COMPREHENSIVE DATA ANALYSIS TOOLKIT DEVELOPMENT FOR LOW INPUT BISULFITE SEQUENCING

A Dissertation

by

YUE YIN

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Yun Huang
Co-Chair,	Roderick Dashwood
Committee Members,	Donald Williams Parsons
	Zhongming Zhao
Head of Program,	Carol Vargas Bautista

August 2021

Major Subject: Medical Sciences

ABSTRACT

The human cell-free DNA (cfDNA) methylation profile in liquid biopsy has been utilized to diagnose early-stage disease and estimate therapy response. However, typical clinical procedures are capable of purifying only very small amounts of cfDNA. Whole-genome bisulfite sequencing (WGBS) is the gold standard for measuring DNA methylation; however, WGBS with small amounts of fragmented DNA introduces a critical challenge for data processing, analysis, and visualization. For data processing, the low mapping ratio of low input bisulfite sequencing samples resulting in genome-wide low sequencing depth and low coverage of CpG sites is a bottleneck for the clinical application of cfDNA-based WGBS assays. We developed LiBis (Low-input Bisulfite Sequencing), a novel augmentation for low-input WGBS data alignment. By dynamically clipping initially unmapped reads and remapping clipped fragments, we judiciously rescued those reads and uniquely aligned them to the genome. By substantially increasing the mapping ratio by up to 88%, LiBis dramatically improved the number of informative CpG sites and the precision in quantifying the methylation status of individual CpG sites. The high sensitivity and cost-effectiveness afforded by LiBis for low-input samples will help the discovery of genetic and epigenetic features suitable for downstream analysis and biomarker identification using liquid biopsy. For data analysis, we present Mmint, a user-friendly comprehensive integrative analysis tool. It generates publication-quality figures with epigenetic data from the following aspects: quality assessment, integrative analysis between BS-Seq and ChIP-seq data, correlation analysis between DNA methylation/Histone modification, and gene expression. Versatile analysis by Mmint can help users to interpret epigenetic data comprehensively and provide potential novel biological insights. To further simplify the data utilization and visualization, especially for researchers who do not specialize in bioinformatics skills, we implement GsmPlot. GsmPlot can simply accept GSM IDs to automatically download NCBI data or accept user's local bigwig files as input to plot the data of interest on promoters, exons or any other user-defined genome locations and generate UCSC visualization tracks. By linking public data repository and in-house data, GsmPlot can spark data-driven ideas and hence promote epigenetic research.

DEDICATION

To my family.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Yun Huang, my committee co-chair, Dr. Roderick H. Dashwood, and my committee members, Dr. Donald Williams Parsons and Dr. Zhongming Zhao, for their guidance and support throughout the course of this research. Thanks also goes to Dr. Deqiang Sun, for his guidance and support.

I am also grateful to all of my colleagues, Dr. Jia Li, Dr. Mutian Zhang, Dr. Jin Li, Dr. Jianfang Li, Dr. Sat byul Seo and Dr. Minjung Lee, for valuable discussion and input about my research.

Finally, and the most importantly, I would never have been able to finish this dissertation without the love of my family. And I would give a special thanks to my parents and my grandparents, for their support and suggestions to my life and career.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professor Roderick H. Dashwood [advisor], Professor Yun Huang of Center for Epigenetics & Disease Prevention, Professor Donald Williams Parsons of Texas Children's Cancer and Hematology Center and Professor Zhongming Zhao of School of Biomedical Informatics in UTHealth.

The sample collection, library construction and sequencing experiments in section 2 are performed by Texas Children's Hospital and Dr. Minjung Lee of Center for Epigenetics & Disease Prevention.

The case studies in section 3 and section 4 were finished under collaboration with Professor Jia Li of Center for Epigenetics & Disease Prevention.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by award RP180131 from the Cancer Prevention and Research Institute of Texas and by start-up funds from Texas A&M University.

NOMENCLATURE

LiBis	Low-input bisulfite sequencing alignment
cfDNA	Cell-free DNA
Bis-seq	Bisulfite sequencing
WGBS	Whole-genome bisulfite sequencing
scWGBS	Single cell whole genome bisulfite sequencing
CpG	Cytosine nucleotide followed by a guanine nucleotide
ctDNA	Circulating tumor DNA
HTML	Hyper Text Markup Language
CSF	Cerebrospinal fluid
HPV	Human papilloma virus
ATAC-Seq	Assay for Transposase-Accessible Chromatin using sequencing
Mmint	Methylation data mining tools
MOABS	model based analysis of bisulfite sequencing
CGI	CpG Island
ChIP-Seq	Chromatin immunoprecipitation sequencing
GEO	Gene Expression Omnibus
GSM	Gene Sample accessions numbers
H3K27ac	Acetylation at the 27th lysine residue of the histone H3 protein
H3K4me3	Tri-methylation at the 4th lysine residue of the histone H3 protein
H3K4me1	Mono-methylation of lysine 4 on histone H3 protein

H3K27me3	Tri-methylation of lysine 27 on histone H3 protein
DNase I	Deoxyribonuclease I
5hmC	5-Hydroxymethylcytosine
NCBI	National Center for Biotechnology Information
RNA-Seq	RNA Sequencing
TF	Transcription Factor
TSS	Transcriptional Start Site

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

1.1. Roles of DNA methylation in cancer genesis and diagnosis

DNA methylation is an important biological process involved in biological development and cancer. It includes the addition of a methyl group to the carbon 5 of the cytosine pyrimidine ring or the nitrogen 6 of the adenine purine ring. Among different types of methylation, the most common one on the mammalian genome is observed on the cytosine of CpG dinucleotides [1]. In the past few decades, a series of researches indicate a direct relationship between CpG methylation and cancer [2]. In various malignant tumors, hypomethylation is observed and the DNMT3b mutations are found in patients which cause chromatin instability [3, 4]. It is speculated that hypomethylation can promote tumorigenesis by activating oncogenes (such as cMYC and H-RAS) or by activating potential retrotransposons [2]. Hypomethylation of the long interspersed nuclear elements also leads to transcriptional activation and is found in many types of cancers (such as bladder cancer) [2, 5]. Such hypomethylation of mobile DNA may also cause the interruption of the expression of neighboring genes during cancer geneses, such as APC in colon cancer and cMYC in breast cancer [6].

Increased DNA methylation level is also reported among human tumors in previous researches. Chromosomes 3p, 11p, and 17p are found as hot spots of hypermethylation on the human genome [7]. The CpG islands in these regions are unmethylated in healthy tissues while getting methylated in disease models [8, 9]. For example, transient transfection experiments show that the retinoblastoma tumor suppressor gene can reduce its expression to 8% of the unmethylated control by applying hypermethylation in its promoter region [10]. The evidence indicate that essential genes such as tumor suppressor genes in normal tissues can be deactivated by hypermethylation in tumors [11, 12]. The inactivation of tumor suppressor genes by hypermethylation can further lead to selective growth of transformed cells [13, 14].

Based on the strong relationship between CpG methylation and tumor genesis, a series of researches are published in the past few years about tumor-specific methylation patterns. A pan-cancer research indicates that tumor-specific methylation patterns can be used to distinguish 10 different cancer types such as breast cancer, lung adenocarcinoma, and colon adenocarcinoma [15]. In the central nervous system, tumor-specific methylation patterns can be observed both among different cancer types and different subtypes of the same cancer type [16]. Moreover, DNA methylation patterns show highly consistency between cfDNA and the genomic DNA from the corresponding origins among tumors, which indicates that cfDNA in liquid biopsies samples such as urine, saliva, blood and cerebrospinal fluid can be used to develop potential clinical applications for cancer diagnosis, prognosis and therapeutics [17].

Facilitated by the evidence, products and algorithms are developed for early cancer diagnosis and prognosis monitoring. For example, Epi proColon is developed to screen colorectal cancer by evaluating the methylation status of the SEPT9 promoter using patient plasma [18]. By applying different algorithms to the methylation patterns in solid tumors and the adjacent liquid biopsy, tumor detecting methods using liquid biopsy for different cancer types are also developed. For example, the generalized linear model is applied to detect renal cell carcinoma [19].

1.2. Whole genome bisulfite sequencing methods for DNA methylome detection

Bisulfite sequencing (BS-Seq) of DNA is considered as the gold standard for analysis of DNA methylation since developed [20]. The treatment of DNA by bisulfite results in the conversion of unmethylated cytosine to uracil while keeping methylated cytosine unchanged [21]. To reveal the methylome conveniently and economically in single-base resolution, bisulfite treatment has been combined with next-generation sequencing to generate reduced representation (RRBS) or whole genome (WGBS) data [20, 22]. After PCR and sequencing, the converted DNA can be mapped to the reference genome as single-base resolution. During the development of next-generation sequencing techniques, the cost keeps decreasing, which allows more researchers to adopt WGBS to investigate genome wide methylome. To adapt to different circumstances of samples such as solid tumor samples, liquid biopsies, and single-cell samples, different WGBS library preparation protocols are developed [23, 24]. Currently, according to the order of bisulfite conversion, these methods are divided into pre-bisulfite sequencing methods and post-bisulfite sequencing methods.

The difference between pre-bisulfite sequencing and traditional next-generation sequencing is that pre-bisulfite sequencing ligates methylated adaptors after end repairing and applies bisulfite conversion after ligation. In 2009, MethylC-Seq, the most widely used pre-bisulfite sequencing method was developed by Ryan Lister [20]. Firstly, the extracted DNA is sonicated to 50-500 bp fragments, followed by end repairing and ligation of methylated adaptors. Bisulfite conversion is applied to the ligated DNA fragments before PCR and next-generation sequencing. However, the bisulfite conversion can damage the ligated DNA fragments and reduce the usability of the DNA fragments [25]. Because of that, pre-bisulfite sequencing method requires at least 1μg starting DNA for the library construction and sequencing [23].

Usage of pre-bisulfite sequencing is limited in cell-free DNA samples from liquid biopsy or single-cell samples since these samples have low DNA concentration and cannot meet basic requirement of pre-bisulfite sequencing method [26, 27]. To determine the methylome of these samples, researchers applies post-bisulfite sequencing to reduce the DNA damage during bisulfite treatment [23]. In post-bisulfite sequencing, bisulfite treatment is directly applied to extracted DNA. The sonication step in pre-bisulfite sequencing is removed since bisulfite treatment can break the DNA chain into short single-strand fragments. After bisulfite conversion, random primers containing any possible combinations of bases is used to catch fragments for adaptor ligation [23]. By applying the only fragmentation step at the beginning of the workflow, post-bisulfite sequencing reduces the DNA cost and only needs less than 100 nanograms of starting DNA [23]. However, applying random priming to converted DNA fragments introduces contamination into the DNA library such as primer self-ligation and repeated primers. These sequences interferes the further data alignment and leads to low mapping ratio and low data efficiency in downstream analysis [25, 28].

1.3. Overview of analytic pipelines and toolkits for bisulfite sequencing data

After BS-Seq, raw data generated by BS-Seq is stored in the text file as FASTQ format, in which every 4 lines contain sequence identifier, raw sequence, and sequence quality value in addition to one raw sequence/reads. Under most circumstances, a sequencing run contains more than one sample and the raw FASTQ file includes reads from all samples. Therefore, to separate reads for each sample, a raw FASTQ file can be demultiplexing to one FASTQ file per sample by indexes [29]. After demultiplexing, the original FASTQ file is divided into multiple FASTQ files and each divided file contains all sequences for one sample. For data quality assessment, FastQC can be used to evaluate the FASTQ files from multiple dimensions such as adaptor content and sequence duplication [30]. TrimGalore and fastp are two widely used software for FASTQ file preprocessing such as adaptor removal and reads trimming when quality issues are identified [31].

After quality assessment, all sequences in FASTQ file need to be aligned to the reference genome to identify their origins. To accomplish the sequence alignment, a few tools are implemented based on different algorithms, which can be categorized into two main groups of aligners by their alignment strategies. The first one is to transform the reference genome into four bisulfite converted genomes (two for each strand), then align the sequence to these reference genomes to find the best match. This type of aligner is called three-letter aligners and some of them are widely used such as Bismark, BS-Seeker and bwa-meth [32-34]. Using a transformed reference genome reduces the engineering workload of implementation and allows researchers to reuse most of the aligner developed for whole genome sequencing. The other type, wildcard aligners directly align the sequence by wildcard to the reference genome. Tools such as BSMAP, GSNAP, and Last are adopting this strategy [35-37]. Compared to three-letter aligners, wildcard aligners have higher genome coverage since it increases the complexity of reference genome and reads. Such complexity also introduces biases in alignment results [38]. Besides general usage, scBS-map is developed especially for single-cell bisulfite sequencing alignment [39]. The output files generated by different BS-Seq alignment tools share the same format: Sequence Alignment/Map (SAM) format. SAM file can also be transferred to a binary format to save storage, which is the BAM format. Each row of the BAM file represents an aligned sequence/read and almost all features about the sequence, including alignment and quality information [40].

To construct features of methylome from aligned reads for downstream research, methylation information in multiple granularities is extracted. Compared to the unsupervised method, the supervised method is the most common method for feature selection which allows researchers to compare the methylation status between the control group and case groups. Among supervised methods, the strategies diverge into two main categories based on whether introducing the spatial relationship between CpGs within the same read. Differentially methylated cytosines and regions are the most widely used strategies without the spatial information, in which the methylation status of CpG is decided by the coverage and the number of methylated cytosine on overlaid reads. Based on statistical models, differentially methylated CpGs (DMCs) can be identified from the control and case group and be concatenated into differentially methylated regions (DMRs). Another feature selection strategy is to include the spatial information among CpGs within the same read. The continuous CpG array represents the co-existence of methylation in the same sequence, which also in the same strand in one of the cells. This high-level information gives researchers more potential to solve deconvolution problems and to identify disease markers.

Instead of only focusing on methylation data, multi-omics analysis such as genome-wide association studies (GWAS) or gene expression are commonly included in the epigenetic research to explain the confounding. These studies revealed more functionalities of methylation in different biological processes and triggered the progress of the integrative analysis toolkits such as deepTools and mint [41, 42].

1.4. Problem statement

Currently, WGBS is broadly applied to purified cfDNA samples searching for new candidate biomarkers. However, the sensitivity of existed or new biomarkers is strongly influenced by the sequencing coverage and the sequencing accuracy. Compared to whole genome sequencing (WGS), bisulfite sequencing brings in more digestion to the DNA fragments, which can introduce more damage and noises into the sequencing process. Low DNA concentration in cfDNA samples also leads to low starting DNA for sequencing, which is overcomed by post-bisulfite sequencing by moving bisulfite conversion to the beginning of the library preparation process and using random priming to ligate the adaptor. However, the downstream data efficiency, namely mapping ratio of post-bisulfite sequencing is reported significantly lower than pre-bisulfite sequencing in both public datasets and in-house generated datasets. Studies also reported that potential self-priming and repeat-priming by random primer reduces the mapping ratio of post-bisulfite sequencing samples. To solve this, we developed LiBis, an ultrasensitive alignment augmentation to remove the contamination of random primer *in silico* and to improve the data efficiency in post-bisulfite sequencing.

Furthermore, during the development of WGBS, the number of samples collected in labs, companies and public Gene Expression Omnibus keep surging. The bisulfite conversion step in WGBS brings new potential problems in sequencing and data analysis compared with WGS, which requires a specific understanding of next generation sequencing, WGBS library preparation and downstream data analysis. The steep learning curve of such knowledge makes the learning time-consumptive and increases the labor of locating and resolving the potential problems. The experience from a single WGBS experiment is also hard to be borrowed by other WGBS experimenters. Moreover, the current analysis pipelines for WGBS data mainly yield numeric and descriptive results, which increases the difficulty of downstream analysis performed by researchers who lack bioinformatic background. To simplify this process, we developed Mmint, a quality assessment and integrative analysis toolkit for WGBS data. For researchers who have limited access to high performance computational resources, we also developed GsmPlot to help them performing analyses on epigenetics datasets such as WGBS and ChIP-Seq.

1.5. Thesis organization

The rest of this dissertation is organized as follows. In section 2, we propose LiBis, a new augmentation method for low-input bisulfite sequencing alignment. LiBis is developed to solve the low mapping efficiency problem in utilizing low-input bisulfite sequencing data. In section 3, we present Mmint, a quality assessment and integrative analysis toolkit for users to measure the WGBS data quality and access the biology meaning conveniently. In section 4, we present GsmPlot, an online analysis tool for users without bioinformatics background to perform analysis on both published and in-house generated epigenetic datasets with no requirement of computing resource. By user cases in sections 3 and 4, we will further explain how Mmint and GsmPlot facilitate the analysis of epigenetic data. In section 5, we present the conclusions of our comprehensive toolkit for low-input bisulfite sequencing data processing, analysis, and visualization.

2. LIBIS: AN ULTRA SENSITIVE ALIGNMENT AUGMENTATION FOR LOW-INPUT BISULFITE SEQUENCING[*]

2.1. Background

DNA methylation abnormalities contribute to tumorigenesis and tumor prognosis [43]. Tumors are characterized with global hypomethylation and focal CpG island hypermethylation [44]. As a simple, economical assay for early cancer diagnosis or monitoring therapeutic response, liquid biopsy has rapidly emerged as an alternative to tumor biopsy due to its minimal invasiveness and ease of repeat sampling [45]. Many studies have reported that tumor DNA methylation status can be accurately detected in circulating cell-free DNA (cfDNA) from blood or other body fluids [46-48].

Compared to the limited number of CpGs detected using microarray technology, whole genome bisulfite sequencing (WGBS) can detect all the CpG sites in the genome, which substantially increases the power for biomarker discovery. However, the extremely low amount of cfDNA in liquid biopsies poses a challenge for whole genome bisulfite sequencing using cfDNA. The amount of cfDNA collected from the plasma of healthy individuals or patients usually ranges from nanograms to several dozen nanograms, and for late-stage cancer patients the amount is usually less than several hundred nanograms, which remains low for WGBS library preparation [49, 50]. To prepare WGBS libraries

[*] Reprinted with permission from “LiBis: an ultrasensitive alignment augmentation for low-input bisulfite sequencing” by Yue Yin; Jia Li, 2020. Briefings in Bioinformatics, bbaa332, Copyright [2020] by Oxford University Press.

using low amounts of DNA, several methods have been developed. Post bisulfite sequencing methods such as post-bisulfite adaptor tagging (PBAT) and single-cell WGBS methods apply adaptor tagging after bisulfite treatment to reduce the loss of tagged DNA fragments during library preparation [23]. Specifically, first, the bisulfite conversion is carried out directly on the cell lysate; second, random oligos are added to the 3’ end of single-stranded bisulfite converted DNA fragments, then adaptors are added to the DNA fragments through two rounds of random priming and extension. This increases the sensitivity of library preparation for low input WGBS. However, such Adaptase reactions can lead to the formation of long synthetic sequences at fragment 3’ ends [51, 52]. Sequencing reads containing variable lengths of synthetic sequence contamination limit the effectiveness of current fixed-length trimming methods [52]. Moreover, random priming may occur not only between the primer and genomic DNA fragments, but also between two primers or two genomic DNA fragments. Mistakenly concatenated DNA fragments may be sequenced as chimeric reads, which cannot be aligned to one correct location but can be separated into two fragments with each mapped precisely to a distinct genome location [28]. These issues of WGBS library preparation from low amounts of DNA result in low mapping ratios (average 40%), which increase the cost of liquid biopsy and restrict its application in clinical settings.

Several mapping programs such as BSMAP, Bismark, BS-Seeker, and scBS-map have been developed to address the above issues [53-55]. Although BS-Seeker utilizes soft-clipping during the mapping procedure and scBS-map adopts a local alignment strategy to improve the mapping ratios, the mapping ratios of samples applying low input

WGBS are still far lower than samples applying traditional WGBS [39]. To get enough data coverage for all CpGs from low-input bisulfite sequencing data, it is imperative to develop a mapping procedure to recover as much data as possible from the reads "unmapped" due to library preparation protocol drawbacks. Here, we developed a novel method, LiBis, to further improve the mapping ratio of low-input bisulfite sequencing data. LiBis applies a dynamic clipping strategy to rescue the discarded information from each unmapped read in end-to-end mapping.

In our simulation study, LiBis achieved the highest mapping ratio improvement with the shortest CPU time among published methods. LiBis also improved the number of detected CpGs and the methylation ratio accuracy in a time-efficient manner. By applying LiBis, we achieved a better cost efficiency using both a public dataset and in-house datasets. The number of informative CpGs increased significantly after using LiBis compared with using a traditional trimming protocol. The precision of bisulfite sequencing was also improved by LiBis for all samples. Furthermore, LiBis was able to identify virus insertion sites in a cervical cancer WGBS dataset, which indicates that bisulfite sequencing data can be used to reveal both genetic and epigenetic changes. LiBis supports a one-command solution for quality control, trimming, mapping, and methylation calling in a reasonable computing time, making it an effective and comprehensive solution to support large-scale single-cell or cfDNA bisulfite sequencing applications.

2.2. Methods

2.2.1. LiBis Implementation

LiBis is available on GitHub (https://github.com/Dangertrip/LiBis). LiBis was developed in Python 3.6 and integrated with published software for infrastructure functionalities, which can be used for processing low-input WGBS data, such as cfDNA methylome and single-cell DNA methylome sequencing data. For first-round mapping and remapping, BSMAP used '-S 123 -n 1 -r 0 -U' as setting parameters for uniquely mapped reads and reproducible results. For second-round mapping, parameters such as window length and the distance from the previous window start (stride) are set by user. After the second-round mapping, contiguously overlapped mapped fragments were combined with the following rules: 1) fragments must be strictly continuous, such that the distance between left ends of two overlapped fragments aligned on the genome must equal to the distance between the two left ends of sequences on the read; 2) only one mismatch is allowed on each fragment; 3) if, after the recombination, multiple overlapped recombined fragments are generated, the program will select the longest fragment with the fewest mismatches (Figure 2.1B); and 4) all recombined fragments that do not overlap are kept (Figure 2.1B). For paired-end reads, all clipped reads are treated as two single-end reads in rescue mapping stage. Then the paired-end information is added to the rescued reads if the mate of the read was also mapped or rescued. To achieve better time and space complexity, the program records the number of mismatches on the last 'S' base pair to accelerate the computation in the bam file. 'S' stands for the length of the stride. The sliding window in the visualization module was generated through "bedtools makewindows".

An HTML report template was provided in the pipeline. Datatable and JQuery were imported for data representation. Numeric results such as mapping ratio were extracted from the original reports of TrimGalore and BSMAP to a TSV file for visualization. Figures were generated during a computation by the Matplotlib package in Python. Principal component analysis plotting used methylation signals in sliding windows along the reference genome.

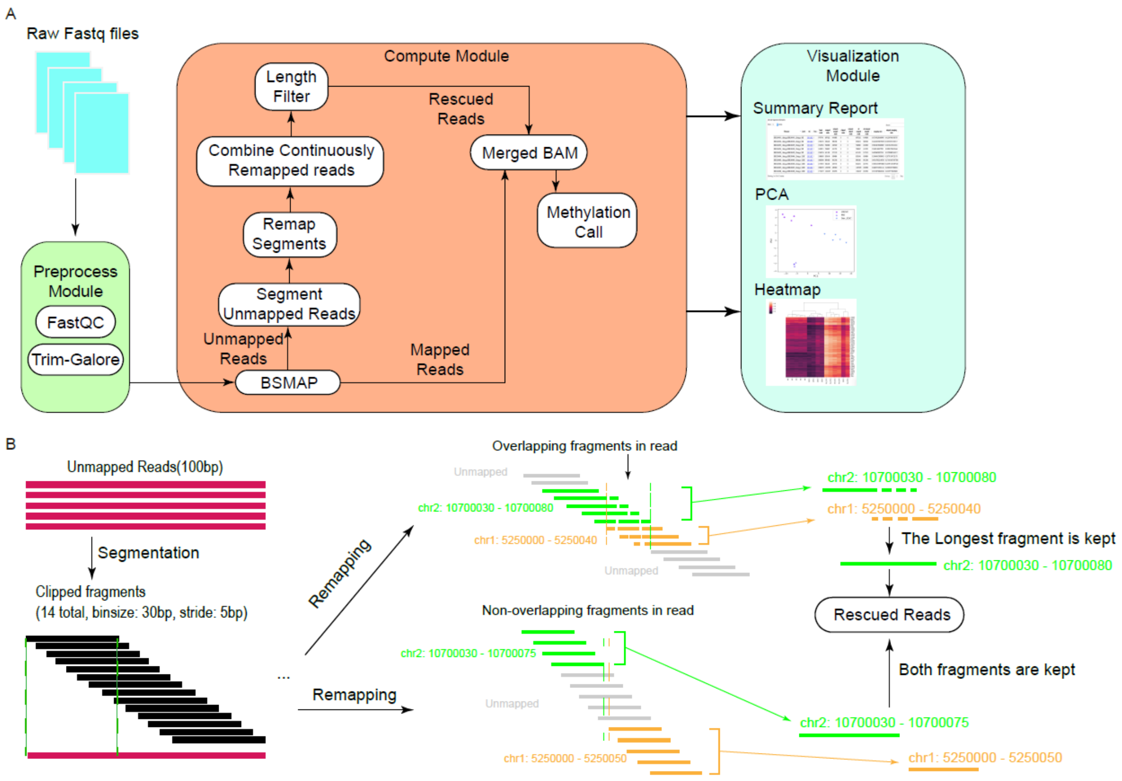


Figure 2.1 Workflow of LiBis (low-input bisulfite sequencing alignment). A. LiBis overview. **B**. Details of the LiBis rescue procedure, including clipping initially unmapped reads (left panel), remapping clipped fragments (middle panel), and recombining contiguous fragments (right panel).

2.2.2. Simulation methods

The random nucleotide was generated by the random function in Numpy. Real sequences (of 110 bases) were copied from the human reference genome with a random chromosome number and a random starting point. To each real sequence was added a head of 1–40 random bases (length m), as well as a random tail of (40-m) bases. All cytosines were randomized as unmethylated or methylated. The chance of being methylated was equal to the methylation ratio of the corresponding CpG site in human embryonic stem cells. For the fully randomized dataset, each nucleotide (A, C, T, or G) had an equal probability to fill each position.

2.2.3. Data collection and process

Signed informed consent was obtained from all patients or their legal guardians prior to sample acquisition in accordance with an institutional review board-approved protocol. CSF samples were obtained from two brain tumor patients at Texas Children's Hospital at the time of clinically indicated lumbar puncture. CSF was processed using a standardized protocol and was then divided into aliquots and stored immediately at -80°C. The cervical tumor tissue was obtained from Southern Medical University as an FFPE sample.

Cell-free DNA (cfDNA) was isolated from 200–400 µL CSF or plasma using a QIAamp Circulating Nucleic Acid Kit (Qiagen) according to the manufacturer's instructions. For tumor tissue, DNA was isolated using an AllPrep DNA/RNA Mini Kit according to the manufacturer's protocol. To process the FFPE cervical tumor sample, we used AllPrep DNA/RNA FFPE Kit (Qiagen, Cat No. 80234). The isolated DNA concentration was measured using a Qubit 4 Fluorometer with the Qubit dsDNA High Sensitivity Assay Kit (Thermo Fisher Scientific).

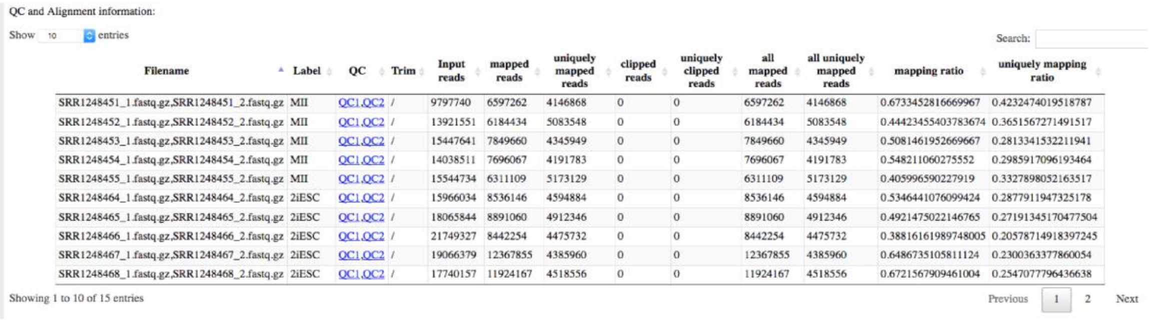
WGBS analysis was used to assess the genome-wide DNA methylation profile. The WGBS libraries were generated using a Pico Methyl-Seq Library Prep Kit (Zymo Research). Briefly, DNA was mixed with 0.1% unmethylated λ-bacteriophage DNA (w/w) (NEB), followed by sodium bisulfite conversion. The bisulfite-converted DNA was then annealed with random primers for initial amplification, followed by adaptor ligation and final amplification with Illumina TruSeq indices. Constructed libraries were run on a 2% agarose gel to assess size distribution, and the library concentration was measured using a Qubit 4 Fluorometer with a Qubit dsDNA High Sensitivity Assay Kit. Normalized libraries were pooled at an equimolar ratio and sequenced on NovaSeq 6000 (Illumina).

All WGBS data mapped by first round mapping with BSMAP are described as "BSMAP", and WGBS data mapped by a second round of clipped mapping are described as "LiBis rescued". The combination of two rounds of mapping are described as "LiBis". scBS-map was applied with default parameters. LiBis used 40 as the window size, five as the stride, and 45 as the filter length. Picard (http://broadinstitute.github.io/picard) was used to remove PCR duplicates from both first round BSMAP results and second round LiBis rescued reads before analysis of clinical samples. The smoothed scatterplots (geneplotter in R package) used the CpG sites in common between the two samples as input. Pearson correlations were calculated using the R cor function. Boxplots were plotted using the Python package seaborn. Fixed length trimming used Trim-Galore with '--clip_R1' and '--clip_R2'. Single cell whole genome bisulfite sequencing data is available at the Gene Expression Omnibus database under accession number GSE56879. Data used to compute correlations between LiBis-specific CpGs and public tumor samples are from the GBM tumor sample methylome at GSE121721 [56].

2.3. Results

2.3.1. Alignment augmentation strategy in LiBis for low-input bisulfite sequencing

LiBis is an integrated Python package for processing low-input WGBS data, such as cfDNA methylome and single-cell DNA methylome sequencing data. FastQC, Trim-galore, BSMAP, mcall, and bedtools are integrated into LiBis for quality control, adaptor trimming, read mapping, methylation calling, and functional analysis, respectively [30, 35, 53, 57]. The LiBis toolkit contains three modules: a preprocess module for quality control and adaptor trimming, a compute module for dynamic clipping and mapping, and a visualization module for report generation (Figure 2.1A, Figure 2.2).



QC and Alignment information:

Show 10 entries

Search:

Filename	Label	QC	Trim	Input reads	mapped reads	uniquely mapped reads	clipped reads	uniquely clipped reads	all mapped reads	all uniquely mapped reads	mapping ratio	uniquely mapping ratio
SRR1248451_1.fastq.gz,SRR1248451_2.fastq.gz	MII	QC1,QC2	/	9797740	6597262	4146868	0	0	6597262	4146868	0.6733452816669967	0.4232474019518787
SRR1248452_1.fastq.gz,SRR1248452_2.fastq.gz	MII	QC1,QC2	/	13921551	6184434	5083548	0	0	6184434	5083548	0.44423455403783674	0.3651567271491517
SRR1248453_1.fastq.gz,SRR1248453_2.fastq.gz	MII	QC1,QC2	/	15447641	7849660	4345949	0	0	7849660	4345949	0.5081461952669667	0.2813341532211941
SRR1248454_1.fastq.gz,SRR1248454_2.fastq.gz	MII	QC1,QC2	/	14038511	7696067	4191783	0	0	7696067	4191783	0.548211060275552	0.2985917096193464
SRR1248455_1.fastq.gz,SRR1248455_2.fastq.gz	MII	QC1,QC2	/	15544734	6311109	5173129	0	0	6311109	5173129	0.405996590227919	0.3327898052163517
SRR1248464_1.fastq.gz,SRR1248464_2.fastq.gz	2iESC	QC1,QC2	/	15966034	8536146	4594884	0	0	8536146	4594884	0.5346441076099424	0.2877911947325178
SRR1248465_1.fastq.gz,SRR1248465_2.fastq.gz	2iESC	QC1,QC2	/	18065844	8891060	4912346	0	0	8891060	4912346	0.4921475022146765	0.27191345170477504
SRR1248466_1.fastq.gz,SRR1248466_2.fastq.gz	2iESC	QC1,QC2	/	21749327	8442254	4475732	0	0	8442254	4475732	0.38816161989748005	0.20578714918397245
SRR1248467_1.fastq.gz,SRR1248467_2.fastq.gz	2iESC	QC1,QC2	/	19066379	12367855	4385960	0	0	12367855	4385960	0.6486735105811124	0.2300363377860054
SRR1248468_1.fastq.gz,SRR1248468_2.fastq.gz	2iESC	QC1,QC2	/	17740157	11924167	4518556	0	0	11924167	4518556	0.6721567909461004	0.2547077796436638

Showing 1 to 10 of 15 entries

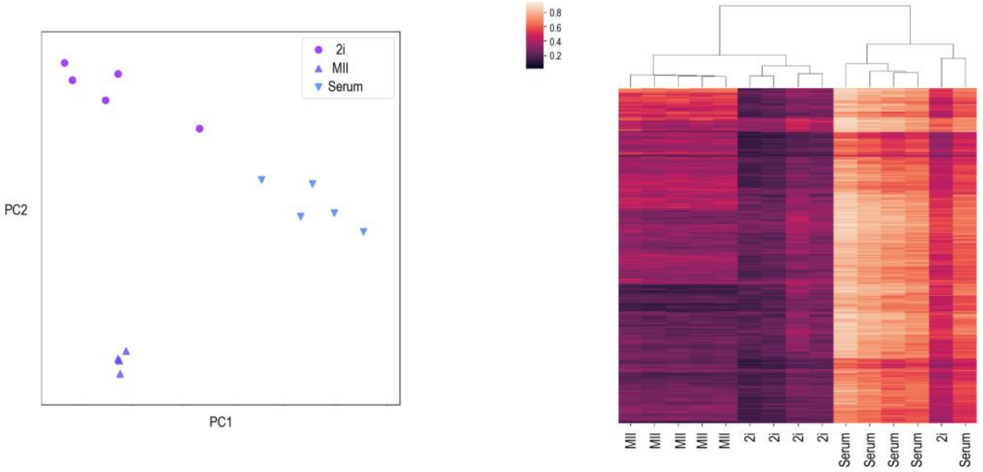


Figure 2.2 Example of the final report generated by the LiBis visualization module. Top panel, table of basic statistics and HTML links to FastQC figures. Bottom panel, principal component analysis scatter plot and heatmap.

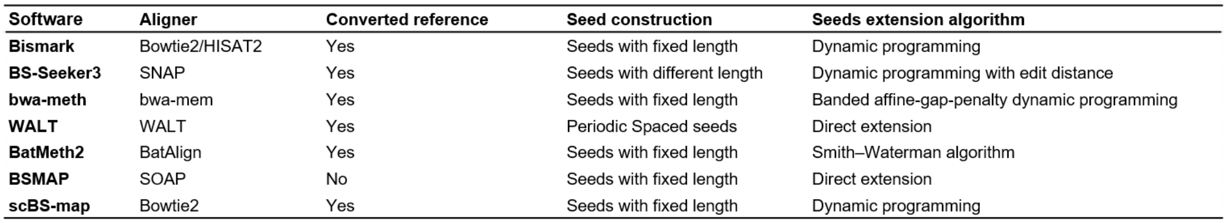
To improve the mapping efficiency for low-input bisulfite sequencing data, we applied a clipping strategy within the compute module to eliminate random base contamination, as follows. First, LiBis maps the trimmed raw data with BSMAP and generated a new fastq file containing unmapped reads. Second, LiBis clips all unmapped reads using a sliding window with a specific width and step size as defined by the user. Third, LiBis remaps all clipped read fragments and keeps only uniquely mapped fragments for subsequent recombination. During recombination, fragments derived from the same unmapped read are recombined only if they are remapped contiguously to the reference genome, and the distance between two adjacent fragments equals the step size. Recombined fragments are required to be above a minimum length to reduce the likelihood of a false-positive alignment. For reads with multiple candidate recombined clipped fragments that overlap, such as that illustrated in the top half of Figure 2.1B, the recombined clip with the greatest mapping confidence (i.e., the longest clip with least mismatches) is kept as a rescued read. If the recombined clipped fragments do not overlap each other, such as the two clips in the bottom half of Figure 2.1B, all recombined clipped fragments will be kept. Through the remapping and recombination steps mentioned above, reads discarded in first-round mapping can be rescued.

The LiBis workflow for bisulfite sequencing data involves five steps. In step 1, the raw reads are examined for quality control by FastQC. FastQC allows assessment of quality features including base quality, base content, and duplication level. These features reveal the overall quality of the library, amplification, and sequencing. Results of FastQC are aggregated in the final report. In step 2, the reads are trimmed by trim-galore, which removes sequencing adaptors and low-quality reads. If random priming was used to generate the library, trimming is recommended but cannot remove random priming-associated amplification artifacts. Step 3 is read mapping by the compute module, which combines initial mapping with the dynamic clipping and remapping strategy to improve the cost efficiency. In step 4, the methylation ratio of CpGs is called by mcall from the MOABS program. In step 5, the data is visualized as various figures. After LiBis analysis, an overview webpage presents summaries of all input samples, heatmaps, and principal component analysis results (Figure 2.1A).

LiBis requires two types of input files, namely the reference genome sequence file in FASTA format and the FASTQ files containing raw reads. For small numbers of samples, a command line can be used to run LiBis. Alternatively, config file reader was developed for large numbers of samples. The LiBis output includes bam files, methylation ratio files, quality control results, stats files and an integrated HTML report.

2.3.2. LiBis improves mapping ratios and mapping sensitivity in simulated data

To compare the mapping effectiveness of LiBis with published approaches, we determined the mapping ratio of BSMAP [35], Bismark [54], BS-Seeker3 [58], scBS-map [39], bwa-meth [59], WALT [60] and BatMeth2 [61] on a single cell WGBS dataset (GSE56879; SRR1248444-SRR1248455) (Table 2.1). Results showed that WALT, Bismark and BS-Seeker3 had a median mapping ratio ranging from 13% to 22%, and that scBS-map, bwa-meth and BatMeth2 had a median mapping ratio ranging from 35% to 55%. Compared to these tools, LiBis had the highest median mapping ratio 65% (Figure 2.3AB).



Software	Aligner	Converted reference	Seed construction	Seeds extension algorithm
Bismark	Bowtie2/HISAT2	Yes	Seeds with fixed length	Dynamic programming
BS-Seeker3	SNAP	Yes	Seeds with different length	Dynamic programming with edit distance
bwa-meth	bwa-mem	Yes	Seeds with fixed length	Banded affine-gap-penalty dynamic programming
WALT	WALT	Yes	Periodic Spaced seeds	Direct extension
BatMeth2	BatAlign	Yes	Seeds with fixed length	Smith–Waterman algorithm
BSMAP	SOAP	No	Seeds with fixed length	Direct extension
scBS-map	Bowtie2	Yes	Seeds with fixed length	Dynamic programming

Table 2.1 Features of collected software.

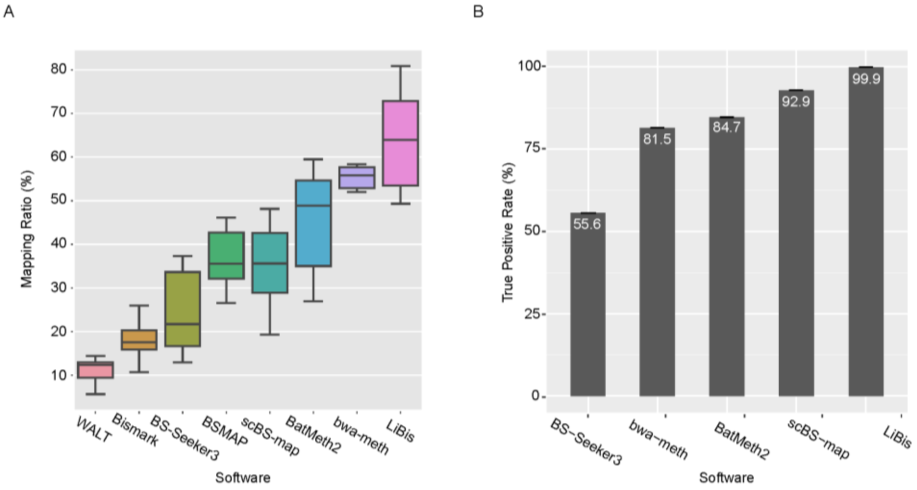


Figure 2.3 Evaluation of mapping efficiency and accuracy of LiBis. A. Mapping ratio of MII oocyte single cell WGBS samples by indicated methods. **B.** True positive rates of mapped reads generated by indicated methods.

To further test other features of LiBis, three simulation datasets (each containing 10 million 150 base pair length reads) were randomly and independently generated in silico. Each 150 base read was composed of a real 110 base DNA sequence (randomly cut from the hg38 human genome), a random artificial head sequence added to the front of the real read, and a random artificial tail sequence added to the end of the real sequence while the head and tail add up to 40 bases. The random heads and tails simulated contamination introduced by the random priming process, which cannot be fully removed by traditional trimming methods such as trim-galore (Figure 2.4A).

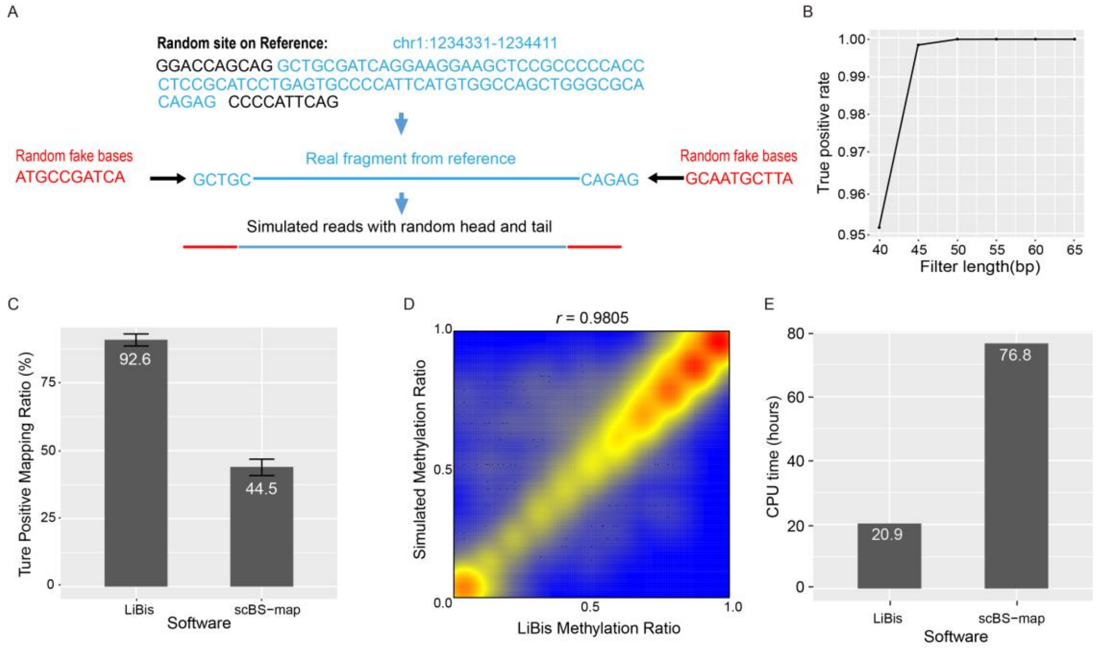


Figure 2.4 LiBis performance on simulation datasets *in silico*. A. Scheme for generating simulated reads with random heads and tails. **B.** True positive rates of LiBis using different filter lengths. **C.** Mapping ratios using simulated reads by LiBis and scBS-map. **D.** Smoothed scatterplot showing the correlation between the DNA methylation ratios generated by LiBis and simulated reads. **E.** CPU times for processing 10 million reads (150 bp) by LiBis and scBS-map.

The LiBis mapping ratio is associated with the minimum length of the rescued reads. We investigated the relationship between the true positive rate and the filter length. A true positive event was defined if the simulated read was aligned to a genomic location which was same as its simulated location. As shown in Figure 2.4B, the true positive rate saturated when the parameter of minimum rescued length was set at 45 bases. Among the different mapping approaches tested, LiBis and scBS-map showed the highest true positive rates (Figure 2.3B). Among these tools, to our knowledge, scBS-map and LiBis are the only two tools designed to rescue WGBS reads with artificial or chimeric sequences introduced by the random priming process, though the original purpose for scBS-map and LiBis are single cell WGBS and cfDNA WGBS, respectively. Thus, we further compared the mapping ratio between LiBis and scBS-map [39]. For simulation data, LiBis achieved a mapping ratio (92.6%) which is higher than scBS-map (44.5%) (Figure 2.4C). The DNA methylation ratios of CpGs identified by LiBis also showed a high correlation (r=0.98) with the ground truth of methylation ratios from the simulation (Figure 2.4D), indicating that LiBis has high accuracy. For real data, LiBis also had a higher mapping ratio (52%) than scBS-map (around 45%) using GSE81233 [62], and a higher median mapping ratio (65%) than scBS-map (35%) using SRR1248444 - SRR1248455. In addition, by analyzing the rescued reads from GSE81233 [62], we found that most of the rescued reads are indeed synthetic reads (Figure 2.5B). These results from simulation data and real data revealed that LiBis could rescue synthetic reads and achieved high detectability and high sensitivity.

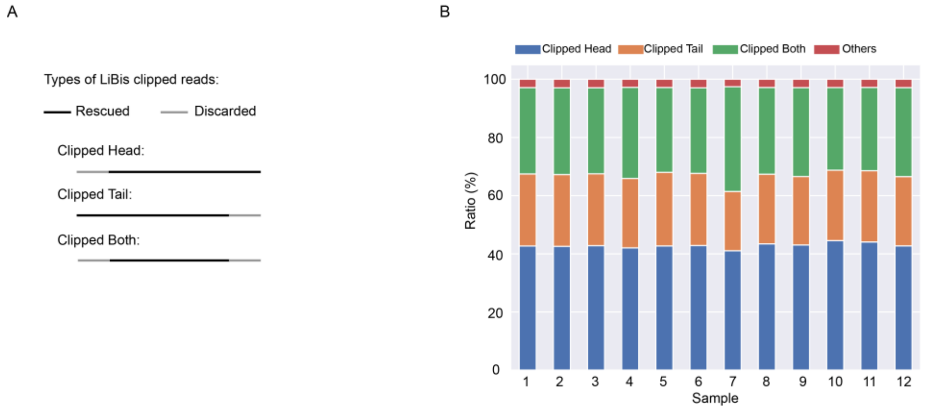


Figure 2.5 Diagram of LiBis clipped reads. A. Clipped types of LiBis rescued reads. **B.** Percentages of different clipped types of LiBis rescued reads in sperm single cell samples in GSE81233.

Regarding efficiency, LiBis required only 20.9 CPU hours in total to process the simulated dataset (10 million 150 bp length reads) from fastq input, which is comparable to BSMAP (10.6 CPU hours) considering the rescue of contaminated reads from simulation dataset. In contrast, LiBis is nearly four times faster than scBS-map (Figure 2.4E).

2.3.3. LiBis improves mapping ratios and mapping sensitivity in a single-cell DNA methylome dataset

To test the efficiency of LiBis using real data, we applied LiBis on single cell mouse ESC WGBS samples from a public dataset (GSE56879). Similar to post bisulfite WGBS, single cell WGBS applies random priming to capture bisulfite truncated fragments. In this dataset, we observed that the mapping ratio for the scWGBS samples increased by as much as 24% with LiBis compared to BSMAP (Figure 2.6A). A large number of CpGs were specifically discovered by LiBis, but failed to be detected by BSMAP (Figure 2.6B). To further gauge the effects of these improvements on downstream analysis, we defined CpGs that were covered at least 5 times as informative CpGs. LiBis recovered as many as 70% more informative CpGs compared to BSMAP, dramatically increasing the number and depth of usable CpG sites (Figure 2.6C). For the CpG sites specifically discovered by LiBis, our result shows that these CpGs had relatively higher methylation ratios compared to the initially discovered CpGs (Figure 2.7A), indicating that reads from highly methylated regions may suffer more severe contamination from random priming, which may partially explain the lower methylation ratios detected in libraries built using random priming versus in libraries prepared using pre-bisulfite sequencing methods [51].

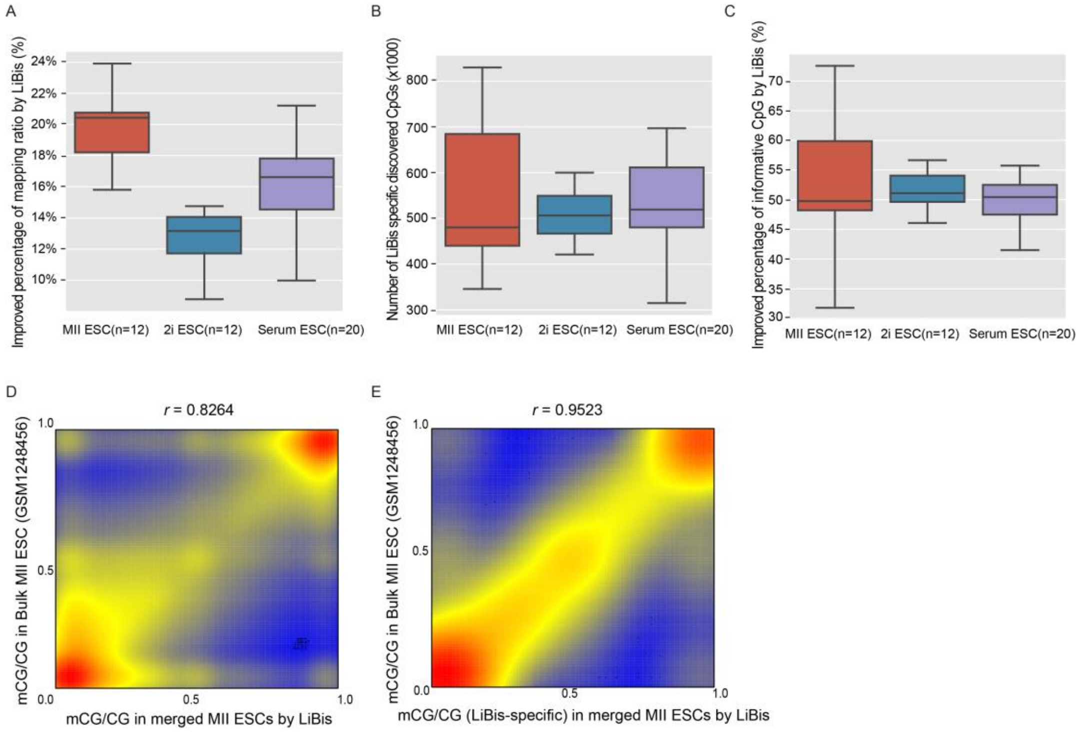


Figure 2.6 LiBis improved both the efficiency and the precision of methylation measurements in scWGBS datasets Diagram of LiBis clipped reads. A. Improvement in mapping ratios by LiBis, compared to BSMAP in three scWGBS datasets; **B.** Number of CpGs specifically detected by LiBis, but not detected by BSMAP, in three scWGBS datasets; **C.** Improved informative CpGs (covered at least 5 times) using LiBis-rescued reads compared to BSMAP. **D.** Smooth scatterplot showing the correlation of the DNA methylation ratio (averaged for 1-kb windows) between bulk and merged single cell DNA methylomes. **E.** Smooth scatterplot showing the correlation of DNA methylation ratio between bulk and merged DNA methylomes at regions specifically rescued by LiBis, but not detected by BSMAP.

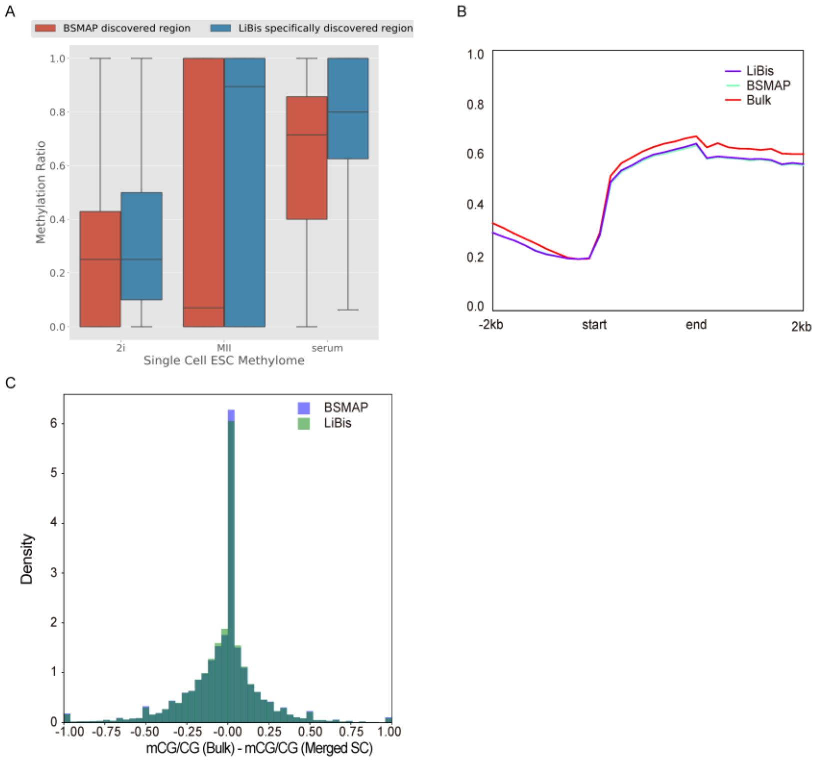


Figure 2.7 Overview of LiBis processed scWGBS dataset. A. Boxplots showing the distribution of DNA methylation ratios calculated using BSMAP reads and LiBis-specific rescued reads (i.e., not mapped by BSMAP) from three scWGBS datasets. **B.** Merged methylomes generated with LiBis or without LiBis exhibit nearly identical methylation ratio distributions over gene regions. We analyzed all genes in the genome beginning 2 kb before the transcription start site (Start) through 2 kb past the transcription end (Stop), and aligned gene positions along the X axis. The Y axis is the methylation ratio. The bulk methylome by BSMAP is in red, the merged methylome with or without LiBis is in purple or blue, respectively. **C.** The distributions of the methylation ratio difference between bulk and merged methylomes are subtly different between BSMAP and LiBis.

To estimate the accuracy of the DNA methylation ratios recovered by LiBis, we performed a comparison between the DNA methylation ratios generated by LiBis using mixed scWGBS data (LiBis-MII ESCs-scWGBS) and the DNA methylation ratios generated by BSMAP using bulk methylome data (BSMAP-MII ESCs-bulk) from the same cells. We observed a high correlation in the DNA methylation ratio (Pearson correlation r = 0.83) between LiBis-MII ESC-scWGBS and BSMAP-MII ESCs-bulk on the 2.5 million CpGs detected in common (Figure 2.6D, 2.7B). This result indicated that LiBis can identify CpGs with accurate DNA methylation ratios. By comparing with BSMAP-MII ESC-scWGBS, we identified 187,804 CpGs that were specifically recovered by LiBis but not by BSMAP in merged single cell samples, while the methylation ratios of these specific CpGs were highly correlated with those from BSMAP-MII ESCs-bulk (Pearson correlation r=0.95) (Figure 2.6E). Additionally, the methylation ratio difference between the BSMAP bulk methylome and the merged scWGBS methylome by BSMAP or LiBis was subtle and comparable with the results using BSMAP (Figure 2.7C). These results strongly demonstrated that the CpG DNA methylation ratios from LiBis rescued reads are accurate and beneficial for downstream analysis.

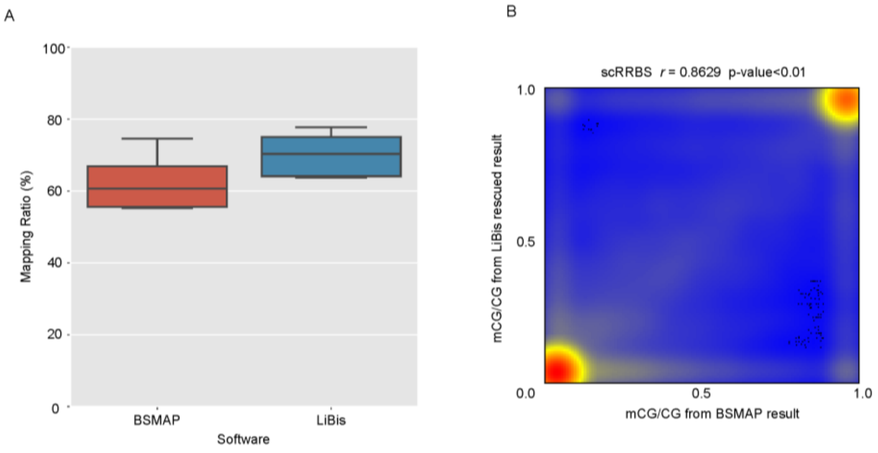


Figure 2.8 LiBis improvement on scRRBS. A. Improvement of mapping ratios by LiBis, compared to BSMAP in scRRBS datasets. **B.** Smooth scatterplots showing the correlation of the DNA methylation ratios of CpGs detected in common by BSMAP and LiBis.

In addition, we also tested the performance of LiBis on single cell reduced representation bisulfite sequencing (scRRBS) data (GSE109085) and we obtained similar results as scWGBS data (Figure 2.8), suggesting that LiBis worked on scRRBS data as designed.

2.3.4. LiBis improves the data efficiency of tumor and cfDNA WGBS experiments with random priming

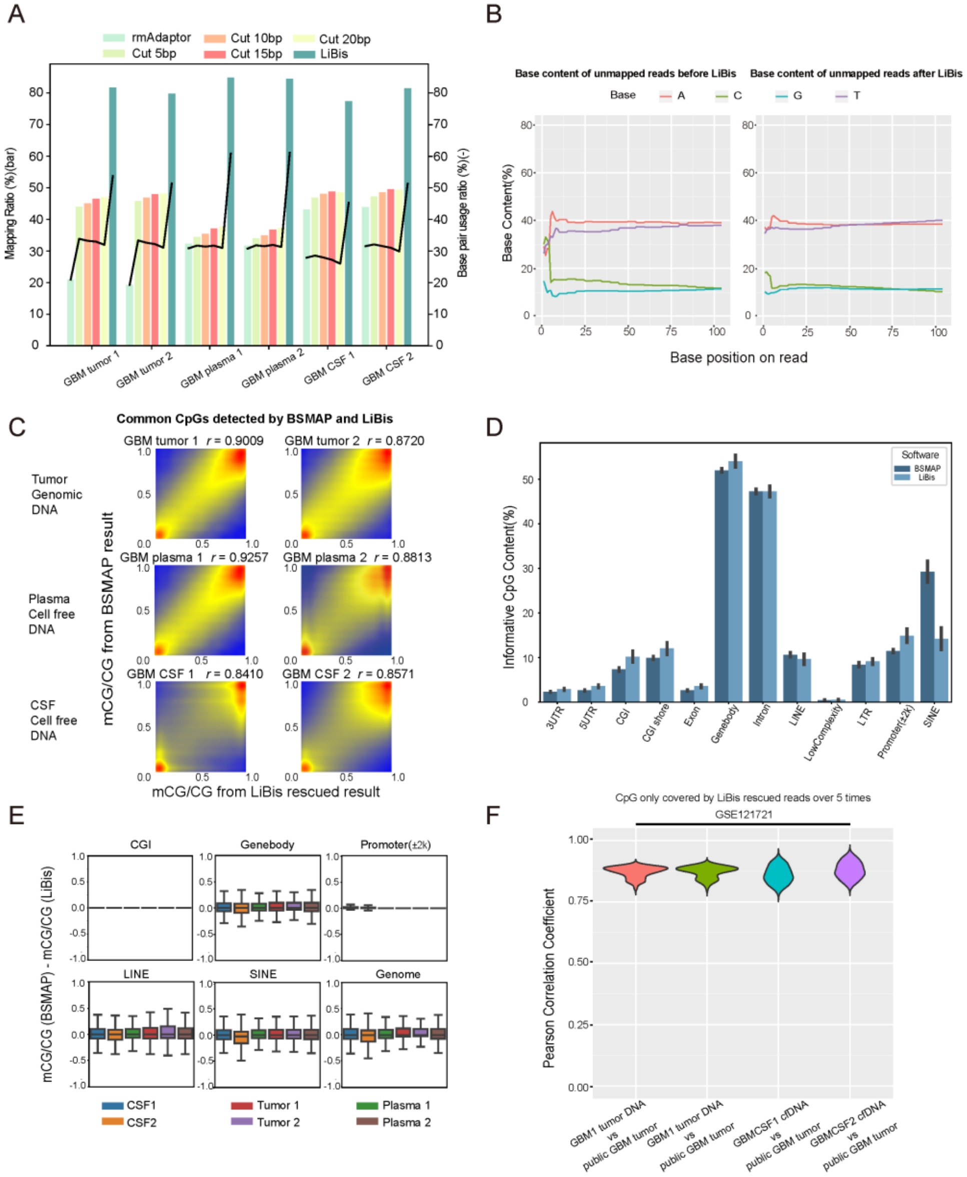


Figure 2.9 LiBis improved the cost efficiency of clinical tumor and cfDNA WGBS data generated using the random priming method. A. Improvement (mapping ratio:

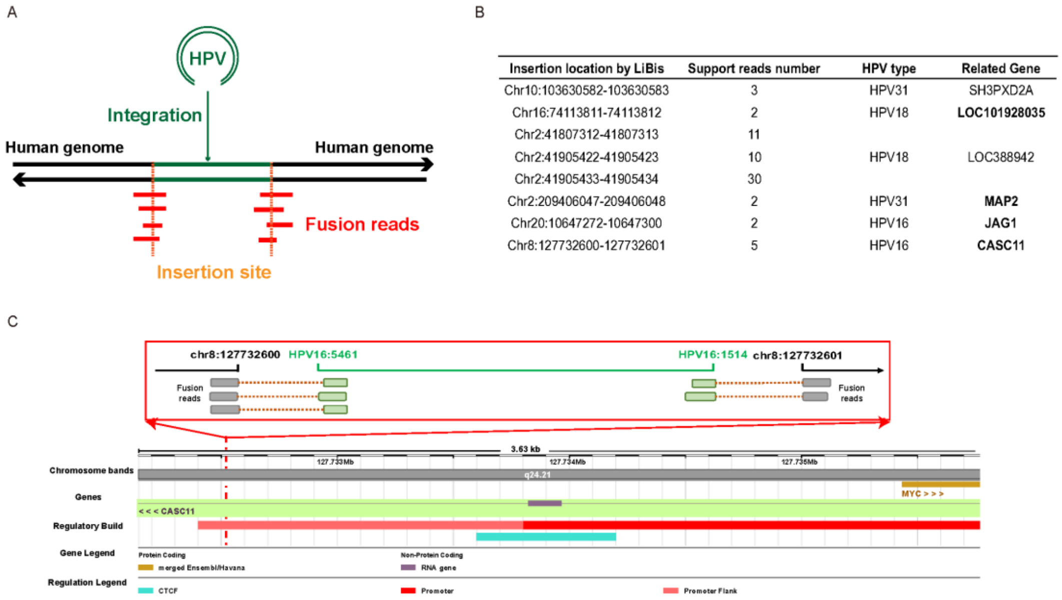
bar plots with left y axis; base pair usage: black lines with right y axis) comparisons between LiBis and fixed length trimming. B. Base content distribution of BSMAP unmapped reads (left panel) and LiBis rescued reads (right panel). C. Smooth scatterplots showing the correlation of the DNA methylation ratios of CpGs detected in common by BSMAP and LiBis. The Pearson correlation coefficients are listed on the top of each scatterplot. D. Distribution of the DNA methylation level across functional genome elements using informative CpGs as determined by mapped reads using BSMAP and the specific recovered reads using LiBis, as indicated. E. The differences in DNA methylation ratios of the CpGs detected in common between BSMAP and LiBis in genomic regions and across the global genome. F. The distribution of DNA methylation ratio Pearson correlations of LiBis-specific read CpGs (compared to BSMAP results) from our data compared to public bulk DNA methylomes (GSE121721) from the same tumors.

To verify the capability of LiBis to rescue bisulfite sequencing data from low input WGBS (such as cfDNA samples), we performed WGBS on six clinical samples and processed the sequencing data using LiBis: 1). Two cerebrospinal fluid (CSF) cfDNA samples from glioblastoma patients; 2). Two plasma cfDNA samples from glioblastoma patients; 3). Two genomic DNA samples from glioblastoma tumor tissues. We adopted random priming-based WGBS to generate libraries of all collected samples [23]. To validate the functionality of LiBis, we first compared LiBis with the traditional fixed length trimming method (Figure 2.9A). Although the fixed length trimming method could improve the mapping ratio, there was only a slight change in the base pair usage rate (right Y-axis). LiBis significantly improved both the mapping ratio and base pair usage over the traditional fixed length trimming method. Furthermore, by analyzing the base content of the sequences, we found that our samples prepared using the random priming library method showed a strong C-bias at the beginning of the reads. This observation was also reported in papers using single-cell bisulfite sequencing due to the dNTP in random priming [24, 52]. LiBis can effectively eliminate the C-bias by dynamically removing random artifactual bases in both ends of the read (Figure 2.9B).

Next, to measure the similarity in the DNA methylation ratios of each CpG comparing reads recovered by LiBis with using the end-to-end mapping reads, we performed a correlation analysis. We observed a high correlation coefficient that ranged from 0.84 to 0.92 (average 0.88) between the DNA methylation ratios from the end-to-end mapping and the LiBis clipped mapping on the CpGs covered at least 10 times by both methods (Figure 2.9C). LiBis recovered informative CpGs share a similar distribution in different genomic regions while having a relatively low recovery rate in repeat regions (Figure 2.9D), which suggests that there is no methylation bias. The reason for the low number of LiBis rescued reads in repeat regions is that short fragments may lead to multiple mapping on repeat regions, which leads to a lower unique mapping ratio in repeat regions. In addition, the majority (>75%) of the CpGs detected in common by BSMAP and LiBis have a DNA methylation ratio difference less than 0.15 across genome elements. (Figure 2.9E). CpGs at regions of low methylation such as CpG islands or promoters showed no methylation ratio difference because as the methylation ratio distribution became more skew, the confidence interval narrowed. These results further showed that LiBis faithfully recovered CpG DNA methylation ratios. CpGs uniquely recovered by LiBis, but not by BSMAP, had DNA methylation ratios that were highly correlated with those from public tumor tissue methylome databases (average Pearson coefficient r=0.87) (Figure 2.9F). These results strongly demonstrated that LiBis can efficiently and correctly extract maximum DNA methylation information from low-input bisulfite sequencing data in clinical samples.

2.4. Discussion

DNA fragments sequenced by post-bisulfite sequencing are reported to undergo self-priming and generate chimeric reads, which leads to low efficiency of utilizing sequencing raw data [28, 52]. In this study, we developed a pipeline called LiBis to reduce this problem in post-bisulfite sequencing by fragmenting, remapping and recombining initially unmapped reads. To the best of our knowledge, LiBis is the first integrated pipeline for the processing of low-input bisulfite sequencing data that includes trimming, initial mapping, clipped mapping, methylation calling, and visualization. By dynamically clipping unmapped reads to remove random priming contamination, LiBis is capable of improving experimental measurement precision and cost efficiency.



Insertion location by LiBis	Support reads number	HPV type	Related Gene
Chr10:103630582-103630583	3	HPV31	SH3PXD2A
Chr16:74113811-74113812	2	HPV18	**LOC101928035**
Chr2:41807312-41807313	11	HPV18	LOC388942
Chr2:41905422-41905423	10		
Chr2:41905433-41905434	30		
Chr2:209406047-209406048	2	HPV31	**MAP2**
Chr20:10647272-10647300	2	HPV16	**JAG1**
Chr8:127732600-127732601	5	HPV16	**CASC11**

Figure 2.10 LiBis identified HPV insertion sites in cervical cancer WGBS experiments using a traditional pre-bisulfite library method. A. Scheme of HPV

insertion. The green line represents HPV genome sequences; black line represents human genome sequences. Vertical orange lines represent insertion sites; red short lines represent the fusion reads which are chimeric reads with human and HPV sequences spanning across an insertion site. **B.** Genomic coordinates of HPV insertion sites in a cervical cancer identified by LiBis. **C.** Visualization of a sample HPV insertion site on chromosome 8. The exact positions of insertion are labeled. Five fusion reads (three on the left most site, two on the right) were identified.

To further explore the capability of LiBis to discover genomic patterns using WGBS data, we collected one cervical cancer tumor sample and prepared one WGBS library using a traditional protocol. The mapping ratio improved from 84% with BSMAP to 90% with LiBis using the 150 bp paired-end sequencing strategy. Although the improvement was relatively small compared to that observed for the cfDNA WGBS data (random primers based), we confirmed that LiBis can identify DNA insertion events, in particular for human papillomavirus (HPV) DNA insertion into the human genome in this case study. In cervical cancer, HPV can cleave human DNA and insert HPV DNA fragments into the human genome in a sense or antisense direction. These HPV insertion events can be detected by fusion reads which contain both HPV and human DNA fragments identified by LiBis (Figure 2.10A). Multiple potential insertion sites in this patient sample were identified using LiBis (Figure 2.10B). Next, we examined the closest gene to each insertion sites to evaluate the potential influence of the viral insertion on human gene expression. Interestingly, four out of the six insertion-proximal genes we identified were previously reported as HPV insertion sites or were differentially expressed in cervical cancer [63-65]. For example, one insertion site was located in the intron of the CASC11 gene and also close to the promoter region of the MYC gene (Figure 2.10C). The expression of CASC11 is upregulated in cervical cancer tissue, compared to normal cervical tissue, which might activate the WNT/β-catenin signaling pathway to promote cervical cancer. And the MYC gene is well known for promoting tumorigenesis [66]. In addition to identifying HPV integration sites, LiBis also can identify differentially methylated CpGs in both human and HPV sequences (if the DNA methylome data is available for control or unintegrated HPV). These results suggest that LiBis can simultaneously identify the viral integration sites and perform differential methylation analysis for both human and viral DNA using WGBS data, which will significantly improve the DNA methylation analysis of viral integration in cancer.

Here, we confirmed that LiBis improved performance in four case studies, namely simulated data, scWGBS data, cfDNA WGBS data from random priming libraries, and tumor WGBS data from random priming libraries. The performance improvement may due to several reasons. First, random priming, adapter dimers, small fragments, or bisulfite conversion may generate artifactual bases at both ends of a read. Our approach reduces the proportion of reads culled due to these issues. Second, through dynamic clipping, LiBis could recognize DNA fusions present in the original cell, including gene fusions, insertions of HPV DNA, fusions that arose during library preparation, and even circular DNA formed by self-priming or other mechanisms.

For data preprocessing, Trimmomatic allows more sophisticated trimming procedures than trim-galore. We applied Trimmomatic and TrimGalore on scWGBS dataset (SRR1248444 - SRR1248455) with sophisticated procedures. Our results showed that more reads were discarded by Trimmomatic (around 30%) than TrimGalore (5.8%) (Figure 2.11A), and that Trimmomatic achieved a higher percentage than TrimGalore for high-quality sequences, where the percentage was defined as bases with quality score larger than 30 divided by all bases (Figure 2.11B). This indicated that Trimmomatic was stricter than TrimGalore in terms of quality score. Next, we compared the base quality distributions for reads mapped by BSMAP, reads rescued by LiBis, and fragments discarded by LiBis. The percentage of high-quality sequences among these three groups were similar (97%-98%), suggesting that LiBis rescue was relatively independent to base quality score (Figure 2.11C). Above all, rescued reads by LiBis were mostly from artificial, synthetic, or chimeric sequences with good quality, which could not be removed by TrimGalore or Trimmomatic at all.

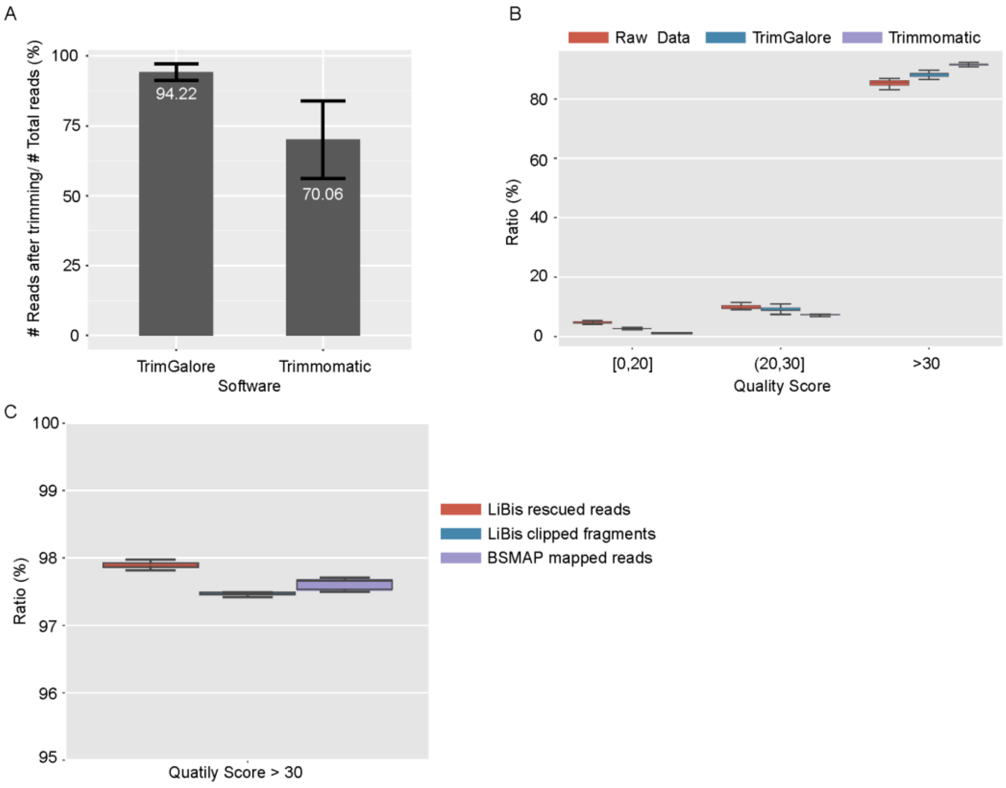


Figure 2.11 Comparison of widely used trimming software for low input bisulfite sequencing. A. Remained reads ratio after trimming by TrimGalore and Trimmomatic. **B.** Distribution of quality score in raw data, data trimmed by TrimGalore and data trimmed

by Trimmomatic. **C.** Percentages of bases with quality score>30 in LiBis rescued reads, LiBis clipped fragments and BSMAP mapped reads.

To maximize the mapping efficiency of post-bisulfite sequencing material, our method shares similar concepts with previous reports. scBS-map applies a local alignment after an end-to-end mapping step to rescue discarded reads [39]. Taurus-MH splits the unmapped reads in end-to-end mapping with a fixed pattern to estimate the mappable parts [67]. From the aspect of implementation, LiBis integrated BSMAP and applied the sliding window approach to segregate reads to sub-sequences and merge the overlapped sub-sequences; while scBS-map used Bowtie2 as aligner and integrated BS-Seeker2 "end-to-end" mode and "local" mode to improve the alignment ratio. Moreover, BSMAP aligned seeds by loop-up hash table and wildcard algorithm; while BS-Seeker2 adapted 3-letter approach and converted the genome into a C-to-T reference and a G-to-A reference. In addition to the differences of algorithm in integrated aligners, local mode in Bowtie2 contained a scoring system for seed extension. However, LiBis aligned segregated reads independently in each sliding window. Since the contamination of random priming was mainly present at the beginning and the end of the reads, direct removal of these bases by sliding window is a more intuitive way, and confirmed to be more efficient. The simple strategy and parameters in LiBis also allow users to adjust the parameters to balance the running time and the rescue efficiency. For example, users can set a smaller stride and window length when the sequencing quality is unsatisfactory to improve the cost efficiency. Compared to previous methods, LiBis has two advantages. First, LiBis had significantly higher efficiency than the traditional trimming method or other mapping approaches in processing WGBS data from a random priming library. Second, LiBis is an integrated pipeline including quality control, trimming, mapping, methylation calling and visualization modules, which can achieve one-command processing of WGBS raw data. However, all methods including LiBis can be further enhanced by applying parallel computing for mapping processes to query more computational capacity. A more precise iteration of parameters for LiBis can also improve the efficiency batch to batch.

3. MMINT: METHYLATION DATA MINING TOOLS

3.1. Background

DNA methylation is one of the most important epigenetic modifications which plays a predominant role in a broad range of biological processes, including development, tumorigenesis and cellular identity. It involves the addition of a methyl group to the carbon-5 position of cytosine residues. Treatment of DNA with bisulfite specifically converts cytosine residues to uracil but leaves 5-methylcytosine residues unaffected. In mammals, DNA methylation is almost exclusively existing on 5-methylcytosine (5-mC) residues in the context of CpG dinucleotides. Whole Genome Bisulfite Sequencing (WGBS) allows investigators to detect the methylation status for each single CpG site in the genome. Besides DNA methylation, techniques are also developed for other epigenetic markers such as TAB-Seq, anti-CMS technique for DNA hydroxymethylation (5hmC) [68, 69], ChIP-Seq for histone modifications and transcription factors (TFs), which all contribute to the genome regulatory landscape. It was established that active regulatory regions (such as promoters and enhancers) are always bound by TFs with low CpG methylation, while high methylation is reported as interference of the TFs binding process [70-72]. However, by characterizing the effect of cytosine methylation on 542 human TFs binding, many TFs bindings were found to be enhanced by cytosine methylation [73]. In addition, the correlation analysis between DNA methylation and histone modification indicated that the demethylation process is associated with H3K4me1 and H3K27ac which marked enhancer regions [74, 75]. These recent findings indicate that DNA methylation plays multiple and complex roles interacting with other epigenetic markers. However, the ability to comprehensively determine the multidimensional epigenetic modifiers' roles for one certain biological process is in unprecedented breadth and detail. Due to lack of appropriate tools, the large-scale crosstalk among DNA methylation, DNA hydroxymethylation, TF bindings and histone modifications remains insufficiently studied.

Moreover, to select representative regions/genes to demonstrate their findings vividly, researchers always spend hours manually browse UCSC genome browser (http://genome.ucsc.edu/) or WashU epigenome browsers (http://epigenomegateway.wustl.edu/browser/) to look for regions presenting region-level coexistence, exclusive or gradually change the relationship among multiple epigenetic data, which is time-consuming and inefficient. To facilitate such a process, genomation [76], ChiPseeker [77], and deepTools2 [78] were developed to visualize certain signals in certain genomic intervals. However, none of them can perform integrative analysis with multiple signals, especially with DNA methylation data. To address these challenges, we developed Mmint, a comprehensive data mining and visualization software centered on WGBS data. The features that make Mmint unique are (I) it conveniently integrates large-scale WGBS data and other epigenetic data for quality assessment and comprehensive analysis; (II) Instead of using descriptive and numeric files, Mmint generates publishable figures to present the results.

3.2. Methods

Mmint is written in Python 3.6 and is available as an open-source software at Github (https://lijiacd985.github.io/Mmint/). To perform as an integrative pipeline, Mmint is also available as the analysis toolkit in MOABS (https://github.com/sunnyisgalaxy/moabs).

All the sequencing data are mapped on the hg19/mm10 genome. MOABS is used to perform analysis of WGBS data [53]. Bowtie2 with default settings is used as the aligner for other types of sequencing data. MACS2 and DESeq2 are used for peak calling and differential peak identification in ChIP-Seq data [79, 80]. The criteria of identified differentially methylated regions (DMRs) are: (I) DMR needs at least 3 CpGs. (II) The absolute mean difference of methylation ratio between is regions are larger than 0.25. (III) The false discovery rate (FDR) of identified DMR is less than 0.05. Such criteria are decided to focus on highly different regions and minimize the influence the false discovery.

3.3. Results

3.3.1. Overview of Mmint

Mmint takes output files from Model-based Analysis Of Bisulfite Sequencing data (MOABS) such as BAM files and BED files [53] and the peak files generated by MACS2 [79]. By taking this information, Mmint can support five types of analysis: (I) Multi-dimensional quality assessment for WGBS data; (II) Versatilely integrative analysis of DNA methylation data and other epigenetic data; (III) Integrative analysis between epigenetic markers and gene expression; (IV) Integration of multiple signals in all interested regions (Figure 3.1). For all types of analysis, Mmint can generate figures and well-defined tables for users, which is convenient and interpretable for users without bioinformatics background. More importantly, Mmint has been embedded into MOABS, which allows users to finish the WGBS data analysis process from raw data to publishable figures in a single step. Besides analysis and visualization, management of increasing WGBS datasets also becomes a pain point of data analysis. Lab information management systems such as HTS-flow and Galaxy are used to manage large batches of next generation sequencing data. However, these lab information management systems lack domain knowledge about sequencing such as library construction and coverage requirement. To fill this gap, Mmint integrates an empirical decision tree to reveal the potential problem in a raw dataset, which allows users to diagnose their sequencing process.

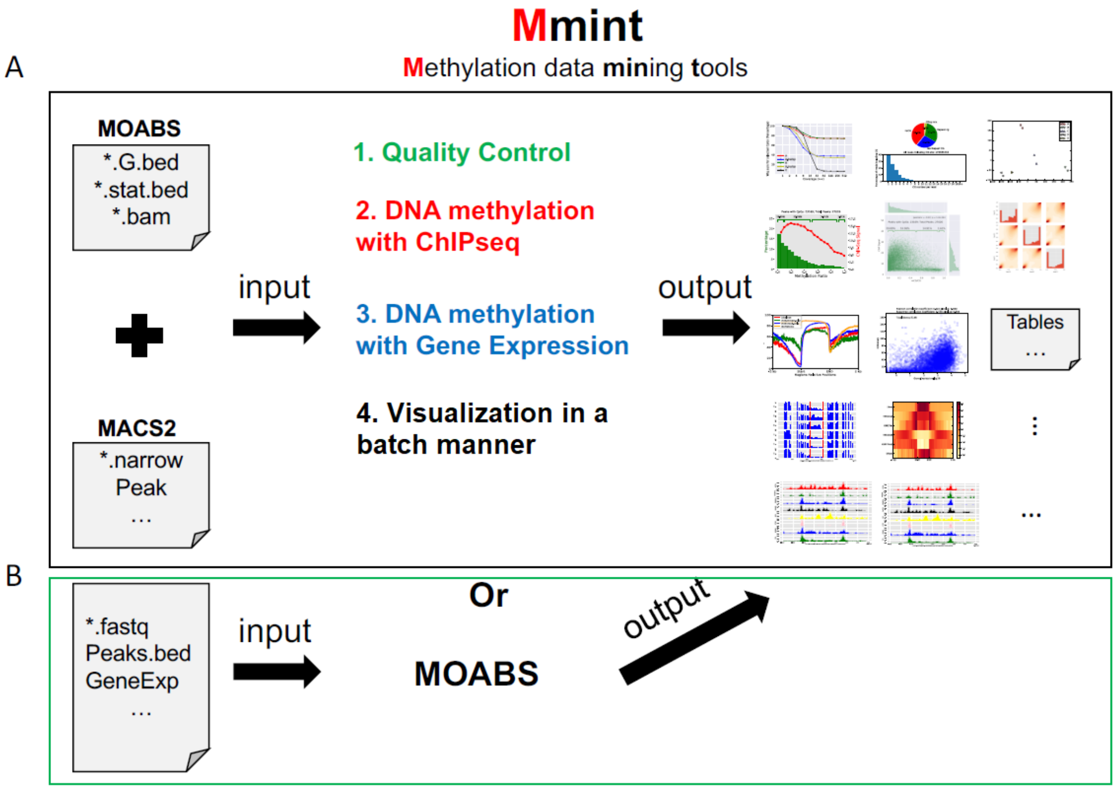

Mmint
Methylation data mining tools
A
MOABS
*.G.bed
*.stat.bed
*.bam
MACS2
*.narrow
Peak
…
input
1. Quality Control
2. DNA methylation with ChIPseq
3. DNA methylation with Gene Expression
4. Visualization in a batch manner
output
Tables
…
B
*.fastq
Peaks.bed
GeneExp
…
input
Or
MOABS
output

Figure 3.1 Overview of Mmint workflow. A. Mmint workflow as an independent tool for analysis. **B.** Mmint workflow as a module of MOABS.

3.3.2. Data quality assessment module in Mmint

Sequencing depth is critical to the DNA methylation analysis, which can be mainly revealed in the relationship among CpG depth, mean methylation ratio and detected CpG numbers. Mmint takes output files from MOABS to generate figures to present the mean methylation ratio and number of CpGs in different depths (Figure 3.2A), which allows users to verify whether there are discrepancies in sequencing depth. Genome coverage and sequencing duplication are another two important indicators for keeping sequencing high-quality, which are also available in Mmint. Figure 3.2B shows the coverage and the percentage of reads at certain coverage. The black curve represents an example with no significant duplication and sequencing bias: There are few reads (<10%) contributing to high coverage (>64) regions. However, in the red or green curve, around 75% of the reads contribute to high coverage regions, which indicates these samples have severe sequence bias or high duplication levels. After deduplication, we can see the percentages of reads in high coverage regions drop significantly (Figure 3.2B). Furthermore, Mmint also focuses on another quality indicator, which is the distribution of CpG number per read. Such indicator represents whether CG bias, one of the potential problems causing bisulfite sequencing bias exists in sequencing. To establish the criteria of the indicator, we simulated the random sequencing process by randomly cutting the human genome to 75bp segments and count the reads with CpG. After the simulation with 100 and 200 million reads, we found around 61.8% of the reads contain CpG (Figure 3.2C). Furthermore, the histogram in Figure 3.2C shows the ratio of reads with different numbers of CpG. Our simulation result shows that about 40% of the reads are including one CGs, which will warn that either the sequencing process or the library preparation has a problem when the deviation is found.

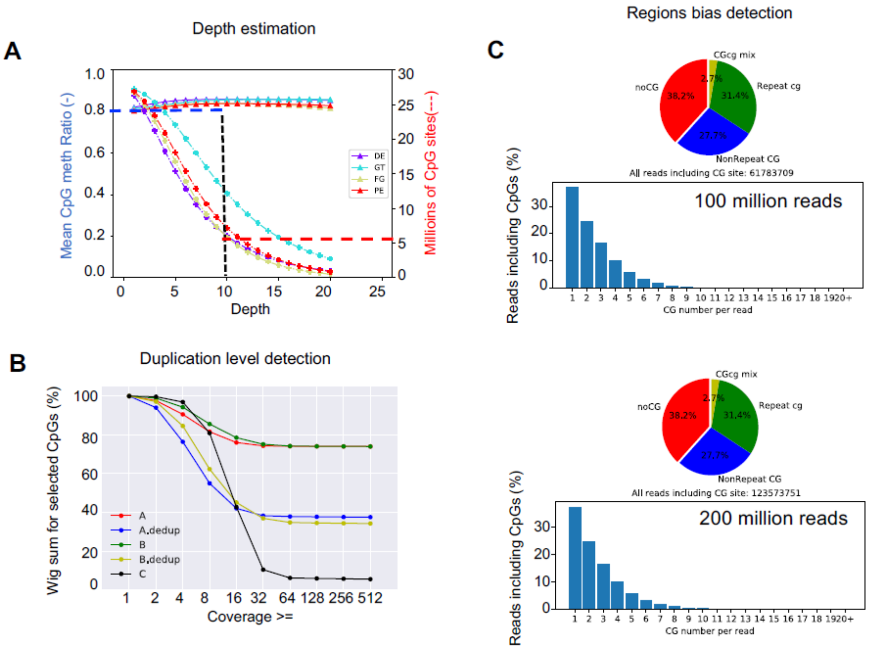


Figure 3.2 Quality assessments for WGBS data. A. Assessment of depth distribution and methylation ratio distribution. X axis represents depth; Y axis in the left represents means DNA methylation ratio under certain depth; Y axis in the right represents the number of CpGs under certain depth. Different colors represent different samples. **B.** Coverage distribution for amplification bias and duplication level estimation. X axis represents the threshold of coverage; Y axis represents the percentage for CpGs have a coverage >= x. **C.** Reads distribution based on whether the reads include CpG sites (simulation data on 100 million / 200 million sequenced reads). Upper panel pie chart shows the ratios of reads with or without CpGs on different locations; The histogram in lower panel shows the percentage of reads that include 1, 2, 3 … CpGs per read.

Other than the single sample quality assessment, Mmint also can accomplish a group-level analysis between samples based on their methylation ratio. By integrating samples methylation ratios with minimum depth limitation, Mmint can apply Principal Component Analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE) and dendrogram cluster analysis to the methylation ratio matrix, which can further reveal the reproducibility of DNA methylation measurements and interrogate the relationship among different groups (Figure 3.3A). To further explore the functionality of Mmint, we used DNA methylation data of four differential stages from H1 embryonic stem cell to pancreatic endoderm (PE) as an example. Figure 3.3A shows the methylation ratio clustering of CpGs with coverage larger than 10 by PCA, in which the samples from the same group are clustered together. The distribution of contributed CpGs in PCA is also recorded by Mmint. In this case, we can estimate that CpGs at promoter, exon, intron and repeat regions have more contribution to the inter-group variations compared with CpGs located in CG islands by counting the overlapping regions with these specific genomic regions (Figure 3.3B). Besides clustering, correlation among biological replicates is another important indicator for the reliability of experiments. In the diagonal of Figure 3.3C, histograms represent the distribution of CG methylation ratios for each sample. The off-diagonal plots are the density plots for pairwise comparisons, from which, users can identify samples with a differently global methylation ratio compared to the others.

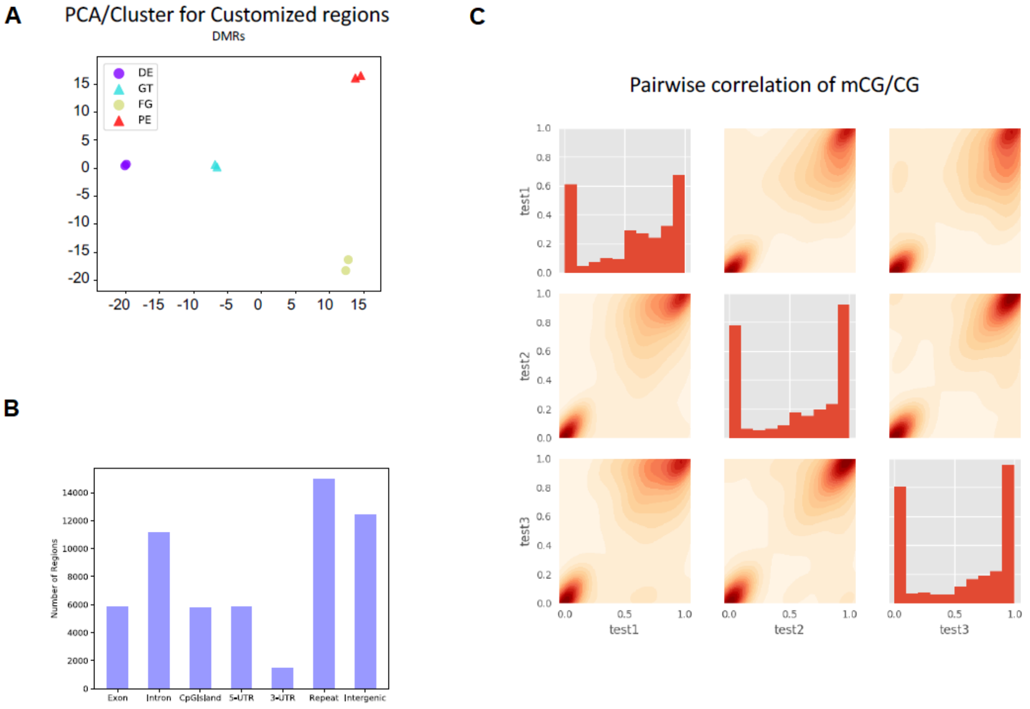


Figure 3.3 Group level quality assessments by Mmint. A. Principal Component Analysis (PCA) on DMRs with CpGs with certain coverage threshold (coverage >= 10). DE, GT, FG and PE are four stages of differentiation from H1 ESC to pancreatic endodermal cell. Each stage has two biological replicates. **B.** Annotation of interested regions in PCA. **C.** Pairwise correlation of DNA methylation ratios among samples. The diagonal histograms represent the distribution of CpGs methylation ratios; The off-diagonal density scatter plot represent the pairwise comparisons of CpGs methylation ratio.

3.3.3. Versatile integrative analysis of multi-omics data by Mmint

To understand the epigenetic mechanism for biological processes such as gene regulation, exploring the interaction between methylation and protein (such as TFs, histone under different modifications) binding sites on the genome is considered as one of the prerequisites for further analysis. To delineate this epigenetic mystery, Meth2ChIP function in Mmint allows users to investigate the distribution of methylation ratio and ChIP-Seq signals on ChIP-Seq peak regions. Figure 3.4A shows the intensity and the signal of ChIP-Seq peaks of different methylation ratio regions (5% as the interval) among all identified peaks. Based on the mean methylation ratio in peaks, we separate the peaks into four categories empirically: Unmethylated Regions (UMR: 0 – 10%), Low Methylated Regions (LMR, 10 – 50%), Mediate Methylated Regions (MMR, 50 – 90%) and High Methylated Regions (HMR, 90 – 100%). The percentages of peaks in these four categories are also marked in the output figure. To interpret our analysis, we take H3K4me3, H3K36me3, H3K4me1 and DNase I (Figure 3.4A, 3.5) data in H1 ESC as examples. As it is shown in Figure 3.4A, 29.89% and 53.58% of the H3K4me3 peaks located in LMR and UMR, these peaks also have higher ChIP-seq signal (red curve) compared with those peaks located in MMR and HMR. The scatter plot also shows that 83.47% (red dash box) of H3K4me3 peaks have an average methylation ratio less than 0.5, the ChIP-seq intensity and average methylation ratio are also negatively correlated (Pearson $r = -0.22$; $p < 0.01$). These observations are consistent with the previous study that H3K4me3 signal marks active promoters with relatively low methylation ratio [81] (Figure 3.4A). For H3K36me3 peaks, 99.07% of them have an average methylation ratio larger than 0.5. This result is consistent with the previous study reporting that H3K36me3 signal marks gene body regions, which have a high methylation level. Moreover, DNase I signal is negatively correlated with DNA methylation (Pearson $r = -0.52$, $p < 0.01$). While there are still 14.59% DNase I peaks have average DNA methylation larger than 0.9. This indicates that 5hmC can be enriched in these peaks. As for H3K4me1, the methylation ratios in H3K4me1 peaks are barely correlated with H3K4me1 ChIP-Seq signals (Figure 3.5).

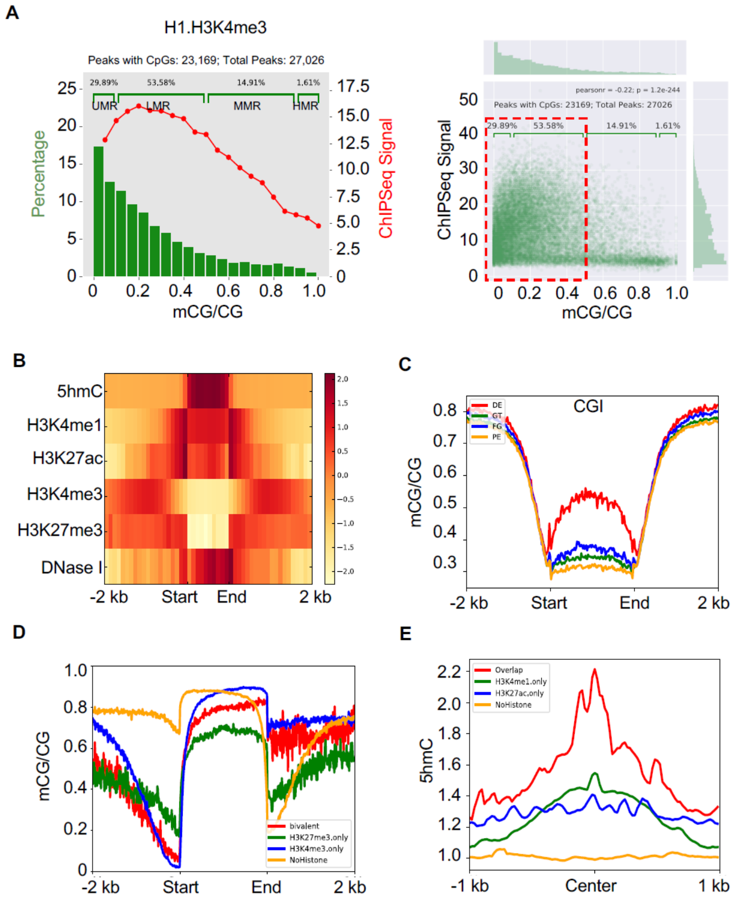


Figure 3.4 Integrative analysis of DNA methylation data with ChIP-seq data. A. Mean DNA methylation ratios in peaks against ChIP-seq intensity distribution (left panel). X axis represents mean DNA methylation in peaks; Left Y axis represents the percentage of peaks at certain mean DNA methylation ratio; Right Y axis represents mean ChIP-seq intensity for those peaks at certain mean methylation ratio. The text on the top of the figure shows the total peaks and the peaks with CpG signals included in the input file. The texts under the top text show the ratios of peaks in each category (Left to right: UMR, LMR, MMR and HMR). The right panel represents the scatter plot by peaks' average DNA methylation ratio against average peaks ChIP-Seq signal. Each dot represents one peak. **B.** Horizontal heatmap for multiple epigenetic signals on interested regions. Each row represents one epigenetic signal and each column represents the signals in one bin with user-defined bin size across interested regions. X axis represents the interested regions (marked by "start" and "end") and upstream/downstream (marked by -2 kb and 2 kb), as the same for the following figures (C, D). The signals are normalized by Z-score in each row. C. DNA methylation distribution on CpG island during the 4 differentiation stages. D. DNA methylation distribution on classified genes based on H3K4me3 and H3K27me3 in transcriptional start site with upstream/downstream 1kb regions. When the interested

region is considered as regions, Mmint will generate result like this example. E. 5hmC distribution on classified peaks based on H3K4me1 and H3K27ac. When the interested region is considered as points, Mmint will generate results like this example.

Besides the output figures, the Meth2ChIP function will also output eight files (regions and annotations for UMRs, IMRs, MMRs and HMRs) for the users' self-usage. In addition, by running Meth2ChIP for different groups, users can compare the distributional difference under different conditions (such as control vs. cancer) to see if the TFs/Histone modification change their dependency on DNA methylation. By utilizing Meth2ChIP, users can understand how DNA methylation ratios correlate with certain epigenetic factors multidimensionally.

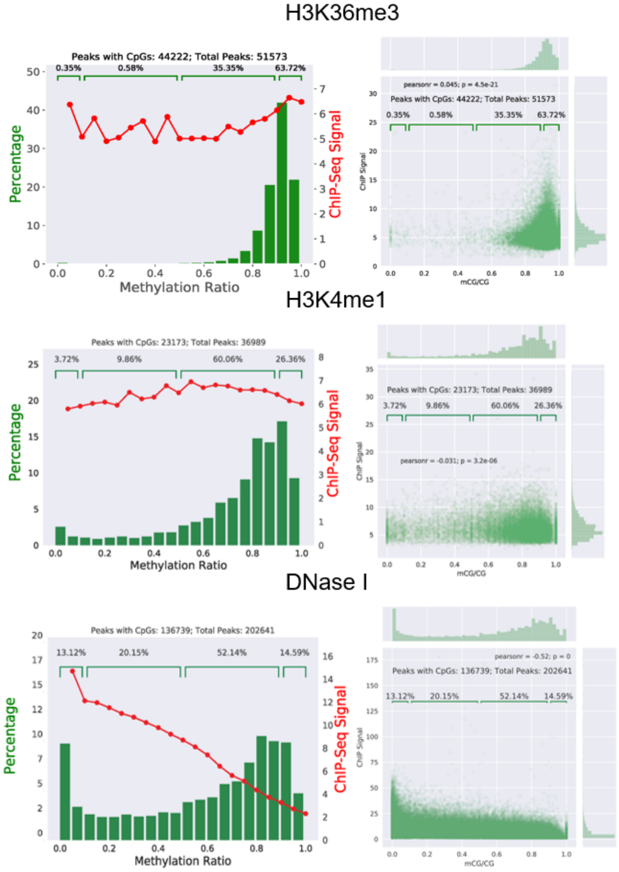


Figure 3.5 DNA methylation in H3K36me3, H3K4me1, DNase I peaks against their intensity.

Another regular task for epigenetic data analysis is to comprehensively visualize multiple signals on certain regions to provide a multi-factorial interaction landscape for the given regions. To fill the gap, we developed HorizontalHeatmap function in Mmint, which generates normalized signal plots to overcome the inconsistency of scales. In Figure 3.4B, each row represents one epigenetic signal. The x axis shows the regions of interest with their upstream/downstream regions (2kb in the example). We can see that in 5hmC enriched regions we selected in H1 embryonic cell, H3K4me1, H3K27ac and DNase I are enriched while H3K4me3 and H3K27me3 are depleted. To investigate the epigenetic signal from another dimension, we developed multiBW2bed and curveDualYaxis to explore the difference of one or more specific signals on certain regions in different groups. As shown in Figure 3.4C, during the four stages (DE, GT, FG and PE) of H1 differentiating to pancreatic endoderm cells, DNA methylation decrease gradually in CpG islands, which indicates that DNA demethylation in CpG island during this process might play pivotal roles. For two signals with different scales, Mmint will automatically use both y axis for visualization.

To further illustrate the interaction among methylation and multiple histone modifications, we developed HistonePeaks and HistoneGenes to comprehensively dissect the complex interactions at peaks and genes level. Based on two histones peaks (For example A, B), Mmint can separate peaks/genes into four categories: A+B+, A+B-, A-B+, A-B- and plot DNA methylation ratios on these four categories peaks/genes (Figure 3.4D, E). As an example, Figure 3.4E shows that 5hmC is the highest enriched on active enhancer regions (coexist H3K4me1 and H3K27ac) which is validated by previous studies [82].

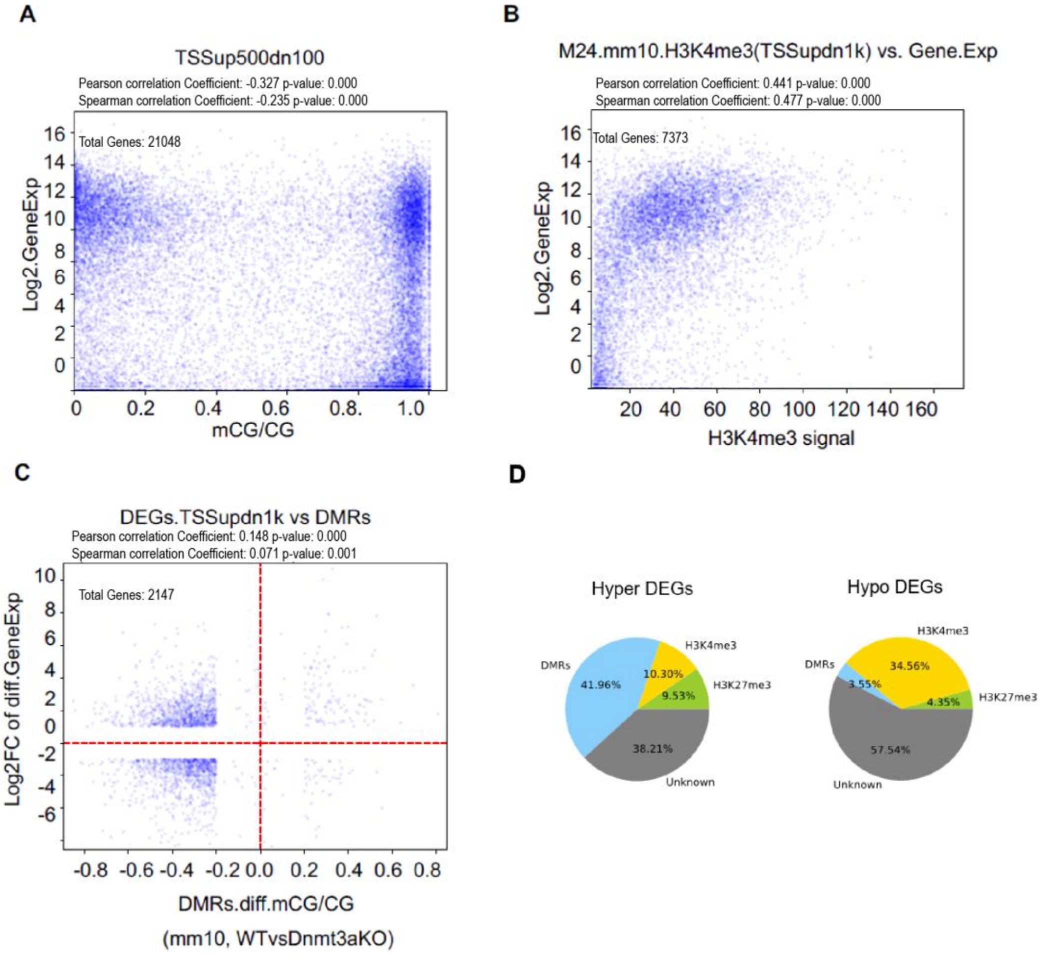


Figure 3.6 Correlation analysis between DNA methylation/histone modification and gene expression. A. Scatter plot for DNA methylation at transcriptional start site with upstream 500bp and downstream 100bp (TSSup500dn100) against gene expression in mouse HSC. **B.** H3K4me3 signals in TSSup1kdn1k regions against gene expression in mouse HSC. **C.** Mean methylation ratio of DMRs located in promoters (TSSup1kdn1k) against gene expression log2Fold Changes; **D.** Pie charts for DEGs explained by histone modifications.

Besides the signal interaction, the relationship between methylation and gene expression level is another central topic in the epigenetic area. Previous studies have shown that DNA methylation in promoter regions negatively regulates gene expression

[83]. However, some genes' changes can be explained by histone modification only and some others are explained by both DNA methylation and histone modification. To reveal the complex relationship between epigenetic signals and gene expression, Mmint can plot a scatter plot of a specific signal in genes' promoter against gene expression level, which can give users a direction of how the signal in promoters influences gene expression. As shown in Figure 3.6A, in mouse hematopoietic stem cell, the DNA methylation in transcriptional start site with upstream 500bp and downstream 100bp (TSSup500dn100) is weakly negative correlated with gene expression (Spearman correlation coefficient = -0.235; $p < 0.01$); however, the H3K4me3 signal in promoter regions is positive correlated with gene expression (Spearman correlation coefficient = 0.477; $p < 0.01$) (Figure 3.6B). Moreover, by plotting the scatterplot for mean methylation ratio of DMRs located in promoter regions against gene expression log2Fold changes, it is clearly showing that most of the DMRs in promoters have a decrease of methylation ratio in Dnmt3a KO mouse hematopoietic stem cell (Figure 3.6C). Among genes with DMRs in their promoter, the change of their expression level is not monotonous, which strongly indicates that DEGs cannot be fully explained by DNA methylation changes. To further investigate the contribution of epigenetic markers in leading to DEGs, Mmint calculates the overlap between DEG's promoter and all differential epigenetic markers between groups. Based on the changing direction of DEG and epigenetic markers, Mmint categorized DEGs into groups by their main influential signal (Figure 3.6D).

During data analysis, another common task for bioinformaticians is to use genome browsers to find proper examples or representing regions. However, all current genome browsers only allow users to view the regions one by one, which may take several hours for users to get a screenshot for publication or summarize a clue for further analysis. To solve this problem, we developed multiTracks to take a list of locations and the bigwig files as input to generate screenshots for each location automatically, which allows users to choose candidate genes or regions for visualization or further study (Figure 3.7A). As an example, we identified 839 DNA methylation canyons in H1 ESC. To characterize the TFs binding in the DNA methylation canyon regions in H1 ESCs, we plotted several TFs signals on selected DNA methylation canyon regions (continuous low methylated regions). CTCF is an architecture protein and RAD21 is a cohesion component complex, Figure 3.7B shows that CTCF (pink) and RAD21 (black) are binding at the same position in the DNA methylation canyon. This result indicates that CTCF and RAD21 co-bind in the “DNA methylation canyon”, which might contribute to the higher order of chromatin. In another user case, to compare the different contexts of the same marker, Figure 3.7C shows that in the gene region, H3K4me3 is decreased in Dnmt3a KO mouse HSC compared with wildtype. These results suggest that multiTracks can screenshot multi-epigenetic factors signal in certain regions effectively and precisely. By using this function, users can easily browse these pictures and simply eye-picking the perfect visualization examples, which significantly simplifies the analysis process, especially for the overview of interested regions.

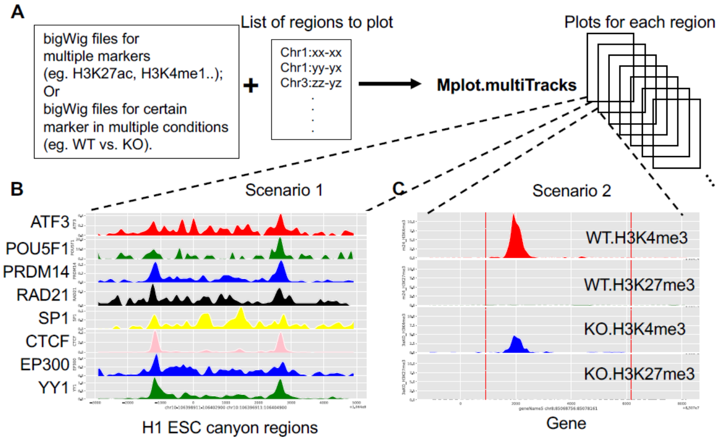


Figure 3.7 Automatically plot multitrack visualization in a batch manner. A. The schematic for multitrack function. **B, C.** The output figure examples generated by Mmint multitrack function.

3.3.4. Case study: Integrative functional analysis of epigenetic signals in H1 embryonic stem cell

To further confirm that Mmint can discover potential novel findings with biological meanings, we performed a detailed analysis on H1 ESC, which has large amounts of epigenetic data contributed by ENCODE project [84]. We downloaded the WGBS data (23,669,925 CpGs have a coverage >= 5), H3K4me1, H3K4me3, H3K27me3, H3K27ac, DNase I data and ten TFs ChIP-seq data from ENCODE official website (https://www.encodeproject.org).

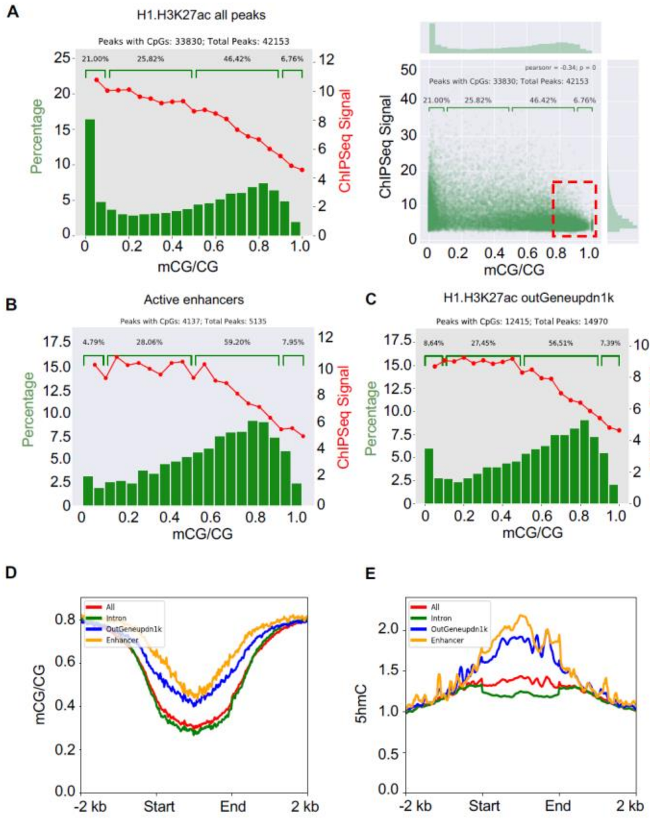


Figure 3.8 Case study for DNA methylation and H3K27ac modification in H1 ESC. A. Mean DNA methylation ratio in H3K27ac peaks regions against mean H3K27ac signals. B. Mean DNA methylation ratio in overlapped regions between H3K4me1 and H3K27ac peaks against mean H3K27ac signals. C. Mean DNA methylation ratio in H3K27ac peaks that located in out of genes upstream and downstream 1kb regions against mean H3K27ac signals. D, E. DNA methylation and 5hmC distribution on All H3K27ac peaks, Intron regions, Out of gene upstream/downstream 1kb regions and enhancer regions.

Figure 3.8A shows the distribution of the average methylation ratio of H3K27ac peaks (total 42,153 peaks, among them, 33,830 peaks include detected CpGs ratio) against H3K27ac peaks' intensity. The H3K27ac peaks' intensity decreases (red dot line) along with DNA methylation increases (Pearson *r* = -0.34; p<0.05). As an active marker, H3K27ac peaks are considered to have a low methylation ratio. Suprisingly, we find that

53.18% of H3K27ac peaks have an average methylation ratio higher than 0.5 in H1 ESC (6.76% of H3K27ac peaks have an average methylation ratio higher than 0.9) (Figure 3.8A). To further explore the locations of H3K27ac peaks with a high methylation ratio, we performed the hotspot analysis on H3K27ac peaks located in specific regions such as gene body, upstream/downstream 1kb of transcriptional starting point, Exons, Introns, CGI and enhancers (Figure 3.9). Except active enhancers and regions excluding upstream/downstream 1kb of genes, all the other regions have a low percentage of peaks with a high methylation ratio (Figure 3.9). As we can see from Figure 3.8B and 3.8C, large portion of H3K27ac peaks located in active enhancers (67.15%) and regions excluding upstream/downstream 1kb of genes (63.9%) have an average methylation ratio higher than 0.5. As the bisulfite treatment cannot distinguish 5mC and 5hmC, we wonder if the H3K27ac peaks with high methylation ratios are enriched for 5hmC. Thus, we include H1 ESC CMS-IP data (GSE97988) for analysis. We firstly used multibw2bed function to plot the DNA methylation ratio and 5hmC level of all interested regions mentioned above. Interestingly, H3K27ac peaks in active enhancers have the highest DNA methylation ratio and the highest 5hmC level, peaks in regions excluding upstream/downstream 1kb of genes have the second-high DNA methylation ratio and second high 5hmC level (Figure 3.8D, E). This result suggests that a high level of 5hmC contributes to the high DNA methylation ratio in H3K27ac peaks located in active enhancers and regions excluding upstream/downstream 1kb of genes.

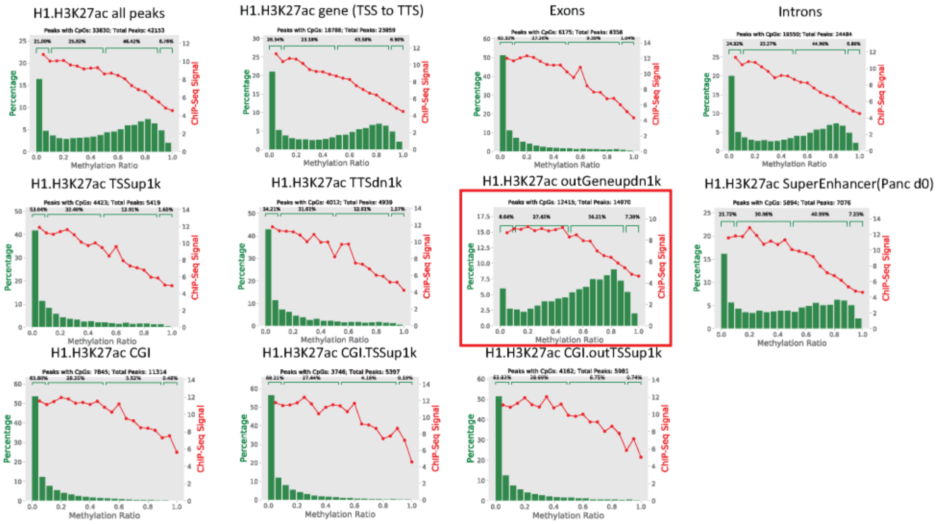


Figure 3.9 DNA methylation in H3K27ac peaks against H3K27ac intensity in various regions in H1 ESC.

We also performed a similar analysis on transcription factors ChIP-Seq data. As shown in Figure 3.10, in EP300, POU5F1 and ATF3 binding regions, about 40% - 50% CpGs have a methylation ratio less than 0.1. And, all of these TFs show the highest ChIP-seq signal in UMRs and the lowest ChIP-seq signal in HMRs. More importantly, in LMRs and MMRs, the ChIP-seq signals do not correlate with methylation ratios. As for CTCF, RAD21, PRDM14 and YY1, they all have a weak negative correlation between ChIP-seq signals and methylation ratio in general. And, in their binding regions, about 11% - 20% peaks are located in HMRs. CTCF and RAD21 are members of the cohesion complex, they both have their highest ChIP-seq signals at LMRs, which is consistent with the study in mice [71]. PRDM14 is a member of PR domain-containing family, and contains multiple zinc fingers. Previous studies show that PRDM14 can interact with polycomb repressive complex 2 (PRC2) and play a critical role in the maintenance and induction of pluripotency [85, 86]. In general, observations about PRDM14 indicates that PRDM14 can either repress methyltransferase through PRC2 or methylation can repress PRDM14s' binding. In YY1 binding regions, as the methylation ratio increases, the ChIP-seq signal of YY1 decreases, which suggests that YY1 binding is sensitive to DNA methylation. Different from the above seven TFs that all prefer to bind in UMRs and LMRs, C/EBP-β, Ring1 and SP1 prefer to bind in HMRs. There are about 70.85% (C/EBP- β), 90.8% (Ring1) and 92.75% (SP1) peaks locate in MMRs and HMRs. For C/EBP- β and SP1, the relatively flat red curve means that the ChIP-seq signals do not correlate with DNA methylation, which is supported by a previous study [87]. These results suggest that the relationship between TFs binding and methylation is complicated rather than a simple linear correlation, which indicates multi-functions of TFs due to their dependency on DNA methylation.

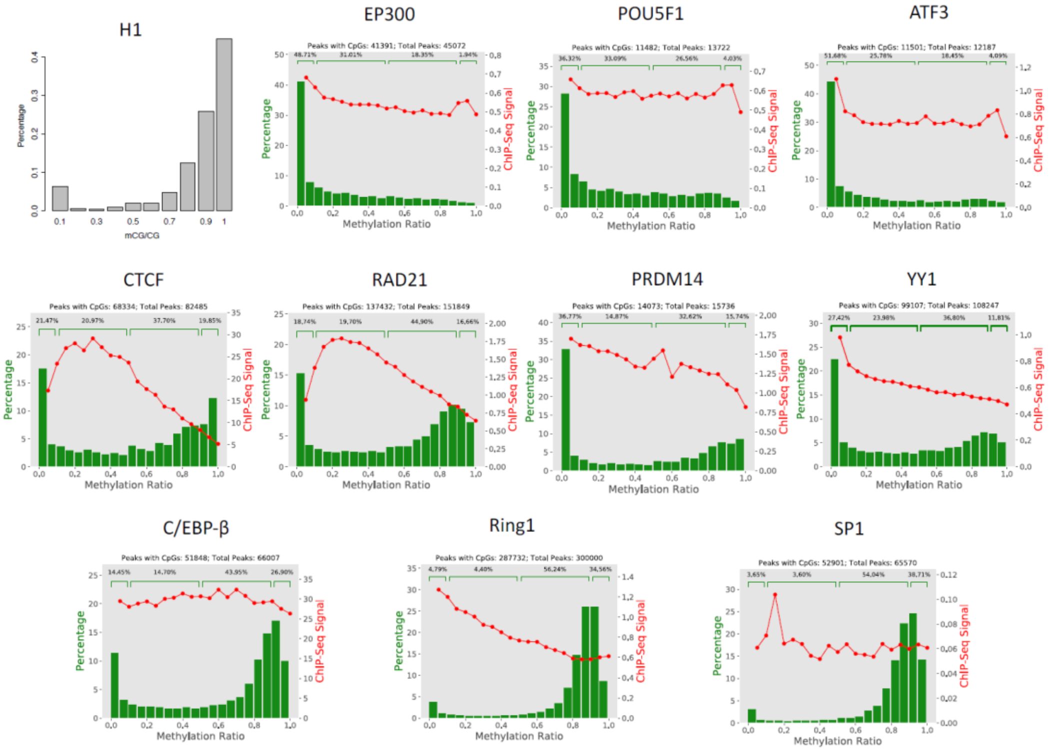


Figure 3.10 DNA methylation in TFs binding regions against TFs binding intensity in H1 ESC.

3.4. Discussion

DNA modifications, including: 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), 5-carboxylcytosine (5caC) are essential epigenetic regulation mechanism for gene expression during mammalian development and disease condition [88-90]. DNA modifications regulate gene expression either through recruiting/inhibiting TFs binding or recruiting/inhibiting histone modification enzyme binding. Most of the studies only focus on one epigenetic mark, ignoring the crosstalk among DNA methylation and other epigenetic markers. Thus, we developed Mmint, a powerful toolkit to perform integrative analysis and to generate publishable visualization among multiple epigenetic markers. Mmint has several advantages over current software: 1. Mmint integrates DNA methylation with other epigenetic markers such as histone modifications, TFs binding and ATAC-seq in a systematical and detailed way. Mmint can category these data into four classes based on their binding regions' DNA methylation level and annotate them with the closest genes, which can associate the combinatory effects of DNA methylation and other epigenetic elements with gene functions; 2. Mmint can directly generates publishable figures from the integrative analysis of methylation and other epigenetic data; 3. Mmint automatically generate multiple "screenshots" (each has multiple tracks) for interested regions with multiple signal tracks at one time.

For bisulfite sequencing data quality control, most of the current tools such as BSeQC [91], BSmooth [92] focus on M-bias, namely the sequencing technical bias, which can produce inaccuracy estimation of methylation ratio in single-base resolution. Mmint, as a complement, provides visualization for other useful quality control, such as sequencing depth, duplication level, statistics for reads including CpGs and PCA. Importantly, the statistic for reads including the number of CpGs can reflect the global bias for sequencing and the location of biases. In addition, PCA and correlation analysis can explore the similarity among biological replicates which can suggest the potential outliers and biases in sample preparation stages.

Currently, there is no visualization tool that can be used to study the global relationship between DNA methylation data and ChIP-Seq data (TFs binding and histone modifications). In a traditional view, DNA methylation inhibits TFs binding in general, which result in disturbs in the certain biological context. Recently, a systematic analysis showed that the DNA methylation could be the negative, positive or mixed relationship to TFs bindings [73]. Specifically, some TFs can interact with methylated DNA without methyl-CpG binding domain (MBD); and some other TFs bind at new regions in certain conditions [93]. Thus, DNA methylation status and the structure of DNA sequence both influence the TFs binding. It is possible that in pathological conditions, such as cancer, the preference for DNA methylation in the TF binding process may be completely different from normal tissues. In addition, DNA methylation also shows a complex relationship with histone modifications. For example, in gene body regions, co-deposit of DNMT3B dependent DNA methylation and SETD2 mediated H3K36me3 prevents spurious transcription initiation [94, 95]. This indicates crosstalk among SetD2, H3K36me3, Dnmt3b and DNA methylation. For both TF binding region detection and histone modification identification, ChIP-Seq is the most popular technique to get the information genome widely. Mmint not only can provide a general overview of the relationship between DNA methylation data and ChIP-Seq data, but also can dissect the details of the crosstalk.

4. GSMPLOT: A WEB SERVER TO VISUALIZE EPIGENOMIC DATA*

4.1. Background

Epigenetic mechanisms alter phenotypes by regulating gene expression patterns without altering the DNA sequences in response to physiological or pathological signals [96]. Due to the high-throughput sequencing technological advances, such as chromatin immunoprecipitation sequencing (ChIP-seq), whole genome-wide sodium bisulfite sequencing (WGBS) [97], anti-CMS immunoprecipitation (CMS-IP)-seq [68] and ATAC-seq [98], an extremely large amount of epigenomic data has been collected and published. Epigenetic factors including histone modifications, TFs bindings, DNA modifications and chromatin accessibilities, always dynamically interact with each other to shape the epigenomic landscape specifically to certain biological processes [75, 99, 100]. Therefore, it is important to compare different epigenetic factors visually from different studies (public data) to ensure a properly comprehensive interpretation. NCBI Gene Expression Omnibus [101, 102] is a primary data source for a high-throughput sequencing data repository, which includes epigenetic data generated from various species, cell types, diseases and experimental conditions. In GEO, every dataset has a GSM ID, which includes process files in multiple formats, such as wig, bigwig and bedgraph. BigWig files are compressed binary indexed files containing genome wide data signals at various resolutions [103], which are easier to manipulate, especially for large size files.

* Reprinted with permission from "GsmPlot: a web server to visualize epigenome data in NCBI" by Jia Li; Yue Yin, 2020. BMC Bioinformatics, 21, 55, Copyright [2020] by Springer Nature.

Although DaVIE [104], Octopus-toolkit [105] and EpiMINE [106] provide visualization of public data, they require to install some necessary software to user's computer and require extensive knowledge of the pipeline from researchers to run the software and analyze the epigenetic data. Both WashU epigenome browser [107] and UCSC genome browser [108] are excellent epigenome data browsers, which allow users to upload bigwig files to visualize data in a "scan" genome manner. However, users are required to set up public URLs for their data which need bioinformatics expertise. Many researchers in the biomedical field do not have bioinformatics expertise and high-performance computer resources to analyze, reform and visualize the public data. Currently, there are no user-friendly tools with convenient visualization function available for non-bioinformatic researchers that do not require any complicated installation step and processing next-generation sequencing data.

To alleviate these limitations, we developed GsmPlot, a user-friendly webserver to easily generate customized visualizations for the public data in GEO and additionally provide interactive explorations. GsmPlot is convenient to use as it needs only GSM IDs or the bigwig file provided by users. GsmPlot can conveniently generate profile plots on functional genome elements (gene, promoter, exon, intron, or any regions defined by user) or visualization on one specifically interested region through UCSC genome browser integration. Moreover, GsmPlot allows interactive selection of regions with specific epigenetic patterns in the heatmap for further explorative study.

4.2. Methods

4.2.1. Components of GsmPlot

GsmPlot server is composed of three parts: web crawler, data process and web interface. (1). Web crawler was coded in Python 3.5 and specifically designed for NCBI to automatically detect the URLs and download files with bigwig, wig and BedGraph format. We also include genome reference version check in web crawler. The data process includes two parts: calculation and visualizations. (2). For data calculation, we wrapped deepTools to calculate the average bigwig signal in bins of user-defined size along concerned regions [78]. A matrix of average bigwig signal with rows as regions and columns as bins are generated, and the column mean values are plotted as aggregated profiles. By transforming the wig signal to z-score, we also plot all the z-scores in one bin as a boxplot and so for all bins. For the z-score matrix, based on each row's z-score standard deviation, the top 5 k most variable regions among all samples were chosen to plot the heatmap. Users can choose regions based on the heatmap patterns to replot and download the selected regions to do further study. For data visualization, we use in-house scripts coded by Python 3.5 (Matplotlib, https://matplotlib.org/) and R (https://www.r-project.org/). (3). GsmPlot web interface is implemented using HTML, CSS (bootstrap, http://getbootstrap.com/2.3.2/), and JavaScript. The backend of GsmPlot is based on Django web framework (https://www.djangoproject.com/). The interactive functions between users and GsmPlot web server are implemented using jQuery (https://jquery.com). For large data which takes a long time to finish the calculation, we include an email alert function by using django.cor.mail function. Due to the limited computing resources, we currently only allow one task for each user at a time. GsmPlot has been tested in Firefox, Chrome, Safari, and Edge.

4.2.2. Flowchart of GsmPlot

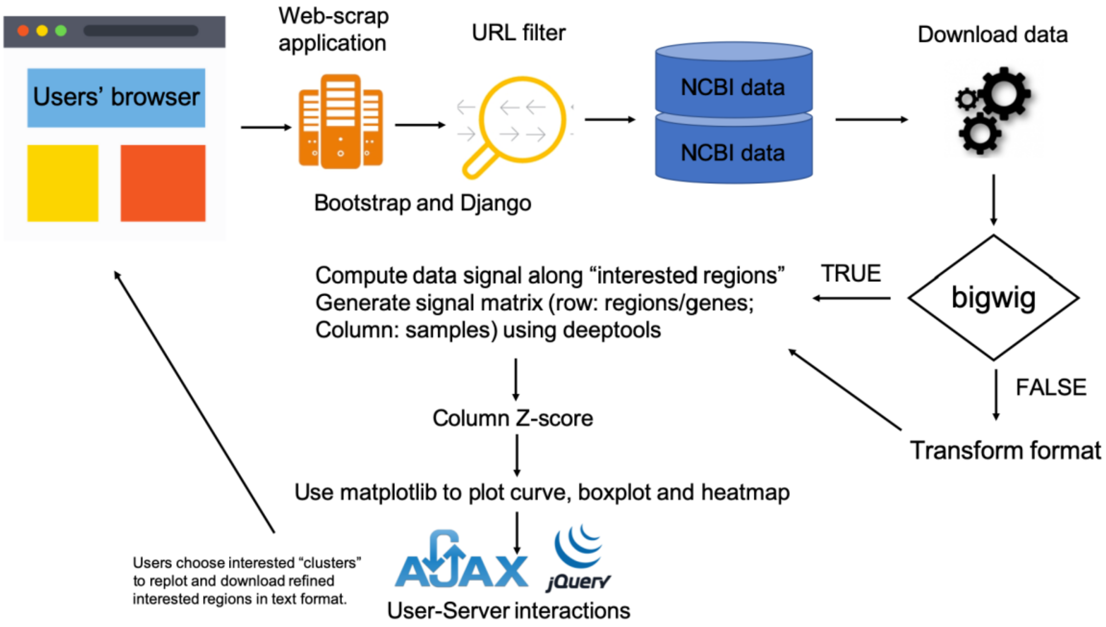

Users' browser
Web-scrap application
URL filter
NCBI data
NCBI data
Download data
Bootstrap and Django
bigwig
TRUE
FALSE
Transform format
Compute data signal along "interested regions"
Generate signal matrix (row: regions/genes;
Column: samples) using deeptools
Column Z-score
Use matplotlib to plot curve, boxplot and heatmap
Users choose interested "clusters"
to replot and download refined
interested regions in text format.
AJAX
jQuery
User-Server interactions

Figure 4.1 Scheme for the structure of GsmPlot web server.

The flowchart of GsmPlot is in Figure 4.1. GsmPlot web server friendly accepts GSM IDs or user uploaded bigwig files as input. If the input is a GSM ID, the web crawler will search NCBI websites to locate bigwig files and automatically download the files. At the same time, web crawler will also try to collect the genome reference version information to double check user input information. If the file format is Wig or BedGraph, GsmPlot will automatically transform them to BigWig format. After downloading the files, wrapped deepTools will calculate the average signals on user provided genome regions according to user provided bin size. The downloaded files will be stored in GsmPlot server for 72 h from last access, which will save the downloading time when users reuse this data frequently. If the input files are uploaded by users, GsmPlot will directly proceed to calculation and visualization. “Reference check” function will aid users to choose the right reference version by collecting the reference information from the NCBI website. Users can select regions with specific epigenetic patterns in the heatmap. Genomic coordinates of these selected regions can be downloaded in text format which could be further studied.

4.3. Results

4.3.1. Overview of GsmPlot

GsmPlot provides two flexible methods for the user to query the data: GSM IDs or bigwig files on the user’s computer. GsmPlot automatically downloads the bigwig/wig/bedgraph file from GEO or from the user computer to the webserver. Users can profile the data along with user-defined genome intervals by providing BED files or gene sets by providing gene names (Figure 4.1). There is no limit on the number of GSM IDs or number of BigWig files, meaning GsmPlot can easily draw RNA-Seq, ChIP-Seq, ATAC-Seq, Bis-Seq or any other type of sequencing data altogether in one plot. We found that more than 65% of ChIP-seq, ATAC-seq and Bisulfite-seq datasets stored in GEO have bigwig, wig or bedgraph files available (Table 4.1), making GsmPlot a significant tool to revisit these large number of datasets in NCBI. Moreover, GsmPlot can automatically perform reference genome sanity check and lift over genome versions whenever necessary to correctly utilize all the data stored in NCBI for the past decades with different genome versions. With the same datasets and same plot setting, GsmPlot is relatively fast in our tests for typical datasets in GEO (Table 4.2, 4.3).

Table 4.1 Statistics for datasets with bigwig/wig/bedgraph files available in GEO.

Organism	Method	Total datasets	GsmPlot usable datasets	Ratio
Human	ATAC-seq	4,010	1,759	43.87%
	ChIP-seq	20,640	17,390	84.25%
	Bisulfite-seq	521	351	67.37%
	Total	25,171	19,500	65.16%
Mouse	ATAC-seq	2,921	1,677	57.41%
	ChIP-seq	11,666	8,540	73.20%
	Bisulfite-seq	552	359	65.04%
	Total	15,139	10,576	65.22%

Table 4.2 Processing time for variable files sizes.

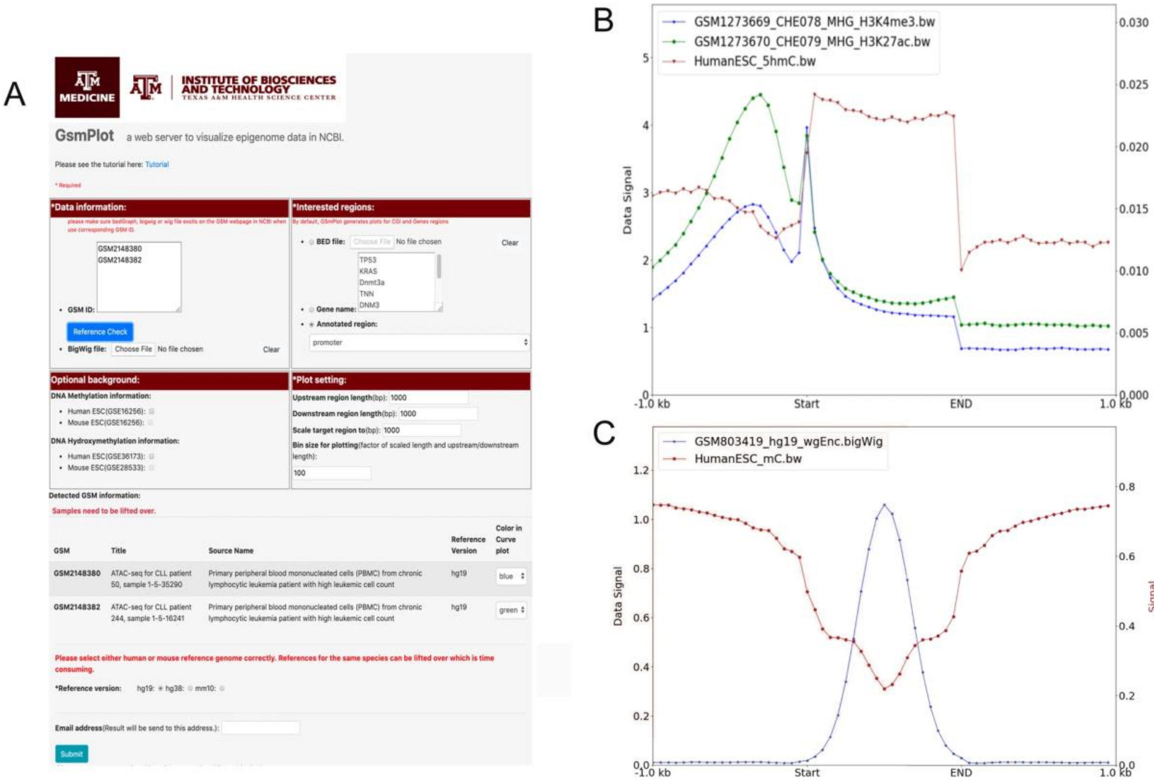
With default GsmPlot setting			
Number of Files	Average file size (Mb)	Total Processing Time	GSM IDs
1	463	2 m 30 s	GSM2535467
1	917	5 m	GSM2535465
2	350	8 m	GSM935297; GSM935299
2	500	8 m 25 s	GSM935305; GSM935282
3	433	9 m 21 s	GSM935282; GSM935298; GSM935301
3	225	11 m 21 s	GSM3073977; GSM3073950; GSM3073962
4	160	14 m 24 s	GSM3444438; GSM3444440; GSM3444437; GSM3444439
5	552	24 m 28 s	GSM2781481; GSM2535467; GSM2535470; GSM2535468; GSM2535464
6	120	20 m 53 s	GSM3073980; GSM3073953; GSM3073981; GSM3073954; GSM3073984; GSM3073966

Table 4.3 Processing time of GsmPlot and EpiMINE.

Reads_number	400 k	2,000 k	4,000 k	8,000 k	9,671 k
Wig file size (MB)	32	115	195	332	353
Bam file size (MB)	23	109	214	419	440
GsmPlot	2 m 31 s	2 m 57 s	3 m 10 s	3 m 26 s	3 m 30 s
EpiMINE	7 m 39 s	7 m 55 s	10 m 20 s	15 m 3 s	15 m 43 s

Furthermore, GsmPlot embedded the public DNA methylation (5mC) and hydroxymethylation (5hmC) data for human and mouse ES cells [109-111]. Therefore, researchers can visualize the 5mC or 5hmC distribution on concerned transcription factor (TF) binding regions, histone modification regions, or any other concerned regions, looking for clues about how DNA modification interacts with TFs, histones, and so on. In addition, the co-binding of TFs is an important gene regulatory mechanism [112]. GsmPlot can also be used to study the co-binding of two or more TFs by integrating the public ChIP-seq data (such as Cistrome [113] and ENCODE database) and the user-provided ChIP-seq data. Such integration of DNA methylation, hydroxymethylation, and TF binding data is extremely useful in terms of interpreting the regulation functions of epigenetic factors. Most importantly, GsmPlot integrated the UCSC genome browser visualization at the end of the analysis pipeline so users can browse to specific genomic locations to visualize these data signals.

4.3.2. Convenient visualization and efficient exploration of public epigenetic data by GsmPlot



GsmPlot a web server to visualize epigenome data in NCBI.

Please see the tutorial here: Tutorial

* Required

***Data information:**

please make sure bedGraph, bigwig or wig file exsits on the GSM webpage in NCBI when use corresponding GSM ID.

- **GSM ID:** GSM2148380
GSM2148382
- **BigWig file:** Choose File No file chosen

***Interested regions:**

By default, GSmPlot generates plots for CGI and Genes regions

- ○ **BED file:** Choose File No file chosen
- ○ **Gene name:** TP53 KRAS Dnmt3a TNN DNM3
- ● **Annotated region:** promoter

Optional background:

DNA Methylation information:

- Human ESC(GSE16256): □
- Mouse ESC(GSE16256): □

DNA Hydroxymethylation information:

- Human ESC(GSE36173): □
- Mouse ESC(GSE28533): □

***Plot setting:**

Upstream region length(bp): 1000

Downstream region length(bp): 1000

Scale target region to(bp): 1000

Bin size for plotting(factor of scaled length and upstream/downstream length): 100

Detected GSM information:

Samples need to be lifted over.

GSM	Title	Source Name	Reference Version	Color in Curve plot
GSM2148380	ATAC-seq for CLL patient 50, sample 1-5-35290	Primary peripheral blood mononucleated cells (PBMC) from chronic lymphocytic leukemia patient with high leukemic cell count	hg19	blue
GSM2148382	ATAC-seq for CLL patient 244, sample 1-5-16241	Primary peripheral blood mononucleated cells (PBMC) from chronic lymphocytic leukemia patient with high leukemic cell count	hg19	green

Please select either human or mouse reference genome correctly. References for the same species can be lifted over which is time consuming.

***Reference version:** hg19: ● hg38: ○ mm10: ○

Email address(Result will be send to this address.):

Figure 4.2 Overview of GsmPlot. A. GsmPlot website interface. **B.** Average H3K27Ac (blue curve), H3K4me3 (red curve), and 5hmC (green curve) ChIP-Seq signals along genes. **C.** Average CTCF ChIP-Seq signal and DNA methylation Bis-Seq signal along CTCF binding sites.

Figure 4.2A shows an example using GsmPlot to investigate the crosstalk between histone modification and DNA methylation. We entered GSM1273669 (H3K4me3 ChIP-Seq) and GSM1273670 (H3K27ac ChIP-Seq) in the "Data information" box and selected "Human ESC" for 5hmC information. We optionally plot the 1000 bases upstream and downstream of the selected regions, and scale all target regions to be 1000 bases. We also set the bin size to be 50 bases to get high-resolution curves. In the result, the blue and green curves in Fig. 1b indicated that the average signal of H3K4me3 and H3K27ac are highly enriched around promoter regions with double peaks, consistent with a previous study [114] and the 5hmC signal is enriched in gene body regions. In an example region shown in the UCSC genome browser in Figure 4.3, the H3K4me3 and H3K27Ac peaks are well aligned with gene promoters. This example confirmed that our program is correct and efficient.

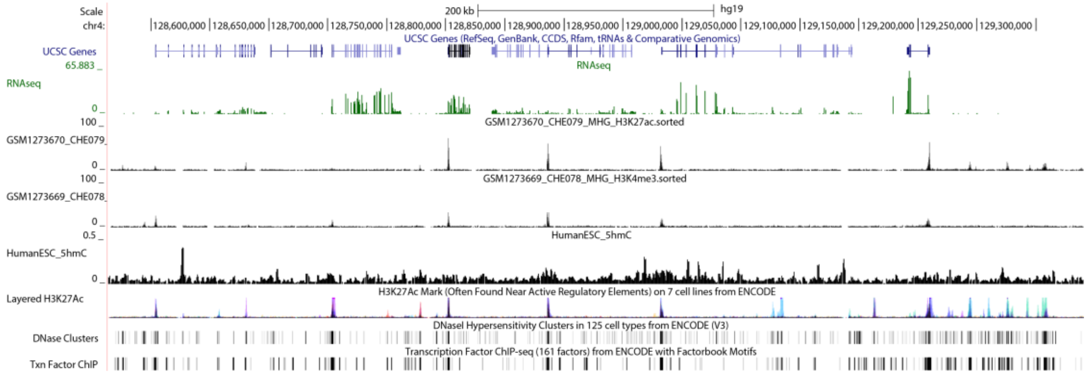

Scale
200 kb
hg19
chr4:
128,600,000
128,650,000
128,700,000
128,750,000
128,800,000
128,850,000
128,900,000
128,950,000
129,000,000
129,050,000
129,100,000
129,150,000
129,200,000
129,250,000
129,300,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
UCSC Genes
65.883 _
RNAseq
RNAseq
0 _
100 _
GSM1273670_CHE079_MHG_H3K27ac.sorted
GSM1273670_CHE079_
0 _
100 _
GSM1273669_CHE078_MHG_H3K4me3.sorted
GSM1273669_CHE078_
0 _
0.5 _
HumanESC_5hmC
HumanESC_5hmC
0 _
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
DNase Clusters
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP

Figure 4.3 UCSC genome browser visualization for RNA-Seq, H3K27Ac, H3K4me3 on an example region for human H1 ESC.

GsmPlot also can be used to investigate the relationship between TFs and DNA methylation or hydroxymethylation. Figure 4.4A shows that the CTCF binding regions in hESC downloaded from GSM803419 generally have a depletion of 5mC but accompanied by complex DNA 5hmC distribution. In the center of the CTCF peak regions, we could observe depletion of the 5mC signal (Figure 4.4B). This result is also consistent with a previous study [115], proving again that GsmPlot can process and plot multiple signals correctly.

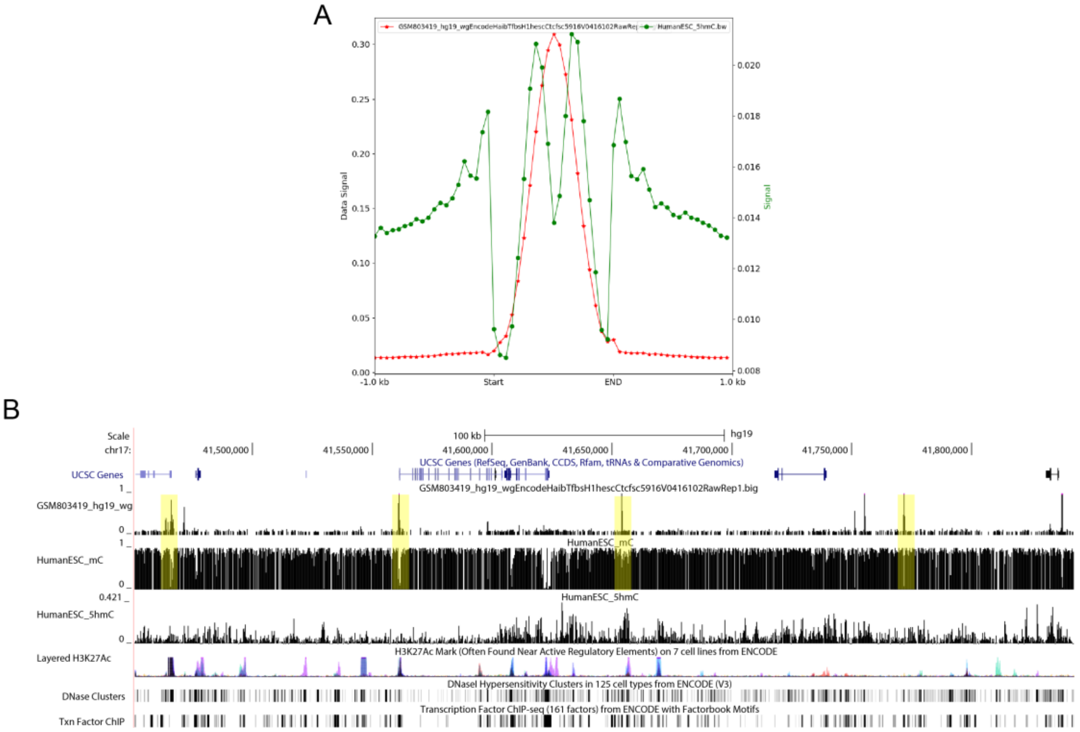


Figure 4.4 Methylation distribution in CTCF binding region of hESC. A. The average ChIP-Seq signal along the CTCF binding sites with red curve for the CTCF signal and blue curve for the 5hmc signal. **B.** The UCSC genome browser visualization for CTCF peak, DNA methylation and DNA hydroxymethylation on an example region. The yellow highlight areas showed the depletion of 5mC at the center of the CTCF peak.

Epigenetic data from different sources are usually generated and normalized differently, preventing such data to be compared directly. To circumvent this problem, we can use z-scores to replace the raw wig signals to allow direct comparison. For each sample, we calculate the average bigwig signal in bins of user-defined size along concerned regions. Then, we calculate z-scores of the corresponding wig values for each bin in each region (Figure 4.6). In the example illustrated by Figure 4.5A and Figure 4.7, we plotted the aggregated profiles on the upper panel and the z-score boxplots on the lower panel for H3K4me3, H3K27ac and H3K27me3 (GSM3444436, GSM3444438 and

GSM3444439) in glioblastoma tissue. From both the average wig profiles and the z-score boxplots, we can see the enrichment of H3K4me3 and H3K27Ac but not H3K27me3 on the selected TSS and CGI regions, and no enrichment on the gene body regions. Furthermore, as a unique feature of GsmPlot, we developed an interactive heatmap to aid users to explore the potentially interesting regions enriched with epigenetic factors. We choose the top 5 k (by default) most variable regions among all samples to plot the heatmap (Figure 4.5B). Cluster 1 represents active genes with both H3K4me3 and H3K27ac enriched in the promoter and cluster 2 represents repressed genes with H3K27me3 enriched in the promoter. Users can slide the sidebar of the heatmap to select the regions with specific patterns. The z-score boxplot for these selected regions will be re-plotted. And the genomic locations of these selected regions can be downloaded as a text file for further study. For example, users can upload this file to GsmPlot as concerned regions to investigate how epigenetic factors distribute on this specific set of regions.

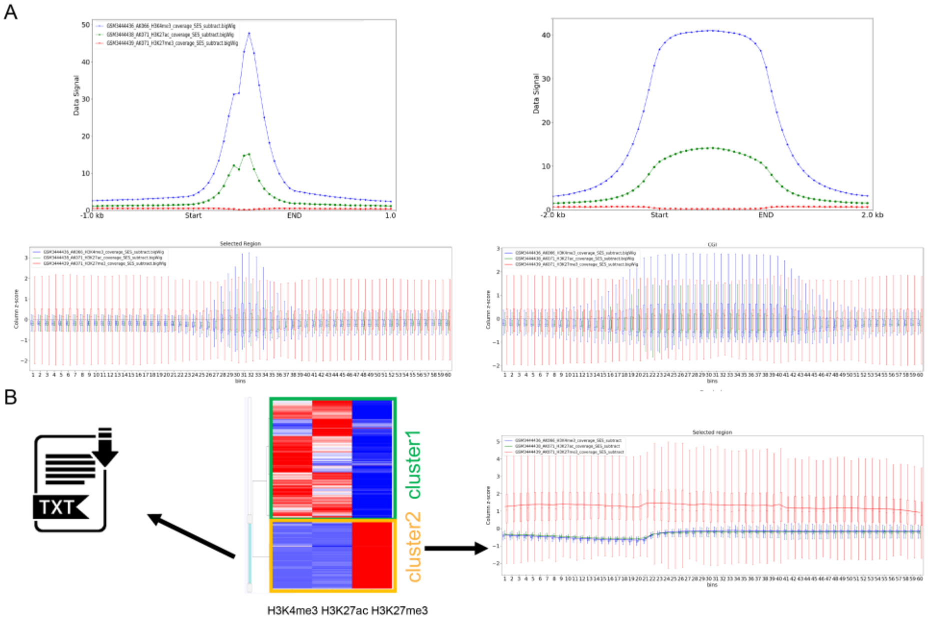


Figure 4.5 Analysis of GsmPlot results. A. GsmPlot default figures for the average signal curve (upper) and the z-score boxplots (lower) along TSS (left) and CpG Island (right) regions. Blue: H3K4me3; Green: H3K27ac; Red: H3K27me3. **B.** GsmPlot interactive heatmap allowing users to choose specific regions to dynamically plot column z-score boxplot and download the selections.

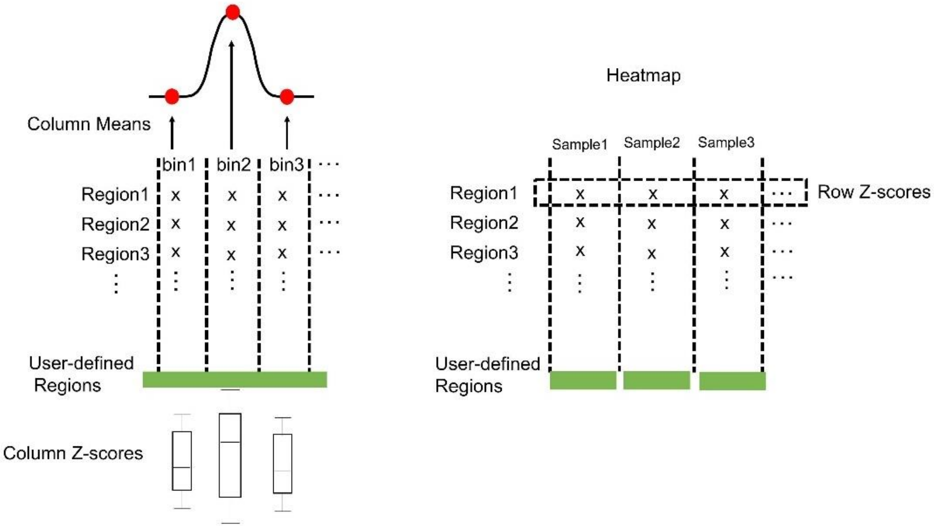


Figure 4.6 Illustration of the data matrix for the profile curve and the z-score boxplots (left), and illustration of the data matrix for the heatmap (right).

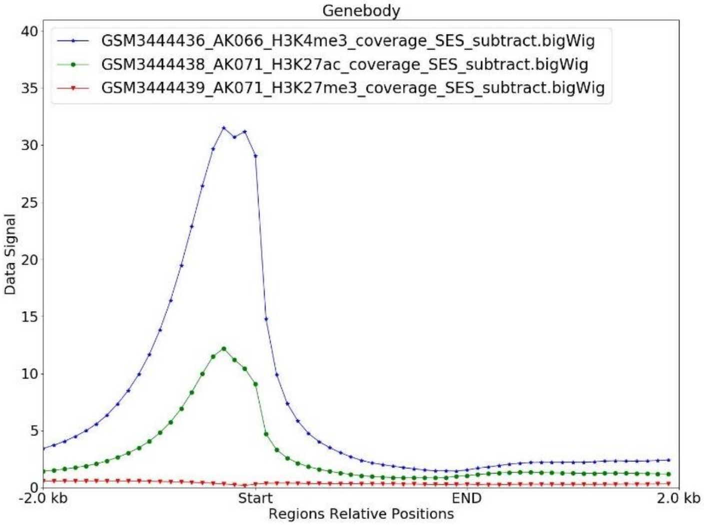


Figure 4.7 GsmPlot default figures for the average signal curve along the gene body regions. Blue: H3K4me3; Green: H3K27ac; Red: H3K27me3.

4.4. Case study: Investigate roles of DNA hydroxymethylation around CGI regions in heart development

As an example, to illustrate that GsmPlot has the potential to shape novel biological hypotheses or discoveries, we explored the potential roles of DNA hydroxymethylation (5hmC) around CGI regions in heart development. We used mouse heart DNA hydroxymethylation data (CMS-IP) from wildtype (GSM3466904) and Tet2/3 knockout (GSM3466906) mice [116]. We also included mouse heart ChIP-seq (GSM3597759) data for Isl1, which is a cardiac progenitor marker gene and is important for heart development [117, 118]. Our GsmPlot results showed that in CGIs with a single transcriptional direction, 5hmC exhibits an unbalanced and directional distribution pattern (Figure 4.8A). On the contrary, the 5hmC level is symmetric on upstream and downstream of CGIs with dual transcriptional directions. Dramatically decreased 5hmC level in Tet2/3 KO mouse hearts are observed in both CGIs with single and dual transcriptional directions (Figure 4.8B). Moreover, Isl1 binding intensity is higher at CGIs with a single transcriptional direction than dual directional transcriptions (dash black line). These results indicate that 5hmC may play different roles in terms of how heart related TFs bind to CGIs with single or dual transcriptional directions.

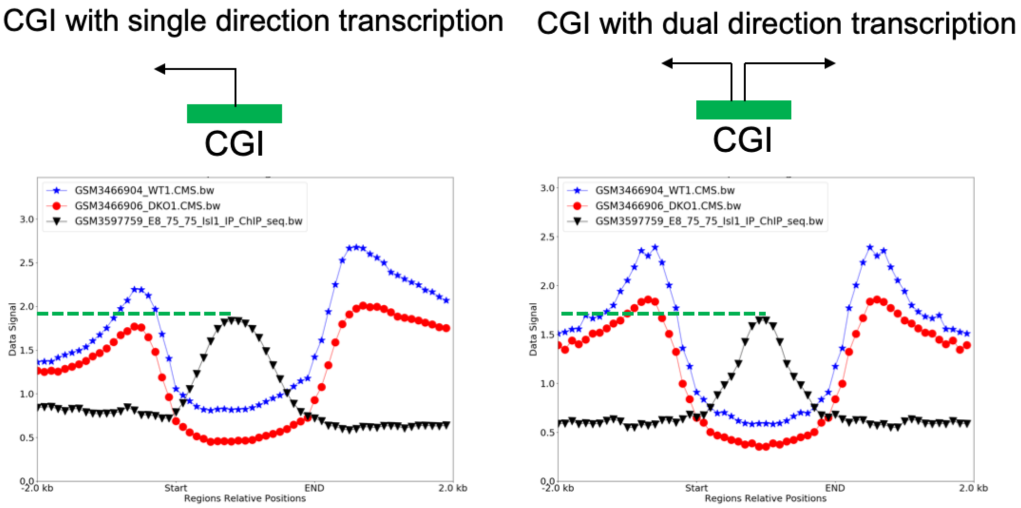


Figure 4.8 Investigate roles of DNA hydroxymethylation around CGI regions in heart development. A. 5hmC signal distribution around CGIs with single transcriptional direction; **B.** 5hmC signal distribution around CGIs with dual transcriptional directions. Blue: 5hmC signal from WT mouse heart; red: 5hmC from Tet2/3 knockout mouse heart; black: mouse heart Isl1 ChIP-seq signal.

4.5. Discussion

Biomedical data stored in NCBI is valuable for biomedical researchers. However, most researchers and physicians do not have computation skills or infrastructure, and hence this “treasure” could not be used immediately. Even for bioinformaticians, complicated procedures including download, computation, aggregation, hosting of data are required to visualize NCBI data. We developed a web server, GsmPlot, which can download, compute, visualize and compare data. The most important feature of GsmPlot is the ability to perform multiple omics integration studies, such as RNA-seq, Bis-seq, ChIP-seq, ATAC-seq with simply GSM IDs from NCBI. Private data sequenced by users in proper visualization format can be fed into GsmPlot to compare with public data. Compared with other good epigenome analysis platforms, such as EpiMINE, GsmPlot have many advantages. GsmPlot does not need users to download public data manually and does not depend on users' computer capacity especially for computation intensive bis-seq data, which cannot be handled on a desktop computer. In addition, installation problems, such as software compatibility, software version, could be a big headache for many researchers, but can be completely avoided using GsmPlot. Moreover, interesting regions with certain epigenetic features can be extracted using an interactive heatmap, which can be fed into GsmPlot again to explore if there are new epigenetic factors in these interesting regions. Importantly, we have successfully proved GsmPlot's reliability and its potential ability to make novel biological ideas from three case studies. Above all, GsmPlot is a user-friendly and reliable tool to investigate public epigenetic data, especially for those biomedical researchers who do not have any computation skills.

Although GsmPlot has an email alert for those large data tasks, GsmPlot will add more CPUs to further improve the speed of calculation in the future depending on the demand. The figure's format, label sizes and other features will be added as user options which will allow users to generate publication quality figures using GsmPlot.

5. CONCLUSIONS AND FUTURE PLAN

In this dissertation, we developed a toolkit to perform comprehensive data analysis for low-input bisulfite sequencing.

In section 2, we proposed an ultrasensitive alignment augmentation for low-input bisulfite sequencing (LiBis) on the idea of segmenting unmapped reads by sliding window. Using FASTQ files as input, LiBis outputs a well-organized HTML report of results from the different modules. LiBis introduces a novel strategy to rescue initially unmapped reads and thereby significantly improved the cost efficiency of analyzing low-input bisulfite sequencing data by providing a larger number of accurate informative CpGs and increasing the sequencing depth of all CpGs for downstream analysis, which also makes the integration of WGBS and liquid biopsy very promising. We have shown this improvement by performing WGBS analysis together with LiBis on cerebrospinal fluid cfDNA. Moreover, LiBis uses a conservative rescue strategy and consequently accuracy is improved. The number of informative CpGs and the overall sequencing depth were both increased according to the simulation and the public data with both bulk and single-cell WGBS experiments. The average correlation coefficient of methylomes between first round BSMAP mapping and LiBis rescued mapping is 0.88 (0.84-0.92). Increased depth and accurate methylation measurements are both crucial for downstream biomarker identification.

In section 3, we proposed Mmint, an integrative toolkit to perform DNA methylation centered epigenetic analysis. By integrating analytical knowledge and data visualization, Mmint can accomplish multiple analysis tasks such as data quality assessment, clustering analysis, correlation analysis, interest region exploration and multiple signal crosstalk analysis. Mmint generates publishable figures as results, which requires less bioinformatics background and skills to interpret the biological meaning and identify the direction of further research. Mmint flattens the learning curve of epigenetics data analysis and facilitates the investigation of epigenetics mechanisms through integrating DNA methylation and other epigenetic markers. Mmint also provides huge potential for the novel epigenetic mechanisms' discovery, which will deepen our understanding of the epigenetic signals in biological processes.

In section 4, we proposed GsmPlot, which is a user-friendly web server for the biologist to analyze in-house epigenetic datasets and re-analyze public datasets on NCBI. To our best knowledge, this is the first webserver that can automatically download data from GEO, transform data, generate images, and support user interaction. Users can easily and quickly visualize and explore any public epigenetic data without requiring any special training or computing resources, and hence can study the epigenetic mechanism efficiently. The user case presented in section 4 confirmed that GsmPlot can be a huge driver to accelerate the research process by providing convenient visualization of both public and private data, and hence promoting data driven ideas. GsmPlot can dramatically improve the efficiency of utilization of public epigenetic data and further promote the research in the epigenetic community.

In conclusion, we proposed 3 tools to improve the data processing, analysis and visualization of low-input bisulfite sequencing data. By utilizing these tools, we can achieve high efficiency in low-input bisulfite sequencing alignment. For downstream analysis, researchers can validate their ideas more easily with little or no computational resources and bioinformatics background. These improvements we made remove the barriers in applying low-input bisulfite sequencing in larger scale researches and will also facilitate the downstream clinical epigenetic biomarker discovery and single cell methylome determination. To further improve data processing for low input bisulfite sequencing, a potential solution would be implementing LiBis in a distributed platform to reduce time consumption. Integrating the analysis and visualization for other sequencing methods such as Hi-C sequencing can also help to extend the usage of our tools.

REFERENCES

1. Tost, J., *DNA methylation: an introduction to the biology and the disease-associated changes of a promising biomarker.* Methods Mol Biol, 2009. **507**: p. 3-20.

2. Das, P.M. and R. Singal, *DNA methylation and cancer.* Journal of clinical oncology, 2004. **22**(22): p. 4632-4642.

3. Nelson, M., et al., *DNA methylation: molecular biology and biological significance*. 1993, Birkhauser-Verlag Press, Basel, Switzerland.

4. Vairapandi, M. and N.J. Duker, *Enzymic removal of 5-methylcytosine from DNA by a human DNA-glycosylase.* Nucleic acids research, 1993. **21**(23): p. 5323-5327.

5. Lam, A.K.Y., *Molecular biology of esophageal squamous cell carcinoma.* Critical reviews in oncology/hematology, 2000. **33**(2): p. 71-90.

6. Akhavan-Niaki, H. and A.A. Samadani, *DNA methylation and cancer development: molecular mechanism.* Cell biochemistry and biophysics, 2013. **67**(2): p. 501-513.

7. Parkin, D.M., et al., *Estimating the world cancer burden: Globocan 2000.* International journal of cancer, 2001. **94**(2): p. 153-156.

8. Ohki, I., et al., *Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA.* Cell, 2001. **105**(4): p. 487-497.

9. Ducasse, M. and M.A. Brown, *Epigenetic aberrations and cancer.* Molecular cancer, 2006. **5**(1): p. 1-10.

10. Ordway, J.M. and T. Curran, *Methylation matters: modeling a manageable genome.* Cell Growth and Differentiation-Publication American Association for Cancer Research, 2002. **13**(4): p. 149-162.

11. Lyko, F., et al., *Quantitative analysis of DNA methylation in chronic lymphocytic leukemia patients.* Electrophoresis, 2004. **25**(10‐11): p. 1530-1535.

12. Wei, S.H., et al., *Methylation microarray analysis of late-stage ovarian carcinomas distinguishes progression-free survival in patients and identifies candidate epigenetic markers.* Clinical Cancer Research, 2002. **8**(7): p. 2246-2252.

13. Flanagan, J.M., et al., *Gene-body hypermethylation of ATM in peripheral blood DNA of bilateral breast cancer patients.* Human molecular genetics, 2009. **18**(7): p. 1332-1342.

14. Moore, L.E., et al., *Genomic DNA hypomethylation as a biomarker for bladder cancer susceptibility in the Spanish Bladder Cancer Study: a case–control study.* The lancet oncology, 2008. **9**(4): p. 359-366.

15. Witte, T., C. Plass, and C. Gerhauser, *Pan-cancer patterns of DNA methylation.* Genome medicine, 2014. **6**(8): p. 1-18.

16. Capper, D., et al., *DNA methylation-based classification of central nervous system tumours.* Nature, 2018. **555**(7697): p. 469-474.

17. Lun, F.M., et al., *Noninvasive prenatal methylomic analysis by genomewide bisulfite sequencing of maternal plasma DNA.* Clinical chemistry, 2013. **59**(11): p. 1583-1594.

18. Lamb, Y.N. and S. Dhillon, *Epi proColon® 2.0 CE: a blood-based screening test for colorectal cancer.* Molecular diagnosis & therapy, 2017. **21**(2): p. 225-232.

19. Nuzzo, P.V., et al., *Detection of renal cell carcinoma using plasma and urine cell-free DNA methylomes.* Nature medicine, 2020. **26**(7): p. 1041-1043.

20. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences.* nature, 2009. **462**(7271): p. 315-322.

21. Grunau, C., S. Clark, and A. Rosenthal, *Bisulfite genomic sequencing: systematic investigation of critical experimental parameters.* Nucleic acids research, 2001. **29**(13): p. e65-e65.

22. Meissner, A., et al., *Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis.* Nucleic acids research, 2005. **33**(18): p. 5868-5877.

23. Miura, F., et al., *Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging.* Nucleic Acids Res, 2012. **40**(17): p. e136.

24. Smallwood, S.A., et al., *Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity.* Nat Methods, 2014. **11**(8): p. 817-820.

25. Olova, N., et al., *Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data.* Genome biology, 2018. **19**(1): p. 1-19.

26. Pan, W., et al., *Brain tumor mutations detected in cerebral spinal fluid.* Clinical chemistry, 2015. **61**(3): p. 514-522.

27. Eberwine, J., et al., *The promise of single-cell sequencing.* Nature methods, 2014. **11**(1): p. 25-27.

28. Miura, F., et al., *Highly efficient single-stranded DNA ligation technique improves low-input whole-genome bisulfite sequencing by post-bisulfite adaptor tagging.* Nucleic acids research, 2019. **47**(15): p. e85-e85.

29. Human, W., *Sequencing Coverage Calculation Methods for Human Whole-Genome Sequencing.*

30. Bioinformatics, B., *FastQC: a quality control tool for high throughput sequence data.* Cambridge, UK: Babraham Institute, 2011.

31. Chen, S., et al., *fastp: an ultra-fast all-in-one FASTQ preprocessor.* Bioinformatics, 2018. **34**(17): p. i884-i890.

32. Krueger, F. and S.R. Andrews, *Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.* bioinformatics, 2011. **27**(11): p. 1571-1572.

33. Chen, P.-Y., S.J. Cokus, and M. Pellegrini, *BS Seeker: precise mapping for bisulfite sequencing.* BMC bioinformatics, 2010. **11**(1): p. 1-6.

34. Pedersen, B.S., et al., *Fast and accurate alignment of long bisulfite-seq reads.* arXiv preprint arXiv:1401.1129, 2014.

35. Xi, Y. and W. Li, *BSMAP: whole genome bisulfite sequence MAPping program.* BMC Bioinformatics, 2009. **10**: p. 232.

36. Wu, T.D., et al., *GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality*, in *Statistical genomics*. 2016, Springer. p. 283-334.

37. Kiełbasa, S.M., et al., *Adaptive seeds tame genomic sequence comparison.* Genome research, 2011. **21**(3): p. 487-493.

38. Bock, C., *Analysing and interpreting DNA methylation data.* Nature Reviews Genetics, 2012. **13**(10): p. 705-719.

39. Wu, P., et al., *Using local alignment to enhance single-cell bisulfite sequencing data efficiency.* Bioinformatics, 2019. **35**(18): p. 3273-3278.

40. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.

41. Ramírez, F., et al., *deepTools: a flexible platform for exploring deep-sequencing data.* Nucleic Acids Research, 2014. **42**(W1): p. W187-W191.

42. Cavalcante, R.G., et al., *Integrating DNA Methylation and Hydroxymethylation Data with the Mint Pipeline.* Cancer Research, 2017. **77**(21): p. e27-e30.

43. Locke, W.J., et al., *DNA methylation cancer biomarkers: Translation to the clinic.* Frontiers in Genetics, 2019. **10**.

44. Saghafinia, S., et al., *Pan-Cancer Landscape of Aberrant DNA Methylation across Human Tumors.* Cell Reports, 2018. **25**(4): p. 1066-1080.e8.

45. Mouliere, F. and N. Rosenfeld, *Circulating tumor-derived DNA is shorter than somatic DNA in plasma.* Proc Natl Acad Sci U S A, 2015. **112**(11): p. 3178-9.

46. Guo, S., et al., *Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA.* Nat Genet, 2017. **49**(4): p. 635-642.

47. Moss, J., et al., *Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease.* Nature communications, 2018. **9**(1): p. 5068.

48. Van Der Pol, Y. and F. Mouliere, *Toward the early detection of cancer by decoding the epigenetic and environmental fingerprints of cell-free DNA.* Cancer cell, 2019. **36**(4): p. 350-368.

49. Yong, W.-S., F.-M. Hsu, and P.-Y. Chen, *Profiling genome-wide DNA methylation.* Epigenetics & chromatin, 2016. **9**(1): p. 26.

50. Newman, A.M., et al., *An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage.* Nat Med, 2014. **20**(5): p. 548-54.

51. Olova, N., et al., *Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data.* Genome Biol, 2018. **19**(1): p. 33.

52. Luo, C., et al., *Robust single-cell DNA methylome profiling with snmC-seq2.* Nat Commun, 2018. **9**(1): p. 3824.

53. Sun, D., et al., *MOABS: model based analysis of bisulfite sequencing data.* Genome Biol, 2014. **15**(2): p. R38.

54. Krueger, F. and S.R. Andrews, *Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.* Bioinformatics, 2011. **27**(11): p. 1571-2.

55. Guo, W., et al., *BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data.* BMC genomics, 2013. **14**(1): p. 774.

56. Hovestadt, V., et al., *Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing.* Nature, 2014. **510**(7506): p. 537-541.

57. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.

58. Huang, K.Y.Y., Y.-J. Huang, and P.-Y. Chen, *BS-Seeker3: ultrafast pipeline for bisulfite sequencing.* BMC bioinformatics, 2018. **19**(1): p. 111.

59. *Fast and accurate alignment of long bisulfite-seq reads.*

60. Chen, H., A.D. Smith, and T. Chen, *WALT: fast and accurate read mapping for bisulfite sequencing.* Bioinformatics, 2016. **32**(22): p. 3507-3509.

61. Zhou, Q., et al., *An integrated package for bisulfite DNA methylation data analysis with Indel-sensitive mapping.* BMC bioinformatics, 2019. **20**(1): p. 1-11.

62. Zhu, P., et al., *Single-cell DNA methylome sequencing of human preimplantation embryos.* Nature Genetics, 2018. **50**(1): p. 12-19.

63. Holmes, A., et al., *Mechanistic signatures of HPV insertions in cervical carcinomas.* NPJ genomic medicine, 2016. **1**: p. 16004.

64. Li, W., et al., *The characteristics of HPV integration in cervical intraepithelial cells.* Journal of Cancer, 2019. **10**(12): p. 2783.

65. Tripathi, R., et al., *Jagged-1 induced molecular alterations in HPV associated invasive squamous cell and adenocarcinoma of the human uterine cervix.* Scientific reports, 2018. **8**(1): p. 1-9.

66. Hsu, W., et al., *LncRNA CASC11 promotes the cervical cancer progression by activating Wnt/beta-catenin signaling pathway.* Biological research, 2019. **52**(1): p. 33.

67. Lee, D.-S., et al., *Simultaneous profiling of 3D genome structure and DNA methylation in single human cells.* Nature methods, 2019: p. 1-8.

68. Huang, Y., et al., *The anti-CMS technique for genome-wide mapping of 5-hydroxymethylcytosine.* Nat Protoc, 2012. **7**(10): p. 1897-908.

69. Yu, M., et al., *Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine.* Nat Protoc, 2012. **7**(12): p. 2159-70.

70. Ziller, M.J., et al., *Charting a dynamic DNA methylation landscape of the human genome.* Nature, 2013. **500**(7463): p. 477-81.

71. Stadler, M.B., et al., *DNA-binding factors shape the mouse methylome at distal regulatory regions.* Nature, 2011. **480**(7378): p. 490-5.

72. Domcke, S., et al., *Competition between DNA methylation and transcription factors determines binding of NRF1.* Nature, 2015. **528**(7583): p. 575-9.

73. Yin, Y., et al., *Impact of cytosine methylation on DNA binding specificities of human transcription factors.* Science, 2017. **356**(6337).

74. Bogdanovic, O., et al., *Active DNA demethylation at enhancers during the vertebrate phylotypic period.* Nat Genet, 2016. **48**(4): p. 417-26.

75. Mahe, E.A., et al., *Cytosine modifications modulate the chromatin architecture of transcriptional enhancers.* Genome Res, 2017. **27**(6): p. 947-958.

76. Akalin, A., et al., *Genomation: a toolkit to summarize, annotate and visualize genomic intervals.* Bioinformatics, 2015. **31**(7): p. 1127-9.

77. Yu, G., L.G. Wang, and Q.Y. He, *ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization.* Bioinformatics, 2015. **31**(14): p. 2382-3.

78. Ramirez, F., et al., *deepTools2: a next generation web server for deep-sequencing data analysis.* Nucleic Acids Res, 2016. **44**(W1): p. W160-5.

79. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome Biol, 2008. **9**(9): p. R137.

80. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 2014. **15**(12): p. 550.

81. Barski, A., et al., *High-resolution profiling of histone methylations in the human genome.* Cell, 2007. **129**(4): p. 823-37.

82. Hon, G.C., et al., *5mC oxidation by Tet2 modulates enhancer activity and timing of transcriptome reprogramming during differentiation.* Mol Cell, 2014. **56**(2): p. 286-97.

83. Jones, P.A., *Functions of DNA methylation: islands, start sites, gene bodies and beyond.* Nat Rev Genet, 2012. **13**(7): p. 484-92.

84. *The ENCODE (ENCyclopedia Of DNA Elements) Project.* Science, 2004. **306**(5696): p. 636-640.

85. Yamaji, M., et al., *PRDM14 ensures naive pluripotency through dual regulation of signaling and epigenetic pathways in mouse embryonic stem cells.* Cell Stem Cell, 2013. **12**(3): p. 368-82.

86. Chan, Y.S., et al., *A PRC2-dependent repressive role of PRDM14 in human embryonic stem cells and induced pluripotent stem cell reprogramming.* Stem Cells, 2013. **31**(4): p. 682-92.

87. Harrington, M.A., et al., *Cytosine methylation does not affect binding of transcription factor Sp1.* Proc Natl Acad Sci U S A, 1988. **85**(7): p. 2066-70.

88. Cantone, I. and A.G. Fisher, *Epigenetic programming and reprogramming during development.* Nat Struct Mol Biol, 2013. **20**(3): p. 282-9.

89. Messerschmidt, D.M., B.B. Knowles, and D. Solter, *DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos.* Genes Dev, 2014. **28**(8): p. 812-28.

90. Song, C.X. and C. He, *Balance of DNA methylation and demethylation in cancer development.* Genome Biol, 2012. **13**(10): p. 173.

91. Lin, X., et al., *BSeQC: quality control of bisulfite sequencing experiments.* Bioinformatics, 2013. **29**(24): p. 3227-9.

92. Hansen, K.D., B. Langmead, and R.A. Irizarry, *BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions.* Genome Biol, 2012. **13**(10): p. R83.

93. Zhu, H., G. Wang, and J. Qian, *Transcription factors as readers and effectors of DNA methylation.* Nat Rev Genet, 2016. **17**(9): p. 551-65.

94. Neri, F., et al., *Intragenic DNA methylation prevents spurious transcription initiation.* Nature, 2017. **543**(7643): p. 72-77.

95. Baubec, T., et al., *Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation.* Nature, 2015. **520**(7546): p. 243-7.

96. Allis, C.D. and T. Jenuwein, *The molecular hallmarks of epigenetic control.* Nat Rev Genet, 2016. **17**(8): p. 487-500.

97. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, 2009. **462**(7271): p. 315-22.

98. Buenrostro, J.D., et al., *ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide.* Curr Protoc Mol Biol, 2015. **109**: p. 21 29 1-21 29 9.

99. Giles, K.A., et al., *Integrated epigenomic analysis stratifies chromatin remodellers into distinct functional groups.* Epigenetics Chromatin, 2019. **12**(1): p. 12.

100. Li, J., et al., *Decoding the dynamic DNA methylation and hydroxymethylation landscapes in endodermal lineage intermediates during pancreatic differentiation of hESC.* Nucleic Acids Res, 2018. **46**(6): p. 2883-2900.

101. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets--update.* Nucleic Acids Res, 2013. **41**(Database issue): p. D991-5.

102. Barrett, T., et al., *NCBI GEO: archive for high-throughput functional genomic data.* Nucleic Acids Res, 2009. **37**(Database issue): p. D885-90.

103. Kent, W.J., et al., *BigWig and BigBed: enabling browsing of large distributed datasets.* Bioinformatics, 2010. **26**(17): p. 2204-7.

104. Fejes, A.P., M.J. Jones, and M.S. Kobor, *DaVIE: Database for the Visualization and Integration of Epigenetic data.* Front Genet, 2014. **5**: p. 325.

105. Kim, T., et al., *Octopus-toolkit: a workflow to automate mining of public epigenomic and transcriptomic next-generation sequencing data.* Nucleic Acids Res, 2018.

106. Jammula, S. and D. Pasini, *EpiMINE, a computational program for mining epigenomic data.* Epigenetics Chromatin, 2016. **9**: p. 42.

107. Li, D., et al., *WashU Epigenome Browser update 2019.* Nucleic Acids Res, 2019. **47**(W1): p. W158-W165.

108. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.

109. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

110. Xu, Y., et al., *Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells.* Mol Cell, 2011. **42**(4): p. 451-64.

111. Yu, M., et al., *Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome.* Cell, 2012. **149**(6): p. 1368-80.

112. Xie, D., et al., *Dynamic trans-acting factor colocalization in human cells.* Cell, 2013. **155**(3): p. 713-24.

113. Mei, S., et al., *Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse.* Nucleic Acids Res, 2017. **45**(D1): p. D658-D662.

114. Ebmeier, C.C., et al., *Human TFIIH Kinase CDK7 Regulates Transcription-Associated Chromatin Modifications.* Cell Rep, 2017. **20**(5): p. 1173-1186.

115. Teif, V.B., et al., *Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development.* Genome Res, 2014. **24**(8): p. 1285-95.

116. Fang, S., et al., *Tet inactivation disrupts YY1 binding and long-range chromatin interactions during embryonic heart development.* Nat Commun, 2019. **10**(1): p. 4297.

117. Cai, C.L., et al., *Isl1 identifies a cardiac progenitor population that proliferates prior to differentiation and contributes a majority of cells to the heart.* Dev Cell, 2003. **5**(6): p. 877-89.

118. Gao, R., et al., *Pioneering function of Isl1 in the epigenetic control of cardiomyocyte cell fate.* Cell Res, 2019. **29**(6): p. 486-501.