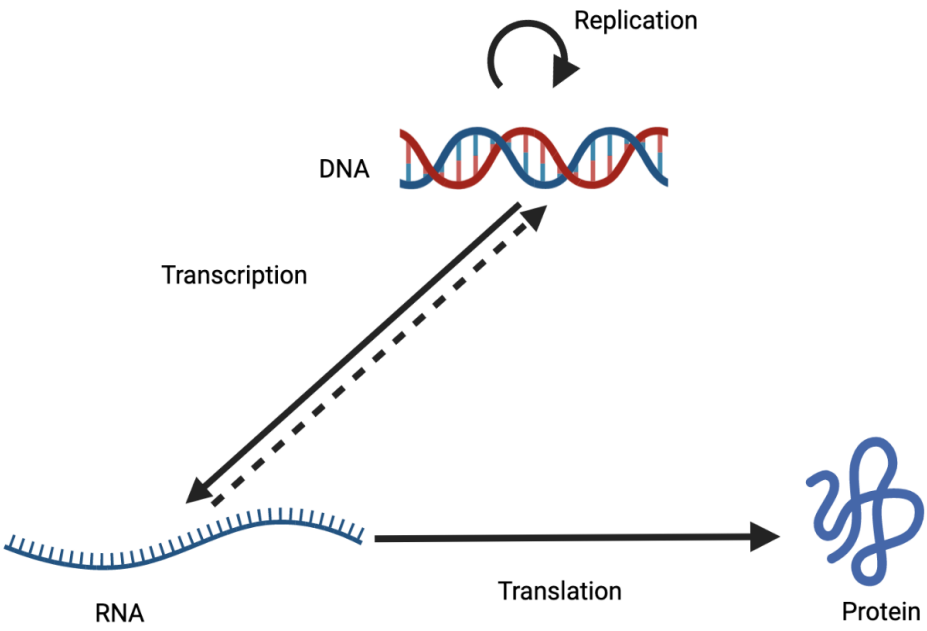

FIGURE 1.1: Central Dogma of Life

It is understood that DNA is the primary source of genetic information, and this information it's later converted to RNA in a process called transcription. RNA contains the code with which proteins are synthesised. This process by which proteins are produced inside a cell with the help of code from RNA is called translation [2]. Proteins are the functional units of cells, and they perform a variety

of functions, including mechanical support, catalysis of chemical processes and also the regulation of gene expression [3].

This flow of information has been accepted widely as the basis of molecular biology. This principle has also been helpful in understanding the molecular basis of how genetic information is inherited and also genetic functions are regulated [1]. Therefore it is important to understand the mechanism by which genetic information flows through these processes. With the help of this knowledge, we could further study complex processes that regulate cellular biology and also develop modern techniques for studying and manipulating them, such as sequencing, gene-editing, cloning, and imaging.

1.1.1 Genome and Epigenome

The genome is defined as the repository of all genetic information contained within a cell. This comprises all the DNA, including the genes and other genetic elements. Every individual has a different genome, and the uniqueness is governed by the number and the nature of the variations present in the DNA sequences [4]. These variations, along with the actual composition of the genome, determine the individual's phenotype, physiology and also their susceptibility and tolerance to various other factors.

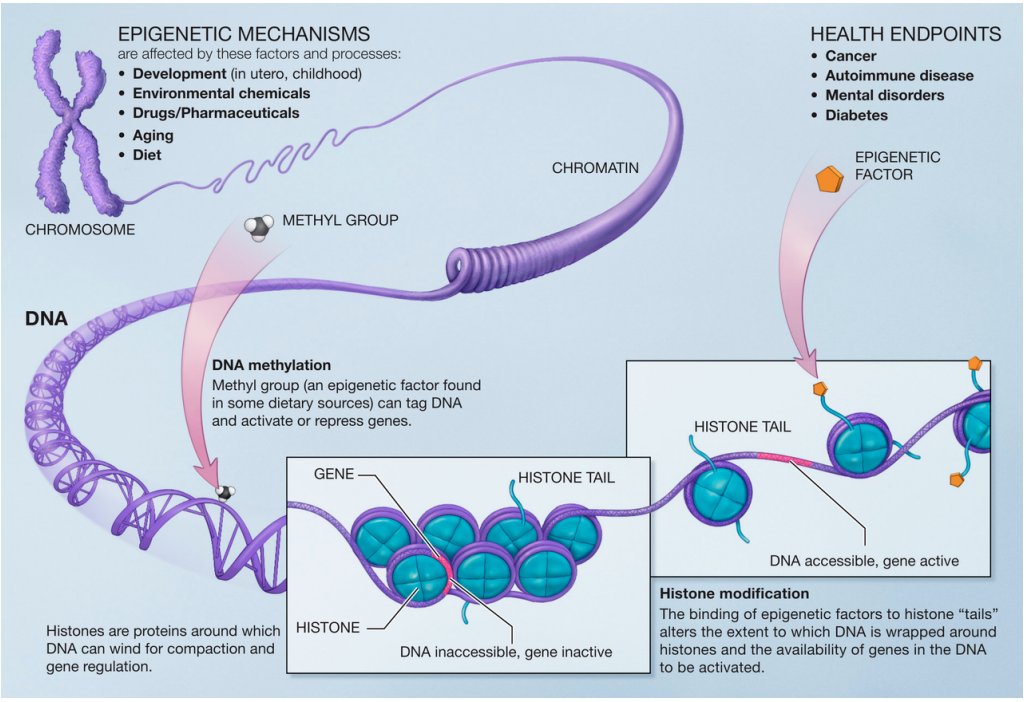

FIGURE 1.2: An illustration of how epigenetic mechanisms might impact health. (Image source: NIH [5])

The variations or mutations in the genome can either be inherited from the parents or acquired during the lifetime of the individual. Such inherited variations with which an individual is born are called germline mutations, and the acquired mutations are called somatic mutations. Germline mutations are globally present in all cells of the individual, whereas somatic mutations will be seen only in those cells which have acquired them. Mutations can either be harmful or benign. Harmful germline mutations are related to germline disorders, hereditary cancers, hereditary syndromes and other rare diseases, whereas somatic mutations are predominantly associated with cancers and are also associated with other health conditions such as Rett syndrome and epilepsy. These harmful somatic mutations can either cause cancer or be a product of genetic alterations caused by cancer itself [6].

Similar to genomic alterations, there are chemical modifications to DNA and histone proteins that happen inside a cell, and these regulate the expression of genes within the genome (Figure 1.2). These modifications are known as epigenetic marks that control gene expression without changing the actual DNA sequence [7]. The epigenome of a cell is the complete description or collection of all these epigenetic marks. Epigenetic marks are reversible and dynamic, meaning that they can be changed in response to environmental and other factors.

The epigenetic marks are unique for each cell type. For example, a skin cell will differ in the number and type of epigenetic marks when compared with a brain cell. It is also important to note that the same cell would have different epigenetic marks in different biological states; for example, a healthy cell will have different epigenetic marks compared to the same cell with a disease. Therefore, understanding the genome and epigenome is essential for many areas of biology and medicine[7].

1.1.2 DNA Methylation

Epigenetics is the scientific discipline which focuses on changes that affect gene expression or cellular phenotype in a heritable manner but do not alter the DNA sequence itself. These changes can be caused by various molecular mechanisms, including DNA methylation, histone modification, and influence by non-coding RNA molecules [7]. DNA methylation is one of the epigenetic marks where the cytosine base at the fifth carbon (C5) position is altered by addition of -CH3 group, which converts cytosine to 5-methylcytosine (5mC). DNA methylation is a common epigenetic process which regulates gene expression [8].

The methyl group is transferred to cytosine by a group of enzymes, the DNA methyltransferases. DNA methylation can take place in various contexts, such as CG and C-N-G (where "N" stands for A, C, or T) and C-N-N. In the case of mammalian genomes, the majority of methylation happens on CG sequences. Two enzymes, DNMT3 and DNMT1, are responsible for this process; DNMT3 is responsible for adding new methylation to cytosines, while DNMT1 ensures that methylation is preserved during cell division [8].

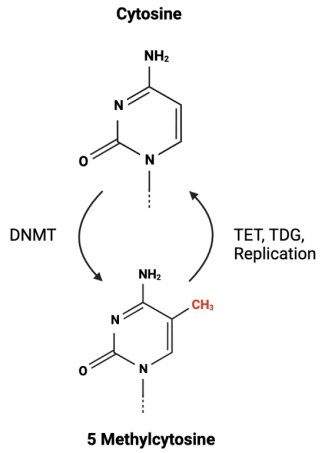

FIGURE 1.3: DNA Methylation

Gene expression is controlled by specific regions in the genome. The most significant ones are promoter regions and enhancer regions. Promoters are located near the start site of a gene and contain specific sequences that bind to the transcriptional machinery, allowing for the initiation of transcription [9]. Enhancers are non-coding regions of DNA that can be located far from the gene they regulate and can be activated by transcription factors to increase the rate of transcription of a specific gene. They can be located 5'-end, 3'-end or in the intronic regions of the gene [9].

Promoters are responsible for initiating transcription, while enhancers enhance or increase the rate of transcription of a specific gene. Both are essential to control the expression of a gene in a cell or tissue-specific manner. Methylation at a promoter region may silence gene expression by blocking transcription initiation or altering the chromatin structure like condensation, thereby resulting in gene silencing [10].

Transcription factor (TF) binding sites within enhancers and promoters are critical for the regulation of gene expression. The binding of TFs to specific sequences within enhancers allows for the recruitment of other proteins, such as co-activators and chromatin remodelling complexes, that ultimately lead to the activation of the gene. Methylation of TF binding sites will directly interfere with this process of TF binding to the enhancer region and thereby repressing gene expression [11]. Additionally, certain transcription factors have been known to bind to methylated DNA and affect the transcription process.

The level of methylation at certain regions of DNA is very much relative and dependent upon various factors like the location, biological state, origin, environmental factors, and disease conditions. Hence, to call that region of DNA as highly methylated (hypermethylated) or less methylated (hypomethylated) compared to the usual methylated state would be highly context-driven.

These processes have been associated with various diseases; for example, hypermethylation of tumour-suppressor genes has been observed in different cancers, including leukaemia, colon cancer, breast cancer, lung cancer, and retinoblastoma [12]. With this information, it is possible to explore possible options for therapy for example, in a type of blood cancer (AML), it has been observed that Azacytidine, a compound that inhibits DNA-methyl-transferase is effective in clinical trials [13].

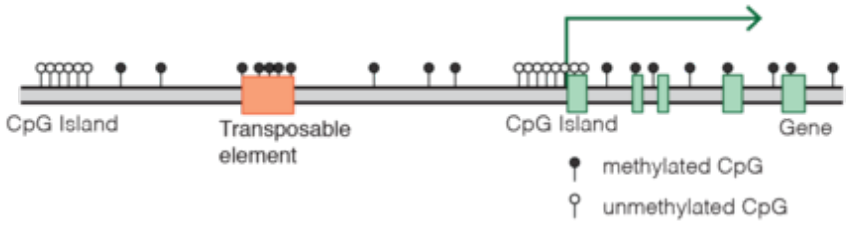

FIGURE 1.4: DNA methylation landscape in mammals genome (Image Source [14])

DNA methylation is reversible through the action of enzymes called DNA demethylases. These enzymes make chemical modifications, due to which the cytosine nucleotide is excised and replaced, thereby reversing the methylation. There are several different DNA demethylases, including Tet1, Tet2, and Tet3, which can demethylate 5-methylcytosine (5mC) in different contexts [8]. The reversibility of DNA methylation allows cells to dynamically regulate gene expression in response to different stimuli and during different stages of development. For example, DNA methylation can be reversible during the process of cell differentiation, in which cells change from a more general to a more specialized state. DNA methylation can also be reversible in response to environmental factors, such as changes in diet or exposure to toxins [8]. DNA methylation can be studied with the help of different next-generation sequencing platforms and technologies.

Next Generation Sequencing (NGS)

The first human genome was sequenced through the Human Genome Project by sanger enzymatic sequencing, in which the sequence of the individual DNA bases is determined one by one using a series of enzymatic processes [15]. Sanger sequencing is applicable only for smaller sequences or for genomic regions of specific interest. Applying this technique to sequence large regions of the genome is time-consuming and very expensive. So massively parallel DNA sequencing or next-generation sequencing (NGS) technology was introduced in 2005 [15], where new NGS equipment combined the enzymatic steps and data

generation into a single operation. This enabled the sequencing of thousands of bases simultaneously using various templates. There has been a lot of advancement in the process of sample preparation for the purpose of sequencing and also advancement in sequencing technologies over the past 20 years. These technological advancements have made it possible for us to sequence not only the genome but also other biomolecules and biological processes within and outside the cell.

S. no.	**Sequencing technology**	**Compartment sequenced**	**Application**
1	Whole Genome (WGS)	Entire genomic DNA	Identification of SNPs, mutations, Structural Variants, CNVs
2	Whole Exome (WES)	Exons and flanking regions	Identification of mutations, Structural Variants
3	RNA sequencing (RNAseq)	Whole Transcriptome	Transcript abundance, Gene Expression, Differential expression analysis
4	Chromatin Immunopreci -pitation (ChIPseq)	Genome-wide	Identify DNA binding sites for a transcription factor or other proteins
5	Whole Genome Bisulfite Sequencing (WGBS)	Genome-wide	Identification for DNA methylation
6	Assay for Transposase Accessible Chromatin with sequencing (ATACseq)	Genome-wide	Determination of Chromatin Accessibility by peak-calling
7	Reduced representation bisulfite sequencing (RRBS)	Reduced Genome	Selected CpG sites methylation level

The epigenome could be profiled by sequencing, and newer methods are being introduced every year. These methods focus on extracting information from various processes and biological compartments to capture different epigenetic marks (Figure 1.5)[17]. In some conditions, DNA methylation is a dominant epigenetic mark and also a relatively constant one that doesn't undergo significant

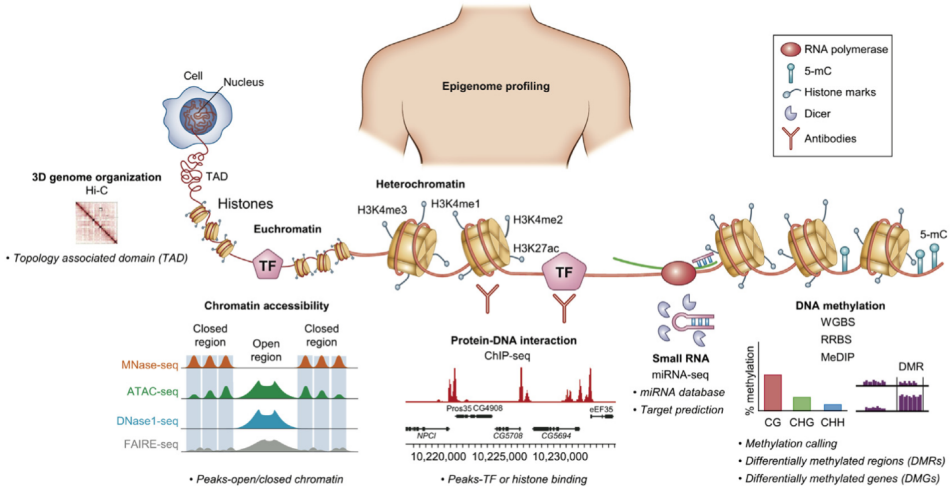

FIGURE 1.5: The next-generation sequencing technologies used for profiling various epigenetic components are shown visually. (Image source: [16])

changes [8]. Therefore, it is considered a good candidate as an epigenetic mark to profile the epigenome.

Several high-throughput sequencing methods can be used to analyse DNA methylation globally throughout the genome. These methods include whole genome bisulfite sequencing, which involves the bisulfite conversion of the entire genome and the subsequent sequencing of the resulting DNA, and sequencing of methylated DNA immunoprecipitation (MeDIP-seq), which involves the enrichment of methylated DNA using an antibody specific for 5mC and the subsequent sequencing of the enriched DNA [18].

Reduced representation bisulfite sequencing (RRBS) is another method for analyzing DNA methylation that is based on bisulfite conversion. RRBS allows for the analysis of methylation at a large number of CpG sites in a cost-effective manner [18]. This method involves the selective enrichment of CpG-rich regions of the genome, followed by bisulfite conversion and sequencing.

1.1.3 Whole Genome Bisulfite Sequencing

A high-throughput technique called whole genome bisulfite sequencing (WGBS) is used to examine DNA methylation across the entire genome. It involves the conversion of unmethylated cytosine bases to uracil using sodium bisulfite, followed by the sequencing of the resulting DNA. The resulting data can be used to identify methylated CpG sites, as well as to quantify the amount of methylation at each site.

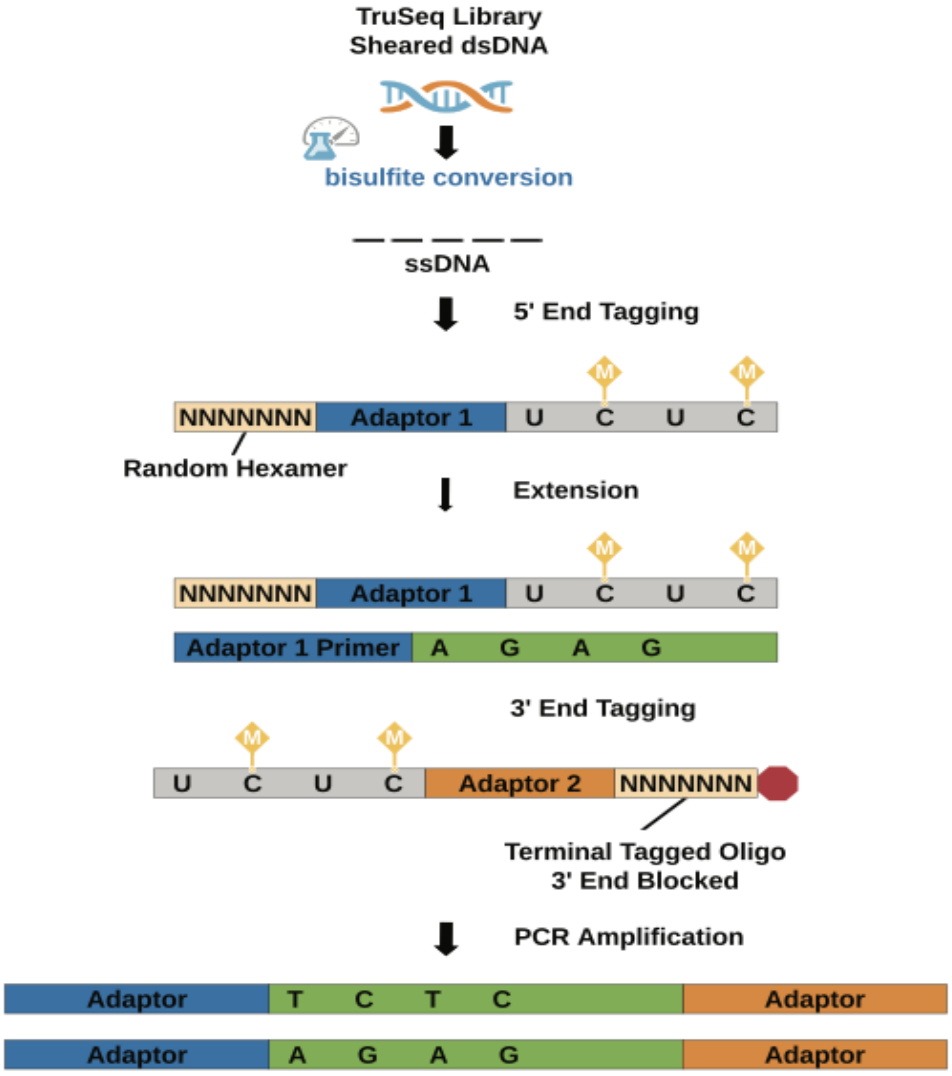

FIGURE 1.6: Illumina TruSeq library preparation for WGBS (Image source: [19])

The WGBS procedure typically begins with the preparation of high-quality DNA from the sample of interest. This DNA is then fragmented into smaller pieces, typically using a restriction enzyme or mechanical shearing. After the DNA is broken into pieces, it is treated with bisulfite, which causes the nonmethylated cytosines to be converted into uracil; however, the methylated cytosines (5mC) remain unaltered. After bisulfite conversion, multiple copies of the DNA are created using PCR, and the resulting sequences are sequenced using high-throughput sequencing techniques, such as Illumina sequencing [19]. The presence or absence of cytosine at a specific position can indicate whether it is methylated or not. The resulting sequence data is then processed and analyzed using specialized software, such as Bismark [20] or MethylC-seq [21], to identify

and quantify methylated CpG sites. Bismark aligns bisulfite-converted DNA to a reference genome and generates methylation calls in a variety of formats (e.g. CpG, CHG, CHH, where H could be C, A, or T). This information is further used to calculate the total number of CpG, CHG, and CHH sites, the number of methylated and unmethylated reads, the percentage of methylation, etc. These metrics are then used to quantify region-wise methylation. Bismark is useful to quantify methylation genome-wide or in areas of interest such as gene bodies, promoters, or enhancer regions.

WGBS has a number of advantages over other methods for analyzing DNA methylation. It allows for the analysis of methylation at every CpG site in the genome, providing a comprehensive view of DNA methylation patterns. It is also highly sensitive, with the ability to detect methylation at low levels. However, WGBS is also a technically demanding and resource-intensive procedure, and it can be costly to perform.

1.1.4 Transcription and Transcription Factor

Transcription can be defined as the process through which a DNA sequence is converted into RNA with the help of the enzyme RNA-polymerase, which reads the DNA template and synthesizes an RNA molecule complementary to one of the strands of the DNA. The RNA molecule produced during transcription is known as a primary transcript, and it may be further processed to produce the messenger RNA (mRNA) molecule[22]. The final mature mRNA molecule codes for protein synthesis. Transcription is regulated at multiple levels, including at the level of transcription initiation, elongation during which the RNA sequence is synthesized from the complementary template single-strand DNA, and termination [22].

Transcription factors are molecules that interact with certain sequences in DNA and manage transcription by controlling the initiation and continuation of the process. Transcription factors can bind to promoter, enhancer or repressor regions of genes, which are sequences located upstream of the gene that controls its expression [9]. Gene enhancers, repressors and promoters form a complex regulatory network along with transcription factors.

Gene enhancers are non-coding DNA sequences that are relatively short (50–1.5k bases) in length and upregulate the gene expression process. They are usually seen at distant locations away from the gene that they control, and they can be found on either the same chromosome or a different chromosome [24]. Enhancers work

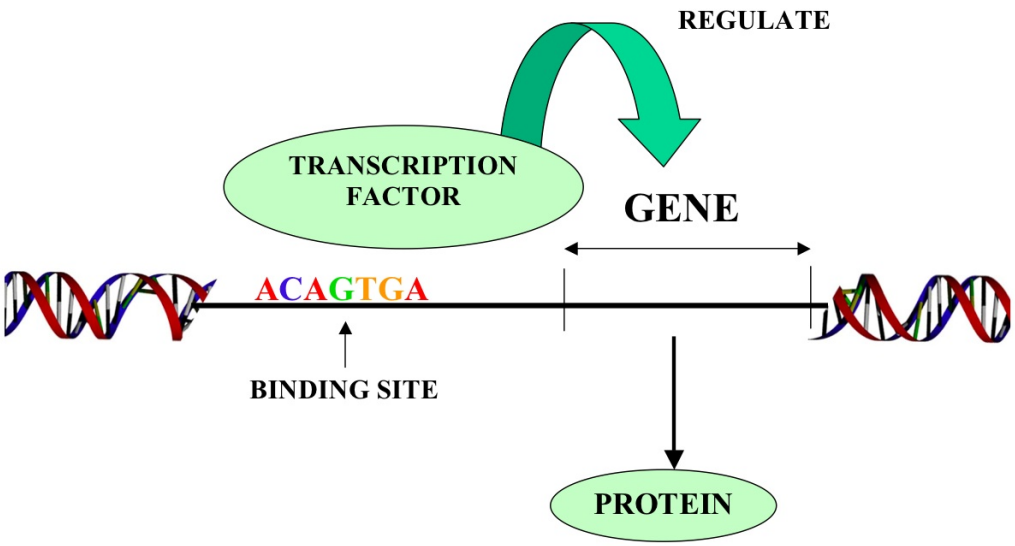

FIGURE 1.7: A transcription factor regulates a gene's ability to produce protein by attaching to its binding site.(Image source: [23])

by binding to transcription factors, which then recruit other proteins necessary for the initiation of transcription.

Enhancers often contain multiple binding sites for transcription factors, which allows for a complex network of interactions that fine-tune the regulation of gene expression. They can also be activated or repressed by specific signals, such as hormones or environmental cues, which can further modulate the activity of the gene[24].

Similar to enhancers, there are other genomic elements called repressors that work by binding to specific transcription factors. However, in contrast to enhancers, this binding will then inhibit the initiation of transcription. So, gene repressors are non-coding sequences of DNA that regulate the gene expression by decreasing their activity. Repressors can be located within the gene they regulate or in faraway regions of the genome, and like enhancers, they can be activated or repressed by specific signals such as hormones or environmental cues [25]. In contrast to enhancers, repressors can bind to different domains of the transcriptional machinery, like the promoter or the enhancer, and inhibit transcription by different mechanisms like direct binding to the RNA polymerase or by recruiting corepressors that change chromatin structure [25].

Enhancers and repressors play a crucial role in the development and in control of cell-specific gene expression, allowing for the regulation of tissue-specific gene

expression. Enhancer and repressor mutations or structural variations have been linked to genetic disorders and cancer [9, 11] .

There are many different types of transcription factors, and they are significant in several biological processes, including development, metabolism, and response to environmental stimuli. Dysregulation of transcription factors can lead to various diseases, including cancer[9].

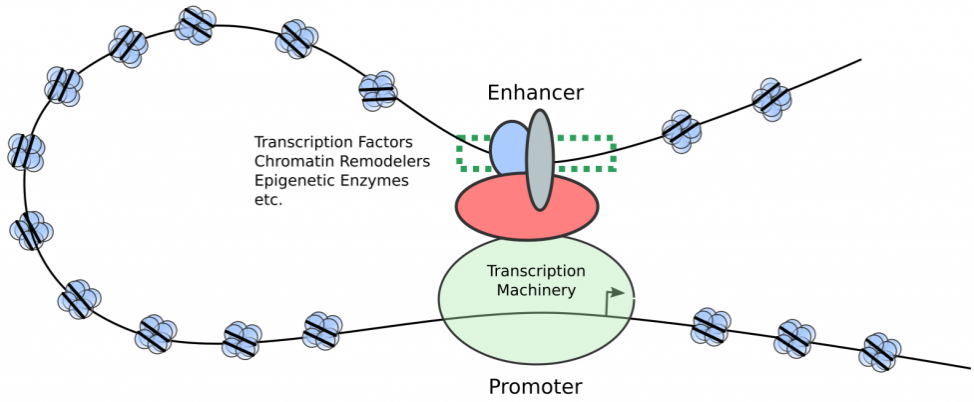

FIGURE 1.8: Enhancers are distal genetic elements that are bound by transcriptional activators to enhance gene transcription. (Image source: [26])

Transcription Factor Binding Sites (TFBSs)

Transcription factor binding sites (TFBSs) are specific DNA sequences which are typically located in the promoter regions of genes. Transcription factors bind with these TFBSs in order to regulate gene expression by controlling the initiation and elongation of transcription. TFBSs are specific DNA motifs, which are short, conserved DNA sequences that are recognized specifically by the transcription factor [9]. Some transcription factors bind to more complex DNA sequences that may contain multiple motifs or may be more flexible in their recognition of DNA sequences.

Identifying TFBSs can provide insight into the regulatory networks that control gene expression and can also help to identify potential targets for therapeutic intervention in diseases such as cancer. There are several methods that can be used to identify TFBSs, including experimental techniques such as chromatin immunoprecipitation (ChIP) and computational methods such as in-silico prediction[9].

ChIP-seq is a molecular biology technique that involves the immunoprecipitation of DNA-protein complexes, followed by the identification of the bound DNA

sequences by high-throughput sequencing[27]. ChIP-seq can be used to identify the DNA sequences that are bound by specific transcription factors, as well as to quantify the relative enrichment of specific DNA sequences.

In-silico prediction methods use computational algorithms to search for DNA sequences that are likely to be bound by specific transcription factors. These methods are based on the understanding that transcription factors bind to specific DNA sequences, and they use this information to identify potential TFBSs in a genome [28]. There are many different in-silico prediction methods available, and they can vary in sensitivity and specificity in the identification of various transcription factors.

1.1.5 Relationship between TFBS and DNA methylation

DNA methylation has been observed to affect the binding of transcription factors to TFBSs and, in turn, gene expression. Methylation at CpG sites within promoter regions or within TFBSs can inhibit the binding of transcription factors and repress gene expression. On the other hand, methylation at CpG sites within gene bodies or within enhancer regions can activate gene expression[29].

The effects of methylation on transcription factor binding and gene expression can be complex and context-dependent. In addition, the effect of methylation on transcription factor binding can be influenced by the specific transcription factor and the DNA sequence to which it binds [29].

Methylation at transcription factor binding sites in enhancers can have a variety of effects, including increasing or decreasing the affinity of the transcription factor for the DNA, influencing the strength of the enhancer, and altering the specificity of the enhancer for a particular gene. This can also affect the recruitment of co-factors which can further increase or decrease the strength of the enhancer [30]. In general, the presence of methylation at transcription factor binding sites has been found to reduce the activity of enhancers. However, this is not always the case and depends on the particular transcription factor and sequence context of the enhancer.

As stated above, there are several methods that can be used to study the effects of methylation on transcription factor binding, including bisulfite conversion, reduced representation bisulfite sequencing (RRBS), and high-throughput sequencing methods like WGBS and MeDIP-seq [31]. These methods allow researchers to identify and quantify methylated CpG sites and to study their effects on transcription factor binding and gene expression.

1.2 Motivation

The significance of TF binding sites in the regulation of the transcription process and gene expression has been studied in depth. Dysregulation of TFBSs can lead to abnormal gene expression and contribute to various diseases, such as cancer. Understanding the role of TF binding sites in transcription can help to identify potential targets for therapeutic intervention and improve our understanding of the underlying mechanisms of gene regulation. TF binding motif sequences are affected by random mutations and DNA methylation.

In this study, we aim to understand the association between known transcription factor binding sites and the methylation signals around them. One example of how methylation changes at TF binding sites affect the gene function was described by ([32]), who characterised the methylation in the promoter region of the tumour suppressor gene p16INK4a. According to them, in normal cells, the p16INK4a gene is silenced by addition of methyl group to its promoter, which prevents the binding of TFs and thereby blocks the initiation of transcription. In cancer cells, however, the promoter region is often demethylated, allowing for the binding of TFs and the activation of p16INK4a expression. This leads to cell cycle arrest and can inhibit the proliferation of cancer cells.

TF footprinting is a common approach to exploring the TF binding sites across the genome, which is used for predicting precise TF binding locations in the genome. TF footprinting is possible when we have data regarding chromatin accessibility. There are several tools like chromVar [33] and ArchR [34] that are used to analyse chromatin accessibility. chromVAR is an R package that can be used to identify genome-wide changes in chromatin accessibility peaks overlapping with motifs from ATAC-seq and DNase-seq data. ChromVar uses a method to gather and analyse the signal of chromatin accessibility across the entire genome at motifs or regulatory regions that contain a specific pattern. It then calculates the deviation score and applies the technical bias correction for each motif in each sample. This score reflects the likely level of activity of the DNA-binding protein linked to the pattern being examined. ChromVar utilises larger regions of accessibility changes, such as peaks, rather than individual nucleotides. ArchR, on the other hand, is very useful when finding chromatin accessibility in large-scale single-cell ATAC-seq data (>80000 cells) [34].

We considered applying the accessibility-based footprinting method suggested in this article [35] to genome-wide methylation data. The advantage of using TF footprinting on methylation data is that we could achieve single nucleotide

base-level resolution compared to ATAC-seq. Measuring CpG methylation status on a whole genome scale is easily achievable using the whole-genome bisulfite sequencing (WGBS) approach.

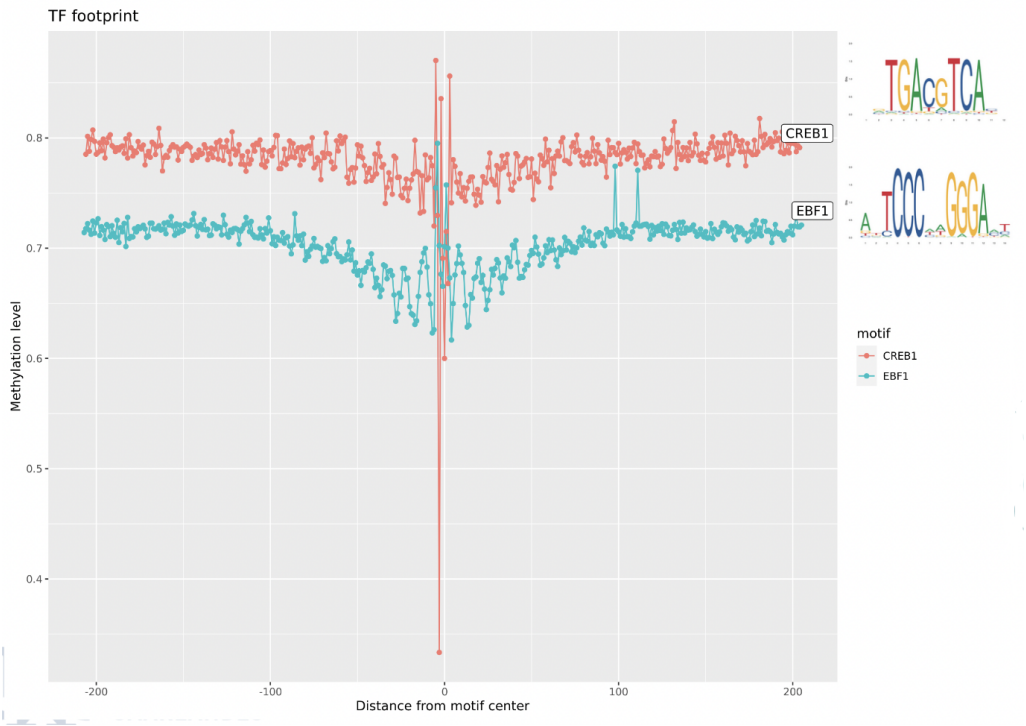

FIGURE 1.9: These TF footprints are based on DNA methylation generated for CREB1 and EBF1, and the figure shows the aggregation of multiple occurrences of these motifs throughout the genome.

With bisulfite-sequencing data from the BLUEPRINT consortium, we implemented TF footprinting for CREB1 and EBF1 gene TFs on one individual B-cell sample. CREB1 (cAMP response element-binding protein 1) and EBF1 (early B-cell factor 1) are transcription factor (TF) motifs that play significant roles in the regulation of gene expression. CREB1 is involved in metabolism, cell growth and differentiation, and stress response. EBF1 is essential for the early development of B cells and influences the differentiation of B cells into immunoglobulin producing cells.

As noted in figure 1.9, the drop in methylation corresponds to the location of the TFBS, where the maximum deviation in methylation occurs exactly at the centre of the motif. Such a drop in methylation at the centre of the TFBS means that the TF has actively bound to the TFBS, which indicates gene transcription activation. We compile a comprehensive score of all such deviations in numerous motifs across the genes of interest and therefore characterise the activation or inactivation of the transcription process itself. We indicate that this method could be used as a precise means to create accurate clustering of different cell types.

Chapter 2

Material and methods

Determining the methylation status in the regulatory regions helps in the characterisation of TF activity. Whole genome bisulfite sequencing data contains information about the exact sites of methylation and hence is well-suited for this purpose. We provide a complete and novel method called methylTFR for the identification of methylation patterns on TFBSs for WGBS data.

2.1 Datasets

Among various international consortia which have benchmarked, standardised and made recommendations for epigenomic profiling. BLUEPRINT consortium is a large-scale European consortium aimed at mapping and understanding the epigenomes of blood cells [36]. The intended outcome of the BLUEPRINT consortium is to decipher how the epigenome changes and responds to disease conditions. Various haematopoietic (blood) cell types from the Blueprint epigenome are employed in this study. Multiple epigenomic data types, including CHIP-seq, RNA-seq, Bisulfite-seq, and DNase-seq, are available for this Blueprint EU project. Whole genome bisulfite sequencing data are being utilised to evaluate our methodology. In addition to the sequencing dataset, interesting parts of the genome, such as proximal and distal enhancers [37], TSS, and others, are used as BED files. The chromosome, start, and end positions of each region are specified in three columns in these bed files. The results discussed in chapter 3 are based on methylation data of B-cell and T-cell samples.

2.1.1 Whole Genome Bisulfite Sequencing Data

In this study, we used bisulfite-seq data from The International Human Epigenome Consortium's (IHEC) Blueprint EU project[38]. There are 206 samples of multiple cell types in it. Before using our method, these samples were analysed using RnBeads. The Epigenome Processing Pipeline, created by Fabian Müller and Christoph Bock (2014), produces bed files in the EPP format. This tab-separated file contains the name of the chromosome, start and end coordinates, the methylation value and coverage as a string ('the number of methylated cytosines M/number

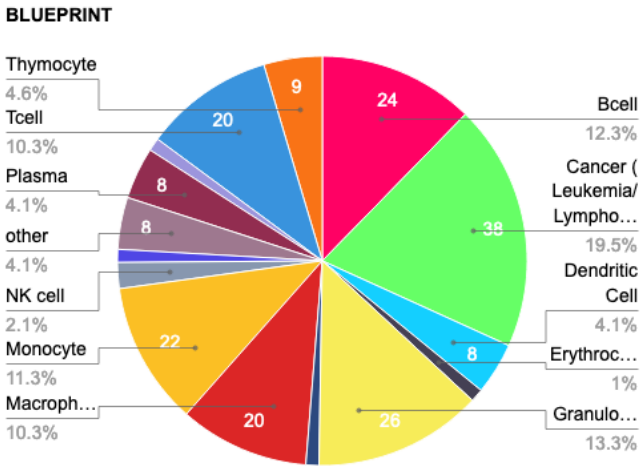

FIGURE 2.1: A summary of Blueprint datasets; Different cell-types and number of WGBS samples for each cell-type

of total reads T'), as well as the methylation level adjusted to the range of 0 to 1000 and the strand. These EPP format bed files are used as Input for our approach.

RnBeads

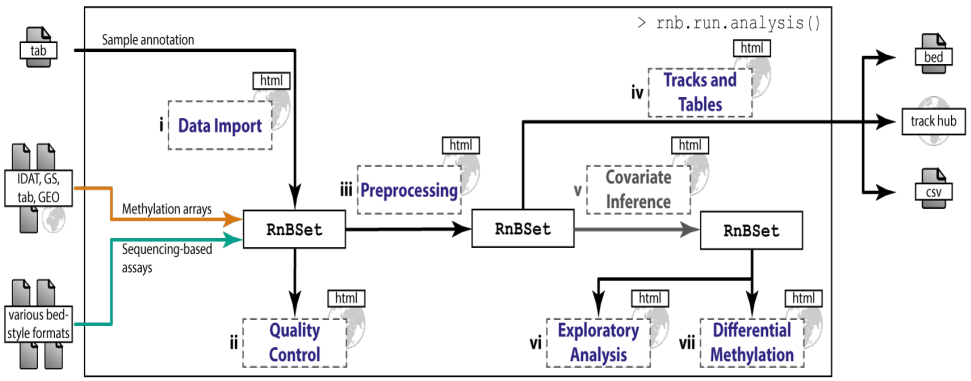

FIGURE 2.2: RnBeads workflow to process the DNA methylation data (Image source: [39])

From data input through filtering, standardisation, and exploratory analytics to identifying differential methylation, RnBeads offers a thorough analytical process. Its modular structure enables targeted, highly-configurable, and simple analysis. It requires only a few parameters, and It handles the rest automatically. There is support for several input formats. Advanced code logging features are also built

into the package. The output may then be displayed as a complete, understandable analysis report in HTML format. The figure 2.2 depicts the typical workflow using RnBeads [40] modules. We used RnBeads to perform quality control and preprocessing of the WGBS data.

2.2 New method development - methylTFR

We developed the methylTFR method, and we built it as an R-package so that it is accessible for the scientific community to test, use and also to collaborate. The method has four steps by which we begin with a set of TF binding regions from databases and use them along with the methylation data to arrive at corrected deviated scores. The four steps are as follows:

1. Obtain catalogues of TF binding regions
2. DNA methylation - TF footprinting
3. Calculating Deviation Score
4. Bias Correction

2.2.1 Obtain catalogues of TF binding regions

Transcription Factors (TFs) can bind to particular DNA regions to regulate the gene or transcriptional activity. These genomic regions are referred to as transcription factor binding sites (TFBS) because they have a strong affinity to bind with TFs and are generally short 9-14 nucleotide sequences. A position frequency matrix (PFM) is a simple and common way to model the interaction between TF and DNA. The PFM summarises the nucleotide frequency distributions of observed TF-DNA interactions at each site. These PFMs were calculated over a period of a few years from various molecular biology experimental methods such as Microarray, CHIP-seq, and SELEX. Developments in sequencing methods increased the coverage of more genomic regions and thereby increasing the number of PFMs. Later these PFMs were collected and organised into multiple databases like JASPAR, CIS-BP, and HOCOMOCO. In this approach, we used the JASPAR CORE, one of the renowned TF-binding profile databases. JASPAR CORE database is accessed programmatically through a RESTful API and a R/Bioconductor package, 'JASPAR2020'.

The first step in our approach is identifying TF binding sites in the whole human genome (hg38). The JASPAR2020 [41] database contains 633 TF binding profiles

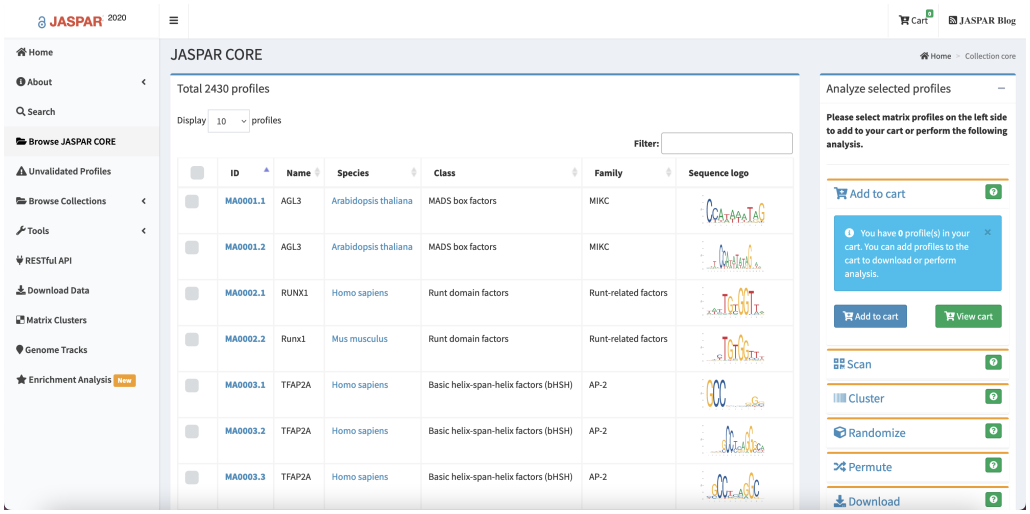

FIGURE 2.3: A screenshot of JASPAR 2020 database homepage

stored as PFMs. These PFMs can identify TF-binding locations in the human genome. To extract binding sites for each TF profile, we used the motifmatchr [42] - R package. 'matchMotifs()' is the primary method of this package, and it is mainly used for short motif searching. It takes a PFM for TF-profile and DNA sequence string as inputs and returns the genomic locations where the particular TF profile or motif can bind with probability scores.

After creating TFBS for each TF profile or motif, we added 250 bp of flanking regions on both sides of binding sites. These flanking regions will be used as background to separate the actual methylation signals around TF binding sites.

Genomic locations of TFBS for each of the 633 TF profiles are combined to create a catalogue of TFBS. This catalogue is stored as a GenomicRanges Object. This catalogue plays a significant role in the identification of the methylation patterns around each TF profile and, therefore, how it could affect the TF-DNA binding affinity to control the gene expression. It has been previously observed that different cell types, cell states, or biological samples can have different methylation patterns around the TFBS in the genome [8].

2.2.2 DNA methylation Footprinting

It has been observed from ATAC-seq data that TF footprinting reveals high DNA accessibility in flanking regions in and around TFBS or the motif sequence facilitating Protein-DNA interaction [35]. Here, we utilise the whole genome bisulfite sequencing (WGBS) data to analyse the TF footprinting. We know that methylated DNA sequences negatively correlate to protein bindings in most cases and suppress transcriptional activity. So, according to the discussion above, TF

binding sites or motifs should have less methylation compared to their flanking regions. This hypomethylation in motifs could show potential affinity towards protein binding in those regions.

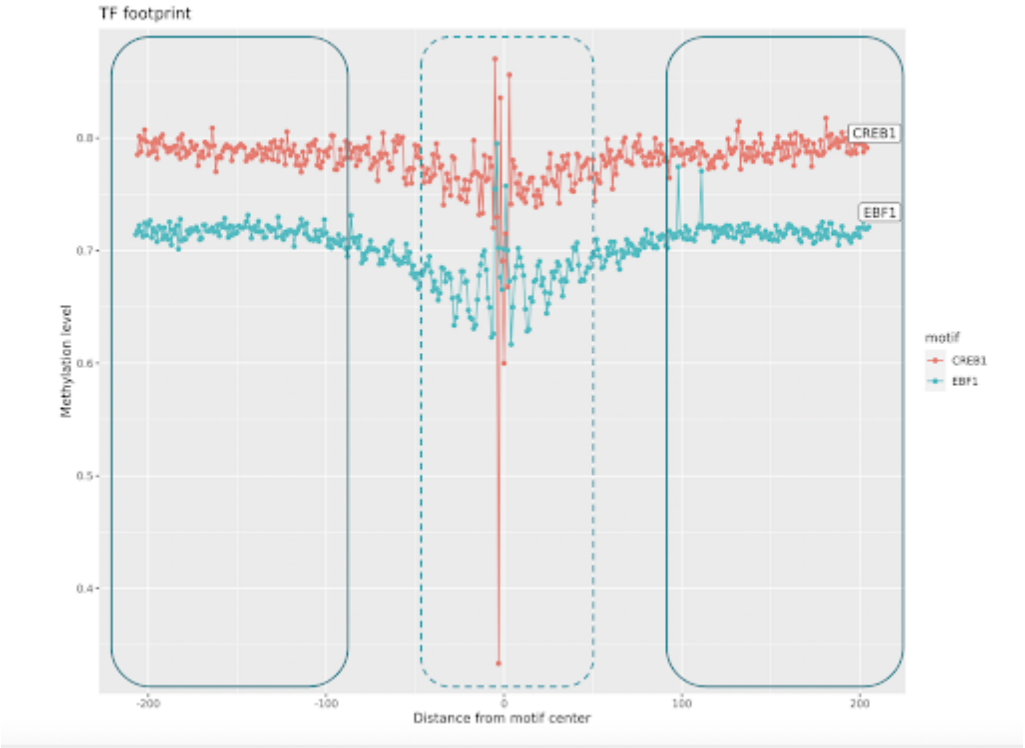

FIGURE 2.4: DNA methylation TF footprint is divided into three components. These are the left background, motif center, and right background regions.

To validate this idea, we can perform a simple analysis to see how the methylation patterns are related to TF footprinting using the findOverlap function from the GenomicRanges. There can be several occurrences of the same DNA motif across the whole genome, and this information is already stored in the TFBS catalogue with 500bp-wide intervals. FindOverlap function assists in determining the overlapped regions of multiple occurrences of motifs and methylation sites. Then, we compute the relative position from the motif center as 0, and the position ranges from -250 to 250 bases. For each position, we calculate the average methylation to show the methylation signal for that particular motif. This shows TF foot printing by DNA methylation (Figure 2.4).

2.2.3 Calculating Deviation Score

After calculating the average methylation at each position, we can see the deviation in the motif center due to hypomethylation. To capture this deviation, we

introduced a scoring system to show how methylation varies around the motif center. The motif center refers to the center position of the actual motif sequences (footprint base) with a length of 9-14 nucleotide bases, with the background containing the flanking regions of 250 bases both sides. The footprint sequence range can be split into three distinct regions: the footprint base (around the motif center) and two backgrounds. The deviation score can be calculated by dividing the average methylation in the motif center by half of the average methylation in the background. This deviation score will provide the overall methylation status for that particular motif.

$$DeviationScore = \frac{Avg.Methylation.Footprintbase(motifcenter)}{\frac{Avg.Methylation.Background.Right + Avg.Methylation.Background.Left}{2}}$$

2.2.4 Bias correction

Sequencing technology has many challenges, even though it has dramatically improved. GC content bias is one of many challenges inherent to the technology; it refers to the relationship identified in sequencing data between the read count (coverage) and GC content. Another factor is the non-random distribution of CG content in the genome. Here we look at DNA methylation GC bias, i.e. GC rich regions tend to have lower methylation levels. We intend to correct for GC bias such that TF motifs in those regions are not affected. We implemented GC bias correction into our method to address this issue. In this bias correction, we used the human genome's quantile values of the GC content distribution to produce five GC bin intervals such as [0, 0.3, 0.36, 0.43, 0.53, 1]. Next, we had to calculate the GC content for each position of the motif, including its flanking region, using a sliding window approach. Based on the GC content calculation, we assigned each position to its corresponding GC bin and created a matrix for each motif occurrence. Aggregating all matrices provided the GC bin frequency table for that particular motif. This GC bin frequency table is stored in annotation packages to reduce the computational cost. Then, we wanted to compute the expected methylation for GC bins. Expected methylation refers to the predicted or expected level of DNA methylation at a specific locus (position) within a genome. The expected methylation vector for GC bins could be estimated when we process the samples. To calculate the bias for a motif, we need to multiply the motif's GC frequency table with the expected methylation vector. Finally, the bias-corrected methylation will be computed by subtracting GC bias from the observed methylation levels.

The deviation score would be calculated from this bias-corrected footprinting. The binding affinity of a TF and its interaction with methylated cytosines would be detectable through this approach genome-wide.

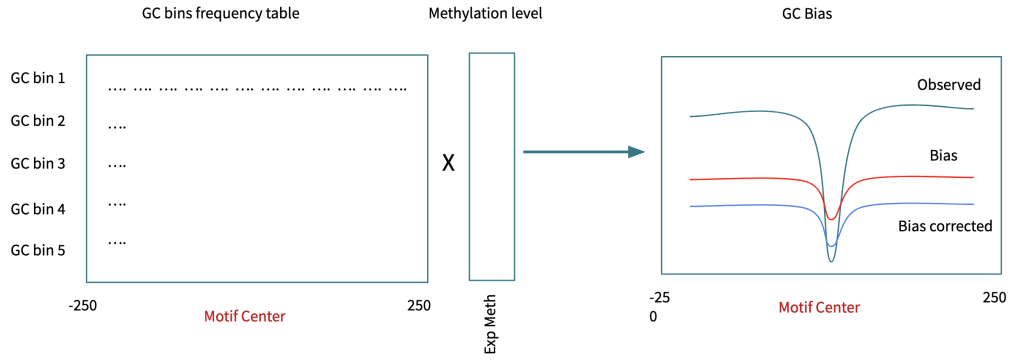

FIGURE 2.5: A diagramatic representation of GC Bias Correction implementation; Bias can be calculated by multiplying GC frequency table and Expected methylation, and It will be subtracted from observed methylation for bias correction

R-package development

We built two R packages to make this method available to other researchers.

- methylTFR: functions required to calculate deviation score, bias correction, plotting footprints, differential testing, and some utility functions
- methylTFRann: Stores required annotation files to run methylTFR, such as Identified TFBS catalogue and GC content information.

Steps followed during the development of R-packages:

1. Planning and implementation: we decided to develop the R package specifically for the purpose of identification of methylation patterns on TFBS using methylation patterns from sequencing data. R package was ideal for this purpose since the supporting functions like GenomicRanges [43], ggplot [44] and dplyr [45] were readily available for import as libraries, thereby reducing development time. The code for various functions was first written separately and functionally unit-tested.
2. Testing: different test cases were written for various conditions and different data states, the code was checked continuously, and the package as a whole was tested to see if it was producing results as expected.
3. Documentation: code functions were commented on and documented during the implementation step. Detailed documentation for the R-package

was written in the form of a vignette. A reference manual that contains instructions on how to download and run the package was also written and published as PDF. This documentation has been updated in the GitHub repository for both our tools.

4. Release: Both the packages are planned to be released on CRAN (the Comprehensive R Archive Network), the central repository for R packages or Bioconductor (Open source R packages archive for bioinformatics analysis), and on GitHub, a popular platform for hosting and sharing code. methylTFR comes as a tool, and methylTFRann contains the reference data as RData objects (rds files) required for annotation.

MethylTFR available on GitHub: `https://github.com/EpigenomeInformatics/methylTFR` and this repository contains detailed documentation with example dataset.

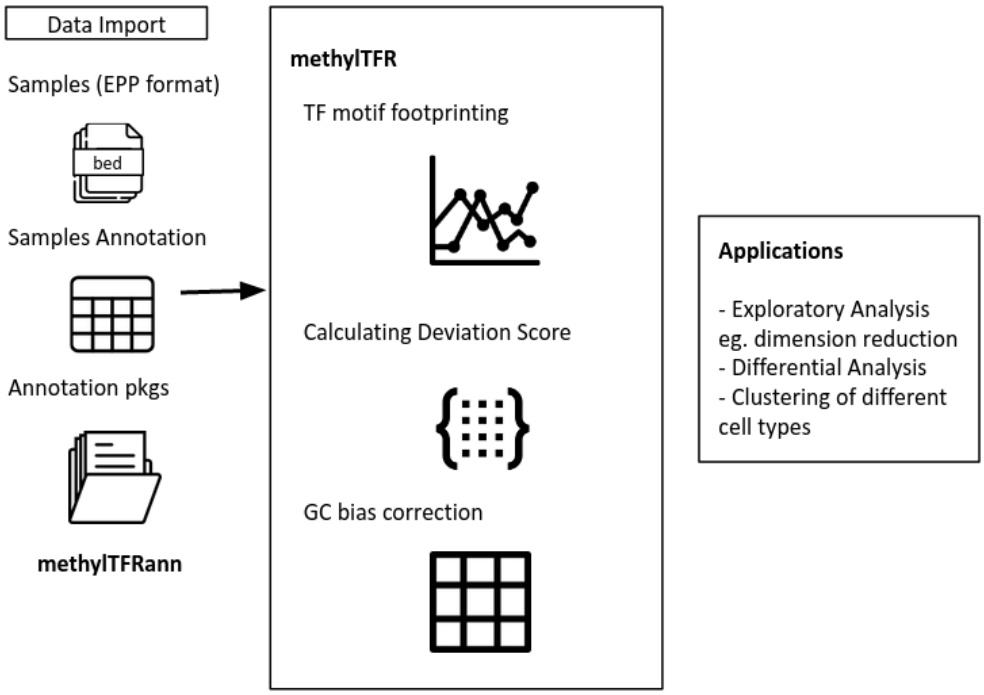

FIGURE 2.6: Schematic diagram of how methylTFR works and its downstream applications

2.3 Statistical Analysis

2.3.1 Differential Analysis

Differential analysis is a statistical method used to identify differences in two or more groups of samples. The goal of the analysis is to identify differences in deviation scores between the statistically significant sequences or to determine how much the groups of deviation scores are spread out across our observations. In a class comparison or differential analysis of methylation deviation patterns, the methylation status of a specific region of DNA is assessed across a group of samples that are divided into two or more classes based on some predetermined criterion, such as disease status or exposure to a particular environmental factor.

To perform this differential analysis, a statistical test is used to compare the deviation scores between the groups. The choice of statistical test will depend on the characteristics of the data, such as the sample size and the distribution of the deviation scores. Once the statistical test has been performed, the results are evaluated for statistical significance, taking into account the multiple testing correction to account for the fact that multiple comparisons are being made. If a difference between the deviation scores is found to be statistically significant, it may be further investigated as a potential factor influencing transcription factor activity in the context of the classes being compared.

In our study, we initially used the **Wilcoxon rank-sum test** [46] (also called the Mann-Whitney test), which is a non-parametric test to compare data. If there are disparities between two sets of data that are not normally distributed, the Wilcoxon signed-rank test is employed. It is typically applied as a non-parametric substitute for the paired- or one-sample t test. It compares the differences between two related samples and creates a pooled ranking of the differences. The two sample medians being equal is the test's null hypothesis. This makes the Wilcoxon rank sum test valuable for data that is not normally distributed or for small sample sizes. In simpler words, the Wilcoxon rank sum test is used as the statistical method to determine the association between two factors.

When we have multiple sample groups in several batches or categories, the Wilcoxon rank sum test may not be applicable since it can handle only two samples at a time. In such cases, we prefer the nonparametric variant of one-way analysis of variance test, or ANOVA, known as the **Kruskal-Wallis Test**. It is used to assess if the medians of two or more different groups diverge statistically significantly from one another [47]. The test determines if independent groups have the same

mean on ranks; instead of using the data values themselves, a rank is assigned to each observation. The Kruskal-Wallis test assesses the differences against the average ranks to determine whether they are likely to have come from samples drawn from the same population.

Calculate the test statistic using the formula:

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3N + 1$$

where,

- N - Total number of bbservations
- n_i - Number of observations in the i-th group
- R_i - Total sum of ranks in the i-th group

Like any other statistical method, the Kruskal-Wallis test also has a limitation. If we do not find a significant difference in the data while doing the test, we cannot say that the samples are the same.

We have calculated P-values using the Wilcoxon or Kruskal Wallis test in differential testing for 633 motifs. Due to multiple statistical tests, the false discovery rate will be high. We used BH correction to calculate the adjusted P-value to control for this multiple-hypothesis testing.

The **Benjamini-Hochberg (BH)** correction is used to control the false discovery rate (FDR) in statistical tests by adjusting the p-values of the individual tests based on the number of tests being performed [48]. This correction method ranks the p-values from smallest to largest and then compares them to a predetermined threshold. The threshold is calculated based on the total number of tests being performed and the desired level of FDR.

The formula for the BH correction is:

$$p^{\mathrm{BH}} = \min\left\{\frac{mp}{i}, 1\right\}.$$

where:

- m is the total number of tests being performed
- p is the original p-value
- i is the rank of the test, orderered by p-value

2.3.2 Clustering Analysis

We did a comparison study of clustering samples between the deviation score matrix and 1kb tiling regions approach (average methylation of 1kb genomic regions). K-means clustering was used to cluster samples based on both data points. We used the Jaccard Index and adjusted Rand Index (ARI) to compare the clustering. The final results show that deviation score-based clustering performed better compared to the 1kb tiling approach.

The Jaccard index

The Jaccard index is calculated as the size of the intersection of two sets divided by the size of the union of the two sets. This index measures the similarity between the two sets, with a higher index indicating greater similarity [49]. The Jaccard index can be used to compare the similarity of any two sets. It is often used in data mining and machine learning applications to evaluate the similarity of sets of items or features.

The formula for the Jaccard index is for A and B (sets):

$$Jaccardindex = \frac{|A \cap B|}{|A \cup B|}$$

Adjusted Rand Index

The adjusted Rand index (ARI) measures the similarity between two clusterings or partitionings of a dataset [49]. It is often used in machine learning and data mining to evaluate the quality of clustering algorithms or to compare different clustering results. The ARI ranges from -1 to 1, with higher values indicating more significant similarity between the two clusterings. A value of 0 indicates that the two clusterings are no better than random, while a value of 1 indicates that the two clusterings are identical.

Calculate the adjusted Rand index using the formula,

$$ARI = (RI - E(RI))/(max(RI) - E(RI))$$

where:

- RI is the raw Rand index
- E(RI) is the expected value of the Rand index under the null hypothesis of random clusterings

- max(RI) is the maximum possible value of the Rand index

The raw Rand Index (RI) formula is:

$$RI = (a + d)/(a + b + c + d)$$

Where:

- a is the number of pairs of elements in the same cluster in both clusterings.
- b is the number of pairs of elements in the same cluster in the first clustering but different clusters in the second.
- c is the number of pairs of elements in different clusters in the first clustering but the same cluster in the second.
- d is the number of pairs of elements in different clusters in both clusterings.

Chapter 3

Results and discussion

Our method, methylTFR, was used to study the methylation patterns and states in the TFBS. Here we will discuss the findings of various studies performed using a deviation score matrix that has been bias-corrected. These analyses are categorised mainly into exploratory analysis, dimensionality reduction, differential analysis, and cluster comparison analysis. All the results in this section are based on the analysis performed on B-cell and T-cell WGBS data from the Blueprint epigenome EU project.

3.1 Exploratory Analysis

The methylation status of transcription factor binding sites is a crucial determinant of their function. Methylation at TFBS in enhancers or promoter regions will competitively inhibit the binding of transcription factors and thereby repress gene expression. However, methylation at TFBS in repressor regions means blocking repressor activity which in turn means up-regulation of the gene expression. Nevertheless, it is not as simple as this inside a living cell. As previously discussed, increased or decreased methylation is purely context-specific, and the regulated gene expression is a resultant effect of the overall methylation status of all related genomic elements.

Based on this hypothesis, we developed our method and applied it to methylation data of B-cells and T-cells. Whole genome bisulfite sequencing data of 25 B-cells and 26 T-cells in BED file format, including the methylation levels for each CpG site. We used the RnBeads pipeline to process these raw methylation bed files for quality control and some preliminary analysis. The output bed files from the rnbeads analysis are in EPP format, described in the 2.1 section. In the initial analysis, we discovered that seven samples cluster (6 T-cells and 1 B-cell) were isolated far from the other samples. It revealed the possible outlier samples in the dataset. We eliminated them from the remaining analysis.

After calculating the deviation scores for the passed B-cells (24) and T-cells (20), we wanted to check if there were any signals to distinguish the cell types. Different transcription factor motifs are responsible for specific cell differentiation and

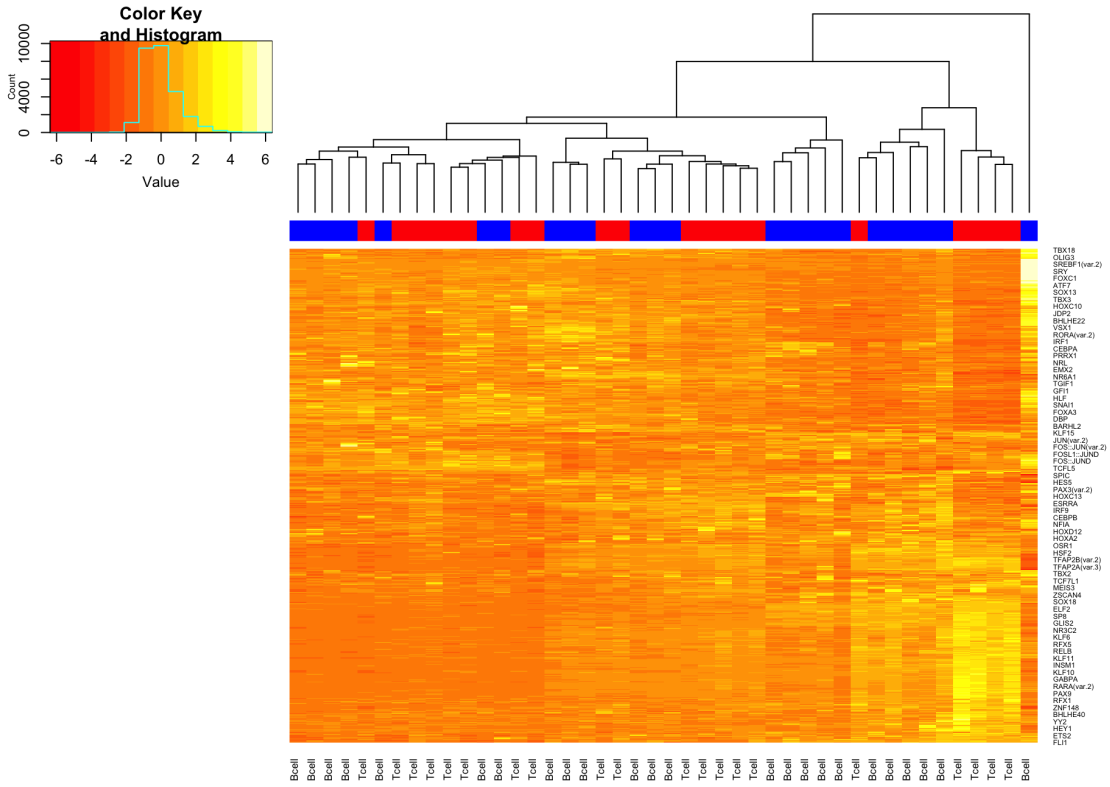

FIGURE 3.1: Heatmap visualisation of raw deviation with z-score, B-cells and T-cells. X and Y axis are represented as samples and TF motifs (633), respectively. The colour key for the z-score with histogram was shown in the top left corner of the heatmap. Hierarchical clustering was applied on these samples on top of this plot; The red colour in the clustering bar indicates the T-cell cluster, and the blue colour indicates the B-cell cluster

maturation processes, so we tried multiple approaches to extract biologically meaningful outcomes in this exploratory analysis. To obtain a glance at the calculated deviation score, we plotted a heatmap shown in Figure 3.1. X and Y axis are represented as samples and known TF motifs respectively.Hierarchical clustering of Bcells (blue) and Tcells (red) was applied on top.

Even though the cell types couldn't be clearly distinguished by the heatmap, we can notice multiple isolated clusters of T-cells and B-cells, and we could identify one B-cell sample entirely away from the rest of the samples. This could have happened due to various reasons. We could observe the same pattern while analysing the methylation levels matrix with tiling regions of 1kb from the rnbeads pipeline. But the 1kb tiling region matrix contains 22 million rows, whereas the deviation score matrix has only 633 TF motifs. Processing 22M data points for multiple samples requires high computational resources and is prone to operational bias. On the other hand, the TF motifs deviation score matrix has

increased the interpretation of biological significance. This overview analysis clearly shows the advantage of using TF motif regions for methylation analysis.

3.1.1 The Ensembl Regulatory Build

The Ensembl regulatory database is a collection of genomic regions which has the potential to regulate gene expressions. We downloaded the regulatory regions as bed files from the Ensembl. This ensembl regulatory build was actually derived from the BLUEPRINT projects' histone modifications datasets. There are five regulatory region sets: a distal enhancer, proximal enhancer, DNase accessible, transcription start site and transcription factor binding sites. These regions are built from the consensus segmentation across histone modifications and other epigenetic marks.

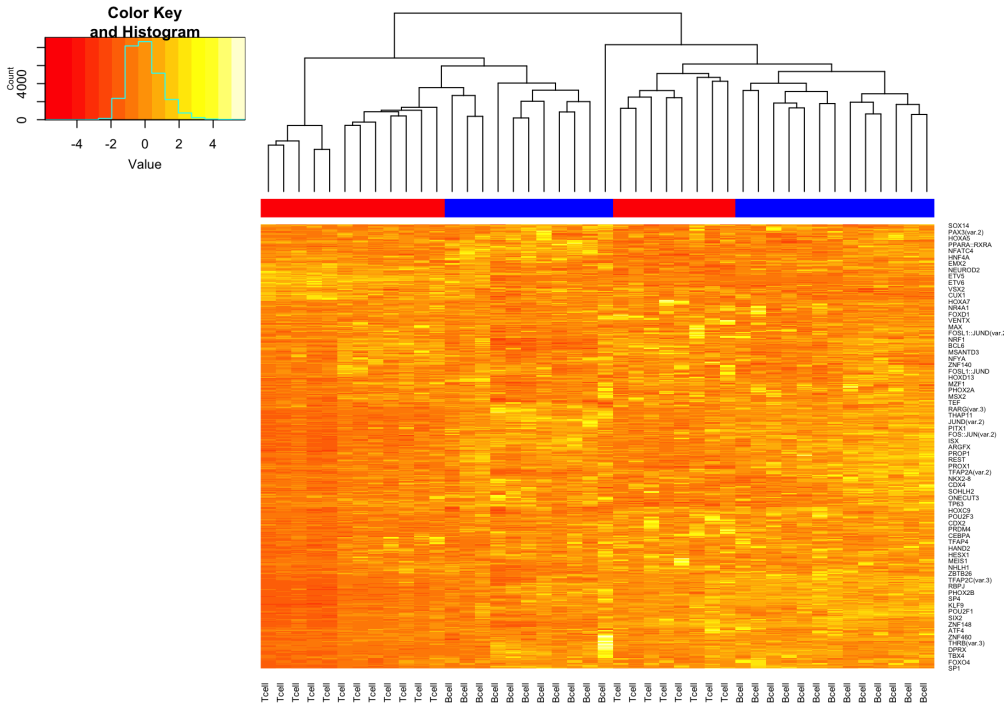

FIGURE 3.2: Heatmap visualisation of distal deviation with z-score, B-cells and T-cells. X and Y axis are represented as samples and TF motifs (633), respectively. The colour key for the z-score with histogram was shown in the top left corner of the heatmap. Hierarchical clustering was applied on these samples on top of this plot; The red colour in the clustering bar indicates the T-cell cluster, and the blue colour indicates the B-cell cluster

From this Ensembl regulation, we used distal enhancer regions to enrich further the performance of deviation score on the predicted transcription factor binding sites. Because a lot of DNA methylation variability actually happens in distal

regulatory regions. We created the exact heatmap visualisation for the distal deviation score (Figure 3.2). The distal deviation score can be computed similarly; however, instead of complete TF binding sites, we use overlapping binding sites with distal enhancer regions. Figure 2 heatmap shows the distal deviation score with hierarchical clustering on the samples.

3.2 Dimensionality Reduction

Dimensionality reduction is a technique used in machine learning and data analysis to reduce the number of features or dimensions in a dataset. This is often useful when dealing with high-dimensional datasets, as they can be challenging to analyse and visualise due to their many features. Dimensionality reduction can also reduce the computational cost of processing massive data and improve performance by removing noise and redundant features. One of the most commonly used dimensionality reduction methods is Principal Component Analysis (PCA). There are other methods, such as t-SNE and UMAP, for better visualisation. Here, we applied a simple principal component analysis (PCA) on the deviation score matrix to visualise in two dimensions.

3.2.1 Principal Component Analysis (PCA)

PCA is a linear dimensionality reduction technique that finds the directions in the data that capture the most variance. It does this by constructing a set of orthogonal axes, known as principal components, which capture the most variance in the data. The first principal component is the axis that captures the most variance, the second principal component is the axis that captures the second most variance, and so on. One of the benefits of PCA is that it can be used to visualise high-dimensional data in two or three dimensions. This can be useful for understanding the underlying structure of the data and identifying patterns or trends. However, it is crucial to remember that PCA is a linear method and may not be suitable for all types of data. Other dimensionality reduction methods, such as non-linear methods like t-SNE, and UMAP, may be more appropriate for specific data types. We observe that there is an apparent separation between B-cell and T-cell samples in figure 3.3. But there is a maximum of 23.2% variance explained in PC1 and PC2 has a 10% variance.

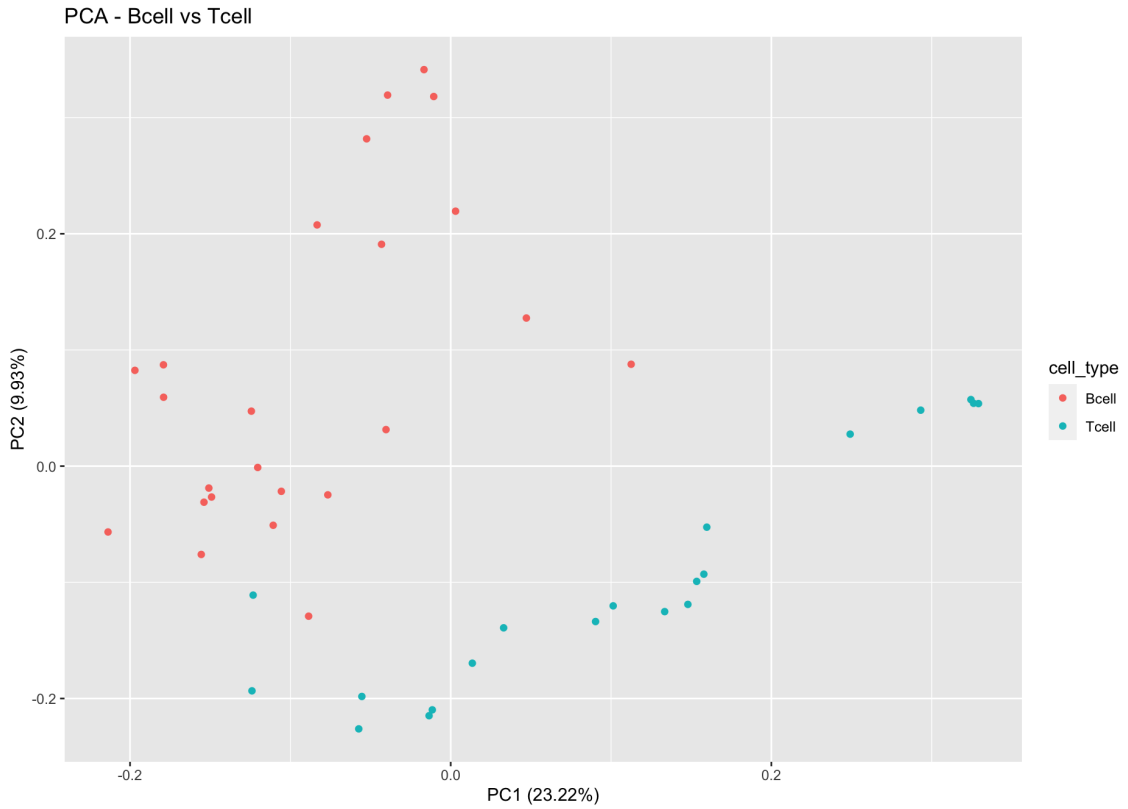

FIGURE 3.3: PCA on distal deviation score to view samples in 2D

3.2.2 Uniform Manifold Approximation and Projection (UMAP)

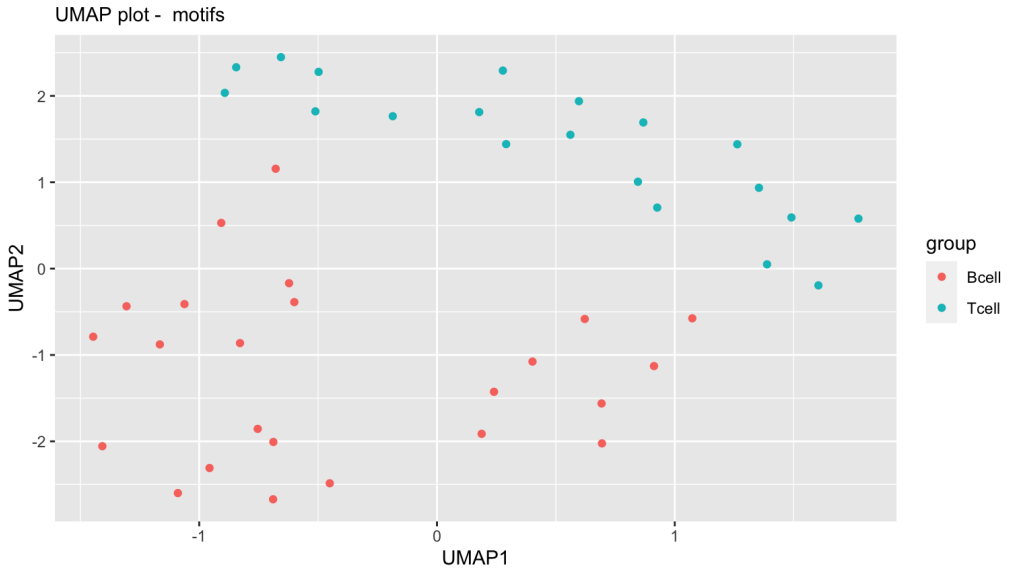

FIGURE 3.4: UMAP on distal deviation score

UMAP is another dimensionality reduction approach that is used to reduce the complexity of high-dimensional data while preserving as much of the original

structure of the data as possible. It does this by constructing a low-dimensional, non-linear manifold that approximates the original high-dimensional data and then projecting the data points onto this manifold. This is useful for visualising high-dimensional data, finding patterns in the data and clustering the data. We could observe the same pattern in the non-linear approach as well from the figure 3.4.

3.3 Differential Analysis

Differential testing is a proper statistical method for comparing the means of two or more groups and assessing whether the differences are statistically significant. The Wilcoxon rank-sum test and the Kruskal-Wallis(KW) rank-sum test are commonly used differential tests. Still, it is essential to carefully consider which test is most appropriate for a given dataset. We implemented these methods in our tool to do the differential analysis on the deviation score matrix.

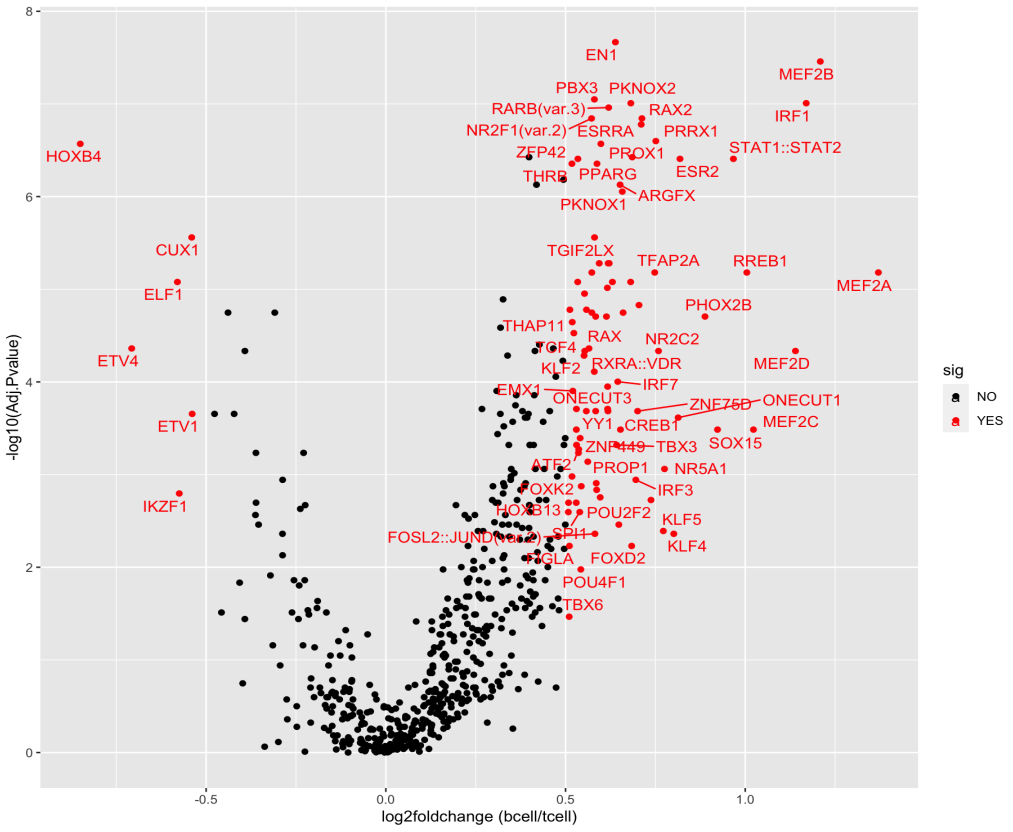

FIGURE 3.5: Volcano plot for differentially deviated TF motifs between B-cells and T-cells; The log2 fold change calculated from T-cell to B-cell (log2foldchange(Bcell/Tcell)). So, the positive value motifs are significantly differentiated in B-cell, and Negative value motifs are from T-cell.

We performed the Wilcoxon test between B-cells (24) and T-cells (20) for each motif and adjusted the P-value using BH correction to reduce the false discovery rate. We created a volcano plot to visualize the differentially deviated motifs and applied a significant cut-off for p-value as 0.05 and log fold change as 0.5. The significantly differentiated motifs are labelled in the volcano plot. SOX15 (SRY-box containing gene 15) and HOXB4 (homeobox B4) are expressed in a variety of tissues and are also involved in the regulation of gene expression during development, including the development of the nervous system, gonads, and cartilage. Moreover, HOXB4 and PBX3 (pre-B cell leukaemia homeobox 3) are involved in the development and function of the blood and immune systems.

3.4 Cell-type specific motif footprints

Motifs can be used to identify specific genes or regulatory elements necessary for the function and identity of a particular cell type. For example, each transcription factor has a unique DNA binding motif that it recognizes, and this motif can be used to identify genes that are regulated by that transcription factor. In this way, cell type-specific motifs can be used to understand the transcriptional regulation of genes in a particular cell type and how this regulation contributes to the function and identity of that cell type.

Understanding cell type-specific motifs are essential for many areas of biological research, including the study of the development, disease, and function of different cell types in the body. By identifying and characterising these motifs, researchers can gain insight into the molecular mechanisms that underlie the function and identity of different cell types and potentially identify new therapeutic targets for the treatment of diseases. For example, EBF1 (early B-cell factor 1) is a transcription factor that plays a critical role in the development of B cells, a type of immune cell that is involved in the production of antibodies and the protection against infections. EBF1 is expressed in early B cell precursors, and its expression is required for the development of B cells from hematopoietic stem cells.

EBF1 activates the transcription of several genes that are important for B cell development, including the gene encoding the B cell receptor (BCR), which allows B cells to recognize and respond to specific pathogens. EBF1 also regulates the expression of other transcription factors, such as Pax5 and Ikaros, which are also required for B cell development.

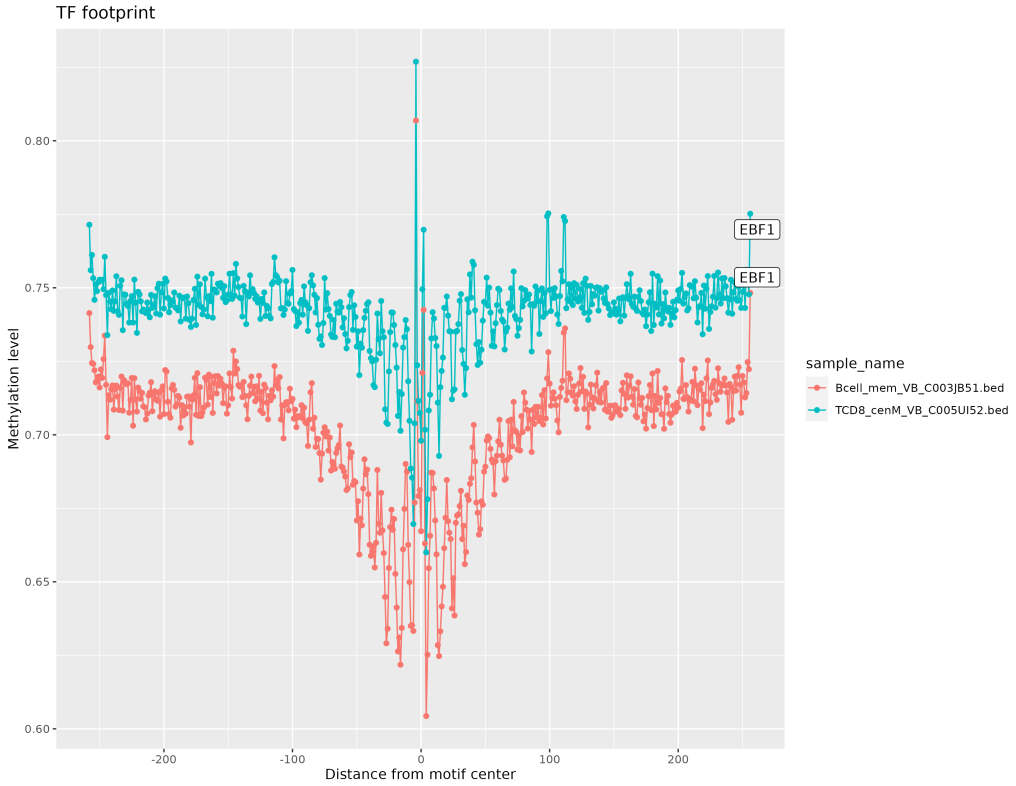

FIGURE 3.6: To visualize the cell-type specific motif, this TF footprint was plotted from one B-cell and one T-cell sample for EBF1 (Early B-cell factor 1); B-cells are represented by the red line, and T-cells by the blue line.

3.5 Cluster Comparison Analysis

With the help of cell-type specific motifs, we aimed to create an accurate clustering of cells. In order to achieve that, we used an unsupervised learning approach such as k-means clustering to check using deviation scores. K-means clustering is a popular and straightforward method for unsupervised learning in which a dataset is partitioned into a predefined number of clusters (k). The goal of the k-means clustering is to group similar data points together and identify patterns or trends in the data. One of the main advantages of the k-means clustering is its simplicity and efficiency. It is a fast and straightforward method for partitioning a dataset into clusters, and it is easy to implement and interpret. However, one of the main drawbacks of the k-means clustering is that it requires the user to specify the number of clusters in advance, which may not always be known or easy to determine.

We applied k-means clustering on the distal deviation score and 1kb tiling region matrix with multiple k values. We computed the Jaccard index and adjusted-rand

index (ARI) to compare the results between k-means clustering and actual cell types. Both the Jaccard index and ARI are measures of similarity. Hence, as previously noted in Methods & Materials, for the Jaccard index, the maximum possible score is one, and the minimum possible score is zero. However, in the case of ARI, the max-min range is 1 to -1.

To identify how competently the clustering happened in terms of cell subtypes, we created two scenarios where the actual cell subtypes are grouped based on their primary morphology and their subtypes.

We performed clustering analysis on these two scenarios and plotted the similarity scores obtained using Jaccard index and ARI on deviation and tiling observations.

Clustering analysis on scenario-1: Primary Cell Types (B-cell, TCD4, TC8) vs Unsupervised K-means clustering (deviations or tiling)

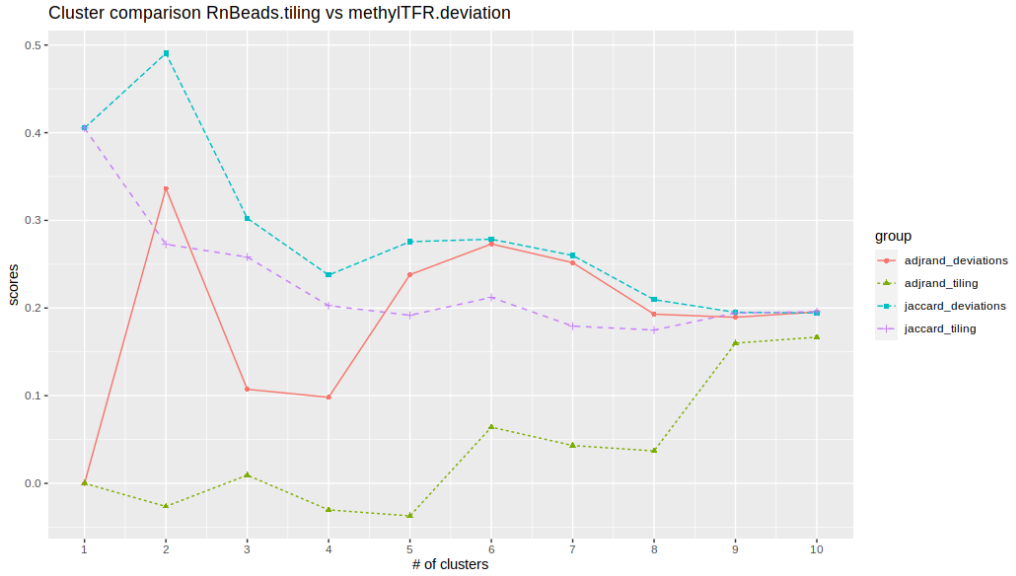

FIGURE 3.7: The line plot shows the result of cluster comparison studies between the primary cell types and k-means clustering (using deviation scores or 1kb tiling methylation scores); X-axis represents K values from K means cluster; Jaccard Index and ARI have been calculated for each K means cluster against primary cell-types (B-cell, TCD4, TCD8)

We observed that the overall deviation-based clustering performs better compared to 1kb tiling regions. We also noticed that 2 and 6 have the highest Jaccard and ARI similarity score for deviations. This infers that they are distinct in their cell type, as indicated by the huge divergence in their deviation scores.

Clustering analysis on scenario 2: cellular subtypes vs Unsupervised K-means clustering (deviations or tiling)

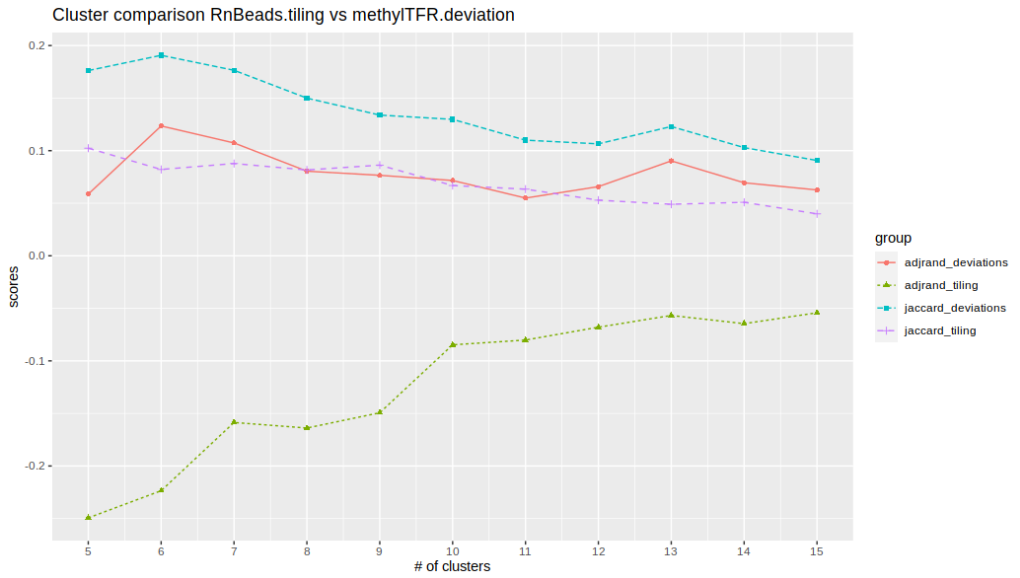

FIGURE 3.8: This line plot shows the result of cluster comparison studies between the cell sub-types and k-means clustering (using deviation scores or 1kb tiling methylation scores); X-axis represents K values from K means cluster; Jaccard Index and ARI have been calculated for each K means cluster against 12 cell sub-types (Bcell-naive, Bcell-pre, Bcell-mem, Bcell-gc, TCD4-cenM, TCD4-effM, TCD4-VB, TCD8-CB, TCD8-cenM, TCD8-effM, TCD8-term, TCD8-VB)

We observed that similar to Scenario 1, the overall deviation-based clustering performs better compared to 1kb tiling regions here as well. Since we have already classified and grouped the cells into several subgroups based on their subtypes, the divergence between their scores is not as clearly evident as we noticed in Scenario 1. However, when we compare the plots between Scenario 1 and Scenario 2, we can notice that the concordance between similarity scores calculated based on deviation scores is much more discernible in Scenario 2.

To summarize, motif deviation scores reflect the presence and abundance of specific DNA binding motifs within a region, which are often associated with specific transcription factors and regulatory elements. On the other hand, tiling methylation levels provide a general measure of methylation levels across a region without taking into account the specific sites of methylation and their potential functional significance. In terms of future studies, the use of motif scores can provide a more nuanced and specific understanding of epigenetic regulation, and help researchers identify potential regulatory elements and target genes with greater accuracy.

Chapter 4

Conclusion and future work

4.1 Conclusion

The primary objective of this project is to develop a computational approach to recognise the methylation marks around motif sequences and provide meaningful insights. In general, TF footprint analysis was performed for chromatin accessibility using ATAC-seq data described in this article [35]. We adopted the TF footprinting approach to bulk bisulfite sequencing data and calculated the deviation score for each motif, as described in chapter two. From our knowledge, this is a novel computational approach to analysing the methylation around motifs.

The deviation score represents differences between motif center and background methylation signals. More deviation around the motif center shows the high possibility of TF binding to that motif sequence. To validate our method, We applied it to human blood cell samples (B-cells and T-cells) from blueprint epigenome data.

Interestingly, we could identify outlier samples with the help of deviation scores. After removing outliers, we used a bias-corrected deviation score for hierarchical clustering with heatmap visualisation. Though clustering appeared as small groups within the same cell types, there is no clear separation between cell types. To improve the accuracy, we utilised the curated distal enhancer regions from ensembl regulatory builds. The distal deviation score clearly shows the divergence between two different cell types. These observations enable us to understand the transcription factor associated with methylation patterns.

The distal deviation score can be used for multiple downstream analyses, such as differential testing and dimensionality reduction (PCA, UMAP). We applied PCA on the distal deviation matrix of b-cell and t-cell samples to visualise the data points in 2D space. There is a 24 % variance explained in principal component one. Differential testing provides significantly differentiated motifs between B-cells and T-cells. The top significant motifs could be responsible for specific cell types and could be employed as a new biomarker for any disease state.

Overall, based on the above findings, we can draw the conclusion that deviation score-based clustering works significantly better in accordance with the cell type annotation and could thus be considered to be more informative than the tiling-based approach. Additionally, it minimises the number of features from a million tiling regions to a TF-focused representation of the data that is easy to comprehend. Furthermore, compared to tiling-based analysis, deviation-based clustering takes less time and uses computing power more efficiently.

4.2 Future work

This study involves various biological and technical elements, providing a platform for numerous modifications, testing, and enhancements over the current model that can be produced with the further time commitment. The ideas that have been suggested but have yet to be implemented are listed below.

- Further testing on different datasets: This approach was applied to only B-cells and T-cells of blueprint epigenome samples. Currently, there is readily available a significant collection of epigenetic data that could be used to evaluate the proposed method and give the results more context.
- Extend to single-cell analysis: This approach was developed and tested on bulk bisulfite sequencing data. But, we can also expand this method for single-cell methylation data sets, where it could particularly help with data sparsity.
- Removing TF motif redundancy: There are a number of motif sequences that are very similar to one another. These comparable motifs can be clustered to improve comprehensibility.
- Integration with RnBeads: We can build a module based on the methylTFR tool and incorporate it into the RnBeads pipeline.
- Filters for TFBS catalogue: we can employ experimentally validated or neural network model-predicted transcription factor binding sites for more accurate representation.
- By adding parallelism at the code level, we may further optimise the runtime and boost performance.

Appendix A

Supplementary figures

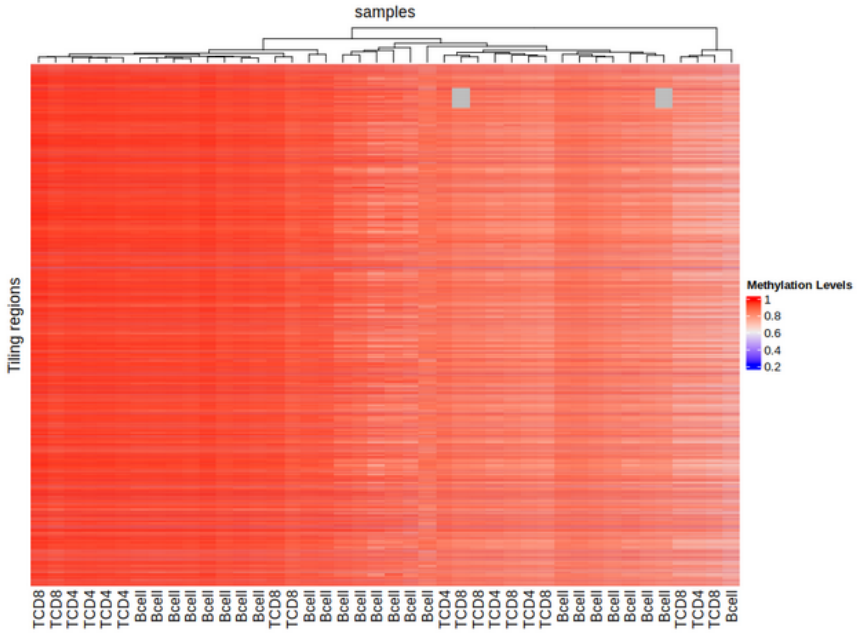

FIGURE A.1: Heatmap Visualization for 1kb Tiling region methylation

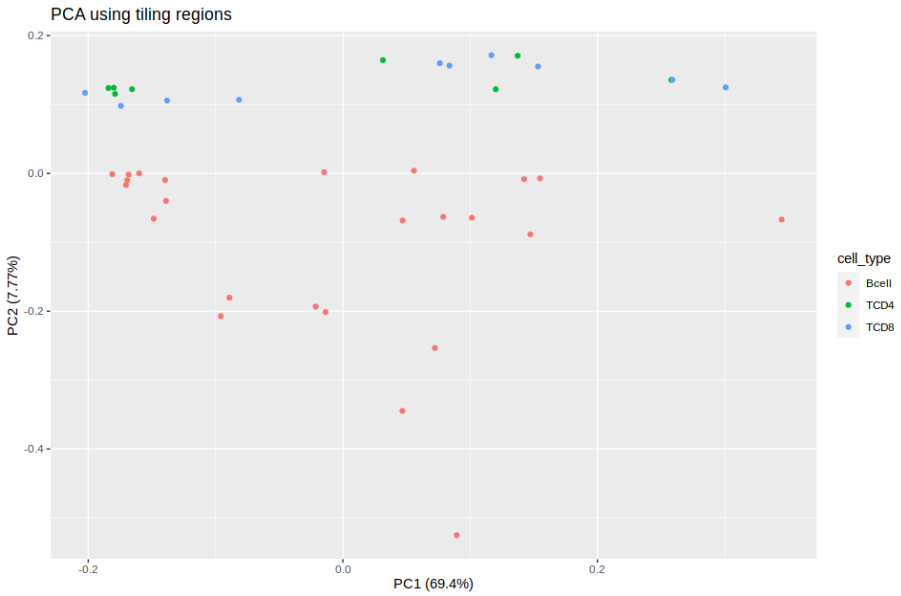

FIGURE A.2: PCA on 1kb Tiling region methylation - B-cell and Tcell

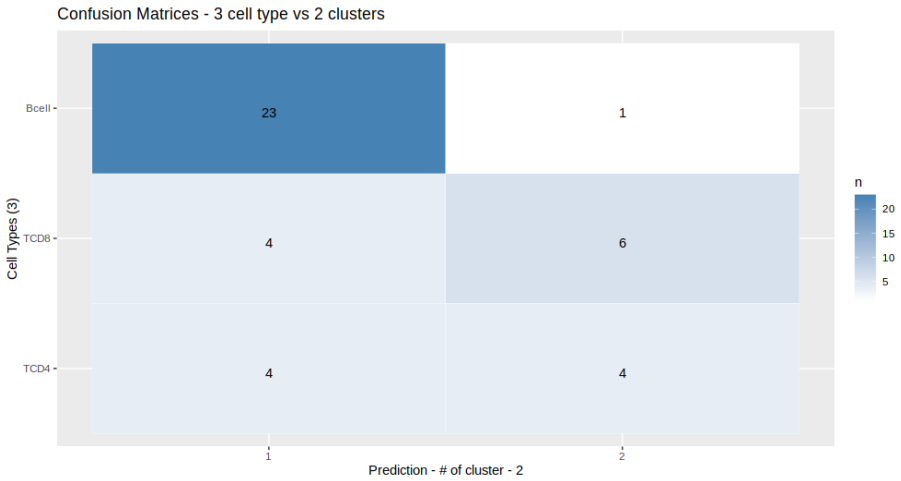

FIGURE A.3: Confusion matrix between primary cell types and predicted clusters by K means (K-2) clustering

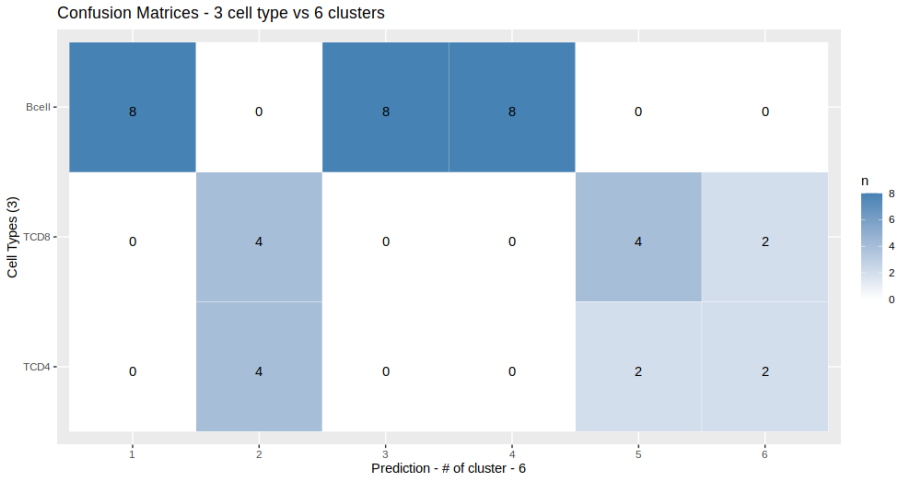

FIGURE A.4: Confusion matrix between primary cell types and predicted clusters by K means (K-6) clustering

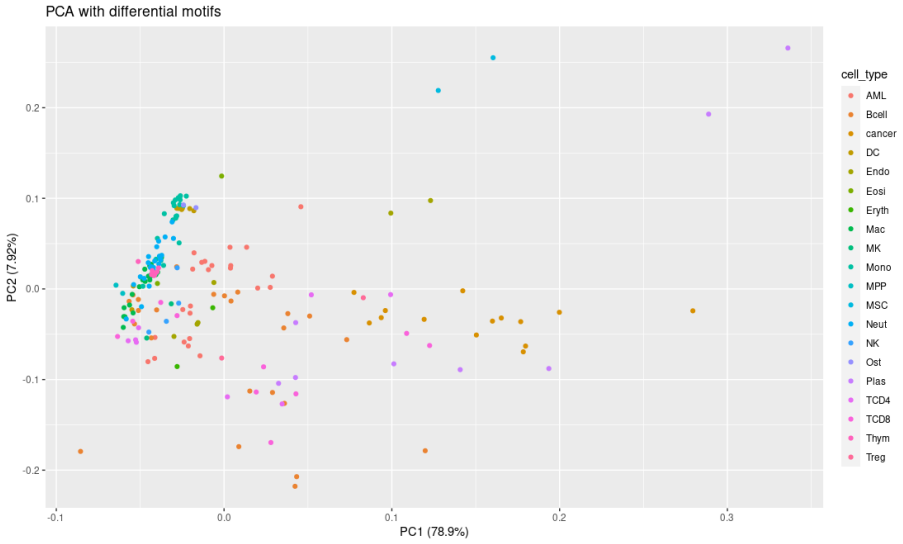

FIGURE A.5: PCA on all WGBS samples from BLUEPRINT data using significantly differentiated motifs among groups

Bibliography

[1] James D Watson and Francis HC Crick. "Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid". In: *Nature* 171.4356 (1953), pp. 737–738.

[2] Francis Crick. "Central dogma of molecular biology". In: *Nature* 227.5258 (1970), pp. 561–563.

[3] Tong Ihn Lee and Richard A Young. "Transcriptional regulation and its misregulation in disease". In: *Cell* 152.6 (2013), pp. 1237–1251.

[4] Bruce Alberts et al. "The structure and function of DNA". In: *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[5] Mar. 2018. URL: http://commonfund.nih.gov/epigenomics/figure.

[6] Stephanie Clare Roth. "What is genomic medicine?" In: *Journal of the Medical Library Association: JMLA* 107.3 (2019), p. 442.

[7] Kimberly E Stephens et al. "Epigenetic regulation and measurement of epigenetic changes". In: *Biological research for nursing* 15.4 (2013), pp. 373–381.

[8] Dirk Schübeler. "Function and information content of DNA methylation". In: *Nature* 517.7534 (2015), pp. 321–326.

[9] Samuel A Lambert et al. "The human transcription factors". In: *Cell* 172.4 (2018), pp. 650–665.

[10] John Newell-Price, Adrian JL Clark, and Peter King. "DNA methylation and silencing of gene expression". In: *Trends in Endocrinology & Metabolism* 11.4 (2000), pp. 142–148.

[11] Inderpreet Sur and Jussi Taipale. "The role of enhancers in cancer". In: *Nature Reviews Cancer* 16.8 (2016), pp. 483–493.

[12] Stephen B Baylin. "DNA methylation and gene silencing in cancer". In: *Nature clinical practice Oncology* 2.1 (2005), S4–S11.

[13] Lewis R Silverman et al. "Randomized controlled trial of azacitidine in patients with the myelodysplastic syndrome: a study of the cancer and leukemia group B". In: *Journal of Clinical oncology* 20.10 (2002), pp. 2429–2440.

[14] Mariuswalter. *DNA methylation landscape in mammals*. [Online; accessed 8-Jan-2023]. 2019. URL: https://commons.wikimedia.org/wiki/File:DNAme_landscape.png.

[15] Taishan Hu et al. "Next-generation sequencing technologies: An overview". In: *Human Immunology* 82.11 (2021), pp. 801–811.

[16] Fei-Man Hsu et al. "Chapter 4 - Bioinformatics of Epigenomic Data Generated From Next-Generation Sequencing". In: *Epigenetics in Human Disease (Second Edition)*. Ed. by Trygve O. Tollefsbol. Second Edition. Vol. 6. Translational Epigenetics. Academic Press, 2018, pp. 65–106. DOI: https://doi.org/10.1016/B978-0-12-812215-0.00004-2. URL: https://www.sciencedirect.com/science/article/pii/B9780128122150000042.

[17] René AM Dirks, Hendrik G Stunnenberg, and Hendrik Marks. "Genome-wide epigenomic profiling for biomarker discovery". In: *Clinical epigenetics* 8.1 (2016), pp. 1–17.

[18] Daniel Beck, Millissia Ben Maamar, and Michael K Skinner. "Genome-wide CpG density and DNA methylation analysis method (MeDIP, RRBS, and WGBS) comparisons". In: *Epigenetics* 17.5 (2022), pp. 518–530.

[19] Ting Gong et al. "Analysis and Performance Assessment of the Whole Genome Bisulfite Sequencing Data Workflow: Currently Available Tools and a Practical Guide to Advance DNA Methylation Studies". In: *Small Methods* 6.3 (2022), p. 2101251.

[20] Felix Krueger and Simon R Andrews. "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications". In: *bioinformatics* 27.11 (2011), pp. 1571–1572.

[21] Mark A Urich et al. "MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing". In: *Nature protocols* 10.3 (2015), pp. 475–483.

[22] Gordon L Hager, James G McNally, and Tom Misteli. "Transcription dynamics". In: *Molecular cell* 35.6 (2009), pp. 741–753.

[23] Hira Mubeen. "In Silico approach to identify transcription factor binding sites and Cis-regulatory elements in tubulin gene promoter". In: 2016.

[24] Muhammad A Zabidi and Alexander Stark. "Regulatory enhancer–core-promoter communication via transcription factors and cofactors". In: *Trends in Genetics* 32.12 (2016), pp. 801–814.

[25] Michael G Rosenfeld, Victoria V Lunyak, and Christopher K Glass. "Sensors and signals: a coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response". In: *Genes & development* 20.11 (2006), pp. 1405–1428.

[26] David Dilworth. *Understanding Enhancer Biology in Cancer*. https://openlabnotebooks.org/project-overview-understanding-enhancer-biology-in-cancer/. [Online; accessed 20-January-2023]. 2023.

[27] Ryuichiro Nakato and Toyonori Sakata. "Methods for ChIP-seq analysis: a practical workflow and advanced applications". In: *Methods* 187 (2021), pp. 44–53.

[28] Martha L Bulyk. "Computational prediction of transcription-factor binding site locations". In: *Genome biology* 5.1 (2003), pp. 1–11.

[29] Yulia A Medvedeva et al. "Effects of cytosine methylation on transcription factor binding sites". In: *BMC genomics* 15.1 (2014), pp. 1–12.

[30] Dvir Aran, Sivan Sabato, and Asaf Hellman. "DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes". In: *Genome biology* 14.3 (2013), pp. 1–14.

[31] Itika Arora and Trygve O Tollefsbol. "Computational methods and next-generation sequencing approaches to analyze epigenetics data: profiling of methods and applications". In: *Methods* 187 (2021), pp. 92–103.

[32] Alfonso Tramontano et al. "Methylation of the suppressor gene p16INK4a: mechanism and consequences". In: *Biomolecules* 10.3 (2020), p. 446.

[33] Alicia N Schep et al. "chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data". In: *Nature methods* 14.10 (2017), pp. 975–978.

[34] Jeffrey M Granja et al. "ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis". In: *Nature genetics* 53.3 (2021), pp. 403–411.

[35] M Ryan Corces et al. "The chromatin accessibility landscape of primary human cancers". In: *Science* 362.6413 (2018), eaav1898.

[36] David Adams et al. "BLUEPRINT to decode the epigenetic signature written in blood". In: *Nature biotechnology* 30.3 (2012), pp. 224–226.

[37] Daniel R Zerbino et al. "The ensembl regulatory build". In: *Genome biology* 16.1 (2015), pp. 1–8.

[38] Hendrik G Stunnenberg et al. "The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery". In: *Cell* 167.5 (2016), pp. 1145–1149.

[39] Yassen Assenov et al. "Comprehensive analysis of DNA methylation data with RnBeads". In: *Nature methods* 11.11 (2014), pp. 1138–1140.

[40] Fabian Müller et al. "RnBeads 2.0: comprehensive analysis of DNA methylation data". In: *Genome biology* 20.1 (2019), pp. 1–12.

[41] Oriol Fornes et al. "JASPAR 2020: update of the open-access database of transcription factor binding profiles". In: *Nucleic acids research* 48.D1 (2020), pp. D87–D92.

[42] Alicia Schep. *motifmatchr: Fast Motif Matching in R. R package version 1.20.0.* 2022. URL: https://bioconductor.org/packages/release/bioc/html/motifmatchr.html.

[43] Michael Lawrence et al. "Software for computing and annotating genomic ranges". In: *PLoS computational biology* 9.8 (2013), e1003118.

[44] Randle Aaron M Villanueva and Zhuo Job Chen. *ggplot2: elegant graphics for data analysis*. 2019.

[45] Hadley Wickham et al. *dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr. 2022.

[46] Patrick E McKnight and Julius Najab. "Mann-Whitney U Test". In: *The Corsini encyclopedia of psychology* (2010), pp. 1–1.

[47] Patrick E McKight and Julius Najab. "Kruskal-wallis test". In: *The corsini encyclopedia of psychology* (2010), pp. 1–1.

[48] Daniel Yekutieli and Yoav Benjamini. "Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics". In: *Journal of Statistical Planning and Inference* 82.1-2 (1999), pp. 171–196.

[49] Silke Wagner and Dorothea Wagner. *Comparing clusterings: an overview*. Universität Karlsruhe, Fakultät für Informatik Karlsruhe, 2007.