Evaluating and Improving Performance of Bisulfite Short Reads Alignment and Identification of Differentially Methylated Sites

Hong Tran

Dissertation submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Genetics, Bioinformatics and Computational Biology
Liqing Zhang, Chair
Xiaowei Wu
Hongxiao Zhu
Hehuang Xie

December, 2017
Blacksburg, Virginia

Keywords: Sequence Analysis, Methylation, Bisulfite short next generation sequence mapping, Bayesian statistics

Evaluating and Improving Performance of Bisulfite Short Reads Alignment and Identification of Differentially Methylated Sites

Hong Tran

Abstract

Large-scale bisulfite treatment and short reads sequencing technology allows comprehensive estimation of methylation states of Cs in the genomes of different tissues, cell types, and developmental stages. Accurate characterization of DNA methylation is essential for understanding genotype phenotype association, gene and environment interaction, diseases, and cancer. The thesis work first evaluates the performance of several commonly used bisulfite short read mappers and investigates how pre-processing data might affect the performance. Aligning bisulfite short reads to a reference genome remains a challenging task. In practice, only a limited proportion of bisulfite treated DNA reads can be mapped uniquely (around 50-70%) while a significant proportion of reads (called multireads) are aligned to multiple genomic locations. The thesis outlines a strategy to improve the mapping efficiencies of the existing bisulfite short reads software by finding unique locations for multireads. Analyses of both simulated data and real hairpin bisulfite sequencing data show that our strategy can effectively assign approximately 70% of the multireads to their best locations with up to 90% accuracy, leading to a significant increase in the overall mapping efficiency.

The most common and essential downstream task in DNA methylation analysis is to detect differential methylated cytosines (DMCs). Although many statistical methods have been applied to detect DMCs, inconsistency in detecting differential methylated sites among statistical tools remains. We adapt the wavelet-based functional mixed models (WFMM) to detect DMCs. Analyses of simulated Arabidopsis data show that WFMM has higher sensitivities and specificities in detecting DMCs compared to existing methods especially when methylation differences are small. Analyses of monozygotic twin data who have different pain sensitivity also show that WFMM can find more relevant DMCs related to pain sensitivity compared to methylKit. In addition, we provide a strategy to modify the default settings in both WFMM and methylKit to be more tailored to a given methylation profile, thus improving the accuracy of detecting DMCs.

Population growth and climate change leave billions of people around the world living in water scarcity conditions. Therefore, utility of reclaimed water (treated wastewater) is pivotal for water sustainability. Recently, researchers discovered microbial regrowth problems in reclaimed water distribution systems (RWDs). The third part of the thesis involves: 1) identifying fundamental conditions that affect proliferation of antibiotic resistance genes (ARGs), 2) identifying the effect of water chemistry and water age on microbial regrowth, and 3) characterizing co-occurrence of ARGs and/or mobile genetics elements (MGEs), i.e., plasmids in simulated RWDs. Analyses of preliminary results from simulated RWDs show that biofilms, bulk water environment, temperature, and disinfectant types have significant influence on shaping antibiotic resistant bacteria (ARB) communities. In particular, biofilms create a favorable environment for ARGs to diversify but with lower total ARG populations. ARGs are the least diverse at 30^0C and the most diverse at 22^0C. Disinfectants reduce ARG populations as well as ARG diversity. Chloramines keep ARG populations and diversity at the lowest rate. Disinfectants work better in bulk water environment than in biofilms in terms of shaping resistome. Network analysis on assembly data is done to determine which ARG pairs are the most co-occurred. Bayesian network is more consistent with the co-occurrence network constructed from assembly data than the network based on Spearman's correlation network of ARG abundance profiles.

General Audience Abstract

Hong Tran

Human genome project has been lately attracting a lot of public attention. With the flood of big genomic data, understanding and extracting valuable information from the data remain challenge. The thesis work first evaluates the performance of different genome analysis tools. After that, the thesis outlies a strategy to improve the overall performance of whole-genome analysis tools, thus contributing to more accurate identification of mutations that are responsible for cancer and diseases. Population growth and climate change leave billions of people around the world living in water scarcity conditions. Therefore, utility of reclaimed water (treated wastewater) is pivotal for water sustainability. Recently, researchers discovered microbial regrowth problems in reclaimed water distribution systems which can worsen the existing problem of antibiotics resistance spread. The thesis identifies fundamental factors that help shape the microbial communities in reclaimed water systems in order to limit the spread of antibiotics resistance.

Acknowledgments

First and foremost, I would like to thank my advisor Dr. Liqing Zhang for letting me who had Mathematics and Statistics background and little coding experience at the time join her lab in the Computer Science department and introducing me to challenges in big data area. I am extremely grateful for her invaluable mentoring, wisdom, intelligence, and wit throughout my Ph.D. She has been a terrific advisor who takes care of and supports her students not only academically but also life. I also would like to thank my committee members, Dr. Xiaowei Wu, and Dr. Hongxiao Zhu for their statistical expertise and guidance and Dr. Hehuang Xie for his insight into biology.

I am also grateful for life time friendship and great collaboration I have with all my lab members Jacob Porter, Mingming, Vinaya, Gustavo, Mohammad, Tithi, and Dhoha who stay up late, work with me, and have patience to answer all my naïve coding questions as well as share their wisdom and their graduate life experience at Virginia Tech. I admire them so much for their hardwork and productivity. I am also grateful for the strong Vietnamese community at Virginia Tech and the close-knit Hollins community that make my academic life and life outside Ph.D. fun and easier.

I thank my parents and sister who love me unconditionally, always make me laugh and provide me a place to fall back to regardless. I thank them for their positive, funny, and happy genes.

Last but not least, I would like to thank my friend Natalie Owen and my partner Ross for the love and support and being part of my ups and downs Ph.D. journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

DNA methylation is the addition of a methyl group (CH_3) at the 5^{th} carbon position of the cytosine ring. Most cytosine methylation occurs in the sequence context of 5'CG3' (also called CpG dinucleotide) in mammalian DNA, but some in CpH dinucleotides (where H=C, T or A). The human genome is not methylated uniformly, and some small regions called CpG islands are usually unmethylated and GC rich. DNA methylation is responsible for regulation of gene expression, silencing of genes on the inactive X chromosome, imprinted genes, and parasitic DNAs [1]. DNA methylation is also a major contributor to the generation of disease-causing germ-line mutations and somatic mutations that cause cancer [2]. Therefore, accurate genome-wide determination of DNA methylation in different cells, tissues, and developmental stages is crucial for identification of causes for phenotype differences and diseases and cancer.

Large-scale characterization of DNA methylation has been made possible by bisulfite conversion of genomic DNA combined with next generation sequencing. After bisulfite treatment of DNAs, unmethylated Cs are converted to Ts and subsequent mapping of the short reads to a

reference genome allows inference of methylated vs. unmethylated Cs. Thus, inference on DNA methylation is highly dependable on the mapping of bisulfite-treated short reads to a reference genome. Similar to regular next generation sequencing analysis, the great challenge is to be able to map thousands of millions of reads in reasonable time and with high mapping efficiency (i.e., the percentage of reads that are mapped to a reference genome).

1.2 Evaluation of bisulfite short read aligners

Many tools have been developed to tackle this computational challenge such as MAQ [3], Bismark [4], BSMAP [5], PASH [6], RMAP [7], GSNAP [8], Novoalign [9], BFAST [10], BRAT-BW [11], Methylcoder [12], CokusAlignment [13], BS-Seeker [14], BS-Seeker2 [15], Segemehl [16], BiSS [17], BatMeth [18], and the latest one ERNE-bs5 [19]. The majority of these bisulfite sequencing mappers first conduct some sequence conversions (e.g. Cs to Ts and Gs to As) either on the reads, the reference genomes, or both, and then use existing regular aligners such as Bowtie [20], Bowtie2 [21], BLAT [22], SOAP [23], and BWA [24] to map short reads to a reference genome. Fonseca et al. [25] classified the tools according to their indexing techniques and supported features such as mismatches, splicing, indels, gapped alignment, and minimum and maximum of read lengths. Stockwell et al. [26] compared Bismark, BSMAP, and RMAPBS in terms of uniquely mapped reads percentages, multiple mapping percentages, CPU running time, and reads mapped per second. They also pointed out that trimming the data before aligning could improve mapping efficiency. However, the study did not examine how setting different parameters might impact program performance.

In this section, we present how modifying default parameters in each program might change the results (i.e., mapping efficiency and CPU time) and the sensitivity of each program to the characteristics of data. Though we examined many software packages, we mainly focused on two mappers: BSMAP and Bismark since they are representatives of two different index algorithms namely Burrows-Wheeler Transform in Bismark and hash table in BSMAP. In general, genome indexing based tools performed better than read indexing tools and read indexing does not provide any significant speed up [27], therefore, we did not include RMAP in our analysis. We also show that trimming data improves mapping efficiency. The paper is organized as follows:

first, we briefly describe the bisulfite sequence mapping problem and mapping techniques used by the tools. Then we describe the datasets used in the study and criteria used to evaluate the performance of the tools. Finally we show results on evaluating the tools using both real and simulated data.

1.2.1 Overview of the Computational Problem, Algorithms, and Tools

Computational challenges of mapping bisulfite short reads

Over the decades, bisulfite sequencing has remained the gold standard for DNA methylation analysis. After bisulfite treatment, unmethylated Cs are converted to thymines (T) whereas methylated Cs unchanged. Several factors make bisulfite short reads more complicated to map than regular reads. Firstly, up to four strands are analyzed from one genomic region. There are two scenarios after Polymerase chain reaction (PCR) amplification. In the first case, if the sequencing library is generated in a directional manner, the strand that the reads are amplified from is known a priori. However, if non-directional, the Watson and Crick strands of bisulfite treated sequences are no longer complementary to each other due to the conversion, and there are four different strands after PCR amplification: BSW (bisulfite Watson), BSWR (reverse complement of BSW), BSC (bisulfite Crick), and BSCR (reverse complement of BSC), all amplified and sequenced at roughly the same frequency [13]. The search space is, therefore, significantly increased relative to the original reference sequence [5]. Secondly, sequence complexity is reduced as all unmethylated Cs are changed into Ts. In the mammalian genome, because C methylation occurs almost exclusively at CpG dinucleotide, the majority of Cs in BSW and BSC strands will be converted to Ts. Therefore, most reads from the two strands will be C-poor. However, PCR amplification will complement all Gs with Cs in BSWR and BSCR strands, so reads from these two strands are typically G-poor and have a normal C content. As a result, we expect the overall C content of bisulfite reads to be reduced by approximately 50% after the two processes (converting Cs to Ts in bisulfite treatment and transcribing Gs to Cs in PCR amplification)[5]. Lastly, C to T mapping is asymmetric. The T in the bisulfite reads could be mapped to either C or T in the reference genome but not vice versa. This complicates the mapping process.

Algorithms and tools for bisulfite short reads mapping

For most of the existing programs, alignment process is to build auxiliary data structures called indices for the reference genome, the reads, or both. The indices are then used to find matching genomic positions for each read. There are many available methods to build the indices [28]. The two most popular techniques are hash tables and suffix/prefix tries [27] reviewed below together with some representative programs (Figure 1.1). A comprehensive comparison of detailed functionalities of the programs is shown in Table 1.1.

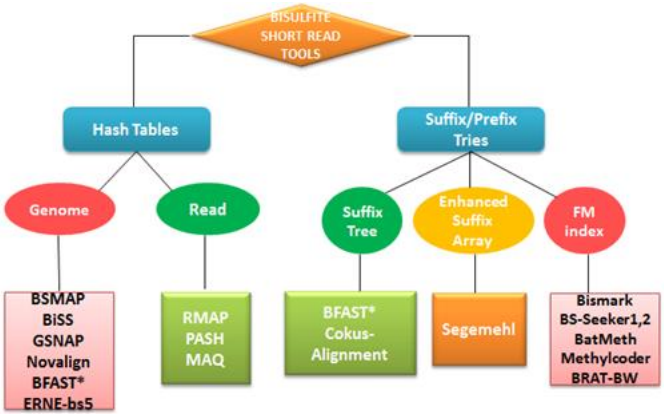


Figure 1.1: Bisulfite mapping tools classification. The tools can be divided into two groups based on indexing strategies: hash tables or Suffix/Prefix tries. Each of the groups are further classified into subgroups where some example programs are shown. Note: BFAST uses multiple index strategies: both hashing and suffix tree

Programs	Year	Algorithmic Technique used	Language	Aligner	Input	Output	Min/ Max read length	Mismatches	Indels	Gaps	Single/ Paired-end	Multi-threaded	Non-directional
ERNE-bs5	2012	Hash genome indexing, use a 5-letter (Cm, Cu) for storing methylation information, use a weighted context-aware Hamming distance to identify a T coming from an unmethylated C.	C++	None	gz/bz2/ fastq/ fasta	BAM/ SAM	up to 600 bp	1 every 15 bp(-errors arg)	Yes	Yes	both	Yes	No
Batmeth	2012	FM index, integrates mismatch Counting, list filtering and mismatch stage filtering and fast mapping onto two indexes.	Perl/C++	None	fasta	NA	NA	up to 5 (-n) in a read	No	No	Yes	Yes	Yes
BiSS	2012	Reference genome hashing, local Smith-Waterman alignment	Perl	None	fasta/ fastq/ gz/ SAM/ BAM	SAM/ BAM/ Next GenMap	up to 4096 bp	(-i from 0 to 1) in a read Default i =65%	Yes	Yes	Yes	Yes	No
Bismark	2011	FM-Index, enumerates all possible T to C conversion	Perl	Bowtie/ Bowtie2	fasta/ fastq	BAM/ SAM	Bowtie: up to 1000 bp Bowtie 2: unlimited	0 or 1 in a seed (-N)	Yes	Yes	both	Yes	Yes
BS-Seeker2	2013	FM-Index, enumerates all possible T to C conversion	Python	Bowtie2/ Bowtie/ SOAP/RMAP	fasta, fastq, qseq, pure sequence	BAM/ SAM/BS-Seeker	50-500bp	up to 4 per read (-m)	Yes	Yes	Single	No	Yes
BS-Seeker	2010	FM-Index, enumerates all possible T to C conversion, converts the genome to a 3 letter and use Bowtie to align reads	Python	Bowtie	fasta, fastq, qseq, pure sequence	BAM/ SAM/ BS_Seeker	50-250bp	up to 3 per read (-m)	Yes	No	Single	No	Yes

BSMAP	2009	hashing of reference genome and bitwise masking, tries all possible T to C combinations for reads	Python	SOAP	fasta/ fastq/SA M	SAM/ txt	up to 144 bp	up to 15 in a read (-v)		up to 3 bp	both	Yes	Yes
RMAP	2008	Wildcard matching for mapping Ts, incorporate the use of quality scores directly into the mapping process	C++		fastq/fasta	BED	unlimited	up to 10 in a read (-m)	No	No	both	No	No
BRAT-BW	2012	Convert a TA referene and CG reerence, Two FM indices are built on the positive strand of the reference genome	C++		Text file with input file names in fastq, sequence only	txt	32 bp-unlimited	unlimited	No	No	both	Yes	Yes
MAQ	2008	Builds multiple hash tables to index the reads, scans the reference genome against the hash tables to find hits	Perl/ C/C++		fastq	maq	Up to 63 bp	up to 3 per read	Yes, -n=2	No	both	No	No
PASH	2010	Implements k-mer level alignment using multi-positional hash tables	C		fastq	Txt/ SAM	NA	Yes	Yes	No	Single	No	No
Novo-align	2010	Hashing genome	C/C++		fastq	SAM/ BAM	up to 8 per read, 16 for paired end reads	Yes	Yes	up to 7bp on single end reads	Both	No	Yes
Methyl-coder	2011	FM-Index, all Cs converted to Ts	C/C++/ Python	GSNAP/ bowtie	fastq/ fasta	BAM/ SAM	Bowtie: up to 1000 bp	Yes	No	Yes	both	No	No
GSNAP	2005	q-mer hashing of reference genome	C/Perl		gzip/ fastq, fasta/ bzip2	SAM/ GSNAP	14-250bp	Yes	Yes	Yes	both	yes	No
BFAST	2009	uses multiple indexing strategies: hashing and suffix array of the reference genome	C		fastq/bz2/ gzip	SAM	NA	Yes	Yes	Yes	both	Yes	Yes

Segemehl	2008	Enhanced suffix arrays to find exact and inexact matches. Align to to read using Myers bitvector algorithm	C/C++	fasta	SAM	unlimited	Yes	(-A *[1])	Yes	both	Yes	No

Table 1.1: Detailed comparison of different bisulfite short reads mapping tools

*BFAST does not have a direct option for bisulfite mapping, users have to convert Cs to Ts in both a reference genome and reads and then align converted reads to the converted reference genome.

*Parenthesis in mismatches column indicates parameter for mismatches in a program.

*[1] A min percentages of matches per read

Indexing using hash tables can be divided into three strategies: hashing the genome, hashing the reads, or a combination of both. All hash table algorithms essentially follow the seed-and-extend technique. The algorithm keeps the positions of each k-mer fragment of the read/genome in a hash table using k-mer as the key and searches the sequence databases for k-mer matches (called seeds) [28]. After this, seeds can be joined without gaps and refined by local sequence alignment. Tools using this indexing technique include: BSMAP (genome hashing) [5], GSNAP (genome hashing) [8], Novalign (genome hashing)[9], BFAST (genome hashing/suffix array)[29], RMAP (read hashing) [7], BiSS (genome hashing) [17], PASH (read hashing) [6], MAQ (read hashing) [3], and ERNE-bs5 (genome hashing) [19].

Specifically, BSMAP is implemented based on SOAP (Short Oligonucleotide Alignment Program) [23]. BSMAP indexes the reference genome for all possible k-mers using hash tables. BSMAP masks Ts in bisulfite reads as Cs (i.e., reverse bisulfite conversion) only at C position in the original reference and keeps other Ts in the bisulfite reads unchanged. Then BSMAP maps the masked BS read directly to the reference genome. By combining bitwise masking and hash table seeding in its algorithm, BSMAP offers fast and good performance [5].

BiSS (Bisulfite Sequence Scorer) is based on Smith-Waterman local alignment with a customized alignment scoring function [17]. BiSS uses NextGenMap [30] to align bisulfite reads to a reference genome. NextGenMap involves three steps. The first step, NextGenMap indexes the reference genome in a hash table. The next step is to identify the genomic region match. NextGenMap only considers regions where the number of k-mer matches exceeds a certain threshold as a match. Unlike other methods, NextGenMap adaptively chooses the threshold, meaning each read has different threshold rather than one threshold for all reads [30].

Indexing algorithm based on suffix/prefix tries essentially converts the inexact string matching to exact matching problem. The algorithm involves two steps: identify exact matches and building inexact alignments supported by exact matches. Several representations for searching exact matches in suffix/prefix tries are suffix tree, enhanced suffix array, and FM-index [28]. Therefore, indexing using suffix/prefix tries can be classified into three subgroups: indexing using suffix tree, enhanced suffix array, and FM-index based on Burrows-Wheeler Transform. Tools falling into this category include Bismark (FM index), BS-Seeker (and BS-Seeker2, FM index),

BatMeth (FM index), Segemehl (enhanced suffix array), Methylcoder (FM index), Cokus Alignment (suffix tree), and BRAT-BW (FM index).

Specifically, in Bismark, bisulfite reads are transformed into a C to T and G to A version (equivalent to a C to T conversion on the reverse strand). Then each of them is aligned to equivalently pre-converted forms of the reference genome using four parallel instances of Bowtie or Bowtie2 [4]. Bowtie starts by building an FM index for the reference genome and uses the modified FM index [31] to find the matching location. Bowtie2 are designed to support reads longer than 50 bps. The two versions of Bowtie performed quite differently [27]. This read mapping enables Bismark to uniquely determine the strand origin of a bisulfite read.

BS-Seeker is very much similar to Bismark. The only difference is that BS-Seeker only works well for single-end reads whereas Bismark can work with both single-end and paired-end reads. Also BS-Seeker can explicitly account for tags generated by certain library construction protocols [14]. BS-Seeker records only unique alignments, defined as those that have no other hits with the same or fewer mismatches in the 3-letter alignment [14].

BRAT-BW is an evolution of BRAT [32]. Two FM indices are built on the positive strand of the reference genome: in the first, Cs are converted to Ts, and in the second, Gs are converted to As. Original reads with C to T conversion are mapped to the first index and reverse-complement reads with all Gs changed to As are mapped to the second index. BRAT-BW uses a multi-seed approach similar to Bowtie2 [32].

Datasets

We evaluated the tools on three types of data, human blood data (GSM791828), human and mouse brain data (GSE47966), and simulated mouse short read data. First, human blood data, including ten datasets (ID: SRR342552, SRR342553, SRR342554, SRR342555, SRR342556, SRR342557, SRR342558, SRR342559, SRR342560 and SRR342561) were downloaded from NCBI's short reads archive [33]. The DNA short read sequences are non-directional. Each file in SRA format contains about 23 million single-end whole genome shot gun bisulfite sequence reads from human hematopoietic stem/progenitor cells (HSPCs). The BS-Seq reads are conventional base call qualities that are Sanger/Illumina 1.9 encoded Phred values (Phred33) and trimmed to 76 bps. Second, human and mouse brain data, including ten datasets from human brain [33] and eight

datasets from mouse brain [33] were downloaded from NCBI's gene expression omnibus [34]. The DNA bisulfite short read sequences are directional. Each file contains around 100 million single-end whole genome shot gun bisulfite sequence reads from human and mouse frontal cortex in SRA format. The BS-Seq reads are conventional base call qualities that are Illumina HiSeq 2000 encoded Phred values (Phred64) and trimmed to 101 bps. Third, simulated bisulfite short reads data were generated from the mouse and human reference genome (version mm10 and hg19 respectively) using Sherman simulator [35]. Parameters such as sequencing error, bisulfite conversion rate for cytosines in CG-context, and CH-context in Sherman, are determined based on literature for the mouse data [36] and cytosine methylation reports from Bismark for the human data. Reads with different read lengths were generated to mimic the real mouse and human data. Specifically, for examining the effect of sequencing error on mapping efficiency, 24 datasets were generated from the mouse reference genome by varying the sequencing error from 0 to 4.75% (The error rate is a mean error rate per bp). Each dataset contained 1 million short reads with length of 101 bps and CG conversion rate of 10% (10% of all CG-cytosines will be converted into thymines) and CH conversion rate of 98.5% (98.5% of all CH-cytosines will be converted into thymines). For examining the effect of read length on mapping efficiency, 28 datasets were generated by varying the read length from 40 to 160 bps with sequencing error of 0.16%, CG conversion rate of 10%, CH conversion rate of 98.5% for the mouse data and with sequencing error of 0.16%, CG and CH conversion rate of 19.73% and 98.9% respectively for the human data. Both human and mouse reference genomes (hg19 and mm10) were downloaded from Ensembl [37].

Important parameters in mapping tools

Programs often have different default settings for the same parameters that can influence their performance. For example, BiSS sets the default mismatch to be 35% of the read whereas Bismark sets the equivalent parameter to zero. It is therefore important and fair to compare them on a common ground. Several important parameters that can greatly influence program performance include, (1). Number of mismatches allowed in the seed (e.g., Bismark); (2). Number of mismatches allowed in the read (e.g., BSMAP, BS-Seeker, BiSS, and BRAT-BW); (3). Directionality of data library (directional or non-directional); (4). Phred quality score (i.e., whether

data have Phred score of 33 or 64). In this study, we examined the effect of these parameters on the performance of the programs and how altering them can influence the final mapping results.

Evaluation criteria

The performance of the tools is evaluated mainly by two aspects: the mapping efficiency (i.e., percentage of uniquely mapped reads) and the CPU time. Uniquely mapped reads are reads that are mapped to only one location. Computationally speaking, most reads have multiple matches and from those matches, alignment scores are determined. An alignment is unique when it has much higher score than all other possible alignments, often determined by some statistics or cutoffs. The greater the difference between the best alignment score and the second-best alignment score, the more unique the alignment is, and the higher its mapping quality should be [38]. Mapping quality is a non-negative integer **Q = -10 log10p**, where **p** is an estimate of the probability that the alignment does not correspond to the read's true point of origin. Mapping quality is sometimes abbreviated MAPQ. (10 log10 Pr{mapping position is wrong}).

Data preprocessing

The original data were processed so reads have better quality scores and consequently can be mapped to reference genomes. Perl programming language was used to trim the tail of a read with residues quality score less than or equal to 2. After removing the tail, if the read length is shorter than 30, the read is also discarded. We use both trimmed and raw data in the analysis for the purpose of comparison of how mapping efficiency can be improved by pre-processing the data.

1.2.2 Results and Discussion

Performance comparison of the programs

Five bisulfite reads mapping tools, BSMAP, Bismark, BS-Seeker, BiSS, and BRAT-BW, were chosen to cover different algorithms discussed in the algorithm overview section (also refer to Table 1.1). BatMeth, Segmenhl, and ERNE-bs5 were not included as BatMeth failed at last step of the reads alignment, Segmenhl consumed too much computer memory (1 TB) and could not be finished in reasonable time, and ERNE-bs5 produced inaccurate results on small test datasets.

The performance is evaluated by considering two factors: mapping efficiency and CPU running time. Mapping efficiency is determined by the number of uniquely mapped reads divided by the

total number of reads. We set the number of mismatches to zero for all the programs and compare mapping efficiency and CPU running time of these programs on ten human blood datasets. Among the five programs, in terms of mapping efficiency (Figure 1.2), Bismark performs the best, achieving the highest mapping efficiency (average around 56% across the ten human blood samples), followed by BiSS (average around 46%) and BSMAP (average around 42%), and finally BRAT-BW (average around 39%) and BS-Seeker (average around 38%) with similar mapping efficiency across samples.

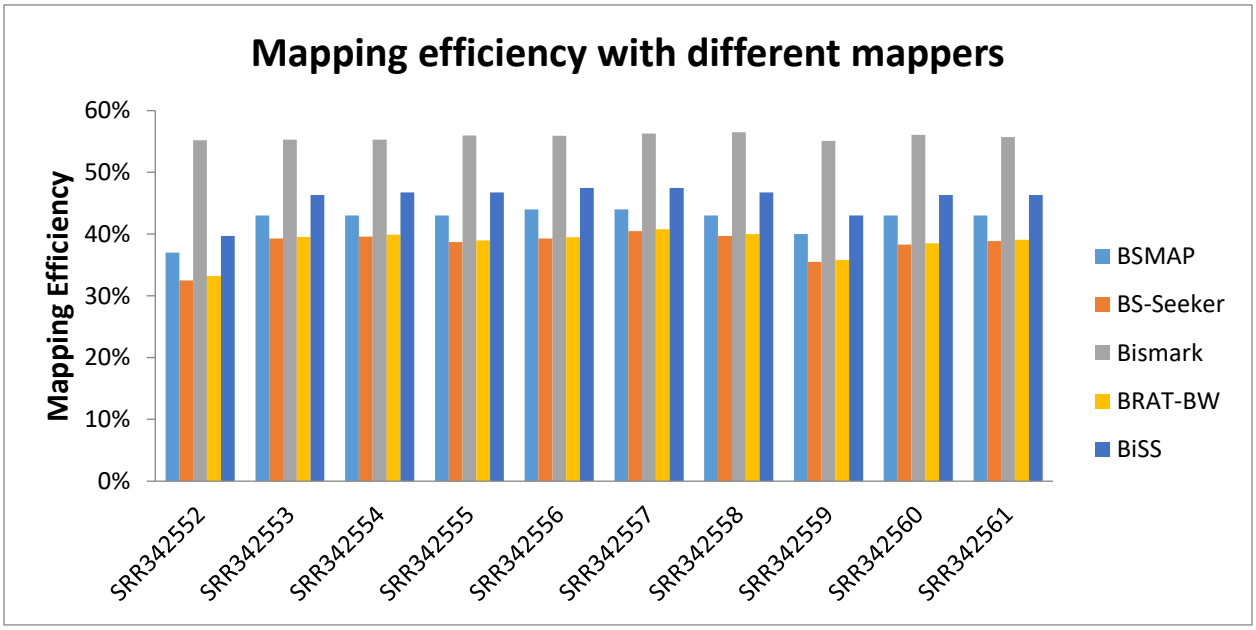


Figure 1.2: Mapping efficiency on ten human blood datasets for BSMAP, Bismark, BS-Seeker, BRAT_BW, and BiSS with zero mismatches allowed between reads and the reference genome.

However, for CPU running time, the trend is almost the opposite (Figure 1.3), with BRAT-BW taking the shortest time (average 16 minutes across samples), followed by BSMAP (average 29 minutes) and BS-Seeker (average 31 minutes). Both BiSS (average 84 hours) and Bismark (average 11 hours) took much longer time than the other three programs, suggesting existence of the tradeoff between mapping efficiency and running time. The observation that BiSS ran the slowest might be because BiSS uses Smith-Waterman local sequence alignment algorithm to align reads to potential genomic locations [17]. Interestingly, although both Bismark (written in Perl) and BS-Seeker (written in Python) use Bowtie (or Bowtie2) for short reads mapping, Bismark ran

much slower than BS-Seeker, but having much higher mapping efficiency. We then used BSMAP and Bismark to map human fetal brain and mouse brain short reads data (refer to Figure 1.5). Consistent with the results for human blood data, Bismark has higher mapping efficiency but longer CPU running time than BSMAP. The mapping percentages are very similar across samples (Figure 1.6). However, mapping efficiency for the human and mouse brain data is higher than those for human blood data, consistent with the original research studies [39], suggesting that mapping efficiency is highly dependent upon the specific experiments producing the data.

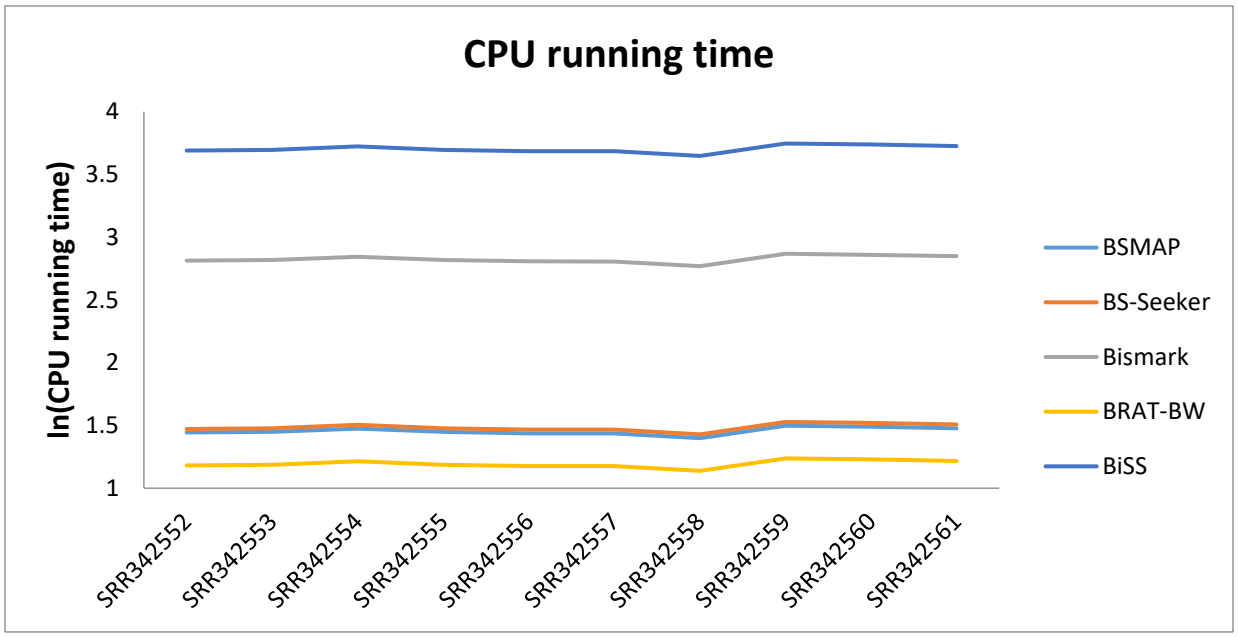


Figure 1.3: CPU running time (on a log scale) on human blood data for BSMAP, Bismark, BS-Seeker, BRAT-BW, and BiSS with zero mismatches allowed between reads and the reference genome.

Even though tools have similar mapping efficiency, reads that are actually mapped (i.e., mapped reads content) might differ among different programs. To examine how much difference the tools have in mapped reads content, we compared uniquely mapped reads from Bismark and BSMAP. On average, for human blood data, uniquely mapped reads shared by both Bismark and BSMAP account for approximately 97% of the total mapped reads by BSMAP and only 69% by Bismark. The numbers change little with different samples. Therefore, most of the mapped reads identified by BSMAP are also identified by Bismark. The difference in mapped reads content between

Bismark and BSMAP can be caused by several factors. First, the two use different string matching strategies. Bismark uses Burrows Wheeler transform and FM-indexes for searching and BSMAP hashes the reference genome for searching. In particular, Bismark uses aligner Bowtie2 whereas BSMAP uses aligner SOAP (older version of SOAP2) to map bisulfite short reads. As a result, difference in mapping algorithms can contribute to difference in mapped read content. According to Hatem et al. [27], Bowtie maintained the best throughput with higher mapping percentages, which could be why Bismark maps more reads than BSMAP. Second, determining whether a read is uniquely mapped is rather arbitrary and program specific [40]. Depending how each program defines "uniquely mapped" computationally, uniquely mapped read content can vary as a result. We also examined whether combining multiple tools to analyze bisulfite short reads could improve the overall mapping efficiency. We used BSMAP and BS-Seeker to align the unmapped reads from Bismark to see how much further BSMAP and BS-Seeker can improve the overall mapping efficiency. Table 1.2 shows that using BSMAP to align the unmapped reads from Bismark improves the overall mapping efficiency slightly better than using BS-Seeker (BSMAP: around 4% improvement; BS-Seeker: only 1%). The lesser improvement from BS-Seeker might be due to the fact that both Bismark and BS-Seeker use Bowtie to align reads although they may have different criteria in post-processing the mapped reads. Overall, results across different datasets indicate that Bismark was able to identify the most uniquely mapped reads, and addition of more programs does not significantly improve mapping efficiency.

File name	Total number of reads	Unmapped reads in BISMARK	Overall Improvement using BSMAP	Overall Improvement using BS-Seeker
SRR342552	23,472,574	10512269	3.72%	0.90%
SRR342553	23,749,583	10610307	4.24%	1.03%
SRR342554	25,232,053	11277407	4.29%	1.07%
SRR342555	23,750,428	10452979	4.23%	1.01%
SRR342556	23,140,352	10204603	4.28%	1.06%
SRR342557	23,089,492	10093756	4.33%	1.05%
SRR342558	21,205,564	9215604	4.26%	1.04%
SRR342560	26,174,056	11491673	4.17%	1.01%
SRR342561	25,457,341	11271400	4.16%	1.02%

Table 1.2: Improvement in mapping efficiency after using BSMAP and BS-Seeker to map unmapped reads from Bismark on human blood data

Effect of varying parameters in different tools

We mainly focus on how changing numbers of allowed mismatches between reads and the reference genome affects mapping efficiency. Different programs have parameters that serve this purpose but sometimes have different meanings. For example, BSMAP has the option of setting the number of mismatches allowed in each short read using the parameter ***v***. If ***v*** is between 0 and 1, it is interpreted as the mismatch rate with respect to the read length. Otherwise it is interpreted as the maximum number of mismatches allowed in a read. The default is 0.08. The maximum number of mismatches allowed is 15 per read. BiSS has the option of setting the number of mismatches allowed in each short read using the parameter ***i*** (minimum identity between a read and a match) ranging from 0 to 1. The default setting is 0.65, meaning 65% of a read and its corresponding match are identical. All reads mapped with an identity lower than this threshold will be reported as unmapped. Our results on changing these parameters show that in general, the mapping efficiency increases with the number of mismatches. The results are consistent across datasets and for all the programs tested. For brevity, only the results from BS-Seeker were used to illustrate (Figure 1.4). BS-Seeker has the option of setting the number of mismatches allowed in each short read using the parameter ***m***. The default is 2 and the maximum number allowed is 3. Figure 1.4 shows that with the number of mismatches allowed increasing from 0 to 3, mapping

efficiency increases by 43%-60%. Worth noting is that with mapping efficiency increases, CPU running time also increases significantly. Therefore, in real practice, though it is desirable to have high mapping efficiency, CPU time is another important aspect that users need to consider before running the programs. Sometimes cost of having high mapping efficiency becomes inhibitive as it takes too much running time. For example, when we changed Bismark's allowed mismatches from 0 to 1, the time it takes to finish the program doubles (e.g., increased from 657 to 1581 minutes to run on sample SRR342553). Another important aspect to consider is that increasing the number of mismatches allowed also runs the risk of increased false positives, although in real practice it is difficult to determine whether mapped reads having mismatches to the mapped location are actually false positives or real variants from the reference genome.

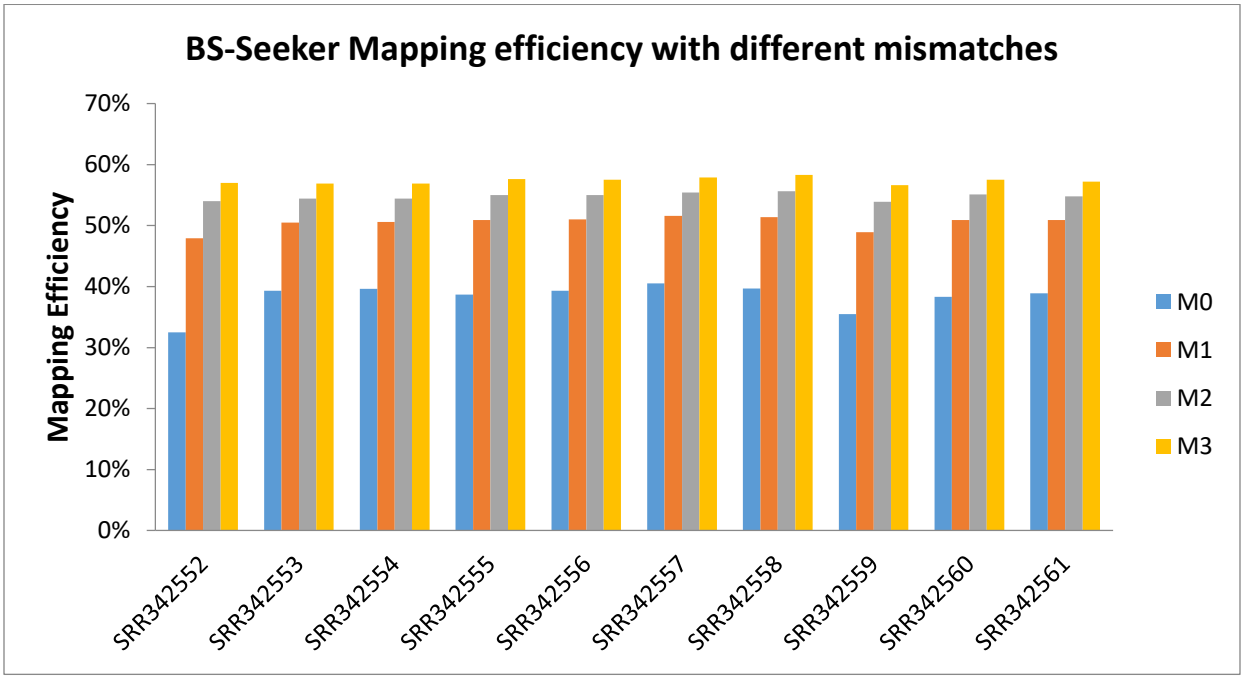


Figure 1.4: Unique mapping efficiency on ten human blood datasets from BS-Seeker with different numbers of mismatches allowed between reads and the reference genome (0, 1, 2, and 3 mismatches)

Effect of data preprocessing

We also preprocessed the reads and used those tools to analyze the trimmed data. Around 2%-4.5% of the blood data and around 1.1%-2.3% were trimmed on the brain data. Figure 1.5 shows that the mapping efficiency increases by around 5% for BSMAP and around 3% for Bismark on the human blood data, and by around 10% for BSMAP and around 6% for Bismark on the human fetal brain and mouse brain data. Therefore, preprocessing reads before mapping is an effective approach to improve mapping efficiency.

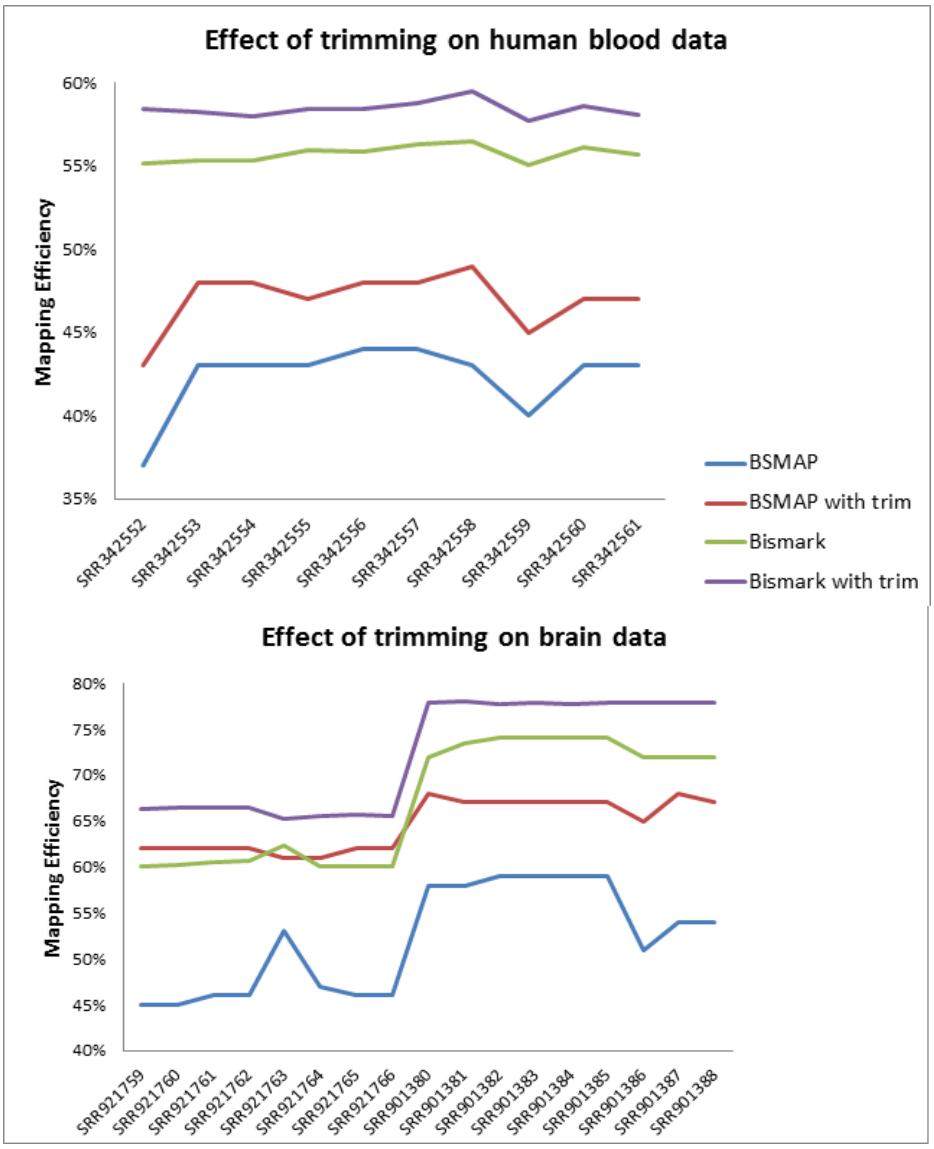


Figure 1.5: The effect of trimming reads on mapping efficiency on ten human blood, ten human brain and eight mouse brain datasets for BSMAP and Bismark

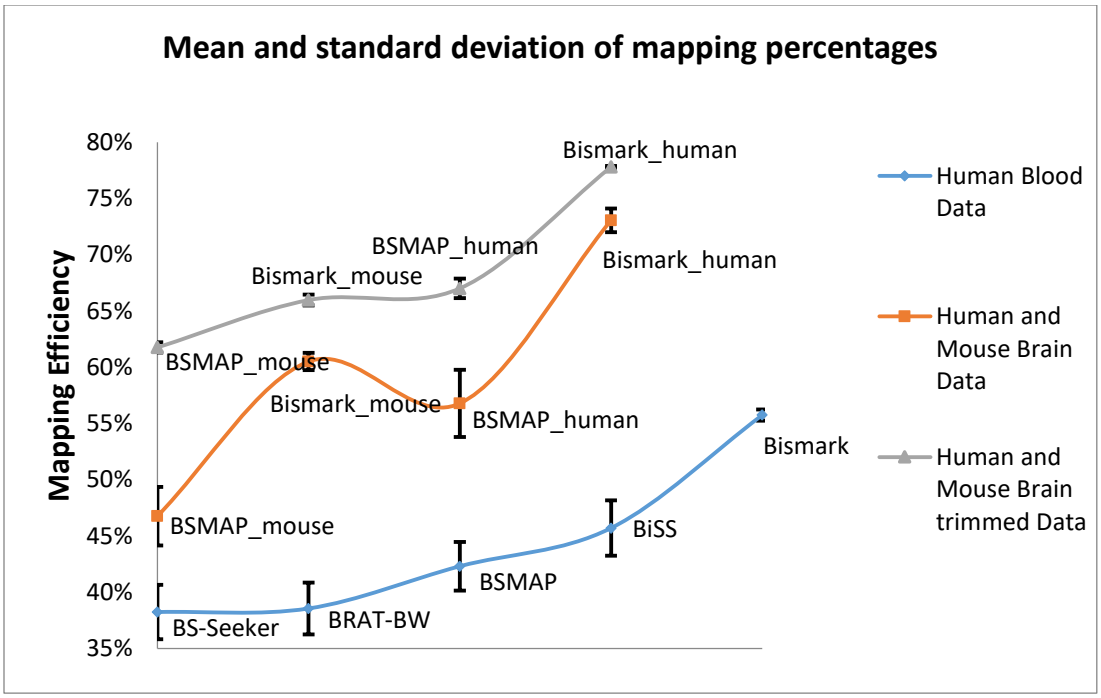


Figure 1.6: Mean and standard deviations of mapping percentages across ten human blood, ten human brain and eight mouse brain datasets

Effect of read length and sequencing error

We used simulated data to see the effect of sequencing error and read length on mapping efficiency. Sequencing error has been found to be an important factor influencing the performance of short reads mapping tools [3]. Consistent with previous finding, our result shows that for both BSMAP and Bismark, as sequencing error increases, mapping efficiency decreases (Figure 1.7). Comparatively, BSMAP is more sensitive to sequencing error than Bismark as the BSMAP's mapping efficiency decay exponentially with the increase of sequencing error, while Bismark's only gradually.

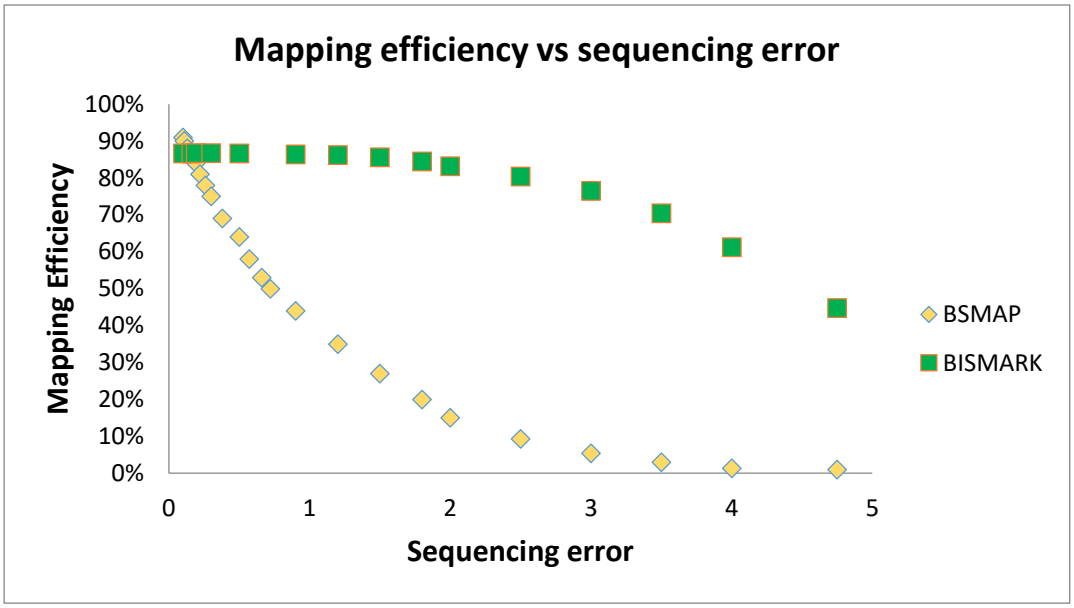


Figure 1.7: The effect of sequencing error on mapping efficiency for BSMAP and Bismark using simulated data generated from Sherman simulator with varying sequencing error from 0.1 to 4.75% (e.g., sequencing error 0.1% means 1 error in every 1000 bases) for read length =101 bp, CG=10% (10% of all CG-cytosines will be converted into thymines) and CH=98.5% (98.5% of all CH-cytosines will be converted into thymines)

Read length is another important factor in short reads mapping. Figure 1.8 shows opposite patterns for BSMAP and Bismark. For BSMAP, as read length increases from 40 to 140 bps, mapping efficiency decreases but with read length above 140 bps, an increase in read length results in an increase in mapping efficiency. On the other hand, unique mapping efficiency from BISMARK increase as read lengths increase consistently. It is unclear what contributes to the pattern exhibited by BSMAP.

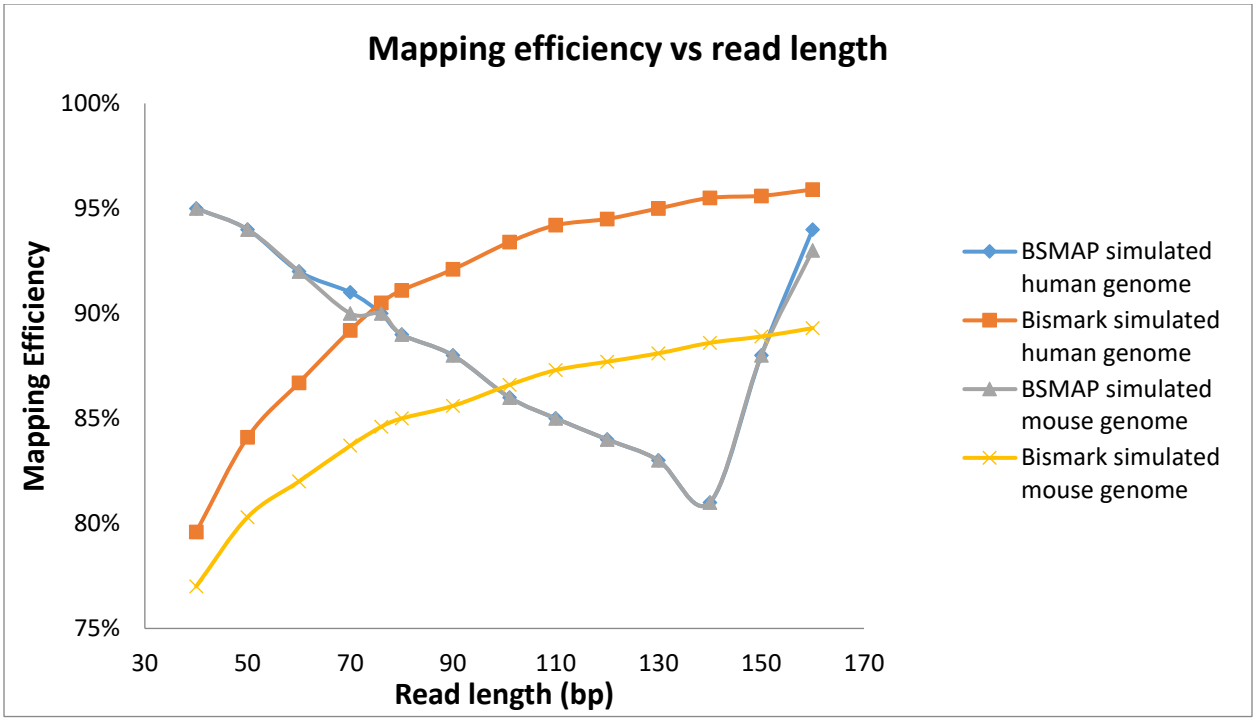


Figure 1.8: The effect of read length on mapping efficiency for BSMAP and Bismark using simulated data generated from Sherman simulator with different read lengths (from 40 to 160 bps) for sequencing error e=0.16, CG=10% and CH=98.5% for mouse and e=0.16, CG=19.73% and CH=98.9% for human data.

1.2.3 Summary

Many bisulfite short read mapping tools are available and choosing the best one among them is a difficult task. In our experiments, even though Bismark produced the highest unique mapping efficiency on real data, its CPU running time was not the shortest. BRAT-BW ran the fastest on real data but with lower mapping efficiency. Also, preprocessing data before mapping can increase mapping efficiency regardless of what tools are used. Changing parameters in the program can affect the mapping results. Overall, as number of mismatches increases, mapping efficiency increases. Short reads length and sequencing error can affect the results. Bismark is more sensitive to read lengths. The longer the read length, the higher the mapping efficiency for Bismark, whereas there is no clear pattern for BSMAP. BSMAP is more sensitive to sequencing error. A small increase in sequencing error can result in significant decrease in mapping efficiency from BSMAP.

1.3 Motivation and Problems

1.3.1 Low mapping efficiencies in bisulfite short reads

Although, numerous alignment software for traditional DNA short reads are available with much faster running time and more accuracy and quite a few DNA bisulfite short reads mappers are adapted from traditional DNA short read mappers (e.g.[4], [5], [11], [15], etc.), the percentage of BS-reads that are mapped uniquely to only one location in the reference genome remains very low (~50%) (refer to Figure 2.2). The rest of the BS sequences (i.e. multireads (BS-short reads that are aligned to multiple locations in the reference genome) and unmapped (no sequence match in the reference genome is found)) are usually removed from downstream analyses. This common practice not only leads to biased information and information loss but also enormous financial cost.

1.3.2 Inconsistency in detecting differentially methylated sites

An essential task following the alignment of bisulfite sequencing data is to detect differentially methylated cytosines among phenotype samples (i.e, disease vs control groups). Although several statistical methods have been applied to DMC detection [41], there are several problems remained. First, individual cytosines are assumed to be independent across genome. However, methylation levels of neighboring cytosines are highly correlated ([42], refer to Figure 3.1). Second, small number of samples for each phenotype coupled with weak methylation effect among different phenotype categories could make it difficult to detect DMRs accurately since most existing statistical methods assume large enough sample sizes and/or normal distribution. Thus, there is little consistency in DMRs detected by these methods.

1.3.3 Identification of fundamental factors shaping microbiome communities

Reusing treated waste water is an essential part of water sustainability. However, microbiome proliferation in RWDs is of concern. Therefore, the third problem in the thesis work involves identifying conditions that affect proliferation of opportunist pathogens and antibiotic resistance genes (ARGs) in simulated RWDs. Once, the factors contributing to ARGs and opportunist pathogens growth are found, researchers will have an insight into manipulating microbial regrowth issues in RWDs.

To address these three mentioned important issues, the thesis work outlines as follows. Chapter 2 describes our Bayesian statistical framework to improve bisulfite sequencing alignment performance. Chapter 3 contains adapting wavelet-based functional mixed models (WFMM) introduced by Morris and Carrol [43] that incorporate correlation among cytosines in estimation to better detect differential methylated sites. Chapter 4 describes how ARG communities change under various conditions, thus providing an insight into microbiome mitigation strategies. Finally, chapter 5 contains summary of new findings from each chapter and concluding remarks.

Chapter 2

BAM-ABS: A Bayesian Assignment Method for Ambiguous Bisulfite Short Reads

2.1 Introduction

DNA methylation is the addition of a methyl group (CH3) at the 5th carbon position of the cytosine ring. Cytosine methylation frequently occurs in the sequence context of 5'CG3' (also called a CpG dinucleotide) in mammalian DNA. Non-CpG methylation at CpH dinucleotides (where H=C, T or A) has been reported in some specific cell types, such as adult brain tissues [44] and stem cells [45]. DNA methylation leads to condensed chromatin and transcriptionally silences genes on the inactive X chromosome, imprinted loci, and parasitic DNAs [1]. It is also a major contributor to the generation of disease-causing germ-line mutations and somatic mutations that cause cancer

[2]. The determination of DNA methylation is crucial for the understanding of phenotype differences among cells or tissues during development and disease.

With the advance of next generation sequencing technology, characterization of genome-wide DNA methylation at single-nucleotide resolution is made possible by whole-genome bisulfite sequencing. After bisulfite treatment of DNA, unmethylated Cs are converted to Ts, whereas methylated Cs remain unchanged. Subsequent mapping of the short reads to a reference genome allows inference of methylated vs. unmethylated Cs. Several factors make bisulfite short reads (BS-reads) more complicated to map than regular short reads. First, due to how BS-reads are generated, after PCR amplification, up to four strands might be produced from one genomic region. The search space is therefore significantly increased. Second, sequence complexity is reduced, as most of the unmethylated Cs are changed into Ts. Third, C to T mapping is asymmetric. The T in the bisulfite reads could be mapped to either C or T in the reference genome but not vice versa [5]. Despite the introduction of several bisulfite short read alignment tools (e.g., Bismark [4], BSMAP [5], BS-Seeker [15], and Batmeth [18]), the mapping efficiency of BS-reads remains very low, that is, a high percentage of BS-reads, nearly 50% are either mapped to multiple genomic locations (called “multireads” or “ambiguous” reads) or unmapped [26].

Most BS-read mapping programs, for instance, Bismark [4], BS-Seeker [15], and Batmeth [18], convert both the genome and the reads to a three-letter alphabet accounting for the C-to-T or G-to-A mismatches caused by bisulfite conversion before applying a regular short read mapper such as Bowtie [21] or BWA [24]. However, due to reduced complexity in C-to-T and G-to-A conversion, this simple strategy causes a greatly increased proportion of reads to be aligned to multiple genomic locations with similar scores, i.e., multireads. The routine practice is to exclude all the multireads and unmapped reads from downstream analyses. This practice leads to not only bias in estimating methylation levels but also financial waste.

In this paper, we present a Bayesian statistical method BAM-ABS to solve the multiread mapping problem so that a great number of ambiguously mapped reads can be allocated to the most probable genomic locations, thus improving the overall mapping efficiency. To this end, we use the mismatch and methylation profiles between multireads and genomic locations, taking advantage of the information gleaned from unique read alignments, prior knowledge of single

nucleotide polymorphisms (SNPs), and context-specific methylation levels at the regions, to assign each multiread to the best location according to the highest posterior probability. Our assignment framework involves two stages. First, we use Bismark - a popular BS-reads mapper [4] to map the BS-reads, and from the mapping results, compile all the multireads with their competing locations as well as all the unique reads overlapping with the multireads. The second stage is refinement, during which we deploy the proposed Bayesian model to assign each multiread to the most likely genomic location (Figure 2.1). We use both simulated data and real data generated with hairpin bisulfite sequencing strategy to evaluate BAM-ABS' performance.

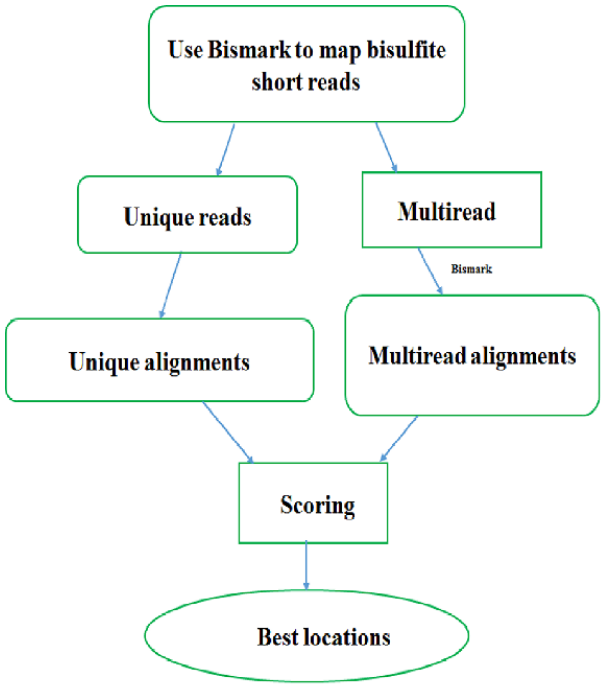


Figure 2.1: Pipeline for assigning multireads to the best locations

2.2 Materials and Methods

2.2.1 Posterior probability calculation

Suppose, for a given multiread X with length K, that there are T competing genomic locations, indexed by $t = 1, \cdots, T$, and that the multiread is mapped with similar fidelity (e.g., equal or similar number of mismatches). For genomic location t, we use M_k to denote the observed base of the

multiread X at position k $(k = 1, \cdots, K)$ of the genomic location and R_k to denote the reference base (i.e., the base that the reference genome has) at that position. The overlapping unique reads are defined as reads that are uniquely mapped with high mapping qualities (usually with MAQ scores greater than 30) and also overlapped with a multiread's mapped location. Assuming that there are r such unique reads, we use $D_k = \{d_{1k}, d_{2k}, \ldots d_{rk}\}$ to denote the observed bases of overlapping unique reads at position k. Given the multiread and genomic location t, the observed data consist of two mismatch profiles, one between the reference genome and the multiread, the other between the reference genome and all the overlapping unique reads. We want to compute the posterior probability of observing M_k given D_k, $P(M_k|D_k)$, based on which decision is made on assigning the multiread.

Applying Bayes' Theorem,

$$P(M_k|D_k) = \frac{\pi(M_k)P(D_k|M_k)}{\pi(M_k)P(D_k|M_k) + \pi(\bar{M}_k)P(D_k|\bar{M}_k)},$$

where $\pi(M_k)$ is the prior probability of observing base M_k and $P(D_k|M_k)$ is the likelihood of observing the overlapping unique reads at position k given the observed M_k. In practice, we would also like to incorporate the reference information R_k into the prior to help improve the inference accuracy. Replacing $\pi(M_k), \pi(\bar{M}_k)$ with $\pi(M_k|R_k), \pi(\bar{M}_k|R_k)$, respectively, and assuming that conditioning on M_k, D_k is independent of R_k, we may write the posterior probability as

$$P(M_k|D_k, R_k) = \frac{\pi(M_k|R_k)P(D_k|M_k)}{\pi(M_k|R_k)P(D_k|M_k) + \pi(\bar{M}_k|R_k)P(D_k|\bar{M}_k)},$$

How the prior probability $\pi(M_k|R_k)$ is computed is given in the next section.

Since the likelihood $P(D_k|M_k)$, as the product of all $P(d_{jk}|M_k)$ for j=1…r, is directly related to the number of overlapping unique reads: the more reads, the smaller likelihood, we calculate $P(D_k|M_k)$ in an average sense instead of using the usual joint probability definition to avoid this bias. Thus we write the likelihood in terms of the base quality of the multiread and unique reads as

$$P(D_k|M_k) = \frac{\sum_{j=1}^{r} P(d_{jk}|M_k)}{r}$$

where

$$P(d_{jk}|M_k) = \begin{cases} 1 - \varepsilon_{jk} - \varepsilon_k + \varepsilon_{jk} \times \varepsilon_k, & \text{if } d_{jk} = M_k \\ \varepsilon_{jk} + \varepsilon_k - \varepsilon_{jk} \times \varepsilon_k, & \text{if } d_{jk} \neq M_k \end{cases},$$

and ε_{jk} is the probability of observing a base miscall in the jth unique read at position k, ε_k is the probability of observing a base miscall in the multiread at position k. It is easy to see that the above calculation follows the general addition rule of probability, that is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Here, A represents the event of having a sequencing error in the jth unique read at position k, and B represents the event of having a sequencing error at the multiread base M_k. Given sequencing errors occur independently in unique reads and in multireads, i.e., $P(A \cap B) = P(A)P(B)$, replacing $P(A)$ with ε_{jk}, and $P(B)$ with ε_k then results in the expression of $P(d_{jk}|M_k)$.

Finally we calculate the posterior probability of observing the multiread X at genomic location t by

$$P(X|\boldsymbol{D}, R) = \prod_{k=1}^{K} P(M_k|D_k, \mathrm{R_k}),$$

where $\boldsymbol{D} = \{D_1, D_2, \ldots, D_K\}$ denotes the set of all observed bases from the overlapping unique reads at positions $1,2,\cdots,K$. The genomic location with the highest posterior probability is chosen, and an assignment score S for the read is calculated by taking the log odds of the posterior probabilities at the best location and at the next best location

$$S = \log \frac{P(X|\boldsymbol{D}) \text{ at best location}}{P(X|\boldsymbol{D}) \text{ at next best location}}. \qquad (1)$$

To assign a multiread, we need to determine a cutoff score S_0. Users can choose a cutoff score suitable to their needs. If a multiread has an assignment score $S \geq S_0$, the read is considered as "assignable" and will be assigned to the best location, otherwise, the read will be labelled as "unassignable". We conducted experiments to determine a cutoff score S_0. Experiments show that BAM-ABS achieves good performance when S is set between 0.005 to 6. We set S_0 to be 0.05 in simulated data and 0.2 in real data. In real data, the sequence coverage is not uniform across the entire genome and some genomic loci may not be covered by any uniquely mapped read. We will assign a multiread to a location that has more unique reads. To increase inference accuracy, we

raise the cut-off in real data to 0.2 and achieved a reasonable efficiency in the multiread assignment.

2.2.2 Prior probability calculation

Given the reference genome, the mutation rate of the organism, the observed multiread sequence, and knowledge on context-specific methylation levels, we can infer the underlying process and compute $\pi(M_k|R_k)$, the prior probability of observing multiread base M_k given the reference genome base R_k at position k. For example, according to NCBI dbSNP [46], transitions are twice as frequent as transversions in many species, such as humans and mice. Also, studies have shown that the methylation rate is about 0.80 at CpG whereas 0.05 at CH (H∈{A,T,C }) in mammals [47]. Such information can be incorporated to compute $\pi(M_k|R_k)$. To illustrate, suppose that the reference genome has a base C at one position of the genomic location that the multiread is aligned to, then there are four possible cases:

1) observing A in the multiread

In this case, we conclude that there is only a C to A mutation occurring and the prior probability of observing A in the multiread given C in the reference genome is

$$\pi(M_k|R_k) = P(\text{C to A mutation}).$$

2) observing C in the multiread

In this case, we conclude that no mutation occurs and the C is methylated. The prior probability of observing C in the multiread given C in the reference genome is

$$\pi(M_k|R_k) = [1 - P(\text{mutation})] \times P(\text{methylation}).$$

3) observing G in the multiread

In this case, we conclude that there is only a C to G mutation occurring and the prior probability of observing G in the multiread given C in the reference genome is

$$\pi(M_k|R_k) = P(\text{C to G mutation}).$$

4) observing T in the multiread

In this case, we conclude that either there is a C to T mutation occurring or there is no mutation and the C in the reference genome is unmethylated and converted to T after bisulfite treatment. Therefore the prior probability of observing T in the multiread given C in the reference genome is the sum of the probabilities of the two disjoint events and can be expressed as

$$\pi(M_k|R_k) = P(\text{C to T mutation}) + [1 - P(\text{mutation})] \times [1 - P(\text{methylation})].$$

The probability of C methylation $P(\text{methylation})$ depends on the sequence context, that is, if the next base in the multiread is G, the probability of C methylation is higher than that if the next base is H (H∈{A,T,C}). The probability of mutation can be computed similarly as in previous methods [48], [49]. For example, if we assume that the SNP rate in the human genome is 0.001 and that the reference allele is C at position k, the prior probabilities of C to A mutation and C to G mutation are 0.00025 and 0.00025, respectively, whereas the prior probability of C to T mutation is 0.0005 and the prior probability of C to C (i.e., no mutation) is 0.999. All other cases are illustrated in Section 2.1 of the Supplementary materials. In a later section of simulation study and real data analysis, we will also consider the "without" prior option, that is, using a uniform prior (equal probabilities for observing different bases on M_k) and make a comparison to illustrate the advantage of using a prior in BAM-ABS.

2.2.3 Bisulfite short read simulation

We aim to generate BS-reads that closely mimic the bisulfite conversion experiment. The simulated data consist of BS-reads generated from the human genome (hg19) and the mouse genome (mm10). First, we randomly assigned a mutation rate of 0.001 to every base in the reference genome, i.e., we randomly changed 0.1% of all current bases in the reference genomes to other bases. As transitions are twice as frequent as transversions, we assigned a higher probability for C↔T and G↔A mutations than other mutations, e.g., $\text{P}(\text{C} \leftrightarrow \text{T}) = 0.0005$ while $\text{P}(\text{C} \leftrightarrow \text{A}) = \text{P}(\text{C} \leftrightarrow \text{G}) = 0.00025$. Second, we randomly assigned a methylation rate to every cytosine in both strands of each chromosome after introducing mutations. We varied the methylation probability at CpG (i.e., 70%, 75%, 80%, 85%, 90%) while maintaining methylation probability at CH (H∈{A,T,C}) 0.5%. To illustrate, we randomly converted C to T at 99.5% of all CH sites and converted C to T at 30%, 25%, 20%, 15% or 10% of all CpG sites to generate different

data sets. After introducing both mutation and methylation, we randomly generated short reads with different read lengths for each data (51 bp, 76 bp, and 101 bp) from the converted reference genome. Finally, we extracted quality score strings from three real datasets SRR980327 (read length=51 bp), SRR342553 (read length=76 bp), and SRR921765 (read length=101 bp) generated by the Illumnia-HiSeq 2000 platform (data downloaded from NCBI's short read archive (http://www.ncbi.nlm.nih.gov/sra) and simulated sequence errors according to the per-base error probabilities of all reads from these datasets. All reads were generated in a directional manner, i.e., only from the top strands of the genome. We simulated 3,000, 40,000, and 100,000 short reads for each methylation probability parameter with varying read lengths.

We used Bismark [4] to align simulated BS-reads and collected all ambiguous reads or multireads. Most of the multireads have two or three mapped genomic locations in both simulated and real data (Figure 2.1S in Supplementary Materials). In this paper, we only examined directional data. However, undirectional data will be addressed similarly, since only methylation and SNP information of uniquely mapped reads from the same DNA strand as a multiread is incorporated in the scoring model.

An important and practical question is how much coverage is required for accurate assignment of multireads using BAM-ABS. To address this problem, for each location that multireads are aligned to, we generated different numbers (i.e., 3x, 5x, 10x, 25x, and 30x) of overlapping unique reads to mimic different depths of coverage. We then introduced sequencing errors for the generated reads using base quality scores from the real data. These reads are treated as overlapping unique reads by BAM-ABS. A detailed pipeline for generating BS-reads and overlapping unique reads is illustrated in supplementary Figure 2.2S.

2.2.4. Real data from hairpin bisulfite sequencing

To validate our model on real data, we used the genome-scale hairpin bisulfite sequencing data for mouse embryonic stem cell (ESC) (NCBI's SRA accession number: GSM1173118) produced in our previous study [50]. The hairpin data are from one sample but generated in five different sequencing lanes (labeled as Lane1, Lane2, Lane3, Lane4, Lane5). In brief, genomic DNA was extracted and then sonicated into fragments of around 200 bp. Then, the DNA fragments were

ligated to the biotinylated hairpin and Illumina sequencing adaptors simultaneously. Following the streptavidin-capture and bisulfite PCR, the fragments linked to both the hairpin adaptor and Illumina sequencing adaptor were amplified for high-throughput paired-end sequencing using Illumina HiSeq 2000 platform. After purification, size selection of 400–600-bp fragments was conducted with LabChip XT DNA Assay (Caliper) to yield longer sequences that are more amenable for unambiguous mapping to the reference sequence. The reads are of 101 bp in length. Unlike traditional bisulfite sequencing methods, which are non-invertible, the hairpin technology allows for recovery of the original sequences; therefore, hairpin data can be used to evaluate the mapping efficiency of BS-reads. The hairpin sequencing approach generates methylation data for two DNA strands simultaneously by putting a linking adaptor between Watson and Crick strands and then using PCR and paired-end technology to sequence short reads [51]. The resulting sequences give paired strands so that the original untreated sequences can be recovered. Taking advantage of this ability, we used Bismark [4] with default parameters and Bowtie2 [21] option (command: ./bismark --path_to_bowtie <path to Bowtie2 folder> --bowtie2 --ambiguous <path to Reference genome folder> <input_short_reads.fastq>) to map approximately 308 million reads generated with genome-scale hairpin bisulfite sequencing. Bismark [4] mapped ~ 50% reads uniquely and 25% ambiguously (Figure 2.2). We collected all the ambiguous reads, recovered their original sequences, and used Bowtie2 [21] with default parameters (command: ./bowtie2 -x <reference.fa> -U <input_short_reads.fastq> -S <output.sam>) to map the original sequences. Here the mapping results of recovered sequences are used as the gold standard to validate our Bayesian assignment model. To ensure the quality of the gold standard, we used only those reads with mapping quality score ≥30. As a measure of the goodness of alignment, mapping quality score is a non-negative integer Q = -10 log10p, where p is an estimate of the probability that the alignment does not correspond to the read's true point of origin. Mapping quality is sometimes abbreviated MAPQ. Approximately 48% of the recovered reads were mapped uniquely and also satisfied our mapping quality requirement, and thus were used to validate our model (Figure 2.2). We randomly sampled 1% and 10% of the reads, respectively, from Lane1, Lane2, Lane3, Lane4 and Lane5. We created ten replicates from 1% random sampling and ten other replicates from 10% random sampling for each of the five lanes. Therefore, we had 100 samples altogether, to generate some of the statistics.

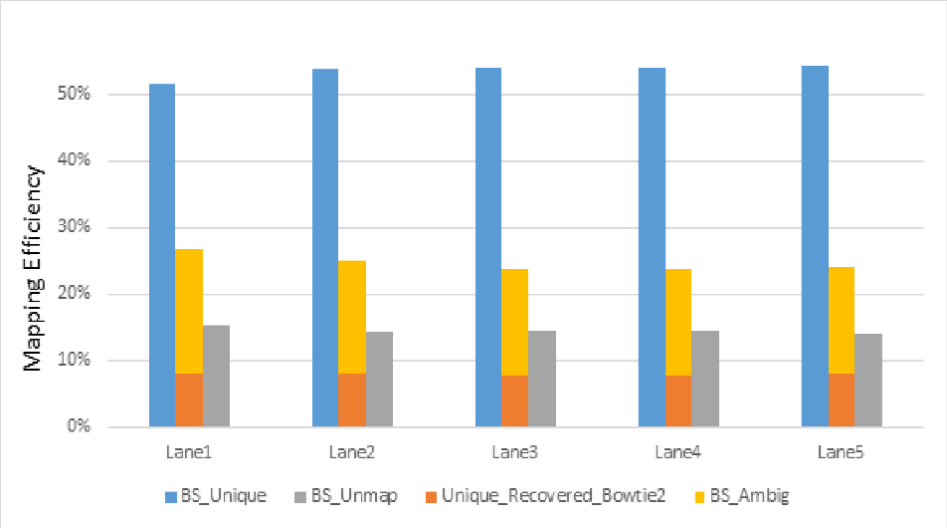


Figure 2.2: Mapping efficiency using Bismark on the mouse embryonic stem cell data for different categories, uniquely mapped reads (blue), multireads (yellow), and unmapped reads (grey). The orange bar is the percentage of multireads that become uniquely mapped with Bowtie2 after recovery to their original sequences using the hairpin bisulfite sequencing technique.

2.2.5 Real data from regular bisulfite sequencing

Although the hairpin bisulfite sequencing data seem ideal as the gold standard from real data, there is still concern that it might differ in some way from data produced by the regular bisulfite sequencing procedure. Therefore, we also applied our assignment model to another real data produced by the regular whole-genome bisulfite sequencing for the human brain (NCBI's SRA accession number: GSM1163695). The human brain data include ten datasets. The DNA bisulfite short read sequences are directional. Each dataset contains around 100 million single-end bisulfite reads for the human frontal cortex. The reads have conventional base call qualities that are Illumina HiSeq 2000 encoded Phred values (Phred64) and have been trimmed to 101 bps. We used Bismark with default parameters to map all the short reads from the ten datasets. Bismark mapped ~75% reads uniquely and ~8% ambiguously. We then used these uniquely mapped reads as "gold standard" to assess the performance of the model. The idea is to shorten these reads so that the original uniquely mapped reads become ambiguously mapped reads, then we apply our model to assign these reads and use the original mapped location as the correct answer to evaluate the assignment accuracy of our model. Specifically, we randomly sampled 1% of the uniquely

mapped reads from the ten datasets and trimmed the reads to shorter ones (i.e., 10 bp shorter than original short reads). After applying Bismark to the trimmed reads, ~50% were uniquely mapped and ~5% multireads. We used our Bayesian model to assign the location of these trimmed multireads and compared the assigned locations with their originally mapped locations.

2.3 Results

2.3.1 Mapping efficiency improvement for simulated data and real data

We simulated 3,000, 40,000, and 100,000 BS-reads for both the human genome and the mouse genome with the setting of read length=76 bp, CG=20% (20% of all CG-cytosines are converted into thymines), CH=99.5% (H can be A, T, or G, 99.5% of all CH-cytosines are converted into thymines), and mutation rate of 0.1% at 30x coverage. We then applied the Bayesian assignment model to score the ambiguously mapped BS-reads and assigned them to their best locations based on the log likelihood ratio *S* (Equation 1). For human BS-reads, the model was able to assign ~ 72% of the multireads to their best locations with an assignment accuracy rate of ~90% for all three datasets (Figure 2.3). The accuracy rate was defined as the percentage of correctly assigned multireads, i.e., the ratio of the number of accurately assigned multireads to the number assigned multireads. For mouse BS-reads, the model was able to assign approximately 53% of all the multireads with an accuracy rate of 80%. Both percentages of assignable multireads and accuracy rates for the mouse data were lower than those for the human. This is likely due to the fact that there are more CTs or TCs in the mouse genome than in the human genome (26.37% vs. 23.87%), consequently, with bisulfite treatment, the mouse genomic DNAs are expected to have a higher frequency of TT posing more challenges to multiread assignment.

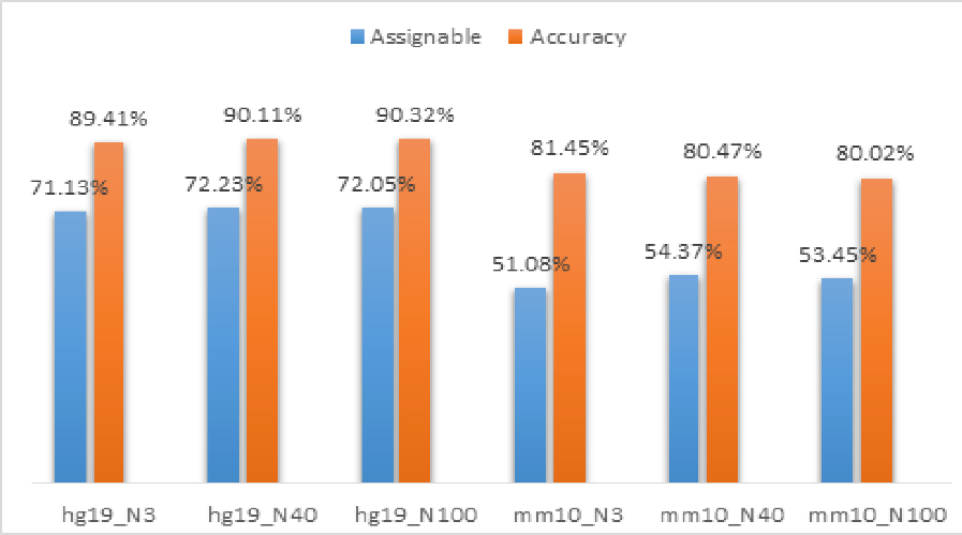


Figure 2.3: Percentages of assignable multireads and accuracy rates of the assigned multireads on six simulated bisulfite datasets generated from the human reference and the mouse reference with read length=76 bp and CG=20% (20% of all CG-cytosines are converted into thymines) and CH=99.5% (99.5% of all CH-cytosines are converted into thymines) and mutation rate of 0.1% at 30x coverage. hg19_N3, hg19_N40, and hg19_N100 denote the datasets with 3k, 40k, and 100k simulated reads respectively for humans; mm10_N3, mm10_N40, and mm10_N100 denote the datasets with 3k, 40k, and 100k simulated reads respectively for mice. All remaining figures use the same notations.

A major challenge in testing the performance of multiread assignment methods on real data is a lack of ground truth for where multireads should be assigned to in the real data. To examine the performance of our Bayesian assignment model on real data, we took advantage of the genome-scale hairpin bisulfite sequencing technique developed recently [52] that allows us to recover the bisulfite converted reads to their original sequences. We assume that once multireads are recovered to their original sequences and these original sequences are mapped to unique locations, the unique locations are indeed true locations. To ensure this assumption to be largely held, we consider only those multireads that are mapped with high mapping quality.

The genome-scale hairpin bisulfite sequencing data for mouse ESC were generated in five sequencing lanes with the Illumina sequencing platforms. For data generated from each of the five lanes, we randomly sampled 1% of the reads and created ten samples per dataset. With assignment

score cut-off of 0.2, in the range of reasonable cut-off point by experiment, 74% of the multireads were assigned to their best locations with ~88% accuracy rates (Figure 2.4). Standard deviations across ten replicates were small, from 0.23-0.42% and from 0.46-0.66% in accuracy rates and assignable percentages, respectively. Thus, 1% random samples were representative of the five datasets.

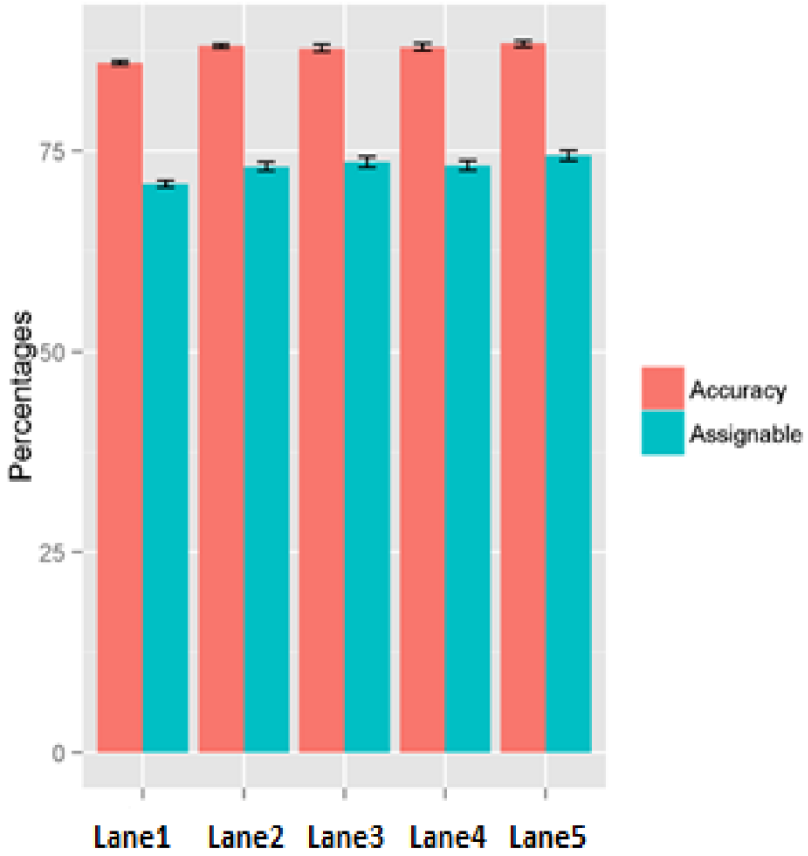


Figure 2.4: Accuracy rates of assigned multireads and percentages of assignable multireads on ten replicates from 1% random samples from five genome-wide hairpin bisulfite sequencing datasets from mouse ESC. The black bar shows the standard deviation.

For human brain whole-genome bisulfite sequencing data, we randomly sampled 1% of the uniquely mapped reads from ten datasets, shortened them so that they "degraded" from previously uniquely mapped reads to multireads. Our model assigned ~75-81% of the multireads to their best locations with ~76-85% accuracy rates (Figure 2.5), therefore, showing similar performance results to that for hairpin sequencing data.

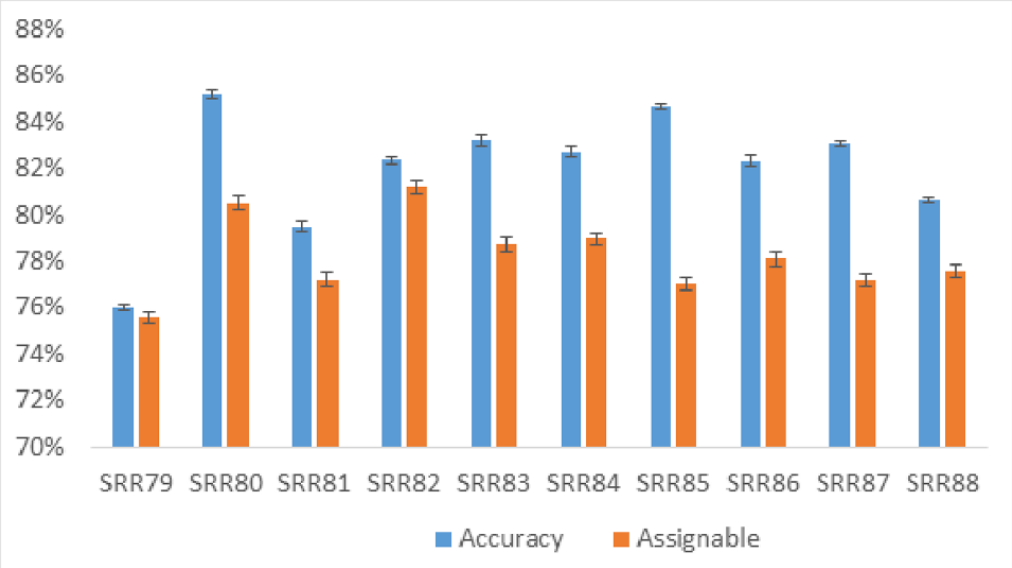


Figure 2.5: Accuracy rates of assigned multireads and percentages of assignable multireads on ten replicates from 1% random samples from ten genome-wide bisulfite sequencing datasets from human frontal cortex (SRA accession number GSM1163695). The black bar shows the standard deviation.

2.3.2 Effect of coverage depth and with/without prior

Table 2.1 shows the effect of sequence coverage on the performance of the model, with and without priors for simulated data. For the simulated human data, the percentage of assignable multireads tends to increase with the coverage depth, and expectedly, the assignment error rate decreases. Compared to simple assignment without a prior, that is, only using observed unique reads to assign multireads, considering prior probability $\pi(M_k|R_k)$ leads to better performance in the model, with much lower error rates (9%-11% compared to 22%-33% for without a prior), although the percentage of assignable multireads decreases at the same time. When the comparison is converted to error rates per read, it is clear that incorporating priors in the method increases the mapping accuracy, with the error rate per read decreasing from 0.01% to 0.005% for the 3x coverage data, and 0.007% to 0.003% for the 30x coverage. The simulated mouse data show a similar pattern, except, in general, has lower percentages of assignable multireads and higher error rates.

Table 2.1: The percentage of assignable multireads and the error rate (ratio of the # of reads assigned to wrong locations to the # of reads that were assigned) as a function of coverage depth and with or without priors for simulated data.

Coverage depth	Without prior		With prior	
	Assignable rate (%)	Error rate (%)	Assignable rate (%)	Error rate (%)
hg19_N40				
3x	96.23	32.55	67.20	10.5
5x	98.10	32.48	69.34	9.96
10x	99.43	27.32	70.55	9.23
25x	99.58	21.95	71.63	9.01
30x	99.37	21.54	72.23	9.00
mm10_N40				
3x	92.56	44.55	49.18	20.68
5x	96.74	44.34	52.44	20.89
10x	98.98	40.67	54.96	19.98
25x	99.43	36.68	54.96	19.81
30x	99.41	36.48	54.37	19.53

For hairpin bisulfite sequencing data, when including prior probabilities, even though the percentages of assignable multireads reduce, the error rates per read decrease (Table 2.2). For example, error rates reduce from 0.00043% to 0.00035% and from 0.00025% to 0.00020% in Lane5_1 and Lane2_10 respectively. Therefore, incorporating priors in the method increases inference accuracy. These results are consistent with simulation results. Compared with simulation results, the accuracy rate improvement in real data is smaller.

Table 2.2: Assignable rates and error rates for assigning multireads with and without priors on 1% and 10% random samples from five genome-wide hairpin bisulfite sequencing datasets from mouse ESC (without priors refers to only using observed unique reads to assign multireads).

Sample ID	Without prior			With prior		
	Assignable rate(%)	Error rate (%)	Error per read (%)	Assignable rate (%)	Error rate (%)	Error per read (%)
Lane1_1	72.17	17.50	0.00043	70.97	14.01	0.00035
Lane1_10	72.27	18.30	0.00004	71.27	13.90	0.00003
Lane2_1	74.60	14.74	0.00239	73.61	11.35	0.00187
Lane2_10	74.67	15.53	0.00025	73.27	12.13	0.00020
Lane3_1	74.44	15.12	0.00275	73.54	12.78	0.00235
Lane3_10	74.54	14.58	0.00026	73.61	12.17	0.00022
Lane4_1	73.24	15.07	0.00282	72.27	12.39	0.00235
Lane4_10	74.35	14.79	0.00027	73.32	12.21	0.00023
Lane5_1	74.76	14.27	0.00251	73.77	12.12	0.00216
Lane5_10	74.23	14.44	0.00025	73.38	12.02	0.00021

We also determined the effect of read coverage on the performance of the assignment model using hairpin sequencing data. Specifically, coverage depth refers to the number of unique reads that overlap with multireads and thus can be used for inference. Table 2.3 shows that as coverage depth increases from 6x to 40x, assignment accuracy increases slightly from 85.92% to 86% in Lane1 and the percentage of assignable reads decreases slightly from 70.9% to 70.82% in Lane1, both at a lower rate than in the simulation study.

Table 2.3: Coverage effect on model performance for 1% random samples from the five hairpin datasets.

Coverage	**Lane 1**		**Lane 2**		**Lane3**	
	Assignable rate (%)	Accuracy rate (%)	Assignable rate (%)	Accuracy rate (%)	Assignable rate (%)	Accuracy rate (%)
6x	70.90	85.92	73.62	88.65	73.41	87.14
10x	70.92	85.92	73.63	88.68	73.37	87.17
20x	70.90	85.95	73.58	88.69	73.45	87.22
30x	70.90	85.99	73.47	88.71	73.42	87.25
40x	70.82	86.00	73.47	88.71	73.53	87.33
Coverage	**Lane4**		**Coverage**	**Lane5**		
	Assignable rate (%)	Accuracy rate (%)		Assignable rate (%)	Accuracy rate (%)	
6x	72.23	87.322	**6x**	73.73	87.79	
10x	72.27	87.329	**10x**	73.77	87.83	
20x	72.28	87.464	**20x**	73.70	87.84	
30x	72.16	87.481	**30x**	73.73	87.84	
40x	72.19	87.505	**40x**	73.74	87.86	

Noteworthy is that the model performs well even with low coverage, for both simulated data and real data. Taken together, the robust performance of the assignment model towards low coverage data makes the model particularly applicable to the current whole genome bisulfite sequencing data (many at 10x coverage).

2.3.3 Effect of read length

To examine the effect of read length on the performance of the Bayesian assignment model, we simulated BS-reads with three read lengths, 51bp, 76bp, and 101bp. All simulated data (3K, 40K, and 100K reads for humans and mice) show similar patterns and only data with 100K BS-reads were used to demonstrate for brevity. Figure 2.6 (left panel) shows that for both human and mouse data, as read length increases, the accuracy rate of assigned multireads to their true locations increases as well as the percentage of assignable multireads. The percentage of increase in accuracy rate is much higher for read lengths increasing from 51bp to 76bp than from 76bp to 101bp.

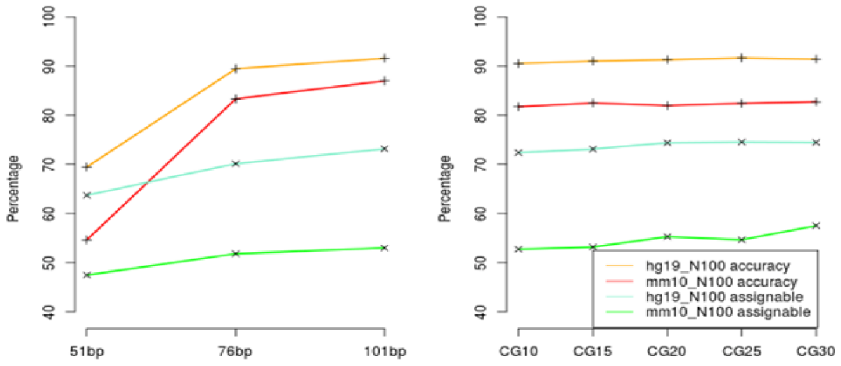


Figure 2. 6: Effect of read length (left panel) and methylation rates at CpGs (right panel, CG10 refers to a methylation rate of 90% at CpGs) on the percentage of assignable multireads and assignment accuracy rates for simulated data generated from hg19 and mm10 at 30x coverage.

In our real data analysis, the hairpin bisulfite sequencing data contain reads with different lengths (Figure 2.3S in supplementary material). This enabled us to determine the effect of read length on our model performance. Reads were classified into 3 groups: short, with read length ≤ 50 bp, moderate, with read length between 50-76 bp, and long, with read length > 76 bp. Figure 2.7 shows that as read length increases, assignable percentages of multireads increase as well as accuracy rates on 1% random samples from the five whole-genome mouse hairpin ESC data. Reads in the long group have highest accuracy rates, around 90% and highest assignable rates, around 75%. Notably, more than a 10% increase in accuracy was observed from the short and moderate groups (i.e., accuracy rate in Lane1 dataset jumps from 75.55% to 85.36%, approximately 10% increase in accuracy).

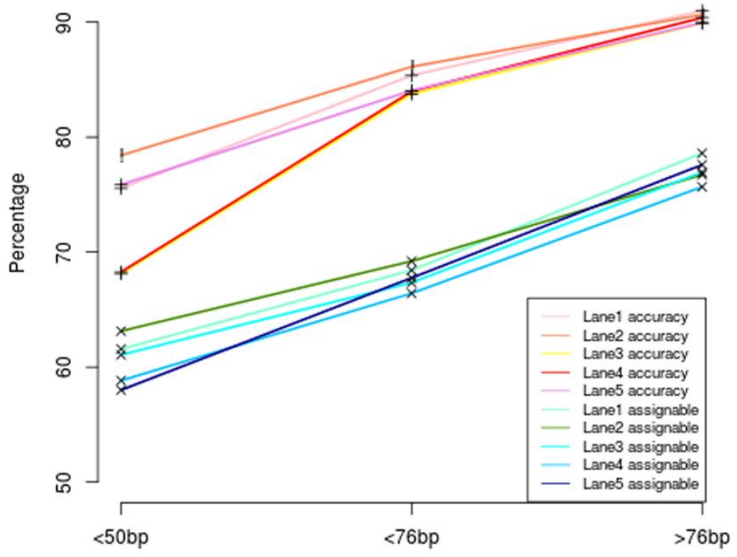


Figure 2.7: Effect of read length on accuracy rates and percentages of assignable multireads on 1% random samples from five genome-wide hairpin bisulfite sequencing datasets from ESC.

2.3.4 Effect of methylation rate at CpGs

As methylation may vary as a function of genomic regions, developmental stages, tissues, species, and so on [47] [53], it is important to examine how the multiread assignment model is

affected by varying methylation rates. We therefore simulated data with different methylation rates (70%, 75%, 80%, 85%, 90%) at CpGs and applied the Bayesian model to assign the multireads in the data. Figure 2.6 (right panel) shows that both the percentage of assignable multireads and assignment accuracy rate change only slightly with respect to different methylation rates, indicating that the method is robust to changes in methylation rates.

2.3.5 Effect of sequencing errors

To examine the effect of sequencing error on the assignment model, we simulated data with different sequencing error rates ranging from 0.002% to 3%. Table 2.4 shows that as sequencing error increases, for both humans and mice, accuracy rate of multiread assignment decreases. However the percentage of assignable ambiguous reads remains similar. Comparatively, sequencing error has a bigger impact on the mouse data than on the human data.

Table 2.4: Effect of sequencing errors on the percentage of assignable reads for simulated data generated from hg19 and mm10 at 30x coverage.

Sequencing error	Accuracy rate (%)		Assignable rate (%)	
	hg19_N40	mm10_N40	hg19_N40	mm10_N40
0.002%	99.31	99.36	71.10	55.05
0.005%	99.12	98.68	71.15	55.19
0.015%	98.97	98.10	71.50	56.60
0.045%	98.62	97.37	71.31	50.87
0.150%	96.97	93.60	71.54	52.23
0.500%	96.31	89.82	72.21	56.06
1.500%	95.30	85.40	72.04	52.28
3.000%	93.23	82.16	72.81	55.56

2.4 Discussion

The whole genome bisulfite sequencing technique allows for determination of C methylation at the whole genome scale and with single nucleotide resolution. Though considered

to be the gold standard for characterizing DNA methylation, its high cost has limited its application to large research laboratories. To make the situation worse, the mapping efficiency of existing tools has been low, mostly 50-70% as compared to over 95% in regular short reads mapping [54]. A large proportion of reads, known as multireads, are routinely discarded from downstream analysis, leading to both biased methylation inference and financial loss. To address the problem, we propose a Bayesian assignment model to help determine the most likely locations the multireads should be mapped to. Results show that the model is effective and can be used to increase the number of uniquely mapped read, and thus allows users to make the best use of the data possible.

Our analysis demonstrates that read length shows a much bigger positive impact on the model performance for real data than for simulated data: both the percentage of assignable reads and the assignment accuracy rate increase much more with read length increase in real data (Figure 2.7) than in simulated data (Figure 2.6). This is likely because reads from real data carry more information than simulated reads giving the assignment model more power to differentiate among the competing locations of multireads, and thus lead to better performance in real data. We note that real whole genome bisulfite sequencing experiments usually generate reads with 100bp or longer. Even after ends trimming, these reads are mostly longer than 76bp. The results here suggest that, with real data, the assignment model is capable of recovering 14-20% of the multireads to their true locations (Figure 2.2), and these reads can be included in downstream analysis to provide more comprehensive information on methylation at the genome level. It might be interesting to conduct a comprehensive survey to examine how these reads that are routinely thrown away affect the downstream inference were they included in the downstream analysis.

Due to the high cost of whole genome bisulfite sequencing, the depth of sequencing coverage is often low, approximately 10X for many experiments. This poses an additional challenge to downstream analyses such as methylation calling and variant calling. For example, Bis-SNP, a program that does methylation calling and SNP calling for bisulfite sequencing data, requires an average of 30X coverage for correctly calling 96% of the SNPs [55]. Our results demonstrate that even with low coverage of ~5X-10X, the Bayesian scoring model performs well and is stable (Tables 2.1 and 2.3).

Our Bayesian scoring model enables a high proportion of multireads to be mapped to unique locations, which in turn increases the overall amount of sequence data suitable for the downstream methylation inference. An interesting issue to examine is whether methylation ratios are affected as a result of changes in the compositions of reads. Thus, we took a set of 50,000 multireads and ~500,000 uniquely mapped reads overlapping with these multireads and another set of ~550,000 uniquely mapped reads in these regions from the human whole-genome bisulfite sequencing data (SRA accession number SRX306253, GSM1163695, see methods for details) and used Bismark for methylation calling. The methylation ratios at CpG sites were very similar between the two datasets. We also took a set of 100,000 multireads and ~300,000 uniquely mapped reads and another set of ~400,000 uniquely mapped reads around these regions and did the same analysis. The methylation ratios were still similar but as expected there were more CpG sites covered in the former dataset. Taken together, the results suggest it depends on data coverage and percentages of multireads. Specifically, CpG methylation ratios are expected to stay similar if the coverage is low, however, more CpG methylation sites will be covered. On the other hand, if the coverage is high, CpG methylation ratios are expected to be more accurate and more CpG sites will be covered. Again, the advantage of multiread mapping is to gain valuable information from "unusable" data by traditional mappers, which benefits the subsequent calling procedure and downstream analysis.

Results for both simulated data and real data (Tables 2.1 and 2.2) show that incorporating prior knowledge such as mutation rates and context specific methylation levels into the assignment model helps improve the accuracy of the assignment. Moreover, for organisms without such prior information, the assignment model can still provide robust assignment, especially reflected by the real data. Comparatively, it is clear that information gleaned from uniquely mapped reads plays a more important role in correctly assigning multireads.

A common problem in the development of tools for bisulfite short read mapping is the lack of a gold standard. We addressed this by taking advantage of the hairpin bisulfite sequencing data that allows the recovery of the original reads (refer to [51] for the mechanism of read recovery), and assuming that the unique locations that recovered reads are mapped to are true locations. Although we required a high mapping quality ($\geq$30), it is still possible that some of the true

locations are false positives. However, the consistency shown between simulated data and real data suggests that even if there are false positives in the gold standard, the number should be very low. Another concern for using hairpin bisulfite sequencing data is that its characteristics might be different from those of the regular bisulfite sequencing data. However, our model performance on regular bisulfite sequencing data is very similar to that on hairpin sequencing data, suggesting that the hairpin sequencing data is representative and can serve as gold standard for real data.

2.5 Conclusion

A major problem in mapping bisulfite short reads is the high percentage of multireads caused by bisulfite conversion. To our knowledge, no program is devoted to address this problem. Here we present a Bayesian model to assign multireads to the best possible locations. Simulation and real data results show that our assignment method is effective in mapping multireads with high accuracy. We investigated several factors that might affect the model performance, including methylation level, coverage, sequencing error, and read length. More specifically, methylation level has little effect, whereas sequencing errors have a negative impact on model performance. Increasing depth of coverage and read length will increase the accuracy of assigning multireads. The model performs quite well even with low read coverage. Therefore, our scoring method can be used to effectively improve the mapping results of bisulfite sequencing data.

Supporting Information

S2.1 Table: Prior calculation

Prior probabilities of all possible cases of alignments on the forward direction

Notation:

Pr(me) is the probability of methylation event occurring at a position

Pr(SNP) is the probability of mutation event occurring at a position

Pr(AB) is the probability of A to B mutation event occurring at a position, i.e. Pr(AT) is the probability of A on the reference genome changes to T on the multiread.

Bases in green are observed bases, in black are unobserved. Cs/Gs in red indicate methylated Cs/Gs, in blue unmethylated Cs/Gs.

Table 2.1a: Prior probabilities at A reference genome of forward alignments

Reference base	A	A	A	A
Unobserved	**A**	**C**	**G**	**T/C**
Multiread base	A	C	G	T
Inference	No mutation	A to C mutation and methylated C	A to G mutation	A to T mutation or A to C mutation and unmethylated C
Prior	1-Pr(SNP)	Pr(AC)xPr(me)	Pr(AG)	Pr(AT)+Pr(AC)x[1-Pr(me)]

Note: 1-Pr(SNP) + Pr(AC)xPr(me)+ Pr(AG)+ Pr(AT)+Pr(AC)x[1-Pr(me)]=1 (sum of all priors is 1)

Table 2.1b: Prior probabilities at C reference genome of forward alignments

Reference base	C	C	C	C
Unobserved	**A**	**C**	**G**	**T/C**
Multiread base	A	C	G	T
Inference	C to A mutation	No mutation and methylated C	C to G mutation	C to T mutation or no mutation and unmethylated C
Prior	Pr(CA)	[1-Pr(SNP)]*Pr(me)	Pr(CG)	Pr(CT)+[1-Pr(SNP)]x[1-Pr(me)]

Table 2.1c: Prior probabilities at G reference genome of forward alignments

Reference base	G	G	G	G
Unobserved	**A**	**C**	**G**	**T/C**
Multiread base	A	C	G	T
Inference	G to A mutation	G to C mutation and methylated C	No mutation	G to T mutation or G to C mutation and unmethylated C
Prior	Pr(GA)	Pr(GC)xPr(me)	1-Pr(SNP)	Pr(GT)+Pr(GC)x[1-Pr(me)]

Table 2.1d: Prior probabilities at T reference genome of forward alignments

Reference base	T	T	T	T
Unobserved	**A**	**C**	**G**	**T/C**
Multiread base	A	C	G	T
Inference	T to A mutation	T to C mutation and methylated C	T to G mutation	No mutation or T to C mutation and unmethylated C
Prior	Pr(TA)	Pr(TC)xPr(me)	Pr(TG)	[1-Pr(SNP)]+Pr(TC)x[1-Pr(me)]

S2.2 Table: Prior probabilities of all possible cases of alignments on the reverse direction

Table 2.2a: Prior probabilities at A reference genome of reverse alignments

Reference base	A	A	A	A
Unobserved	**A/G**	**C**	**G**	**T**
Multiread base	A	C	G	T

Inference	No mutation or A to G mutation and unmethylated G	A to C mutation	A to G mutation and methylated G	A to T mutation
Prior	[1-Pr(SNP)]+Pr(AG)x[1-Pr(me)]	Pr(AC)	Pr(AG)xPr(me)	Pr(AT)

Table 2.2b: Prior probabilities at C reference genome of reverse alignments

Reference base	**C**	**C**	**C**	**C**
Unobserved	**A/G**	**C**	**G**	**T**
Multiread base	**A**	**C**	**G**	**T**
Inference	C to A mutation or C to G mutation and unmethylated G	No mutation	C to G mutation and methylated G	C to T mutation
Prior	Pr(CA)+Pr(CG)x[1-Pr(me)]	1-Pr(SNP)	Pr(CG)xPr(me)	Pr(CT)

Table 2.2c: Prior probabilities at G reference genome of reverse alignments

Reference base	**G**	**G**	**G**	**G**
Unobserved	**A/G**	**C**	**G**	**T**
Multiread base	**A**	**C**	**G**	**T**
Inference	G to A mutation or no mutation and unmethylated G	G to C mutation	No mutation and methylated G	G to T mutation
Prior	Pr(GA)+[1-Pr(SNP)]x[1-Pr(me)]	Pr(GC)	[1-Pr(SNP)]xPr(me)	Pr(GT)

Table 2.2d: Prior probabilities at T reference genome of reverse alignments

Reference base	**T**	**T**	**T**	**T**
Unobserved	**A/G**	**C**	**G**	**T**
Multiread base	**A**	**C**	**G**	**T**
Inference	T to A mutation or T to G mutation and unmethylated G	T to C mutation	T to G mutation and methylated G	No mutation
Prior	Pr(TA)+p(TG)x[1-Pr(me)]	Pr(TC)	Pr(TG)xPr(me)	1-Pr(SNP)

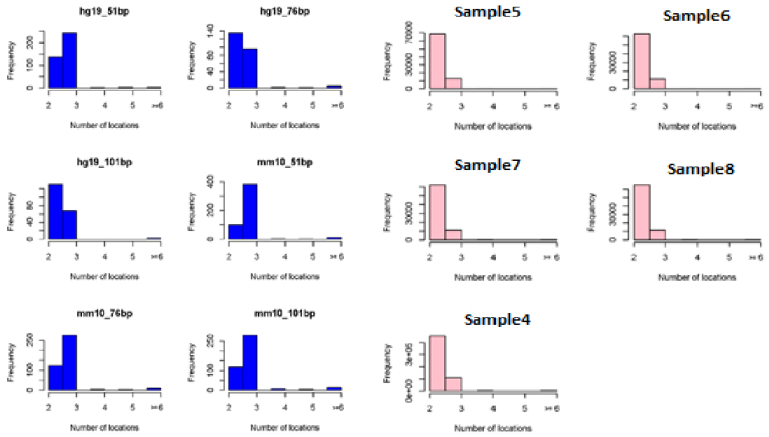


Figure 2.1S: Histogram of number of genomic locations Bismark found for multireads in simulated data (left) and in real hairpin data (right)

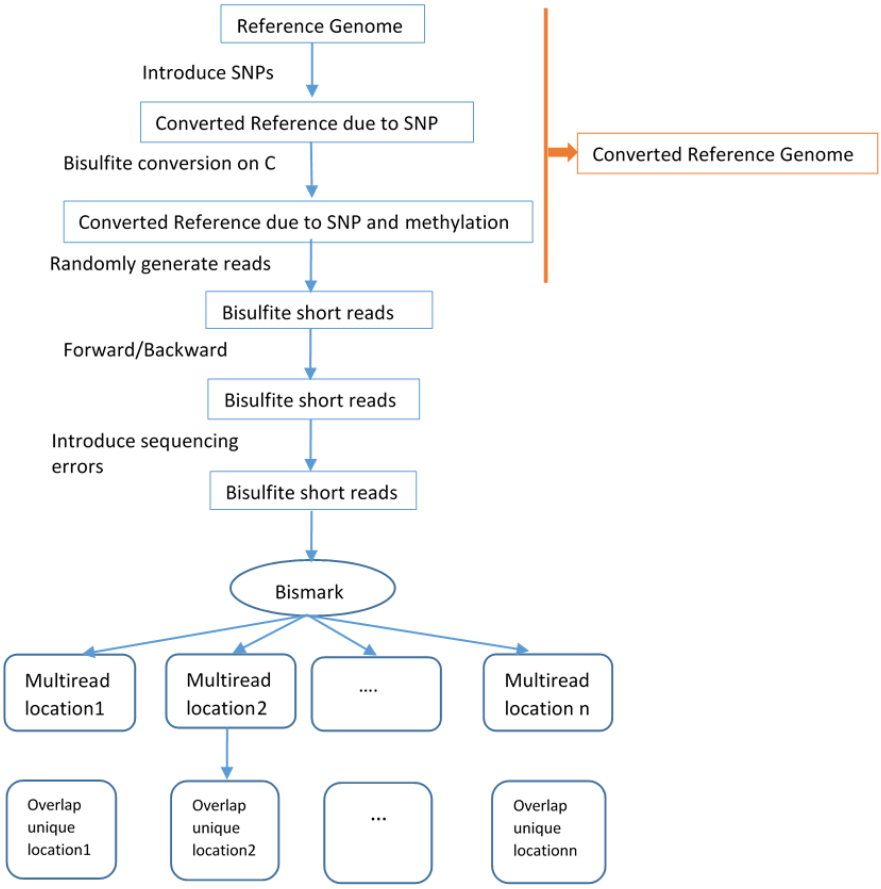


Figure 2.2S: Pipeline for generating bisulfite short reads, multireads, and overlap unique reads.

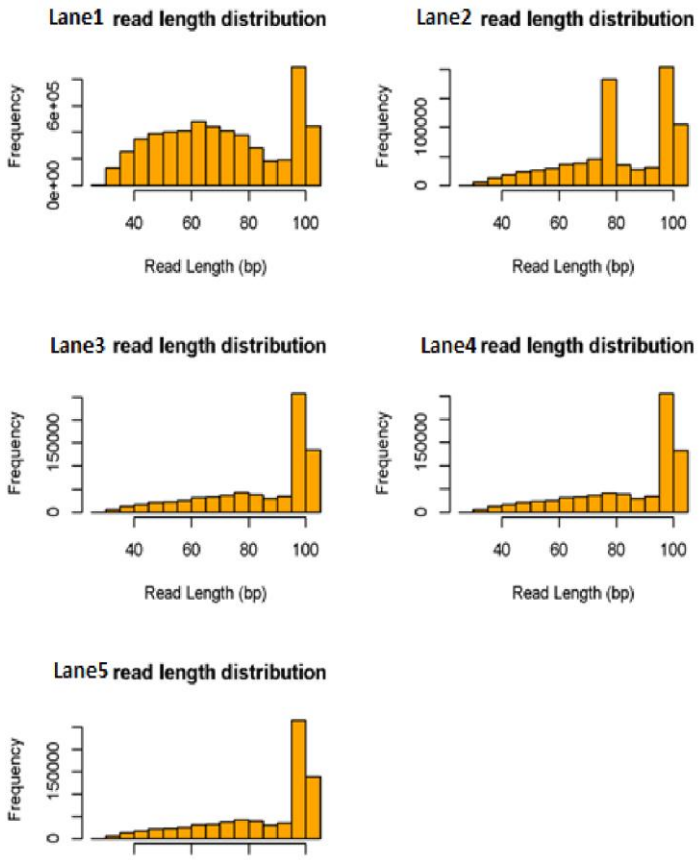


Figure 2.3S: Histograms of read length from Lane1, Lane2, Lane3, Lane4, and Lane5 hairpin data

Chapter 3

Identification of Differentially Methylated Sites from Weak Methylation Effect

Abstract

Motivation: DNA methylation is an epigenetic alteration crucial for differentiating normal and stress responses. In order to better understand phenotype changes among cells or tissues during development and stress response stages, it is essential to accurately characterize genome-wide DNA methylation. Whole genome bisulfite sequencing has made it possible to characterize large-scale DNA methylation at the single nucleotide resolution. An essential task following the generation of bisulfite sequencing data is to detect differentially methylated cytosines (DMCs) between different samples. Many statistical methods for DMC detection ignore the dependency of methylation patterns across the genome, which could lead to inflated type I error, i.e., identifying

DMCs that are not truly significant. Furthermore, small sample sizes and weak methylation effect among different phenotype categories make it difficult for these methods to accurately detect DMCs. To address these issues, we adopt the wavelet-based functional mixed model (WFMM) approach to detect DMCs and compare its performance to that of the most popular DMC detection tool methylKit.

Results: Analyses of simulated data based on a reference data set that measure the effects of herbicide glyphosate on *Arabidopsis thaliana* show that WFMM results in higher sensitivity and specificity in detecting DMCs compared to methylKit especially when the methylation differences among phenotype groups are small. Moreover, the performance of WFMM depends less on read coverage and is robust to sample sizes, making it particularly attractive considering the prohibitive cost of bisulfite sequencing. The analysis of the *Arabidopsis thaliana* data under varying herbicide glyphosate dosages and the analysis of monozygotic twins who have different pain sensitivities (both datasets have weak methylation effect, i.e. average methylation differences between two phenotype groups is less than 0.01) show that WFMM can find more relevant DMCs related to the phenotype of interest compared to methylKit.

[#]Differentially methylated regions (DMRs) are genomic regions with different DNA methylation status across biological samples. DMRs and DMCs are the same concepts, with the only difference being how methylation information across genome is summarized. If methylation levels are determined by grouping neighboring cytosine sites, then they are DMRs; if methylation levels are calculated based on single cytosines, they are DMCs.

3.1 Introduction

DNA methylation is an important epigenetic mechanism in controlling gene expression, silencing of genes on the inactive X chromosome, imprinted genes, and parasitic DNAs [48]. Accurate characterization of DNA methylation is essential for understanding genotype-phenotype association, gene-environment interaction, diseases, and stresses [44]. Genome-wide bisulfite treated DNA sequencing has enabled the measurement of DNA methylation at the single nucleotide resolution. After DNA is treated with sodium bisulfite, unmethylated Cs are converted to Ts, whereas methylated Cs remain unchanged. At a single cytosine site, methylation levels are estimated by taking the ratio of C/(T+C) where C and T are the counts of cytosines and thymines

respectively from all aligned reads at the site. The count of Ts represents the number of unmethylated Cs and the count of Cs represents the number of methylated Cs. The most common task is to detect differentially methylated cytosine sites across different phenotype samples (e.g., dosage vs. non-dosage samples, and patients vs. healthy people). Although numerous statistical methods such as Fisher's exact test and logistic regression have been used for the detection of DMCs [41], several challenges remain. First, most current methods make the assumption that individual cytosine methylation levels are independent across the genome. This assumption is questionable as it has been shown that methylation levels of nearby cytosine sites are highly correlated ([42], Figure 3.1). Assuming independence across cytosine sites can lead to underestimation of the p-values and inflated type-I error, resulting in mistakenly identifying more significant DMCs than the underlying truth Second, due to the high cost of whole genome bisulfite sequencing, studies are often done with only a small number of samples for each phenotype, which makes it difficult to detect small methylation differences. To address these issues, Lee and Morris [56] adapted the wavelet-based functional mixed model (WFMM) developed by Morris and Carroll [43] to identify differentially methylated sites. In this paper, we validate the effectiveness of WFMM by analyzing two different methylation data sets and compared its performance with the commonly used approach methylKit. We introduced an empirical approach to setting the tuning parameters to specific methylation profiles in real data to detect more relevant DMCs that related to phenotype changes under different stresses. Our results showed that WFMM has advantages over methylKit when there is weak methylation effect and sample sizes are small. When methylation effect is large enough, WFMM and methylKit are comparable. The paper is organized as follows. First, we describe the methodology. Then we describe the simulation studies based closely on our herbicide glyphosate experiments with *A. thaliana* [57]. Finally we evaluate the WFMM method on simulated and real datasets from whole genome bisulfite sequencing of *A. thaliana* leaves and whole genome methylation profiles of monozygotic (MZ) twins and make comparison with the methylKit program [58].

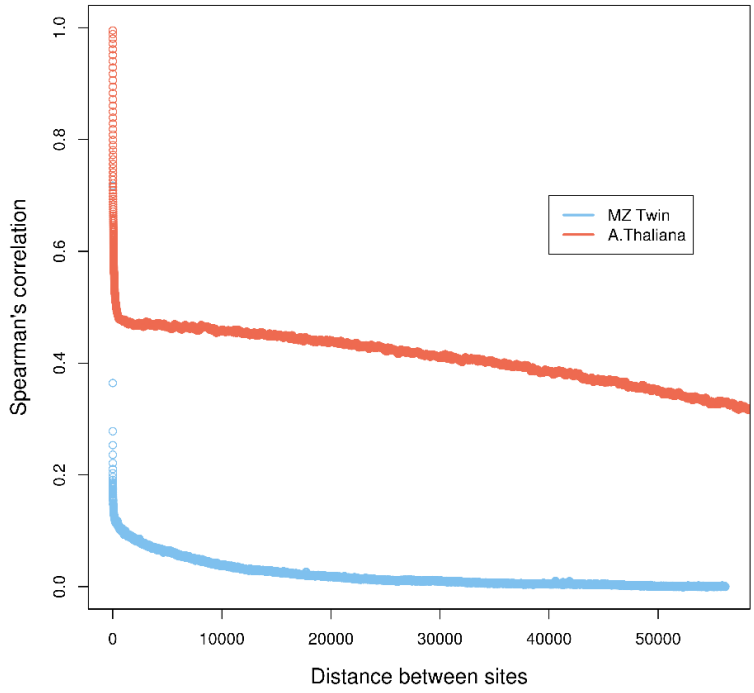


Figure 3.1: Correlation of methylation levels of neighboring cytosine regions in monozygotic twin and neirboring cytosines in *A. thaliana* datasets.

3.2 Methods

3.2.1 Wavelet based functional mixed models

Assume that all methylation measurements come from N individuals across all $\mathcal{T}$ genomic locations. A functional mixed effect model can be represented by

$$y_i(t) = \sum_{j=1}^{J+1} X_{ij} B_j(t) + \sum_{m=1}^{M} Z_{im} U_m(t) + E_i(t), t \in \mathcal{T} \quad (1)$$

where $y_i(t)$ represents the logit-transformation of methylation levels at a genomic location $t \in \{t_l; l = 1, \dots, \mathcal{T}\}$ for the ith individual, $i = 1, \dots, N$. $X_{ij} = 1$ if individual i belongs to treatment j and 0 otherwise, for $1 \leq j \leq J$ The function $B_j(t)$ represents the fixed effect corresponding to treatment and other covariates of interest). Z_{im} is a random covariate that takes into account variations in $y_i(t)$ that are caused by potential multilevel structures in the measurements (e.g.,

when multiple subjects from the same family were measured, then each family will introduce its own random effect and $Z_{im} = 1$ if individual i is from family m and $U_m(t)$ is the random effect of family m). $E_i(t)$ is a residual error function. Using vectorized formulation, we may write model (1) as

$$\boldsymbol{Y}(t) = \boldsymbol{X}\boldsymbol{B}(t) + \boldsymbol{Z}\boldsymbol{U}(t) + \boldsymbol{E}(t), t \in \mathcal{T} \quad (1a)$$

where $\boldsymbol{Y}(t) = [Y_1(t), \dots, Y_N(t)]^T$, $\boldsymbol{B}(t) = \left[B_1(t), \dots, B_J(t)\right]^T$, $\boldsymbol{U}(t) = [U_1(t), \dots, U_M(t)]^T$, and $\boldsymbol{E}(t) = [E_1(t), \dots, E_N(t)]^T$. Here, $\boldsymbol{Y}$ is a $N \times \mathcal{T}$ matrix across all $\mathcal{T}$ genomic locations for all N individuals. $\boldsymbol{X}$ is an $N \times J$ design matrix that indicates which treatment group the N individuals belong to or other covariates of interest (e.g., a phenotype), the $\boldsymbol{B}$ $(J \times \mathcal{T})$ matrix contains the fixed effects of the covariates. The tth column of $\boldsymbol{B}$, denoted by $\boldsymbol{b}_t$, is a J-dimensional vector describing the effects the J covariates on $\boldsymbol{Y}$ at genomic location t.

For example, if we let the ith row of $\boldsymbol{X}$ be a 1/0 vector to indicate which of the herbicide glyphosate dosage groups the ith plant was treated, $i = 1, \dots, N$, then $\boldsymbol{b}_t$ corresponds to the effect of dose levels on $\boldsymbol{Y}$ at genomic location t. In equation (1a), $\boldsymbol{Z}$ is a design matrix for random effects that takes into account variations in $\boldsymbol{Y}$ that are caused by potential multilevel structures in the measurements; $\boldsymbol{U}$ contains the corresponding random effects; and $\boldsymbol{E}$ is an $N \times \mathcal{T}$ matrix of residual errors. We assume that $\boldsymbol{E}$ is multivariate normal with mean 0 and variance-covariance matrix **S**. For example, in our *A. thaliana* experiment, there are four plants for each of the 0%, 5%, 10% glyphosate-treated group. Therefore, the $\boldsymbol{X}$ design matrix is a 12×3 and $\boldsymbol{B}$ is a $3 \times \mathcal{T}$ matrix, where $\mathcal{T}$ is the number of cytosine locations. Since the *A. thaliana* data does not involve multilevel structures, the random effect term in equation (1a) is omitted. The resulting functional model can be rewritten as

$$\boldsymbol{Y}(t) = \boldsymbol{X}\boldsymbol{B}(t) + \boldsymbol{E}(t), \quad t \in \mathcal{T} \quad (1b)$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{B} = [\boldsymbol{b_1}\ \boldsymbol{b_2}\ \boldsymbol{b_3} \ldots \boldsymbol{b_{\mathcal{T}}}],$$

each $\boldsymbol{b_t}$ is a column vector consisting of p=3 elements/groups giving the mean methylation profiles for each group at a given genomic location t.

To incorporate nearby methylation correlations across all genomic locations $\mathcal{T}$ into the model, we first use a basis function transform to transform model (1b) from the original data space into the basis space, and then fit the basis space model to estimate parameters. Finally, we transform results back to the original data space for inference. In particular, we apply the discrete wavelet transform (DWT) to each row of $\boldsymbol{Y}$ to obtain a $N \times \mathcal{T}^*$ matrix of wavelet coefficients $\boldsymbol{D}$. The corresponding wavelet space model can be obtained by post-multiplying both sides of Equation (1b) by $\boldsymbol{\Phi'}$, the wavelet transformation operator**:**

$$\boldsymbol{Y\Phi'} = \boldsymbol{XB\Phi'} + \boldsymbol{E\Phi'} \qquad (1b)$$

$$\boldsymbol{D=XB^*+E^*} \qquad (2)$$

where $\boldsymbol{\Phi'}$ is a $\mathcal{T} \times \mathcal{T}^*$ wavelet transformation operator, $\boldsymbol{D=Y\Phi'}$, $\boldsymbol{B^*=B\Phi'}$ and $\boldsymbol{E^*=E\Phi'}$. The model (2) is a wavelet space model with $\boldsymbol{D}$, $\boldsymbol{B^*}$ and $\mathbf{E^*}$ representing the wavelet coefficients of $\boldsymbol{Y}$, $\boldsymbol{B}$, and $\boldsymbol{E}$ respectively. We adopt a Bayesian approach to fit model (2) following Morris and Carroll (2006) [43]. The posterior samples of the parameters in (2) are obtained by employing a Markov chain Monte Carlo (MCMC) algorithm. Inverse DWT is finally applied to the posterior samples of $\boldsymbol{B^*}$ to obtain posteriors for $\boldsymbol{B}$ in the data domain, which were subsequently used to identify DMCs following a Bayesian false discovery rate approach.

3.2.2 Bayesian false discovery rate (FDR)

Based on the posterior samples of $\boldsymbol{B}$, we can identify significant regions either on $\boldsymbol{B}$ or on the contrast effects that contains the differences between covariate effects in $\boldsymbol{B}$. For example, in the *A. thaliana* data example, since we are interested in identifying DMCs with different dosage effects, we will calculate the contrast effects by pre-multiplying $\boldsymbol{B}$ with a contrast effect operator $\begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & 1 \end{pmatrix}$, which transforms the effect of each dosage level to the contrast effects of level 2 vs. level 1, level 3 vs. level 2, and level 3 vs. level 1 respectively. We will apply this operator to all posterior samples of $\boldsymbol{B}$ to obtain the posterior samples of the contrast effects. Denote $C_a(t)$, $t \in \{t_l; l = 1, \dots, T\}$ the ath contrast effect, identifying significant DMCs on $C_a(t)$ amounts to identifying locations on $C_a(t)$ that are large in magnitude. We achieve this by performing a Bayesian multiple testing that controls the overall false discovery rate following Morris et al. [43], Zhu et al. [59], and Lee and Morris [56].

Specifically, in the Bayesian FDR approach, we detect locations in $\{t_l; l = 1, \dots, T\}$ that has $C_a(t)$ values greater than some threshold δ (in absolute value) based on G posterior samples of $C_a(t)$ for all contrast effects. We first calculate the pointwise posterior probability of at least δ difference at t_l by calculating $\hat{p}_a(t_l) = \Pr\{|C_a(t_l)| > \delta | \boldsymbol{Y}\} \approx \frac{\sum_{g=1}^{G} I\{|C_a(t_l)^{(g)}| > \delta\}}{G}$, where $C_a(t_l)^{(g)}$ denotes the gth sample of C_a at t_l. Then, we find a cut-point ϕ_α for $\hat{p}_a(t_l)$ so that the expected global Bayesian FDR is less than or equal to a pre-specified level α. We claim all of the t_l on which $\hat{p}_a(t_l) > \phi_\alpha$ as genomic locations with $C_a(t_l)$ greater than δ.

3.3 Data and Simulation

3.3.1 *A. thaliana* treated with herbicide glyphosate experiment

We previously investigated methylation profiles of twelve *A. thaliana* plants induced by herbicide glyphosate at different dosage concentrations [57]. Blocks of four *A. thaliana* plants were randomly assigned to glyphosate treatment at three different dosages, 0%, 5%, and 10%. Following glyphosate treatment, these plants were transferred to a growth chamber with a 12-hour light cycle and light intensity of 90 μmol m^{-2} s^{-1} and let grow for approximately 2 weeks for the 0% and 5%

glyphosate-treated plants and 8 weeks for the 10% glyphosate-treated plants until fully-developed siliques were formed [57]. The tissue samples from these twelve plants were sent to Genomics Research Laboratory at Biocomplexity Institute of Virginia Tech for bisulfite sequencing. First, the sequenced reads' quality was checked using FastQC [60] to eliminate adapter sequences and barcodes using Trimmomatic [61] and FastX Tookit [62]. Low quality reads (quality score Q<30) were exclueded. After all quality checks, bisulfite short sequences were aligned to *A. thaliana* (TAIR 10) reference genome using Bismark aligner (v 0.14.5) using default parameters (-n 1 -l 50) [4]. Cytosine methylation level information was extracted from aligned reads using Bismark [4] methylation extractor. In total, there are 3,348,756 cytosines in the dataset for detecting significant methylated cytosines differentiating glyphosate dosage groups.

3.3.2 Methylation level simulation

We aimed to generate methylation profiles that closely mimic the real data collected from our experiment ([57], Figure 3.1S). We generated two sets of methylated cytosines, one set with correlation among nearby cytosine sites and the other set without methylation correlation. For uncorrelated dataset, we first randomly selected 10,000 out of 100,000 cytosine sites as DMCs (~10% of all cytosine sites are differentially methylated). The average methylation levels for each of the three dosage groups, i.e., no treatment (0%) or two different sub-lethal doses (5% and 10%) of herbicide glyphosate were generated from estimating the real *A. thaliana* dataset. To illustrate, from the real *A. thaliana* dataset, for each cytosine site, pairwise mean methylation differences between 0% vs 5%, 5% vs. 10% and 0% vs 10% were calculated. If one of the mean methylation differences was greater than 0.04, cytosine sites were considered differentially methylated, therefore the methylation levels at these cytosine sites were used to generate methylation profiles for differentially methylated sites (true positive methylation differentiation) in simulated data. If none of the mean methylation differences between any of the two groups were greater than 0.04, cytosine sites were considered nondifferential. Thus, methylation levels at these nondifferential sites were used to generate not differentially methylated sites (true negative methylation differentiation) in simulated data (Figure 3.2S).

To generate correlated simulated datasets, we first divided the real *A. thaliana* dataset into blocks of 100,000 cytosine sites and we randomly chose blocks to generate methylation profiles

for simulated data. For each random block, if one of the mean methylation differences was greater than 0.04, cytosines were considered differentially methylated, therefore the methylation levels at these cytosine sites were used to generate methylation profiles for differentially methylated sites in simulated data with correlation. Otherwise, sites were considered nondifferential and used to simulate true negative methylation profiles (Figure 3.2S). Individual methylation levels for each of the three dosage groups from both correlated and uncorrelated datasets were generated from truncated normal distribution ranged from 0 to 1 with mean and standard deviations calculated from the real *A. thaliana* dataset.

We changed methylation difference profiles by changing cutoff value for a cytosine site to be considered differentially methylated by increasing 0.04 to 0.08, 0.1, 0.12, 0.15, 0.2, and 0.25. To illustrate, with the cutoff value 0.25, only cytosines with at least one of the pairwise mean methylation differences greater than 0.25 are considered differentially methylated. We also increased sample sizes for each dosage group from 4 to 10, 20, 30, and 40 to examine how the WFMM method performs under different scenarios and compared its performance to the commonly used program methylKit [58].

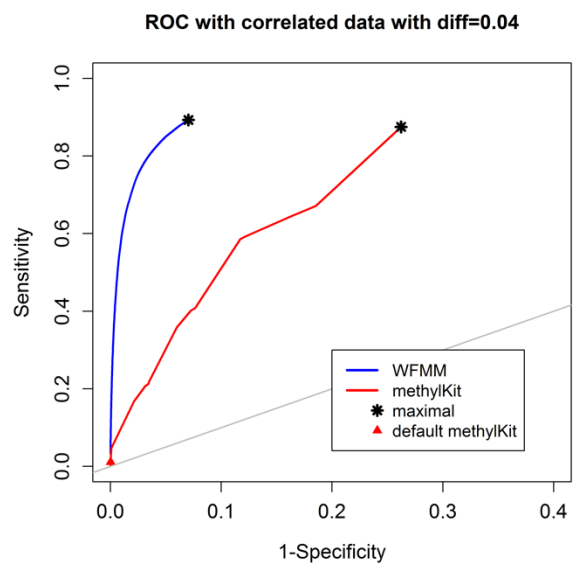
3.4 Results

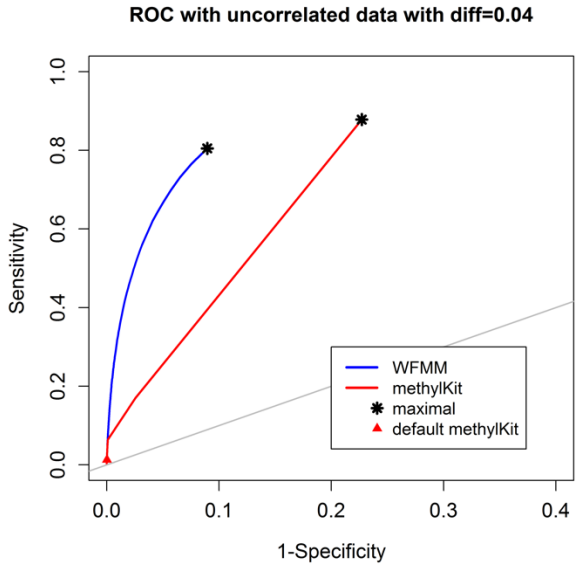
3.4.1 Simulation results

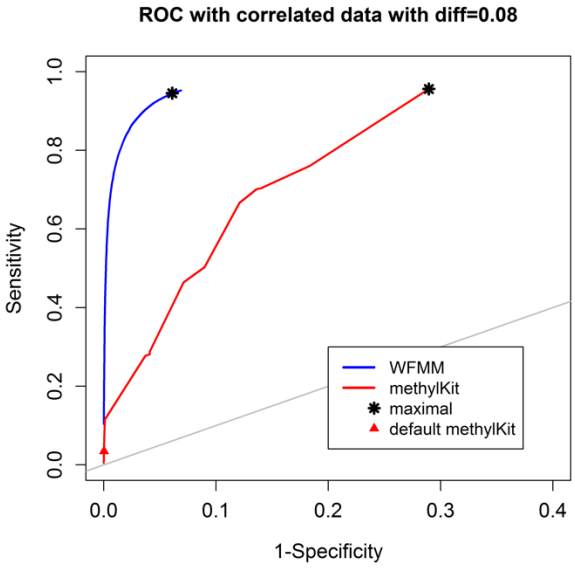
(1) Effect of the degree of methylation difference

The degree of methylation difference between different phenotypes is an obvious factor to consider when examining the performance of tools for detecting differentially methylated cytosines. Therefore, we examined the performance of the WFMM method through receiver operating characteristic (ROC) curve analysis and compared it to methylKit [58] for different degree of methylation difference. Figure 3.2 shows the performance of the two methods with different methylation difference cutoffs. We used Youden's rule to find the optimal threshold for the delta parameter (δ) in WFMM and the qvalue parameter in methylKit. MethylKit uses qvalues, the adjusted P-values for multiple testing correction. According to Youden's rule, the optimal threshold is where the sum of sensitivity and specificity is maximized. Figure 3.2 shows that overall WFMM performs better than methylKit with higher sensitivity and specificity in both

correlated and uncorrelated scenarios. When differentially methylated cutoff is 0.04 or 0.08 and in both correlated and uncorrelated cytosines, the optimal value for delta δ parameter in WFMM is 0.01 and the optimal value for qvalue parameter in methylKit is 1.00. Noteworthy is that there is an improved performance in WFMM, i.e., higher specificity and slightly higher sensitivity in correlated data compared to uncorrelated data whereas methylKit performance is similar in both scenarios.







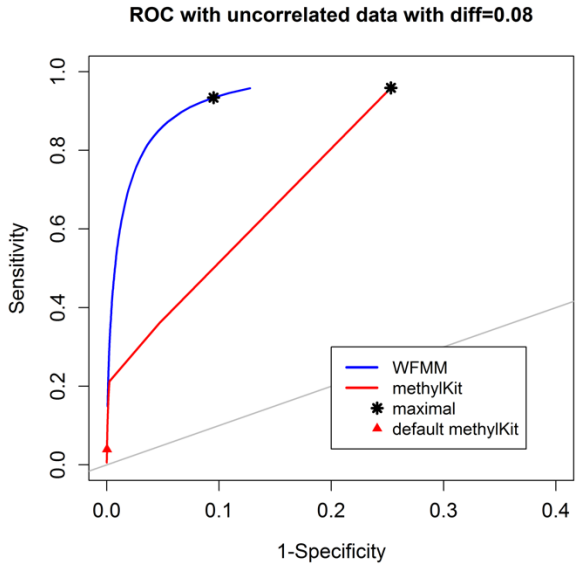
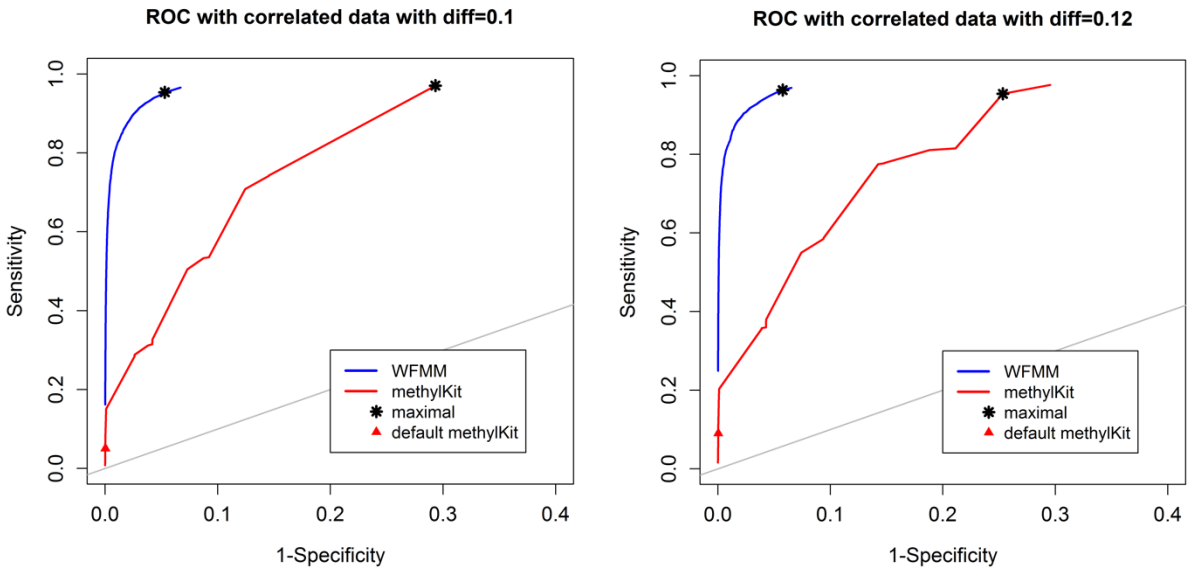


Figure 3.2: ROC curve comparison between WFMM (blue curve) and methylKit (red curve) when differentially methylated cutoff is 0.04 in correlated cytosines (top left), uncorrelated cytosines (top right) and when differentially methylated cutoff is 0.08 in correlated cytosines (bottom left), uncorrelated cytosines (bottom right).

Figure 3.3 shows that as increasing differentially methylated cutoff from 0.1, 0.12, 0.15, 0.2 and 0.25, the gaps in ROC curves between WFMM and methylKit become narrower. Specifically, there is little improvement in WFMM whereas the performance of methylKit improves with increasing differentially methylated cutoff values. When differentially methylated cutoff is 0.2 or 0.25, WFMM and methylKit perform similarly. To illustrate, when differentially methylated cutoff = 0.25, at an optimal threshold δ=0.013 in WFMM, and at an optimal threshold qvalue=0.76 in methylKit, WFMM has higher sensitivity (0.953 vs. 0.806) but lower specificity (0.696 vs. 0.828) than methylKit. Therefore, there is a trade-off between two methods.



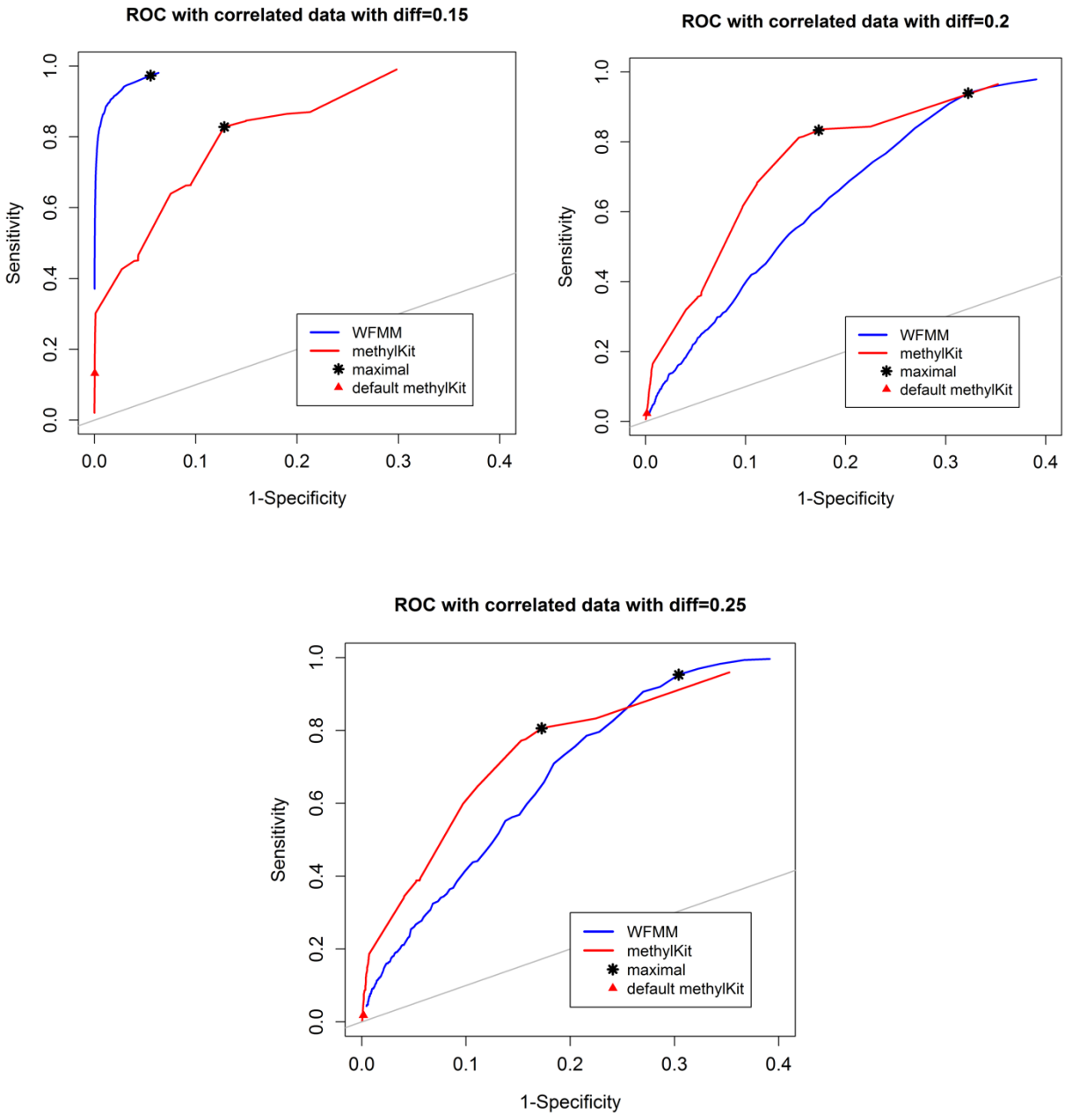


Figure 3.3: ROC curve comparision in ROC curve comparison between WFMM (blue curve) and methylKit (red curve) as differentially methylated cutoff increases from 0.1, 0.12, 0.15, 0.2 and 0.25.

(2) Effect of sample sizes

Overall, when sample sizes increase from 4, 10, 20, 30, to 40, WFMM performance remains robust (Figure 3.4). There is a moderate improvement in sensitivity and specificity when sample size increases from 4 to 10. There is only slight improvement in sensitivity and specificity in sample sizes of 10 or greater. In contrast, increase in sample sizes results in dramatic improvement in specificity in methylKit while sensitivity only improves slightly (Figure 3.4). Therefore, increase in sample sizes significantly improves methylKit's performance whereas only slightly for WFMM. It can be inferred that increased sample sizes give methylKit more power to detect small methylation differences across different phenotype groups whereas WFMM performance is more stable because the method incorporates methylation levels of nearby cytosines to make inference rather than solely relies on sample sizes.

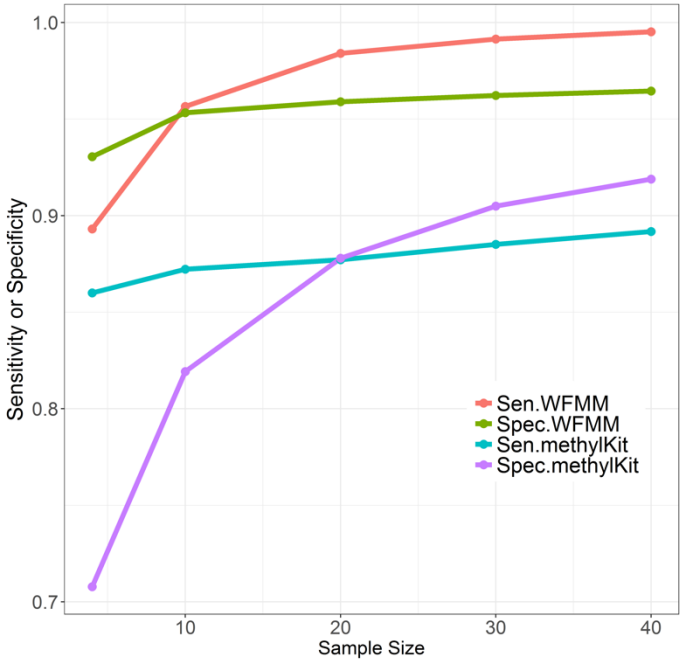


Figure 3.4: Effect of different sample sizes on WFMM with δ=0.01 and methylKit with adjusted setting (qvalue=1.00 and difference=4) performance on correlated simulated data when differentially methylated cutoff is 0.04.

3.4.2 Real data from herbicide glyphosate treatment of Arabidopsis thaliana

We applied WFMM and methylKit on dataset generated from our herbicide glyphosate treatment experiment on *A. thaliana*. WFMM was able to detect 557,664 DMCs (~17% of all cytosines in

A. thaliana genome) corresponding to 15,823 TAIR genes recognized from DAVID [63]. In contrast, methylKit detected only 48,041 DMCs (~1.43% of all cytosines in *A. thaliana* genome)

corresponding to 12,166 TAIR genes with default settings (qvalue=0.01 and difference=25), and 1,338,219 DMCs (~40% of all cytosines in *A. thaliana* genome) corresponding to 30,947 TAIR genes with adjusted settings (qvalue=1.00 and difference=4). Table 3.1 shows the breakdown of the number of significant DMCs and TAIR genes for each chromosome in *A. thaliana* genome. Chromosomes 1 and 5 have the most number of genes responding to herbicide glyphosate stress. Analysis of overlapping DMCs between WFMM and methylKit (Figure 3.5) shows that there are 33.6% and 21.7% common DMCs detected by both WFMM and methylKit in simulated and real dataset respectively. The similarity in proportions of common DMCs detected by both methods and of DMCs detected by only one of the two methods shows that simulation is reflective of real data.

Table 3.1: Number of significant DMCs, genes recognized by DAVID by applying WFMM with δ=0.01 and methylKit with default setting (difference=25, qvalue=0.01) and methylKit with adjusted setting (difference=4, qvalue=1.00) on real *A. thaliana* dataset.

Chromosome	WFMM δ=0.01, Number of DMCs	methylKit default, qvalue=0.01, difference=25, Number of DMCs	methylKit qvalue=1.00, difference=4, Number of DMCs	WFMM δ=0.01, Number of significant genes	MethylKit default, qvalue=0.01, difference=25, Number of significant genes	MethylKit qvalue=1.00, difference=4, Number of significant genes
Chr1	133,512	12,048	294,153	4,041	3,098	7,760
Chr2	87,488	7,627	244,683	2,417	1,887	5,129
Chr3	113,229	9,863	274,382	3,180	2,459	6,254
Chr4	91,327	7,708	227,539	2,563	1,943	4,815
Chr5	123,027	10,776	290,090	3,622	2,779	6,989
ChrC	9,081	19	7306	0	0	0
ChrM	0	0	66	0	0	0
Total	557,664	48,041	1,338,219	15,823	12,166	30,947

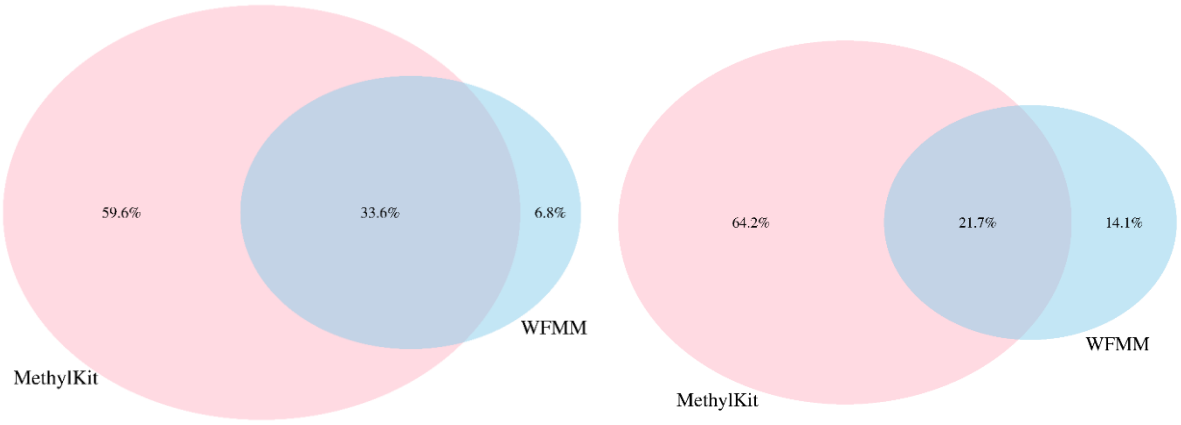


Figure 3.5: Percentages of overlapping DMCs from methylKit with adjusted settings (difference=4, qvalue=1.00) and WFMM with δ=0.01 in correlated simulated data when differentially methylated cutoff is 0.04 (left panel) and in real data (right panel).

Functional annotation results of significant genes detected by WFMM and methylKit show similar results between both methods (Figure 3.6). The most significant gene ontology (GO) terms in WFMM are also in top 50 significant methylKit GO terms. Malay Das et. al [64] did similar experiment applying herbicide glyphosate to *A. thaliana* plants and identified 484 genes that might be responsive to glyphosate stress. Comparatively, methylKit with default settings identified 12,166 genes, 181 of which overlap with Malay Das et al., and with adjusted settings (difference=4, qvalue=1.00), identified 30,947 genes, 466 of which overlap with Malay Das et al.. WFMM with δ=0.01 identified 12,166 genes, 238 of which overlap with Malay Das et. al. (Table 3.2). Thus, WFMM is slightly better than methylKit with default settings by identifying slightly more relevant genes related to glyphosate responses. For a fair comparison, of 3000 top most significant genes, methylKit with default settings has 39 overlapped genes, methylKit with adjusted settings (difference=4, qvalue=1.00), 41 overlapped genes and WMFF with default setting δ=0.01, 51 overlapped genes which also identified by Malay Das et. al [64] (Table 3.2). Though there are minor differences in gene clusters from methylKit and WFMM δ=0.01, the GO analysis between two methods are very similar (Figure 3.6, Figure 3.7).

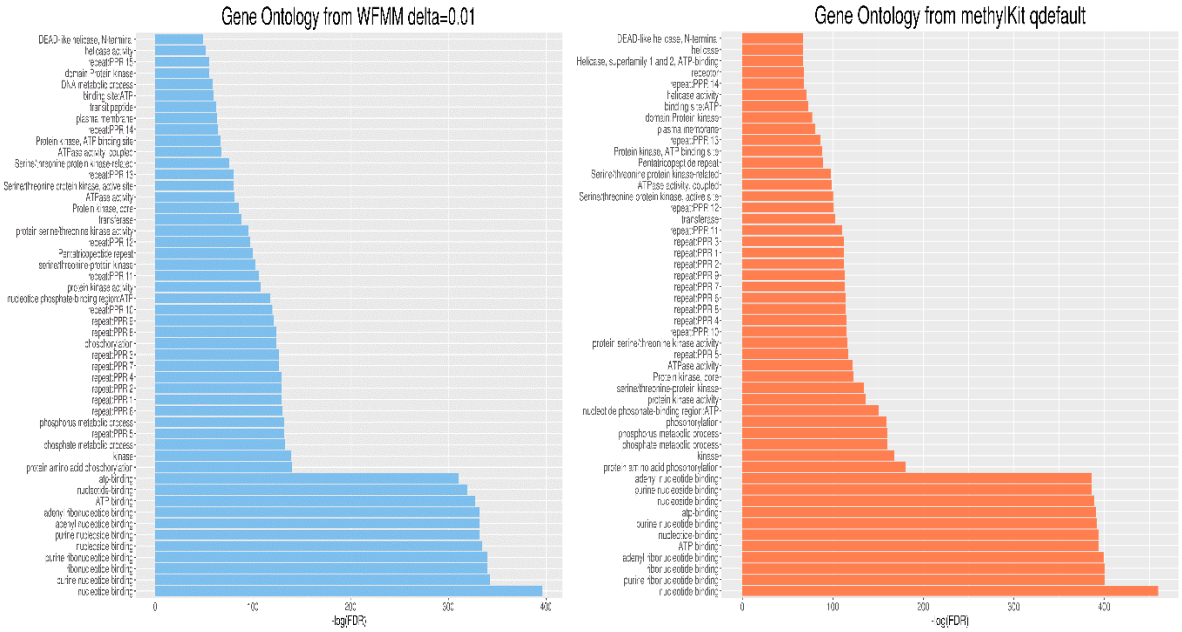
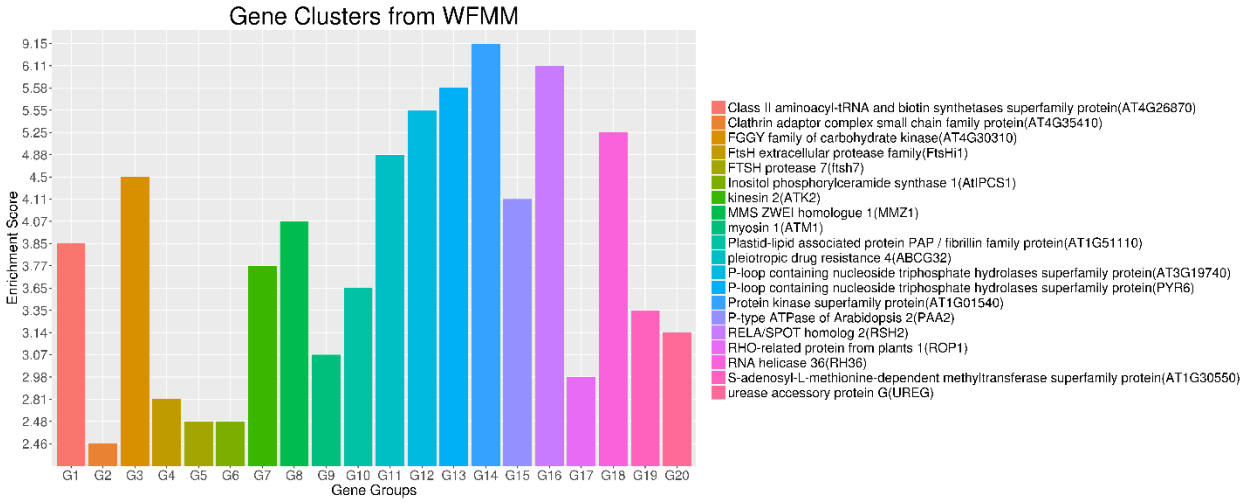


Figure 3.6: Gene Ontology for significant differentially methylated TAIR genes detected by WFMM with δ=0.01 (left panel) and methylKit with default settings (difference=25, qvalue=0.01) (right panel).



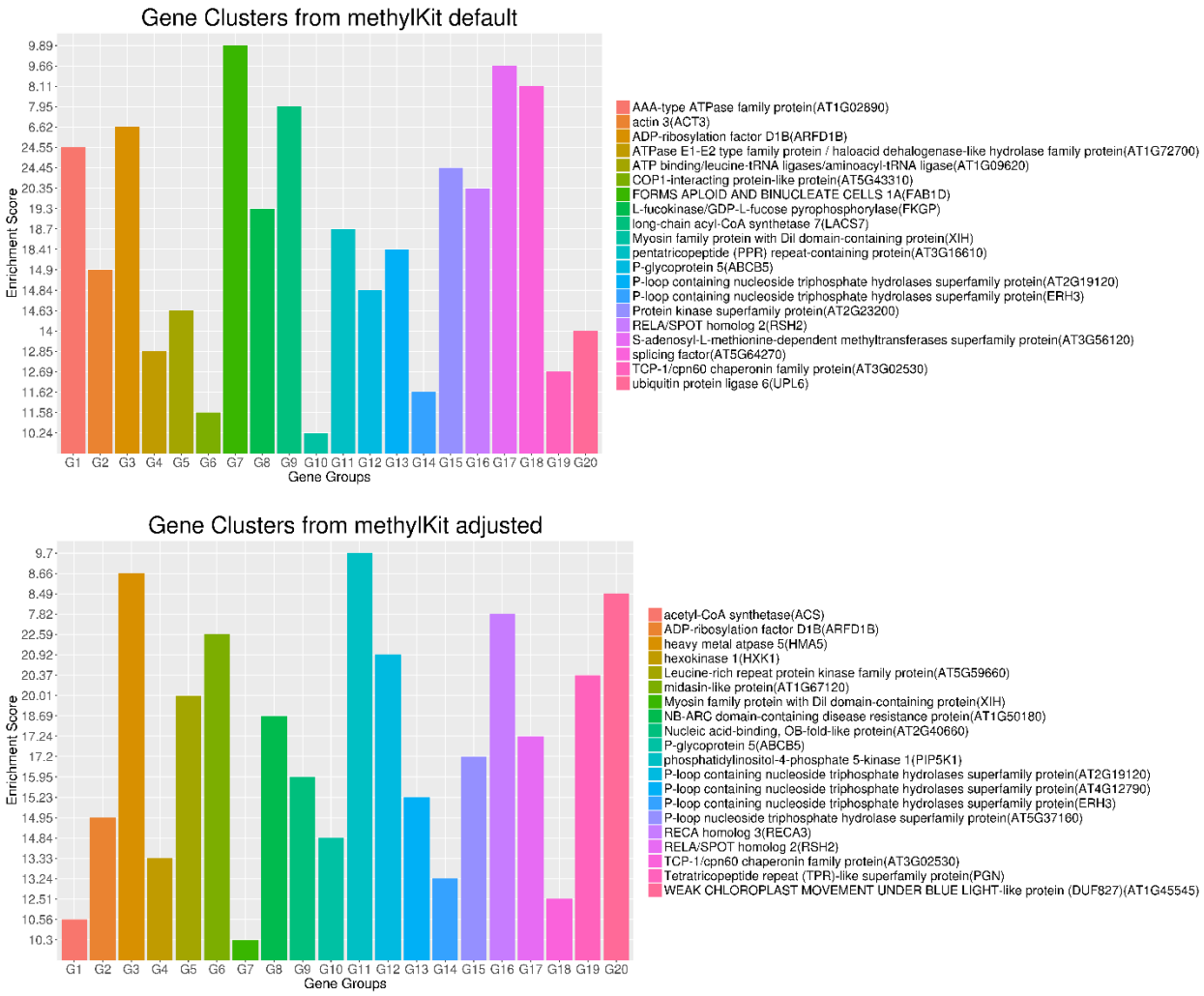


Figure 3.7: Gene Clusters of the top 3,000 most significant genes from WFMM with δ=0.01 (top panel), methylKit with default settings (difference=25, qvalue=0.01) (middle planel), and methylKit with adjusted settings (difference=4, qvalue=1.00) (bottom panel).

Table 3.2: Number of intersecting genes between 484 genes identified by Malay Das et al. [64] that are related to herbicide glyphosate stress and significant genes identified by WFMM and methylKit.

Methods	Number of significant genes	Number of shared genes in all significant genes	Number of shared genes in top 3000 most significant genes
WFMM δ=0.01	15,823	238	51
methylKit default, qvalue=0.01, difference=25	12,166	181	39

methylKit adjusted, qvalue=1.00, difference=4	30,947	466	44

3.4.3 Real data from monozygotic twin data with different pain sensitivity scores

We used the methylation profiles generated from MeDIP-sequencing data of 25 MZ twin pairs (50 MZ twins) who were discordant for heat pain sensitivity for model comparison. Heat pain tolerance between twins was determined experimentally using quantitative sensory testing. Datasets were downloaded from [65] with sample IDs from GSM1278649 to GSM1278698. The methylation levels in these datasets were summarized by combining cytosine regions rather than single cytosine sites. In total, there are 5,735,431 DMRs in these datasets. We assigned MZ twins in each of 25 MZ pairs to two groups (high vs low pain temperatures) according to MZ twins' pain sensitivity temperatures. For example, for a MZ twin pair from family ID 1, MZ twin 1 and MZ twin 2 have pain sensitivity temperature of 44.7 and 47.8 respectively. Therefore, we assigned MZ twin 1 to the low pain sensitivity temperature group and MZ twin 2 to the high pain sensitivity temperature group. We then applied WFMM and methylKit to the 50 MZ twins' methylation profiles with high vs. low pain sensitivity temperatures as phenotype groups. There were no significant DMRs detected by both WFMM with δ=0.01 and methylKit default settings or methylKit adjusted settings (difference=0.04, qvalue=1.00). This can be explained that the mean methylation differences between high vs. low pain temperature groups are very small (~4.1% of all mean methylation differences across DMRs < 10^{-5}) (Figure 3.3S). Therefore we adjusted parameter settings in both WFMM with δ=3.44x10^{-5} and methylKit (difference=4.34x10^{-5}, qvalue=1.00). These parameter settings from both methods were determined by empirical function applied on the real twin data and further described in the discussion section. For the 769 significant DMRs detected by WFMM with δ=3.44x10^{-5}, there were 236 genes recognized by the gene function enrichment program DAVID (Table 3.3). These genes were clustered into 5 groups by DAVID (Figure 3.8 left panel). For 2,023 significant DMRs from MethylKit (difference=4.34x10^{-5}, qvalue=1.00), there were 892 genes recognized by DAVID (Table 3.3) that were clustered into 32 clusters (Figure 3.8 right panel). The most important gene groups were ranked by enrichment scores (EASE scores). EASE scores are calculated from geometric mean of all enrichment P-values for each annotation term of all gene members in a gene group [66]. Two gene clusters that

have the highest EASE scores from significant differentially methylated genes detected by WFMM contain myelin transcription factor 1 like (MYT1L, enrichment score=1.19) and transient receptor potential cation channel subfamily C member 1(TRPC7, enrichment score=0.90). MYT1L functions in the developing mammalian central nervous system. TRPC7 was identified by Bell et.al [65] responsive to heat pain sensitivity. In comparison, methylKit was not able to capture relevant gene clusters pertaining to pain sensitivity in its first top 17 clusters. In the 18^{th} cluster, two genes (out of the 112 genes in this cluster) ST6GALNAC1 and TRPC7 were also found involved in heat pain sensitivity by Bell et al. [65]. It is remarkable that WFMM was able to capture the significant gene groups related to pain sensitivity using only the 25 MZ twin pairs' methylation profiles whose methylation differences are very small whereas Bell et al. [65] had to use the methylation profiles of 25 MZ twin pairs together with 50 unrelated individuals in a meta-analysis to capture the genes responsible for heat pain sensitivity.

Table 3.3: Number of significant DMCs, genes recognized by DAVID by applying WFMM with δ=3.44x10^{-5} and difference=4.34x10^{-5}, qvalue=1.00 on 25 monozygotic twin pairs with different pain sensitivity temperature.

Methods	Number of significant DMRs	Number of significant genes using DAVID
WFMM δ=3.44x10^{-5}	769	236
methylKit adjusted, qvalue=1.00, difference=4.34 x10^{-5}	2023	892

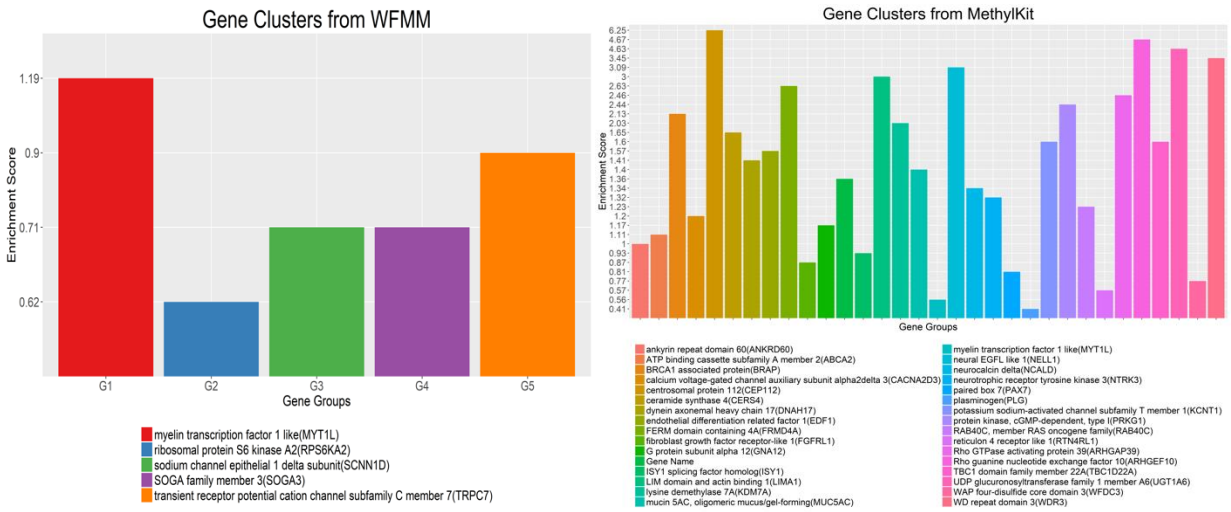


Figure 3.8: Gene clusters of significant genes detected by WFMM with δ=3.44x10^{-5} (left panel) and methylKit (difference=4.34x10^{-5}, qvalue=1.00) (right panel).

3.5 Discussion

Though there are numerous statistical methods for detecting differentially methylated cytosine, small sample sizes and small methylation differences in methylation data across phenotype groups remain a challenge for the commonly used methods [56]. Our analysis of the datasets demonstrated that the wavelet-based functional mixed model has several advantages. Firstly, the method is flexible and can be applied to different experimental designs and does not depend on coverage depth. Secondly, simulation results show that the WFMM method is robust even when sample sizes are small. Thirdly, the method is particularly effective for cases where methylation differences across phenotype groups are relatively small, for example, as demonstrated in our MZ twin pair analysis, the method is able to capture significant regions that are relevant to the phenotype of interest. Fourthly, if there is strong methylation correlation in the data, the method is able to take it into account for the inference, thus having more power in calling of DMCs/DMRs, as illustrated in the *A. thaliana* data and MZ twin data analysis. Finally, default settings of DMR analysis tools might not be the most suitable for some methylation profiles as shown in the *Arabidopsis* and twin datasets. We recommend some empirical rules to adjust the default settings so that the method can be better adapted to different methylation profiles of real datasets. For methylKit, we suggest to set the "diff" parameter to be at the 100(1-E)th quantile of the absolute pairwise methylation level differences between two phenotype groups across the whole genome,

where E is an expected percentage of methylation differences across all cytosines for a particular dataset based on prior knowledge. For example, in our *Arabidopsis* data, we expect ~10% (E=10%) of cytosines to be DMCs. Therefore, we set diff = 0.04 (corresponding to the 90^{th} quantile of the absolute pairwise methylation level differences between phenotype categories). In the twin dataset, we expect E=0.3%, therefore, we adjust diff in methylKit to $4.34x10^{-5}$ (i.e., the 99.7^{th} quantile of the absolute pairwise methylation level differences across whole human genome). In methylKit, the qvalue parameter should also be adjusted accordingly. If diff is very small (<0.1), set qvalue =1.00 to collect all significant DMRs. Similarly, we can adapt WFMM to be more tailored to different methylation profiles by controlling the δ parameter, setting δ to be the difference between the $100(1\text{-}E)^{th}$ quantile of the absolute pairwise methylation differences between two phenotype groups across the whole genome and the standard deviation of the methylation differences. For example, in our *A. thaliana* dataset, the 90^{th} quantile of the absolute pairwise methylation level differences between dosage categories is 0.04 and the standard deviation of pairwise methylation level differences between phenotype categories is 0.03, therefore, $\delta = 0.04 - 0.03 = 0.01$. In the twin dataset, the corresponding 99.7^{th} quantile and standard deviation are $4.34x10^{-5}$ and $9.2x10^{-6}$, respectively, therefore, we use $\delta = 4.34x10^{-5} - 9.2x10^{-6} = 3.44x10^{-5}$. In this way, a better DMC detection result can be achieved based on different methylation datasets.

Supporting Information

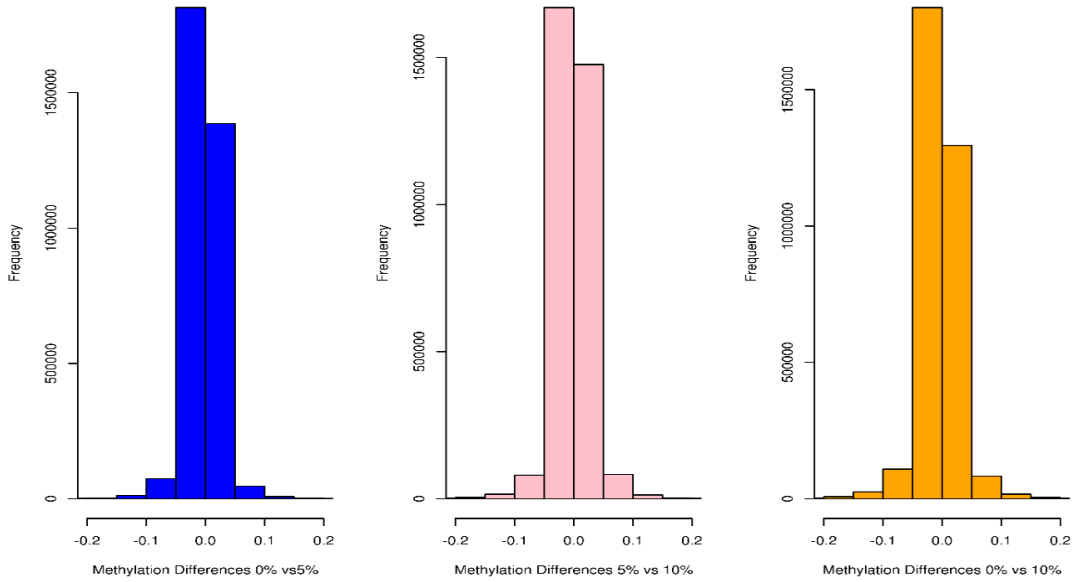


Figure 3.1S: Pairwise mean methylation Difference Profile of 12 *A. Thaliana* plants after glyphosate treatment

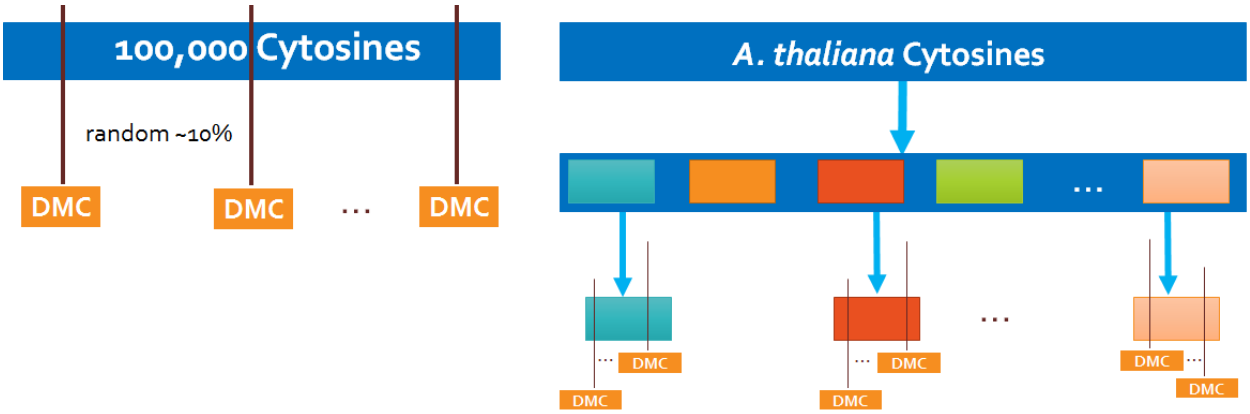


Figure 3.2S: Methylation level simulation at cytosine sites. Uncorrelated methylated cytosine simulated data (left panel) and correlated methylated cytosine simulated data (right panel)

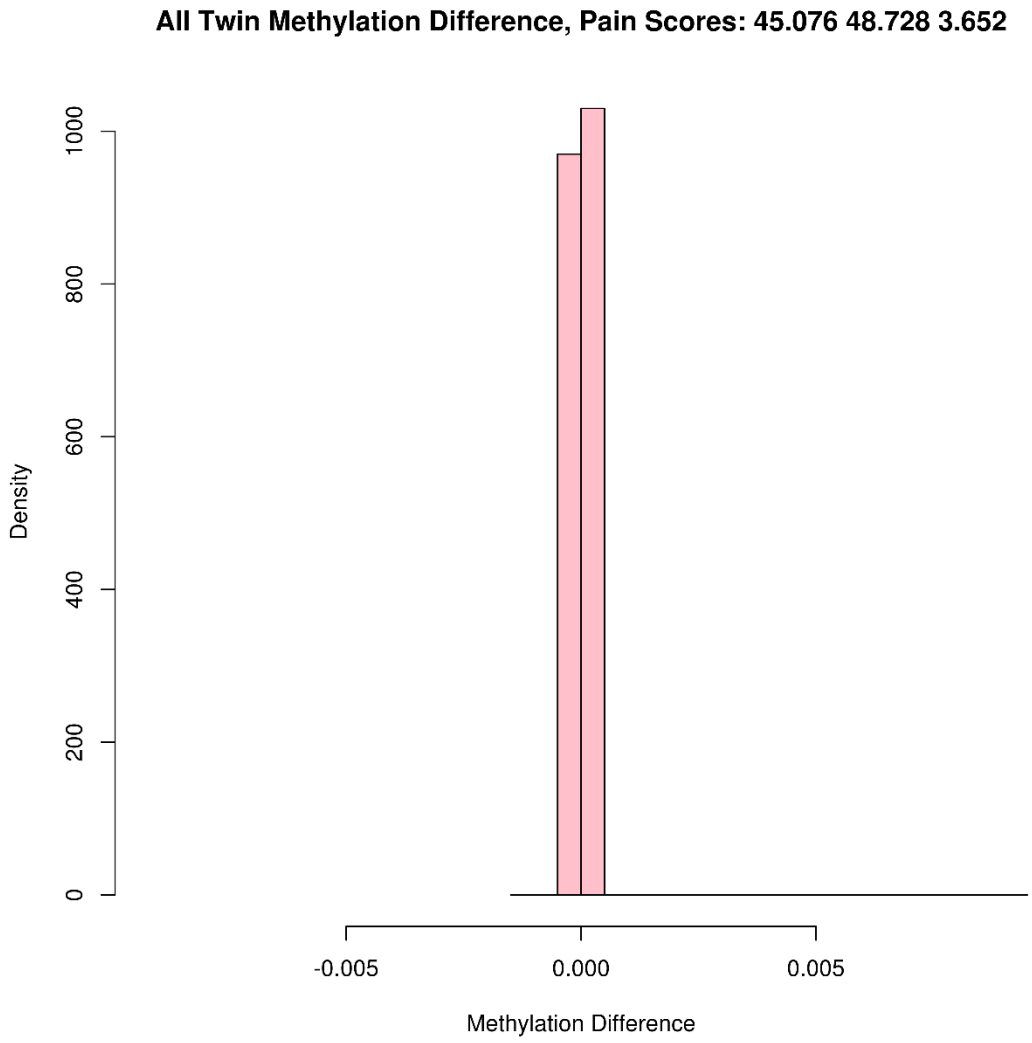


Figure 3.3S: Mean methylation profiles between higher and lower pain temperature group in 25 MZ twin pairs

Table 3.1S Number of significant DMCs, genes recognized by Ensemble by applying WFMM $\delta=4x10^{-5}$ and qvalue=1.01, difference=0.07 on 25 monozygotic twin pairs with different pain sensitivity temperature for each chromosome.

Chrom	WFMM δ=4x10^{-5}, Number of DMRs	MethylKit, qvalue=1.01, difference=0.07, Number of DMRs	WFMM δ=4x10^{-5}, Number of significant genes from Ensemble	MethylKit, qvalue=1.01, difference=0.07, Number of significant genes from Ensemble
Chr1	53	59	21	35
Chr2	23	28	9	23
Chr3	3	3	1	2
Chr4	25	17	10	9
Chr5	10	16	3	8
Chr6	40	21	11	8
Chr7	31	25	19	15
Chr8	36	33	11	12
Chr9	22	21	5	7
Chr10	50	40	11	9
Chr11	20	20	9	11
Chr12	0	15	0	9
Chr13	0	6	0	2
Chr14	7	13	4	4
Chr15	8	11	1	3
Chr16	78	54	21	25
Chr17	27	24	10	13
Chr18	11	15	5	7
Chr19	12	45	4	21
Chr20	10	30	5	11
Chr21	9	20	4	10
Chr22	19	30	3	9
Total	494	546	167	253

Chapter 4

Identification of factors contributing to microbiome regrowth in Simulated Reclaimed Water Distribution Systems

4.1 Introduction

Population growth and climate change leave billions of people around the world living in water scarcity conditions. Therefore, utility of reclaimed water (highly treated wastewater) plays an essential role in water sustainability [67]. Recently, researchers discovered microbial regrowth problems in potable water distribution systems (PWDs) [68]. In particular, microbial populations including opportunistic pathogens are observed to regrow in PWDs. Studies have shown that though many microbes in PWDs are benign, some microbes, including opportunist pathogens such as *Legionella pneumophila, Acanthamoeba polyphaga*, *Mycobacterium avium, Naegleria fowleri* and *Pseudomonas aeruginosa* can be a public health threat, especially for immunocompromised population [69]. Reclaimed water distribution systems (RWDs) share some similar characteristics of PWDs, thus we hypothesize that RWDs would encounter the same issues as PWDs. Previously, researchers have shown that it is impossible to remove all microbes from PWDs. Rather, we can only shift the microbial community to more favorable for humans [70]. This can be done through controlling various fundamental factors pertaining to the water system, for example, disinfect

types, limiting nutrients (N, C and P), water age, dissolved oxygen, temperature or pH in PWDs. Our knowledge of PWDs can help understand RWDs. For example, opportunistic pathogens are naturally occurring in PWDs and thrive under certain conditions. Biofilm is likely where opportunistic pathogens live and bulk water is likely where they spread and come into contact with humans. Several other fundamental factors can affect regrowth of opportunistic pathogens in PWDs, such as nutrients, water age, and disinfectants. We should take these fundamental factors into consideration to shape a healthy microbiome in RWDs. However, these factors might not have the same impact on RWDs. To illustrate, in PWDs, assimilable organic carbon (AOC) with concentration of 10-20 ug/L is reported to limit regrowth of opportunistic pathogens. Chloramine or chlorine, in some cases, is reported to control *Legionella spp* [68]. However, AOC has much higher concentrations in RWDs than PWDs [68], therefore it might no longer control opportunistic pathogen regrowth in RWDs. Moreover, no impact of chloramine or chlorine on control of *Legionella spp.* is found in RWDs [70].

In addition, RWDs is treated-waste water, there are concerns that go beyond the existing problems in PWDs. RWDs might have added new and/or worsen the existing problems, contributing to potential public health threats. There are two main reasons for these rising concerns. First, RWDs' microbial compositions (e.g., viruses, bacteria, and archaea) are mostly uncharacterized. Second and more importantly, RWDs contain more antibiotic resistant bacteria (ARBs), and antibiotic resistance genes (ARGs). During wastewater treatment, both residual antibiotics and ARBs are injected into wastewater, and certain conditions are imposed and further the spread of antibiotics resistance [71]. For example, in highly concentrated bacterial areas during sludge treatment, sharing ARGs among bacteria is facilitated through horizontal gene transfer [72]. These bacteria can persist or even multiply through wastewater treatment [71], thus contributing to the spread of antibiotic resistance. In summary, there are three objectives in this chapter 1) evaluate effects of several factors on shaping microbial communities, 2) identify the interplay of water chemistry, water age and microbial regrowth, and 3) characterize co-occurrence of ARGs and mobile genetics elements (MGEs), i.e., plasmids in simulated RWDs.

This chapter is largely contributed by Ni (Joyce) Zhu who constructed, maintained as well as collected and analyzed water samples from the simulated reclaimed distribution systems. The author performed all the statistical analysis in this chapter.

4.2 Materials and Methods

Experimental Design: The reclaimed water was collected twice weekly in Blacksburg-Christiansburg area. The wastewater was treated with activated sludge followed by nitrification/denitrification. High AOC and low AOC source water were collected via aerobic biological filtration as describe by Wang et al. [73]. After chlorinated to remove ammonia, the water was subject to no secondary disinfection, chloramination, or chlorination at varying concentrations. All treated water was kept in a constant temperature room at 14°C, 22°C, 30°C, 22°C and 14°C for a period of two months at each temperature. Bulk water and biofilm water were collected at two different water ages at the end of each two months. Joyce Zhu designed a network of simulated RWD to investigate how disinfection behaves differently under different conditions. Six 4-inch in diameter PVC pipe connected by narrow 3/8-inch in diameter tubing provided a hybrid design that enabled examination of extended water ages and collection of biofilms under fast shear conditions. The conditions investigated were: 1) high and low organic carbon levels; 2) three disinfectant conditions (no residual, 4 mg/L of chlorine, 4 mg/L of chloramine); and 3) four water ages (0, 1, 2.5 and 5 days). A temperature cycle of from 14°C-22°C-30°C-22°C-14°C was implemented to simulate a seasonal effect (Figure 4.1).

Low AOC High AOC

Figure 4.1: Simulated reclaimed water distribution investigating different behavior of disinfection under varying water conditions. Photos taken by Joyce Zhu.

Collection of water chemistry: total cell counts (counts/uL) were collected by flow cytometer technology. Disinfectant concentrations (mg/L), dissolved oxygen (mg/L), and nitrite concentrations (mg/L) were also measured.

Quantification of microbiome profiles: A whole sample metagenomics approach was applied to all water samples collected at the end of the experiment to generate relative abundances and diversity of ARGs in RWDs. Metagenomic DNA library was prepared using SwiftBio amplification. The samples then were subject to deep sequencing using an Illumina HiSeq 2500 at the Biocomplexity Institute of Virginia Tech facility. Data processing and normalization were done using Metastorm pipeline [74]. Samples were then submitted to deepARG [75] and ACLAME [76] for assembly analysis for ARGs and MGEs. The GreenGenes [77] database was used for phylogenetic and functional analysis for the bacterial community. In the end, there are 36 water samples for analysis.

Statistical Analysis: First, outlier detection methods are conducted to check if the data is consistent among replicates. Then, diversity analysis and multivariate analysis are conducted to identify significant population shifts of ARBs and ARGs under varying conditions. Specifically, the nonparametric method, analysis of similarity (ANOSIM) is applied along with nonmetric multidimensional scaling (NMDS) plots to determine which factors significantly affect ARG

populations. Negative binomial mixed models are also applied to examine the effect of water chemistry and water age on bacterial regrowth with water chemistry and water age. Finally, network analyses are conducted to find communities and co-occurrence patterns among ARGs and plasmids. We first construct a network based on Spearman's correlation for any ARG pairs or ARG-plasmid pairs. Edges are created between an ARG pair or an ARG-plasmid pair if their Spearman's correlation ρ is greater than 0.8 and the adjusted p-value for multiple testing is less than 0.01. Bayesian networks are also constructed based on abundance profiles of 655 ARGs and 100 genuses and 100 species across all water samples using the max-min hill-climbing algorithm. Assembled data are also used to construct the ARG-ARG and ARG-plasmid co-occurrence networks as a way to validate the two network models.

4.3 Results

4.3.1 Consistency of the simulated RWDs

For water samples at 30^0C and treated with chlorine, three replicated measurements from the simulated RWDs for each sample were collected to make sure that the system runs consistently. If the system runs consistently, we expect to get similar measurements for each replicate in each ARG class in a given sample. We can simply check if there is any outlier in all three replicates and if replicates have any effect on ARG abundance. Data used for this analysis is the normalized count data for each ARG class. We performed Grubbs' outlier detection test for all three replicates for each ARG class. Grubbs' test gives p-values indicating if there is any extreme value among there replicates, for example, if p-value < 0.05, there is an outlier among there replicates. We also adjust p-values for multiple testing using the false discovery rate. If adjusted p-values>0.05, there would be no outlier, otherwise there would be an outlier among the three measurements. The ANOVA analysis on samples for each ARG class is performed to determine if replicates have any effect on ARG abundance. P-values from the ANOVA analysis are also adjusted for multiple testing. There are no suspected outliers since all adjusted p-values from Grubbs' test are greater than 0.05 (Table 4.1 left panel). Though there are small variations in some samples in the system, the replicate measurements are quite consistent within each ARG class. There is also no replicate effect on any of the ARG classes since there are no p-values and adjusted p-values<0.05 (Table 4.1 right panel). Finally, we apply ANOSIM on all replicated samples treated with chlorine at 30^0C. Replicate is

included as a factor in the model. Replicate has an R-value of -0.055 and p-value of 0.777. Therefore, replicate has no effect on ARG composition. From the above outlier detection, ANOVA, and ANOSIM results, we can conclude that the system is consistent.

Table 4.1: p-values and p-values adjusted for multiple tests from Grubb's outlier test (right panel). p-values for replicate effect from the ANOVA test (right panel)

ARG class	Sample IDs	p-values	p-values adjusted
rifampin	30.H.5.W.Cl2	0.003	0.324
tetracycline	30.H.5.W.Cl2	0.013	0.576
aminocoumarin	30.H.2.W.Cl2	0.016	0.576
bacitracin	30.H.5.W.Cl2	0.023	0.621
multidrug	30.H.2.W.Cl2	0.029	0.626

ARG class	p-values
glycopeptide	0.121
peptide	0.264
pleuromutilin	0.308
tetracycline	0.349
thiostrepton	0.398

4.3.2 Decay pattern of disinfectant types in RWDs

Overall, there is a decay pattern for both disinfectant types as temperature increases under both high and low AOC (Figure 4.2). Chlorine tends to decay faster than chloramines. Furthermore, under low AOC, decay occurs slower in both disinfectant as temperature increases compared to high AOC.

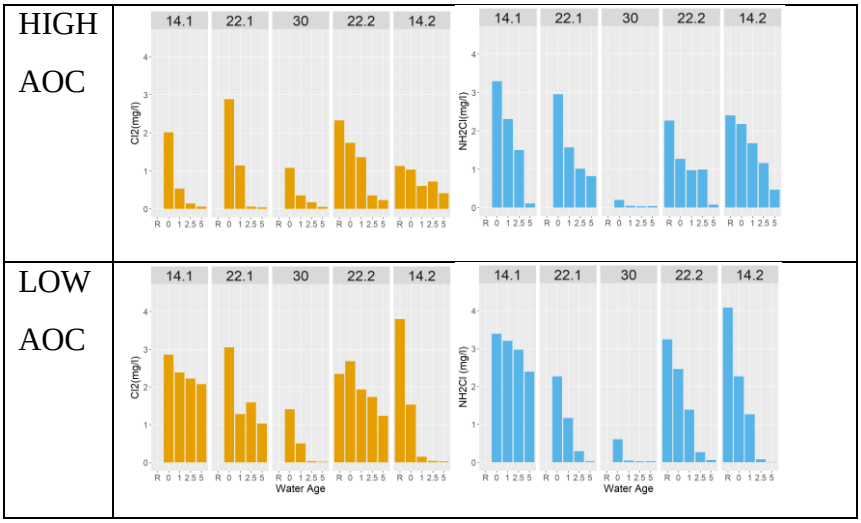


Figure 4.2: Decay pattern of chloramines and chlorine disinfectant in RWDs

4.3.3 Relationship of water chemistry, water age, and microbial regrowth

We also examine how water chemistry (AOC, disinfectant types, etc.) affects cell counts. We tried to fit the total cell counts with Poisson regression, Poisson mixed models, negative binomial, and negative binomial mixed models. We chose negative binomial mixed models since it has the lowest Akaike Information Criterion (AIC=3913.716). Pearson's chi-square test for goodness of fit gives p = 0.805, suggesting the negative mixed model fits the total cell counts data. By applying negative binomial mixed models on cell counts, water chemistry factors that have the most effect on bacterial regrowth are identified. Most regrowth occurs at the high temperature (30^{o}C). Chlorine condition has significantly negative effect on cell counts whereas chloramine does not (Table 4.2). AOC is marginally significant in determining microbial regrowth (p=0.067). Low AOC nutrients keep regrowth of resistome at a lower rate (80.2% lower compared to high AOC). The results shown in Table 4.2 are consistent with findings shown in Figure 4.3.

Table 4.2: The effect of water chemistry and water age on bacterial regrowth

Water Chemistry		Coefficient Estimate	P-value
Temperature		0.105	0.031*
AOC	Low	-0.802	0.067.
Disinfectant Types	Chlorine(Cl2)	-3.034	1.71e-08 ***
	Chloramine (NH2Cl)	0.035	0.948
Water Age	1	0.012	0.947
	2.5	-0.123	0.509
	5	0.030	0.869
	F	-0.109	0.608

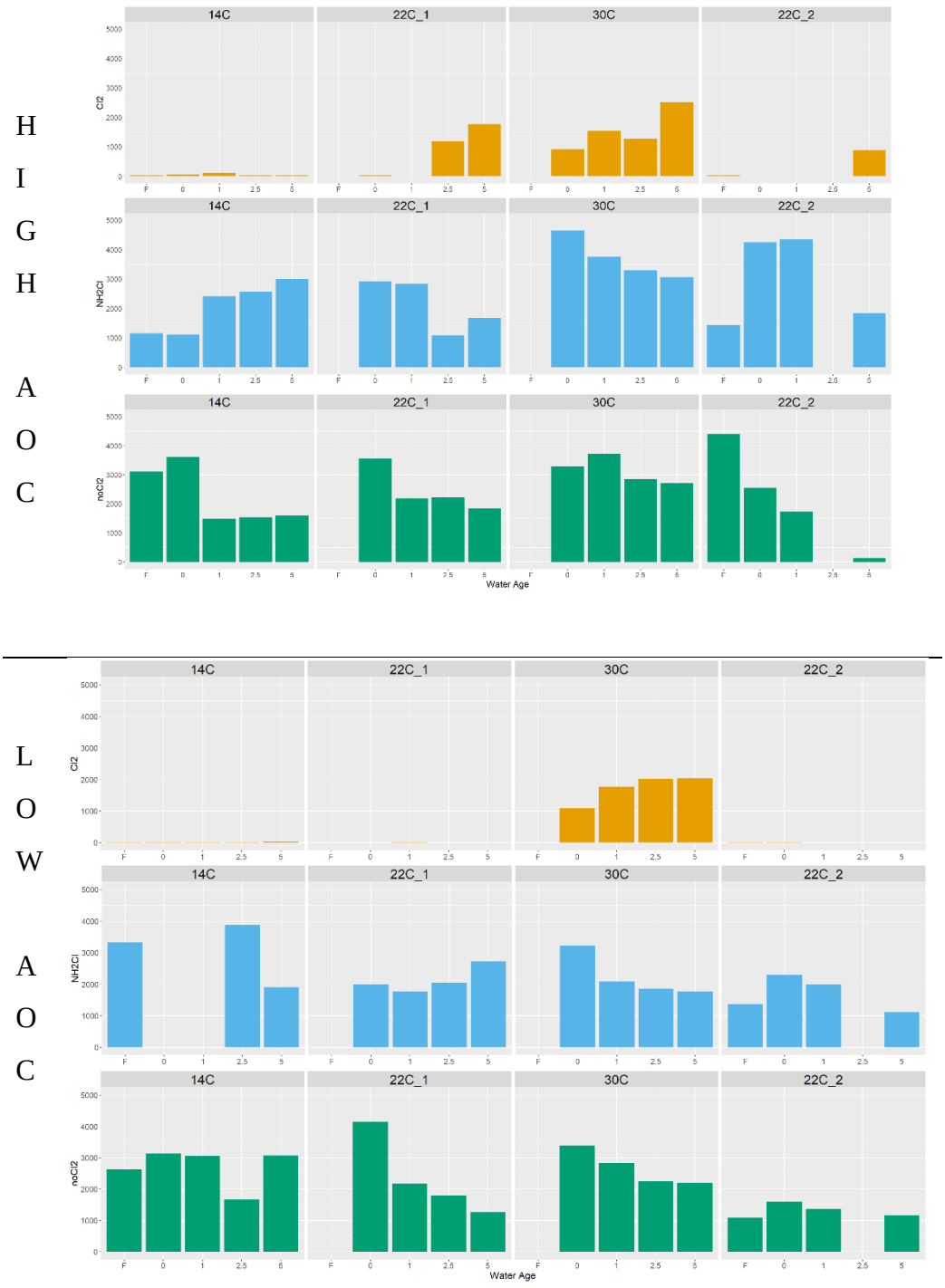


Figure 4.3: The effect of water chemistry and water on observed total cell counts

4.3.4 Factors that affect ARG profiles in the simulated RWDs

Based on Bray-Curtis dissimilarity on all water samples, the ARG compositions are significantly different for three condition comparisons, biofilm vs. bulk water (R=0.313, p-value=0.001),

different disinfectant types (R=0.228, p-value=0.003), and different temperatures (R=0.258, p-value=0.002). The ARG compositions are similar under two water age conditions (R=-0.013, p-value=0.607). At 30^0C, the ARG communities are even more distinctive under biofilm vs. bulk water (R=0.6, p-value=0.001), and different disinfectant types (R=0.374, p-value=0.003). There is still no difference in ARG compositions under two different water age conditions (R=-0.046, p-value=0.707). It is noteworthy that at 30^0C disinfectant types are influential in shaping the resistome especially in water, i.e., there are clear groupings of disinfectant types in bulk water (Figure 4.4 left). NMDS plot also reveals that at 30^0C and in bulk water, MSBA, TETA, MDTC, MUXC, LLMA in tetracycline and aminocoumarin categories are the most affected by chlorine (Cl2); BACA, MUXB, MDTB, MUXA (belong to bacitracin and aminocoumarin categories) by chloramines (NH2Cl); TET33, NJ69_08675, ILES2, VANRC, TETX, SUL2, and ARNA prominently in polymyxin, macrolide-lincosamide-streptogramin and chloramphenicol categories by no disinfectant. At 30^0C, biofilm environment has more diverse ARG compositions. Also both disinfectants reduce diversity in ARG communities. Chloramines keep ARGs the least diverse (Figure 4.4 right).

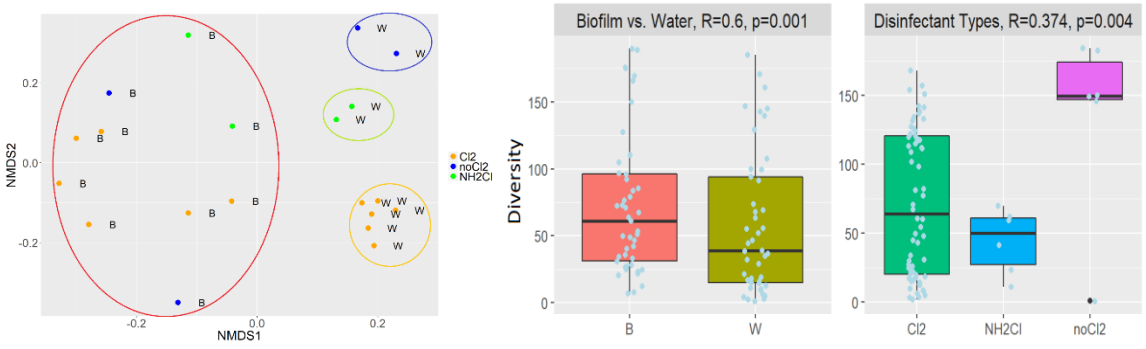


Figure 4.4: NMDS plot on biofilms (B) vs. bulk water (W) and disinfectant types at 30^0C (left) and ANOSIM plots on diversity under biofilms (B) vs bulk water (W) and under different disinfectant types at 30^0C (right)

In addition, Simpson diversity indices of ARG communities are calculated and compared across all samples. Mixed ANOVA on diversity indices is applied to determine which factors significantly affect ARG diversity. Results from the mixed ANOVA show that biofilms vs. bulk

water have significant effect on ARG diversity (p-value<0.05), whereas disinfectant types, temperature, and water age have no significant impact on ARG diversity. ARGs are more diverse in biofilms environment than in bulk water. At 30ºC, ARGs tend to be less diverse than at 22 ºC. Disinfectants tend to keep ARG communities less diverse. Chlorine tends to keep ARGs at least diversity (Figure 4.5). These univariate results are mostly consistent with ANOSIM multivariate results.

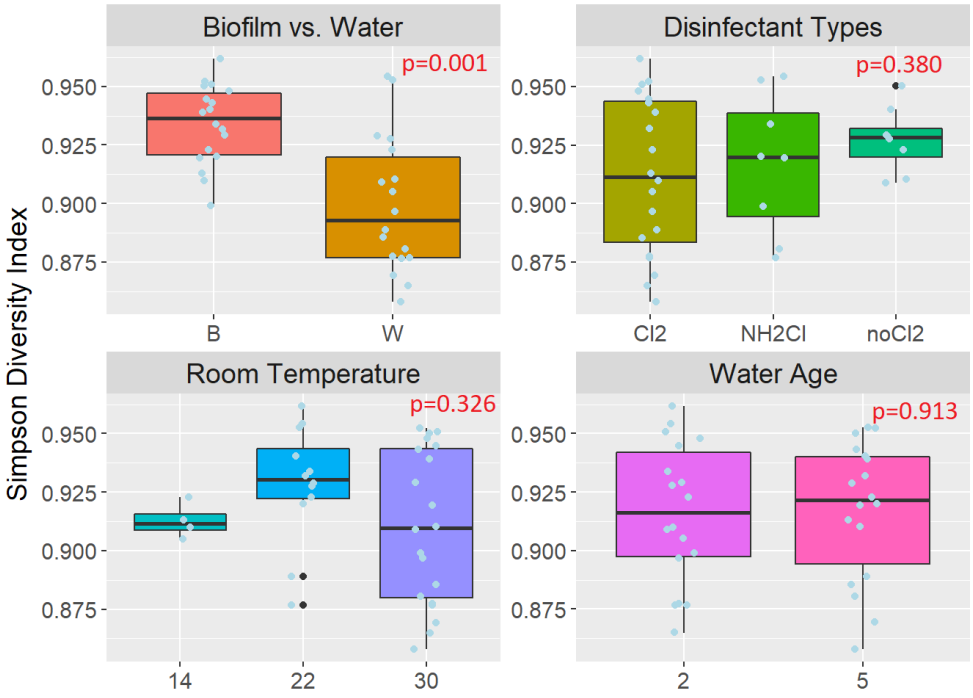
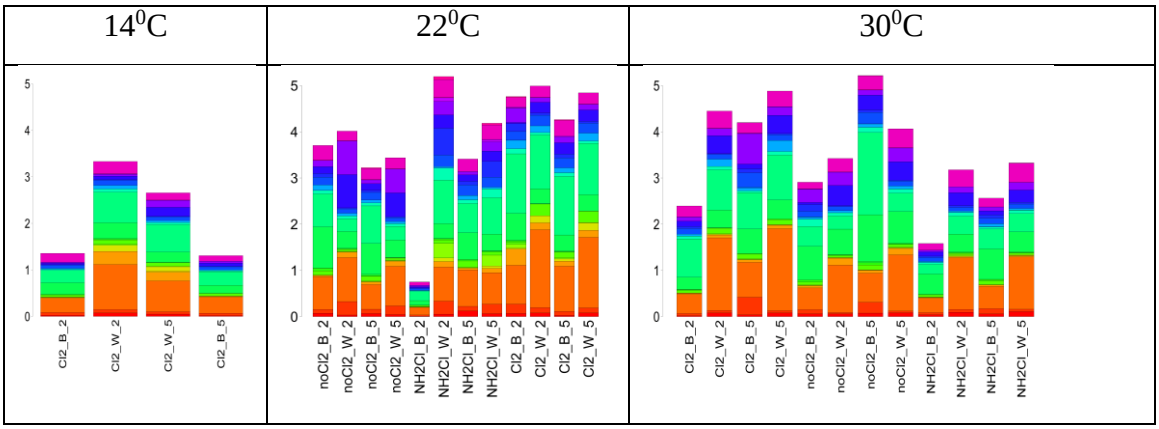


Figure 4.5: Simpson diversity plots across all samples

Wilcoxon Mann-Whitney tests are conducted on abundance to determine factors that may control the total ARG abundance across samples. Total ARG abundance is significantly smaller in biofilms compared to bulk water (average total abundance in biofilms and water are 53.713 and 75.645 respectively, with p-value=0.003) (Figure 4.6 upper) while temperature, disinfectant types and water age are not significant in controlling total ARG abundance, i.e., p-value>0.05. Average total ARG abundance tends to increase with temperature even though average total ARG abundance is more similar between 22^0C and 30^0C. Average total abundance tends to be the lowest

under chloramines followed by no disinfectant (24.187 and 29.986 respectively) and average total abundance tends to be the highest in chlorine (75.185) (Figure 4.6 upper).

ARG composition plots (Figure 4.6 lower) reveal that ARG compositions differ across disinfectant types. At 30^0C and under chloramines, ARG compositions are the least diverse followed by chlorine. These observations are consistent with ANOSIM analysis. In summary, the multivariate nonparametric method, ANOSIM, works the best for the data because it does not assume any particular distribution imposed for ARG abundance and it takes into consideration all ARG abundance profiles rather than collapses all ARG profiles into one diversity index per sample. ANOSIM reveals that water/biofilm and disinfectant types play a significant role in shaping the resistome. In particular, the selective effect of disinfectants is the most pronounced in the water phase where the bacteria are in the most direct contact with the disinfectants. Higher ARG diversity observed in biofilms suggests that biofilm tends to serve as a reservoir for ARG exchange and accumulation, even though lower ARGs abundance is observed. Disinfectants reduce ARG abundance as well as ARG diversity. Disinfectants are more in shaping the resistome of bulk water than that of biofilms. It is remarkable that bacitracin and multidrug are the most abundant in all water samples.



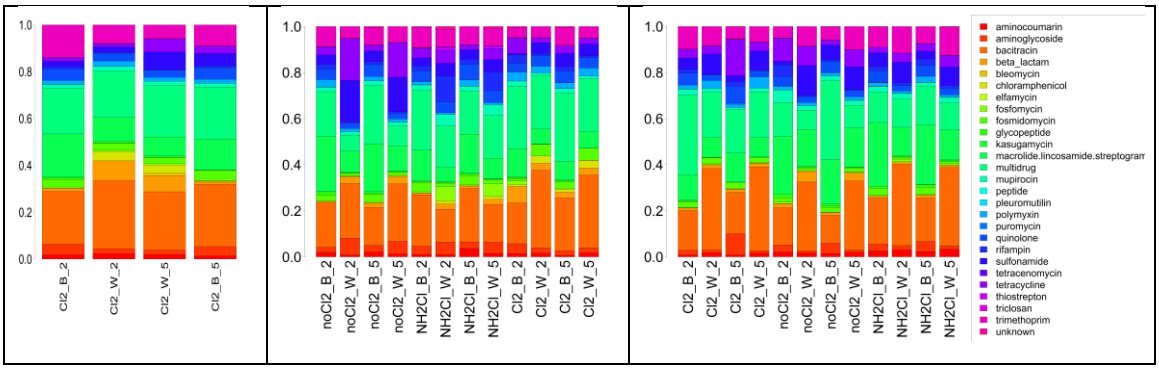


Figure 4.6: Absolute abundances (upper) and relative abundances (lower) of ARG classes in RWDs

4.3.5 Correlation analysis of water chemistry and ARG abundance

Spearman's correlations between ARG abundance and concentrations of dissolved oxygen and two disinfectant types across samples were calculated. We reported the most significant correlations (p-value<0.05) with the highest Spearman's correlation values. ARG abundance across samples are positively correlated to dissolved oxygen and chlorine concentrations (Table 4.3a, b). ARGs in multidrug category have the highest correlation with dissolved oxygen (Table 4.3a). It is likely that the antibiotic resistance mechanism is also shared as a defense mechanism to cope with lower dissolved oxygen. Most ARGs that are highly correlated with chlorine belong to the beta_lactam group (Table 4.3b). ARG abundances across samples are observed to have the strongest positive or negative correlation with chloramines.

Table 4.3a: Spearman correlation between ARG abundance and dissolved oxygen

ARGs	Spearman correlation	ARG class
ACRB	0.671	multidrug
MDTD	0.654	multidrug
ACRA	0.646	multidrug
OPRN	0.640	multidrug
MEPB	0.633	multidrug

Table 4.3b: Spearman correlation between ARG abundance and chlorine

ARG	Spearman correlation	ARG class
VANRI	0.656	glycopeptide
MSRB	0.654	MLS
ARR.2	0.654	rifampin
AB182_22595	0.633	beta_lactam

CATB9	0.600	chloramphenicol
EMRA	0.599	quinolone
MEXA	0.597	multidrug

OXA.129	0.633	beta_lactam
AADK	0.633	aminoglycoside
WM16_03815	0.633	beta_lactam
AADA7	0.601	aminoglycoside
DSX2_1119	0.600	beta_lactam
MEXP	0.593	multidrug

Table 4.3c: Spearman correlation between ARG abundance and chloramines

ARG	Spearman's correlation (ARGs vs. Chloramines)	ARG class
ADER	0.777	tetracycline
AAC-3'-IIB	0.742	aminoglycoside
ADEB	0.740	multidrug
MAB_2875	0.726	unknown
FLOR	0.721	multidrug
TETZ	0.719	tetracycline
AAV95_16190	0.714	beta_lactam
CCNA_03676	0.714	aminoglycoside
EFPA	0.708	multidrug
MACB	-0.857	macrolide-lincosamide-streptogramin
TURPA_2231	-0.785	macrolide-lincosamide-streptogramin
AMS22_10315	-0.757	macrolide-lincosamide-streptogramin
DFRA3	-0.748	trimethoprim
RAHAQ2_0060	-0.744	macrolide-lincosamide-streptogramin
ARNA	-0.730	polymyxin
MPHA	-0.722	macrolide-lincosamide-streptogramin
CCC_03326	-0.713	beta_lactam
OTRA	-0.713	tetracycline

The multivariate regression tree (MRT) analysis is first carried out on ARG profile for all samples and adjusted for all water chemistry parameters, dissolved oxygen, cell counts, nitrite

concentrations and disinfectant types to determine the tree splits. Of all the factors considered, only disinfectant types are significant in determining the tree splits of the ARG abundance. We further carry out the MRT analysis on ARG profile for samples from chlorine and chloramine separately to determine whether there exists critical concentration values for each of the disinfectant types that effectively shape the ARG profile. For chlorine samples, the tree split is at chlorine concentration of 0.09 mg/L. Samples with chlorine concentration of 0.09 mg/L or greater have higher average ARG abundance (5.61 compared to 1.09 in lower concentration samples). For chloramine samples, the tree split is at chloramine concentration of 0.43 mg/L. Samples with this concentration or greater have lower average ARG abundance (0.535 compared to 0.584 in samples with lower concentration) (Figure 4.7). It is noticeable that there is a bigger difference in average ARG abundance in the tree split in chlorine samples with even small concentration compared to chloramine samples (4.51 vs. 0.049). The multivariate regression tree analysis for chloramines gives counterintuitive results since we expect that higher disinfectant concentrations lower total ARG abundance. This is likely due to small sample sizes and further validation is needed.

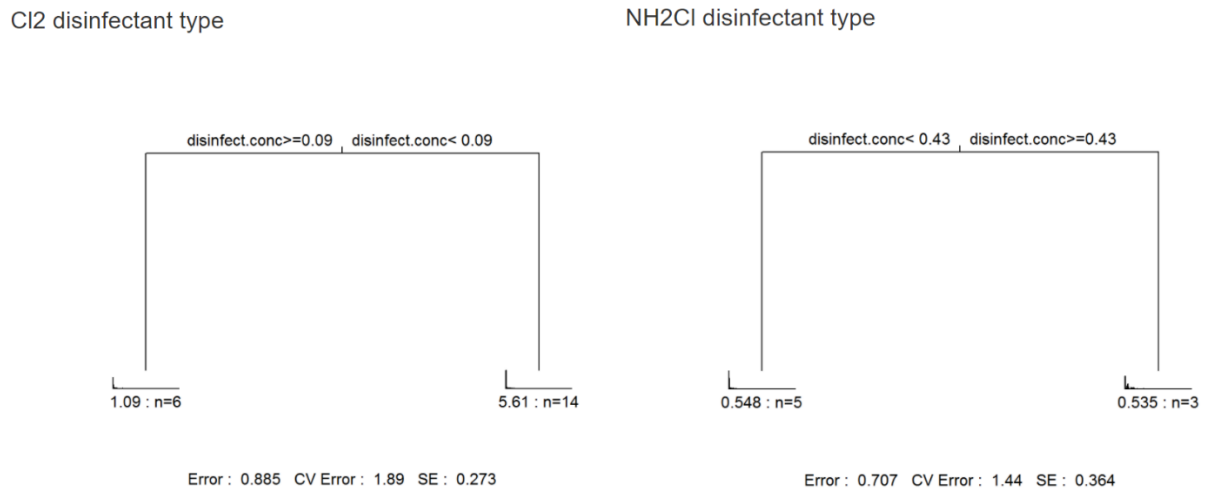


Figure 4.7: Multivariate Regression Tree on ARG profile

4.3.6 Network analysis on assembled data

To investigate co-occurrence of ARGs with mobile genetic elements and pathogens, we assembled the metagenomic short reads into contigs using the assembly pipeline from MetaStorm [74]. The assembled contigs were analyzed for plasmids using the ACLAME database [76]. The GreenGenes database [77] is used to for bacterial phylogenetic and functional analyses. The read matches were

filtered out if e-value > 10^{-10}, coverage<90%, or sequence identity < 60% to ensure high quality co-occurrence from the contigs. In total, there are 45,066 scaffolds, 557 (1.24%) of which contain one or more ARGs and 501 (1.11%) contain both ARGs and plasmids. Of all 577 scaffolds from deepARG [75], 92 (15.95%) scaffolds have ARG co-occurrence. UPPP (of bacitracin class) and DFRE (of trimethoprim class) are the most co-occuring across all samples (24 occurences), followed by UPPP and CEOB (multidrug) (13 co-occurences) (Figure 4.8). Other most significant connections include DFRE and BACA (8 co-occurences), DFRE and CEOB (7 co-occurences), UPPP and ACRB (7 co-occurences), and UPPP and MEXF (7 co-occurences). Most of the high frequency co-occuring ARGs are found to be general housekeeping genes that have no specific targeted mechanism towards individual antibiotics. Other commonly co-occurred ARGs include SUL1 and TET genes. Co-occurrence of these antibiotic-specific genes with broad-spectrum resistance genes could enhance the propagation of these genes.

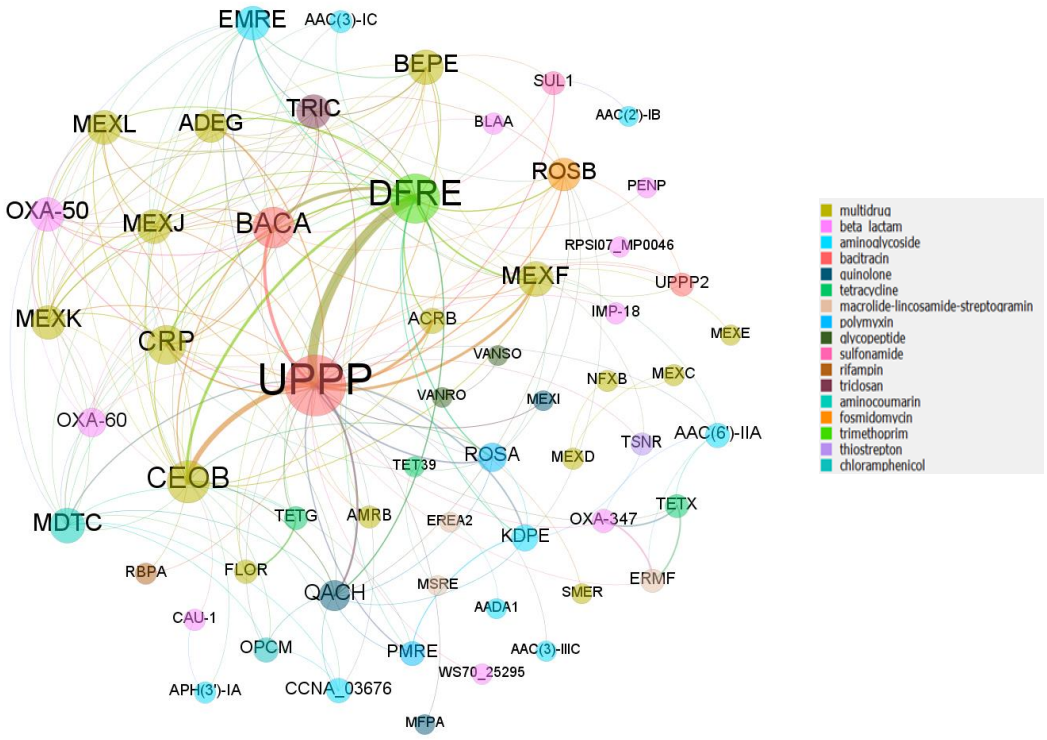


Figure 4.8: The ARG co-occurrence network constructed based on ARGs occurring on the same scaffolds using de novo assembly of metagenomic sequences. The sizes of ARG nodes correspond

to degrees of the nodes. The thickness of edges reflects the number of ARG connections occurring on the same scaffolds across all water samples.

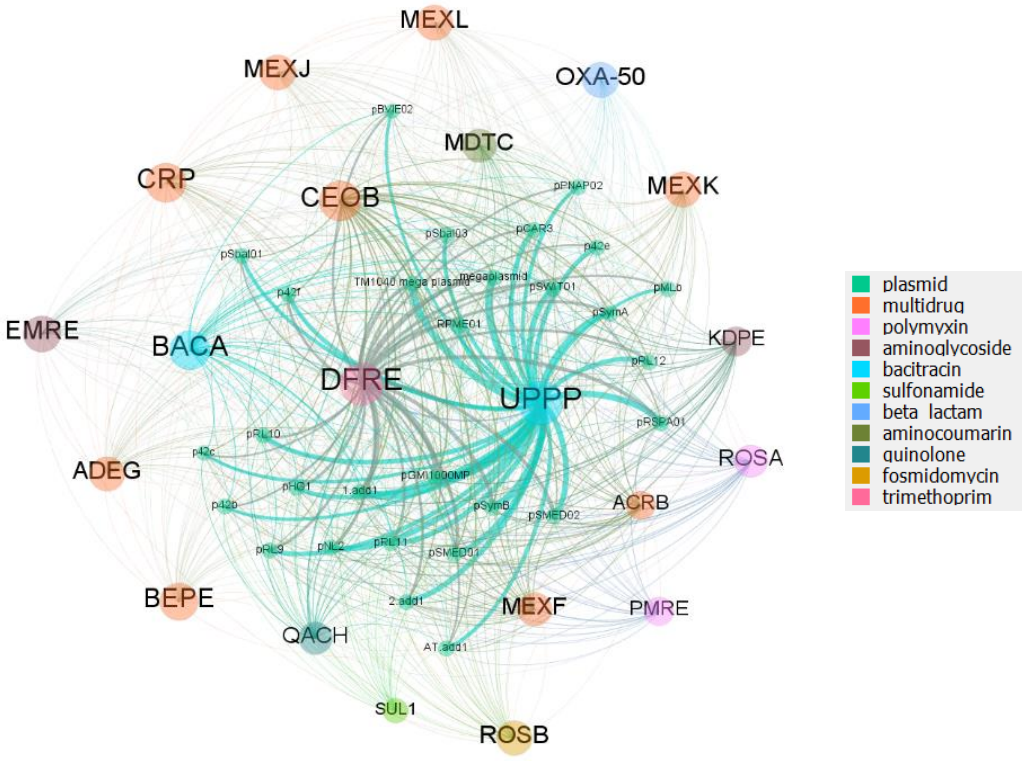


Figure 4.9: Network constructed based on ARGs and plasmids occurring on the same scaffolds using de novo assembly of shotgun metagenomic sequences. The size of the node corresponds to the degree of the node. The thickness of the edge reflects the number of ARG-plasmid co-occurrence on the same scaffolds across all water samples. For clarity, only the top 50 ARGs or plasmids with the highest connections with other ARGs or mobile genetic elements are shown.

Network analysis from the assembled data shows that UPPP, BACA (bacitracin), DFRE (trimethoprim), CEOB, MEXF, ACRB (multidrug), and ROSA (polymyxin) are ARGs most frequently found in plasmids while pGMI1000MP, 1, pSymB, megaplasmid, and pRSPA01 are plasmids that are most frequently associated with ARGs. Association of these ARGs with mobile genetic elements such as plasmids indicates the likelihood of these ARGs to spread to susceptible bacteria through horizontal gene transfer and conferring resistance.

4.3.7 Modeling co-occurrence of ARGs based on ARG abundance

Several networks are constructed on ARG abundance. One network is constructed based on the pairwise Spearman's correlations between 655 ARG subtypes across all water samples. Only ARG pairs with Spearman's correlation $\rho>0.8$ and the adjusted p-value for multiple tests <0.01 are retained in the network. In addition, a Bayesian network is constructed using the max-min hill climbing algorithm.

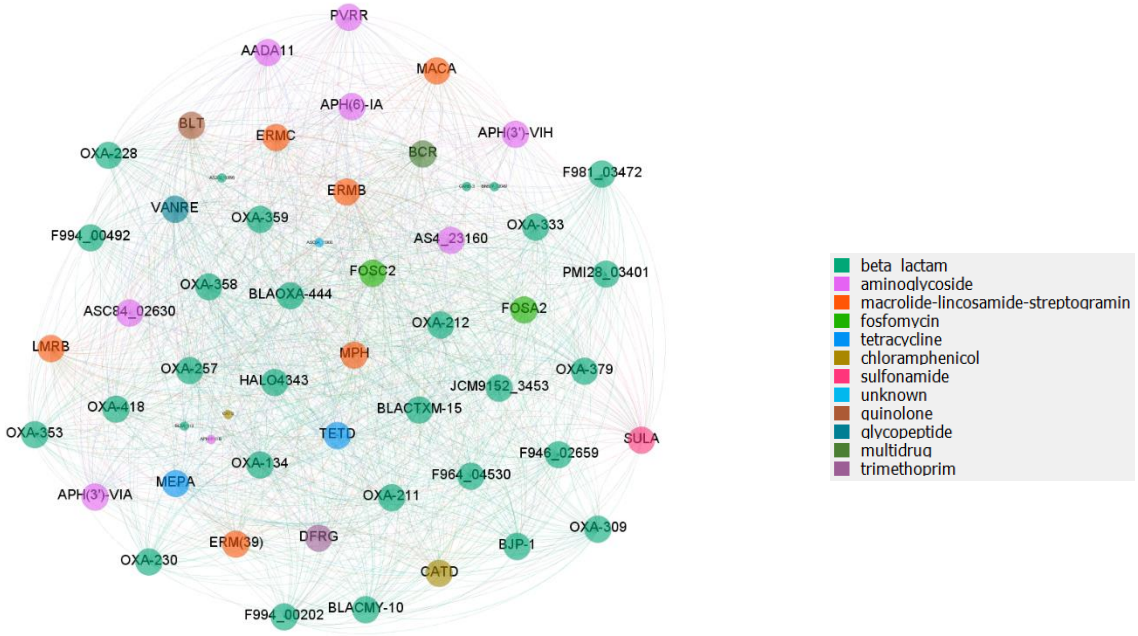


Figure 4.10: The network constructed on Spearman's correlations on ARG abundance with $\rho>0.8$ and the adjusted p-value for multiple tests <0.01. For clarity, only the top 58 ARGs with the highest connections with other ARGs are shown.

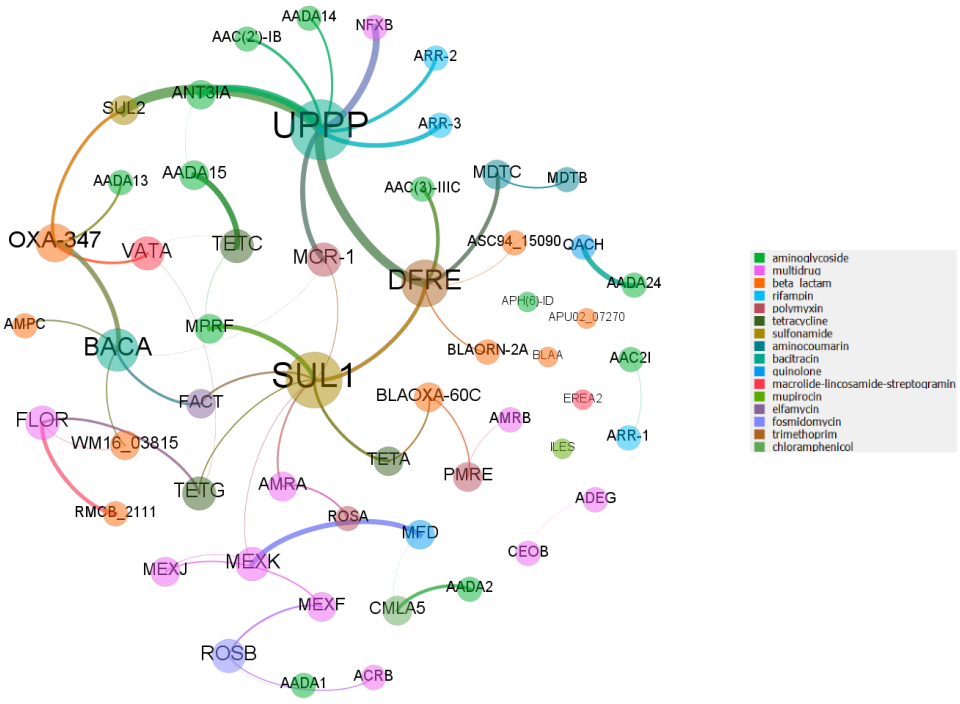


Figure 4.11: The Bayesian network constructed on ARG abundance using the max-min hill climbing algorithm. Only ARGs' edge connections with p>0.8 are kept in the network. In addition, the top 58 ARGs with highest connections with other ARGs are retained.

The Bayesian ARG network is more consistent with the network constructed from assembled data than Spearman's correlation network (Figure 4.10) as 43.10% of the most connected ARGs from the Bayesian network agree with the top ARG hubs from the network based on assembled data compared to 5% for the network based on Spearman's correlation). The Bayesian network also identifies UPPP and DFRE as the most connected ARGs (Figure 4.11)

4.3.8 Modeling co-occurrence of ARGs and microbial taxa based on abundance data

As the above analysis shows that the Bayesian network is a better fit for our data; subsequent networks were constructed based on the Bayesian network on abundance profiles of ARGs and microbial taxa. The similarity between ARGs and microbial taxa profiles across samples indicates co-occurrence of ARGs and microbial taxa. In addition, co-occurrence of ARGs and microbial taxa may provide ARG-host information, i.e., the kinds of microbial species that carry ARGs. For example, at the genus level, Sphingomonas stands out as having the most connections with

different types of ARGs. It is a host for glycopeptide resistant genes (VAND and vanRN) and beta lactam resistant genes (e.g., IMP-47, mrdA, blaTER-2, and WS70_25295) and aminoglycoside resistant genes (e.g., LHA_1702, AAC(3)-Id, and aadA15).

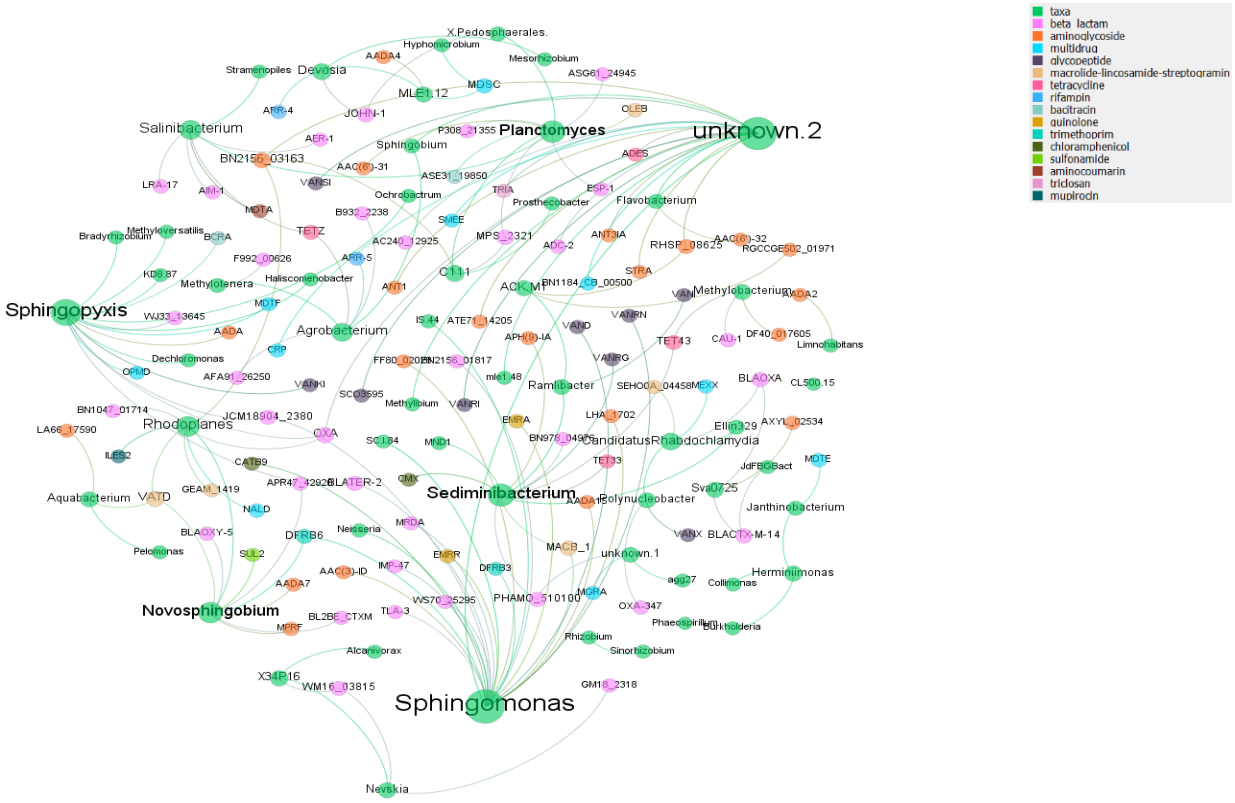


Figure 4.12: The Bayesian network of ARGs and bacterial taxa constructed based on ARG abundance profiles and taxa profiles at the genus level. Edge connections with p>0.8 are kept in the network.

At the species level, Gelatinosus hosts the most ARGs (13 ARGs), including glycopeptide resistant genes (VAND and VANRC), polymyxin (BASR), beta_lactam (OR214_01020, BL2C_PSE3), aminoglycoside (ANT1_1), and trimethoprim (DFRK). Analysis at both genus and species levels showed consistent results that glycopeptide resistant genes, beta-lactam genes, and aminoglycoside are likely to be preferred by microbial hosts.

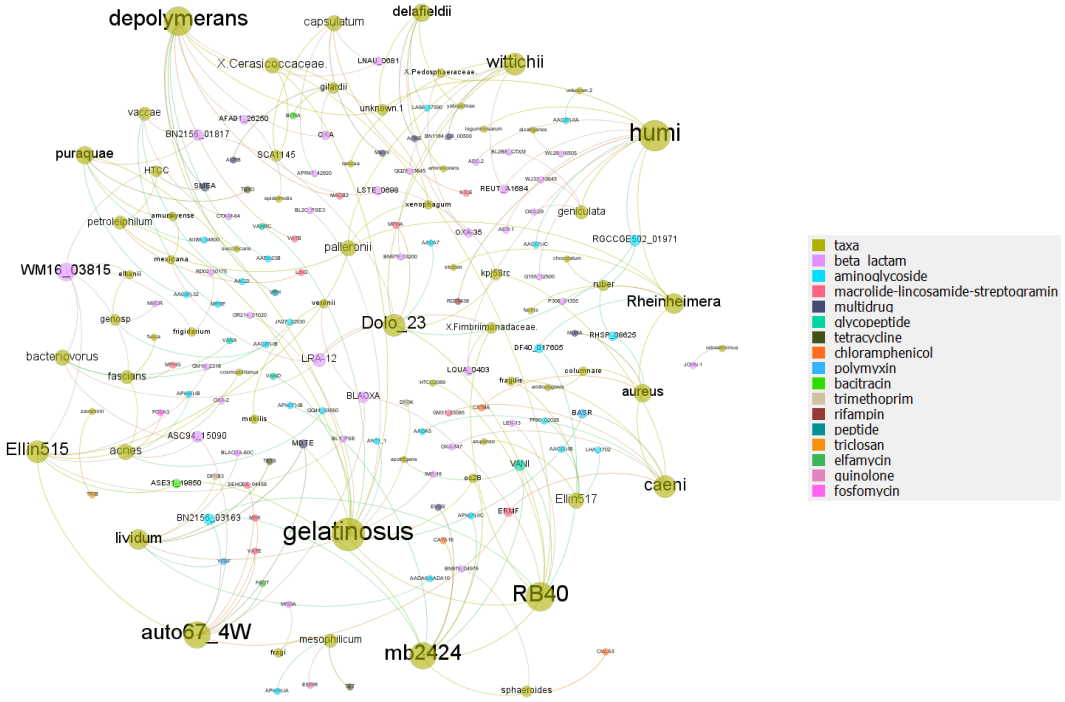


Figure 4.13: The Bayesian network of ARGs and bacterial species constructed based on ARG abundance profile and taxa profiles. For clarity, only edges with p>0.8 are kept.

4.4 Conclusions

Overall several observations can be made from the study of the simulated water distribution system. First, temperature and disinfectant types are important factors in influencing microbial regrowth. In particular, as temperature increases, regrowth occurs more often. Chlorine reduces microbial regrowth but chloramine does not. Second, biofilm/bulk water, disinfectant types, and temperature have significant contribution to the shaping of microbial communities. At 30^0C, biofilm environment has more diverse microbial compositions than water. Also disinfectants can reduce the diversity of the microbial communities and comparatively, chloramines has a stronger effect on the diversity reduction than chlorine. ARGs are the most diverse at 22^0C and the least diverse at 30^0C. Network analysis on assembly data reveals that UPPP and DFRE are the most co-

occurred ARG pair. The Bayesian networks constructed for ARG and genuses/species abundances reveal important host information of ARGs.

Chapter 5

CONCLUSION

Mapping whole-genome bisulfite short reads is challenging because of reduced complexity in genome sequences due to bisulfite treatment and increased search space after PCR amplification in the experiment. This thesis evaluates different bisulfite short read mapping tools and develops a framework for bisulfite sequencing analysis. First, we compared five different bisulfite short read mappers. Though Bismark is not the fastest mapper, it has the highest mapping efficiency and is highly recommended for bisulfite short reads alignment. Pre-processing data, i.e., trimming bad quality bases in short reads improves mapping efficiency. Sequencing errors have a negative impact on mapping efficiency for all the mappers. Second, we developed a Bayesian framework that takes advantage of uniquely mapped reads to differentiate ambiguously mapped short reads. By applying the Bayesian scoring model BAM-ABS on simulation and real hairpin mouse data, we showed that up to 70% of the ambiguously mapped short reads were assigned to unique locations with 90% accuracy. Thus, BAM-ABS is effective in mapping multireads to unique locations. Moreover, BAM-ABS showed robust performance for data with different methylation rates. As expected, increase in depth coverage and read length improves the performance of BAM-ABS while sequencing error decreases its performance. BAM-ABS assumes most of the variants

between uniquely mapped reads and multireads are homozygous. However, this may not be applicable to all cases. Therefore, for future work, improvement can be made by incorporating heterozygous variants into the scoring model.

The subsequent step after bisulfite short read alignment is to detect differentially methylated sites between phenotype groups. The traditional techniques for DMR detection do not take correlation among cytosine sites into consideration and inaccurately detect DMRs when there are small methylation differences with small sample sizes between phenotype categories. This thesis evaluates the traditional (methylKit) and Bayesian WFMM methods for DMR detection using simulated and real data with small methylation effect. Results show that WFMM has higher sensitivity and specificity than methylKit when methylation effect is small (i.e., average methylation differences between phenotype categories <0.01). We also suggest empirical rule to tune parameters to be reflective of specific methylation profiles. The method can be easily turned into a classifier for general machine learning purpose that can incorporate spatial and temporal correlation in the data.

Reusing treated waste water is essential for water sustainability. Our study of simulated RWDs shows that biofilm/bulk water, temperature, and disinfectant types play an important role in shaping microbiomes. ARGs are the most diverse under biofilm environment or at 22^{o}C. Increasing temperature to 30^{o}C and injecting disinfectant in water reduce microbial diversity. Network analysis of assembled data on ARG abundance shows that UPPP and DFRE are the most co-occurred. This network serves as a validation of network modeling. Results show that Bayesian networks fit our ARG profile data better than the network based on simple Spearman's correlation coefficients.

Bibliography

1. Das, P.M. and R. Singal, *DNA methylation and cancer.* Journal of Clinical Oncology, 2004. **22**(22): p. 4632-4642.
2. Jones, P.A. and D. Takai, *The role of DNA methylation in mammalian epigenetics.* Science, 2001. **293**(5532): p. 1068-1070.
3. Li, H., J. Ruan, and R. Durbin, *Mapping short DNA sequencing reads and calling variants using mapping quality scores.* Genome research, 2008. **18**(11): p. 1851-1858.
4. Krueger, F. and S.R. Andrews, *Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.* bioinformatics, 2011. **27**(11): p. 1571-1572.
5. Xi, Y. and W. Li, *BSMAP: whole genome bisulfite sequence MAPping program.* BMC bioinformatics, 2009. **10**(1): p. 232.
6. Coarfa, C., et al., *Pash 3.0: A versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel DNA sequencing.* BMC bioinformatics, 2010. **11**(1): p. 572.
7. Smith, A.D., et al., *Updates to the RMAP short-read mapping software.* Bioinformatics, 2009. **25**(21): p. 2841-2842.
8. Wu, T.D. and S. Nacu, *Fast and SNP-tolerant detection of complex variants and splicing in short reads.* Bioinformatics, 2010. **26**(7): p. 873-881.
9. Hercus, C., *Novocraft short read alignment package.* Website http://www. novocraft. com, 2009.
10. Homer, N., *Bfast: Blat-like fast accurate search tool.* 2009.
11. Harris, E.Y., et al., *BRAT: bisulfite-treated reads analysis tool.* Bioinformatics, 2010. **26**(4): p. 572-573.
12. Pedersen, B., et al., *MethylCoder: software pipeline for bisulfite-treated sequences.* Bioinformatics, 2011. **27**(17): p. 2435-2436.
13. Cokus, S.J., et al., *Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning.* Nature, 2008. **452**(7184): p. 215-219.
14. Chen, P.-Y., S.J. Cokus, and M. Pellegrini, *BS Seeker: precise mapping for bisulfite sequencing.* BMC bioinformatics, 2010. **11**(1): p. 203.
15. Guo, W., et al., *BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data.* BMC genomics, 2013. **14**(1): p. 774.
16. Hoffmann, S., et al., *Fast mapping of short sequences with mismatches, insertions and deletions using index structures.* PLoS computational biology, 2009. **5**(9): p. e1000502.
17. Dinh, H.Q., et al., *Advanced methylome analysis after bisulfite deep sequencing: an example in Arabidopsis.* PloS one, 2012. **7**(7): p. e41528.
18. Lim, J.-Q., et al., *BatMeth: improved mapper for bisulfite sequencing reads on DNA methylation.* Genome Biol, 2012. **13**: p. R82.
19. Prezza, N., et al. *ERNE-BS5: aligning BS-treated sequences by multiple hits on a 5-letters alphabet*. in *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*. 2012. ACM.

20. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
21. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature methods, 2012. **9**(4): p. 357-359.
22. Kent, W.J., *BLAT—the BLAST-like alignment tool.* Genome research, 2002. **12**(4): p. 656-664.
23. Li, R., et al., *SOAP: short oligonucleotide alignment program.* Bioinformatics, 2008. **24**(5): p. 713-714.
24. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows–Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-1760.
25. Fonseca, N.A., et al., *Tools for mapping high-throughput sequencing data.* Bioinformatics, 2012. **28**(24): p. 3169-3177.
26. Chatterjee, A., et al., *Comparison of alignment software for genome-wide bisulphite sequence data.* Nucleic acids research, 2012. **40**(10): p. e79-e79.
27. Hatem, A., et al., *Benchmarking short sequence mapping tools.* BMC bioinformatics, 2013. **14**(1): p. 184.
28. Li, H. and N. Homer, *A survey of sequence alignment algorithms for next-generation sequencing.* Briefings in bioinformatics, 2010. **11**(5): p. 473-483.
29. Homer, N., B. Merriman, and S.F. Nelson, *BFAST: an alignment tool for large scale genome resequencing.* PloS one, 2009. **4**(11): p. e7767.
30. Sedlazeck, F.J., P. Rescheneder, and A. von Haeseler, *NextGenMap: fast and accurate read mapping in highly polymorphic genomes.* Bioinformatics, 2013. **29**(21): p. 2790-2791.
31. Ferragina, P. and G. Manzini. *Opportunistic data structures with applications*. in *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*. 2000. IEEE.
32. Harris, E.Y., et al., *BRAT-BW: efficient and accurate mapping of bisulfite-treated reads.* Bioinformatics, 2012. **28**(13): p. 1795-1796.
33. Esteller, M., et al., *A gene hypermethylation profile of human cancer.* Cancer Res, 2001. **61**.
34. Baylin, S.B. and J.G. Herman, *DNA hypermethylation in tumorigenesis: epigenetics joins genetics.* Trends Genet, 2000. **16**.
35. http://www.bioinformatics.babraham.ac.uk/projects/sherman/.
36. Minoche, A.E., J.C. Dohm, and H. Himmelbauer, *Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems.* Genome Biol, 2011. **12**(11): p. R112.
37. http://www.ensembl.org/info/data/ftp/index.html.
38. http://www.biostars.org/p/59281/.
39. Lister, R., et al., *Global epigenomic reconfiguration during mammalian brain development.* Science, 2013. **341**(6146).
40. Trapnell, C. and S.L. Salzberg, *How to map billions of short reads onto genomes.* Nature biotechnology, 2009. **27**(5): p. 455.
41. Robinson, M.D., et al., *Statistical methods for detecting differentially methylated loci and regions.* Frontiers in genetics, 2014. **5**: p. 324.
42. Leek, J.T., et al., *Tackling the widespread and critical impact of batch effects in high-throughput data.* Nature Reviews Genetics, 2010. **11**(10): p. 733-739.

43. Morris, J.S. and R.J. Carroll, *Wavelet-based functional mixed models.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006. **68**(2): p. 179-199.
44. Guo, J.U., et al., *Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain.* Nature neuroscience, 2014. **17**(2): p. 215-222.
45. Lorincz, M.C. and M. Groudine, *CmC (a/t) GG methylation: A new epigenetic mark in mammalian DNA?* Proceedings of the National Academy of Sciences, 2001. **98**(18): p. 10034-10036.
46. Zhao, Z. and E. Boerwinkle, *Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome.* Genome research, 2002. **12**(11): p. 1679-1686.
47. Ziller, M.J., et al., *Charting a dynamic DNA methylation landscape of the human genome.* Nature, 2013. **500**(7463): p. 477-481.
48. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nature genetics, 2011. **43**(5): p. 491-498.
49. Li, R., et al., *SNP detection for massively parallel whole-genome resequencing.* Genome research, 2009. **19**(6): p. 1124-1132.
50. Zhao, L., et al., *The dynamics of DNA methylation fidelity during mouse embryonic stem cell self-renewal and differentiation.* Genome research, 2014. **24**(8): p. 1296-1307.
51. Porter, J., et al. *Improving bisulfite short-read mapping efficiency with hairpin-bisulfite data*. in *Computational Advances in Bio and Medical Sciences (ICCABS), 2014 IEEE 4th International Conference on*. 2014. IEEE.
52. Laird, C.D., et al., *Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules.* Proceedings of the National Academy of Sciences, 2004. **101**(1): p. 204-209.
53. Hansen, K.D., et al., *Increased methylation variation in epigenetic domains across cancer types.* Nature genetics, 2011. **43**(8): p. 768-775.
54. Tran, H., et al., *Objective and Comprehensive Evaluation of Bisulfite Short Read Mapping Tools.* Advances in bioinformatics, 2014. **2014**.
55. Liu, Y., et al., *Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data.* Genome Biol, 2012. **13**(7): p. R61.
56. Lee, W. and J.S. Morris, *Identification of differentially methylated loci using wavelet-based functional mixed models.* Bioinformatics, 2016. **32**(5): p. 664-672.
57. Kim, G., et al., *Herbicide injury induces DNA methylome alterations in Arabidopsis.* PeerJ, 2017. **5**: p. e3560.
58. Akalin, A., et al., *methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles.* Genome Biology, 2012. **13**(10): p. R87.
59. Zhu, H., P.J. Brown, and J.S. Morris, *Robust, adaptive functional regression in functional mixed model framework.* Journal of the American Statistical Association, 2011. **106**(495): p. 1167-1179.
60. Andrews, S., *FastQC: a quality control tool for high throughput sequence data*. 2010.
61. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014: p. btu170.
62. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet. journal, 2011. **17**(1): p. pp. 10-12.

63. Huang, D.W., et al., *DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists.* Nucleic acids research, 2007. **35**(suppl 2): p. W169-W175.
64. Das, M., et al., *A composite transcriptional signature differentiates responses towards closely related herbicides in Arabidopsis thaliana and Brassica napus.* Plant molecular biology, 2010. **72**(4-5): p. 545-556.
65. Bell, J., et al., *Differential methylation of the TRPA1 promoter in pain sensitivity.* Nature communications, 2014. **5**.
66. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature protocols, 2009. **4**(1): p. 44-57.
67. Park, L., et al., *The Role of Recycled Water In Energy Efficiency and Greenhouse Gas Reduction.* California Sustainability Alliance, available at: www. sustainca. org Also see: California Energy Commission (2005). Integrated Energy Policy Report, 2005.
68. Blackburn, B.G., et al., *Surveillance for waterborne-disease outbreaks associated with drinking water—United States, 2001–2002.* MMWR Surveill Summ, 2004. **53**(8): p. 23-45.
69. Pruden, A., et al., *Research Needs for Opportunistic Pathogens in Premise Plumbing: Methodology.* Microbial Ecology, and Epidemiology. Water Research Foundation Project, 2013. **4379**.
70. Wang, H., et al., *Probiotic approach to pathogen control in premise plumbing systems? A review.* Environmental science & technology, 2013. **47**(18): p. 10117-10128.
71. Rizzo, L., et al., *Urban wastewater treatment plants as hotspots for antibiotic resistant bacteria and genes spread into the environment: a review.* Science of the total environment, 2013. **447**: p. 345-360.
72. Szczepanowski, R., et al., *Detection of 140 clinically relevant antibiotic-resistance genes in the plasmid metagenome of wastewater treatment plant bacteria showing reduced susceptibility to selected antibiotics.* Microbiology, 2009. **155**(7): p. 2306-2319.
73. Wang, H., et al., *Effect of disinfectant, water age, and pipe material on occurrence and persistence of Legionella, mycobacteria, Pseudomonas aeruginosa, and two amoebas.* Environmental science & technology, 2012. **46**(21): p. 11566-11574.
74. Arango-Argoty, G., et al., *MetaStorm: A Public Resource for Customizable Metagenomics Annotation.* PloS one, 2016. **11**(9): p. e0162442.
75. Arango-Argoty, G.A., et al., *DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data.* bioRxiv, 2017: p. 149328.
76. Leplae, R., et al., *ACLAME: a CLAssification of Mobile genetic Elements.* Nucleic acids research, 2004. **32**(suppl_1): p. D45-D49.
77. DeSantis, T.Z., et al., *Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB.* Applied and environmental microbiology, 2006. **72**(7): p. 5069-5072.