

---

國立成功大學資訊工程學研究所  
碩士論文

可視化轉錄因子結合位點與甲基化資訊之  
實現

**Computational pipeline for visualizing  
transcription factor binding site with  
methylation information**

研究生：郭昱伶

**Student: Yu-Ling Guo**

指導老師：賀保羅

**Advisor: Paul Horton**

National Cheng Kung University,

Tainan, Taiwan, R.O.C.

Thesis for Master of Science Degree

July, 2023

中華民國122年7月

---

# 可視化轉錄因子結合位點與甲基化資訊之 實現

郭昱伶\* 賀保羅†

國立成功大學資訊工程學研究所

## 摘要

MethylSeqLogo是一種涵蓋DNA甲基化資訊的序列標誌，由Fei-Man和Paul2位學者在去年提出來，像這樣的可視化圖表呈現，用意是讓我們可以透過轉錄因子結合位來量化轉錄因子在細胞內的功能以及基因調控的情形，然而MethylSeqLogo無法讓研究人員輸入自己實驗後的結果檔案來產生序列標誌，因此我研究的目標就是要為MethylSeqLogo建構一個自動化的計算管道，另外本研究也為背景模型提出了一個新的選項，命名為“Flanking 區域”，如此一來可以確保統計差異是歸因於結合位點本身或側翼區鹼基。本研究的實驗結果(視覺化圖表呈現)皆與多位學者提出之研究報告結果皆一致相符，說明自動化計算管道建立的方法是正確的且自動化也加速了結果的產生。透過本研究建立的自動化的計算管道能更讓研究者方便且更有效率和清楚的觀察細胞內的基因調控情況，進而用於學術研究的相關分析與交流，甚至是做疾病的預防或是治療。

**關鍵詞：**甲基化序列標誌、轉錄因子、DNA 甲基化

---

\*學生

†指導教授

---

# **Computational pipeline for visualizing transcription factor binding site with methylation information**

Yu-Ling Guo\* Paul Horton†

Institute of Medical Informatics, National Cheng Kung University

## **Abstract**

MethylSeqLogo, a type of sequence logo that encompasses DNA methylation information, was introduced by scholars Fei-Man and Paul last year. The purpose of this visualization is to quantify the functionality of transcription factors and the state of gene regulation within cells based on transcription factor binding sites. However, MethylSeqLogo does not allow researchers to input their own experimental raw data for generating sequence logos. Therefore, the goal of my research is to establish an automated computational pipeline for MethylSeqLogo. Additionally, this study proposes a new option for the background model called "Flanking Region," which ensures that statistical differences are attributed to the binding sites themselves or adjacent nucleotides. The experimental results of this study, presented in visualized charts, are consistent with previous research reports from multiple scholars, confirming the accuracy of the automated pipeline and the accelerated generation of results. The establishment of this automated computational pipeline through my research enables researchers to conveniently, efficiently, and clearly observe gene regulation within cells. It can be utilized for academic research, related analysis, and communication, and even for disease prevention and treatment.

---

\*Student

†Advisor

---

**Keywords:** MethylSeqLogo, Transcription Factor, DNA methylation

---

## 誌謝

首先我想在這裡感謝我的指導教授賀保羅老師。對於大學期間沒有涉略生物相關基礎知識的我來說，透過老師的幫忙讓我了解基因序列分析常用的技巧與困難，也才衍生出進行這份研究的動機。在討論的過程中，也很感謝老師給予的指教，讓我在面對困難時更快速的找到解決問題的方向。本論文的完成同時還要感謝原論文 EAGLE 的主要作者之一 Tony Kuo，除了和我分享當時研究時的寶貴經驗，還提供了當初實驗的測試資料，使我們在本論文中能更客觀地針對結果進行比較與分析。最後，也要感謝這兩年來在實驗室一起奮鬥過的所有同學們，面對大家普遍陌生的生物資訊領域，彼此教學相長，透過不斷的交流討論更快速的習得各方面的專業知識。

# CONTENTS

<b>中文摘要</b>	<b>i</b>
<b>Abstract</b>	<b>ii</b>
<b>誌謝</b>	<b>iv</b>
<b>Contents</b>	<b>v</b>
<b>List of Tables</b>	<b>vii</b>
<b>List of Figures</b>	<b>viii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Motivation & Research objective . . . . .	3
<b>2 Related Work</b>	<b>5</b>
2.1 MethylSeqLogo . . . . .	5
2.2 pybedtools . . . . .	7
2.3 Database . . . . .	8
2.3.1 JASPAR . . . . .	8
2.3.2 ReMap . . . . .	9
2.3.3 ENCODE . . . . .	10
<b>3 Methods</b>	<b>12</b>

3.1	Pipeline Overview . . . . .	12
3.2	automatic computational workflow . . . . .	13
3.2.1	Finding TFBSS . . . . .	13
3.2.2	Data preprocessing . . . . .	14
3.2.3	Calculate the methylation probability of the background model . . . . .	18
3.2.4	About Flanking region . . . . .	18
3.2.5	The execution time and memory usage of the automated computational pipeline . . . . .	19
<b>4</b>	<b>Results</b>	<b>22</b>
4.1	MYC in H1-hESC cell line . . . . .	22
4.2	CEBPB in H1-hESC cell line . . . . .	23
<b>5</b>	<b>Discussion &amp; Future Work</b>	<b>25</b>
5.1	Discussion . . . . .	25
5.2	Future Work . . . . .	27
<b>6</b>	<b>Conclusion</b>	<b>29</b>
	<b>Bibliography</b>	<b>30</b>

# List of Tables

2.1	Example of bedMethyl file from ENCODE . . . . .	11
3.1	Example of dataframe with seqdata . . . . .	15
3.2	Example of dataframe with methylation condition . . . . .	16
3.3	Example of dataframe with unmethylation read counts record .	17
3.4	Example of dataframe with methylation read counts record . .	17
3.5	Detailed information about the WGBS files from the ENCODE	19

# List of Figures

2.1	Design of MethylSeqLogo(Figure is from [14],without changes)	7
2.2	interesct method from BEDTools(Figure is from [17],without changes) . . . . .	8
2.3	Sequence logo of MYC(Figure is from [23],without changes) .	9
2.4	schema data of ReMap(Figure is from [24],without changes) .	10
2.5	The principle of WGBS . . . . .	11
3.1	Pipeline overview . . . . .	13
3.2	Illustration of Finding Transcription Factor Binding Sites in Cell 14	
3.3	Workflow of automatic computational pipeline . . . . .	21
4.1	Seqlogo of MYC in H1-hESC cell in whole-genome background model with Kullback-Liebler . . . . .	23
4.2	Seqlogo of CEBPB in H1-hESC cell in flanking region back- ground model with Kullback-Liebler . . . . .	24
5.1	Sequence logo of different versions of CEBPB in JASPAR (with- out no change) . . . . .	26
5.2	MethylSeqLogo of CEBPB transcription factor in H1-hESC cell with different background models . . . . .	26
5.3	Comparative analysis of the number of MethylSeqLogos gen- erated in whole-gnome background within the same timeframe	27

# Nomenclature

## General

bg                      Background

## Gene/Protein      Names

bp                      Base Pair

DNA                    Deoxyribonucleic Acid

H1-hESC               Human Embryonic Stem Cells

meC                    Methylcytosine

TF                      Transcription Factor

TFBS                   Transcription Factor Binding Site

# **Chapter 1**

## **Introduction**

### **1.1 Background**

During the process of biological growth, which includes cell growth, development, differentiation, apoptosis, and death, meticulous orchestration is required. It is akin to a production line where each step and pipeline of manufacturing and production influences the quality of the final product. Transcription factors (TFs) play a crucial role as key players in the growth of living organisms. TFs are proteins that regulate gene transcription, and upon interaction with DNA, they can control cellular logic by inhibiting or activating gene expression. This is why, despite all cells utilizing the same set of genes, there can be a wide variety of cell types (e.g., muscle cells, skin cells) with distinct functions. This diversity arises because only specific genes can be recognized and transcribed by TFs in different organs and tissues [1].

DNA methylation refers to the addition of a methyl group to the fifth carbon atom of cytosine, resulting in the formation of methyl cytosine (meC). It is often classified as the fifth DNA base [2] [3] and is known to cause gene silencing and the abnormal functioning of genes, leading to inhibitory effects [4]. DNA methylation is an epigenetic modification that can alter cellular gene function without changing the DNA sequence. Its effects can be inherited by the next generation. Numerous studies have shown a strong association between DNA

methylation and cancer [5] [6].

A dinucleotide sequence may be considered too short and lack biological interest. However, CpG dinucleotides are an exception. In most vertebrate genomes, CpG dinucleotides are generally rare, occurring at a frequency of approximately one-fifth [7]. This is because they are prone to methylation by DNA methyltransferases. However, the situation is different in promoter regions. In many promoter regions, there is often a high density of consecutive CpG dinucleotides, forming gene sequences known as CpG islands. CpG islands are characterized by a relatively low or nearly absent level of methylation compared to the rest of the genome. This implies that genes within these regions are typically active and capable of transcription [8].

Given that transcription is a complex and crucial process, it is of great importance to enable researchers to analyze and compare data efficiently and clearly. One early successful method in the field of molecular biology was the invention of "sequence logos" [9]. Sequence logos provide a visual representation of transcription factor binding site (TFBS) preferences, allowing researchers to understand the binding preferences of transcription factors. A sequence logo consists of a stack of letters at each position (e.g., A, C, G, T for DNA sequences), where the height of each letter stack is proportional to the information content of the corresponding nucleotide distribution at that position [9]. The widespread application of this method has demonstrated its utility in summarizing and visualizing binding sites, facilitating comparisons and communication of transcription factor binding preferences among researchers. In fact, sequence logo methods have been expanded in various ways, such as increasing the resolution of compositional enrichment/depletion (e.g., Seq2logo [10] and EDlogo [11]), displaying higher-order sequence motifs [12], or demonstrating relationships between binding site positions [13].

Sequence logos help biologists understand the sequence preferences of transcription factors, but they cannot explain cell-type-specific selection of TFBS (transcription factor binding sites). Therefore, additional information that af-

fects transcription, such as DNA methylation and CpG dinucleotides, is required to gain a more comprehensive understanding of TF (transcription factor) function and quantify gene regulation within cells. Here, we introduce MethylSeqLogo [14], a software tool that display sequence logo with DNA methylation and CpG depletion as epigenetic information. Unfortunately, MethylSeqLogo only provides a limited set of pre-processed example files and does not allow researchers to input their own raw experimental data. Therefore, based on this idea, this study aims to provide computational pipelines for MethylSeqLogo, enabling researchers to utilize it extensively.

## 1.2 Motivation & Research objective

The binding of transcription factors to DNA is a crucial mechanism for controlling gene expression in cells. For instance, humans possess approximately 1600~1700 different transcription factors, and through dynamic interactions with DNA, these transcription factors enable precise regulation of cellular processes such as growth, development, and response to the environment. Therefore, visualizing the binding sites of transcription factors to quantify gene regulation within cells has consistently been an important and necessary topic in life science research [15] [16].

In a nutshell, this study explores the development of an automated computational pipeline for MethylSeqLogo [14] using tools and methods such as BEDTools [17], Pandas [18], seaborn [19], and others. The motivation behind this work stems from the observation that MethylSeqLogo currently only provides a limited set of pre-processed example files and lacks the functionality to accept researchers' own experimental raw data to generate results(sequence logo).

In addition to providing an automated computational pipeline, we propose a novel background model called the "flanking region." The background model is utilized to assess statistical differences associated with a set of binding sites,

indicating the amount of relative entropy [20] we can extract from the logos that produced by software .MethylSeqLogo currently provides the "whole-genome" and a set of predefined "promoter" regions as background models. However, this study further focuses on the establishment of background regions specific to individual transcription factor binding sites, thereby tailoring customized background models for each transcription factor. The approach involves considering a defined range (e.g., 100 bp) of regions surrounding each binding site. Ideally, this process is performed independently for each binding site, resulting in the inclusion of  $x$  binding sites near genomic positions  $x$  times in the background model statistics. This ensures that the statistical differences depicted in the MethylSeqLogo visualization primarily arise from the binding sites themselves or their adjacent bases, rather than being influenced by large-scale trends in methylation and/or CpG frequency across the whole-genome.

Here, we aims to enable biologists to input their own experimental raw data, and then generate sequence logos with methylation information and CpG depletion. Moreover, We provide a novel background model option called "flanking regions." The objective is to facilitate efficient and comprehensive observation of gene regulation within cells, including the extent of methylation, thereby providing valuable insights for academic research, analysis, and scientific communication. Furthermore, this research has potential implications in disease prevention and treatment.

# Chapter 2

## Related Work

In this chapter, we will first give an overview of MethylSeqLogo [14], and then introduce pybedtools [21] and the data required for this study to visualize seqlogo of transcription factor binding sites and methylation information, which are supported and provided by various databases. The following subsections provide a detailed explanation.

### 2.1 MethylSeqLogo

MethylSeqLogo [14] is an innovative tool known as DNA methylation smart sequence logos. It was proposed by Fei-Man Hsu and Paul Horton in 2022 as a method to incorporate DNA methylation information into sequence representation, providing insights into transcription factor functionality and gene regulation in cellular contexts.

Building upon traditional sequence logos, MethylSeqLogo intuitively represents the sequence conservation and preferences at each position in the DNA sequence. However, it expands this representation by incorporating DNA methylation information and highlighting CpG dinucleotide depletion in additional tracks. This enables researchers to gain a comprehensive understanding of the interplay between DNA methylation, transcription factor functionality, and gene regulation.

A detailed description of MethylSeqLogo shown in 2.1. The column height of each track in MethylSeqLogo is determined by calculating the relative entropy (unit:bits), which represents the statistical distribution difference between a set of binding site sequences and the background model. Relative entropy, also known as the Kullback-Leibler directed divergence [20], is equivalent to information content [22] when a uniform distribution background is used. The formula from MethylSeqLogo shown in equation 2.1:

$$D(M||B) \stackrel{\text{def}}{=} E \left[ \lg \left( \frac{P[s|\text{Motif model M}]}{P[s|\text{Background model B}]} \right) \right] \quad (2.1)$$

Here, an example illustrating how to calculate the relative entropy from methylation. Suppose we want to calculate the relative entropy of mCG in a set of transcription factor binding sites (TFBSs) at position 1 in the whole-genome background (plus strand). First, we need to calculate the probability of CG methylation level at the first position in all binding sites( $P(^mC|CG)$ ), and the same calculation applies to the background model(whole-genome). Furthermore, it is necessary to incorporate the information content of un-methylation, as it completes the full event. Lastly, since there are multiple binding sites, we also need to calculate the probability of the expected occurrence of CG dimer the first position across all binding sites ( $P(C|CG)$ ).The complete calculation process is as follows:

$$\begin{aligned} \text{Entropy} = & P_1(C|CG) * (P_1(^mC|CG) * \lg(P_1(^mC|CG)/P_{bg}(^mC|CG))) + \\ & ((1 - P_1(^mC|CG)) * \lg((1 - P_1(^mC|CG))/(1 - P_{bg}(^mC|CG)))) \end{aligned}$$

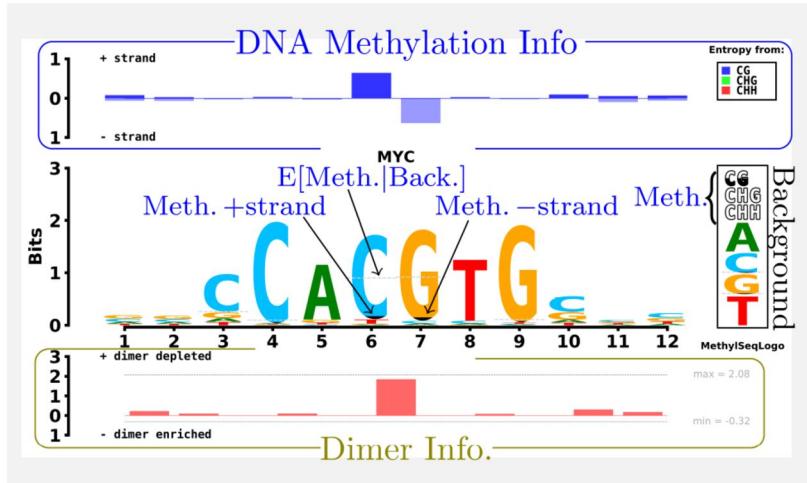


Figure 2.1: Design of MethylSeqLogo(figure is from [14],without changes)

In summary, MethylSeqLogo provides a powerful and intuitive visualization tool for studying the functional role of DNA methylation in gene regulation. It is a valuable tool for researchers in the field of epigenetics and genomics, aiding in the identification of potential regulatory regions, examining the impact of DNA methylation on transcription factor binding, and gaining deeper insights into the complex gene regulatory processes occurring within cells.

## 2.2 pybedtools

pybedtools [21] is a powerful Python library used for manipulating and analyzing genomic interval data. It serves as an interface to the BEDTools [17] suite, offering a wide range of functions to handle genomic intervals, including operations such as intersection, merging, and complementation between genomes. By leveraging the underlying BEDTools tools, pybedtools optimizes performance and memory usage, making it well-suited for efficient processing of large-scale genomic data. It supports various file formats, such as BED, BAM, VCF, and GTF, enabling easy reading and writing of data in different formats. Additionally, it provides advanced features like computing overlaps, calculating coverage statistics, and performing set operations. One significant advantage is its seamless integration with popular libraries like pandas and numpy, facilitating

data analysis and calculations. Hence, pybedtools was chosen as a valuable tool for studying transcription factor binding sites in this research. Overall, pybedtools is a versatile and efficient tool for working with genomic interval data, offering an intuitive interface, extensive functionality, and integration with other Python libraries. The functions utilized in this study include "intersect" and "getfasta".

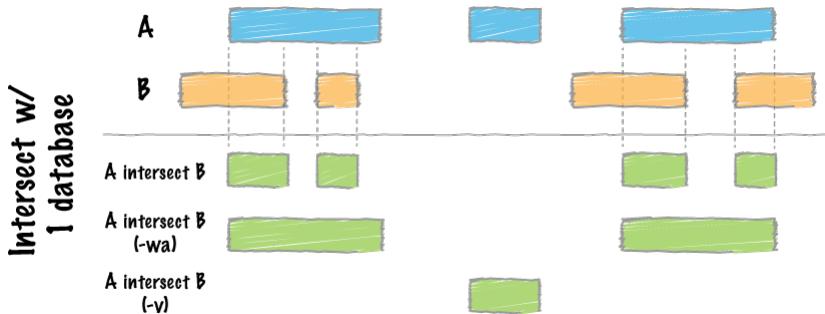


Figure 2.2: intersect method from BEDTools(figure is from [17],without changes)

The figure shown in 2.2 represent two BED files, labeled as A and B. In this study, BED file A represents the data obtained from JASPAR [23], while BED file B corresponds to the data obtained from ReMap [24].

## 2.3 Database

### 2.3.1 JASPAR

Transcription factor binding sites, also known as motifs, refer to specific DNA sequences where transcription factors bind and initiate transcription. Typically, the length of these sequences ranges from 5 to 20 nucleotides. JASPAR [23] is a freely accessible database that provides information on transcription factor binding sites. It serves as a valuable resource for conducting transcription factor-related analyses. JASPAR collects and analyzes transcription factor binding sites from six major species classifications, including Fungi, Insecta, Nematoda, Plantae, Urochordata, and Vertebrata. The database offers not only frequency

matrices for transcription factor binding site locations but also provides additional data in the form of BED files, FASTA files, external links (such as ReMap and DRV), and more for researchers to download and utilize. In this study, we utilized the BED files from JASPAR, which provide the coordinates of transcription factor binding sites for each transcription factor across various cells or tissues. Furthermore, it is worth noting that in JASPAR the Matrix ID format typically follows the pattern of MAxxxx.x (e.g., MA0147.3) , where the number after decimal point represents the version of the datasets. A higher numerical value indicates a more recent version.

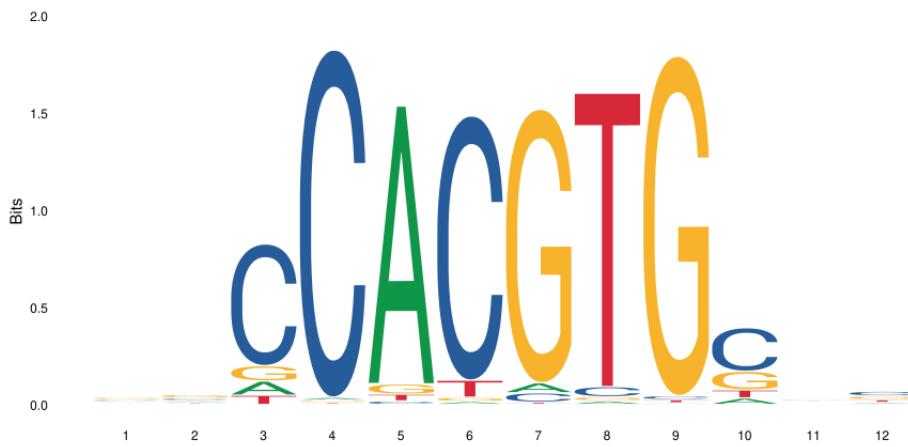


Figure 2.3: Sequence logo of MYC(Figure is from [23],without changes)

### 2.3.2 ReMap

If we are interested in identifying and predicting transcription factor binding sites in different cells or tissues, we can perform ChIP-seq experiments. In these experiments, we obtain numerous peaks, with each peak representing a potential transcription factor binding site within a specific region. Fortunately, ReMap [24] has organized this information for us. ReMap is a large-scale database of transcriptional regulatory peaks, which collects and analyzes data from ChIP-seq, ChIP-exo, and DAP-seq experiments. It covers various species, including humans, mice, fruit flies, and Arabidopsis, and provides analysis for a total of 1210 transcription factors. ReMap offers the data in both BED file

and FASTA file formats, allowing users to download experiment-related information. For this study, we utilized the BED files containing ChIP-seq data from ReMap.

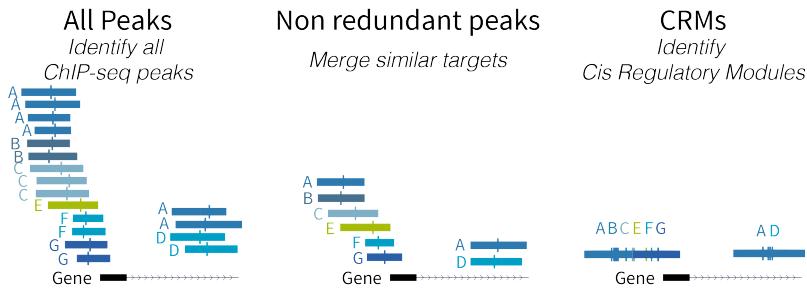


Figure 2.4: schema data of ReMap(Figure is from [24],without changes)

### 2.3.3 ENCODE

ENCODE [25] (Encyclopedia of DNA Elements) is a research project initiated by the National Human Genome Research Institute (NHGRI) in the United States in 2003. Its aim is to establish a comprehensive catalog of functional elements on the human genome, including DNA hypersensitive sites, DNA methylation, and transcriptional regulatory regions, among other experimental aspects. ENCODE researchers employ various assays and methods to identify functional elements, and all generated data is made available for download through their website for user analysis. In this study, DNA methylation data from ENCODE was utilized. Through sodium bisulfite sequencing, the binding of bisulfite to Cytosine (C) bases can be examined, leading to deamination. After the experimental process, unmethylated Cytosine bases are converted to Uracil (U) bases, while methylated Cytosine bases remain as Cytosine. This enables the acquisition of methylation information from the sequence, which is recorded in BED file format, but is distinct from JASPAR and ReMap, not only includes the basic sequence coordinate information but also records the read count and methylation level, thus referred to as bedMethyl, and shown in Table 2.1.

chromosome	start coordinates	end coordinates	...	read coverage	methylation level(%)
chr1	869	870	...	8	50
chr1	1450	1451	...	2	0
chr2	520	521	...	4	38
chr2	14755	14756	...	5	90
chr3	15449	15450	...	8	0

Table 2.1: Example of bedMethyl file from ENCODE

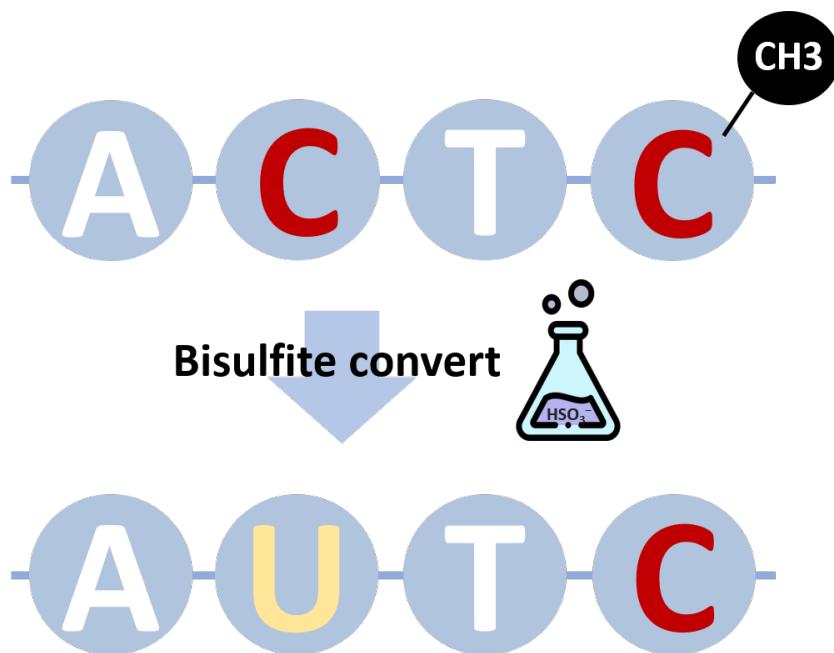


Figure 2.5: The principle of WGBS

# **Chapter 3**

## **Methods**

In this chapter, the framework and workflow of my research methodology will be presented and discussed.

### **3.1 Pipeline Overview**

Pipeline overview shown in 3.1,The user is required to input files obtained from various experiments or databases. Files from ReMap and JASPAR contain relevant information about transcription factors and their binding sites, while files from ENCODE contain information about cellular methylation levels. Next, we will search for all binding site information of transcription factors in the cells and define our background model. We then calculate the methylation levels and the probabilities of base occurrence for each of the three methylation scenarios. Subsequently, we employ Relative Entropy to assess the statistical differences between the binding site distribution and the background model. Finally, we visualize the entropy on a chart as the final output. In the following subsection, I will provide a more detailed and concrete explanation of the research methodology using the example of observing the gene regulation of the transcription factor MYC in human embryonic stem cells(H1-hESC).

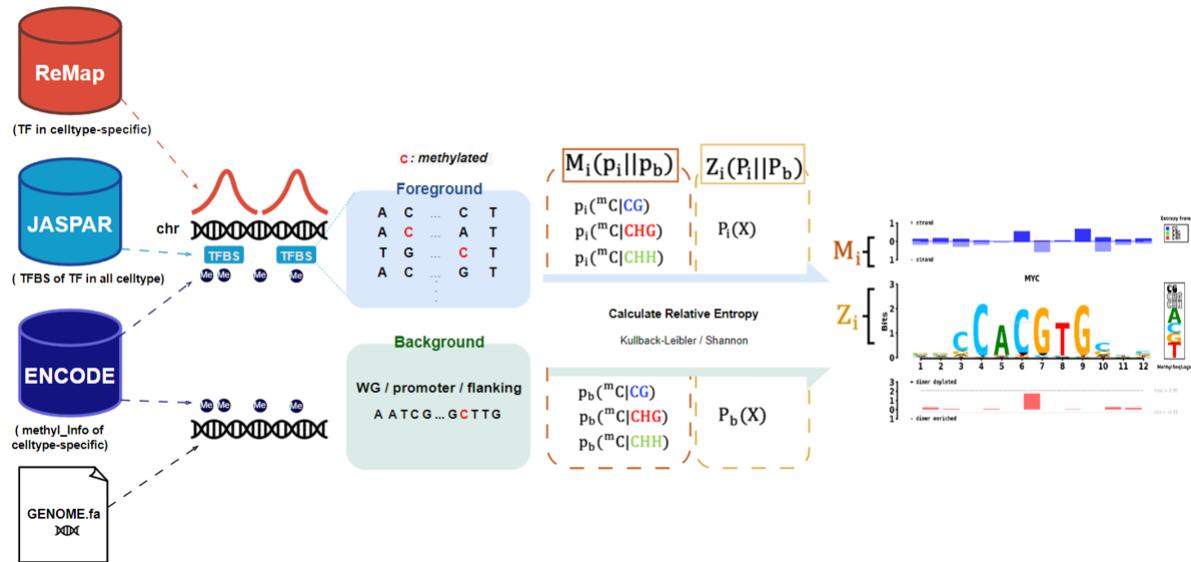


Figure 3.1: Pipeline overview

## 3.2 Detail workflow of automatic computational pipeline

### 3.2.1 Identifying the binding site information of TF in cell.

In this step, two input files are required. The first file contains the regions obtained from the Chip-seq experiment, which potentially contain the binding sites of MYC in human embryonic stem cells. These regions are represented as fuzzy genomic coordinates and typically have a range of 200-500 base pairs. This data can be obtained from ReMap, and the specific file I used is "ENCSR000EBY.MYC.WA01.bed". The second file can be obtained from JASPAR and contains the precise binding site coordinates of the MYC transcription factor across various cells or tissues. The file I used is "MA0147.3.bed". Both files obtained from ReMap and JASPAR are in the .bed file format. We can then intersect the genomic coordinates from these two files to find the binding site information of MYC in human embryonic stem cells(H1-hESC). In the automated computational pipeline, this study used the "intersect" function provided by the pybedtools library in BEDTools to perform this operation. A total of 2088 transcription factor binding sites of MYC were found. It is important

to note that both files must be mapped to the same human reference genome. Genomic coordinates generated based on different human reference genomes can vary significantly. The commonly used human reference genomes are hg19 and hg38, with hg38 being more prevalent. This is because hg38 has improved genome coverage compared to hg19, resulting in more accurate positioning and annotation of gene sequences. Furthermore, some structural errors in chromosomes have been corrected in hg38. Fortunately, we can quickly convert the genomic coordinates between hg19 and hg38 using the "Lift Genome Annotations" website provided by UCSC.

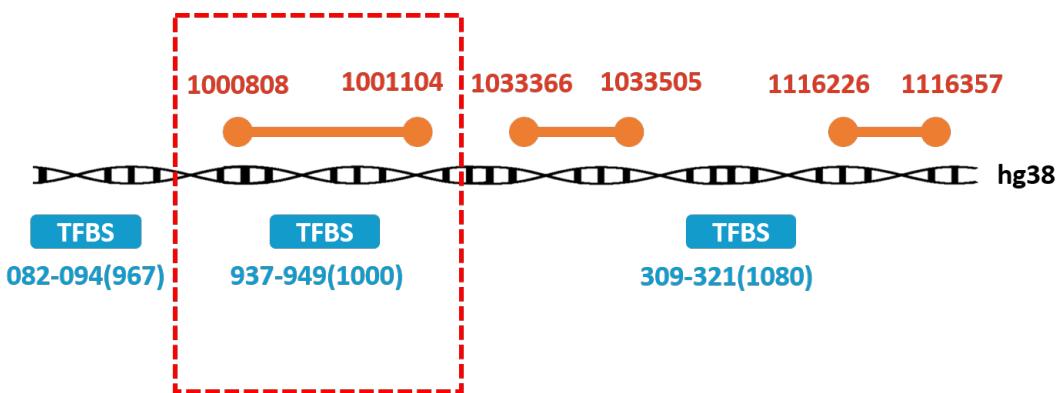


Figure 3.2: Illustration of Finding Transcription Factor Binding Sites in Cell

In 3.2, the orange and blue colors represent data from ReMap and JASPAR, respectively. The numbers indicate the coordinates of the gene sequence. The interpretation of "082-094(967)" is that the gene sequence spans from coordinate 967082 to coordinate 967097. From the figure, we can see that the red dashed box represents the intersection, indicating that the gene fragment from coordinate 1000937 to 1000949 is the binding site of MYC in human embryonic stem cells(H1-hESC).

### 3.2.2 Data preprocessing

Once the binding sites are identified, we need to convert the coordinate data into sequence data. In our automated computational pipeline, we also utilize pybedtools to accomplish this task. The specific command used is "tfbs.sequence(fi=fasta)".

In this command, "tfbs" refers to the .bed file containing the binding sites of MYC in human embryonic cells that we just found, and "fasta" refers to the human reference genome file(version: hg38), it's a FASTA format file, which can be downloaded from the Illumina iGenomes website.

	1	2	3	4	5	6	7	8	9	10	11	12
chr1:1000935-1000951(+)	C	G	C	C	A	C	G	T	G	C	T	T
chr1:1232506-1232522(-)	A	G	C	C	A	C	G	T	G	C	T	G
...	..	..	..	..	..	..	..	..	..	..	..	..
chrX:154763520-154763536(-)	C	A	C	C	A	C	G	T	G	G	C	T
chrX:154763547-154763563(+)	C	G	A	C	A	C	G	T	G	C	A	G

Table 3.1: Example of dataframe with seqdata

Once we obtain the sequence data, we can calculate the probability of occurrence for each { A, C, G, T } nucleotide at each position within a set of binding sites using the ppm calculation. By comparing this probability distribution to the background statistical distribution and calculating the Relative Entropy, we can generate the sequence logo, which represents the transcription factor's binding site preference information displayed in the central track. It is worth mentioning that in our automated computational pipeline, we use the sequence data to construct a dataframe that records the methylation condition for C/G base, shown in 3.2. This dataframe is used for convenient calculation of methylation levels in subsequent analyses.

	1	2	3	4	5	6	7	8	9	10	11	12
0	X	x	Z	Z	-	X	x	-	z	Z	-	-
1	-	X	x	y	-	X	x	-	X	x	-	X
...	..	..	..	..	..	..	..	..	..	..	..	..
2062	z	z	z	z	-	z	Z	-	Y	X	x	z
2088	X	x	-	Z	-	X	x	-	z	Y	-	y

Notes: X: CG, Y: CHG, Z: CHH(Uppercase and lowercase letters represent the plus and minus strands, respectively)

Table 3.2: Example of dataframe with methylation condition

Regarding the methylation of TFBSS, the combination of {CG, CHG, CHH} is sufficient to describe the DNA methylation condition. Here, H represents any nucleotide except for G in plus strand. However, during the WGBS experiment, some experimental errors or bias may occur. We downloaded WGBS data from the ENCODE website for the H1-hESC cell line. The ENCODE IDs for the downloaded data are ENCFF601NBW, ENCFF918PML, ENCFF524BMX, ENCFF379ZXG, ENCFF086MMC, and ENCFF417VRB. These files represent methylation at CG, CHG, and CHH contexts, with two files for each context. To enhance the reliability of methylation level estimation.

To calculate the methylation of TFBSS, we utilized the intersect function in pybedtools to intersect the WGBS BED file with the BED file containing MYC binding sites in the H1-hESC cell line (generated in Section 3.2.1). The process is similar to the TFBS identification. After that, we can obtain the methylation levels of TFBSS.

Here, we will explain how the study handles the two replicates WGBS files for each methylation context. The calculation formula used to merge two replicates files as shown in 3.1. In the automated computational pipeline, the study organizes the six WGBS files into two dataframe tables. These tables record the counts of methylated reads and unmethylated reads, respectively. Table 3.1 serves as the main table, while Tables 3.2 and 3.3 provide supplementary

information. By referencing these tables, the methylation levels and probabilities of { CG, CHG, CHH } at each position within a set of binding sites can be calculated using a table lookup approach. The relative positions of each dataframe are interconnected, enabling the calculation based on their respective locations.

$$mlevel = \frac{\text{methylated reads}_A + \text{methylated reads}_B}{(\text{methylated reads}_A + \text{unmethylated reads}_A) + (\text{methylated reads}_B + \text{unmethylated reads}_B)} \quad (3.1)$$

	1	2	3	4	5	6	7	8	9	10	11	12
0	13.0	10.0	13.0	13.0	NaN	13.0	11.0	NaN	11.0	11.0	NaN	NaN
1	NaN	0.0	24.0	14.0	NaN	13.0	23.0	NaN	24.0	24.0	NaN	0.0
...	..	..	..	..	..	..	..	..	..	..	..	..
2062	0.0	20.0	30.0	30.0	NaN	30.0	28.0	NaN	28.0	28.0	0	28.0
2088	22.0	13.0	NaN	23.0	NaN	23.0	13.0	NaN	13.0	23.0	NaN	13.0

Notes 1: The number represents sum of unmethylated read counts from two replicate WGBS files for each {CG, CHG, CHH} condition.(NaN:the base in that field is A/T (adenine/thymine))

Table 3.3: Example of dataframe with unmethylation read counts record

	1	2	3	4	5	6	7	8	9	10	11	12
0	0.0	0.0	0.0	0.0	NaN	0.0	0.0	NaN	0.0	0.0	NaN	NaN
1	NaN	0.0	0.0	0.0	NaN	0.0	0.0	NaN	0.0	0.0	NaN	0.0
...	..	..	..	..	..	..	..	..	..	..	..	..
2062	0.0	0.0	0.0	0.0	NaN	0.0	0.0	NaN	0.0	0.0	0.0	0.0
2088	1.0	0.0	NaN	0.0	NaN	0.0	0.0	NaN	0.0	0.0	NaN	0.0

Notes 1: The number represents sum of methylated read counts from two replicate WGBS files for each {CG, CHG, CHH} condition.(NaN:the base in that field is A/T (adenine/thymine))

Table 3.4: Example of dataframe with methylation read counts record

### 3.2.3 Calculate the methylation probability of the background model

In the previous subsection, we discussed the methylation probabilities of binding sites. Therefore, here we will explain the calculation of methylation probabilities in the background model. For this, WGBS (Whole-Genome Bisulfite Sequencing) data from the H1-hESC cell line are required, as described in subsubsection 3.2.2. The approach involves reading two WGBS files with the same methylation status, and they must be sorted. Since the failure rate of WGBS experiments is very low, the content of the files obtained from the same cells will have highly similar or even identical coordinates. The differences lie in the read coverage and methylation ratio, which may vary slightly due to experimental variations. Based on this idea, we can merge the methylation levels of the same coordinates in the two files (the pseudocode for the condition should include an OR statement). The same approach applies to different background models, with the only difference being the size of the data used. Various background models are filtered using the pybedtool intersect function.

### 3.2.4 About Flanking region

The new background model option proposed in this study, called the "Flanking region," is designed to tailor a background model specifically for a set of binding sites. The approach involves extending the starting and ending genomic coordinates of a set of binding sites by a certain range on both sides, for example, 100 base pairs. The total length of the range is equal to the length of the binding site itself plus twice the length of the extension range (due to both the left and right sides). This ensures that the statistical differences described in MethylSeqLogo can be attributed to the binding site or its neighboring flanking nucleotides, rather than large-scale trends. Large-scale trends can sometimes amplify effects and lead to confusion in the interpretation of the information we are interested in.

### 3.2.5 The execution time and memory usage of the automated computational pipeline

In the automated computational pipeline established in this study, in addition to addressing the lack of automation in the original MethylSeqLogo and providing a new background model option, we have also made improvements to address two additional issues. The first issue pertains to insufficient memory space. As shown in table 3.5, the six WGBS files from ENCODE have a combined size of approximately 100GB for CHH methylation data. When calculating the methylation probabilities for the background model, it would require occupying 80The second issue relates to the excessively long execution time due to the large file sizes, which is the second aspect of improvement. To address this problem, I employed the multiprocessing package to implement parallel computing, thereby accelerating the calculation of methylation probabilities.

Methyl-condition	Celltype	File capacity(GB)	original execution time(min)	after implementing parallel(min)
ENcff417VRB	CHH	50.92	144	27
ENcff086MMC		51.3		
ENcff524BMX	CHG	14.69	44	8
ENcff379ZXG	H1-hESC	14.57		
ENcff601NBW	CpG	3.44	6	2
ENcff918PML		3.47		

Table 3.5: Detailed information about the WGBS files from the ENCODE

In conclusion, this chapter have provided detailed explanations of the methods and procedures for finding binding sites, a serial of data preprocessing, and various probability calculations(base frequency and methylation), and the workflow shown in 3.3. It is important to note that since DNA is double-stranded, the methylation probability calculation should consider the information from both the forward and reverse strands. With the calculated probabilities, we can now proceed to calculate the relative entropy between a set of binding sites and the background model, enabling the generation of MethylSeqLogos chart. The calculation method and graphical presentation of relative entropy follow the

specifications and design of the original MethylSeqLogo.

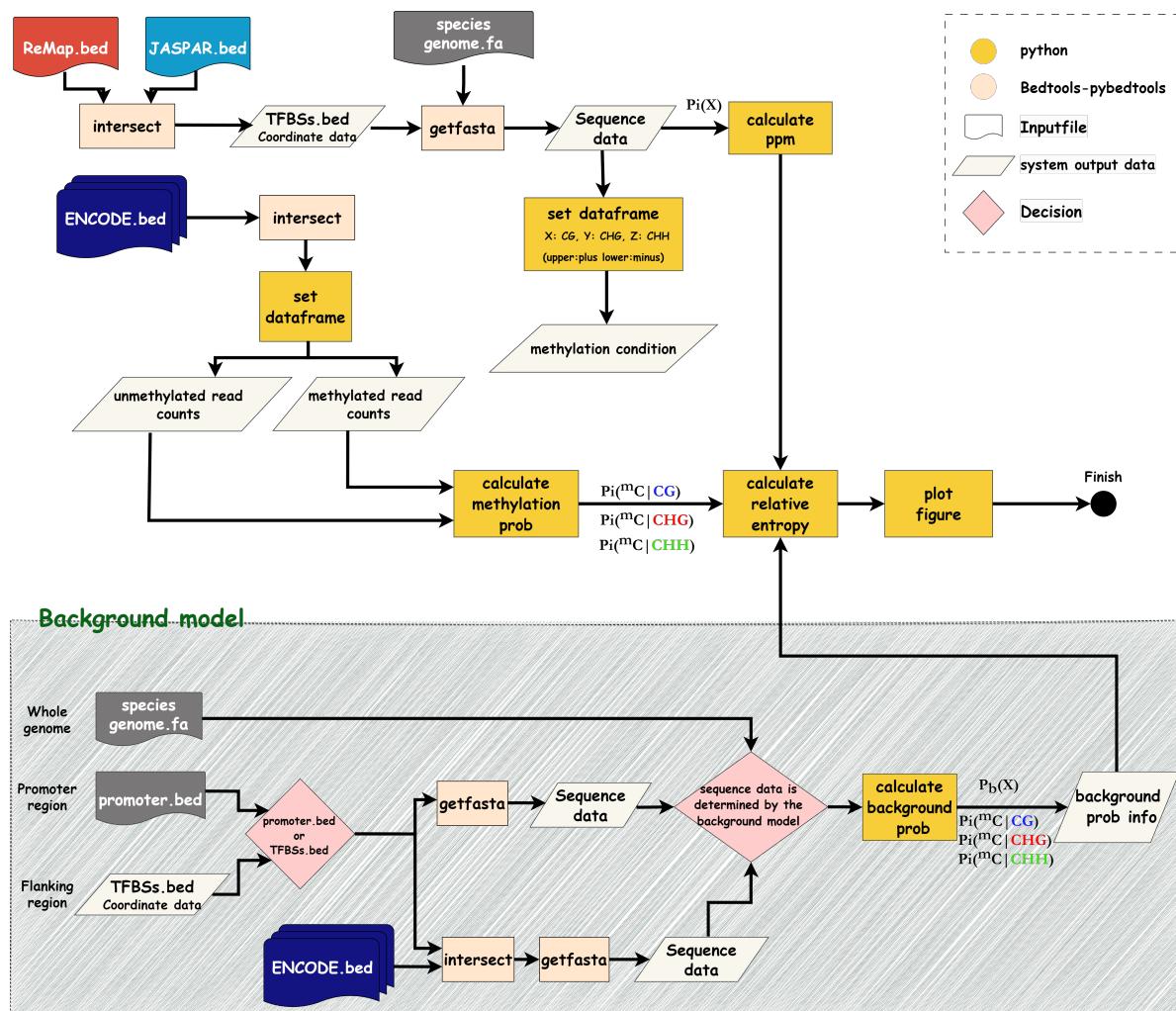


Figure 3.3: Workflow of automatic computational pipeline

# Chapter 4

## Results

In this chapter, we will discuss whether the content mentioned in the literature or reports regarding transcription factors aligns with the visual presentation of our study's sequence logo. Tests were executed on a Linux server with a single AMD RyzenTM 9 5950X desktop processor, consisting of 16 hyper-threaded physical cores running at their base clock 3.4GHz, and 128GB of random access memory.

### 4.1 MYC in H1-hESC cell line

The transcription factor MYC belongs to the helix-loop-helix family and possesses characteristic structural features of a basic region and a leucine zipper. It is an important transcription factor that plays a crucial role in processes such as cell growth, metabolism, angiogenesis, and immune response. It is considered a significant oncogene, as its overexpression has been associated with abnormal cell proliferation in many tumors. According to the research conducted by Michael Allevato et al., the preferred sequence for the binding site of the MYC transcription factor is CANNTG, with the CACGTG sequence being the most highly favored and preferred [26]. The visualization in 4.1 demonstrates consistency with the reported findings.

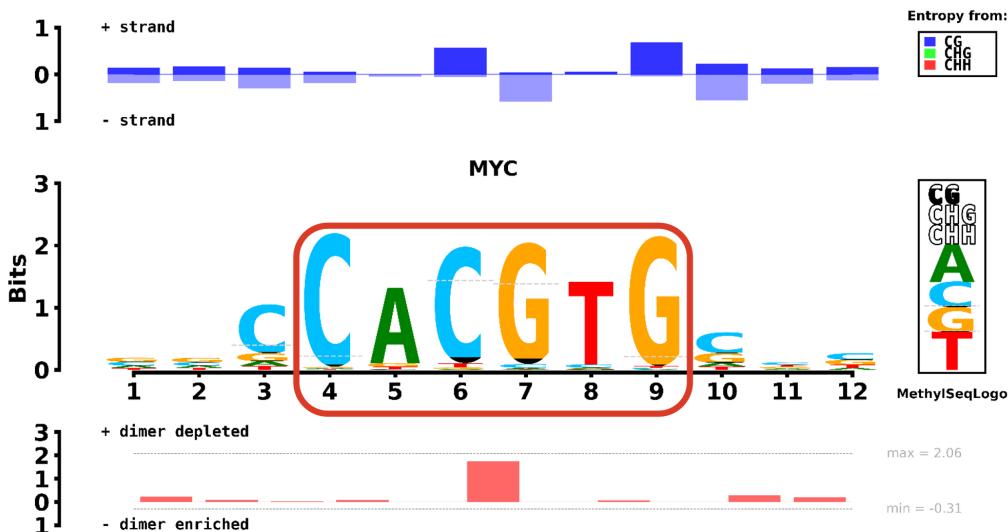


Figure 4.1: Seqlogo of MYC in H1-hESC cell in whole-genome background model with Kullback-Liebler

## 4.2 CEBPB in H1-hESC cell line

CEBPB is a transcription factor belonging to the C/EBP (CCAAT/enhancer-binding protein) family. It is widely expressed in mammalian cells, particularly at higher levels in tissues such as the liver, spleen, kidney, and bone marrow. The CEBPB protein contains a basic region that enables it to bind to specific DNA sequences. Additionally, CEBPB transcription factor includes a leucine zipper structure following the basic region, which is a helix-loop-helix motif facilitating protein-protein interactions to form dimers or multimeric complexes. CEBPB is involved in regulating the production of inflammatory factors, playing a crucial role in immune response and inflammation processes. Furthermore, CEBPB is closely associated with cell proliferation and differentiation, regulating the expression of specific genes in various cell types, thus influencing cell differentiation status and function. In adipocytes, CEBPB is considered one of the key transcription factors involved in adipocyte differentiation. In the study conducted by Ximei Luo et al., they mentioned that CEBPB binding sites were located in highly methylated regions in H1-hESC cell [27]. The dotted lines represent the degree of background methylation, while the black

shading indicates the level of methylation at the binding sites. In Figure 4.2, the methylation condition of CG at positions 6 and 7 in motif is prominently displayed, with a substantial area of black shading exceeding the dotted line. The visual representation of the data in the figure aligns with the findings of the study.

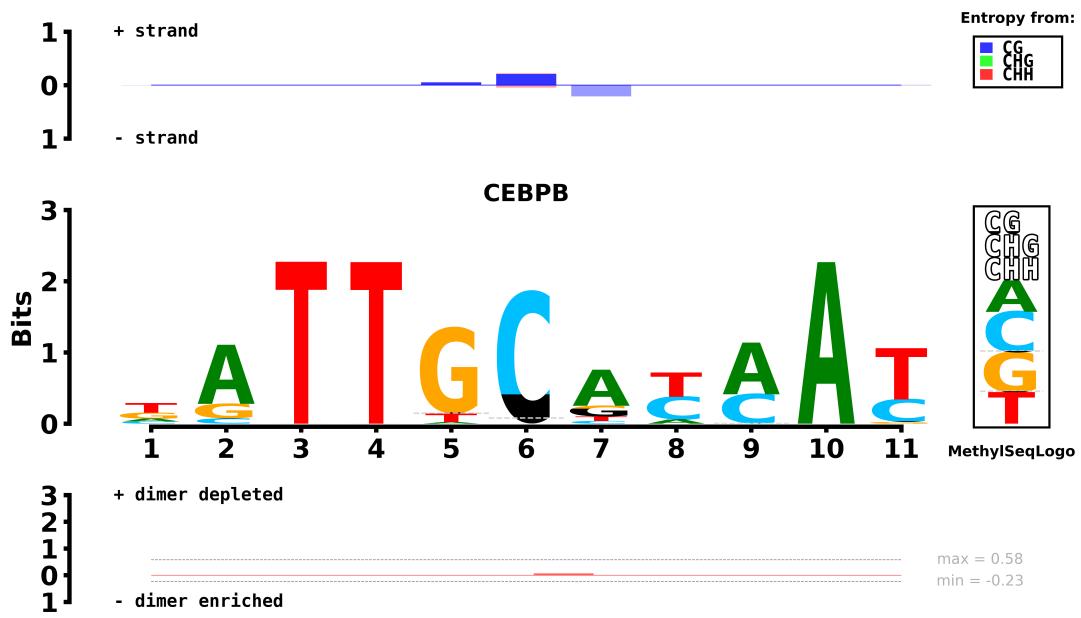


Figure 4.2: Seqlogo of CEBPB in H1-hESC cell in flanking region background model with Kullback-Liebler

# **Chapter 5**

## **Discussion & Future Work**

### **5.1 Discussion**

In this study, the input data utilized a BED file containing specific binding site coordinates of transcription factors in all cells or tissues, which was curated by JASPAR. For each included transcription factor, JASPAR always provides a position weight matrix (PWM) but may not necessarily offer a corresponding BED file. Additionally, there may be multiple versions of the data (as detailed in the section 2.3.1), so special attention is required. In the case of the CEBPB transcription factor, this study encountered different versions of the data. The current version available is the 3rd version, which only provides a PWM matrix. The BED file, on the other hand, is only available in version 1. However, the data in versions 1 and 3 differ significantly(shown in Figure 5.1), leading to different results in the sequence logo analysis. In a research report, it was shown that the preferred binding sites of the CEBPB transcription factor is ATTGCGCAAT [28]. Figure 4.2 presents the results obtained using version 1 (Figure 5.1a) of the BED file. The common bases at positions 7 and 8 in motifs on the figure differ from the research findings of scholars Peter F.Johnso. At first glance, this might lead us to believe that there is an error in the computational pipeline established in this study. However, upon observing the Sequence logo provided by JASPAR in version 3, it becomes evident that the disparities

arise from differences in the content of the JASPAR versions. Details like these are worth paying close attention to.

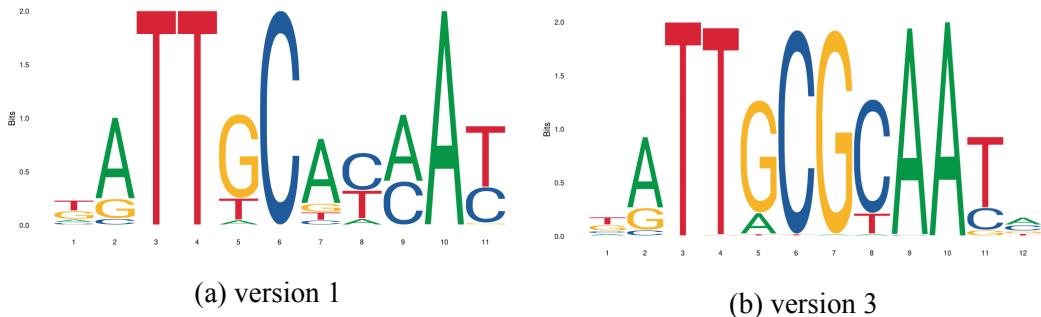


Figure 5.1: Sequence logo of different versions of CEBPB in JASPAR (without no change)

In Figure 4.2, we used the flanking region background model to illustrate the high methylation status of the binding sites of the CEBPB transcription factor in human embryonic stem cells. Now, we have switched to using the whole-genome as the background model for statistical distribution calculations. The results, as shown in Figure 5.2b, reveal that the visualization of high methylation status under the whole-genome background model is not as prominent as when using the flanking region background model. Therefore, choosing an appropriate background model is crucial depending on the purpose of our observation. In this case of examining the high methylation status of CEBPB binding sites in human embryonic stem cells, the flanking region would be a preferable choice.

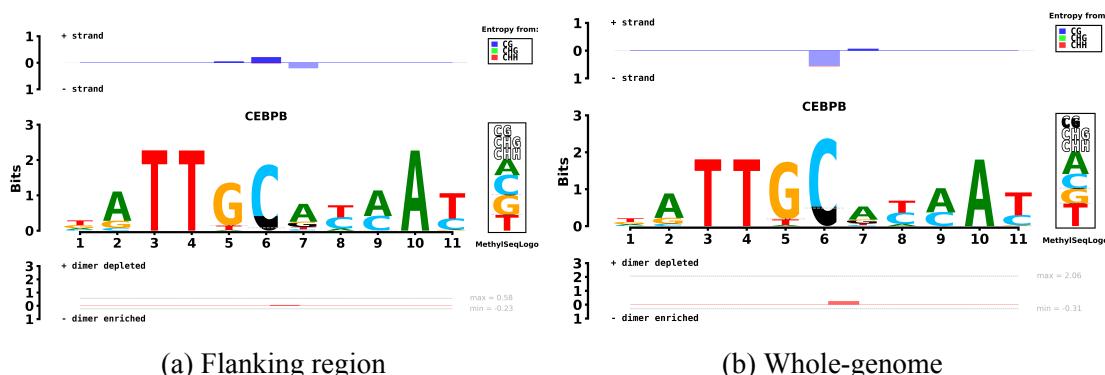


Figure 5.2: MethylSeqLogo of CEBPB transcription factor in H1-hESC cell with different background models

Lastly, I would like to discuss the establishment of an automated computational pipeline in this study. The design of automation not only brings conven-

nience but, more importantly, allows for the rapid generation of results. Prior to the implementation of the automated computational pipeline, it was estimated that it would take approximately 3-4 hours to obtain a single MethylSeqLogo depicting the binding of MYC to human embryonic stem cells using the whole-genome background model. However, with the implementation of the automated computational pipeline, this study can now generate approximately 10 MethylSeqLogos depicting the binding of various transcription factors to human embryonic stem cells in the whole-genome background model within the same timeframe.

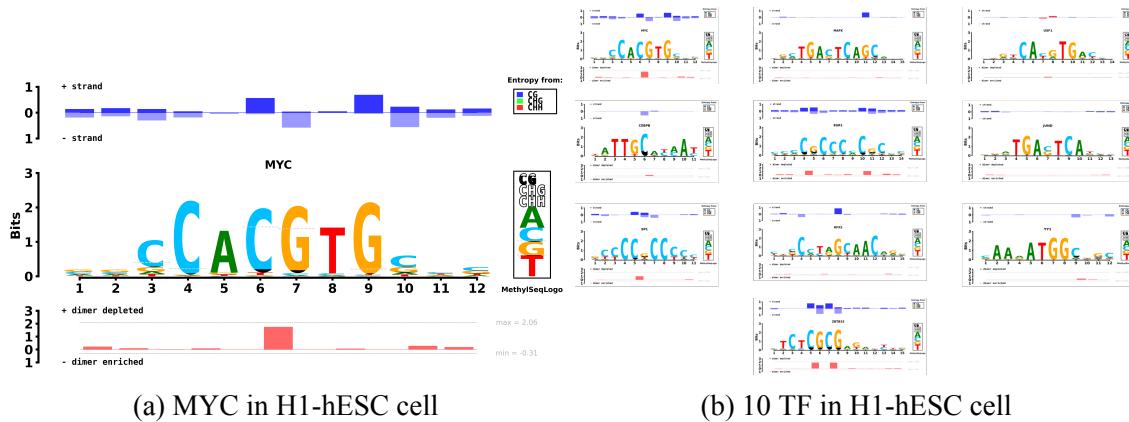


Figure 5.3: Comparative analysis of the number of MethylSeqLogos generated in whole-genome background within the same timeframe

## 5.2 Future Work

The accuracy of the input data is a crucial factor influencing MethylSeqLogo. To ensure accurate and reliable visualization, it is important to obtain the latest version of the BED file from JASPAR or similar curated datasets containing transcription factor binding site coordinates. By utilizing the automated pipeline developed in this study, more comprehensive and accurate visual representations can be generated. Furthermore, in the future, it is possible to incorporate additional information onto the sequence logo, such as the information on 5-hydroxymethyl cytosine and other epigenetic modifications (e.g., histone modifications) as proposed by the author of MethylSeqLogo. This approach would

enable a multifaceted understanding of transcription factor functionality.

# **Chapter 6**

## **Conclusion**

In this paper, we establish an automated computational pipeline for MethylSeqLogo, which not only provides convenience but also accelerates the generation of results. While previously only one MethylSeqLogo result could be obtained, this study is now capable of producing 10 MethylSeqLogos within the same timeframe. Additionally, a novel background model option, namely the "flanking region," has been introduced to ensure that statistical differences are attributed to binding sites or, at most, adjacent flanking nucleotides.

The experimental results of this study, as presented in the visualized charts, are consistent with findings reported by multiple researchers. In future studies, researchers utilizing the automated computational pipeline established in this study can efficiently and clearly observe gene regulation within cells, such as the extent of methylation occurrence and binding site preferences. This can contribute to academic research, related analyses, and discussions, and even aid in disease prevention and treatment.

# Bibliography

- [1] Samuel A Lambert et al. “The human transcription factors”. *Cell* 172.4 (2018), pp. 650–665.
- [2] Ecker JR Lister R. “Finding the fifth base: genome-wide sequencing of cytosine methylation”. *Genome Res* 19.6 (2009), pp. 959–66.
- [3] Walker N et al Viner C Johnson J. “Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet”. *bioRxiv* (2016).
- [4] P. A. Jones. “Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Reviews Genetics”. *Nat Rev Genet* 13.7 (2012), pp. 484–492.
- [5] Keith D Robertson. “DNA methylation and human disease”. *Nat Rev Genet* 6 (2005), pp. 597–610.
- [6] Vogelstein B Feinberg A. P. “Hypomethylation distinguishes genes of some human cancers from their normal counterparts”. *Nature* 301 (1983), pp. 89–92.
- [7] Serge Saxonov, Paul Berg, and Douglas Brutlag. “A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters”. *Proceedings of the National Academy of Sciences of the United States of America* 103 (Feb. 2006), pp. 1412–7. DOI: 10.1073/pnas.0510310103.
- [8] Bird A Deaton AM. “CpG islands and the regulation of transcription”. *Genes Dev* 25.10 (Dec. 2011), pp. 1010–22. DOI: 10.1101/gad.2037511.
- [9] Thomas D Schneider and R Michael Stephens. “Sequence logos: a new way to display consensus sequences”. *Nucleic acids research* 18.20 (1990), pp. 6097–6100.

- [10] Nielsen M Thomsen MC. “Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion”. *Nucleic Acids Research* 40 (May 2012).
- [11] Stephens M Dey KK Xie D. “A new sequence logo plot to highlight enrichment and depletion”. *BMC bioinformatics* 19.473 (2018).
- [12] András Micsonai et al. “BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra”. *Nucleic Acids Research* 46.W1 (June 2018), W315–W322.
- [13] Söding J Siebert M. “Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences”. *Nucleic Acids Res* 44.13 (2016), pp. 6055–69.
- [14] Paul Horton Fei-Man Hsu. “MethylSeqLogo: DNA methylation smart sequence logos”. *bioRxiv* (2022). DOI: 10.1101/2022.11.05.515271. eprint: <https://www.biorxiv.org/content/early/2022/11/05/2022.11.05.515271.full.pdf>. URL: <https://www.biorxiv.org/content/early/2022/11/05/2022.11.05.515271>.
- [15] Juan M Vaquerizas et al. “A census of human transcription factors: function, expression and evolution”. *Nature Reviews Genetics* 10.4 (2009), pp. 252–263.
- [16] Yin Liu et al. “Transcriptional landscape of the human cell cycle”. *Proceedings of the National Academy of Sciences* 114.13 (2017), pp. 3473–3478.
- [17] Aaron R. Quinlan and Ira M. Hall. “BEDTools: a flexible suite of utilities for comparing genomic features”. *Bioinformatics* 26.6 (Jan. 2010), pp. 841–842.
- [18] Wes McKinney et al. “Data structures for statistical computing in python”. *Proceedings of the 9th Python in Science Conference*. Vol. 445. Austin, TX. 2010, pp. 51–56.

- [19] Michael L. Waskom. “seaborn: statistical data visualization”. *Journal of Open Source Software* 6.60 (2021), p. 3021. DOI: 10.21105/joss.03021. URL: <https://doi.org/10.21105/joss.03021>.
- [20] Solomon Kullback and Richard A Leibler. “On information and sufficiency”. *The annals of mathematical statistics* 22.1 (1951), pp. 79–86.
- [21] Ryan K. Dale, Brent S. Pedersen, and Aaron R. Quinlan. “Pybedtools: a flexible Python library for manipulating genomic datasets and annotations”. *Bioinformatics* 27.24 (Sept. 2011), pp. 3423–3424. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr539. eprint: [https://academic.oup.com/bioinformatics/article-pdf/27/24/3423/48862135/bioinformatics\\_27\\_24\\_3423.pdf](https://academic.oup.com/bioinformatics/article-pdf/27/24/3423/48862135/bioinformatics_27_24_3423.pdf). URL: <https://doi.org/10.1093/bioinformatics/btr539>.
- [22] Thomas D Schneider et al. “Information content of binding sites on nucleotide sequences”. *Journal of molecular biology* 188.3 (1986), pp. 415–431.
- [23] Jaime A Castro-Mondragon et al. “JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles”. *Nucleic Acids Research* 50.D1 (Nov. 2021), pp. D165–D173. ISSN: 0305-1048. DOI: 10.1093/nar/gkab1113. eprint: <https://academic.oup.com/nar/article-pdf/50/D1/D165/42058061/gkab1113.pdf>. URL: <https://doi.org/10.1093/nar/gkab1113>.
- [24] Pierre De Langen Fayrouz Hammal, Fabrice Lopez Aurélie Bergon, and Benoit Ballester. “ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments”. *Nucleic Acids Research* 50 (Jan. 2022).
- [25] ENCODE Project Consortium et al. “An integrated encyclopedia of DNA elements in the human genome”. *Nature* 489.7414 (2012), p. 57.
- [26] Michael Allevato et al. “Sequence-specific DNA binding by MYC/MAX to low-affinity non-E-box motifs”. *PloS one* 12.7 (2017), e0180147.
- [27] Ximei Luo et al. “Effects of DNA methylation on TFs in human embryonic stem cells”. *Frontiers in Genetics* 12 (2021), p. 639461.

- [28] PETER F Johnson. “Identification of C/EBP basic region residues involved in DNA sequence recognition and half-site spacing preference”. *Molecular and cellular biology* 13.11 (1993), pp. 6919–6930.