ABSTRACT

Title of dissertation: ESTIMATION AND ANALYSIS OF CELL-SPECIFIC DNA METHYLATION FROM BISULFITE-SEQUENCING DATA

Faezeh Dorri, Doctor of Philosophy, 2018

Dissertation directed by: Professor Héctor Bravo Corrada
Department of Computer Science

DNA methylation is the best understood heritable gene regulatory mechanism that does not involve direct modification of DNA sequence itself. Cells with different methylation profiles (over temporal or micro-environmental dimensions) may exhibit different phenotypic properties. In cancer, heterogeneity across cells in the tumor microenvironment presents significant challenges to treatment. In particular, epigenetic heterogeneity is discernible among tumor cells, and it is believed to impact the growth properties and treatment resistance of tumors.

Existing computational methods used to study the epigenetic composition of cell populations are based on the analysis of DNA methylation modifications at multiple consecutive genomic loci spanned by single DNA sequencing reads. These approaches have

yielded great insight into how cell populations differ epigenetically across different tissues. However, they only provide a general summary of the epigenetic composition of these cell populations without providing cell-specific methylation patterns over longer genomic spans to perform a comprehensive analysis of the epigenetic heterogeneity of cell populations.

In this dissertation, we address this challenge by proposing two computational methods called methylFlow and MCFDiff. In methylFlow, we propose a novel method based on network flow algorithms to reconstruct cell-specific methylation profiles using reads obtained from sequencing bisulfite-converted DNA. We reveal the methylation profile of underlying clones in a heterogeneous cell population including the methylation patterns and their corresponding abundance within the population.

In MCFDiff, we propose a statistical model that leverages the identified cell-specific methylation profiles (from methylFlow) to determine regions of differential methylation composition (RDMCs) between multiple phenotypic groups, in particular, between tumor and paired normal tissue. In MCFDiff, we can systematically exclude the tumor tissue impurities and increase the accuracy in detecting the regions with differential methylation composition in normal and tumor samples. Profiling the changes between normal and tumor samples according to the reconstructed methylation profile of underling clone in different samples leads us to the discovery of de novo epigenetic markers and a better understanding about the effect of epigenetic heterogeneity in cancer dynamics from the initiation, progression to metastasis, and relapse.

ESTIMATION AND ANALYSIS OF CELL-SPECIFIC DNA METHYLATION FROM BISULFITE-SEQUENCING DATA

by

Faezeh Dorri

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Héctor Bravo Corrada, Chair/Advisor
Professor Sridhar Hannenhalli
Professor Mark Leiserson
Professor Stephen Mount
Professor Mihai Pop

Dedication

To my daughter, Helia, who brought life to my life!

Acknowledgments

First and foremost I wish to thank multitude of people who helped me throughout my journey as a PhD student. I would like to express my sincere gratitude to my advisor, Prof. Héctor Bravo Corrada, for giving me an invaluable opportunity to work on challenging and extremely interesting projects over the past 6 years. He has always made himself available for help and gave me the freedom much required to develop my own research agenda. He never doubted my capabilities and always had confidence in my ability to complete my PhD. I am thankful for his understanding, support and belief in me when I had to leave for a while for family reasons. It has been a pleasure to work with and learn from such an extraordinary individual. His effort and patience will never be forgotten. I would like to extend my appreciation to my committee members, Prof. Mihai Pop, Prof. Sridhar Hannenhalli, Prof. Stephen Mount, and Prof. Mark Leiserson for serving on my dissertation committee and providing me with informative comments.

I believe that grad school is indeed one of the most interesting pages of one's life. I would like to thank all my friends Vahid Liaghat, Ferreshteh Radaei, Mahfuza Sharmin, Ali Shafahi, Kiana Roshanzamir, Ehsan Behnam Ghader, Sara Samkhani, Mohsen Asgari, Mohammad Rashidian, Hamidreza Mahmoudi, Leila Fotoohi, Atefeh Kashani-Nejad, MohammadReza Khani, Azam Rivaz, Omid Naeimaei, and many others for filling this page of my life with the most wonderful memories. Thank you all!

As always it is impossible to mention everybody who had an impact to this work however there are those whose spiritual support is even more important. I feel a deep sense of gratitude for my mother and father, who formed part of my vision and taught me good things that really matter in life. Their infallible love and support has always been my strength. Their patience and sacrifice will remain my inspiration throughout my life. I would like especially thank my twin sister, Fatemeh Dorri and my brother-in-law, Mohsen Keshavarz Akhlaghi for their sincere help and support. They are always thanked in my papers for helpful discussion and proofreading, and in my heart for their enormous love. Last but not least, I am deeply grateful to the love of my life and my husband, Hamid Mahini, not only for his affection and emotional support, but also for his inspiration and contribution to my research. I would like to thank my husband and my daughter for their company during every single step of this journey.

It is impossible to remember all, and I apologize to those I've inadvertently left out. Lastly, thank you all and thank God!

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Overview

DNA as the molecule that encodes the genetic information of life is very stable and robust. During cell division, daughter cells inherit almost exact DNA sequence. However, scientists observed that cells with identical genomic information may have different phenotypic properties. Conard Waddington, in 1939, first introduced the concept of epigenetics [Wad42] to describe the interaction between the cell and its environment. Today, epigentics refers to any heritable modification of DNA (e.g methylation) that does not change the genome [DAB09]. These modifications affect gene activity and leads to emergence of cells with different phenotypes. For example, differences in monozygotic twins are partially explained by epigenetic modifications on genome [PGK+03]. However, epigenetic modifications are also frequently associated in tumor cells to the modulation of their malignant transcriptome [Ced88]. Different phenotypic properties in different tumor tissues and cell types (inter-tumor heterogeneity), and also the subclonal diversity within a single tumor tissue (intra-tumor heterogeneity) correspond to different epigenetic mod-

ifications to the genome [BMBS13]. EWAS (Epigenome-Wide Association Studies) have investigated the genome-wide differences of the epigenetic markers in different individuals with various diseases or traits [RDBB11]. Epigenome-Wide Association Studies are carried out under the assumption that methylation modifications vary by time due to environmental stimuli. These modifications can change the healthy regulation of gene expression to a disease pattern in cancerous cells.

Among different types of epigenetic modifications (that are discussed comprehensively in Section 1.1), DNA methylation has a significant role in different cancers [RDBB11, Bir02]. DNA methylation is the covalent binding of a methyl group to the fifth carbon of cytosine, forming 5-methylcytosine. DNA methylation modification mainly happens at CpG sites and less frequently in non-CpG context. It regulates gene expression by recruiting proteins involved in gene repression or by inhibiting the transcription factor(s) binding to DNA. If the methyl group binding happens at the promoter sites of a gene, it leads to inactivation of a gene by decreasing its expression. EWAS confirm the association of numerous cellular processes, such as transcriptional repression [KPW97], X chromosome inactivation [MSS81], embryonic development [RBL+00], genomic imprinting [LBJ93], the alteration of chromatin structure, and transposon inactivation with DNA methylation modification across genome [Jon12]. None of those findings would be revealed without the development of both experimental and computational approaches.

There are several experimental approaches that are available to catalogue the genome-wide DNA methylation in epigenome-wide association studies (explained thor-

oughly in Section 1.2). The main differences between these approaches is the procedure they use to reveal the methylated sites across genome, their output resolutions, and information content. The chemical procedure for identifying methylated and unmethylated sites can be categorized as:

(i) enzyme digestion based methods,

(ii) affinity enrichment based methods,

(iii) bisulfite conversion based methods.

Any of the above procedures could be followed by microarray analysis or sequencing procedure that implies different outputs as a result. For example in MeDIP-chip, an affinity enrichment based method [JBE08], the methylation status for a region (depending on the probes design in the array) is reported. In other methods, the methylation status at a single base resolution is reported. Such methods include whole genome bisulfite sequencing techniques (WGBS), reduced representation bisulfite sequencing techniques (RRBS) [MGB+05a], Illumina's Infinium Methylation assay [BLB+09] as bisulfite conversion based approaches, and MeDIP-seq that is an affinity enrichment based method. The methods that utilize microarrays technology, such as those from Illumina's HM450K, report intensity. Intensity reflects the methylation status of the samples within each locus and is equal to the ratio between the abundance of methylated and unmethylated CpGs. On the other hand, methods like WGBS, RRBS reflect the methylation status of every single base and its coverage within the sample. Note that methods followed by sequenc-

ing techniques usually have higher cost than those followed by microarray techiniques. In addition, all these techniques could be customized to single-cells, introducing a new generation of methylation profiling at single-cell resolution. Among those, third-generation sequencing technology employs sequencing techniques to reveal the methylation profile of a single cell. This opens the doors to a lot of discoveries about DNA methylation and also improves our understanding about tumorigenesis and tumor molecular heterogeneity. Regardless of the promises, third-generation sequencing technologies still suffer from higher error rate, higher cost, and lower throughput than second-generation sequencing technologies like WGBS and RRBS.

The choice of the techniques selected for different analysis is determined based on:

- budget/cost;
- the goal of the study-detecting de novo epigenetic marks or the investigation of known methylation site;
- the amount, type and quality of the DNA sample(s), number of samples, tissue types(human, mouse) and specimen amount;
- the required performance in terms of sensitivity and specificity;
- the availability of computational tools.

Among all the techniques, bisulfite sequencing technique is considered the gold standard method to assess genome-wide DNA methylation in spite of the expense. In order to detect de novo differential methylation loci or regions when few samples are available, whole genome bisulfite sequencing technique is the most comprehensive one

among all existing methods.

With the access to this increasing amount of data retrieved from the above technologies, different computational approaches are developed to analyze DNA methylation profiles across various tissues and samples. The analysis are mainly focused on visualizing and finding differentially methylated loci or regions that are tissue or disease-specific. In Section 1.3, we review different analytical methods and techniques implemented to find differential methylation loci and regions as a major problem in EWAS. Most of existing methods that are concentrated on detecting differentially methylated loci and regions, utilize data at a single base resolution which induces an enormous computational cost on the algorithm. In addition, most of the existing methods utilize the mean and the variance of marginal methylation percentage as the input signal to find the differentially methylated loci or regions. Using this type of data, we lose the spatial correlation of the methylation status of adjacent CpG sites. In addition, tumor tissue impurity and the sequencing noise introduce error in the analysis.

In this dissertation, we propose a statistical and computational approach to study the epigenetic diversity of a heterogenous cell population comprehensively. In our approach we do not process data at single base resolution and systematically remove the noises due to tumor impurity and sequencing errors. We (i) reconstruct the methylation profiles of underlying clones within a heterogeneous cell-type population; and (ii) find the regions with the significant difference in the composition of the methylation profiles of underlying clones comparing normal and tumor samples by utilizing high-throughput bisulfite-

converted sequencing data. In fact, our method could be considered as a computational complement to second-generation sequencing techniques to extract the information about the DNA methylation profiles at single-cell single-base resolution. It is similar to that of the third-generation sequencing techniques that reveal the underlying DNA methylation clones of a heterogenous population while keeping more coverage across different regions in genome. It also gives an estimate about the abundance of each clone in the heterogenous cell-type population. The latter cannot be obtained by third-generation sequencing techniques.

1.1 Introduction to Epigenetics

Definition of epigenetic term reveals that we are more than sum of our genes. The current and our past situations affect the phenotypes of our cells. Epigenetic modifications generate phenotypic variations within cells with identical DNA sequence. These variations could be established in genome and even inherited by the next generations.

In order to understand epigenetic modifications, first we need to know about chromatin. Chromatin is a condensed structure of genomic DNA and its associated proteins in the nucleus [Kou07]. Chromatin organizes DNA and wraps it around histone proteins such that two meters of DNA is packed inside a cell nucleus with a diameter of 6 um. At the same time chromatin allows DNA to become accessible easily to a variety of proteins that are involved in DNA transcription, replication and repair processes [Est08, Est07].

1.1.1 Histone modification

A complex of eight histone proteins that include two copies of each of the core histones H2A, H2B, H3 and H4 with a stretch of double-stranded DNA (that is wrapped around the histone core) form a nucleosome. A histone modification is a covalent post-translational modification (PTM) to histone proteins which includes methylation, phosphorylation, acetylation, ubiquitylation, and sumoylation. Amino acid residues in histone tails can be acetylated (lysines), methylated (lysines) or phosphorylated (serines). All these modifications can make changes to the shape of the histones and might cause incomplete unwounding of DNA during replication. If the modified histones carried into new copies of DNA, they can act as templates for initiation of new histones and alters the shape of new histones. This is how histone modifications inherited from ancestors.

Studying the effect of different histone modification on genome shows that acetylation is generally associated with active transcription [BK11], but for most other modifications the effects are less predictable. These modifications on histones generally impact gene expression by altering chromatin structure or recruiting histone modifiers.

1.1.2 DNA methylation

DNA methylation modification is described as binding of methyl group to 5 carbon of cytosine that occurs most frequently at CpG locations. DNA methylation is an epigenetic modification that is associated with many of the biological processes such as tran-

scriptional repression [KPW97], X chromosome inactivation [MSS81], embryonic development [RBL+00], genomic imprinting [LBJ93], the alteration of chromatin structure and transposon inactivation with DNA methylation modification across genome [Jon12].

This process affects the expression of genes by changing the chromatin structure. Hypermethylation usually causes silencing of genes by inhibiting the binding of transcription factors to the promoter region of genes and in case of hypomethylation, the expression and production of corresponding proteins increases because the limitations disappear, and more transcription factors bind to the promoter regions of DNA.

DNA methylation pattern of a cell changes over time and during developmental stage of cells or it can also remain from the germ line of one of the parents into the zygote. Hence, the methylation status of some regions is inherited from one parent or the other. DNA methylation varies from tissue to tissue and from cell to cell, make a population of cells in a tissue heterogenous. DNA methylation modification is also sequence-context dependent. It is mainly found in CpG dinucleotides and less in non-CpG regions. In vertebrates $60 - 80\%$ of CpGs are methylated in somatic cells. However in mammals this ratio is less in CpG islands, i.e. these regions are mainly unmethylated [CFJ+11]. CpG islands mainly refer to regions that has a length greater than 200bp and a G+C content greater than 50%. This epigenetic modification is also trans-generationaly heritable. In each generation, DNA methylation patterns are cleared and established again during development except for the trans-generational epigenetic regions that their clearance process is incomplete.

All these confirm the necessity of studying DNA methylation modification and profiling its pattern across genome. Also detection of differentially methylated regions helps to find the association between DNA methylation profiles and disease or tissue specific expression. In Section 1.2 we overview different experimental methods invented to profile DNA methylation status of tissue- and disease specific samples.

1.1.3 Chromatin remodeling and nucleosome repositioning

Chromatin remodelling and nucleosome repositioning are other epigenetic mechanisms. Chromatin may have different compaction state. The more compact the chromatin, the harder it is for transcription factors and other DNA binding proteins to access DNA and accomplish their job. This is similar to what happens after DNA methylation modification and histone modification. The accessibility of some regions across genome alters due to DNA methylation or histone modifications and chromatin remodeling complexes that are responsible for changing the compaction state of chromatin. When chromatin is condensed, and not actively being transcribed it is called heterochromatin. When chromatin is more loosely packed, and therefore accessible for transcription factors and other DNA binding proteins, it is called euchromatin. Loosening up the chromatin allows access to DNA and facilitates DNA processes such as transcription, replication and repair.

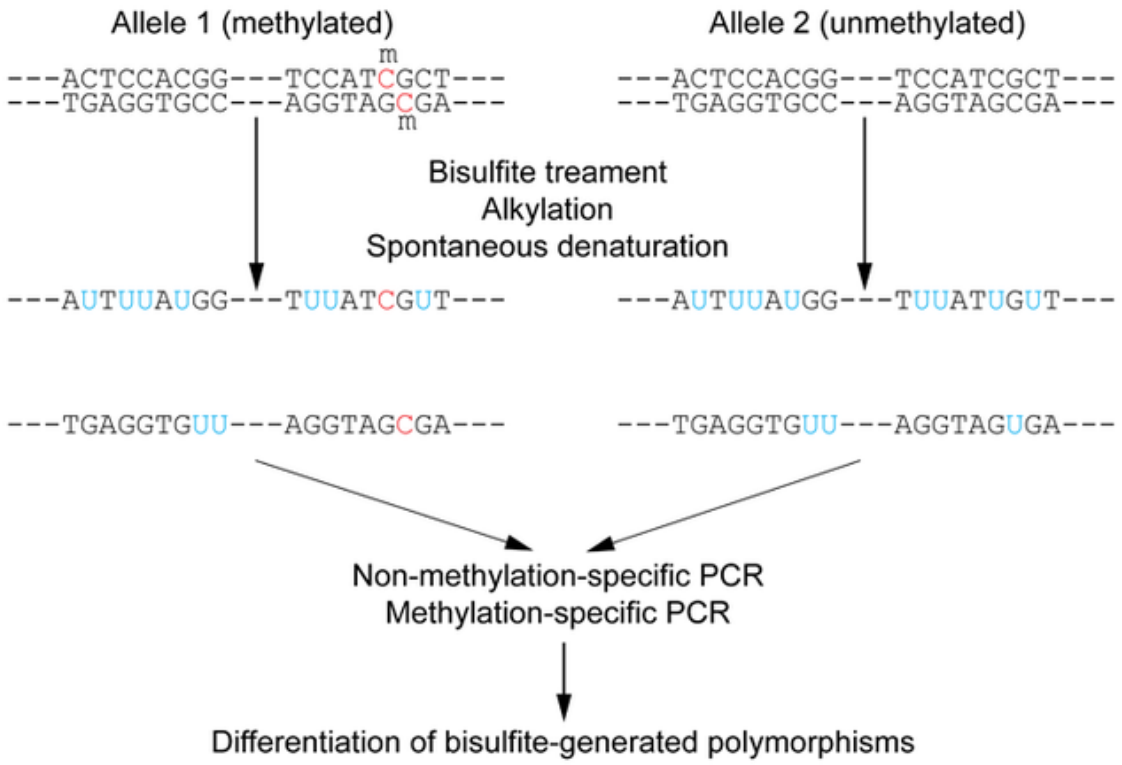

Figure 1.1: DNA methylation analysis methods not based on methylation-specific PCR by Toeng from

1.2 Experimental methods for DNA methylation profiling

As mentioned above, DNA methylation is one of the epigenetic gene regulatory mechanism where silencing of gene expression is established by bonding of methyl groups to DNA at specific genomic regions [HP75]. It is the best understood heritable mechanism for gene regulation that does not involve direct modification of DNA sequence itself. Various studies confirmed the role of DNA methylation modification in many biological processes and its association in many diseases like diabetes and cancers. Scientists discovered all these findings by the advent of techniques to profile DNA methylation modifications in various tissues and samples. Previous standard molecular biology techniques, such as PCR, cloning, and hybridization does not distinguish between methylated and unmethylated cytosines. Initial studies mainly focused on obtaining the methylation status of selected regions and of the genes of interest. Later with the invention of microarrays and microarray hybridization technology, profiling methylation status scaled up to the genome-wide level and finally high-throughput sequencing technologies and new probe design for arrays enable profiling of DNA methylation modifications at single base resolution. DNA methylation profiling methods use different strategies in revealing the methylation pattern information. Here, we briefly describe the three aforementioned categories.

1.2.1 Enzyme digestion based methods

Enzyme digestion based methods take advantage of those enzymes that are methylation-sensitives. Some of the restriction enzymes like such as BstUI, HpaII, NotI and SmaI, cleve only the unmethylated DNA and some like McrBC digest the methylated DNA.

- *MRE-seq* In MRE-seq DNA is first digested by methylation-sensitive restriction enzyme. Then it is sequenced and compared with a non-digested DNA. Deep sequencing makes the accurate profiling of DNA methylation at single base resolution by reporting the relative DNA methylation levels. However, this method has relatively low coverage of the genome with more reliable results at CpG-containing recognition sites [MNB+10].
- *CHARM* The comprehensive high-throughput arrays for relative methylation (CHARM) method [ILAC+08a] uses McrBC, an enzyme that digests methylated DNA. Then without PCR, and differential hybridization to an array, It uses a new tiling array specifically designed to maximize the number of assayed CpGs. McrBC cleave half of the methylated DNA and can detect DNA genome-wide methylation.

1.2.2 Affinity enrichment based methods

The methylated sites of genome can bind to methyl-CpG-binding domain proteins (MBDs) or antibodies specific for 5mC to enrich methylated DNA regions. Afterwads

CHIP or sequencing techniques with various platforms are available to profile DNA methylations. Depending on the technique and the type of protein used to profile the DNA methylation, the analysis of DNA methylation data and genome coverage varies. GC content, the extent of DNA amplification, and copy number should be considered in the subsequent analysis.

- *Methylated DNA immunoprecipitation (MeDIP-chip, MeDIP-seq)* In Methylated DNA immunoprecipitation (MeDIP) method [ZWH+14], the methylated DNA is fractionated using an antibody; then hybridized with a differentially labelled DNA control to an oligonucleotide array. The relative levels of hypermethylation and hypomethylation is shown by producing a ratio of green fluorescence to red fluorescence in the array, while immunocaptured DNA and control genomic DNA that are both labelled with Cy5 and Cy3 fluorescent dyes.

  The drawback of MeDIP includes their inability to pinpoint DNA methylation changes at a single base resolution and its biased toward hypermethylated regions. It also uses an antibody to 5-methyl-cytosine that is targeting single-stranded DNA.

- *MethylCap-seq and MBDCap-seq* These techniques use MBD protein which relies on the capacity of MBD proteins to bind specifically to methylated DNA sequences. The advantage of this method is the ability of MBD protein to discriminate between 5mc and 5hmc. Later, microarray (MBD-chip) or sequencing (MBDCap-seq/MethylCap-seq) [BSM+10] can be used to profile DNA methylation. Also, in MBDCap approach the methyl-CpG binding domain of the MBD2 protein capture

double-stranded DNA [SLT09].

1.2.3 Bisulfite conversion based method

DNA is treated with sodium bisulfite to reveal the methylation status of single nucleotides. Sodium bisulfite is a chemical compound that converts unmethylated cytosines into uracil and methylated cytosines remains the same. Different high-throughput sequencing techniques of bisulfite-converted DNA is employed to measure DNA methylation modifications at base-pair level. This approach has led to deeper understanding of DNA methylation and its role in the development of cells [LPD+09] and many disease [HTB+11]. In methylation array techniques the bisulfite-treated DNA is hybridized to arrays that contain predesigned probes to distinguish between methylated and unmethylated Cs. Although whole genome bisulfite sequencing technique has high cost compare to other methods, bisulfite sequencing quickly considered as the method of choice for bulk DNA methylation analysis due to its high coverage.

- *Whole Genome Bisulfite Sequencing* (WGBS)

  Whole genome bisulfite sequencing is a next-generation sequencing technology used to determine the DNA methylation status of single cytosines. DNA is treated with sodium bisulfite to reveal the methylation status of single nucleotides. Sodium bisulfite is a chemical compound that converts unmethylated cytosines into uracil and methylated cytosines remains the same. Converted DNA is then sequenced in a process similar to standard DNA sequencing methods. The treated DNA is se-

quenced using non-methylation- specific PCR methods. The thymine appears in sequencing instead of uracil representing the unmethylated cytosines while cytosines sequenced for methyl cytosine residues. Figure 1.1 from (https://en.wikipedia.org) shows the process. Thus, sequencing the treated DNA with bisulfite reveals information about the methylation status of DNA. Various analyses can be performed to reveal the underlying information of methylation status of DNA among different samples.

- *Reduced Representation Bisulfite Sequencing* (RRBS) To cut the cost and reduce the amount of nucleotides that need to be sequenced, some restriction enzymes, typically with MspI which is methylation insensitive, are used to cut the genome at CCGG sites and extract the high CpG density regions of genome. After repairing the ends, the rest of process follows the same strategy as whole genome bisulfite sequencing. The CpG-rich segmented DNA is treated by sodium bisulfite and is sequenced using non-methylation-sensetive PCR methods. In reduced representation bisulfite sequencing method the number of reads need to be sequenced reduced dramatically and only 1% of genome is sequenced. The major problem with this method is that only about 85% of CpG islands and 60% of promoters and CpG island shores are captured [GSB+11], [LJD+14], [MGB+05a].

- *Methylation Arrays* . To enable cost-effective DNA methylation profiling, Illumina offers a robust methylation profiling platform [BLB+09]. Illumina's Infinium HumanMethylation450 BeadChip (HM450K) protocol involves the treatment of ge-

nomic DNA with bisulfite and followed by amplification and hybridization of the converted DNA to arrays containing predesigned probes featuring comprehensive gene regions and CpG island coverage to distinguish between methylated and unmethylated Cs.

These method has no PCR and hence no bias toward short fragments. The main disadvantage of this method is that not every gene of interest is included in the design of assay.

1.2.4 Single cell DNA methylation profiling

Profiling DNA methylation is extended to the single cell level by recent technological innovations. We can categorize different developed strategies in to bisulfite-base and bisulfite-free methods. Profiling DNA methylation at single cell level give us the opportunity to answer many biological questions and obtain new and deeper insights about some biological events like tumor heterogeneity and evolution of cancers. These information could be added to other level of information like gene expression, mutation. One of the main advantage of single-cell DNA methylation profiling method is its application toward clinical purposes.

scRRBS [GZW+13, GZG+15], Q-RRBS [WLD+15], MID-RRBS [MRS+18] are reduced representation bisulfite sequencing based methods. scWGBS [FSN+15], scBS-seq [SLA+14a], scPBAT [MI14] are whole genome bisulfite sequencing techniques at single cell level. There are also some bisulfite-free methods like scCGI-seq [HWZ+17],

RGM [SSS+15], RSMA [KKH+11] and SCRAM [LCB+13, CQBM15].

However, single-cell DNA methylation profiling methods are not able to give a reliable estimate about the abundance of each cell in a heterogenous cell population which is also a key factor in DNA methylation profile. They also suffer from higher cost, lower throughput compared with second generation sequencing methods.

1.3 Analysis of DNA methylation data

Recent technologies made an enormous and ever-growing amount of DNA methylation data to analyze. Development of statistical models and visualizing DNA methylation data are the core of analysis to identify DNA methylation differences across different samples, tissues or different patients with different diseases.

The advent of various techniques provides the opportunity to explore whole genome DNA CpG methylomes. Here, we focus on methods that analyze DNA methylation profiles at single base resolution. Analyzing sequencing based data usually includes alignment and post-alignment data processing to more accurately reveal important biological associations.

1.3.1 Alignment

Many softwares have been developed to align bisulfite sequencing reads. Since unmethylated cytosines are sequenced as thymines, this complicates the alignment process. Increasing the number of the mismatches increase noise and error in the alignment. On

the other hand, there is no reverse compliment for the reads with unmethylated cytosine. Hence, the followings are some error sources that could affect the quality of mapping reads.

- *Wrong alignment of the reads*
- *Existence of Single Nucleotide Variants (SNV)*
- *Sequencing errors*
- *Bisulfite failure*

Given these facts, some new alignment algorithms are developed for aligning bisulfite sequencing reads, such as BatMeth [LTL+12], Bismark [KA11], Bison [CSRM12], [KRCL+14], bisReadMapper [DPG+12], BSMap [XL09], BRAT and BRAT-BW [HPL+10], BS-Seeker [CCP10], and BSMooth-align [HLI12a].

1.3.2 Detection of differentially methylated loci and regions

Aligning the bisulfite sequencing reads helps to detect different types of variation in DNA methylation across genome. Some of known variation in DNA methylation are listed below [RDBB11].

- *Methylation variable position (MVP)*: A CpG site that shows different methylation level between different groups. Given recent findings on non-CpG methylation, potentially all Cs could be MVPs.

- *Differentially methylated region (DMR)*: A region of the genome at which multiple adjacent CpG sites show different methylation level. DMRs can occur in many different contexts, such as:
  - *iDMR*: imprinting-specific differentially methylated region,
  - *tDMR*: tissue-specific differentially methylated region,
  - *rDMR*: reprogramming-specific differentially methylated region,
  - *cDMR*: cancer-specific differentially methylated region,
  - *aDMR*: aging-specific differentially methylated region,
- *Variably methylated region (VMR)*: These regions are defined by increased variability rather than gain or loss of DNA methylation.
- *Allele-specific methylation (ASM)*: These are loci or regions that vary in DNA methylation because of different parent-of-origin, the presence of a polymorphism or due to a stochastic event.
- *Haplotype-specific methylation (HSM)*: This is a differentially methylated region that is defined by a set of co-inherited SNPs (a haplotype).

In Tables 1.1, 1.2, and 1.3 we gather a list of existing methods to find differentially methylated loci or regions with a brief description of their features.

Different high-throughput sequencing techniques are used to quantify the methylation state of a specific bisulfite-treated locus at single nucleotide resolution. Discovering the association between the cell mixture methylation profiles and a cancer type is a well studied area in EWAS. Some methods identify the differentially methylated Loci (DML)

Category	Software	Description	Ref
DML	MethylSig	- β-binomial modeling - Likelihood ratio test - Bisulfate sequencing data	[BJMV14]
	MethylKit	- Logistic regression based approach - Fisher test - RRBS	[AKL+12]
	DiffVar	- Emperical bays model - Z-test - DNA methylation array	[PO14]
	DMAP	- Mixed statistical test approach - Fisher exact test - Bisulfate sequencing data	[SCRM14]
	DSS	- β-binomial modeling - Wald test - Bisulfate sequencing data	[FCW14]

Table 1.1: Various methods for finding differentially methylated loci.

Category	Software	Description	Ref
DMR (L→R)	Metilene	- Binary segmentation based approach - 2D-KS test - Bisulfate sequencing data	[JKB+15]
	BSmooth	- Smoothing based approach - Log-ratio test - Bisulfate sequencing data	[HLI12b]
	Biseq	- Smoothing based approach - Hierarchical testing - Bisulfate sequencing data	[HDK13]
	Bisulfighter	- HMM based approach - Log-ratio test - Bisulfate sequencing data	[STM14]
	MOABS	- β-binomial modeling - Using CDIF instead of p-value	[SXR+14]
	RADmeth	- β-binomial modeling - Weighted Z-test - Bisulfate sequencing data	[DS14]

Table 1.2: Various methods for finding differentially methylated regions based on previously identified differentially methylated loci.

Category	Software	Description	Ref
DMR	DMRseq	- Regression based approach - FDR is controlled using [BH95] - Bisulfate sequencing data	[KCBI17]
	Getis-DMR	- β-binomial modeling - likelihood ratio test - Bisulfate sequencing data	[WCZ+16]
	Epipolymorphism	- using epipolymorphism statistic - high rate of epipolymorphism - Bisulfate sequencing data	[LCM+12b]
	ProbeLasso	- gather neighboring significant loci - feature2 - DNA methylation array	[BB15]
	QDMR	- Entropy based approach - using a methylation probability model - Platform free input data	[ZLL+11]
	DMAP	- Mixed statistical test approach - Fisher exact test - Bisulfate sequencing data	[SCRM14]
	swDMR	- Mixed statistical test approach - Integrated multiple hypothesis tests - Bisulfate sequencing data	[WLJ+15]

Table 1.3: Various methods for finding differentially methylated regions without finding differentially methylated loci.

between two groups [AKL+12, CHL+13, HCH13, SWZ+12, SW12, SW13] and some others identify differentially methylated regions (DMRs) that are associated with cancer or any other disease [ELC+06, ILAC+08a].

In Section 1.2 we review different experimental methylation profiling methods either report the ratio of methylated and unmethylated CpGs within sample per loci or the number of methylated and unmethylated reads per loci.

The ratio of methylated and unmethylated reads per loci is used to derive the marginal methylation percentage. In most of studies, the mean of marginal methylation percentage within a phenotypic group is considered to identify DMLs or DMRs [MSS14, LGBA+13, AKL+12, BB15, DS14, HLI12a, ZLL+11]. There are studies, using array data or sequencing data, in which other statistics like the variance of marginal methylation percentage is used to identify DMLs or DMRs [PO14, TGJ+16, AW13, RSS+16, SWC+17].

Many methods that concentrate on finding DMLs model read counts with a beta-binomial regression model [PFRS14, FCW14], while there are studies that do not consider within group variation [MSS14, LGBA+13, AKL+12, BB15, DS14, HLI12a, ZLL+11]. All of them estimate the mean and variance of read counts per loci and suffer from the following issues:

- High-dimensionality: in order to find DMLs they check every CpG site for computing the desired statistics. This bring a huge computational cost for proposed methods,
- Low sample size: to test the significance of a differentially methylated loci, a large

sample size is needed. Also, in order to model the count data per loci, the predictions will lose their accuracy, because the methods do not account for spatial correlation across genome,

- Interpretability: a single loci doesn't have an epigenetic effect, and biochemical changes in multiple locations might have an epigenetic effect on the expression of corresponding genes. This means that differentially methylated regions are more biologically relevant than a single loci.

Methods like [HLI12a, JML+12a, BB15, JKB+15, HSR15, WCZ+16] find DMRs by chaining DMLs in order to obtain inference over regions. But, the significance testing procedure in these methods faces challenges, and in particular they don't have any control on the false discovery rate (FDR) at the region level. The last category of methods, find DMRs using a region-based statistic. A hidden Markov model based method also define differentially methylated regions [SM15, STM14] based on the methylation profiles explained above. These methods mainly have the challenge of finding DMR boundaries across genome.

Besides, the current epigenome studies face two main challenges. First, comparing the methylation profiles of a cell-type population at single nucleotide resolution does not completely reflect the heterogeneity of DNA methylation modifications of single cells within the population. The spatial correlation information of the methylation status of adjacent CpGs is lost and only distribution of methylation status at single nucleotide resolution is reported. As shown in Figure 1.2 changes in the methylation pattern of single

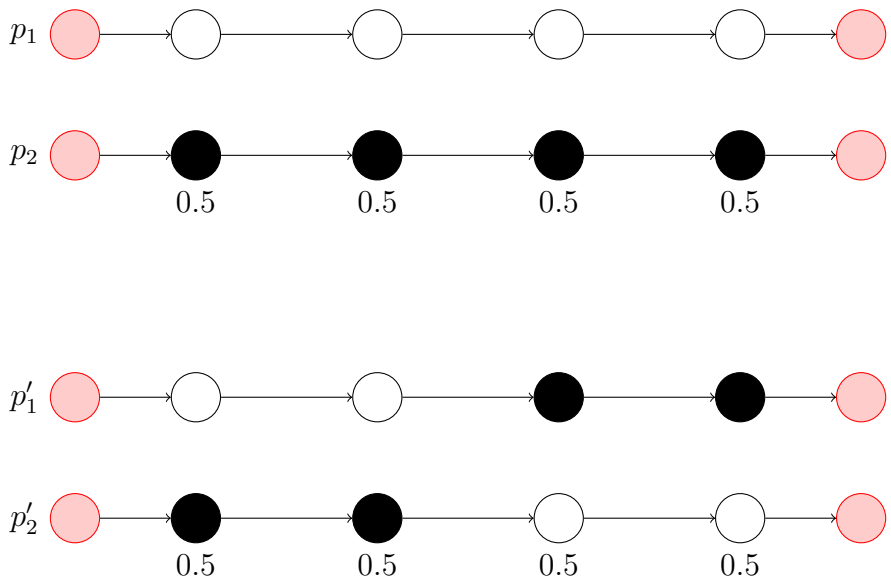

Figure 1.2: Methylation percentage per loci does not reflect the changes in the underlying pattern. Black and white circles shows methylated and unmethylated CpG sites respectively. Red circles represent start and end loci of selected region.

cells within a sample does not reflected by its methylation percentage per loci. Second, in order to call differentially methylated regions between different tissue types, i.e normal and tumor samples, we might face the tumor cell impurity because of inaccurate tissue sampling. This means there might be normal cells within tumor samples when tumor samples are collected. The impurity add some noise to the methylation information and complicates the process. This raise the need to infer the cell mixture more accurately.

1.4 Research contribution

In this dissertation, we develop two computational models to (i) reconstruct the underlying methylation profiles of normal and tumor samples, and (ii) detect regions with significant difference in their methylation profiles. These models are briefly outlined below and provided in detail through next chapters.

1.4.1 Reconstructing underlying methylation patterns using bisulfite-converted sequencing data

In chapter 2, we propose a method called methylFlow to infer the underlying methylation patterns and calculate their abundances across genome. It incorporates the spatial correlation between methylated sites and low coverage regions to improve its performance.

1.4.2 Finding regions with significant difference in their methylation profiles between normal and tumor samples

In chapter 3, we propose a method called MCFDiff to find regions that their methylation profiles are significantly different in normal and tumor samples. It uses the patterns inferred by methylFlow and their abundances. we systematically consider tumor content impurities to capture tissue or disease specific variations across genome.

CHAPTER 2

Reconstructing Methylation Patterns from High-throughput Bisulfite Sequencing Data

2.1 The general problem of DNA methylation composition

While single-cell methods to sequence bisulfite-converted DNA are currently under development [SLA+14b], the most reliable current method to measure DNA methylation at the base-pair level across the entire methylome is to bisulfite-convert and sequence DNA from a population of cells. A number of existing computational methods may then be used to calculate the percentage of DNA fragments that harbor a DNA methylation modification at specific genomic loci [HLI12b]. In many normal human tissues, for example, these percentages vary from the expected levels in a population of diploid cells

with identical DNA methylation modifications: 100% (where all cells in the population are methylated at a specific locus), 0% (where all cells in the population are unmethylated) or 50% (where only one chromosome in all cells in the population are methylated). For example, in the normal colon methylome, the majority of the methylome is partially methylated at a level of roughly 70-80% [HTB+11]. Similar patterns are observed in other human tissues [TBM+14], and tissues in other eukaryotes.

An obvious observation that follows from this is that cell populations in normal tissues are composed of epigenetically heterogenous cells. Furthermore, when comparing DNA methylation across different tissues, for example, colon normal tissue and colon tumor, Figure 2.1, or a population of stem cells to a population of somatic cells, e.g., fibroblast [LPD+09], differences in DNA methylation percentages at a specific locus is indicative of a shift in the epigenetic composition of these cell populations.

Computational and statistical methods to study the epigenetic composition of cell populations have been proposed based on the analysis of DNA methylation modifications at multiple consecutive genomic loci spanned by single sequencing reads [LCM+12b], where they analyzed DNA methylation modifications at each group of four contiguous CpG dinucleotides using sequencing reads that span all four CpGs. They then calculate the proportion of reads compatible with each of the 2^4 possible DNA methylation modifications over these four positions. They summarize these 2^4 proportions to define the *epipolymorphism* of each set of four contiguous CpGs.

While these approaches have yielded great insight into how cell populations differ

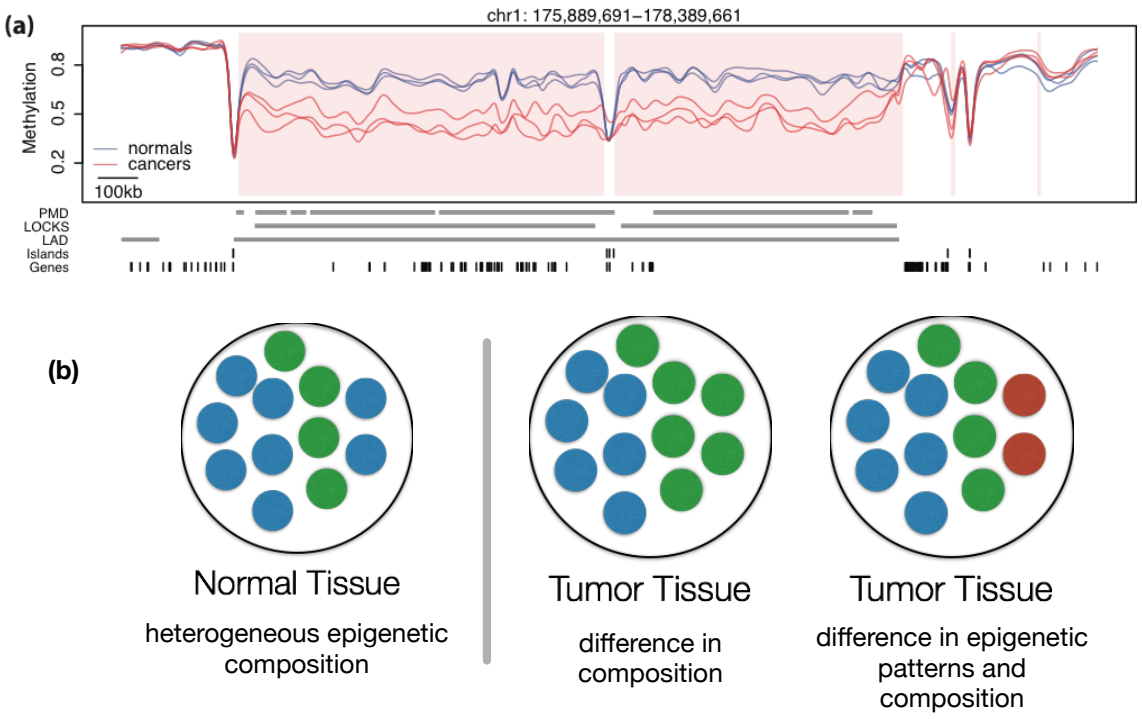

Figure 2.1: Differences in DNA methylation percentage at a given locus are indicative of a shift in the epigenetic composition of cell populations. (a) Base-pair level DNA methylation percentage estimate for three colon tumors and paired normal tissue (Figure from [HTB+11]). (b) Different shifts in the epigenetic composition of the cell population in a tissue lead to identical marginal differences of DNA methylation percentage at the base-pair level.

epigenetically across different tissues, they only provide a general summary of the epigenetic composition of these cell populations. For instance, distinguishing between the two types of cell population shifts illustrated in 2.1 is limited to those differences observed over four contiguous CpGs. In order to perform a comprehensive analysis of these cell population shifts, the ability to reconstruct cell-specific methylation patterns over longer genomic spans is required.

In this chapter we present methylFlow, a novel computational method to reconstruct cell-specific patterns using reads obtained from sequencing bisulfite-converted DNA based on network flow algorithms. We report on a simulation study characterizing the behavior of our method. We then present an application of this method using ultra-high coverage targeted sequence in a colon cancer study [HTB+11], and on whole genome sequencing of fully differentiated B-cells and KSL and CLP progenitor cells [KKTM+13]. We also perform a validation study using bisulfite-converted DNA from single-cells [SLA+14b]. We believe that this method will allow for increased understanding of the role of epigenetic heterogeneity at the cell population level in gene regulation. This work is in press in Bioinformatics [DMCB16].

2.2 methylFlow algorithm

Our method uses sequencing reads from bisulfite converted DNA to reconstruct heterogeneous cell populations by assembling cell type-specific methylation patterns spanning multiple CpGs from read overlaps (Figure 2.2). It jointly reconstructs these methy-

lation patterns and quantifies their abundance in heterogenous cell populations.

Our method assumes a set of aligned reads from a bisulfite converted DNA sequencing run sorted by genomic starting location. For this paper, we only analyze on cytosine methylation so that each CpG overlapped by a given aligned read can be determined to be methylated (M) or unmethylated (U). Each aligned read r is thus associated with a starting genomic position l_r and a specific methylation pattern over the CpGs it spans. The latter is defined by set $p_r = \{\langle \text{offset}_i, m_i \rangle\}$ where offset_i specifies the location of the CpG based on the read start position l and $m_i \in \{M, U\}$ specifies the methylation status of the ith CpG covered by the read.

2.2.1 Overlap graph

Following existing methods from viral population reconstruction [EPM+08], we build a read overlap graph based on read starting location and compatibility of methylation patterns. Read overlap graph $G_o = \{V_o, E_o\}$ contains a node (l_r, m_r) for each aligned read r (as described above) originating from position l_r with methylation pattern m_r. A directed edge (r, r') between from source node r to target node r' is included in the graph if it satisfies the following:

1. $l_r < l_{r'}$: the starting position of the source is to the left of the starting position of the target, and

2. methylation patterns m_r and $m_{r'}$ are equal on overlapping cpgs (if any), and

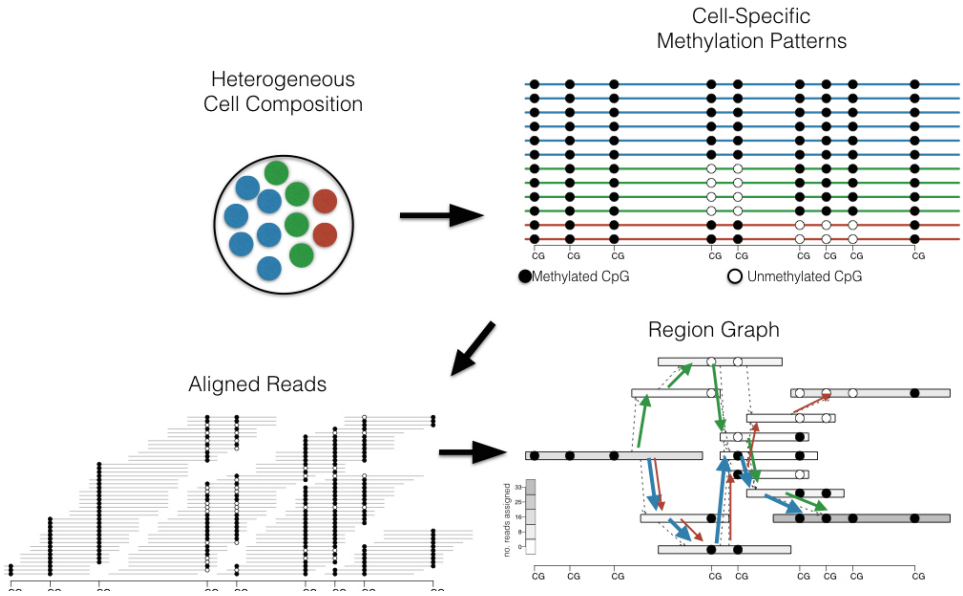

Figure 2.2: **Overview of methylation pattern estimation**: We assume that samples are obtained from cell populations (top left) that are epigenetically heterogeneous as determined by distinct CpG methylation patterns along a genomic region (top right). Reconstruction is based on the overlap of bisulfite converted reads to a reference genome (bottom left). Read overlaps and methylation calls are used to define a region graph (bottom right). Based on coverage (the number of reads originating in each region), a minimum cost network flow problem to estimate the number and abundance of methylation patterns (paths in the graph).

3. there is no path between r and r' in the graph unless this edge is present, so there are no paths between ancestors of the target node.

We denote the number of reads originating at position l with methylation pattern m as $c_{l,m}$. This is the same construction as [EPM+08] with methylation patterns taking the place of variants in reads obtained from virus sequencing.

2.2.2 Coverage normalization

To build a statistical model, we first normalize the coverage in the overlap graph to account for variability introduced by non-uniform sequencing coverage and copy number variations. The number of reads for node (l, m) is normalized as follows:

Let $c_l = \sum_m c_{l,m}$ be the total number of reads originating in position l, and let $\mu = \text{median}_l c_l$ across a connected component of the graph. The normalized number of reads for node (l, m) is defined as $y_{l,m} = \frac{c_{l,m}}{c_l} \times \mu$. After normalization, all positions l have total normalized number of reads equal to s.

2.2.3 Region graph

Building a statistical model over position-specific coverage is difficult due to variability in low coverage experiments. To alleviate this issue, we use the fact that DNA methylation modifications show high spatial consistency [HP75] and convert the read overlap graph G_o to a region graph $G = \{V, E\}$ by collapsing non-branching paths in the overlap graph G_o so that nodes now span multiple genomic loci. The total, normalized,

number of reads originating in region $v \in V$ is defined as $y_v = \sum_{(l,m)\in v} y_{l,m}$. We define the starting position l_v of region $v \in V$ as $\min_l\{(l,m) \in v\}$, i.e., the smallest position l over nodes of the overlap graph G_o contained in region v. We also merge read methylation patterns into region methylation patterns (since by definition these are consistent), so that each region also defines a methylation pattern $p_v = \bigcup_{(l,m)\in v} m$.

To complete the region graph we add a source node s connected to every region in the graph without an incoming edge, and a sink node t connected to every region in the graph without an outgoing node. Cell-specific methylation patterns p are defined by paths in the region graph from start node s to end node t each with a specific methylation pattern defined by the methylation patterns of the regions in the path. We denote the abundance of cell-specific methylation pattern p, equivalently path p, as θ_p.

Given this notation, the total abundance of methylation patterns consistent with region $v \in V$ is given by the sum of the abundances of paths that include v: $\sum_{\{p:p\ni v\}} \theta_p$. Note that by construction the following three sets are equal

$$\begin{aligned}\{p : p \ni v\} &= \bigcup_{\{u:(v,u)\in E\}} \{p : p \ni (v,u)\} \\ &= \bigcup_{\{u':(u',v)\in E\}} \{p : p \ni (u',v)\}\end{aligned}$$

This just states that the set of paths going through node $v \in V$ can be enumerated as the union of all paths going through all outgoing edges $\{(v,u) \in E\}$, or as the union of all paths going through all incoming edges $\{(u',v) \in E\}$. This implies

$$\sum_{\{p:p\ni v\}} \theta_p = \sum_{\{u:(v,u)\in E\}} \sum_{\{p:p\ni(v,u)\}} \theta_p \tag{2.1}$$
$$= \sum_{\{u':(u',v)\in E\}} \sum_{\{p:p\ni(u',v)\}} \theta_p$$

We will use relationship 2.1 in our estimation procedure.

2.2.4 Statistical model

We introduce a statistical model that motivates our reconstruction algorithm based on fitting the normalized observed number of reads y_v originating in region $v \in V$ of the region graph. This is similar to statistical models used in viral population reconstruction methods [EPM+08], or RNA-seq [BJMV14].

Our goal is to estimate $\mathbb{E}y_v$, the *expected number of reads originating from region* v as a function of the abundances θ_p of unobserved methylation patterns p. In order to do so, we need to define the *effective length* of region v in pattern p which we denote ℓ_{vp}. Since every methylation pattern p corresponds to a path p through region graph G, the effective length of region $v \in V$ within pattern p is determined by outgoing edge $(v,u) \in p$. Specifically, the effective length $\ell_{vp} = \ell_{vu} = l_u - l_v$ for every path p such that $(v,u) \in p$ and l_u and l_v are the starting positions of regions u and v respectively. As $\mathbb{E}y_v$ corresponds to the expected number reads originating in region v, it is proportional to the effective length of the region.

Using this notation we model

$$\mathbb{E}y_v = \sum_{\{p:v\in p\}} \ell_{vp}\theta_p$$
$$= \sum_{\{u:(v,u)\in E\}} \ell_{vu} \sum_{p:p\ni(v,u)} \theta_p.$$

Using a regularized method of moments, we estimate parameters θ_p corresponding to every possible path p through region graph G by minimizing loss function

$$\min_{\theta_p} \sum_{v\in V} \left| y_v - \sum_{\{u:(v,u)\in E\}} \sum_{\{p:p\ni(v,u)\}} \theta_p \right| + \lambda \sum_p \theta_p \tag{2.2}$$

where p ranges over all paths in the region graph and λ is a regularization term. This formulation is similar to the IsoLasso [LFJ11] model defined for RNA-seq transcript assembly and quantification. In our case, we use absolute loss to implement robust median regression (instead of least squares regression).

The regularized method of moment estimator yields a linear optimization problem over a large number of unknowns, namely, the number of possible paths through the region graph $G = (V, E)$. We follow the idea behind the FlipFlop method [BJMV14] developed for transcript assembly from RNA-seq data using regularized loss functions. We do not explicitly solve over all possible paths p, instead we introduce variables f_{vu} for each edge $(v, u) \in E$ defined as $f_{vu} = \sum_{\{p:p\ni(v,u)\}} \theta_p$ and rewrite the method of moments estimating equation 2.2 as

$$\min_{f_{vu}} \sum_v \left| y_v - \sum_{\{u:(v,u)\in E\}} \ell_{vu} f_{vu} \right| + \lambda \sum_{u:(u,t)\in E} f_{ut} \tag{2.3}$$

with regularization term $\lambda \sum_p \theta_p$ in Equation 2.2 rewritten using edge variables f_{ut} where t is the sink node in G. Since all paths p end at sink node t we have $\sum_p \theta_p = \sum_{\{u:(u,t)\in E\}} \sum_{\{p:p\ni(u,t)\}} \theta_p = \sum_{\{u:(u,t)\in E\}} f_{ut}$.

To ensure variables f_{uv} correspond to the sum of methylation pattern abundances, equivalently paths, that include edge (v, u), we add the following constraints which follow directly from Equation 2.1:

$$\sum_{\{u:(v,u)\in E\}} f_{vu} = \sum_{\{u':(u',v)\in E\}} f_{u'v} \tag{2.4}$$

Since we are using absolute deviation as our method of moments estimating criterion we obtain a linear optimization program with linear constraints. It corresponds to a network flow problem where variable f_{uv} is the flow assigned to edge uv and constraints in Equation 2.4 correspond to standard network flow balance constraints.

Our software takes as input a set of aligned bisulfite-converted reads which may be obtained using existing bisulfite-aware read mappers [KA11, HLI12b]. It assumes the input is in SAM files as produced by the *Bismark* [KA11] aligner. We solve the dual problem [Lue73] of the above linear optimization problem using the GLPK [Mak08] linear programming (LP) solver and the LEMON [DJK11] C++ library to represent and manipulate the read overlap and region graphs. Source code is freely available at http://github.com/hcorrada/methylFlow as C++ source code, and includes a small R package for reading, visualizing and manipulating resulting methylation patterns and their abundances.

2.3 Evaluation

Error metrics are based on first matching each simulated pattern with one or more of the patterns estimated by our method, and then determining error in abundance estimates or methylation calls for the estimated patterns based on this matching. We note that these error metrics are only applicable in simulation settings where true patterns and abundances are known.

To match estimated patterns to simulated patterns we build a bipartite graph $(\{S, T\}, E)$: each node in S represents simulated pattern with abundance θ_i while set T has a node for each estimated pattern with abundance θ_j. Each edge connecting node $i \in S$ to node $j \in T$ has weight w_{ij} equal to the total number of methylation call differences between patterns i and j. w_{ij} equals the number of overlapping CpGs with different methylation status plus the number non-overlapping CpGs. w_{ij} is zero if pattern $i \in S$ exactly matches pattern $j \in T$ in all their methylation calls and have no non-overlapping CpG sites. We then solve a minimum weight matching problem on the bipartite graph so that the matching node of simulated pattern i is the estimated pattern j in set T which has the smallest weight w_{ij} among neighbors of node i. Below we use indicators x_{ij} equal to 1 if i and j are matched and is equal 0 otherwise.

To better understand the behavior of error metrics, we report errors for multiple thresholds based on weights w_{ij}. If the number of methylation call errors between estimated and simulated patterns is above the threshold, then the match is not used when

calculating error metrics below.

2.3.1 Abundance error

Based on the resulting matches for each estimated pattern, we determine a score to evaluate how well our algorithm predicts the average abundance of patterns as follows:

$$\text{Average Abundance error} \quad = \quad \frac{1}{n} \sum_{i \in S} \sum_{\forall j \in T : x_{ij}=1} \left(\frac{\theta_i - \theta_j}{\theta_i} \right)^2$$

This error metric shows how well our algorithm predicts the abundance of simulated patterns. Since the abundance of patterns are different in different settings, we compare the abundance of true patterns by their matched estimated patterns and scale them by the abundance of true patterns. This gives us the relative error between the abundance of true and estimated patterns.

2.3.2 Methylation call error

Our second error metric evaluates the prediction of methylation calls for estimated patterns. We use the same bipartite graph and same matching problem we did for calculating the average abundance error. Hence, based on our bipartite graph and the matches for every simulated pattern, we determine a score to evaluate how well our algorithm predicts the methylation patterns as follows:

$$\text{Average Methylation call error} \quad = \quad \frac{1}{n} \sum_{i \in S} \sum_{j \in T : x_{ij} = 1} w_{ij}$$

The average methylation call error shows how well the estimated patterns are matched to true patterns. It is equal to the average number of methylation status errors between simulated and their matched estimated patterns. Since we are using these weights to discard matched patterns, we expect that the methylation call error is less than the corresponding threshold.

2.3.3 Minimum cost network flow error

Our third error metric evaluates performance based on both methylation call error and pattern abundance estimates. In this metric there is no threshold is used to filter pattern matches between simulated and estimated patterns. Instead, we run a minimum cost network flow problem that matches every true pattern to a set of estimated patterns on the same bipartite graph.

However, we add constraints such that the sum of outgoing flows from every node in true pattern set is equal to the abundance of corresponding true pattern and the sum of incoming flow to every node of estimated pattern is also equal to the abundance of corresponding estimated pattern. f_{ij} is the amount of flow goes from true pattern i to pattern j such that the sum of all f_{ij} s are minimized. Then we compute the expected methylation call error by multiplying the probability of pattern i and j being matched, i.e., what percentage of the abundance of pattern i is covered by pattern j (Figure 2.3(right))).

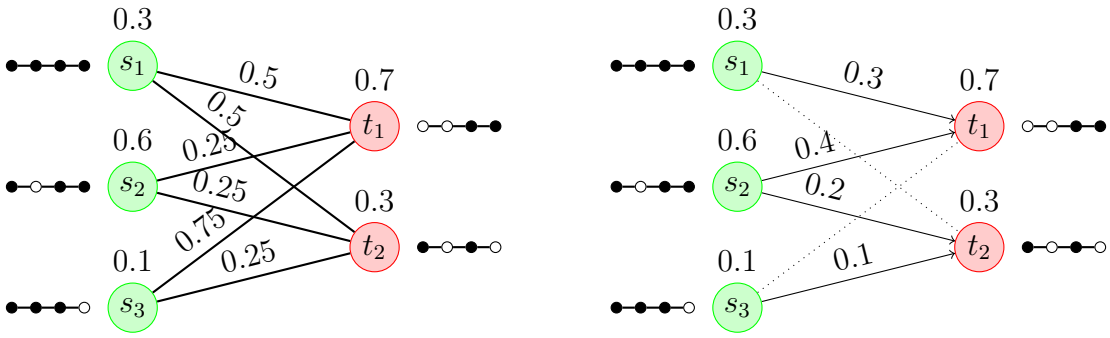

Figure 2.3: Bipartite graph built to solve a minimum weight matching between simulated patterns and estimated ones(Left), Minimum weight matching solution (Right)

The cost of our network flow in our bipartite graph is equal to sum of the weights of every pair (i, j) multiplied by the amount of the corresponding flow. This metric evaluates how well our algorithms predict both methylation calls and the abundances.

$$\begin{aligned} \text{cost of network} &= min \sum_{ij} w_{ij}.f_{ij} \\ \text{s.t} \sum_{j} f_{ij} &= \theta_i, \forall i \in S \\ \sum_{i} f_{ij} &= \theta_j, \forall j \in T \\ f_{ij} &\geq 0, \forall i \in S, \forall j \in T \end{aligned}$$

Here f_{ij} is the amount of flow from node $i \in S$ to node $j \in T$ and corresponds to the fraction of the abundance of simulated pattern i matched to estimated pattern j. The cost of our network flow in our bipartite graph is equal to sum of the weights of every pair (i, j) multiplied by the amount of the corresponding flow. It is a measurement to evaluate

how well our algorithms predict both methylation calls and the abundances.

2.4 Simulation study

We performed a simulation study to evaluate the performance of our algorithm based on how well it predicts the number of cell-specific patterns, how many methylation calls are reconstructed correctly in each pattern and how well it predicts the abundance of each pattern.

Our simulation has two separate steps: first, we simulate n cell-specifc methylation patterns over a genomic region and then simulate the sequencing process to produce short reads using uniform samples across the simulated pattern. We call these simulated patterns as true patterns.

2.4.1 Simulating true patterns

We use three different settings of increasing difficulty to simulate the cell-specific true patterns:

- *Simple:* Number of true patterns is $n = 2$, one with 75% of abundance and the other with 25% of abundance. The two patterns share almost no CpGs with the same methylation status.
- *Moderate:* Number of true patterns is $n = 4$, with 15%, 15%, 30% and 30% of abundances respectively. Patterns share a moderate number of CpGs with the same methylation status.

- *Hard:* Number of true patterns is $n = 10$, all with 10% of abundances and only a small number of CpGs have distinct methylation status across patterns.

Methylation patterns are simulated such that nearby CpG sites are likely to have similar methylation status based on their genomic distance. Specifically, two CpG sites with distance d have the same methylation status with probability:

$$f(d) = 1 - \frac{1}{1 + e^{-10 \times (d - \text{corrDist})}} \tag{2.5}$$

where $\text{corrDist} = 20$ is a parameter that controls methylation status correlation between two consecutive CpG sites. To simulate a pattern, the methylation status of the first CpG is set uniformly at random, and each subsequent site is set to the same status as the previous CpG with probability $f(d)$, otherwise it is set uniformly at random. CpG locations for 1750 CpGs were obtained from a whole genome bisulfite sequencing dataset [10].

2.4.2 Simulating short reads

For simulating the sequencing process, we randomly select a pattern with probability proportional to its abundance. Read start position is uniformly chosen at random. Every CpG site is sequenced without any error with probability $1 - \text{error}$. If parameter $\text{error} = 0$, then methylation pattern of every short read is exactly the same as its true pattern. A total of $\frac{\text{coverage} \times \text{dnaLength}}{\text{readLength}}$ short reads are generated in each simulation setting.

Parameters for coverage, short read length, number of CpG sites in the simulation region are varied over a specified range for each simulation setting as we test the behavior of the algorithm as a function of these three parameters. Otherwise, these parameters are

held to constant values 20 for coverage, 100 for number of CpGs, and 70 bp for short read length.

When testing the effect of each of these parameters on the performance of our algorithm, coverage is varied from 5 to 20, short read length varies from 50 to 250, the number of CpG sites varies from 75 to 150.

2.5 Results

2.5.1 Simulation results

To evaluate the performance of our algorithm, we need to consider the average abundance error and average methylation call error simultaneously as abundance errors may increase as more stringent thresholds are placed on methylation call error. In Figure 2.4 (A-C) average abundance error versus average methylation call error is shown for the *moderate* simulation setting as we test the effect of coverage, number of CpGs and read length on the reconstruction algorithm. We show the effect of using different methylation call error thresholds on matched patterns and the error metrics.

We observed that abundance error decreases and methylation call error increases as read length and coverage increases. Increasing coverage and read length help to decrease the problem complexity and have a more accurate reconstructed pattern. In particular, we observed that while doubling coverage from 5x to 10x significantly decreases error, doubling coverage from 10x to 20x has much less pronounced effect. Increasing the number of CpGs increases the complexity of the problem and both average abundance

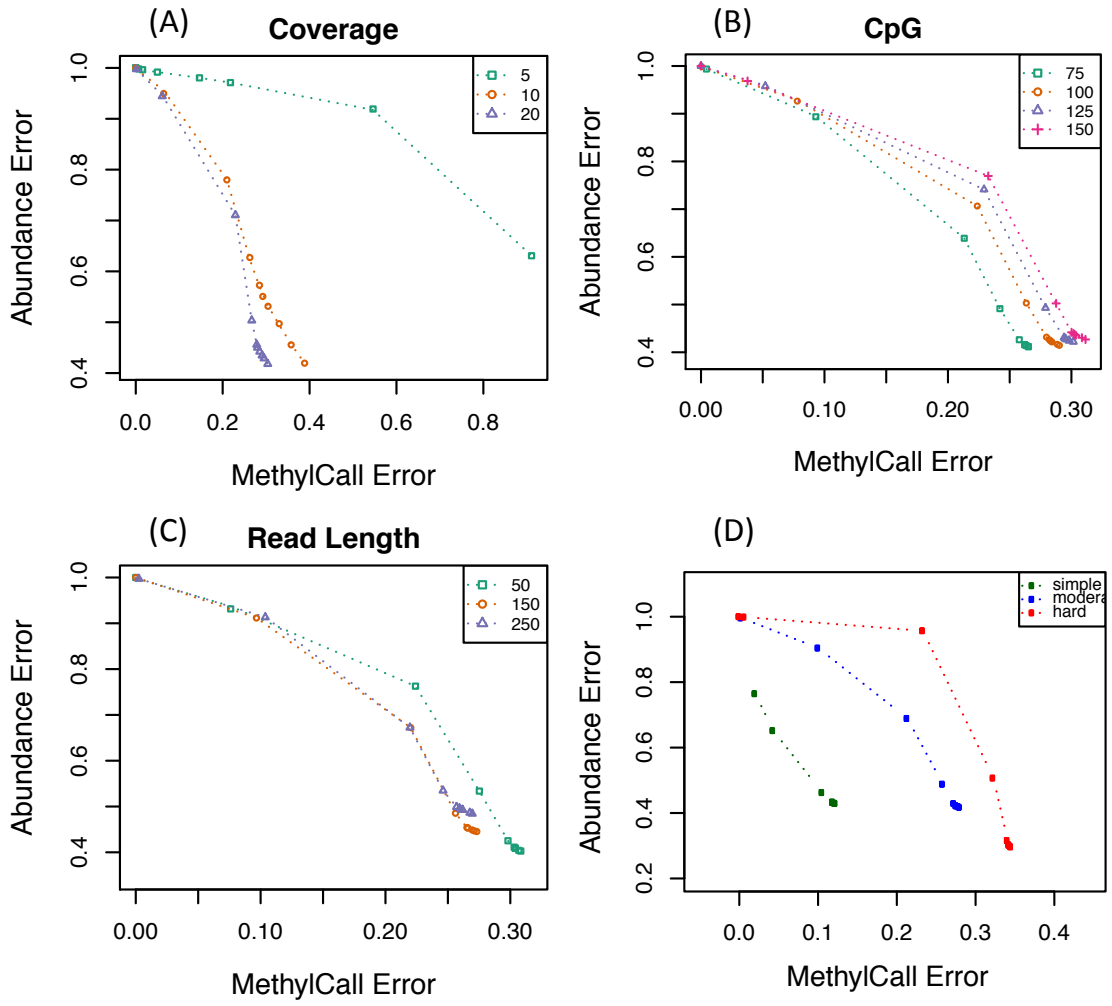

Figure 2.4: Average abundance error vs. average methylation call error in different setting of simulation and various thresholds in moderate complexity of patterns. Points correspond to increasing threshold on methylation error between matched patterns. Panels show the effect of different (A) coverage (B) number of CpG sites, and (C) short read length on error. (D) Average abundance error vs. average methylation call error in different simulated pattern complexity with fixed coverage, number of CpG and short read length.

and average methylation call error increase. In this Figure 2.4D, CpG, read length and coverage are fixed, while pattern complexity is varied. Methylation call error and average abundance error increases as the complexity of patterns increases.

Figure 2.6 shows the minimum cost network flow error metric for our three simulation settings as a function of coverage, number of CpGs and read lenght. By increasing coverage and read length, as we expect, the complexity of reconstruction decreases and the error decreases consequently with error decreasing sharply from 5x to 10x coverage, with slower decrease after that. Since we are reading CpG positions from a real data set, increasing the number of CpG sites are expanding the genomic region but the density of CpG sites remains almost the same. Hence, we see a slight increase in the error.

Our algorithm is less dependent on the number of CpGs. The performance of our algorithm slightly decreases by increasing number of CpGs and that is mainly because of the extension in the length of reconstructed region. Coverage has more significant effect on the performance of our algorithm. We see more errors in low coverage regions. Also our algorithm performs better if we could have longer sequencing reads and less ambiguity between cell-specific patterns.

We also evaluate sensitivity of our algorithm to error in sequencing CpG methylation status. Figure 2.5 shows that the minimum cost flow error increases by increasing the probability of noise. Note that when the noise level is 50, i.e $p(error) = 0.5$ in sequencing, then the short reads are random, and thus the output will be random, i.e., the minimum cost flow error is around 0.5. The regularization parameter λ indirectly con-

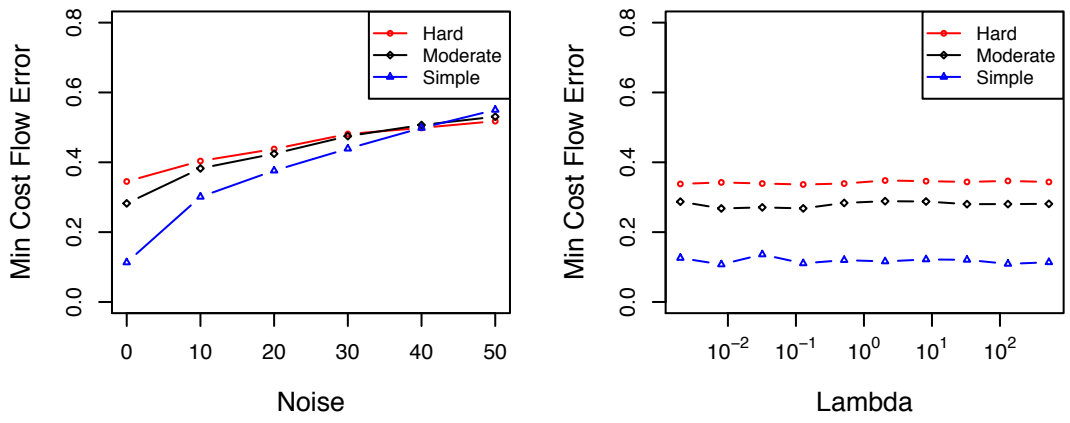

Figure 2.5: (Left) Sensitivity to the noise level in the input. Minimum cost flow error for various noise levels, probability error in sequencing, of the input data. (Right) Sensitivity to regularization parameter λ. Minimum cost flow error for various values of the regularization parameter.

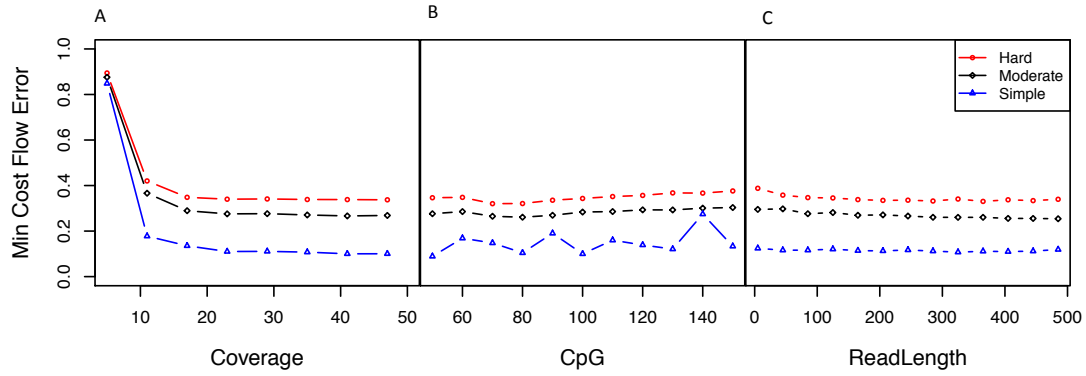

Figure 2.6: Minimum cost flow (MCF) error for three different simulation settings with different complexity. (A) The effect of coverage on MCF error. (B) The effect of the number of CpG sites on MCF error. (C) The effect of short read length on MCF error.

trols the number of estimated patterns. As can be seen in Figure 2.5, the methylFlow algorithm is not sensitive to regularization parameter in a wide range of λ. In particular, methylFlow achieves consistently good performance with λ varying from 0 to 100 for different types of simulated data. This is an interesting property because we do not need to tune the regularization parameter very precisely in the real data sets.

2.5.2 Single cell results

We also evaluated our algorithm using a single-cell bisulfite sequencing (scBS-seq) dataset [SLA+14b]. Smallwood et al performed scBS-Seq on mouse embryonic stem cells (ESCs) cultured either in $2i$ ($2i$ ESCs) or serum (serum ESCs) conditions to determine whether scBS-Seq can reveal DNA methylation heterogeneity at the single-cell level. In order to evaluate the performance of our algorithm, we ran our algorithm separately on 2i and serum single cell data sets that were aligned to GRCm38 mouse genome using *Bismark* in single-end mode. We observed that our algorithm reconstructed a single pattern for 93% of the regions covered and obtained two patterns for 6% of covered regions. We also ran methylFlow on a mixture of $2i$ and Serum samples. For 79% of regions, methylFlow recovered exactly the same number of patterns as expected from the mixture. For 17% of the regions, methylFlow recovered one fewer pattern than expected from the mixture. This result suggests that methylFlow is capable of identifying long-range methylation patterns in a epigenetically heterogeneous cell population. Unfortunately, coverage for single-cell sequencing data is too low to reliably estimate our

algorithm's performance in estimating the abundance of patterns in a heterogeneous cell population.

2.5.3 Whole genome bisulfite sequencing results

We also applied our method to a whole genome bisulfite sequencing data on mouse wild-type activated B cells and mouse CLP and KSL cells [KKTM+13] aligned using *bismark_v0.11.1*. The length of short reads are 50 bp and all analyses were done relative to the mm10/GRCm38 assembly of the mouse genome. We were able to reconstruct cell-specific methylation patterns with median length 200 to 750bp (Figure 2.7). Patterns longer than 750bp were reconstructed in each sample.

Again, we report the performance of our method using the marginal methylation percentage of estimated patterns at CpGs to those obtained from short reads directly. a Figure 2.7, panels C and D, shows that marginal methylation estimates from patterns to those obtained from short reads (correlation 0.92 and 0.91).

2.5.4 Targeted bisulfite sequencing

We applied our method to data from a targeted bisulfite sequencing experiment on three colon tumors and matched normal tissue [HTB+11]. Read lengths in this dataset are either 73bp or 80bp, and we used the provided read alignments with a post-processing script (available upon request) to resolve strand-aware methylation status as reported by the alignment tool before constructing the read overlap graph. We were able to reconstruct

cell-specific methylation patterns with median length 110 to 200bp. Patterns longer than 350bp were reconstructed in each sample.

Since the true cell type methylation patterns are not known, the error metrics presented in Section 2.3 are not applicable. In real datasets, we instead report performance by comparing marginal methylation percentage of estimated patterns at CpG level to those estimated from short reads directly. Since we are not using this information in reconstructing patterns, similar beta value (marginal methylation percentage) could evaluate the performance of our method.

As illustration of the type of inference provided by our method, we show in Figure 2.8 the patterns estimated for a differentially methylated region. We obtained the most differentially methylated region in chromosome 13 using *bumphunter* software [JML+12b]. The figure depicts the estimated patterns in every sample along with their abundances in this hyper-methylated region. We observed that populations in tumor are more heterogenous than in normal (which itself is heterogenous to a small degree), and that dominant patterns in the normal population are present in the tumor population.

2.6 Discussion and conclusion

We have presented an algorithmic method to reconstruct cell-specific methylation patterns using overlap and coverage of sequencing reads of bisulfite-treated DNA. Our method allows researchers to probe intra-cellular epigenomic heterogeneity from a standard sequencing experiment of pooled cells. This work opens new avenues in the analysis

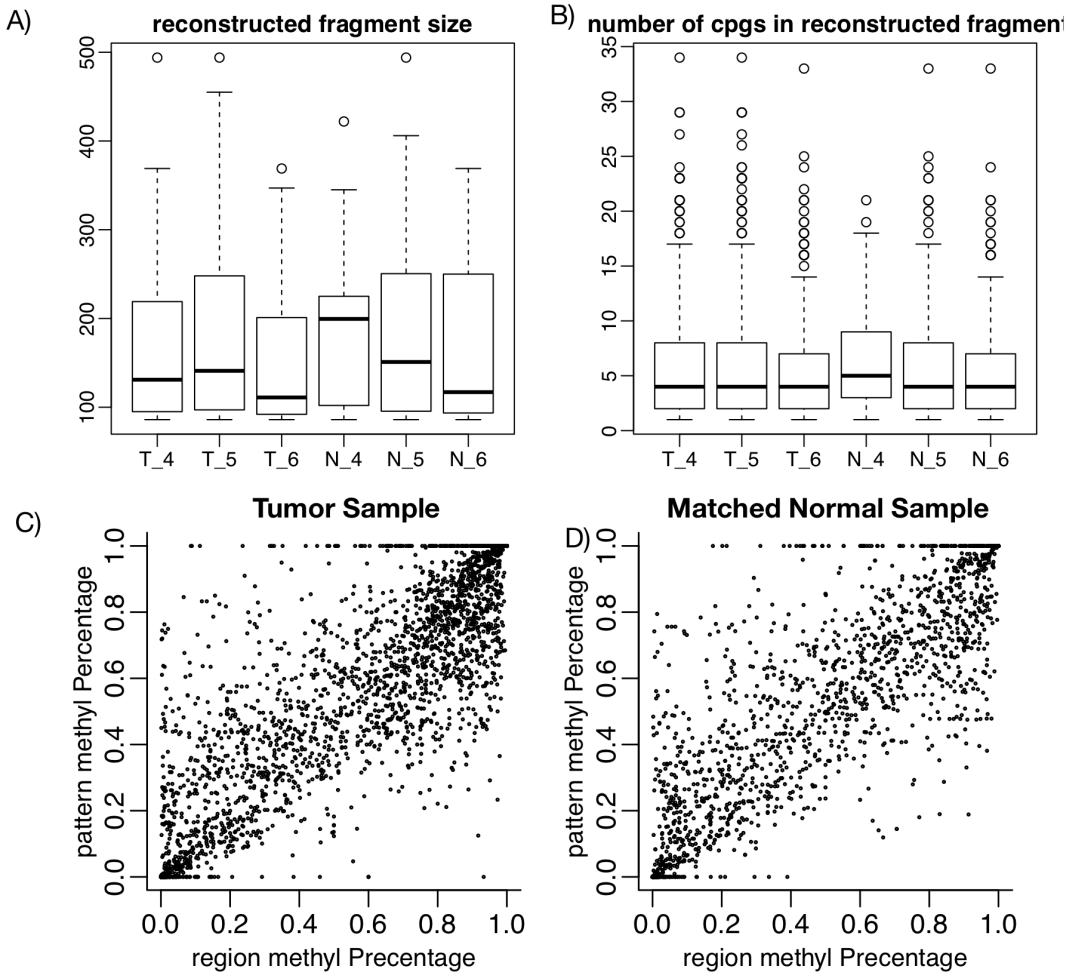

Figure 2.7: Pattern estimation in targeted bisulfite sequencing of three colon tumors and matched normal tissue in chromosome 13. (A) Length distributions of reconstructed cell-specific methylation patterns. (B) Distributions of the number of CpGs per reconstructed cell-specific methylation patterns. (C and D) CpG methylation percentage estimated from reconstructed cell-specific methylation patterns *(pattern methyl Percentage)* vs. observed CpG methylation percentage (region methyl Percentage) for a single tumor sample and matched normal.

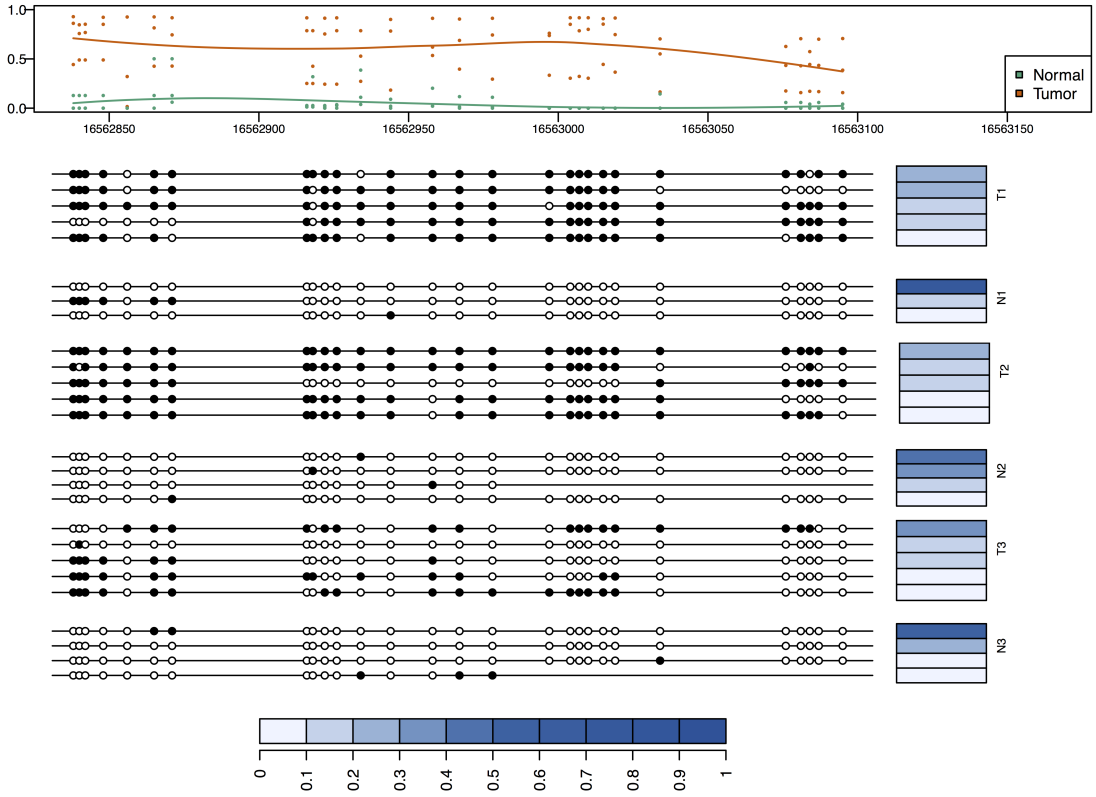

Figure 2.8: Differentially methylated region between colon tumors and matched normal pairs with corresponding patterns and their abundances across different samples. The top panel shows the marginal methylation percentage and the average curve of marginal methylation percentage as estimated by *bumphunter*. The bottom panel depicts the methylation patterns of samples. Blue bars represent the abundance of corresponding patterns. The abundances are normalized by sum of the abundances of all patterns in selected region.

of epigenomes as statistical extensions to our work here can start addressing questions of differential presence of cell-specific methylation patterns across phenotypes of interest, and begin to understand specific changes in the epigenomic complexity of cell communities.

Some cell-devoncolution methodologies like methylPurify [ZZW+14] uses regions with bisulfite reads showing discordant methylation levels to infer tumor purity from tumor samples alone. They do not assume any genomic variation information or prior knowledge from other datasets. Some restrictions in their method is that they infer the fraction of normal cells within tumor samples by assuming that there are only two component of normal and tumor cells. They also detects differentially methylated regions from tumor and normal cell lines, under assumption of homogeneous tumor and normal cell lines. Since they only consider CpG sites, they expect to see consistent methylation level within short intervals(300bp). [HAK+12] also present a statistical method to infer the distribution of different cells in a subpopulation and similarly, methylMix [Gev15] developed a computational algorithm to identify differentially methylated genes that are also predictive of transcription. The two latter methods used the Illumina Infinium HumanMethylation $27k$ or $450k$ BeadChip.

Our simulation study shows that our methodology is sensitive, as other similar methods for sequencing data, to sequencing depth. Figure 2.4 indicates that that our approach works well at depths of 10x or greater. Our software outputs total coverage per connected component in the region graph. In practice, regions that have less than 10x

average coverage should be removed for downstream analysis.

While we have not applied our method to Reduced Representation Bisulfite Sequencing data [MGB+05b], it should directly apply as presented in this manuscript, under the same caveats regarding coverage discussed above. RRBS is designed for high density regions, and usually tends to yield higher coverage than WGBS, which makes it suitable for our methodology. Our method requires single fragment methylation calls as input as provided by sequencing assays, which makes it unsuitable for array-based assays, tiling [ILAC+08b] or based on Bisulfite conversion as signal in this case depends on the number of methylated and unmethylated fragments in a pool of cells [BBT+11]. While we believe that our normalization method somewhat alleviates coverage biases stemming from sequence or amplification effects, a normalization model that incorporates relevant technical covariates could significantly improve any instability in our estimation method stemming from coverage biases.

As presented here, our method only performs reconstruction of patterns for single samples (e.g., a single tumor sample). A consideration for future work is to establish an algorithm that jointly estimates cell-specific methylation patterns across samples. However, our graph matching procedures described in Section 2.3 can be used to associate estimated patterns across individual samples in subsequent analyses.

CHAPTER 3

Finding Regions with Differential Methylation Composition using Bisulfite Sequencing Data

3.1 Introduction

Cells in our body have almost identical DNA sequence since they stem from a similar single embryonic cell. However, they have different DNA methylation modification profile depending on the cell-type, age, disease or other biological states. Different phenotypic properties in different cell-types and in different tumor tissues (inter tumor heterogeneity), as well as the subclonal diversity within a cell-type or tumor tissue (intra tumor heterogeneity) correspond to their epigenetic modification profiles within a cell population. These modifications are not fixed and vary over time [LCM+12a]. DNA

methylation is identified as one of the most important epigenetic modifications. Variations in the DNA methylation modification profiles are not consistently reflected by phenotypic variations while some make distinct phenotypic changes. Studies have found that variations in methylation levels (hyper-methylation) in promoter-related CpG islands leads to silencing of downstream tumor suppressor genes in many type of cancers. Furthermore, instability of epigenome in some regions has been associated with cancer heterogeneity. This influences the functionality of cells and disease state of the cell-type population as a result. Therefore, a comprehensive understanding of cell mixture methylation profile and its association to health and disease, age or other state of interest has become a priority in epigenome-wide association studies (EWAS).

In this chapter we propose a method, which is called MCFDiff, to find the regions of differential methylation composition (RDMC). We infer the underlying DNA methylation patterns within a heterogeneous cell population using methylFlow algorithm, which has been defined in Section 2.2. We compare the composition of underlying patterns using a new statistical method. We consider both the underlying patterns of DNA methylation modification and the abundance of each pattern within samples in our model. Any significant changes in the abundance, the pattern or both is considered as RDMC.

We improve the prediction of RDMC by applying two different strategies. First, we employ the reconstructed underlying pattern to retain the spatial correlation of DNA methylation modifications in cells within a heterogeneous population. Second, regarding the underlying patterns, we infer the impurity of the cell-type population by leveraging our

proposed similarity metric, and improve the prediction of RDMC. Both of the strategies improve false negative rate along as sensitivity and specificity. See Section 3.2 for more details about MCFDiff method.

We evaluate the accuracy of the proposed method for finding RDMC by comparing its results with the results of DMRseq method which is introduced in [KCBI17]. We first evaluate the performance of our method by the simulation. We employ the simulation data including the simulated patterns for normal and tumor replicates and also the information of the regions that have different methylation pattern composition. The results discussed in Section 3.4.1 show the accuracy of our proposed method. In particular, we improve the sensitivity and specificity along with smaller type I and type II error compared with other methods. We also evaluate our method using real data. The results given in Section 3.4.2 show that in regions with experimental evidence, MCFDiff performs meaningfully better than DMRseq in predicting regions with significant change in their DNA methylation modification.

3.2 Materials and method

3.2.1 Constructing methylation patterns

In our proposed method for finding RDMC, we first use methylFlow algorithm [DMCB16] for reconstructing methylation profiles. Let's assume we have k number of normal tissues and their matched tumor tissues. We first use their bisulfite sequencing read counts to reconstruct their underlying methylation profiles across

genome for every tissue. Then, we exploit the composition of inferred methylation patterns from the first step in the same cell-type population to capture the variation in composition of patterns, namely their epigenetic profiles. In particular, the result of running methylFlow is DNA methylation profiles $\{N_1, N_2, \cdots, N_k\}$ for normal tissues and $\{T_1, T_2, \cdots, T_k\}$ for tumor tissues. Furthermore, each tissue profile $\mathcal{P} \in \{N_1, \cdots, N_k, T_1, \cdots, T_k\}$ is a set of patterns $\{p_{\mathcal{P},1}, p_{\mathcal{P},2}, \cdots, p_{\mathcal{P},n_\mathcal{P}}\}$ with their corresponding abundances $\{\theta_{\mathcal{P},1}, \theta_{\mathcal{P},2}, \cdots, \theta_{\mathcal{P},n_\mathcal{P}}\}$ where $n_\mathcal{P}$ denotes the number of patterns in tissue profile $\mathcal{P}$.

3.2.2 Similarity metric

Assume a region $\mathcal{R}$ with start position $l_\mathcal{R}$ and end position $e_\mathcal{R}$ is given, and our proposed algorithm needs to find if it is a RDMC. Region $\mathcal{R}$ spans DNA methylation profiles $\{N_1, N_2, \cdots, N_k\}$ for normal tissues and $\{T_1, T_2, \cdots, T_k\}$ for tumor tissues. Let methylation profiles $\{N_1^\mathcal{R}, N_2^\mathcal{R}, \cdots, N_k^\mathcal{R}\}$ and $\{T_1^\mathcal{R}, T_2^\mathcal{R}, \cdots, T_k^\mathcal{R}\}$ be the methylation profiles of normal tissues and tumor tissues in region $\mathcal{R}$ respectively. Furthemore, for each tissue profile $\mathcal{P}$ and the selected region $\mathcal{R}$, let $\{p^\mathcal{R}_{\mathcal{P},1}, p^\mathcal{R}_{\mathcal{P},2}, \cdots, p^\mathcal{R}_{\mathcal{P},n_\mathcal{P}}\}$ be the set of patterns of profile $\mathcal{P}$ in region $\mathcal{R}$, and $\{\theta^\mathcal{R}_{\mathcal{P},1}, \theta^\mathcal{R}_{\mathcal{P},2}, \cdots, \theta^\mathcal{R}_{\mathcal{P},n_\mathcal{P}}\}$ be their corresponding abundances. Note that for all $1 \leq i \leq n_\mathcal{P}$ we know $\theta^\mathcal{R}_{\mathcal{P},i} = \theta_{\mathcal{P},i}$. Moreover, for the sake of simplicity we might use $p_{\mathcal{P},i}$ instead of $p^\mathcal{R}_{\mathcal{P},i}$.

In order to find RDMC, we measure the pairwise similarity between any of two profile. For the selected region $\mathcal{R}$ and tissue profiles $\mathcal{P}, \mathcal{P}' \in$

$\{N_1^{\mathcal{R}}, N_2^{\mathcal{R}}, \cdots, N_k^{\mathcal{R}}, T_1^{\mathcal{R}}, T_2^{\mathcal{R}}, \cdots, T_k^{\mathcal{R}}\}$, we build a wighted bipartite graph $G^{\mathcal{R}}_{\mathcal{P},\mathcal{P}'} = \{\{S^{\mathcal{R}}_{\mathcal{P},\mathcal{P}'}, R^{\mathcal{R}}_{\mathcal{P},\mathcal{P}'}\}, E^{\mathcal{R}}_{\mathcal{P},\mathcal{P}'}\}$ where each node in $S^{\mathcal{R}}_{\mathcal{P},\mathcal{P}'}$ represents a methylation pattern that belongs to methylation profile $\mathcal{P}$ and each node in $R^{\mathcal{R}}_{\mathcal{P},\mathcal{P}'}$ represent a methylation pattern that belongs to profile $\mathcal{P}'$. For the sake of simplicity we might use $G_{\mathcal{P},\mathcal{P}'} = \{\{S_{\mathcal{P},\mathcal{P}'}, R_{\mathcal{P},\mathcal{P}'}\}.E_{\mathcal{P},\mathcal{P}'}\}$ instead of $G^{\mathcal{R}}_{\mathcal{P},\mathcal{P}'} = \{\{S^{\mathcal{R}}_{\mathcal{P},\mathcal{P}'}, R^{\mathcal{R}}_{\mathcal{P},\mathcal{P}'}\}, E^{\mathcal{R}}_{\mathcal{P},\mathcal{P}'}\}$ in the rest of this chapter.

The edge from node $i \in S_{\mathcal{P},\mathcal{P}'}$ with methylation patterns $p_{\mathcal{P},i}$ and the abundance of $\theta_{\mathcal{P},i}$ to node $j \in R_{\mathcal{P},\mathcal{P}'}$ with methylation patterns $p_{\mathcal{P}',j}$ and abundance of $\theta_{\mathcal{P}',j}$ has weight w_{ij} and equals to the number of overlapping CpGs with different methylation status plus the number of non-overlapping CpGs divided by the total number of CpGs spanning the region. The value of w_{ij} is equal to zero if both methylation patterns match on their overlapping region and have not any non-overlapping CpGs.

Then, to measure the similarities between the composition of cell-specific tissue profiles $\mathcal{P}$ and $\mathcal{P}'$, we solve an optimization problem that captures the variation in composition of patterns. Considering the weights of edges in our bipartite graph and the abundances of every node in $S_{\mathcal{P},\mathcal{P}'}$ and $R_{\mathcal{P},\mathcal{P}'}$, we want to match every node in $S_{\mathcal{P},\mathcal{P}'}$ to a set of nodes in $R_{\mathcal{P},\mathcal{P}'}$. In other words matching of $p_{\mathcal{P},i} \in S_{\mathcal{P},\mathcal{P}'}$ to a set of patterns in $R_{\mathcal{P},\mathcal{P}'}$ represent the changes that happened to pattern $p_{\mathcal{P},i}$ from tissue profile $\mathcal{P}$ to patterns in tissue profile $\mathcal{P}'$.

In particular, we use MCF error for measuring the difference between any two tissues with methylation profile $\mathcal{P}$ and $\mathcal{P}'$ as follows:

$$
\begin{aligned}
MCF_{\mathcal{P}\mathcal{P}'} = \quad & min \sum_{ij} w_{ij}.f_{ij} \\
\text{s.t.} \sum_{j=1}^{n_{\mathcal{P}'}} f_{ij} = \quad & \theta_{\mathcal{P},i} && 1 \leq i \leq n_{\mathcal{P}} \\
\sum_{i=1}^{n_{\mathcal{P}}} f_{ij} = \quad & \theta_{\mathcal{P}',j} && 1 \leq j \leq n_{\mathcal{P}'} \\
f_{ij} \geq \quad & 0 && 1 \leq i \leq n_{\mathcal{P}} \text{ and } 1 \leq j \leq n_{\mathcal{P}'},
\end{aligned}
$$

where $f_{i,j}$ is the amount of flow from ith node in profile $\mathcal{P}$ that represent pattern $p_{\mathcal{P},i}$ with the abundance of $\theta_{\mathcal{P},i}$ to the jth node in profile $\mathcal{P}'$ that represent pattern $p_{\mathcal{P}',j}$ with the abundance of $\theta_{\mathcal{P}',j}$. The value of $f_{i,j}$ corresponds to the fraction of the abundance of pattern $p_{\mathcal{P},i}$ that is matched to pattern $p_{\mathcal{P}',j}$. The MCF error between profile $\mathcal{P}$ and $\mathcal{P}'$ is equal to the sum of the weights of every pair $w_{i,j}$, multiplied by the amount of the corresponding flow. Also the sum of the abundance of all patterns in methylation profile $\mathcal{P}$ is equal to one and that is similar to the sum of the abundances of patterns in methylation profile $\mathcal{P}'$. We use this metric as our main similarity metric throughout this chapter.

3.2.3 Significance testing methods

In order to measure the similarities between the composition of cell-specific methylation profiles for a given region $\mathcal{R}$, we define a distance-based similarity matrix $\mathcal{A}^{\mathcal{R}}$ of dimension $2 \times k$ described as follows:

$$\mathcal{A}^{\mathcal{R}} = \begin{cases} a_{i,j+k} = MCF_{N_i,T_j} & \text{for } 0 < i \leq k\ ,\ 0 < j \leq k \\ a_{i+k,j} = MCF_{T_i,N_j} & \text{for } 0 < i \leq k\ ,\ 0 < j \leq k \\ a_{i,j} = MCF_{N_i,N_j} & \text{for } 0 < i \leq k\ ,\ 0 < j \leq k \\ a_{i+k,j+k} = MCF_{T_i,T_j} & \text{for } 0 < i \leq k\ ,\ 0 < j \leq k, \end{cases} \tag{3.1}$$

Here, MCF_{N_i,T_j} is MCF error between patterns of normal tissue $N_i^{\mathcal{R}}$ and pattern of tumor tissue $T_j^{\mathcal{R}}$, MCF_{N_i,N_j} is MCF error between patterns of normal tissues $N_i^{\mathcal{R}}$ and $N_j^{\mathcal{R}}$, MCF_{T_i,T_j} is MCF error between patterns of tumor tissues $T_i^{\mathcal{R}}$ and $T_j^{\mathcal{R}}$.

Given the distance-based similarity matrix $\mathcal{A}^{\mathcal{R}}$ for the region of our interest, the significance of each RDMC is assessed via a permutation procedure using MiRKAT method. We run a t-test and use the false discovery rate (FDR) control procedure introduced by [BH95]. The results show that the accuracy of our method is improved by running t-test compared to applying MiRKAT (See Section 3.4.1 for more details).

- **MiRKAT method:** High-throughput sequencing technologies provide a great opportunity to improve population-based studies in order to find the association between population profile and many various outcomes. MiRKAT [ZCC+15] is a regression-based kernel method to test the association of the human microbiome and exposure response. It also allows multiple outputs and using different distance-based scores with the goal of choosing the best distance. It uses a variance-component score statistic to test for the association with analytical p-value calculation. We apply MiRKAT to find if there is any association between the methylation

profiles and the health status of our samples.

- **t-test method:** T-test provides an analytical framework that is used to assess whether the mean of two groups are different from each other. We also use t-test to study the association of DNA methylation profiles of region $\mathcal{R}$ in different samples and desired output, i.e., health status of samples. We apply t-test on the similarity scores in our distance-based similarity matrix $A^{\mathcal{R}}$. In particular, we run a t-test to compare the distribution of similarity scores of normal-normal methylation profiles, i.e., MCF_{N_i,N_j} for $1 \leq i, j \leq k$, and the distribution of similarity scores of normal-tumor methylation profiles, i.e., MCF_{N_i,T_j} for $1 \leq i, j \leq k$. We use FDR (false discovery rate) control procedure to control the rate of false discoveries in our results.

3.3 Evaluation frameworks

3.3.1 Synchetic data

In order to evaluate our method, we first simulate synthetic data and compare MCFDiff with other methods. Consider a DNA methylation profile $\mathcal{P} \in \{N_1, \cdots, N_k\}$ from a heterogeneous normal tissue sample, such as a colon tissue. It contains $n_{\mathcal{P}}$ number of different DNA methylation patterns such as $p_{\mathcal{P},i}$ for $i \in \{1, \cdots, n_{\mathcal{P}}\}$. For the simulation studies, we first set the methylation profile of a normal tissue to $\mathcal{P}$, i.e. the reconstructed methylation patterns of a normal sample that we get from running methylFlow on

reduced representation of bisulfite sequencing data of normal colon samples. For a given number of normal replicates r, we simulate the rest of normal profiles for $r-1$ normal samples and also r tumor samples using methylation profile $\mathcal{P}$. The methylation status of the CpGs in each pattern is randomly changed for different replicates. The probability of change in the methylation status is around 0.02 for regions for which the composition of methylation patterns between normal and tumor replicates is not different. This parameter is called *non-RDMC mutation probability*. Otherwise, the methylation status of each CpG changes with a larger probability around 0.8 if the region of interest supposed to be a RDMC. This parameter is called *RDMC mutation probability*.

However, we assume that the epigenetic heterogeneity within a cell-type population is reflected either by variations in methylation patters and initiation of new patterns, or by changes in the abundance of the methylation patterns. Here, we describe how to make RDMC by making changes in the abundance of patterns and also in the methylation status of CpGs for the region of interest. This procedure is slightly modified depending on the complexity of the corresponding region. The complexity of a region is a notion we use to measure how distributed are the composition of patterns within a region. We define a pattern as highly abundant (HA), if the abundance of the corresponding pattern is more that 15%. Then we count the number of HA patterns within the region of interest and categorize our regions into two different types:

- **Low HA**: regions with at most one HA pattern. Algorithm for constructing a RDMC for "Low HA" regions is described in Algorithm 1.

- **High HA**: regions with more than one HA pattern. Algorithm for constructing a RDMC for "High HA" regions is described in Algorithm 2.

Once the DNA methylation profiles are simulated (original patterns for two categories of samples containing the regions with differential methylation compositions), we simulate the sequencing process. The output of this step is a set of short reads with their alignment information. The length of short reads and the sequencing error are the parameters that vary between different sequencing simulation scenarios. We set the error equals to 1 and the length of short reads is equal to 50 in our simulation studies. Short reads from this step are used later as the inputs for the evaluation process.

3.3.2 Experimental data

We applied our proposed algorithm on an experimental data from a targeted bisulfite sequencing experiment (RRBS study) [HTB+11]. The experiment is on three colon tumor tissues and their matched normal tissues. The length of the reads are 73bp or 80bp long. In their RRBS study [HTB+11], they introduced the concept of cancer-specific differentially DNA-methylated regions (cDMRs). They showed large hypomethylated genomic blocks and small DMRs. Authors further confirmed that these regions distinguish normal tissue types from each other and also distinguish between normal and tumor tissues. They also proposed that these cDMRs might be common across various cancer types.

Algorithm 1 Procedure for constructing a RDMC for **Low HA** regions.

Input: tissue profile $\mathcal{P}$ with methylation patterns $\{p_{\mathcal{P},1}, \cdots, p_{\mathcal{P},n_{\mathcal{P}}}\}$ and their abundances $\{\theta_{\mathcal{P},1}, \cdots, \theta_{\mathcal{P},n_{\mathcal{P}}}\}$.

for each HA pattern p with abundance θ in tissue profile $\mathcal{P}$ **do**
    construct a clone of p and call it p'
    let q be a random number in $[0, 1]$
    if $q \leq \frac{1}{3}$ **then**
        decrease abundance of p to $\frac{2\theta}{3}$ and set abundance of p' to $\frac{\theta}{3}$
    else if $\frac{1}{3} < q \leq \frac{2}{3}$ **then**
        decrease abundance of p to $\frac{\theta}{3}$ and set abundance of p' to $\frac{2\theta}{3}$
    else
        decrease abundance of p to 0 and set abundance of p' to θ
    end if
    change the methylation status with probability equal to the non-RDMC mutation probability for each CpG in pattern p
    change the methylation status with probability equal to RDMC mutation probability for each CpG in pattern p'
end for
keep the methylation patterns exactly the same for non-HA patterns.

Algorithm 2 Procedure for constructing a RDMC for **High HA** regions.

Input: tissue profile $\mathcal{P}$ with methylation patterns $\{p_{\mathcal{P},1}, \cdots, p_{\mathcal{P},n_{\mathcal{P}}}\}$ and their abundances $\{\theta_{\mathcal{P},1}, \cdots, \theta_{\mathcal{P},n_{\mathcal{P}}}\}$.

Randomly select half of HA patterns of tissue profile $\mathcal{P}$.

for each selected HA pattern p with abundance θ **do**

    construct a clone of p and call it p'

    let q be a random number in $[0, 1]$

    if $q \leq \frac{1}{2}$ **then**

        decrease abundance of p to $\frac{\theta}{3}$ and set abundance of p' to $\frac{2\theta}{3}$

    else

        decrease abundance of p to 0 and set abundance of p' to θ

    end if

    change the methylation status with probability equal to the non-RDMC mutation probability for each CpG in pattern p

    change the methylation status with probability equal to RDMC mutation probability for each CpG in pattern p'

end for

keep the methylation patterns exactly the same for the rest of HA patterns and also non-HA patterns

3.4 Result

We evaluated MCFDiff for finding RDMC on both synthetic and real data.

3.4.1 Synthetic data

Synthetic data was generated (as described in Section 3.3.1) for $r = 3$ number of matched normal and tumor samples. the RDMC mutation probability is set to 0.8, and non-RDMC mutation probability is set to 0.02.

3.4.1.1 Evaluating different approaches for testing the significance

We evaluated the performance of MCFDiff on the generated synthetic data. MiRKAT and t-test were used to detect regions with significance difference in normal and tumor methylation profiles. Figure 3.1 demonstrates the distribution of the performance of MCFDiff based on t-test and MiRKAT by varying different simulation parameters including number of replicates, the RDMC mutation probability and non-RDMC mutation probability.

We compared the area under the ROC curve (called AUC) of MiRKAT and t-test results for different simulation parameters. In order to compute the accuracy distribution, we did the experiment 100 times and for more than 1000 regions. Figure 3.1 illustrates that the average of AUC of ROC curve is higher based on t-test for different simulation parameters. Table 3.1 summarizes the mean of the performance of MCFDiff in terms of

sensitivity, specificity, average type I and type II error for more than 1000 regions over 100 times and various number of replicates. T-test shows higher sensitivity and specificity and lower rate of type I and type II error compared with MiRKATto detect RDMC.

Therefore, in the following analysis, we will use t-test to detect regions with significant difference in methylation profile of normal and tumor samples.

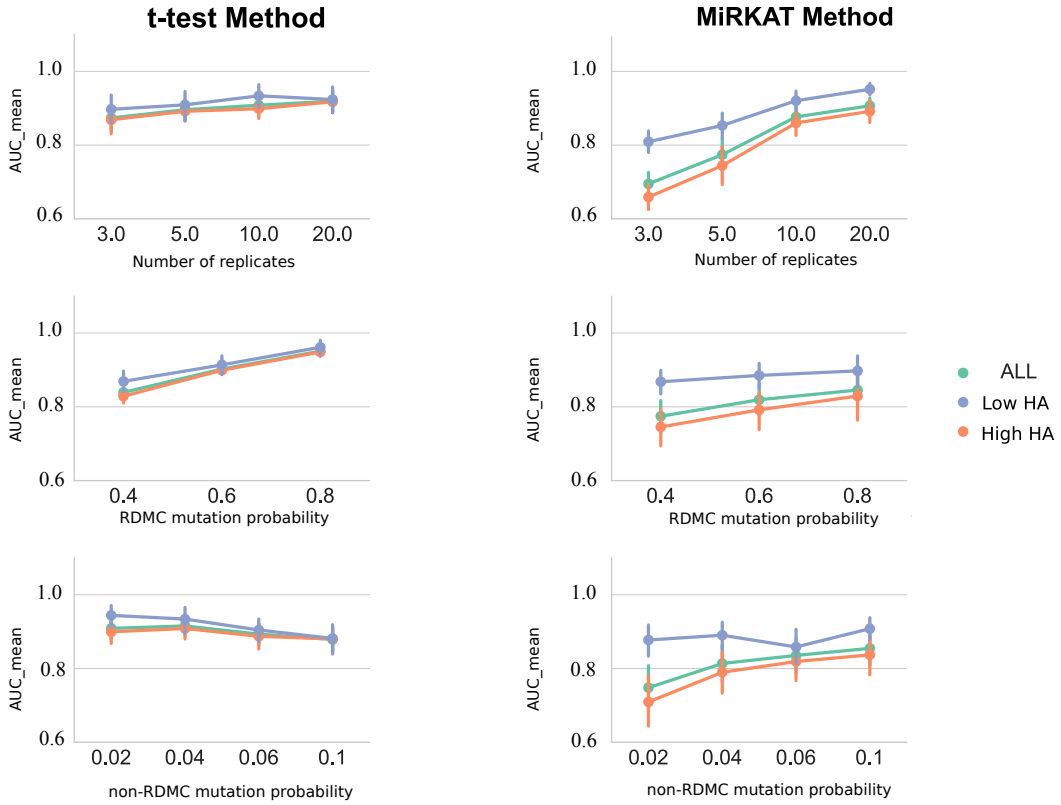

Figure 3.1: The performance in terms of AUC mean for different number of replicates, the RDMC mutation probability and non-RDMC mutation probability in terms of the area under ROC curve, using MiRKAT or t-test method.

In Figure 3.2, we evaluate MCFDiff by monitoring sensitivity, specificity and Youden parameter with various threshold in rejecting null hypothesis of t-test. Youden's

Method	Sensitivity	Specificity	type I error	type II error
MiRKAT	0.65	0.78	0.19	0.09
t-test	0.96	0.91	0.02	0.05

Table 3.1: MCFDiff performance (sensitivity, specificity, type I error, and type II error) when using t-test or MiRKAT as the significance testing method

index is equal to sensitivity + specificity − 1 and is used as a measurement for monitoring the performance of our method. Results confirm that MCFDiff improves the sensitivity significantly along with higher Youden index for the threshold of 0.01 at rejecting null hypothesis in t-test method. However, we use FDR control method to control the number of false discoveries in MCFDiff.

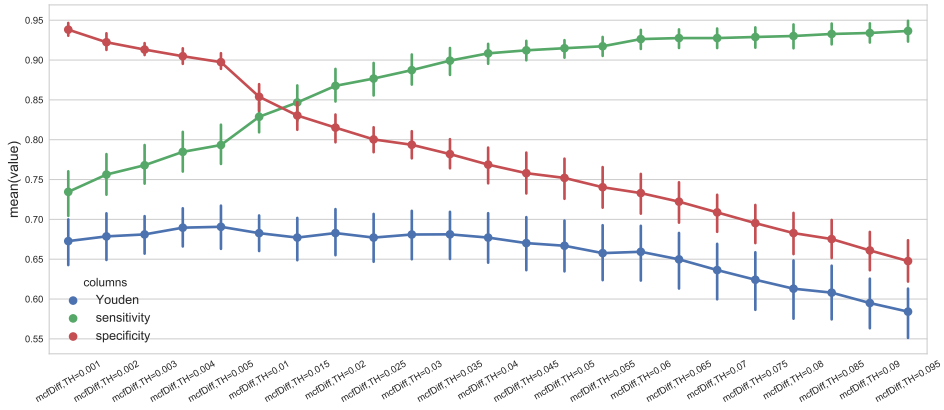

Figure 3.2: MCFDiff results base on different thresholds in rejecting null hypothesis of t-test using simulated data.

3.4.1.2 Different methods comparison

We used synthetic data as described in Section 3.3.1 to compare the performance of different methods including DMRseq [KCBI17], DSS [FCW14] and Metilene [JKB+15] with our proposed method, MCFDiff. In synthetic data, RDMC are defined such that the methylation profile composition is significantly different between normal and tumor samples. A significantly different region either (i) has different set of methylation patterns in normal and tumor samples; (ii) the abundance of patterns are different; or (iii) have different set of patterns and different abundances. Note methylation profile in non-RDMC regions are not significantly different in normal and tumor. We first explained the details of each method and then compare their performance.

- **DMRseq:** DMRseq We ran DMRseq using the R package version 1.0.14 from Bioconductor. DMRseq [KCBI17] detects differentially methylated regions (DMRs) using bisulfite sequencing short reads as input. It provides a permutation-based approach using a generalized least squares model that considers inter-individual and inter-CpG variability to generate region-level statistics. These statistics are utilized later in order to find DMRs across genome. FDR control procedure is used to decrease the number of false positives. All parameters left as default except the minimum number of CpGs per DMRs is set to 3 and the gap between them are kept to be less than 1000 base pairs.

- **DSS:** We ran DSS using the R package version $2.28.0$ from Bioconductor. It is based on a Bayesian hierarchical model to model the mean of marginal methylation percentage per loci. A Wald test is developed to test the similarity of distributions between samples and to estimate the p-values per loci. Then, DSS detects DMRs according to a set of criteria that includes (i) minimum number of CpGs in DMRs (default is 3); (ii) the threshold of rejecting the null hypothesis; (iii) the length of DMRs (default is 100); and (iv) the minimum number of CpGs rejecting null hypothesis in the region (considered as 80% of number of CpGs in the region). We set DSS's 'smoothing' parameter to be "False" in our study according to DSS's manual suggestion.
- **Metilene:** We ran Metilene version $0.2 - 7$ downloaded from http://www.bioinf.uni-leipzig.de/Software/metilene/Downloads/. Metilene is based on a scoring model in which they use a binary segmentation algorithm and 2D-KS test to find DMRs. Metilene can be run on both RRBS and WGBS data. The defined DMRs boundaries have a minimum number of CpGs (We set this parameter to 3) and there are less than 300 base pairs between adjacent CpGs.

In MCFDiff, we use FDR control procedure to limit the rate of false discoveries at region level. FDR at region level controls the number of regions that do not have significantly different composition of methylation profiles between normal and tumor samples and are categorized as RDMC . FDR at loci level controls the number of loci without significant difference in their marginal methylation percentage between normal and tu-

mor samples and are considered as DML. Note that FDR control at the loci level doesn't imply the FDR at region level.

In Figure 3.3, we investigate the change of sensitivity versus FDR for DMRseq and MCFDiff with different FDR thresholds; and for DSS and Metilene for with different single-loci level thresholds. It is shown that MCFDiff achieves significantly higher sensitivity in predicting RDMC than the alternative methods at similar observed FDR levels or alternative single-loci thresholds. The sequencing error rate is 1% in our synthetic data.

Note that the FDR control is not valid at a region level for DSS and Metilene. Although higher sensitivity was observed in the simulation study for DSS compared to DMRseq, individual loci thresholds can not be recognized for corresponding false discovery rate (one must select a loci threshold using default setting or by checking various thresholds).

In Figure 3.4, we compared MCFDiff and DMRseq using synthetic data with FDR control rate of 0.01 for various sequencing error rates. (Only MCFDiff and DMRseq can be compared using the FDR rate). MCFDiff improves both the sensitivity, specificity and Youden's index compared to DMRseq.

We claim that the most important advantage of MCFDiff lies in reconstructing the underlying patterns within a profile and estimate their abundances. This keeps the spatial correlation information of adjacent CpG sites, since the methylation patterns within a profile are not reflected completely using the marginal methylation percentage profile.

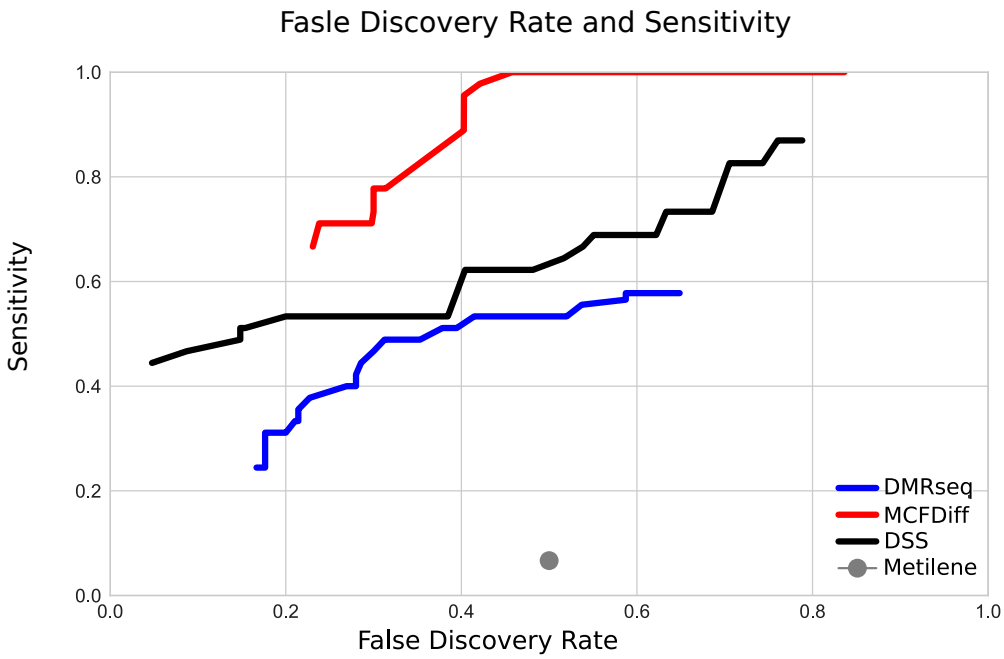

Figure 3.3: Comparing the performance of different methods on synthetic data. different region level thresholds are varied in MCFDiff and DMRseq; different loci level thresholds are varied in DSS and Metilene.

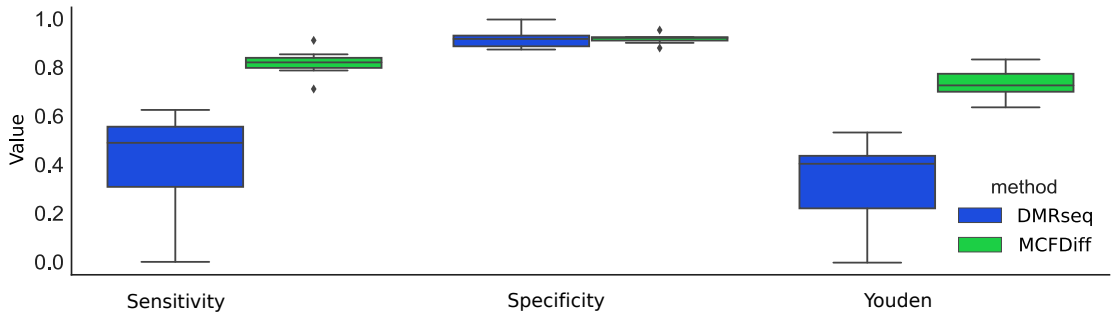

Figure 3.4: Comparing sensivity, specificity, and Youden's index of DMRseq and MCFDiff method using synthetic data with different sequencing error rate. The FDR control rate is set to 0.01. Note that Youden index equals sensitivity plus specificity minus 1.

Figures 3.5 and 3.6 compare the performance of MCFDiff and DMRseq.

Figure 3.5 compares the frequency of CpGs in regions detected by MCFDiff and the frequency of CpGs in regions detected by DMRseq having similar absolute difference of average marginal methylation percentage in normal and tumor samples. The results confirm that more CpGs with less difference in their average marginal methylation percentage reported by MCFDiff rather than DMRseq. This confirms our assumption about losing the methylation profile information by taking the average across loci.

Figure 3.6 compares the frequency of the difference between marginal methylation percentage in normal and tumor samples for regions with significant differentially methylated regions in MCFDiff and DMRseq.

The result shows that the number of regions with small difference in their marginal methylation percentage of normal and tumor samples are higher in DMRseq compared

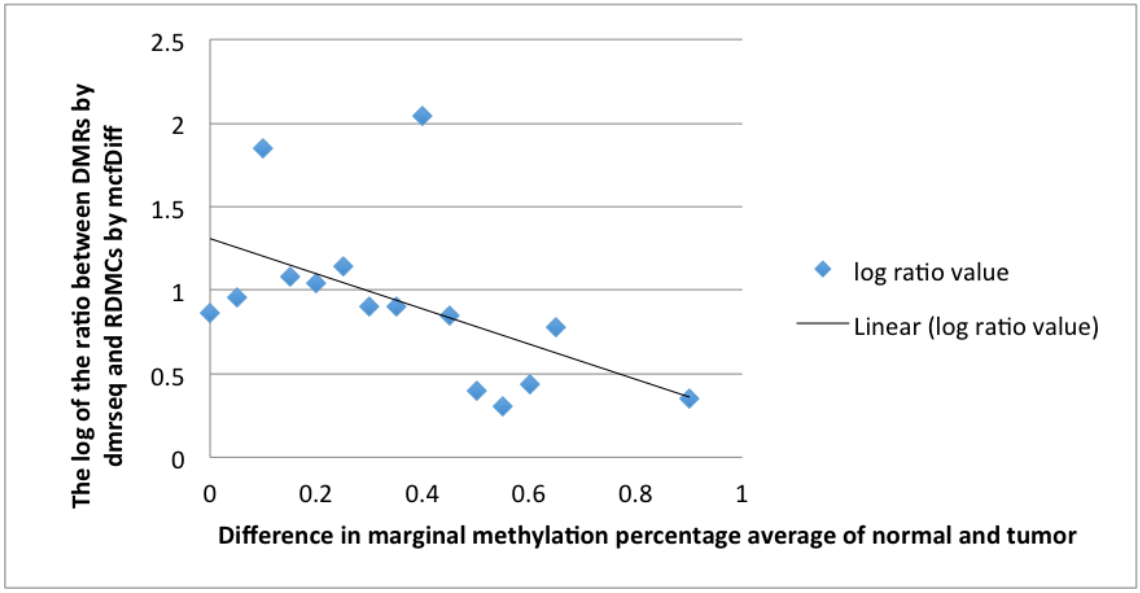

Figure 3.5: Comparing the number of CpGs reported in RDMC detected by MCFDiff to the number of CpGs in DMRs detected DMRseq with similar difference in marginal methylation percentage of normal and tumor samples. The Y axis is shown at the log ratio scale.

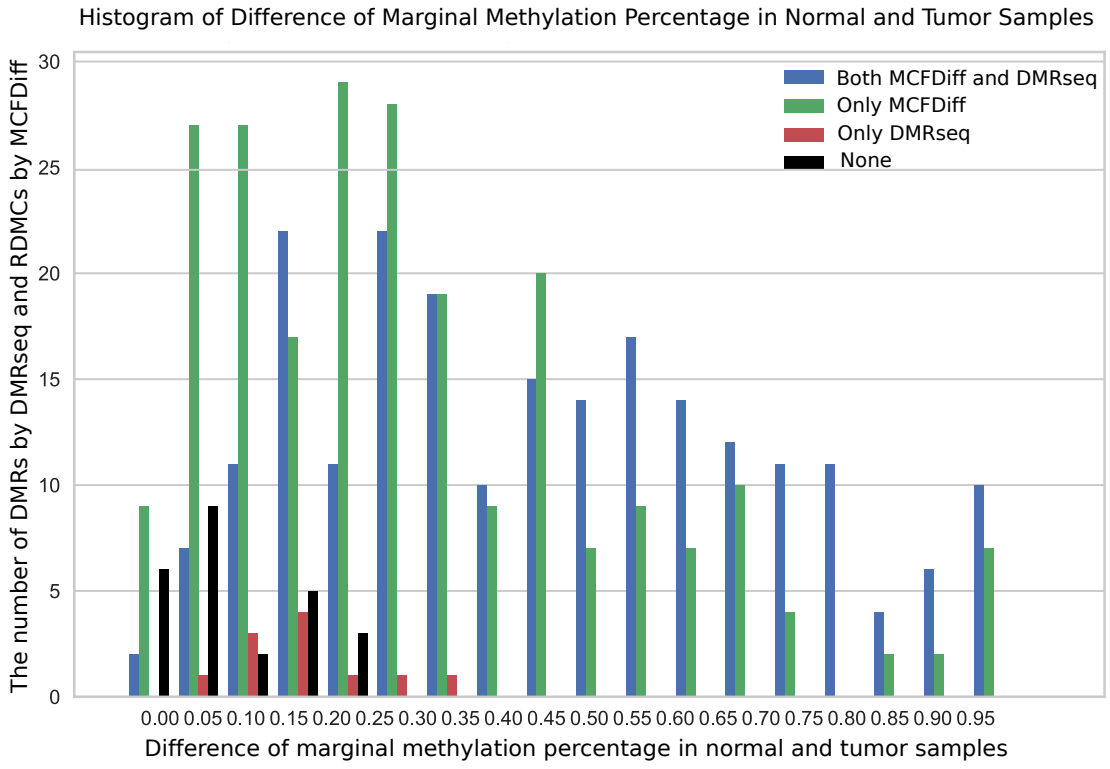

Figure 3.6: Histogram for number of regions detected by DMRseq, MCFDiff, both or none that are categorized by the average of absolute difference between marginal methylation percentage of normal and tumor samples within the region.

to MCFDiff.

3.4.2 Real data

We also evaluated our proposed method using RRBS data of colon tumor samples from [HTB+11]. In their study, they identified multiple cancer-specific differentially methylated regions.

For the experimentally identified regions, MCFDiff detects RDMC based on the reconstructed underlying pattern (given by methylFlow). Next, we compared the RDMC detected by MCFDiff with the results of DMRseq method. Figure 3.8 and 3.7 show that for cDMRs characterized by [HTB+11] in chromosome 1 and 3 MCFDiff predicts a larger subset of regions as RDMC while DMRseq predicts a smaller subset of regions as DMRs. Note that MCFDiff is sensitive to both changes in the methylation patterns of subclonal cells and changes in their abundance, while DMRseq detects variations in the marginal methylation percentage of samples.

3.5 Discussion and conclusion

In this chapter we proposed a statistical method to call the regions of differential methylation composition (RDMC). We inferred underlying DNA methylation pattern within a heterogeneous cell population using methylFlow and compared the composition of underlying patterns using a new statistical method that captures the variation in composition of methylation profiles. We considered both the underlying patterns of DNA

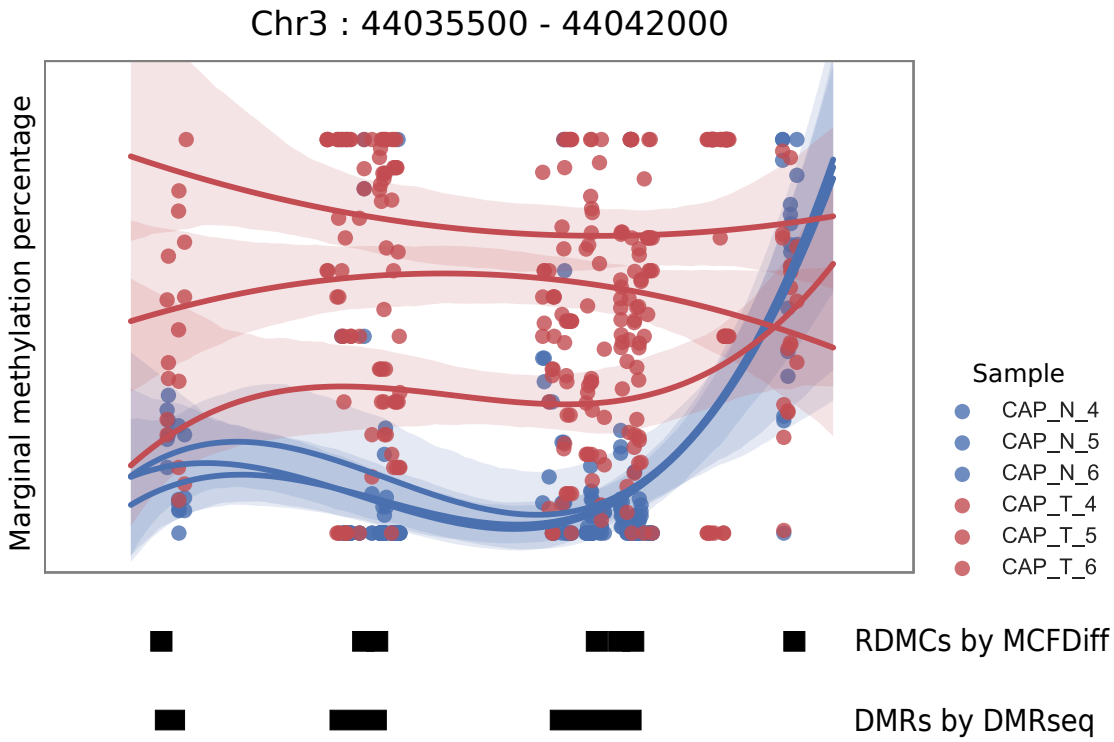

Figure 3.7: Comparison between DMRseq and MCFDiff method using real data for known RDMC on region of interest in chromosome 3. Blue and red circles represent normal and tumor data respectively.

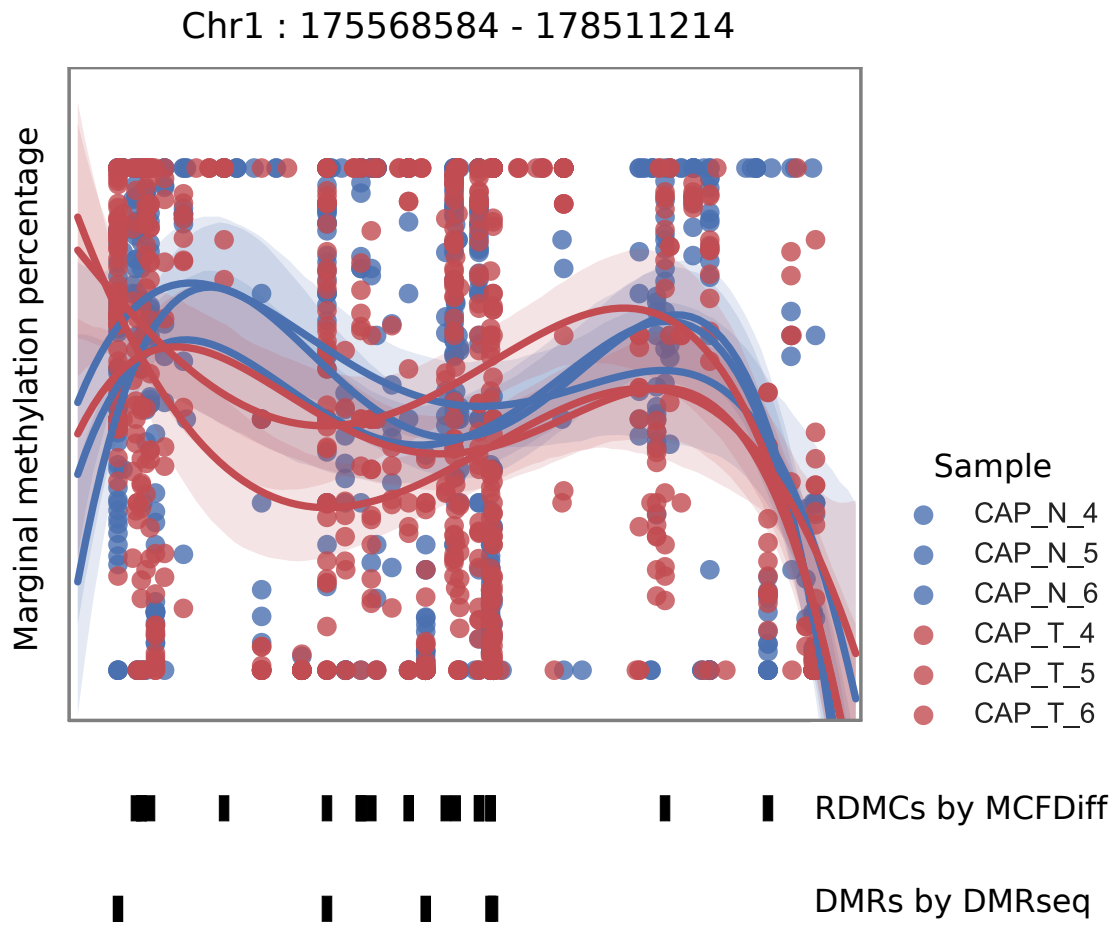

Figure 3.8: Comparison between DMRseq and MCFDiff method using real data for known RDMC on region of interest in chromosome 1. Blue and red circles represent normal and tumor data respectively.

methylation modification and the abundance of each pattern within our samples in order to detect the regions with differential methylation composition. Any significant changes in the abundance, the pattern, or both, is considered as RDMC.

We improved our prediction of RDMC by applying two different strategies. First, we employed the reconstructed underlying patterns to retain the spatial correlation of DNA methylation modifications across genome. This information is lost when marginal methylation percentage is applied. Second, regarding the underlying pattern, we inferred the impurity of the cell-type population and improved our prediction about RDMC. Both strategies improved false negative rate along with sensitivity and specificity. The results in Section 3.4.1.2 confirm that there is a higher sensitivity and specificity rate as well as a lower rate of type I and type II error using t-test and FDR control procedure for finding RDMC compared with other methods.

In real life scenario, we don't know if the methylation profiles of a region has meaningful differences between normal and tumor samples. A meaningful difference would imply the association of the methylation profiles with the desired output, i.e. health status of samples. Hence, we used simulation studies to evaluate the performance of our method. We used the synthetic data including the simulated patterns for normal and tumor replicates and including the regions that have significant difference between their methylation pattern composition. Figure 3.3 and 3.4 shows the improvement in sensitivity and specificity along with smaller rate of type I and type II error in our proposed method compared with other methods. We also evaluated our method using real data. For the experimen-

tally identified regions, MCFDiff performs better that DMRseq in predicting regions with significant change in their DNA methylation profiles.

CHAPTER 4

Conclusion and Future Directions

4.1 Summary of contributions

This dissertation contributes two computational methods to improve our understanding on cancer cells dynamics through the analysis of DNA methylation data.

Reconstructing underlying methylation patterns using bisulfite-converted sequencing data

We developed a novel computational method, called methylFlow that reconstructs the methylation profiles at single-cell level and to infer the abundance of underlying clones for different heterogeneous cell populations across genome. The method, methylFlow, provides an exceptional opportunity to (i) analyze DNA methylation data at single-cell resolution within a population by inferring the methylation patterns and their abundances over longer genomic spans, and (ii) profile the dynamics (beginning to extinction) of methylation changes across different tissues.

Finding regions with significant differences in their methylation profiles between normal and tumor samples

We developed a novel computational method, called MCFDiff that utilizes the valuable information provided by methylFlow (the methylation profile of underlying clones and their abundances in a heterogeneous population to capture tissue or disease specific variations across genome. MCFDiff increases the accuracy of detecting regions with differentially methylated profiles by (i) comparing different tissue types and penalizing any changes in the methylation profiles of underlying clones including the patterns and their abundances; and (ii) systematically considering the tumor cell impurities that might happen because of inaccurate tissue sampling. Profiling the changes between normal and tumor samples utilizing the reconstructed methylation profiles of the underling clones in different samples help us to discover *de novo epigenetic markers and to investigate known methylation sites or any specific genes of interest.*

4.2 Conclusion and future directions

My goal in this dissertation is to study the epigenetic repertoire of a heterogeneous cell population. With the advance of high-throughput sequencing techniques, we developed statistical and computational methods to reconstruct the methylation profiles of underlying clones and estimate their abundances within a cell population. We detected regions with significant differences in their methylation profiles when comparing normal and tumor samples. The proposed methods can be considered as a computational com-

plement to second-generation sequencing techniques to gain the information about the DNA methylation profiles at single-cell single-base resolution. The methylation profiles of single-cells within a cell-type population, without the limitation of single-cell studies, provide a great opportunity to compare and analyze the differences between different tissue types.

This thesis has been restricted to interpreting methylation changes in cancer. An important area of future research is to integrate methods, which will jointly profile these aberrations along with methylation changes. We expect that the joint inference of multiple classes of aberrations will provide more accurate information that can help us to stratify clonal populations and to better understand the dynamics of cancer genome.

The underlying clonal populations are related by a phylogeny. The inference of this phylogeny is an important issue that allows us to have a better insight about a tumor. It can help us to identify where an aberration is acquired or lost. It is of high interest to investigate whether the joint phylogenetic analysis from different class of aberration is beneficial to better understand the cancer genome dynamics.

Single cell sequencing is emerging as a promising tool for studying clonal populations for both genomic and methylation data. For future research (and when the technology is more matured, the cost is decreased and the quality of data is increased), a major area of research could be to develop methods that group cells within a clonal population in order to infer the underlying methylation profiles. Grouping cells to share statistical strength will reduce the noise in measurements similar to the idea of jointly inferring the

clonal population across multiple classes of aberrations.

The inference of clonal population and their phylogeny is an important step in answering fundamental questions about the underlying populations of tumor, their dynamic, their resistance to different treatments, metastatic potential and the effect of environmental stimuli. Thus, we see a valuable opportunity in collaborations between computational biologists, population geneticists, evolutionary biologists and mathematicians to tackle these problems.

Bibliography

[AKL+12] *Altuna Akalin, Matthias Kormaksson, Sheng Li, Francine E Garrett-Bakelman, Maria E Figueroa, Ari Melnick, and Christopher E Mason. methylkit: a comprehensive r package for the analysis of genome-wide dna methylation profiles.* Genome biology, *13(10):R87, 2012.* 20, 23

[AW13] *Surin Ahn and Tao Wang. A powerful statistical method for identifying differentially methylated markers in complex diseases. In* Biocomputing 2013, *pages 69–79. World Scientific, 2013.* 23

[BB15] *Lee M Butcher and Stephan Beck. Probe lasso: a novel method to rope in differentially methylated regions with 450k dna methylation data.* Methods, *72:21–28, 2015.* 22, 23, 24

[BBT+11] *Marina Bibikova, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M Le, David Delano, Lu Zhang, Gary P Schroth, Kevin L Gunderson, Jian-Bing Fan, and Richard Shen. High density DNA methylation array with single CpG site resolution.* Genomics, *98(4):288–295, October 2011.* 54

[BH95] *Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the royal statistical society. Series B (Methodological), *pages 289–300, 1995.* 22, 61

[Bir02] *Adrian Bird. Dna methylation patterns and epigenetic memory.* Genes & development, *16(1):6–21, 2002.* 2

[BJMV14] *Elsa Bernard, Laurent Jacob, Julien Mairal, and Jean-Philippe Vert. Efficient RNA isoform identification and quantification from RNA-Seq data with network flows.* Bioinformatics (Oxford, England), *30(17):2447–2455, September 2014.* 20, 35, 36

[BK11] *Andrew J Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications.* Cell research, *21(3):381, 2011.* 7

[BLB+09] *Marina Bibikova, Jennie Le, Bret Barnes, Shadi Saedinia-Melnyk, Lixin Zhou, Richard Shen, and Kevin L Gunderson. Genome-wide dna methylation profiling using infinium® assay.* Epigenomics, *1(1):177–200, 2009.* 3, 15

[BMBS13] *Rebecca A Burrell, Nicholas McGranahan, Jiri Bartek, and Charles Swanton. The causes and consequences of genetic heterogeneity in cancer evolution.* Nature, *501(7467):338, 2013.* 2

[BSM+10] *Arie B Brinkman, Femke Simmer, Kelong Ma, Anita Kaan, Jingde Zhu, and Hendrik G Stunnenberg. Whole-genome dna methylation profiling using methylcap-seq.* Methods, *52(3):232–236, 2010.* 13

[CCP10] *Pao-Yang Chen, Shawn J Cokus, and Matteo Pellegrini. Bs seeker: precise mapping for bisulfite sequencing.* BMC bioinformatics, *11(1):203, 2010.* 18

[Ced88] *Howard Cedar. Dna methylation and gene activity.* Cell, *53(1):3–4, 1988.* 1

[CFJ+11] *Pao-Yang Chen, Suhua Feng, Jong Wha Joanne Joo, Steve E Jacobsen, and Matteo Pellegrini. A comparative analysis of dna methylation across human embryonic stem cell lines.* Genome biology, *12(7):R62, 2011.* 8

[CHL+13] *Zhongxue Chen, Hanwen Huang, Jianzhong Liu, Hon Keung Tony Ng, Saralees Nadarajah, Xudong Huang, and Youping Deng. Detecting differentially methylated loci for illumina array methylation data based on human ovarian cancer data.* BMC medical genomics, *6(1):S9, 2013.* 23

[CQBM15] *Lih Feng Cheow, Stephen R Quake, William F Burkholder, and Daniel M Messerschmidt. Multiplexed locus-specific analysis of dna methylation in single cells.* Nature protocols, *10(4):619, 2015.* 17

[CSRM12] *Aniruddha Chatterjee, Peter A Stockwell, Euan J Rodger, and Ian M Morison. Comparison of alignment software for genome-wide bisulphite sequence data.* Nucleic acids research, *40(10):e79–e79, 2012.* 18

[DAB09] *Cathérine Dupont, D Randall Armant, and Carol A Brenner. Epigenetics: definition, mechanisms and clinical perspective. In* Seminars in reproductive medicine, *volume 27, page 351. NIH Public Access, 2009.* 1

[DJK11] *Balázs Dezső, Alpár Jüttner, and Péter Kovács. Lemon–an open source c++ graph template library.* Electronic Notes in Theoretical Computer Science, *264(5):23–45, 2011.* 37

[DMCB16] *Faezeh Dorri, Lee Mendelowitz, and Héctor Corrada Bravo. methylflow: cell-specific methylation pattern reconstruction from high-throughput bisulfite-converted dna sequencing.* Bioinformatics, *32(11):1618–1624, 2016. 30, 57*

[DPG+12] *Dinh Diep, Nongluk Plongthongkum, Athurva Gore, Ho-Lim Fung, Robert Shoemaker, and Kun Zhang. Library-free methylation sequencing with bisulfite padlock probes.* Nature methods, *9(3):270–272, 2012. 18*

[DS14] *Egor Dolzhenko and Andrew D Smith. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments.* BMC bioinformatics, *15(1):215, 2014. 21, 23*

[ELC+06] *Florian Eckhardt, Joern Lewin, Rene Cortese, Vardhman K Rakyan, John Attwood, Matthias Burger, John Burton, Tony V Cox, Rob Davies, Thomas A Down, et al. Dna methylation profiling of human chromosomes 6, 20 and 22.* Nature genetics, *38(12):1378, 2006. 23*

[EPM+08] *Nicholas Eriksson, Lior Pachter, Yumi Mitsuya, Soo-Yon Rhee, Chunlin Wang, Baback Gharizadeh, Mostafa Ronaghi, Robert W Shafer, and Niko Beerenwinkel. Viral Population Estimation Using Pyrosequencing.* PLOS Computational Biology, *4(5):e1000074, May 2008. 31, 33, 35*

[Est07] *Manel Esteller. Cancer epigenomics: Dna methylomes and histone-modification maps.* Nature Reviews Genetics, *8(4):286–298, 2007. 6*

[Est08] *Manel Esteller. Epigenetics in cancer.* New England Journal of Medicine, *358(11):1148–1159, 2008. 6*

[FCW14] *Hao Feng, Karen N Conneely, and Hao Wu. A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data.* Nucleic acids research, *42(8):e69–e69, 2014. 20, 23, 70*

[FSN+15] *Matthias Farlik, Nathan C Sheffield, Angelo Nuzzo, Paul Datlinger, Andreas Schönegger, Johanna Klughammer, and Christoph Bock. Single-cell dna methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics.* Cell reports, *10(8):1386–1397, 2015. 16*

[Gev15] *Olivier Gevaert. Methylmix: an r package for identifying dna methylation-driven genes.* Bioinformatics, *31(11):1839–1841, 2015. 53*

[GSB+11] *Hongcang Gu, Zachary D Smith, Christoph Bock, Patrick Boyle, Andreas Gnirke, and Alexander Meissner. Preparation of reduced representation bisulfite sequencing libraries for genome-scale dna methylation profiling.* Nature protocols, *6(4):468–481, 2011. 15*

[GZG+15] *Hongshan Guo, Ping Zhu, Fan Guo, Xianlong Li, Xinglong Wu, Xiaoying Fan, Lu Wen, and Fuchou Tang. Profiling dna methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing.* Nature protocols, *10(5):645, 2015.* 16

[GZW+13] *Hongshan Guo, Ping Zhu, Xinglong Wu, Xianlong Li, Lu Wen, and Fuchou Tang. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing.* Genome research, *2013.* 16

[HAK+12] *Eugene A Houseman, William P Accomando, Devin C Koestler, Brock C Christensen, Carmen J Marsit, Heather H Nelson, John K Wiencke, and Karl T Kelsey. Dna methylation arrays as surrogate measures of cell mixture distribution.* BMC bioinformatics, *13(1):86, 2012.* 53

[HCH13] *Hanwen Huang, Zhongxue Chen, and Xudong Huang. Age-adjusted nonparametric detection of differential dna methylation with case-control designs.* BMC bioinformatics, *14(1):86, 2013.* 23

[HDK13] *Katja Hebestreit, Martin Dugas, and Hans-Ulrich Klein. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data.* Bioinformatics, *29(13):1647–1653, 2013.* 21

[HLI12a] *Kasper D Hansen, Benjamin Langmead, and Rafael A Irizarry. Bsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions.* Genome biology, *13(10):R83, 2012.* 18, 23, 24

[HLI12b] *Kasper D Hansen, Benjamin Langmead, and Rafael A Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions.* Genome Biology, *13(10):R83, October 2012.* 21, 27, 37

[HP75] *R Holliday and J E Pugh. DNA modification mechanisms and gene activity during development.* Science (New York, NY), *187(4173):226–232, January 1975.* 11, 33

[HPL+10] *Elena Y Harris, Nadia Ponts, Aleksandr Levchuk, Karine Le Roch, and Stefano Lonardi. Brat: bisulfite-treated reads analysis tool.* Bioinformatics, *26(4):572–573, 2010.* 18

[HSR15] *Nina Hesse, Christopher Schröder, and Sven Rahmann. An optimization approach to detect differentially methylated regions from whole genome bisulfite sequencing data. Technical report, PeerJ PrePrints, 2015.* 24

[HTB+11] *Kasper Daniel Hansen, Winston Timp, Héctor Corrada Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G McDonald, Bo Wen, Hao Wu,*

Yun Liu, Dinh Diep, Eirikur Briem, Kun Zhang, Rafael A Irizarry, and Andrew P Feinberg. Increased methylation variation in epigenetic domains across cancer types. Nature Genetics*, 43(8):768–775, August 2011.* vii*,* 14*,* 28*,* 29*,* 30*,* 49*,* 64*,* 77

[HWZ+17] Lin Han, Hua-Jun Wu, Haiying Zhu, Kun-Yong Kim, Sadie L Marjani, Markus Riester, Ghia Euskirchen, Xiaoyuan Zi, Jennifer Yang, Jasper Han, et al. Bisulfite-independent analysis of cpg island methylation enables genome-scale stratification of single cells. Nucleic acids research*, 45(10):e77–e77, 2017.* 16

[ILAC+08a] Rafael A Irizarry, Christine Ladd-Acosta, Benilton Carvalho, Hao Wu, Sheri A Brandenburg, Jeffrey A Jeddeloh, Bo Wen, and Andrew P Feinberg. Comprehensive high-throughput arrays for relative methylation (charm). Genome research*, 18(5):780–790, 2008.* 12*,* 23

[ILAC+08b] Rafael A Irizarry, Christine Ladd-Acosta, Benilton Carvalho, Hao Wu, Sheri A Brandenburg, Jeffrey A Jeddeloh, Bo Wen, and Andrew P Feinberg. Comprehensive high-throughput arrays for relative methylation (CHARM). Genome Research*, 18(5):780–790, May 2008.* 54

[JBE08] Filipe V Jacinto, Esteban Ballestar, and Manel Esteller. Methyl-dna immunoprecipitation (medip): hunting down the dna methylome. Biotechniques*, 44(1):35–43, 2008.* 3

[JKB+15] Frank Jühling, Helene Kretzmer, Stephan H Bernhart, Christian Otto, Peter F Stadler, and Steve Hoffmann. metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. Genome research*, 2015.* 21*,* 24*,* 70

[JML+12a] Andrew E Jaffe, Peter Murakami, Hwajin Lee, Jeffrey T Leek, M Daniele Fallin, Andrew P Feinberg, and Rafael A Irizarry. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. International journal of epidemiology*, 41(1):200–209, 2012.* 24

[JML+12b] Andrew E Jaffe, Peter Murakami, Hwajin Lee, Jeffrey T Leek, M Daniele Fallin, Andrew P Feinberg, and Rafael A Irizarry. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. International Journal of Epidemiology*, 41(1):200–209, February 2012.* 50

[Jon12] Peter A Jones. Functions of dna methylation: islands, start sites, gene bodies and beyond. Nature Reviews Genetics*, 13(7):484, 2012.* 2*,* 8

[KA11] Felix Krueger and Simon R Andrews. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics (Oxford, England)*, 27(11):1571–1572, June 2011.* 18*,* 37

[KCBI17] *Keegan Korthauer, Sutirtha Chakraborty, Yuval Benjamini, and Rafael A Irizarry. Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing.* bioRxiv, *page 183210, 2017.* 22, 57, 70

[KKH+11] *Martin Kantlehner, Roland Kirchner, Petra Hartmann, Joachim W Ellwart, Marianna Alunni-Fabbroni, and Axel Schumacher. A high-throughput dna methylation analysis of a single cell.* Nucleic acids research, *39(7):e44–e44, 2011.* 17

[KKTM+13] *Kyong-Rim Kieffer-Kwon, Zhonghui Tang, Ewy Mathe, Jason Qian, Myong-Hee Sung, Guoliang Li, Wolfgang Resch, Songjoon Baek, Nathanael Pruett, Lars Grøntved, Laura Vian, Steevenson Nelson, Hossein Zare, Ofir Hakim, Deepak Reyon, Arito Yamane, Hirotaka Nakahashi, Alexander L Kovalchuk, Jizhong Zou, J Keith Joung, Vittorio Sartorelli, Chia-Lin Wei, Xiaoan Ruan, Gordon L Hager, Yijun Ruan, and Rafael Casellas. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation.* Cell, *155(7):1507–1520, December 2013.* 30, 49

[Kou07] *Tony Kouzarides. Chromatin modifications and their function.* Cell, *128(4):693–705, 2007.* 6

[KPW97] *Stefan U Kass, Dmitry Pruss, and Alan P Wolffe. How does dna methylation repress transcription?* Trends in Genetics, *13(11):444–449, 1997.* 2, 8

[KRCL+14] *Govindarajan Kunde-Ramamoorthy, Cristian Coarfa, Eleonora Laritsky, Noah J Kessler, R Alan Harris, Mingchu Xu, Rui Chen, Lanlan Shen, Aleksandar Milosavljevic, and Robert A Waterland. Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing.* Nucleic acids research, *42(6):e43–e43, 2014.* 18

[LBJ93] *En Li, Caroline Beard, and Rudolf Jaenisch. Role for dna methylation in genomic imprinting.* Nature, *366(6453):362, 1993.* 2, 8

[LCB+13] *Chanchao Lorthongpanich, Lih Feng Cheow, Sathish Balu, Stephen R Quake, Barbara B Knowles, William F Burkholder, Davor Solter, and Daniel M Messerschmidt. Single-cell dna-methylation analysis reveals epigenetic chimerism in preimplantation embryos.* Science, *341(6150):1110–1112, 2013.* 17

[LCM+12a] *Gilad Landan, Netta Mendelson Cohen, Zohar Mukamel, Amir Bar, Alina Molchadsky, Ran Brosh, Shirley Horn-Saban, Daniela Amann Zalcenstein, Naomi Goldfinger, Adi Zundelevich, et al. Epigenetic polymorphism and*

the stochastic formation of differentially methylated regions in normal and cancerous tissues. Nature genetics, *44(11):1207, 2012.* 55

[LCM+12b] *Gilad Landan, Netta Mendelson Cohen, Zohar Mukamel, Amir Bar, Alina Molchadsky, Ran Brosh, Shirley Horn-Saban, Daniela Amann Zalcenstein, Naomi Goldfinger, Adi Zundelevich, Einav Nili Gal-Yam, Varda Rotter, and Amos Tanay. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues.* Nature Genetics, *44(11):1207–1214, October 2012.* 22, 28

[LFJ11] *Wei Li, Jianxing Feng, and Tao Jiang. Isolasso: a lasso regression approach to rna-seq based transcriptome assembly.* Journal of Computational Biology, *18(11):1693–1707, 2011.* 36

[LGBA+13] *Sheng Li, Francine E Garrett-Bakelman, Altuna Akalin, Paul Zumbo, Ross Levine, Bik L To, Ian D Lewis, Anna L Brown, Richard J D'Andrea, Ari Melnick, et al. An optimized algorithm for detecting and annotating regional differential methylation. In* BMC bioinformatics, *volume 14, page S10. BioMed Central, 2013.* 23

[LJD+14] *Yew Kok Lee, Shengnan Jin, Shiwei Duan, Yen Ching Lim, Desmond PY Ng, Xueqin Michelle Lin, George SH Yeo, and Chunming Ding. Improved reduced representation bisulfite sequencing for epigenomic profiling of clinical samples.* Biological procedures online, *16(1):1, 2014.* 15

[LPD+09] *Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, Lee Edsall, Jessica Antosiewicz-Bourget, Ron Stewart, Victor Ruotti, A Harvey Millar, James A Thomson, Bing Ren, and Joseph R Ecker. Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, *462(7271):315–322, November 2009.* 14, 28

[LTL+12] *Jing-Quan Lim, Chandana Tennakoon, Guoliang Li, Eleanor Wong, Yijun Ruan, Chia-Lin Wei, and Wing-Kin Sung. Batmeth: improved mapper for bisulfite sequencing reads on dna methylation.* Genome Biol, *13(10):R82, 2012.* 18

[Lue73] *David G Luenberger.* Introduction to linear and nonlinear programming, *volume 28. Addison-Wesley Reading, MA, 1973.* 37

[Mak08] *Andrew Makhorin. Glpk (gnu linear programming kit), 2008.* 37

[MGB+05a] *Alexander Meissner, Andreas Gnirke, George W Bell, Bernard Ramsahoye, Eric S Lander, and Rudolf Jaenisch. Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis.* Nucleic acids research, *33(18):5868–5877, 2005.* 3, 15

[MGB+05b] *Alexander Meissner, Andreas Gnirke, George W Bell, Bernard Ramsahoye, Eric S Lander, and Rudolf Jaenisch. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis.* Nucleic acids research, *33(18):5868–5877, 2005.* 54

[MI14] *Fumihito Miura and Takashi Ito. Highly sensitive targeted methylome sequencing by post-bisulfite adaptor tagging.* DNA research, *22(1):13–18, 2014.* 16

[MNB+10] *Alika K Maunakea, Raman P Nagarajan, Mikhail Bilenky, Tracy J Ballinger, Cletus Dsouza, Shaun D Fouse, Brett E Johnson, Chibo Hong, Cydney Nielsen, Yongjun Zhao, et al. Conserved role of intragenic dna methylation in regulating alternative promoters.* Nature, *466(7303):253, 2010.* 12

[MRS+18] *Sai Ma, Mario Fuente Revenga, Zhixiong Sun, Chen Sun, Travis W Murphy, Hehuang Xie, Javier González-Maeso, and Chang Lu. Cell-type-specific brain methylomes profiled via ultralow-input microfluidics.* Nature Biomedical Engineering, *2(3):183, 2018.* 16

[MSS81] *T Mohandas, RS Sparkes, and LJ Shapiro. Reactivation of an inactive human x chromosome: evidence for x inactivation by dna methylation.* Science, *211(4480):393–396, 1981.* 2, 8

[MSS14] *Tom R Mayo, Gabriele Schweikert, and Guido Sanguinetti. M3d: a kernel-based test for spatially correlated changes in methylation profiles.* Bioinformatics, *31(6):809–816, 2014.* 23

[PFRS14] *Yongseok Park, Maria E Figueroa, Laura S Rozek, and Maureen A Sartor. Methylsig: a whole genome dna methylation analysis pipeline.* Bioinformatics, *page btu339, 2014.* 23

[PGK+03] *Arturas Petronis, Irving I Gottesman, Peixiang Kan, James L Kennedy, Vincenzo S Basile, Andrew D Paterson, and Violeta Popendikyte. Monozygotic twins exhibit numerous epigenetic differences: clues to twin discordance?* Schizophrenia bulletin, *29(1):169–178, 2003.* 1

[PO14] *Belinda Phipson and Alicia Oshlack. Diffvar: a new method for detecting differential variability with application to methylation in cancer and aging.* Genome biology, *15(9):465, 2014.* 20, 23

[RBL+00] *Bernard H Ramsahoye, Detlev Biniszkiewicz, Frank Lyko, Victoria Clark, Adrian P Bird, and Rudolf Jaenisch. Non-cpg methylation is prevalent in embryonic stem cells and may be mediated by dna methyltransferase 3a.* Proceedings of the National Academy of Sciences, *97(10):5237–5242, 2000.* 2, 8

[RDBB11] *Vardhman K Rakyan, Thomas A Down, David J Balding, and Stephan Beck. Epigenome-wide association studies for common human diseases.* Nature Reviews Genetics*, 12(8):529–541, 2011. 2, 18*

[RSS+16] *Peifeng Ruan, Jing Shen, Regina M Santella, Shuigeng Zhou, and Shuang Wang. Nepic: a network-assisted algorithm for epigenetic studies using mean and variance combined signals.* Nucleic acids research*, 44(16):e134–e134, 2016. 23*

[SCRM14] *Peter A Stockwell, Aniruddha Chatterjee, Euan J Rodger, and Ian M Morison. Dmap: differential methylation analysis package for rrbs and wgbs data.* Bioinformatics*, page btu126, 2014. 20, 22*

[SLA+14a] *Sébastien A Smallwood, Heather J Lee, Christof Angermueller, Felix Krueger, Heba Saadeh, Julian Peat, Simon R Andrews, Oliver Stegle, Wolf Reik, and Gavin Kelsey. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity.* Nature methods*, 11(8):817, 2014. 16*

[SLA+14b] *Sébastien A Smallwood, Heather J Lee, Christof Angermueller, Felix Krueger, Heba Saadeh, Julian Peat, Simon R Andrews, Oliver Stegle, Wolf Reik, and Gavin Kelsey. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity.* Nature methods*, 11(8):817–820, August 2014. 27, 30, 48*

[SLT09] *David Serre, Byron H Lee, and Angela H Ting. Mbd-isolated genome sequencing provides a high-throughput and comprehensive survey of dna methylation in the human genome.* Nucleic acids research*, 38(2):391–399, 2009. 14*

[SM15] *Yutaka Saito and Toutai Mituyama. Detection of differentially methylated regions from bisulfite-seq data by hidden markov models incorporating genome-wide methylation level distributions.* BMC genomics*, 16(12):S3, 2015. 24*

[SSS+15] *Yonatan Stelzer, Chikdu Shakti Shivalila, Frank Soldner, Styliani Markoulaki, and Rudolf Jaenisch. Tracing dynamic changes of dna methylation at single-cell resolution.* Cell*, 163(1):218–229, 2015. 17*

[STM14] *Yutaka Saito, Junko Tsuji, and Toutai Mituyama. Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions.* Nucleic acids research*, 42(6):e45–e45, 2014. 21, 24*

[SW12] *Hokeun Sun and Shuang Wang. Penalized logistic regression for high-dimensional dna methylation data with case-control studies.* Bioinformatics*, 28(10):1368–1375, 2012. 23*

[SW13] *Hokeun Sun and Shuang Wang. Network-based regularization for matched case-control analysis of high-dimensional dna methylation data.* Statistics in medicine, *32(12):2127–2139, 2013.* 23

[SWC+17] *Hokeun Sun, Ya Wang, Yong Chen, Yun Li, and Shuang Wang. petm: a penalized exponential tilt model for analysis of correlated high-dimensional dna methylation data.* Bioinformatics, *33(12):1765–1772, 2017.* 23

[SWZ+12] *Jing Shen, Shuang Wang, Yu-Jing Zhang, Maya Kappil, Hui-Chen Wu, Muhammad G Kibriya, Qiao Wang, Farzana Jasmine, Habib Ahsan, Po-Huang Lee, et al. Genome-wide dna methylation profiles in hepatocellular carcinoma.* Hepatology, *55(6):1799–1808, 2012.* 23

[SXR+14] *Deqiang Sun, Yuanxin Xi, Benjamin Rodriguez, Hyun Jung Park, Pan Tong, Mira Meong, Margaret A Goodell, and Wei Li. Moabs: model based analysis of bisulfite sequencing data.* Genome biology, *15(2):R38, 2014.* 21

[TBM+14] *Winston Timp, Héctor Corrada Bravo, Oliver G McDonald, Michael Goggins, Chris Umbricht, Martha Zeiger, Andrew P Feinberg, and Rafael A Irizarry. Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors.* Genome medicine, *6(8):61, 2014.* 28

[TGJ+16] *Andrew E Teschendorff, Yang Gao, Allison Jones, Matthias Ruebner, Matthias W Beckmann, David L Wachter, Peter A Fasching, and Martin Widschwendter. Dna methylation outliers in normal breast tissue identify field defects that are enriched in cancer.* Nature communications, *7:10478, 2016.* 23

[Wad42] *Conrad H Waddington. The epigenotype.* Endeavour, *1:18–20, 1942.* 1

[WCZ+16] *Yalu Wen, Fushun Chen, Qingzheng Zhang, Yan Zhuang, and Zhiguang Li. Detection of differentially methylated regions in whole genome bisulfite sequencing data using local getis-ord statistics.* Bioinformatics, *32(22):3396–3404, 2016.* 22, 24

[WLD+15] *Kangli Wang, Xianfeng Li, Shanshan Dong, Jialong Liang, Fengbiao Mao, Cheng Zeng, Honghu Wu, Jinyu Wu, Wanshi Cai, and Zhong Sheng Sun. Q-rrbs: a quantitative reduced representation bisulfite sequencing method for single-cell methylome analyses.* Epigenetics, *10(9):775–783, 2015.* 16

[WLJ+15] *Zhen Wang, Xianfeng Li, Yi Jiang, Qianzhi Shao, Qi Liu, BingYu Chen, and Dongsheng Huang. swdmr: a sliding window approach to identify differentially methylated regions based on whole genome bisulfite sequencing.* PloS one, *10(7):e0132866, 2015.* 22

[XL09] *Yuanxin Xi and Wei Li. Bsmap: whole genome bisulfite sequence mapping program.* BMC bioinformatics, *10(1):1, 2009.* 18

[ZCC+15] *Ni Zhao, Jun Chen, Ian M Carroll, Tamar Ringel-Kulka, Michael P Epstein, Hua Zhou, Jin J Zhou, Yehuda Ringel, Hongzhe Li, and Michael C Wu. Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test.* The American Journal of Human Genetics, *96(5):797–807, 2015.* 61

[ZLL+11] *Yan Zhang, Hongbo Liu, Jie Lv, Xue Xiao, Jiang Zhu, Xiaojuan Liu, Jianzhong Su, Xia Li, Qiong Wu, Fang Wang, et al. Qdmr: a quantitative method for identification of differentially methylated regions by entropy.* Nucleic acids research, *39(9):e58–e58, 2011.* 22, 23

[ZWH+14] *Ming-Tao Zhao, Jeffrey J Whyte, Garrett M Hopkins, Mark D Kirk, and Randall S Prather. Methylated dna immunoprecipitation and high-throughput sequencing (medip-seq) using low amounts of genomic dna.* Cellular Reprogramming (Formerly Cloning and Stem Cells), *16(3):175–184, 2014.* 13

[ZZW+14] *Xiaoqi Zheng, Qian Zhao, Hua-Jun Wu, Wei Li, Haiyun Wang, Clifford A Meyer, Qian Alvin Qin, Han Xu, Chongzhi Zang, Peng Jiang, et al. Methylpurify: tumor purity deconvolution and differential methylation detection from single tumor dna methylomes.* Genome biology, *15(7):419, 2014.* 53