Statistical algorithms in the study of mammalian DNA methylation

Meromit Singer

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2012-272
http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-272.html

December 21, 2012

Statistical algorithms in the study of mammalian DNA methylation

by

Meromit Singer

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Lior Pachter, Chair
Professor Richard M. Karp
Professor Jasper Rine

Fall 2012

Statistical algorithms in the study of mammalian DNA methylation

Abstract

Statistical algorithms in the study of mammalian DNA methylation

by

Meromit Singer

Doctor of Philosophy in Computer Science

and the Designated Emphasis

in

Computational and Genomic Biology

University of California, Berkeley

Professor Lior Pachter, Chair

DNA methylation is a dynamic chemical modification that is abundant on DNA sequences and plays a central role in the regulatory mechanisms of cells. This modification can be inherited across cell divisions and generations, providing a "memory mechanism" for regulatory programs that is more flexible than that coded in the DNA sequence. In recent years, high-throughput sequencing technologies have enabled genome-wide annotation of DNA methylation. Coupled with novel computational machinery, these developments have enabled unperceivable insight to the characteristics, biological function and disease association of this phenomenon. The collaborations between experimental and computational researches who take part in these efforts has been closer than ever before due to the need to involve computational methodologies throughout the entire research pipeline, from experimental design through bias correction to the analysis of large datasets.

In the first part of this thesis we present contributions to the field of high-throughput DNA methylation. We introduce statistically sound criteria for the detection of methylation signatures in DNA sequence, and present an algorithm for the annotation of an informative non-overlapping subset of such regions that is optimal under biologically motivated assumptions. Our method outputs a sequence-generated list of regions that are of interest with respect to their methylation states. We then present a Bayesian network to infer corrected site-specific methylation states from a favorable but biased experimental method, and describe its incorporation in a software package. Along with site-specific methylation calls our package annotates experiment-specific regions of interest by considering both the methylation state inferences and the genomic sequence. These regions can serve as a basis for comparative methylation studies. In the last chapter of this section we bring results from a genome-scale comparative study conducted on humans, chimpanzees and an orangutan, providing evidence of DNA methylation differences that propagate through generations and distinguish these closely related species.

The second part of this thesis concerns error correction in high-throughput sequencing datasets. In the course of studying DNA methylation with high-throughput sequencing we discovered a systematic error that results in false-positive variant detection and can significantly affect biological inferences in a variety of genomic studies. We present a classifier to correct for such errors and show that it performs very well with respect to both sensitivity and specificity.

To my parents.

Contents

List of Figures

List of Tables

Acknowledgments

This thesis would not have been possible without continuous support from my advisor Lior Pachter. The encouragement I received from Lior to pursue a broad range of research interests has led to fascinating years at Berkeley that I will always cherish. Lior's sincere love for science and the beauty of exploring new frontiers is contagious, and I feel honored to have been under this influence during these years. I thank Lior for his guidance and insightful advice throughout the years, at times when I needed it most.

I have been fortunate to meet wonderful colleagues over these years. I am thankful to my committee, Richard Karp, Jasper Rine and Michael Jordan, for their attention and advice. I thank Dario Boffelli and David Martin for contributing to my knowledge of experimental biology through our close collaboration, and for their valuable input during the early years of my PhD. I thank Yael Mandel-Gutfreund and Idit Kosti for an exciting ongoing collaboration and for hosting me in their group. I thank Zohar Yakhini for being an invaluable mentor and collaborator. I thank Alex Engström for many enlightening and entertaining discussions we have had on math, biology, and areas in between. Of the many wonderful people that have contributed to the making of this thesis and to my progress as a researcher I would specifically like to thank Sharon Aviran, Eran Halperin, Gadi Kimmel, Bonnie Kirkpatrick, Limor Leibovich, Frazer Meacham, Alex Schönhuth, Ron Shamir, Sriram Sankararaman, and Rotem Sorek.

I thank the bioinformatics group of Illumina at Hayward for an exciting summer internship, and specifically Steffen Durinck and Gary Schroth. I am grateful for the generous funding received from Microsoft through the Graduate Women's Scholarship that has supported me during the third year of studies. I thank Lior's lab for being great and the theory group at Berkeley for its hospitality towards theory-oriented folks.

I thank Berkeley for being itself. I am forever grateful for the unique and colorful atmosphere, the wonderful food and the good coffee, all of which have contributed to my optimism and spirit. I have also had the extraordinary gift of meeting friends for life during my time here, with whom philosophical late-night conversations have contributed to my sanity at times. I also thank my childhood friends from Israel for their sincere engagement and interest in my path.

I would like to wholeheartedly thank my parents Ayelet and Gadi for their unconditional support throughout my life and specifically during these years. I thank them for their dedication, for granting me with a sense of self confidence, and for always accepting me as I am. I thank my brother, Tal, for tolerating me always, and for listening and offering his input whenever I needed it. I would also like to thank my son, Ilai, for his patience while I was writing this thesis.

And to Yaron, thank you for being the perfect partner to share this journey with. I thank you for your astonishing devotion to my success, that has left me speechless at times; for encouraging me to test my limits, but also urging that I relax once in a while. For being there every step of the way. I cannot thank you enough.

Chapter 1

Introduction

1.1 Introduction to epigenetics

Epigenetic mechanisms can be defined as mechanisms that influence phenotype[1] through heritable, but potentially reversible, regulation of gene expression [30]. A number of different mechanisms have been determined as epigenetic (e.g. DNA methylation, histone modifications, nucleosome positioning and replication timing [30]), and have been shown to be heritable across cell divisions and to differ across different tissues and cell-types [41; 54; 55; 58; 121; 147]. We summarize below the two most studied epigenetic features: DNA methylation and histone modifications.

DNA methylation. DNA methylation relates to the covalent attachment of a methyl group (-CH_3) to a cytosine nucleotide, in place of the hydrogen atom on the 5 position of the cytosine's pyrimidine ring, or to the addition of a methyl group to an adenine nucleotide. Adenine methylation has been found only in bacterial genomes [54], and the term "DNA methylation" usually, and throughout this thesis, refers to cytosine methylation. DNA methylation has been shown to be present in many different branches of the tree of life (some examples are its presence in human, rice, honeybee, green algae and fungi [177]), indicating that it is an ancient feature, predating the divergence of plants and animals [44; 177]. But it is also missing, and assumed to have been lost, in several well studied model organisms. The spread and characteristics of DNA methylation can differ greatly across organisms [158]. For example, while the methylation in animals is mostly restricted to CpG sites[2] (cytosines that are followed by guanines), in plants it is observed in a broader collection of contexts [88]. Major events in the cell, such as cell differentiation, X-chromosome inactivation and retrotransposon silencing to name a few, are characterized by DNA methylation, and it has been shown to be curtail for development: inhibition of the DNA methylating enzymes in mice results in embryonic lethality [89; 158]. Broadly speaking, DNA methylation is associ-

[1] See Appendix A for definitions of the biological terms used throughout this thesis.

[2] Annotating cytosines that are followed by guanines by the CpG notation is common in the literature and will be used throughout this thesis. The "p" stands for the phosphate group that binds the two bases.

ated with a heterochromatin state and silencing of transcription. It has been shown to be causal of transcription states in some cases, and associated with, but not causal of, transcription activity in others [109].

Histone modifications. Histones - proteins that function as the basic units around which the DNA is packed - can incorporate various types of covalent chemical modifications. These modifications take place at specific locations on the histones, called “histone tails”, and include lysine and arginine methylation, lysine acetylation, ubiquitination and serine phosphorylation [83]. A histone may harbor a number of different modifications at the same time, giving rise to many possible configurations, sometimes related to as a histone code [155; 164]. Histone modifications can influence the packing assembly of the DNA by moderating a histone’s DNA-binding affinity and by recruiting further remodeling complexes [83]. By doing so, these modifications can affect gene expression, where some modifications are associated with transcriptional repression, and others associated with transcriptional activation. Both types of modifications can be present at either promoter or intragenic regions [28].

Epigenetic phenomena can affect gene regulation and be inherited between cell divisions through a mechanism different than that by which they were initiated. This results in an ability to maintain a regulatory program across cell divisions, enabling a form of cell “memory”. Epigenetics is therefore at the forefront of cell differentiation studies [33; 75; 103], and has been shown to be dynamic across different developmental stages [49; 153], and to differ between different tissues [71]. Epigenetic states have also been associated with various diseases [141]. For example, many studies have reported association between altered methylation states and various cancers [9; 40; 76].

Another active field of research concerns deciphering the affects of environmental factors on epigenetic states, and the extent to which changes in epigenetic states are stochastic. For example, it has been shown that monozygotic twins accumulate differences in DNA methylation throughout their lives, and that the accumulated differences affect their gene expression portrait [47]. Another study has showed that prenatal tobacco smoke exposure affects DNA methylation of the fetus [23].

An interesting direction of research involves the detection and annotation of transgenerational inheritance of epigenetic factors. In plants for example, changes in DNA methylation that are inherited across generations have been shown to be rather frequent and seem to be a common way of adapting gene regulation to a changing environment [57; 65; 143; 173]. In mouse transgenerational inheritance of DNA methylation has been observed in several loci [20; 107; 131], despite wide-spread reprogramming that occurs in the zygote [152]. The implications of these phenomena for theories of inheritance and evolution are currently being explored [151].

In this thesis we focus on DNA methylation research in the era of high-throughput sequencing. In the next section we describe in more detail DNA methylation and its characteristics in the mammalian genome. We then describe, in section 1.3, high-throughput sequencing technologies and their contribution to various fields of biology research. Then, in section 1.4 we will describe in detail how high-throughput sequencing can be used in the study of DNA methylation.

1.2 DNA methylation in the mammalian genome

DNA methylation is an ancient phenomenon. While the genomes of central model organisms, such as C. elegance (worm), Drosophila melanogaster (fruit fly) and S. cerevisiae (yeast), are not methylated[3], the phenomenon has been observed throughout the tree of life [11; 44; 177], and is assumed to have been lost in those model organisms. While found in many organisms throughout the tree of life, DNA methylation exhibits different pattens, and adheres different roles, in different animals [158]. For example, invertibrate genomes tend to have mosaic methylation patterns in which there are large domains that are either highly methylated or lack methylation, while vertebrate genomes are highly methylated throughout with "islands" of unmethylated sites [158]. In animals DNA methylation occurs mostly at CpG sites, and in plants it occurs at all of CpG, CpH and CpHpH sites (where H is a nucleotide different from G). In this thesis we are concerned with DNA methylation in mammals and we dedicate this section to describe the known behavior and functionality in that group.

In the mammals DNA methylation occurs predominantly at CpG sites [158], with some non CpG methylation recently detected in pluripotent cell types [94; 181]. In this thesis we restrict ourselves to the discussion of CpG methylation, since non-CpG methylation has thus far been observed only in pluripotent cells and its basic characteristics are still being investigated [181]. The genomes of mammals are generally methylated, with unmethylated sites mostly occurring in clusters, many times referred to as unmethylated islands [158]. DNA methylation (or the lack of, at certain regions) has been show to be associated with many central mechanisms (e.g. gene expression and silencing, imprinting, suppression of viral genes and transposons [158]). Methylation states have been shown to be associated with diseases (e.g. type 2 diabetes [162]), and many studies of cancer have shown that the genome becomes hypomethylated, with some unmethylated regions becoming highly methylated [40]. There is a lot of interest in deepening our understanding in this respect for risk-assessment and treatment generation.

As mentioned above, DNA methylation is structured in mammalian genomes: the majority of the genome is methylated, and incorporates scattered "unmethylated islands" - regions in which the large majority of CpGs lack methylation. This structure has influenced the design of experiments and the analysis of datasets. Efforts have centered on determining the locations and boundaries of these islands as well as in determining their methylation status across different conditions, because although they are called "unmethylated islands", a region that is unmethylated in one tissue or condition may be methylated in a different environment. The genome sequence can hold information regarding the whereabouts of such islands because methylated CpGs tend to mutate into TpG or CpA sites (through deamination of the methylated cytosine). Therefore, the accumulation of CpG sites in a region of the genome indicates it is unmethylated or that the CpGs there are maintained by selection. Regions of the genome that are rich in CpGs are called CpG islands, and

[3] There have been several studies that report low levels of DNA methylation in Drosophila [56], but homologs of the canonical methylating enzymes have not been found, and the DNA methylation has been observed in motifs that are different from those prevalent in animals [11]. This leads to the conclusion that any mechanism that establishes or maintains DNA methylation in Drosophila is fundamentally different than that observed in other organisms to date.

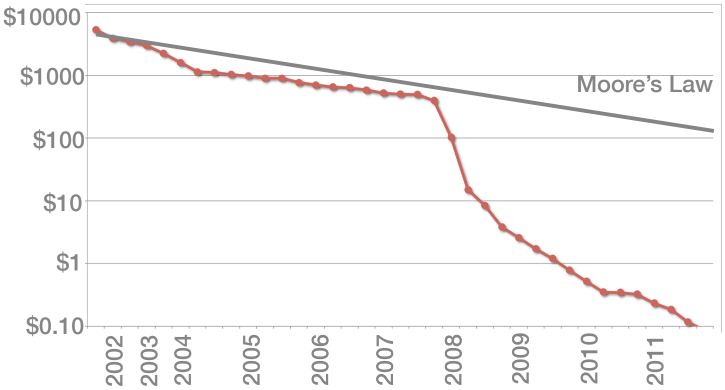


Figure 1.1: DNA sequencing costs per megabase (`genome.gov/sequencingcosts/`).

chapter 2 includes a detailed discussion of this subject in its introduction. It has been shown that the methylation of islands at promoter regions correlates with transcription silencing [172]. The recently introduced possibility of obtaining DNA methylation data at whole-genome scale (see sections 1.3 and 1.4) has enabled research of the methylation status of large sets of islands across different conditions. This has resulted in interesting observations regarding the islands' boundaries and definitions as well as their methylation states [73].

DNA methylation is currently a very active field, and has gained significant momentum over the past few years due to high-throughput sequencing methods that were adjusted for DNA methylation studies (see section 1.4). Given these recently developed abilities to annotate methylation states at genome-wide scale, and the novel findings such studies have resulted in so far, it is expected that this field will continue to flourish, alongside the development of bioinformatic methods for bias correction and analysis.

1.3 High-throughput sequencing

In the past decade technologies for sequencing genomic fragments at high throughput and low cost have been introduced, resulting in a revolution that has changed the way in which hypotheses can be generated and tested in a wide range of fields within biology. These technologies, known as "high-throughput sequencing" (HTS) or "next-generation sequencing" technologies, differ from classic Sanger sequencing by two main qualities: they produce short sequences (tens of base-pairs as apposed to hundreds) and do so at incredibly low cost (Figure 1.1) and high speed. This, combined with the ability to adjust these DNA-sequencing protocols for the study of various fields in biology research (see further description in section 1.3.2), has resulted in an explosion of research and progress, the depth and breadth of which are yet to be realized.

Technologies for HTS have been introduced by several different companies in parallel, and over the years improvements and novel technologies altogether have been introduced to the market (see [104] for a detailed discussion). While the protocols for sequencing differed across manufactures, all introduced new sequencing methods that were relatively simple to perform and produced extraordinary amounts of sequencing (from hundreds of thousands to hundreds of millions of DNA molecules per run) at extremely low costs, leading to vigorous competition. It is perhaps due to this competition that HTS technologies have been improving at an astonishing rate, both in throughput per-run and in price of sequencing (Figure 1.1). HTS techniques have introduced many new challenges for computational biologists. This is due in part to the massive amounts of data generated and to the different error profiles and biases compared to the earlier sequencing methods [38].

In this thesis we focus on Illumina sequencing (originally named "Solexa Sequencing", after the company that initiated the technology). Illumina sequencing is currently the most prevalent method for HTS, and has been dominating the market since 2007. The experimental methods and datasets used in this thesis make use of Illumina sequencing, and we therefore give a more detailed description of the technology in the next section.

1.3.1 Illumina sequencing

In this section we describe the basics of Illumina sequencing, in order to give the necessary background for understanding the experimental protocols used in this thesis.

In Illumina's HTS protocol the first step is fragmentation of the DNA, either randomly by sonication or by one of various non-random methods (e.g. digestion with restriction enzymes). The DNA fragments are then size-selected, reducing the sample so that the remaining fragments' lengths are distributed around 200 base-pairs (bps) in length. This step is usually performed by a run on an agarose gel. The fragments are then distributed onto a chip called a flow cell, each fragment "sticking" at some location. A PCR amplification step is conducted to the fragments on the flow cell to produce clusters; each cluster being a collection of (nearly) identical sequences. It is this PCR step that requires the size-selection of the fragments: since the amplification is done on the flow cell using a bridging technique, fragments that are too short cannot bridge well, and fragments that are too long form overlapping bridges, resulting in overlapping clusters of sequences. In the next step, the DNA molecules on the flow cell are sequenced one base-pair at a time and at each cycle the base-pair added at each cluster is read and recorded. Illumina sequencing is done in 4-color space, reading off the base that was added to each cluster by the spectrum of light extracted. The sequences produced are called *sequenced reads* (or simply reads) and are typically shorter than the length of the fragment. The current standard read length for Illumina sequencers is 100bps. In the sequencing step, either one or both ends of the fragments can be sequenced to produce reads. In the latter case, called paired-end sequencing, the sequenced reads are paired in the output file, such that a fragment is represented by a pair of sequenced reads.

The protocol described above is simple to execute and can produce large amounts of data cheaply, but suffers from various biases and error-generating mechanisms. Some examples are the non-uniform elongation of sequences in a cluster resulting in a noisy signal that leads to erroneous base calls [77], or bias in the extent to which a region of the genome is sequenced due to sequence composition [7; 38]. The rapid rate at which the technologies and protocols change results in changes in the types and characteristics of errors and biases. However, the annotation and correction for these behaviors is extremely important when separating true signal from noise, to avoid propagation of error to downstream analysis (e.g. [123]). An active field of research involves the annotation of such biases and the generation of solutions. Our contribution to the correction of error in Illumina HTS data is detailed in chapter 5 of this thesis.

1.3.2 The role of HTS in molecular biology

The technology used for HTS allows the sequencing of the ends of DNA fragments, with no restrictions on the sequence composition of the fragments to be sequenced, or on how those fragments were generated. It is this ability, coupled with the low cost of sequencing, that has resulted in a proliferation of experimental methods that make use of HTS to study different aspects of molecular biology and functional genomics.

Such methods consist of several steps:

1) **Reduction:** Reducing a biological question to a counting-of-fragments problem, through a protocol that generates an informative set of fragments.

2) **Data generation:** Sequencing of the fragments with HTS.

3) **Data analysis:** Designing a pipeline for answering the initial question from the sequencing output.

Example: Suppose our intention is to find the locations at which a specific transcription factor is bound to the genome (in some tissue). We can do this using HTS by:

1) Isolating from a population of cells DNA fragments that were bound by the transcription-factor, using the CHiP protocol (see [79] for more details).

2) Sequencing of the fragments isolated in step 1.

3) Applying a bioinformatic analysis. This usually incorporates mapping the sequenced reads to the appropriate reference genome (an established representative genome sequence for the species) and running a specialized peak-detection algorithm on the stacked reads. The algorithm attempts to find (in an unbiased manner) locations on the genome that generated a significant amount of sequence due to the transcription factor binding.

The protocol incorporating steps (1) and (2) above is called CHiP-Seq, and various methods have been proposed over the past few years for carrying out step (3), providing an answer to the initial question [48; 120]. While step (3) might seem simple given a protocol for step (1), the proliferation of methods available for analyzing CHiP-Seq data hint that this is not the case.

Indeed, noisy sampling, sequencing bias of various sorts, and the fact that a population of cells is being sequenced (rather than a single cell), all contribute to the peak-detection task on the reads that were mapped to a reference genome being far from trivial. Moreover, the initial step of mapping the reads onto a reference genome is in itself a challenging task, for which several methods have been proposed [61; 85; 90].

High-throughput sequencing has been incorporated into different protocols to study various phenomena, in a similar manner as the example above. In particular, in the field of DNA methylation HTS methods have flourished, and the main methods of interest are discussed in the next section. Other fields that have benefited from protocols incorporating HTS include cancer genomics, annotating rare variants in populations, gene expression studies, non-coding RNA annotation, metagenomics, histone modifications, splicing variants, open chromatin regions, and transcription elongation rate, to name a few.

As has been elicited in this section, over the past few years HTS has been deeply integrated into experimental pipelines in the fields of genomics and molecular biology. In the great majority of the HTS studies preformed it is the cost of analysis and interpretation, not data generation, that is the bottleneck [95]. In many of the cases the contributions of computational biologists have been critical for the success of such studies. First, the amounts of data generated are enormous. The need to handle large datasets with respect to storage, computation speed, and statistical analysis is currently being addressed by many fields, and within the field of computer science it is being acknowledged that achieving effective use of large datasets is of high priority. Second, the technology of HTS has biases that need to be found, characterized, and corrected. Third, analysis of the data generated by the different protocols requires development of protocol-specific algorithms to analyze the sequencing data and generate probabilistic answers to the initial experimental questions. In addition, the ability to conduct HTS studies adds relevance to existing genomic bioinformatic studies, due to the availability of larger numbers of genomes and to the fact that the results of the bioinformatic analysis can be used in the analysis of HTS studies. For example, a list of enhancers generated from genomic sequences using conservation patterns can be an interesting set to use in the analysis of HTS data aiming at mapping these enhancers.

It is currently clear that the potential of HTS is far from being exhausted. Companies that support HTS technology are constantly releasing new and improved products, and new companies are being launched. There is no doubt that the past decade has witnessed a revolution in the breadth of data produced and the role computational biologists play in experimental design and analysis. In the next years it is expected that researches build upon this accumulated knowledge to conquer additional goals, specifically in fields of regulatory genomics and personalized medicine, through the continued integration of experimental and computational efforts.

1.4 High-Throughput study of DNA methylation

We have described the transformative impact of HTS technologies on many fields in biology in the previous section. One of the fields in which these sequencing technologies have resulted in an explosion of novel research and in extraordinary progress is epigenetics, and specifically DNA methylation. In this section we describe different methods that use HTS to study DNA methylation, focusing on their advantages and challenges, and explain why there is a relatively large number of methods in use today. We will first outline several techniques that enable measurement of DNA methylation on a genome-wide scale, and then describe the main methods that make use of these techniques.

There are several ways to detect DNA methylation, and a detailed description of the different techniques can be found in [16; 84]. The major techniques currently available are illustrated in Figure 1.2 and include:

- **Enzyme digestion.** This technique uses restriction enzymes that in addition to being sequence-specific are either methylation sensitive (cut only if their recognition site is unmethylated) or methylation dependent (cut only if their recognition site is methylated). A notable example is *HpaII*, which digests at CCGG sites at which the second cytosine is unmethylated (in this thesis we denote a cytosine that is followed by a guanine as CpG, and to ease notation denote a pair of cytosines followed by a pair of guanines as CCGG, rather than CpCpGpG). The pattern of digestion by such enzymes can provide a read-out of the DNA methylation. A disadvantage of this technique is the possibility of non-complete digestion and for some enzymes, the lack of site-specific resolution.

- **Sodium bisulfite conversion.** Treatment of DNA with sodium bisulfite converts unmethylated cytosines to uracils [170; 66]. Followed by a PCR step, the uracils are sequenced as thymines. This procedure enables the conversion of an epigenetic mark into a modification in the genomic sequence. Disadvantages of this technique include a reduction in sequence complexity (see details below), the need to chemically treat the DNA which can result in degradation, and the possibility of non-complete conversion.

- **Affinity enrichment.** In this process, the DNA is digested, and the digest is enriched for methylated regions. This is achieved through the recognition of methylated regions by either antibodies [105] or methyl-CpG binding domain proteins [179] and the separation of DNA regions that were bound from those that were not. Disadvantages of this technique include binding biases and the lack of site-specific resolution.

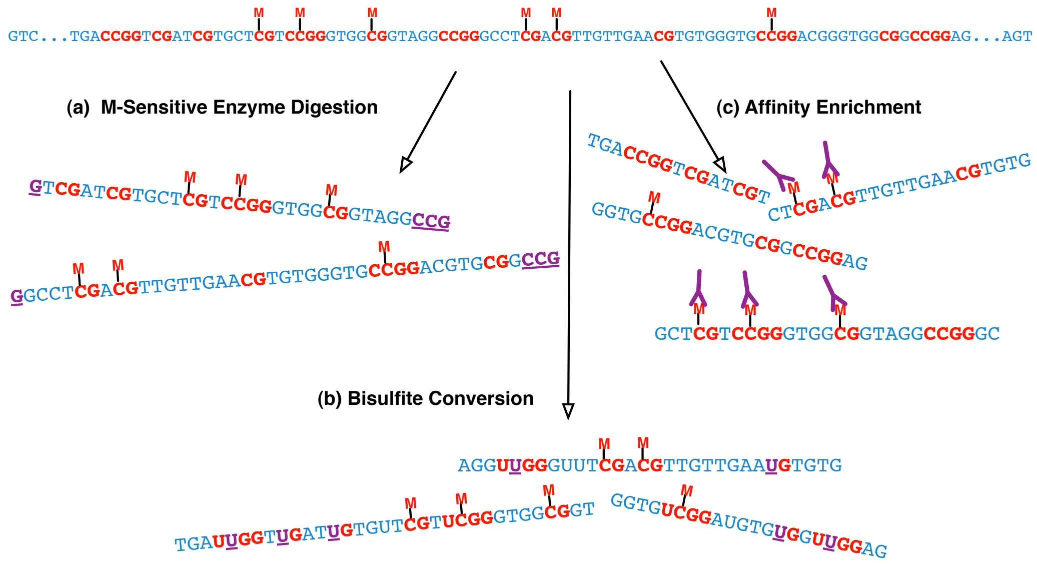

GTC...TGACCGGTCGATCGTGCTCGTCCGGGTGGCGGTAGGCCGGGCCTCGACGTTGTTGAACGTGTGGGTGCCGGACGGGTGGCGGCCGGAG...AGT
(a) M-Sensitive Enzyme Digestion
GTCGATCGTGCTCGTCCGGGTGGCGGTAGGCCG
GGCCTCGACGTTGTTGAACGTGTGGGTGCCGGACGTGCGGCCG
(c) Affinity Enrichment
TGACCGGTCGATCGT
CTCGACGTTGTTGAACGTGTG
GGTGCCGGACGTGCGGCCGGAG
GCTCGTCCGGGTGGCGGTAGGCCGGGC
(b) Bisulfite Conversion
AGGUUGGGUUTCGACGTTGTTGAAUGTGTG
TGAUUGGTUGATUGTGUTCGTUCGGGTGGCGGT
GGTGUCGGAUGTGUGGUUGGAG
M

Figure 1.2: Three common techniques for genome-scale annotation of DNA methylation. (a) Enzyme digestion (b) Bisulfite conversion (c) Affinity enrichment

1.4.1 Methylomes and methyltypes

Methods for genome-wide measurement of DNA methylation generally use one of the aforementioned techniques coupled with either high throughput sequencing or array hybridization. Potentially, methods may incorporate more than one of the techniques, but most current approaches do not do so. The methods can be broadly classified as *methyltyping* versus *methylome sequencing*, in analogy with *genotyping* versus *genome sequencing* for DNA. In genotyping, only a small subset of an individual's nucleotides are assayed (SNP locations), while in genome sequencing the whole genome of an individual is sequenced. The advantage of genotyping is in its significantly lower cost compared to whole genome sequencing, allowing for the inclusion of many more individuals in an experiment of a fixed cost. The locations sampled can be used to infer the sequence at adjacent locations. Similarly, methyltyping technologies allow surveying genome-scale methylation patterns by sampling a subset of CpG sites, and emphasize low cost at the expense of high resolution. In this section we discuss different methods for methylome sequencing and methyltyping, along with the advantages and disadvantages associated with each method.

The different methods used for high throughput genomics, including those for measuring DNA methylation, reduce to the ability of counting DNA fragments in a digest, obtained using either arrays or sequencing. Arrays are considered less accurate than sequencing for quantification purposes, mainly due to biases introduced in by the variability in hybridization. Each of the techniques described can be followed by array hybridization, but affinity enrichment is by far the technique that is best suited for arrays [84]. The advantage of using arrays is the low cost, but disadvantages include the biases introduced in the hybridization step, as well as genome-scale array based

	Site-specific	Pre-selected regions	Coverage of human genome	Coverage of CpG islands	# CpGs sampled	Analysis challenges
methyl-Seq	Yes	No	9.2%	92.9 %	$\sim$1.4M	Inference procedure needed
RRBS	Yes	No	8.1%	69.8 %	$\sim$1.4M	Low complexity + PCR bias
Affinity based Array	No	Yes	pre-selected	pre-selected	-	Binding biases
WGBS	Yes	No	whole genome	whole genome	$\sim$28M	Low complexity + PCR bias
Affinity based Seq	No	No	whole genome	whole genome	$\sim$28M	Binding biases + Array biases

Table 1.1: A summary of the characteristics of commonly used methods for methyltyping (top three) and whole methylome annotation (bottom two) in human. For further information on the derivation of the values in the table see [148] and [84].

methods not being site specific (there are array based methods that use either enzyme digestion or bisulfite conversion and are site-specific, but since they sample a considerably smaller number of CpGs, we do not consider them as genome-scale methyltyping methods).

Advances in high throughput sequencing in the past several years have brought about a large increase in the number of available methods for mapping DNA methylation on genome-wide scale. Whole-genome bisulfite sequencing (WGBS) involves random digestion of the genome followed by bisulfite treatment, amplification, and sequencing, offering the ability to measure absolute levels of DNA methylation as single-nucleotide resolution. While this procedure has been used in different studies [34; 93; 94; 177], its use is limited due to it being the most expensive method for DNA methylation annotation, because it requires the sequencing of whole genomes. Moreover, sequencing a whole methylome is considerably more expensive than sequencing a whole genome because of the continuous nature of the methylation phenomenon (we would like to determine the proportion of cells in the digest in which a cytosine was methylated), requiring significantly higher coverage than genome sequencing. Therefore, this method cannot, in the foreseeable future, be used for large-scale comparison studies such as population studies and epigenetic association studies. Other disadvantages of the method that are not present in whole-genome sequencing include biases introduced in the initial PCR amplification step (prior to distribution on the flow cell) and a reduction in sequence complexity. The PCR amplification step introduces biases that are due to unmethylated instances introducing AT-rich sequences, since AT-rich sequences are known to be amplified by PCR at a larger pace than CG-rich sequences. The sequence complexity is reduced because every “T” sequenced in a read could be mapped to either a “T” or a “C” in the reference genome, reducing the number of locations producing uniquely mapping reads. A second method for annotating methylation across the whole genome is that of affinity enrichment followed by high throughput sequencing. This method has been used in [98], but requires extensive sequencing, is not site-specific, and is prone to binding biases.

The use of high throughput sequencing has enabled several methods for methyltyping, two of which are discussed here: methyl-Seq and RRBS. In methyl-Seq the genome is digested with the methylation sensitive restriction enzyme *HpaII* that digests unmethylated CpGs that are within CCGG sites. This is followed by size selection of the fragments, a PCR amplification step and sequencing of the fragments' ends. Mapping sequenced reads back to a reference genome reveals unmethylated CpGs. methyl-Seq is a convenient methyltyping strategy because it is cost-effective due to the sequencing of only unmethylated sites (the minority of CpG sites in vertebrates [158]), requires only small amounts of material and avoids bisulfite conversion. However, although the experiment is relatively simple, interpretation of the sequencing data is not straightforward due to the protocol resulting in a non-random segmentation of the genome which is followed by a size selection step. We describe the methyl-Seq protocol and the bias that is associated with it in greater detail in chapter 3 of this thesis.

A second methyltyping method, Reduced Representation Bisulfite Sequencing (RRBS), is based on digestion with a methylation insensitive enzyme followed by bisulfite sequencing [102]. The procedure is enzyme digestion followed by size selection of the fragments, bisulfite treatment, PCR amplification and sequencing of the fragments' ends. The digestion is commonly performed with *Msp*I [60; 103], which digests CCGG sites, to enrich for regions of the genome that are rich in CpG sites. RRBS is favorable due to being significantly less expensive to perform than whole-genome bisulfite sequencing, but suffers from the same disadvantages related to methods that use bisulfite conversion: bias introduced by the PCR amplification step and reduced sequence complexity.

When designing an experiment in which DNA methylation is measured on genome-wide scale, one must consider the tradeoffs between the cost of the method and the type of coverage it produces. In addition to this, the application and analysis of the different methods are complicated by a number of other issues. The major complications of the different methods that originate from the methylation-detection technology used were discussed above. On top of this, the analysis requirements for different assays vary in difficulty, and the rapid development of the sequencing field calls for frequent re-assessment of the comparison between methods. In [84] it was suggested that methyl-Seq is the method with the most favorable profile of pros and cons, with respect to the measures chosen for comparison (see Table 2 of [84]). We compare here the different methods presented, given the current stage of the field, but the reader should keep in mind that the comparison criteria will change as the field develops and as new techniques are introduced.

Table 1.1 summarizes the main features by which methods are compared. It is divided into methyltyping methods and whole-genome methods. We have omitted a column for a cost comparison due to the changing nature of the field (sequencing costs are rapidly decreasing), but it seems certain that in the foreseeable future whole-genome methods will require significantly more sequencing than methyl-Seq and RRBS, and that RRBS will require more sequencing than methyl-Seq.

1.5 Outline of this thesis

This thesis has two parts. The first part centers on statistical algorithms and techniques used for the study of DNA methylation in mammals, in the era of HTS data. Chapter 2 presents a statistical method for annotating regions of interest with respect to their methylation states based on genomic sequence, and is based on joint work with Alexander Engström, Alexander Schönhuth and Lior Pachter [149]. Chapter 3 presents a Bayesian network for bias correction and first analysis of output from a methyl-Seq experiment, and is based on joint work with Lior Pachter [150] and on joint work with Dario Boffelli, Joseph Dhabhi, Alexander Schönhuth, Gary Schroth, David Martin and Lior Pachter [148]. In chapter 4 we discuss a genome-wide comparative study of DNA methylation across primates, based on joint work with David Martin, Joseph Dhahbi, Guanxiong Mao, Lu Zhang, Gary Schroth, Lior Pachter and Dario Boffelli [97].

The second part of this thesis centers on error-correction techniques for HTS. In chapter 5 we give a characterization of a novel type of error in Illumina datasets and present a procedure to correct for it. This chapter is based on joint work with Frazer Meacham, Dario Boffelli, Joseph Dhahbi, David Martin and Lior Pachter [100].

We give definitions and descriptions of the biological and sequencing terms used in this thesis in appendix A.

Part I

Statistical algorithms in the study of mammalian DNA methylation

Chapter 2

Using Markov chains to annotate CpG islands

2.1 CpG islands

The ability to annotate DNA methylation on genome-wide scale calls for the development of new strategies for the analysis of this type of data. Studies that make use of this data usually incorporate an analysis approach that is either site-specific (considers the methylation states of specific cytosines), regional (considers some statistic of the methylation states in a region, the mean methylation state for example), or some combination of the two. Considering the methylation states across regions is natural for genomes that are overall methylated and contain "unmethylated islands" such as the mammalian genomes. As detailed in section 1.2, in many genomes the majority of the CpG sites are methylated and occurrences of unmethylated CpGs tend to cluster together, in so-called unmethylated islands. The extent of methylation at such islands has been shown to be associated with the expression levels of nearby genes [172], and in many cases aberrant methylation of these islands is associated with diseases such as cancer [76; 154]. Computational methods based on evolutionary principles have been used to annotate these functional regions and are based on the fact that deamination of a methylated cystosine in a CpG dinucleotide results in a TpG (or, in the complementary strand, a CpA dinucleotide). Since only methylated cytosines are deaminated in such a manner, and in many organisms, in particular in mammals, cytosines are methylated at CpG sites, one observes an overall depletion of CpGs and frequently observed C $\rightarrow$ T mutations in alignments of related species. Unmethylation in the germline and selection are seen as the dominating mechanisms which shield functional regions from CpG depletion [22; 142].

Due to the connection between CpG depletion and methylation, it is possible to identify unmethylated genomic regions based on sequence criteria alone. Interest in sequence-defining features of unmethylated regions which are over-represented in CpG dinucleotides, known as *CpG islands (CGIs)*, is also motivated by the need for CGIs as ground sets in a variety of more specific studies on the functional impacts of methylation (e.g. [71; 156; 172]). While the recent advances in high-throughput sequencing have been used to probe methylation states at genome-wide resolution [64; 94; 148], providing tools for unprecedented progress in the study of regulatory methylation, such methods cannot annotate a complete set of regions which are unmethylated at some developmental stage or in some tissue. This is particularly relevant in intragenic regions, where the functionality of unmethylated islands is likely to be different from that of unmethylated islands overlapping promoters [72]. While most CGIs at promoters are observed to be consistently unmethylated across different cell types and conditions [72], this is not the case for islands in intragenic regions [98], and it is hypothesized that methylation states of CGIs in intragenic regions are highly tissue specific and are mostly involved in the "fine tuning" of expression [72]. Therefore, annotating regions of interest based on genomic features (rather than experimental methylation states) is of great relevance in this setting. Investigation of the extent of conservation to measure methylation (*pro-epigenetic selection*) has recently been undertaken in this intragenic setting with results suggesting extensive unmethylation of CpGs in developmentally related genes [22; 101].

In this chapter we present a new approach to annotate regions of interest with respect to methylation from the genome sequence (CpG islands). We give a direct and simple criterion for defining these islands—arguably the simplest criterion possible. The simplicity of our definition lies in the fact that it only requires specification of a (Markov based) null model. We define CpG islands to be regions where the number of CpGs significantly deviates from this null model according to a p-value threshold. We demonstrate the utility of the p-value of a region in assessing the functional significance of a CpG island by showing that p-values can be used to prioritize existing CpG island predictions according to functional significance.

We then describe an algorithm for selecting a set of non-overlapping islands from among all possible islands, which is optimal for biologically motivated criteria. This result is proved in Theorem 1. Our algorithm is greedy, and has a stopping criterion based on the Benjamini-Hochberg theory for controlling the false discovery rate. Coupled with a dynamic programming procedure which allows to efficiently perform the hypothesis tests required, this algorithm yields a powerful and utmost flexible new approach to finding CpG islands: it can easily be applied to different genomes, or to regions subject to selective pressures that affect nucleotide composition. The only parameter to be specified in our approach is the false discovery rate — the model parameters are estimated directly from the data.

We showcase our approach by predicting *coding* CpG islands in the human genome (CpG islands that are within regions of the genome that code for proteins). To our knowledge, we are the first to predict such CGIs utilizing a statistical model that appropriately accounts for the excess of CpGs expected due to coding constraints. We show that in coding regions different rules should be applied and illustrate this by showing that previous methods miss coding CGIs in some exons while predicting non-significant islands in others.

Related Work

The presence of unmethylated regions in mammalian genomes was reported during the 1980s [12]. These works also noted the relative abundance of CpG sites within unmethylated regions compared to the rest of the genome. The term CpG-rich islands was initially used in [13] and shortly afterwards Gardiner-Garden and Frommer proposed thresholds for sequence-based annotation of CpG islands [50]: $\geq$ 200bps, GC % $\geq$ 50 and number of observed CpGs/number of expected CpGs ≥ 0.6.

Over the years a variety of novel statistical approaches to genome-wide prediction of CGIs have been proposed, leading to improvements and allowing genome-wide computation of CGIs in a variety of genomes (Table 2.1). It is noteworthy that the Gardiner-Garden and Frommer (GF) definition depends heavily on the precise implementation of the search algorithm. Since there is no canonical choice for a search algorithm many alternative options have been suggested (see e.g. [80; 125; 160; 171] for different sliding-window algorithms). However, it has recently been shown that these differences can lead to nearly arbitrary differences in the sets of CGIs selected [68; 72]. Indeed, a variety of recent epigenomic studies [39; 71; 148; 156; 172; 180] have pointed out that the popular GF criteria, and also the later, more stringent definition due to Takai and Jones [160], have to be adjusted. Recently, Hsieh et al. [68] and Wu et al. [175] have proposed statistical definitions of CGIs and their methods also provide statistical approaches to identifying non-overlapping islands. However the sophisticated statistical techniques underlying these methods make it difficult to succinctly describe the characterizing properties of the CpG islands they predict, and it is perhaps for this reason that researchers, despite the advantages of these statistical approaches, continue to use the Gardiner-Garden-Frommer based CpG islands available from the UCSC genome browser.

In the case of coding regions, the problems discussed above are compounded by the selective pressures on nucleotides (resulting from the translation of codons into amino acids) that affect nucleotide composition [22; 98; 142]. The problem of modifying heuristics such as the GF definition in order to account for coding constraints is non-trivial. In principle, current statistical approaches to predicting CpG islands could be adapted to explicitly incorporate the characteristics of coding regions, but we do not know of any method that has been proposed for doing so.

The approach we present in this chapter assumes as a null hypothesis that DNA sequences are generated by Markov chains of higher order, and we will make “CpG island calls” if CpG counts significantly deviate from the background model assumption. Short range dependencies among nucleotides have been reported early on [2; 14; 122] where [14] even suggests that higher-order Markov chains are, to a certain degree, “realistic” genomic sequence models. In this work we center on 5-th order Markov models due to the benefits of such higher order models in analyzing exonic sequences. This approach has been used before in prominent gene finding programs. One example is GENSCAN [27], which uses “phased” 5-th order models that were in turn inspired by earlier work on the topic [21; 51]. Use of “phased” 3-periodic inhomogeneous Markov chains necessitates a significant increase in the number of model parameters, which we have opted to avoid here and leave as interesting future work. There are different methods by which model order can be chosen, such as the Bayesian or Akaike Information Criterion (BIC and AIC), as suggested in [116], or by means of Chi-square tests, as suggested in [136]. Here, also following [136, p. 2], we agree that model order “from a practical point of view ... also depends on the composition of the biological sequence one wants to take into account.”, in that the 5-mer previous to a nucleotide

	Year	Island model needed	Parameters: statistical / ad hoc	Applicability of method	# Parameters (learned : specified)	Min length requirement	MHT correction
Our approach	2011	No	Statistical	Easy	17 (16:1)	No	Yes
Wu et al. [175]	2010	No	Both	Medium	($\sim 10^7$:1)	No	No
Hsieh et al. [68]	2009	Yes	Both	Weak	3 (2:1)	No	No
Straussmanet et al. [156]	2009	Yes	Both	Weak	12 (11:1)	Yes	No
Sujua et al. [157]	2008	No	Ad hoc	Weak	5 (0:5)	Yes	No
Glass et al. [53]	2007	No	Both	Easy	5 (2:3)	No	No
Hackenberg et al. [62]	2006	No	Ad hoc	Easy	5 (3:2)	No	No
UCSC Genome Browser [78]	2004	No	Ad hoc	Weak	5 (0:5)	Yes	No
Wang et al. [171]	2004	No	Ad hoc	Weak	7 (1:6)	No	No
Takai et al. [160]	2002	No	Ad hoc	Weak	6 (0:6)	Yes	No
Ponger et al. [125]	2002	No	Ad hoc	Weak	4 (0:4)	Yes	No
Gardiner-Garden et al. [50]	1987	No	Ad hoc	Weak	4 (0:4)	Yes	No

Table 2.1: Comparison of different methods for CpG island annotation. “Applicability of method” - how easily the method can be used for different sequences (such as different organisms, genomic regions etc.). “MHT” - multiple hypothesis testing. The large number of parameters learned for [175] is due to inferring ‘smooth deviations’ of CpG counts, individually for every genomic segment of length 16.

in a coding region always contains the entire codon of the previous amino acid.

Modeling exonic sequences by more involved methods such as employing hidden Markov models [31], variable length Markov models [26], drifting Markov models [167] or, as mentioned above, 3-periodic inhomogeneous (phased) Markov chains, may further reduce the number of false-positives. We leave it as future work to explore the use of such methods for this purpose, and to overcome the associated technical difficulties, such as parameter estimation and the rapid and efficient computation of p-values.

2.2 Methods

2.2.1 Notation

In the following, we denote the set of nucleotides by $\Sigma := \{A,C,G,T\}$. Let Σ^k be the set of strings over Σ of length k (k-mers) and let Σ^* be the set of all strings of finite length (oligomers of arbitrary length). We write $v \in \Sigma^k, w \in \Sigma^l$ for strings and $v \cdot w \in \Sigma^{k+l}$ (or simply vw) for their concatenation. For $B \subset \Sigma^{n-k}$ and $v \in \Sigma^k$ we write $Bv \subset \Sigma^n$ for the set of strings from B extended by the string v. In this language, genomes, or parts of the genome are collections of strings over Σ.

We write $\mathbf{G}_s^e$ for the substring of the string $\mathbf{G}$ which stretches from position s to e, including the positions s and e. Let $v,w \in \Sigma^*$ be strings and define $\#(v,w)$ to be the number of (overlapping) occurrences of the string w as a substring in v. For example, $\#(CGACG,CG) = 2$.

Let $\mathbf{P}_n$ be some probability distribution over strings of length n, $w = CG$ and $v \in \Sigma^n$ be a random string drawn according to $\mathbf{P}_n$. We denote by

$$p_{n,m} := \mathbf{P}_n(\#(v,w) \geq m) \tag{2.1}$$

the tail probability that an n-mer randomly drawn according to P_n contains at least m occurrences of CG. A small value for $p_{n,m}$ indicates that a substring with m CGs contains an unusually high number of CGs.

Let P_n and w be as before and let $u \in \Sigma^n$ be a random string drawn according to P_n that begins and ends with a CG. We denote by

$$p_{n,m}^{cg\to cg} := \mathbf{P}_n(\#(u,w) \geq m \mid u \in CG\Sigma^{n-4}CG) \tag{2.2}$$

the tail probability that an n-mer randomly drawn according to P_n *that begins and ends with CG* contains at least m occurrences of CGs.

For a given substring $\mathbf{G}_s^e$ where $n = e - s + 1$ and $m = \#(\mathbf{G}_s^e, CG)$ we denote $p(\mathbf{G}_s^e) := p_{n,m}$ and $p^{cg\to cg}(\mathbf{G}_s^e) := p_{n,m}^{cg\to cg}$.

2.2.2 CpG islands and coding CpG islands

As discussed above, CpG islands may be defined in terms of a region's functionality, but in this study we propose a purely sequence based definition. Sequence based definitions have the advantage of being applicable to any sequenced genome without the need for further experimental information, and can pinpoint CpG islands that are unmethylated only in particular tissues (being maintained as CGIs due to selective pressure), and therefore hard to find by experimental mapping of DNA methylation in a subset of tissues.

We begin by giving a statistical definition of CpG islands, and return to questions about association with function in the Results section. Due to the higher mutation rate of methylated cytosines to thymines, and the methylation of cytosines in mammals being mostly at CpG sites, regions that are methylated in the germline and at which there is no selection to maintain CpG sites will be sparser in CpG sites than regions at which the previous conditions do not apply. We postulate, as in [50], that CpG islands should be defined in terms of the number of CpGs. Specifically, we define CpG islands (CGIs) and coding CpG islands (CCGIs) as regions with a significantly high number of CpGs.

Definition 1. *Given a threshold for significance q, a region G_s^e is a CpG island iff $p_{(e-s+1),m} < q$, where $m = \#(G_s^e, CG)$.*

Definition 2. *Given a threshold for significance, q, a region G_s^e is a coding CpG island iff $p_{(e-s+1),m} <$ q, where $m = \#(G_s^e, CG)$, and G_s^e is completely contained within a coding exon.*

There are two immediate advantages to using our method for direct annotation of coding CpG islands. The first being that coding exons are under strong selective pressure resulting in different mutation rates than in the noncoding sequence and therefore in different underlying null models. In the Results section we show the importance of choosing an accurate null model. The second being that by limiting ourselves to a small fraction of the genome we test less hypotheses, reducing the extent of the multiple hypothesis testing problem.

2.2.3 A Markov chain model

The definition of CpG islands depends on the probability distributions $\mathbf{P}_n$, representing the null model. In the case of CGIs we use the genomewide mutation rates to construct these distributions. In the case of CCGIs, we choose these distributions to reflect the coding mutation rate (CMR) present in coding exons (rather then genomewide mutation rates).

Our model needs to reflect k-mer statistics on genomic regions, and it is therefore convenient to assume that sequences are generated by a (stationary) Markov chain, denoted as $(X_t)_{t\in\mathbb{N}}$. A k-th order Markov chain emitting symbols from Σ is parametrized by a transition probability matrix

$$M = (m_{uv})_{u\in\Sigma^k, v\in\Sigma} \quad \text{where} \quad m_{uv} = \mathbf{P}(u \to v) := \mathbf{P}(X_t = v \mid X_{t-k}...X_{t-1} = u)$$

and an initial probability distribution $\pi \in [0,1]^{\Sigma^k}$ that is defined as $\pi(v) = \mathbf{P}(X_1...X_k = v)$, where π is chosen such that the resulting stochastic process is stationary (in case of $k = 1$ this translates to $\pi^T M = \pi^T$). According to the rules of Markov chains one computes probabilities as

$$\mathbf{P}(v = v_1...v_n) = \mathbf{P}(X_1...X_n = v_1...v_n) = \pi_{v_1...v_k} \cdot m_{v_1...v_k,v_{k+1}} \cdot ... \cdot m_{v_{n-k}...v_{n-1},v_n}.$$

Our objective is to use this way of modeling sequences to compute $p_{n,m}$ values, and throughout this section we explain our approach to doing so. We first describe how we have chosen the order of the Markov chain to use and what the parameters of our model are. We continue by describing how we efficiently compute, for sequences generated by a k-th order Markov chain, the probability of a sequence of length n to have $\geq m$ occurrences of CG, and to end with a sequence $v \in \Sigma^k$. Lastly, we show that by using this computation and summing over all possibilities for v, we can compute $p_{n,m}$. In this section we also explain how such recursive computations are used to compute $p_{n,m}^{cg\rightarrow cg}$, which is used in the algorithm described in the next section.

In the coding regions, nucleotide sequences are organized into codons, each codon consisting of three nucleotides. Since the different codon positions (first, second and third) have different mutation rates, there are different transition rates between nucleotides at different positions in the codon [22; 142]. We therefore opted for 6-mer statistics (a 5th order Markov model), in which the entire codon of the previous amino acid is taken into account when determining the probability of transition, giving indication of the position of the codon being predicted.

We would like the parameters of our model to reflect the CMR. In our model, we would like $\pi(u)$ to be the probability that the k-mer u is randomly drawn from a coding region and m_{uv} to be the probability that in the coding regions a k-mer u is followed by the nucleotide v. The resulting Markov chain encodes the $k+1$-mer statistics of the coding regions. We denote by $\mathbf{P}_{CMR}(v)$ the probability of observing a string v in a coding region under the Markov chain model described.

2.2.4 Computing p-values

Given our assumption that the underlying probability distribution follows a 5-order Markov model, we would like to compute $p_{n,m}^{cg\rightarrow cg}$ and $p_{n,m}$ for different values of n and m. Pattern counting statistics have been extensively studied [4; 96; 113; 116; 134], and the asymptotic cases of probabilities such as (2.1,2.2) have been considered in large deviations theory. We used a dynamic programming approach to efficiently compute (see section 2.2.6 for a further discussion on efficiency) the exact probabilities of (2.1,2.2) for Markov chains of order k.

We denote the set of n-mers that end with the string v of length k by $\Sigma^{n-k}v$. Let $u \in \Sigma^n$ be a random string generated by the Markov chain (X_t), $w = CG$, and $v \in \Sigma^k$ be some fixed string.

Define

$$\pi_{n,m}(v) := \mathbf{P}_X(\{\#(u,w) \geq m\} \cap \{u = \Sigma^{n-k}v\})$$

as the probability that the Markov chain (X_t) generates a string of length n that contains at least m *CG*s and ends with the k-mer v. Define

$$\pi_{n,m} := [\pi_{n,m}(v)]_{v\in\Sigma^k}$$

to be the vector of values $\pi_{n,m}(v)$ for all $v \in \Sigma^k$.

We recall that a k-th order Markov chain (X_t) where $k > 1$ can be transformed into a 1-st order Markov chain (Y_t) which acts on the alphabet Σ^k rather than Σ by defining

$$\mathbf{P}_Y(v_1...v_k \to w_1...w_k) := \begin{cases} \mathbf{P}_X(v_1...v_k \to w_k) & v_2...v_k = w_1...w_{k-1} \\ 0 & \text{otherwise} \end{cases}. \tag{2.3}$$

This generates a transition probability matrix for the new 1-st order model by using M - the transition matrix of the k-th order model. We denote this new transition probability matrix by M_Y.

We recall that (Y_t) generates symbols from Σ^k rather than Σ. Let $w = w_1...w_k \in \Sigma^k$ and $I^w := [m_{vw}]_{v\in\Sigma^k}$ be the w-column in M_Y. We define

$$M_Y^w := (\mathbf{0},...,\mathbf{0},I^w,\mathbf{0},...,\mathbf{0}), \quad M_Y^{CG} := \sum_{w \text{ s.t. } w_{k-1}w_k = CG} M_Y^w \quad \text{and} \quad M_Y^{\neq CG} := \sum_{w \text{ s.t. } w_{k-1}w_k \neq CG} M_Y^w$$

that is, M_Y^w results from putting all entries in M_Y to zero apart from those in the w-column I^w and M_Y^{CG} ($M_Y^{\neq CG}$) results from summing over all matrices M_Y^w where w ends (does not end) with CG. Note that $M_Y = M_Y^{CG} + M_Y^{\neq CG}$. We then obtain that

$$(\pi_{n,m})^T = (\pi_{n-1,m-1})^T \cdot M_Y^{CG} + (\pi_{n-1,m})^T \cdot M_Y^{\neq CG}. \tag{2.4}$$

Given $\pi_{n,m}$, one can now compute $p_{n,m}$ by the identity $p_{n,m} := (\pi_{n,m})^T\mathbf{1}$. We can use this to compute the probability that a random string of length n generated by $\mathbf{P}_X$ has at least m CGs and ends with a CG by summing over the appropriate v instances of $\pi_{n,m}$, and we can compute $p_{n,m}^{cg\to cg}$ in the case of the CMR 5-order model by initializing with

$$\hat{\pi}_{5,m}(v) = \begin{cases} \pi_{5,m}(v) & v \in CG\Sigma^3 \\ 0 & \text{otherwise,} \end{cases}$$

and summing over the appropriate v instances of $\pi_{n,m}$.

Example

Suppose $\Sigma = \{G, C\}$ is the alphabet at hand, and that we use a 2-nd order Markov chain. Suppose that the parameters for our model are the following transition matrix and initial probability distribution:

$$M = \begin{array}{c} \\ CC \\ CG \\ GC \\ GG \end{array} \!\!\begin{array}{c} \begin{array}{cc} C & G \end{array} \\ \begin{pmatrix} 0.5 & 0.5 \\ 0.2 & 0.8 \\ 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix} \end{array} \qquad \pi = \begin{pmatrix} \pi(CC) \\ \pi(CG) \\ \pi(GC) \\ \pi(GG) \end{pmatrix} = \begin{pmatrix} 0.28 \\ 0.155 \\ 0.155 \\ 0.41 \end{pmatrix}.$$

For example, $m_{GC,C} = 0.9$. We build M_Y from M for the recursive computation using (2.3) and get:

$$M_Y = \begin{array}{c} \\ CC \\ CG \\ GC \\ GG \end{array} \!\!\begin{array}{c} \begin{array}{cccc} CC & CG & GC & GG \end{array} \\ \begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.2 & 0.8 \\ 0.9 & 0.1 & 0 & 0 \\ 0 & 0 & 0.3 & 0.7 \end{pmatrix} \end{array}.$$

We have that $(\pi_{2,0})^T = (0.28, 0, 0.155, 0.41)$ and $(\pi_{2,1})^T = (0, 0.155, 0, 0)$ and that

$$M_Y^{CG} = \begin{pmatrix} 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \text{ and } M_Y^{\neq CG} = \begin{pmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0.8 \\ 0.9 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0.7 \end{pmatrix}.$$

We can now compute $p_{n,m}$ recursively. For example,

$$(\pi_{3,1})^T = (\pi_{2,0})^T \cdot M_Y^{CG} + (\pi_{2,1})^T \cdot M_Y^{\neq CG} = (0, 0.155, 0.031, 0.124),$$

and $p_{3,1} = (\pi_{3,1})^T \mathbf{1} = 0.3105$.

2.2.5 Determining non-overlapping CCGIs: the algorithm

Our algorithm for determining CCGIs, given a set of coding exons as input, begins with the estimation of the null model $\mathbf{P}_{CMR}$. Maximum likelihood estimates for the Markov chain parameters are determined simply by counting k-mer transitions in the set of coding exons at hand. The algorithm takes as input these maximum likelihood estimates and a threshold, q, for determining significance and proceeds as follows:

1. Compute probability tables $(p_{n,m})_{1\leq n\leq N, 0\leq m\leq n/2}$ and $(p_{n,m}^{cg\to cg})_{1\leq n\leq N, 0\leq m\leq n/2}$ for a reasonably large N, the maximum possible length for an island.

2. Order all exonic substrings $\mathbf{G}_s^e$ which are bounded by CG from both ends by their tail probabilities $p^{cg\to cg}(\mathbf{G}_s^e)$. We define p_i to be the i-th smallest among these tail probabilities and $G_i = \mathbf{G}_{s_i}^{e_i}$ the corresponding substring. That is

$$\begin{aligned} p_1 = p^{cg\to cg}(G_1) \leq p_2 = p^{cg\to cg}(G_2) \leq \cdots \\ \leq p_{K-1} = p^{cg\to cg}(G_{K-1}) \leq p_K = p^{cg\to cg}(G_K) \end{aligned}$$

   where K is the number of all exonic substrings that are bounded by CGs from both ends.

3. Set a threshold $k^* = \max\{i : p_i \leq \frac{i}{K}q\}$ and choose all G_i s.t. $i \leq k^*$ to construct a set of candidate islands. Discard all other substrings from further analysis.

4. Order the set of candidate islands by their tail probabilities $p(\mathbf{G}_s^e)$, s.t. p_i is the i-th smallest among these tail probabilities and G_i is the matching candidate island.

5. Select a set of non-overlapping islands from the candidate island set by running:

```
GREEDYCCGI
CCGI $\leftarrow \emptyset$, CAND $\leftarrow \{G_i, 1 \leq i \leq k^*\}$
while CAND $\neq \emptyset$ do
    $i^* \leftarrow \underset{i}{\mathrm{argmin}}\ \{p(G_i) \,|\, G_i \in \mathrm{CAND}\}$
    CCGI $\leftarrow$ CCGI $\cup \{G_{i^*}\}$, CAND $\leftarrow$ CAND $\setminus \mathcal{N}(G_{i^*})$
Output CCGI as the set of coding CpG islands.
```

   where $\mathcal{N}(G_i)$ is the set of candidate islands that overlap G_i.

Figure 2.1 presents an overview of the algorithm.

Details about the different steps of the algorithm are below:

Step 1 [Computation of probabilities]

$\mathbf{P}_{CMR}$ is constructed from the maximum likelihood estimates, and two tables are constructed using the dynamic programming approach described in section 2.2.3.
We observe that, using the notation as in the previous section,

$$\begin{aligned}(\pi_{n,m})^T - (\pi_{n-1,m})^T &= [(\pi_{n-1,m-1})^T - (\pi_{n-2,m-1})^T] \cdot M_Y^{CG} \\ &\quad +[(\pi_{n-1,m})^T - (\pi_{n-2,m})^T] \cdot M_Y^{\neq CG}\end{aligned}$$

which by induction on n and m gives the intuitive result

$$\pi_{n,m} \geq \pi_{n-1,m} \quad \text{for all} \quad v \in \Sigma^k \qquad \text{and therefore} \qquad p_{n,m} \geq p_{n-1,m} \quad \forall n,m \geq 0.$$

Lemma 1. *Let $v \in \Sigma^n, w \in \Sigma^k, k \leq n$ be two sequences with equal content of CGs, that is $m := \#(v,CG) = \#(w,CG)$. Then $p_{k,m} \leq p_{n,m}$ and $p_{k,m}^{cg \to cg} \leq p_{n,m}^{cg \to cg}$, i.e., the shorter one is more significant.*

Step 2 [Ordering of p-values]

$p_{n,m}^{cg \to cg}$ values are computed for each substring within an exon that begins and ends with a CG, setting the stage for step 3. We can restrict ourselves to those substrings because Lemma 1 guarantees that only such substrings will be selected in step 5 of the algorithm. This results in a large decrease in the number of hypotheses tested, significantly reducing both the multiple hypothesis testing problem and the algorithm's running time. We make use here of $p_{n,m}^{cg \to cg}$ because when applying an FDR correction to a set of hypothesis tests the distribution of the p-values for the hypotheses generated by the null model needs to be uniform.

Step 3 [Benjamini-Hochberg correction]

In this step a set of candidate islands is chosen by using the Benjamini-Hochberg correction [6; 8] with the significance threshold q. This guarantees that under our null model the expected proportion of false-positives in our set of candidate islands is less than q. Note that our hypotheses are dependent, for example in the case of overlap between substrings. For our case the positive regression dependency condition required for multiple testing under dependency [8] translates to that CG content in substrings which overlap with one another is positively correlated—which is an obvious observation.

Step 4 [Reordering of p-values]

The candidate islands are sorted by their p-values, not conditioning on having a CG at the beginning and end (which solely served to ensure uniformity of the p-value distribution for step 3). This step is a preparatory for step 5.

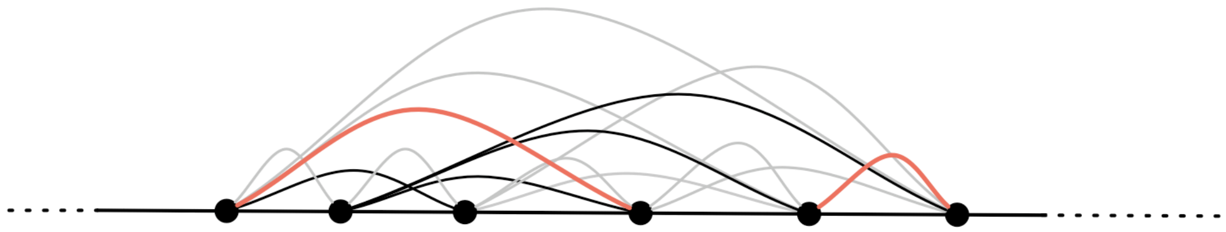


Figure 2.1: Illustration of the result of the CCGI algorithm. Nodes marked along the genome correspond to CpG sites. An edge between every two nodes corresponds to an interval, and represents a possible CCGI. A p-value is associated with every edge (computed by dynamic programming). Light grey edges are intervals with p-values that do not pass the FDR threshold. Red edges denote non-overlapping intervals chosen by the algorithm to maximize the product of the p-values subject to the constraints in Theorem 1 .

Step 5 [Greedy algorithm]

The algorithm GREEDYCCGI receives a collection of exonic substrings $\mathbf{G}_s^e$ and iteratively selects regions $\mathbf{G}_s^e$ with minimal $p(\mathbf{G}_s^e) = p_{n,m}$ (where $n = e - s + 1$ and $m = \#(\mathbf{G}_s^e, CG)$), removing from the candidate set any substrings overlapping the chosen one. Theorem 1 shows that the output is of maximal significance under a biologically reasonable constraint. Before we present the theorem, we give some notation.

Let $\mathbf{H}$ be a superstring (in the following an exon) and $\{H_l\}_{l=1}^L$ be a set of non-overlapping substrings of $\mathbf{H}$, ordered by position in $\mathbf{H}$: $H_1 = \mathbf{H}_{s_1}^{e_1} < H_2 = \mathbf{H}_{s_2}^{e_2} < \cdots < H_L = \mathbf{H}_{s_L}^{e_L}$, where ordering translates to $s_1 \leq e_1 < s_2 \leq e_2 < \cdots < s_L \leq e_L$. Let $S_1 < S_2 < \cdots < S_L < S_{L+1}$ be the strings in between, that is

$$S_1 \cdot H_1 \cdot S_2 \cdot H_2 \cdots S_l \cdot H_l \cdot S_{l+1} \cdots S_L \cdot H_L \cdot S_{L+1} = \mathbf{H}$$

is the entire superstring. Based on this notation, we write

$$\begin{array}{rrcl} & \overline{H_l H_k} & := & H_l \cdot S_{l+1} \cdot H_{l+1} \cdots S_k \cdot H_k \\ \text{and} & \overleftrightarrow{H_l H_k} & := & S_l \cdot \overline{H_l H_k} \cdot S_{k+1}. \end{array}$$

Theorem 1. *Let $\{G_i\}_{i=1}^K$ be a set of possibly overlapping substrings of a superstring $\mathbf{G}$ and $\mathbf{S}(G_i) := \log p(G_i)$ be a score for a substring G_i. Applying* GREEDYCCGI *to $\{G_i\}_{i=1}^K$ selects L non-overlapping substrings $\{H_l\}_{l=1}^L$ from among the substrings $\{G_i\}_{i=1}^K$ such that $\sum_{l=1}^L \mathbf{S}(H_l)$ is minimized among all choices of L non-overlapping substrings, subject to the constraints*

$$S(G) \geq \min_{l \leq j \leq k} \mathbf{S}(H_j) \text{ for all } 1 \leq l,k \leq L, \text{ and } G \subset \overleftrightarrow{H_l H_k}. \tag{2.5}$$

In the setting of this chapter, the constraints in (2.5) translate to the property that the region between any two CpG islands H_l, H_k does not contain a sub-region with a p-value better than the

p-values of the islands in between H_l, H_k.

Proof. By design of GREEDYCCGI, its output clearly satisfies the constraints. Assume that there is a set of non-overlapping substrings $\{J_l\}_{l=1}^{L}$ which satisfy the constraints in (2.5) such that

$$\sum_{l=1}^{L} \mathbf{S}(J_l) < \sum_{l=1}^{L} \mathbf{S}(H_l)$$

where $\{H_l\}_{l=1}^{L}$ are the islands returned by the GREEDYCCGI. Let $\mathbf{S}^* = \min_i\{\mathbf{S}(G_i)\}$, and $W = \{G_i \mid \mathbf{S}(G_i) = \mathbf{S}^*\}$ be the set of substrings with the lowest score. We prove the result by induction on L. We show in Lemma 2 that there are no erroneous overlaps among the islands of W [1]. We distinguish between the following two cases:

- $L > |W|$. In this case $\{H_l\}_{l=1}^{L}$ will contain all substrings of W.
    - If one of the substrings from $\{J_l\}_{l=1}^{L}$ is in W, upon removal of that substring from $\{J_l\}_{l=1}^{L}$ and $\{G_i\}_{i=1}^{K}$, and removal of their overlapping neighbors from $\{G_i\}_{i=1}^{K}$ we obtain a new set of substrings $\{\tilde{G}_i\}_{i=1}^{\tilde{K}}$. By induction, GREEDYCCGI will return a set of islands from $\{\tilde{G}_i\}_{i=1}^{\tilde{K}}$ which is optimal, resulting in a contradiction.
    - If non of the substrings from $\{J_l\}_{l=1}^{L}$ are in W, let $\hat{J}$ be some island in $\{J_l\}_{l=1}^{L}$. When considering the entire superstring $\mathbf{G}$ condition (2.5) is violated w.r.t. $\hat{J}$, in contradiction.
- $L \leq |W|$. $\{H_l\}_{l=1}^{L}$ consists entirely of elements of W, achieving the minimum score.

□

Lemma 2. *Let $\{H_i\}_{i=1}^{K}$ be a set of substrings of a set of superstrings $\mathbf{H}$, $p^* = \min_i p(H_i)$ and $W = \{H_i \mid p(H_i) = p^*\}$. Then there are no two islands in W that overlap and are not contained in each other, i.e., $H_{s_1}^{e_1}, H_{s_2}^{e_2}$ s.t. $s_1 < s_2 < e_1 < e_2$.*

Proof. Assume that there are two islands $H_{s_1}^{e_1}, H_{s_2}^{e_2} \in W$ s.t. $s_1 < s_2 < e_1 < e_2$. We will show that in this case $p(H_{s_1}^{e_2}) < p(H_{s_1}^{e_1}) = p(H_{s_2}^{e_2})$, in contradiction.

Let n_1, n_2 and n_3 be the lengths of $H_{s_1}^{e_1}$, $H_{s_2}^{e_2}$ and $H_{s_1}^{e_2}$, respectively, and let f_s^e be the frequency of CG sites (the number of CGs divided by $e - s + 1$) in the substring starting at s and ending at e, inclusive. Suppose w.l.g. that $n_1 \leq n_2$. Since the two substrings have the same p-values, $f_{s_2}^{e_2} \leq f_{s_1}^{e_1}$ (a shorter substring must have a higher frequency of CGs to reach the same significance), and therefore $f_{e_1+1}^{e_2} \leq f_{s_1}^{s_2-1}$. By concatenating both $H_{s_1}^{s_2-1}$ and $H_{e_1+1}^{e_2}$ to $H_{s_2}^{e_1}$ we concatenate two strings to $H_{s_2}^{e_1}$ which together are longer and have more (or equal) frequency of CGs than concatenating just $H_{e_1+1}^{e_2}$. Therefore, since $f_{s_2}^{e_2} \leq f_{s_1}^{e_2}$ and $n_2 < n_3$, $p(H_{s_1}^{e_2}) < p(H_{s_2}^{e_2})$, in contradiction.

□

[1] For any two islands in W that overlap, one island is strictly contained in the other. In case the algorithm encounters overlaps, it chooses the shorter island and it can be easily seen that the proof holds for such cases.

2.2.6 Running time and implementation

We compute two tables holding values of $p_{n,m}$ and $p_{n,m}^{cg\to cg}$ for $1 \le n \le N := 10000$ and $0 \le m \le N/2$ where N is an upper bound on the maximum length of a CpG island in an exon. Using dynamic programming in the implementation of (2.4), the running time is $O(|\Sigma|^{k+1} \cdot N^2)$ for alphabets Σ and model order k in general where here $\Sigma = \{A,C,G,T\}$ and $k = 5$. Note that each recursive iteration $(\pi_{n,m})^T = (\pi_{n-1,m-1})^T \cdot M_Y^{CG} + (\pi_{n-1,m})^T \cdot M_Y^{\neq CG}$, due to the sparsity of M_Y^{CG} and $M_Y^{\neq CG}$, requires only $O(|\Sigma| \cdot |\Sigma|^k)$ runtime. Storage of tables requires $O(N^2)$ space. Note that [115] suggests a routine for fast computation of a single $p_{n,m}$ which requires $O(|\Sigma| \cdot L \cdot n \cdot m)$ where $L = O(|\Sigma|^k)$ depends on the model order k which coincides with the runtime of our approach.

A different algorithm presented in [138] requires a runtime of $O((|\Sigma|^k)^2 \cdot N \log N)$ which yields advantages for small model orders in particular. In our case, due to evaluating powers of matrices which yields the factor $\log N$, one would have to call the respective routine $O(N)$ times, yielding an overall runtime of $O((|\Sigma|^k)^2 \cdot N^2 \log N)$ which equally establishes no advantage.

Subsequently, we traverse each coding region and assign both $p_{n,m}$ and $p_{n,m}^{cg\to cg}$ to each substring which is bounded by CGs. In human, this results in $10.8M$ substrings with $p_{n,m}^{cg\to cg} \le 0.01$. In order to allow a rapid FDR treatment, we designed and implemented a routine which yields an estimation of the FDR threshold that is stricter than the exact FDR (guaranteeing our expected FDR to be less than or equal to the specified α). The estimated threshold is obtained by storing for each substring G_i, $\lfloor \log p^{cg\to cg}(G_i) \rfloor$, assigning

$$i^* := \underset{i}{\operatorname{argmax}} \{\lfloor \log p^{cg\to cg}(G_i) \rfloor + 1 \le \log(i\frac{\alpha}{K})\},$$

and declaring any G_i with $\log p^{cg\to cg}(G_i) < \lfloor \log p^{cg\to cg}(G_{i^*}) \rfloor$ as a candidate CCGI. Since $\lfloor \log(p_i) \rfloor \le \log(p_i) \le \lfloor \log(p_i) \rfloor + 1$, we are guaranteed to be taking a threshold which is at least as stringent as that attained from the Benjamini-Hochberg procedure. Then, the greedy algorithm picks a final list of CCGIs from the candidates.

The methods were implemented in C and the running time on a standard MacPro desktop was 7 hours for generating the p-value tables and 30 minutes for finding the CCGIs for all of the coding exons in the human genome. The software is available upon request from the authors.
Our coding CpG islands Genome Browser track for hg18 is available at

`http://bio.math.berkeley.edu/CCGI/.`

2.3 Results

2.3.1 The *p*-value statistic

We began by investigating whether the *p*-values are indeed a good measure for determining the extent of unmethylated CpGs, by comparing the *p*-value measure to the epigenetic score as defined in [17] and to the observed/expected number of CpGs used in [50] ($\frac{\#CpG}{\#C \times \#G} \times N$, where N is the length of the island).

Because genome-wide methylation has not yet been characterized for a significant number of developmental stages or cell types, we used two different datasets that are believed to be correlated with functional unmethylated regions to determine our true-positive sets (whether a region is unmethylated or not). The first dataset consisted of regions experimentally determined by the ENCODE project as open chromatin regions in non-malignant cell types [2] [52]. Overlap of an examined island with an open-chromatin region determined the island at hand as functional. The second dataset consisted of genomewide site-specific methylation measurements for two cell types [94]. A methylation score of between 0 and 1 was computed for a region as the mean of the methylation scores of the CpG sites within it, and the region was considered as unmethylated if the score was ≤ 0.3 (this is an arbitrary threshold but since the island scores are well partitioned it does not affect our results). Only sites with coverage of at least 8 reads and islands with at least 5 such sites were considered. An island was considered functional if it was unmethylated in both cell types.

To test whether our *p*-value statistic is informative we computed the *p*-values for the set of bona fide CpG islands annotated in [17] and compared how well they indicate functionality compared to the islands' epigenetic scores and the islands' observed/expected scores. We used the set of bone fide islands for this comparison because they are annotated along with scores, allowing us to compare the ordering of the assigned *p*-values to a different scoring method.

The *p*-values were computed using a first order Markov model estimated from statistics on the *entire genome* (and not only on coding regions). A receiver operating characteristics (ROC) curve for the *p*-values, the epigenetic scores and the observed/expected score, using the ENCODE open chromatin data to asses functionality (Figure 2.2.A), shows that the *p*-value statistic achieves better overall performance - area under curve (AUC) of 0.76 - than the two other scores (AUC of 0.73 and 0.754 for epigenetic score and observed/expected score, respectively). When using the methylation scores from [94] to asses functionality (Figure 2.2.B), we again observe better overall performance of the *p*-values: AUC of 0.865 as opposed to 0.852 and 0.844 for the epigenetic score and observed/expected scores, respectively. This is surprising since the epigenetic scores incorporate functional information such as the presence of promoter activity (not from the ENCODE project), whereas our *p*-values are computed from sequence alone.

[2] One file of open chromatin was compiled from: ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeChromatinMap/ using the files: wgEncodeUncFAIREseqPeaks{H1hesc,Nhek,Gm12878V2,Huvec,Panislets}.narrowPeak

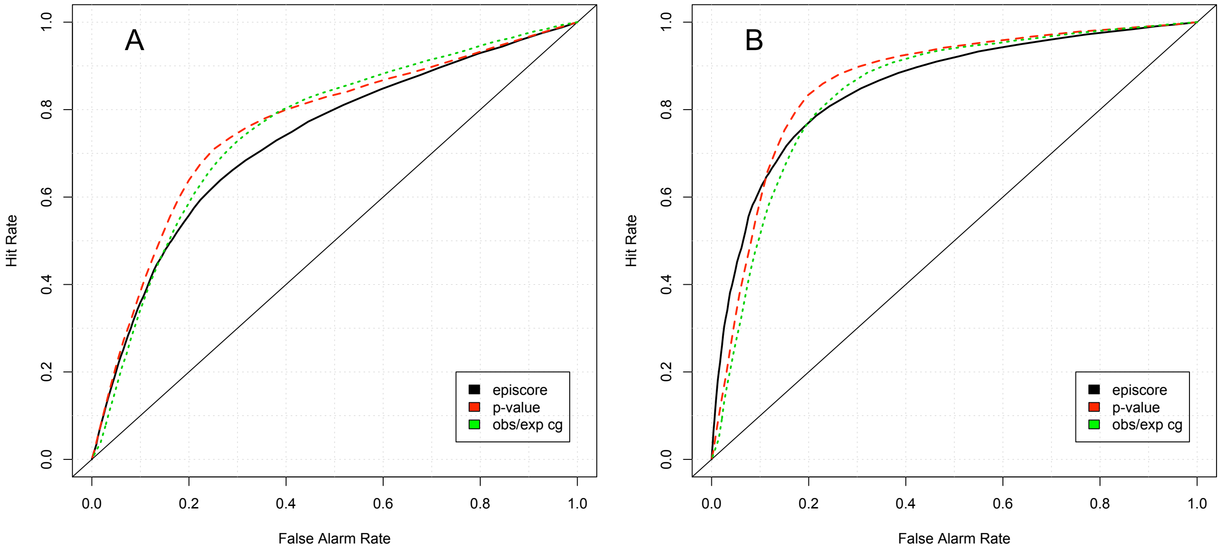


Figure 2.2: ROC plots showing correlation of different scoring schemes (epigenetic score, p-value, and observed over expected CG) with (A) FAIRE-Seq open chromatin regions and (B) low methylation in H1 and IMR90 cell types.

2.3.2 Choice of null model

In order to asses the importance of the null model choice on the p-value scores we examined the effect of estimating the null model parameters using only coding sequence as opposed to the entire genome sequence. We learned parameters for a 5-th order Markov chain for the two types of sequences, and calculated the log ratio of the p-values from the two models for the $194,586$ unique, but possibly overlapping, coding exons from [128] (Figure 2.3). For all exons the p-value from the genome null model was lower than that from the coding null model. This was to be expected, given the higher frequency of guanine and cytosine residues in coding regions.

We then tested whether the difference in the p-values has an effect on the set of coding exons considered significant. At an FDR threshold of 0.05, 54,596 coding exons were determined significant when using the genome model, while 9,639 were determined significant when using the coding model. At an FDR of 0.01, 37,597 and 7,050 coding exons were determined as significant for the genome model and coding model, respectively. The drastic reduction in the number of coding exons predicted as significant with respect to their CpG content shows the importance of the specified null distribution.

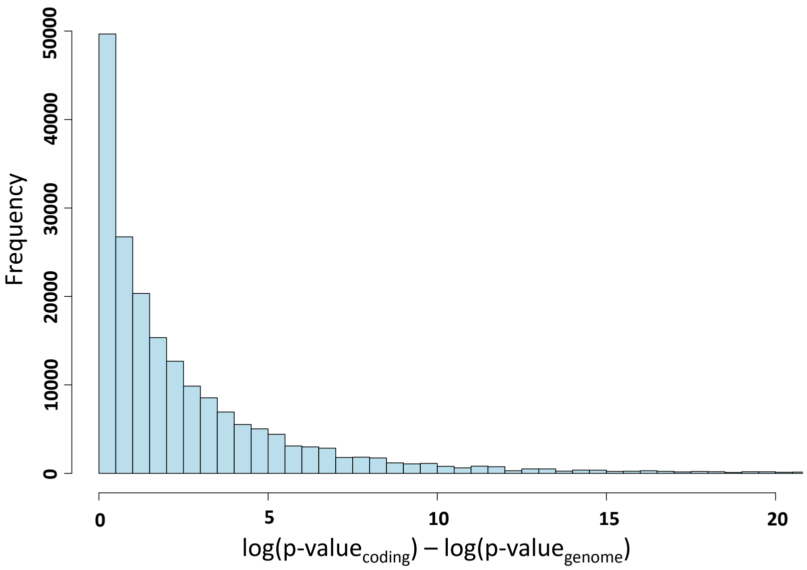


Figure 2.3: Fold change in $\log p$-value for exons when computed using a *genome* null model versus a *coding* null model. 82,505 exons have p-values that differ by more than a factor of 2.

2.3.3 Coding CpG islands

Using our algorithm on the coding regions of the human genome with FDR threshold 0.01, we determined a Coding CpG island (CCGI) set for the human coding exons, that consisted of 12,445 islands.

As a first validation of the sensitivity of the CCGIs we examined a region rich in HOX genes in chr7 (Figure 2.4.a). This region was studied in [22] where it was shown that CpGs in protein coding exons of HOX genes are subject to pro-epigenetic selection. Our results not only define CpG islands in every one of these HOX genes (with the exception of HOXA10 in which the $5'$ exon consists of only 2 amino acids), but also provide quantitative information about the extent of over-representation of CpGs via p-values.

We compared our CCGIs to two alternative CpG island sets: the Gardiner-Garden-Frommer CpG island set [50] available from the UCSC genome browser (GFCGIs) and islands predicted by the method in [74] (HMMCGIs). We chose to compare to GFCGIs, because despite its use of arbitrary thresholds, this set of islands is very popular. The HMMCGIs represent recent work applying state-of-the-art statistical methods to CpG island prediction. Testing the specificity of the CCGI set is not trivial because one cannot validate that a region determined as a CCGI is not unmethylated and functional in some developmental state or cell type. We therefore chose to test the extent to which the CCGIs were enriched for known functional regions (Table 2.2).

	Total	Open Chromatin	17-Cons Track	First Exon	Near A-TSS
CCGIs	12445	2734(0.21)	11041(0.88)	5539(0.44)	7248(0.58)
CCGIs \ GFCGIs	3000	189(0.06)	2687(0.89)	433(0.14)	1248(0.41)
CCGIs \ HMMCGIs	802	25 (0.03)	706 (0.87)	82 (0.10)	286 (0.35)

Table 2.2: Overlap of CCGI sets with different functional regions. The proportion for each computation is in parentheses. “Open Chromatin” - The same regions used in section 2.3.1. “17-Cons Track” - Regions of the UCSC 17-way Conservation track. “First Exon” - First exons of RefSeq genes. “Near A-TSS” - Regions of ±500 bps from alternative transcription start sites [168].

	Total	H1 (r1)	H1 (r2)	IMR90 (r1)	IMR90 (r2)
CCGIs	12445	0.075 (73/961)	0.108 (439/4047)	0.168 (663/3912)	0.164 (565/3429)
GFCGIs in exons	1143	0.053 (14/262)	0.052 (43/817)	0.072 (60/826)	0.065 (48/737)
HMMCGIs in exons	3617	0.029 (26/895)	0.023 (55/2312)	0.039 (86/2177)	0.037 (77/2066)

Table 2.3: Proportion of islands completely within exons determined as unmethylated from the data in [94]. The absolute counts in parentheses are the number of unmethylated islands over the total number of islands for which there was sufficient data do determine the methylation state.

In a comparison to the GFCGIs, we found 3,000 CCGIs that did not overlap any GFCGI. In contrast, of the 1,143 CGIs that were completely within some exon (and therefore could have been detected by our algorithm) only 34 did not overlap a CCGI. Although many of our CCGIs are located in the first exon of genes (44%), of the set not overlapping GFCGIs we found only 14% in first exons. This suggests that the GFCGIs are missing coding CpG islands in exons not located adjacent to large CpG islands in promoters. To better understand why the GFCGIs were missing in many exons, we examined the properties of the CCGIs that they did not overlap. We found that these CCGIs were mostly shorter than the 200bp cutoff for GFCGIs (2,689 out of the 3,000). Moreover, the constant threshold for accepting an island as a GFCGI (#CpG observed / expected > 0.6) implicitly leads to more stringent requirements for longer regions to be declared islands, and can lead to a higher rate of false-negatives for the longer lengths than for the shorter lengths (for example, an island of length 1,500 bps with #CpG observed / expected ratio of 0.5 is much more likely to be an unmethylated island than a 200 bps island with the same ratio, but the constant threshold does not account for this). To verify this, we calculated the correlation between the log of the p-values of GFCGIs and their length, and found it to be -0.78 (with a fitted line of slope -5.164).

The CCGIs we found not overlapping GFCGIs did mostly overlap HMMCGIs. In fact, of our 12,445 CCGI predictions, only 802 did not overlap HMMCGIs. However the comprehensive coverage of HMMCGIs comes at a price of specificity. We examined the set of 1,636 HMMCGIs that are completely within some exon (and therefore could have been predicted by our method) but do not overlap with a CCGI, and found that their average p-value was 0.105. This surprisingly large number suggests that many of these islands are being predicted on the basis of an incorrect null model.

Differential Methylation. In order to further investigate the characteristics of the CCGIs we used the dataset from [94], in which methylation was measured for two different cell types - H1 (human embryonic stem cells) and IMR90 (fetal lung fibroblasts). We computed a methylation score for each CCGI and determined unmethylated CCGIs in the same manner as described in section 2.3.1. In all experiments (two replicated for each cell type) the large majority of the CCGIs was found to be methylated (Table 2.3). To validate that this is a general phenomenon for CpG islands inside exons we examined the GFCGIs and HMMCGIs that were completely within exons and found that the large majority of those CpG islands were methylated (Table 2.3). This supports the hypothesis that CpG islands present within exons play an important role in tissue specific regulatory mechanisms. Interestingly, in all island sets tested, the percentage of unmethylated CpG islands in IMR90 was consistently larger than the percent of unmethylated CpG islands in H1 cells, suggesting that in the stages *after* differentiation the intragenic CpG islands have a greater regulatory role. We compared the methylation status from the two different cell types (using replicate 2 of H1 and replicate 1 from IMR90 since they produced data for a larger number of islands) and found that of the 2,956 CCGIs for which there was sufficient data in both experiments, 249 islands were unmethylated in both samples and 103 were differentially methylated (38 were unmethylated

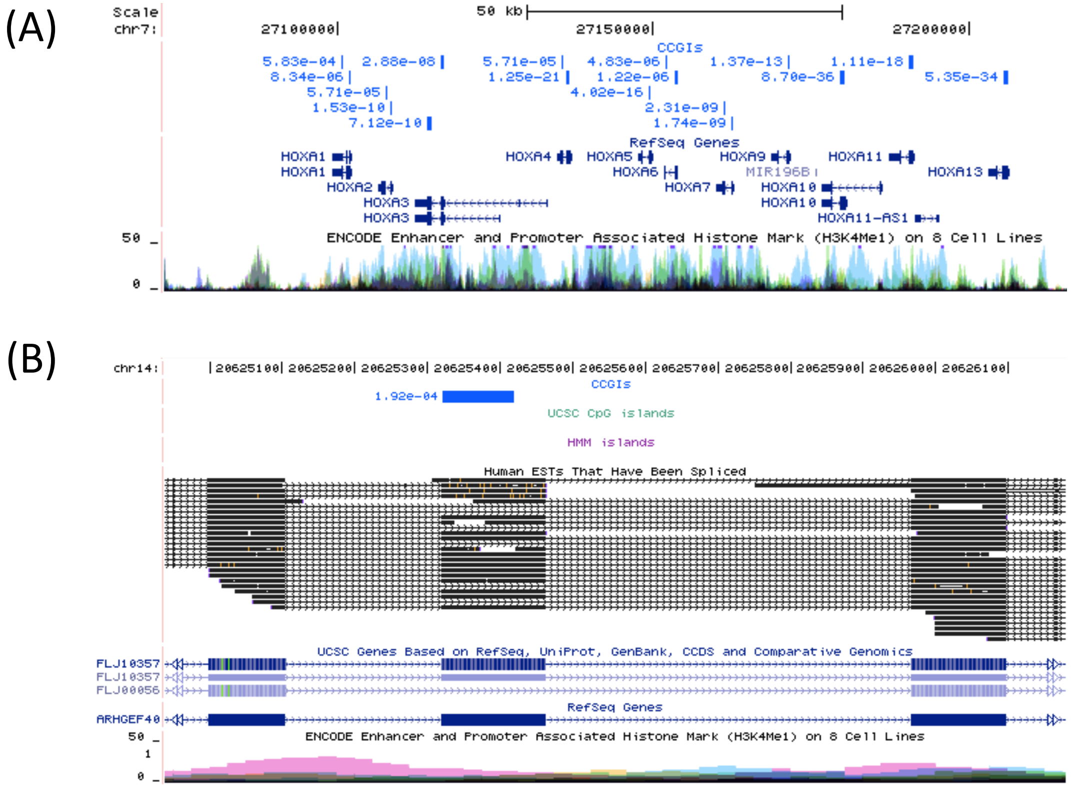


Figure 2.4: UCSC genome browser screen shots. (A) A HOX gene cluster showing our CCGI predictions. (B) A differentially methylated CCGI on an alternatively spliced exon was not detected as a GFCGI or a HMMCGI. Histone marks associated with promoters and enhances (bottom track) are seen mostly in the NHLF cells (seen in pink, normal human lung fibroblasts). This aligns well with the CCGI being unmethylated in IMR90 cells and methylated in H1 cells.

in H1 and 65 were unmethylated in IMR90). Figure 2.4.b presents an interesting example of such a CCGI which was not detected as a GFCGI or an HMMCGI.

2.4 Conclusions

We have presented a novel method for annotating CpG islands in coding regions. Our method is based on a simple, statistically natural, criterion: the statistical significance of the CpG content of a genomic region. Based on this idea, we have developed an algorithm which provably optimizes a biologically motivated criterion for selecting islands while controlling the false discovery rate.

We found that we can reduce the number of false positives in existing annotations while discovering previously undetected coding CpG islands which are likely to contain unmethylated CpGs.

Chapter 3

Bayesian networks for methyl-Seq analysis

3.1 Introduction

New methods that use high-throughput sequencing to assay epigenetic modifications at whole genome scale reveal insights into cell differentiation, development and gene expression regulation [5; 18; 10; 25; 42; 71; 76; 94; 103; 117; 119; 130; 156; 172; 174; 180]. Moreover, incorporation of genome-scale epigenetic data into case-control studies is now becoming feasible, and has the potential to be a powerful tool in the study of disease [132; 165]. Recent evidence has suggested that epigenetic variation is heritable, and may underlie phenotypic variation in humans [97; 99; 106]. Such comparative studies rely both on the ability to obtain genome-scale epigenomic information cheaply and efficiently, and on the availability of methods for analysis of the data produced.

High-throughput sequencing technologies have catalyzed the development of new methods for measuring DNA methylation, enabling the study of this phenomenon on genome-wide scale. These methods can be broadly classified as *methyltyping* versus *methylome sequencing*, in analogy with *genotyping* versus *genome sequencing* for DNA (see section 1.4.1 for further discussion). Methyltyping technologies allow for the assessment of genome-scale methylation patterns, while emphasizing low cost at the expense of high resolution. In contrast, whole-genome bisulfite sequencing offers the ability to measure absolute levels of DNA methylation at single-nucleotide resolution [34; 93; 94], but it is expensive because it requires sequencing of whole genomes. A recent analysis (Table 2 in [84]) suggested that methyl-Seq is the method with the most favorable profile of pros and cons, with respect to the measures chosen for comparison.

methyl-Seq is a convenient methyltyping strategy because it is cost-effective, requires only small amounts of material, and avoids bisulfite conversion. However, although the experiment is relatively simple, interpretation of the sequencing data is confounded by the dependence of read depth at a given site on the methylation status of neighboring sites. This has limited the use of methyl-Seq and previous studies either pointed out the need for a method of site-specific normalization [25], or attempted to deal with the bias by removing problematic sites from the analysis [5] (resulting in the loss from the analysis of more than 19% of the potentially informative sites in CpG islands, see Methods).

In order to make effective use of methyl-Seq for genome-scale methyltyping we designed a Bayesian network that models the process by which the reads are generated from the experiment, incorporating biases inherent in methyl-Seq experiments. We could then make use of this model to infer the methylation states at individual CpG sites. We have implemented this method in a freely available software package, called MetMap. An additional important feature of MetMap is the annotation of strongly unmethylated islands (SUMIs) which, as opposed to the current definition of CpG islands, incorporate information from both a reference sequence and genome-scale methylation data. We have validated MetMap's site-specific analysis, as well as its unmethylated-island annotation, with bisulfite sequencing of specific regions.

We demonstrate the use of methyl-Seq with MetMap by methyltyping neutrophils from four male human individuals, and annotating their unmethylated islands. We show that the picture revealed by such analysis is sufficient to survey methylation states across the genome. Such analysis gives significant insight into the methylome of each specimen, inside and outside of CpG islands, at site specific resolution. MetMap identifies numerous unmethylated regions, of varying lengths, which have not previously been annotated as CpG islands and are associated with other features indicative of transcriptional function. We conclude that our approach leverages the cost-efficiency and practical ease of methyl-Seq to produce informative genome-scale methylation annotations (methyltypes) that are suitable for both region- and site-specific comparative studies.

This chapter is organized as follows. We begin by examining the methyl-Seq method in detail in section 3.2, focusing on significant biases that are specific to the method. In Section 3.3 we introduce a Bayesian network for the analysis of methyl-Seq data. A recurring theme is the interplay between the model used to glean information from the technology, and the view of methylation that drives the model specification. In section 3.4 we describe the implementation of our model in a the MetMap software package and in section 3.5 we describe the validation of MetMap's procedure, using the methyltypes of neutrophils from four human individuals. In section 3.6 we discuss the contributions of our introduced procedure and the implications of the novel findings presented of methylation states in the human neutrophil, and in section 3.7 we give further details of the methods used throughout this chapter.

3.2 The methyl-Seq method

In this section we describe the methyl-Seq experiment and the experimental bias it introduces. An outline of the methyl-Seq experiment is given in figure 3.1. In this experiment, the genomic DNA is digested with a methylation-sensitive restriction enzyme, in our case *HpaII*. *HpaII* digests at CCGG sites in which the second cytosine is unmethylated. This step obtains a digest of DNA fragments such that at all of the fragment ends there is an unmethylated *HpaII* site[1] and all *HpaII*

[1] We use *HpaII* site and CCGG site interchangeably throughout this thesis.

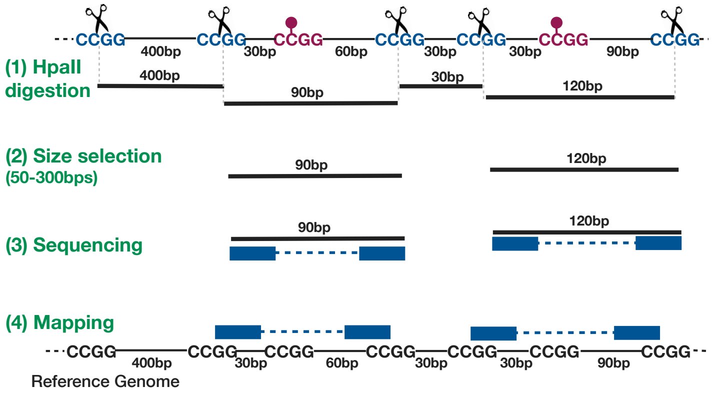


Figure 3.1: An outline of a methyl-Seq experiment. Genomic DNA is digested with *HpaII*, a methylation-sensitive restriction enzyme. Unmethylated *HpaII* sites (in blue) are digested and thus found at the ends of restriction fragments, while methylated *HpaII* sites (in red) are not digested. Restriction fragments are size-selected according to the Illumina protocol; fragments that are either too long or too short are removed. Fragments that pass the size-selection are sequenced, producing paired-end reads from each fragment. Finally, the reads are aligned against the reference genome by a software of choice.

sites within a fragment (not including the ends) are methylated (assuming complete digestion). Then, the fragments are size-selected using a run on an agarose gel, where the common length accepted is between 50 and 300 bps. This size selection is important to achieve good sequencing throughput when using Illumina machines. A library is constructed from the fragments that pass the size selection step and, followed by a PCR amplification step, the ends of the fragments are sequenced. The paired sequenced reads are then mapped to a reference genome using an aligner such as Bowtie [85]. The library constructed in this experiment can be either paired- or single-end. Notice that when constructing a paired-end library the sequencing experiment returns pairs of sequences, where each pair is the product of sequencing the ends of one fragment. Using this protocol one can conclude which *fragments* of DNA where present in the digest. Throughout this chapter we will assume the construction of a paired-end library, and will describe how we deal with single-end sequencing in section 3.4.

After the digestion step has completed, all unmethylated *HpaII* sites are present at the ends of digested fragments (Figure 3.1), but the size selection step required by the sequencing protocol limits sequencing to fragments of a narrow size range. This results in many cases in which one cannot determine the extent to which a site is methylated based on its read counts alone. Figure 3.2 shows three different scenarios of epialleles, and the fragments generated by them that pass the size selection step and are sequenced. When determining the methylation state of the highlighted site, cases (2) and (3) are indistinguishable if one considers only sequenced fragments originating at that site: in both cases, there are no fragments sequenced that have the highlighted site at either end.

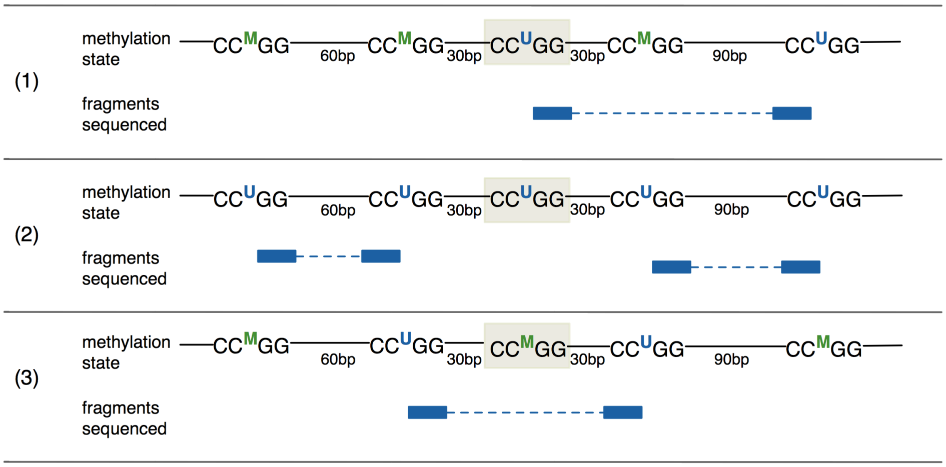


Figure 3.2: In many cases the methylation state of a site cannot be determined from the extent to which it was present at the end of sequenced fragments, but can be determined by integrating sequencing data from its neighborhood.

However, in case (2) we see that the sites on both sides of the highlighted site are unmethylated, and therefore if the highlighted site was methylated we would have sequenced a fragment of length 60 bps in which it was in the middle (as sequenced in case (3)). Since such a fragment was not sequenced, we can conclude that the site is unmethylated. In case (3), since the highlighted site was present in the interior of a fragment, we can conclude that it was methylated. While the examples in this figure assume binary methylation states, when relating to a population of cells, methylation states are assumed to be continuous variables, because the methylation states can be heterogeneous even within a population of a single cell type [5; 180]. It can easily be seen that the examples above can be generalized to the case of continuous variables, where one cannot determine the extent to which a site is methylated in a sample given the read counts at that site alone, but can do so when making use of the fragments generated from the site's neighborhood.

3.3 A Bayesian network for site-specific inference

We have seen in the previous section that while methyl-Seq is a favorable method for methyltyping due to its cost-effectivness and simplicity, the bias present in the method needs to be corrected for. We recall that the main bias present is due to the non-random digestion that is followed by a size selection step. When performing a methyl-Seq experiment, we would like to gain knowledge of the extent to which each *HpaII* site is methylated. More precisely, we would like to know for each *HpaII* site the proportion of cells in the digest that were methylated at that site. What we observe however is the number of times each fragment was sequenced in the experiment. Using our knowledge of the experimental procedure, we take a generative approach to model the procedure by which the methylation states of the *HpaII* sites impact the number of times each fragment is sequenced. On the basis of this generative model we can infer the expected methylation states of the *HpaII* sites, given the observed fragment counts.

In this section we discuss how the methyl-Seq experiment can be modeled as a Bayesian network, and how that Bayesian network can be used to infer the methylation states of the *HpaII* sites given the fragment counts. We begin by introducing a generative model for our data. We continue by describing a Bayesian network that models bias in the methyl-Seq setting; we explain how prior knowledge of the behavior of DNA methylation in mammals can be incorporated into the model through the addition of random variables and dependencies. We then explain how the parameters of such a model can be learned using the expectation maximization (EM) algorithm and describe how the learned parameters together with the observed fragment counts can be used to infer the methylation states of individual *HpaII* sites.

3.3.1 Notation

In this chapter we will denote random variables by uppercase letters, and a specific assignment to a random variable by the corresponding lowercase letter. For example, three coin tosses can be modeled by the random variables X_1,X_2 and X_3 for the first, second and third toss, respectively, and an assignment to X_1 is denoted by x_1. We will denote sets of random variables by bold uppercase letters, and an assignment to the set of random variables by the corresponding bold lowercase letter. For example, we would denote by $\mathbf{X}$ the three random variables X_1,X_2 and X_3, and by $\mathbf{x}$ an assignment to all three variables.

3.3.2 A generative model

In this section we describe a generative model for the methyl-Seq experiment. Given a reference genome, let F be the set of genomic sequences that have a *HpaII* site at each end. Notice that these regions may also have *HpaII* sites within them. We note that although theoretically F is the set of fragments that may be present in the digest of a methyl-Seq experiment, some fragments have probability of being sequenced which is essentially zero, like fragments spanning whole chromosomes, but we disregard that now to ease notation. For every *HpaII* site we assign a random variable $\{Y_i\}_{i=1,\dots,N} \in [0,1]$, where N is the number of *HpaII* sites in the reference genome. Y_i represents the proportion of cells from the digest in which site i was methylated.

In our generative model, we assume that at each step a cell is drawn at random, and the methylation state of each *HpaII* site is determined as 1 (methylated) or 0 (unmethylated) using Bernoulli draws with probability y_i (in the case of diploid genomes this sampling is done twice, once for each of the chromosomes). The methylation assignment to all *HpaII* sites of the chromosome determines what fragments will be generated from this cell. Let $\{Z_i\}_{i=1,\dots,|F|} \in [0,1]$ be random variables for the proportion of instances from which fragment f_i was generated. In other words, let $\{Z_i\}_{i=1,\dots,|F|}$ be random variables for the proportion of times that the methylation configuration of the *HpaII* sites of a chromosome generated fragment f_i. In this generative approach,

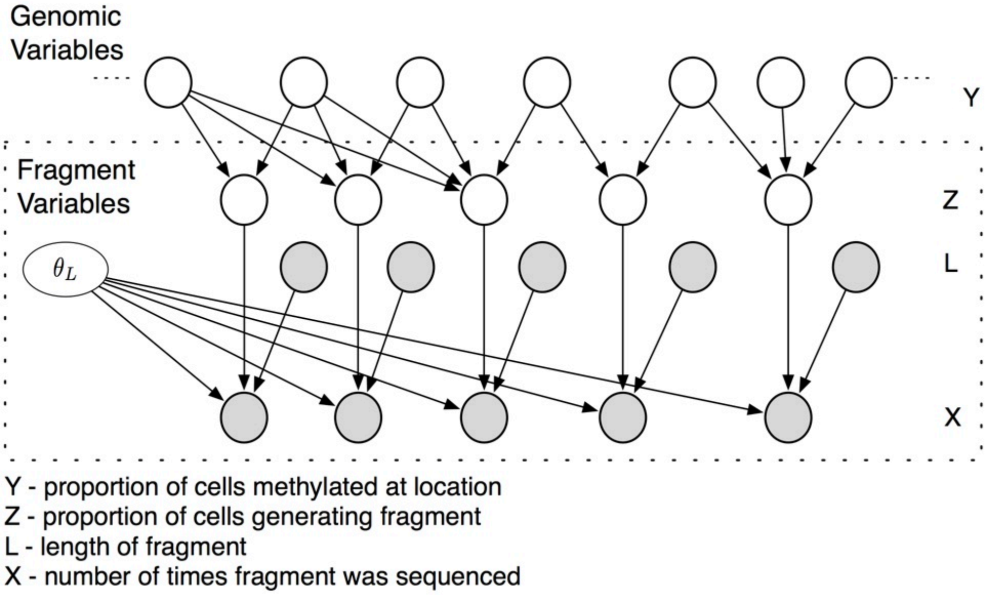

Genomic Variables
Y
Fragment Variables
Z
θ_L
L
X
Y - proportion of cells methylated at location
Z - proportion of cells generating fragment
L - length of fragment
X - number of times fragment was sequenced

Figure 3.3: A Bayesian network representation for the generative model. Variables that are observed in the methyl-Seq experiment are filled in grey.

$z_i \sim \mathcal{N}(\mu_i, \frac{\mu_i(1-\mu_i)}{wq})$ where $\mu_i = (1-y_s)(1-y_e)\Pi y_m$ and y_s and y_e are the sites on the upstream and downstream boundaries of fragment f_i, y_m are the sites between y_s and y_e, q is the number of cells in the digest and w is the number of chromosome copies in each cell (for human $w = 2$)[2].

Let $\{X_i\}_{i=1,...,|F|}$ be random variables for the number of times fragment f_i was sequenced (the number of paired-end reads that were sequenced from fragment f_i). In our generative model X_i follows a Poisson process in which the expected number of reads to be sequenced is $\lambda_i = z_i\theta_{l_i}$, where θ_{l_i} is a factor determined by the length of the fragment, and depends on the specific experiment technicalities, such as the total amount of sequencing in the experiment. We condition X_i on the length of f_i since the size selection step in the experiment is not precise, and moreover, fragments of different lengths have different probabilities of being sequenced, due to technicalities of the sequencing machinery. Figure 3.3 shows the dependencies of our generative model in a graphical model representation.

We have made a few simplifying assumptions in this generative model that greatly simplify the dependencies among variables and result in a tractable model that is convenient to work with for inference purposes. The first simplifying assumption we make is that the methylation states of neighboring *HpaII* sites are independent. Second, unlike many generative models of RNA-Seq [118], in this model we assume X_i, the number of times fragment f_i is sequenced, depends on

[2] The number of occurrences of fragment f_i in the solution follows a Binomial distribution $B(wq, \mu_i)$, for which $\mathcal{N}(wq\mu_i, wq\mu_i(1-\mu_i))$ is a good approximation. Therefore, the proportion of samples from which f_i was generated can be approximated by the distribution $\mathcal{N}(\mu_i, \frac{\mu_i(1-\mu_i)}{wq})$.

the proportion of cells from which f_i was generated (Z_i), but not on the relative proportion to which f_i is present in the digest ($\frac{z_i}{\sum_{j=1}^{|F|} z_j}$). We allow ourselves to make this assumption in the methyl-Seq case because, unlike RNA-Seq, the contribution of each fragment to the digest is bounded by the number of cells (each cell can contribute at most $w \cdot f_i$ fragments), making the differences between fragment-frequencies much smaller than in RNA-Seq experiments. This, together with the fact that in methyl-Seq the number of possible fragments (contributors) is much larger than the number of transcripts in RNA-Seq experiments, brought us to relax the dependency on the relative proportion of f_i in the solution.

3.3.3 Genomic structure as a prior on methylation status

In the generative model described in the previous section we did not assume any prior knowledge about distribution of the methylation states at different *HpaII* sites. However, it is well known that one can make use of the genomic sequence to gain some information regarding the probability that a specific *HpaII* site is methylated. This is because in vertebrates unmethylated sites tend to cluster together, and CpG sites that are constitutively unmethylated are more conserved than other CpG sites. This results in regions that tend to be unmethylated being CpG rich compared to regions that tend to be methylated. There are several methods to annotate such CpG-rich regions, based on genomic sequences, in an effort to find regions that have interesting DNA methylation behavior. Such regions are called “CpG islands” (see chapter 2 for a detailed discussion).

The precise definition of CpG island regions and their methylation states is a challenging task that has been the subject of much research throughout the years [13; 50; 53; 68; 149; 156; 157; 160; 175], including chapter 2 of this thesis. This is not surprising if we recall that DNA methylation states are more dynamic than the DNA sequence. For example, DNA methylation states may be different across the different tissues of an individual [71; 156]. Thus, a purely sequence based definition of a “CpG island” is problematic. Nonetheless, in the case of experimental annotation of unmethylated sites, clusters of CpGs provide evidence against methylation that should ideally be incorporated into our model.

We therefore add a hidden random variable for each *HpaII* site, $\{W_i\}_{i=1,...,N} \in \{1,0\}$, indicating whether the i^{th} *HpaII* site is in an unmethylated cluster or not. We note that in our setting we assume that there are unmethylated clusters in the *experiment*, and the **W** variables denote the presence of a *HpaII* site in such a cluster. Different methyl-Seq experiments, on different tissues for example, may have a different set of unmethylated clusters. In addition, we add for each *HpaII* site two observed variables, denoted by C and D. $\{C_i\}_{i=1,...,N} \in \{1,...,c_{max}\}$ indicates the number of CpG sites between the $(i-1)^{\text{th}}$ *HpaII* site and the i^{th} *HpaII* site (any CpGs that are not part of a CCGG site). $\{D_i\}_{i=1,...,N} \in \{1,...,d_{max}\}$, indicates the distance in bps between the $(i-1)^{\text{th}}$ *HpaII* site and the i^{th} *HpaII* site. A graphical model representation of the model augmented with these additions can be seen in Fig. 3.4.

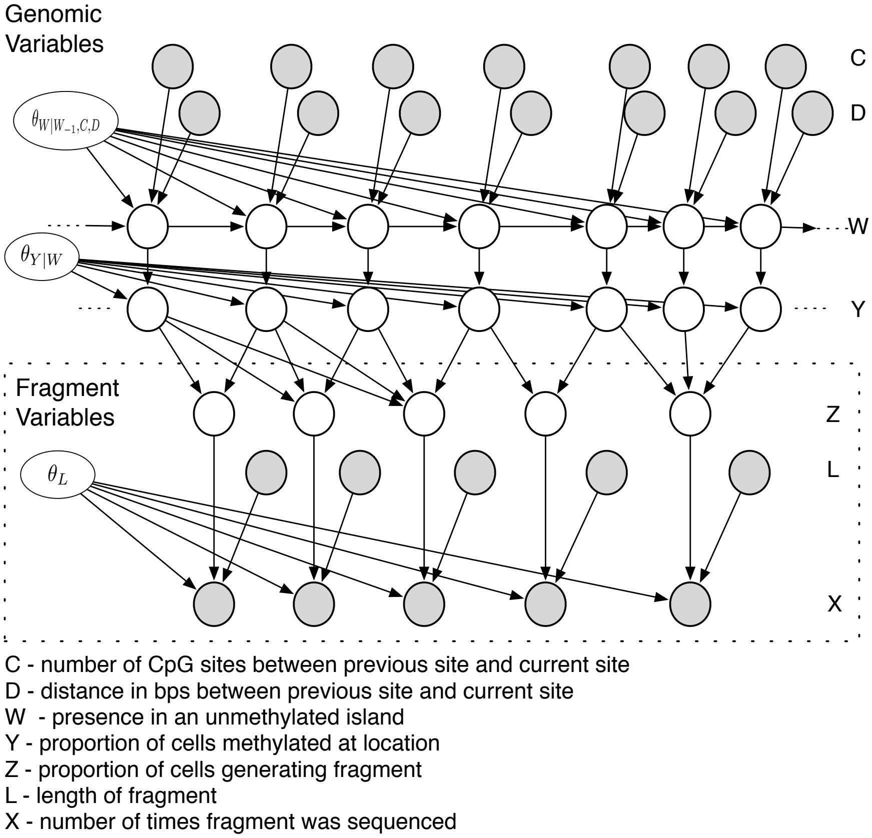

Genomic Variables
$\theta_{W|W_{-1},C,D}$
$\theta_{Y|W}$
C
D
W
Y
Fragment Variables
θ_L
Z
L
X
C - number of CpG sites between previous site and current site
D - distance in bps between previous site and current site
W - presence in an unmethylated island
Y - proportion of cells methylated at location
Z - proportion of cells generating fragment
L - length of fragment
X - number of times fragment was sequenced

Figure 3.4: A Bayesian network representation for our generative model incorporating genomic structure. Variables that are observed in the methyl-Seq experiment are filled in grey.

3.3.4 Parameter learning and inference of posterior probabilities

Now that we have a model in place, recall that our objective is to infer the posterior probability for each Y_i, given the observed variables. In this section we will describe how we can use learning and inference algorithms to attain estimates of the hidden **Y** and **W** variables given the experimental results. For this purpose we will make a few simplifying changes to our model: we discretize the **Y** variables such that $\{Y_i\}_{i=1,\ldots,N} \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and we assume that the **Z** variables are determined deterministically by the **Y** variables, such that $z_i = \mu_i$.

To summarize, the list of variables, parameters and dependencies in our model is as follows:

Variables:

$W_1, \ldots, W_N \in \{1, 0\}$: indicates for *HpaII* site i whether it is present in an unmethylated cluster (1) or not (0).

$Y_1, \ldots, Y_N \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$: for each *HpaII* site, the proportion of cells in the digest that are methylated.

$Z_1, \ldots, Z_{|F|} \in [0, 1]$: the proportion of cells from which fragment f_i was generated.

$L_1, \ldots, L_{|F|} \in \{1, \ldots, l_{max}\}$: the length in bps of fragment f_i.

$X_1, \ldots, X_{|F|} \in \{1, \ldots, x_{max}\}$: the number of times fragment f_i was sequenced.

$C_1, \ldots, C_N \in \{1, \ldots c_{max}\}$: the number of CpG sites between the previous and current *HpaII* sites.

$D_1, \ldots, D_N \in \{1, \ldots, d_{max}\}$: the distance between the previous and current *HpaII* sites.

Parameters:

$\theta_L = (\theta_{1-20}, \theta_{21-40}, \ldots, \theta_{381-400}, \theta_{\geq 400})$.
$\theta_{Y|W} = \{\theta_{y^0|w^0}, \ldots, \theta_{y^4|w^0}, \theta_{y^0|w^1}, \ldots, \theta_{y^4|w^1}\}$,
where $0 \leq \theta_{y^i|w^j} \leq 1$, $\sum_i \theta_{y^i|w^0} = 1$ and $\sum_i \theta_{y^i|w^1} = 1$.

$\theta_{W|W_{-1},C,D} = \{\theta_{w^a|w^b_{-1}, c_{(l_C,h_C)}, d_{(l_D,h_D)}}\}$ where
$a, b \in \{0, 1\}$,
$(l_C, h_C) \in \{(0,2), .., (2^{k-1}+1, 2^k), .., (513, 1024), (\geq 1025)\}$,
$(l_D, h_D) \in \{(0,5), .., (5 \cdot 2^{k-1}+1, 5 \cdot 2^k), .., (1280, 2560), (\geq 2561)\}$,
and w_{-1} denotes the states of the W variable that is the parent of the W variable at hand.
For all indexes, $0 \leq \theta_{w^a|w^b_{-1}, c_{(l_C,h_C)}, d_{(l_D,h_D)},} \leq 1$, and
$\sum_{t=\{0,1\}} \theta_{w^t|w^b_{-1}, c_{(l_C,h_C)}, d_{(l_D,h_D)}} = 1$.

Dependency functions:

$z_i = (1 - y_{s_i})(1 - y_{e_i}) \Pi_{j=s_i+1}^{e_i-1} y_j$, where s_i and e_i are the indexes of the locations at which f_i begins and ends.

$P(x_i|z_i, l_i) = \frac{\lambda^{x_i} exp\{-\lambda\}}{x_i!}$, where $\lambda = z_i f(l_i)$, and $f : \{1, \ldots, l_{max}\} \rightarrow \theta_L$ is a function that takes in l_i and returns θ_{s-e} such that $s \leq l_i \leq e$.

The rest of the dependency functions are fully described by the model parameters.

In practice, fragments that are very short or very long relative to the boundaries of the size selection step have probability zero of being observed, and we can omit the corresponding variables from the model. This significantly reduces the number of variables, and the complexity of parameter learning and inference procedures.

Although we listed $\{Z_i\}_{i=1,...,|F|} \in [0,1]$ for convenience, we note that when the **Y** variables are discrete, so are the **Z** variables. The number of values a specific Z_i can take is directly dependent on the number of *HpaII* sites that are on the fragment Z_i represents. Due to the relatively small size-selection bounds (50-300 bps) and the *HpaII* restriction site being of length four (CCGG), the majority of **Z** variables are directly dependent on a small number of **Y** variables. This is important for practical reasons, because when conducting the inference step in the EM approach (see below), the Bayesian network is moralized[3], and the running time of the procedure is exponential in the size of the largest clique in the moralized graph. To avoid our moralized graph from encompassing large cliques we determine *a priori* a maximal number of parents we allow for any Z_i variable. Variables with more parents than allowed are not incorporated into the model. While this results in ignoring data from fragments that encompass too many *HpaII* sites, it has a crucial impact on lowering the computational complexity of the inference procedure.

The direct output of a paired-end experiment is a list of paired sequenced reads. Each such pair is mapped to a reference genome to determine which genomic fragment it originated from (see section 3.2). In practice, some of the pairs cannot be mapped uniquely to one fragment of the genome. To resolve this, one may add another layer to the model, taking into account the uncertainty regarding which fragment the reads originated from. However a simpler approach is to disregard from the model fragments that can generate pairs of reads that do not uniquely map.

In the same manner that we have a variable determining the length of each fragment, we can consider additional variables for characteristics that affect the extent of sequencing (such as GC content [7; 38]). However one must take into account that in the current setting this will result in an exponential growth in the number of parameters.

We would like to compute for each Y_i and W_i the posterior distribution given the observed data. Given some parameter assignment, we can compute these posterior probabilities using the junction tree algorithm [82]. This computation however requires a parameter assignment, which we do not know. On the other hand, if we were to observe a full assignment of all variables in the Bayesian network (including **Y**, **Z** and **W**), let such a dataset be $\mathscr{D}$, we could learn a set of parameters for our model using the maximum likelihood approach. This is a good setting to use the expectation maximization (EM) algorithm [37], which returns a parameter assignment for models with hidden variables, given the observed data. In the interest of clarity, we will first describe the application of the EM algorithm to the simplified model (Figure 3.3) and then describe the application to the full model (Figure 3.4).

[3] In the moralization step edges are added between all parents of each node, forming cliques, and all edges of the graph are changed to be non-directed. For further information on moralization and its importance for exact inference procedures see [82].

The EM algorithm

From the Bayesian network structure of the generative model (Figure 3.3) we can construct the likelihood function:

$$\begin{aligned} L(\theta_L : \mathscr{D}) &= \left[\Pi_{j=1}^{N} P(y_j)\right] \cdot \left[\Pi_{i=1}^{|F|} P(l_i) P(z_i|\mathbf{y}) P(x_i|z_i, l_i)\right] \\ &\propto \Pi_{i=1}^{|F|} \frac{(z_i f(l_i))^{x_i}}{x_i!} e^{-z_i f(l_i)}, \end{aligned}$$

assuming a uniform prior over **Y** and **L** and that the probability of $\mathscr{D}$ is not zero (that all z_i assignments are as expected given the y_i assignments). $f(l_i)$ is as defined in the "Dependency function" annotation. The log-likelihood is therefore:

$$\ell(\theta_L : \mathscr{D}) = \sum_{j=1}^{L} \left[\sum_{i=1}^{|F|} (x_i \log \theta_{L(j)} - z_i \theta_{L(j)}) \, \mathbf{1}\{f(l_i) = \theta_{L(j)}\} \right] + C,$$

where $\theta_{L(j)}$ is the j^{th} element of θ_L and C is the part of the equation with elements that do not include instances from θ_L. We would like to find θ_L which maximizes the likelihood function. Each $\theta_{L(j)}$ can be optimized separately, and by differentiating the log-likelihood, setting to zero and solving for $\theta_{L(j)}$ we get the maximum likelihood parameters :

$$\hat{\theta}_{L(j)} = \frac{\sum x_i \, \mathbf{1}\{f(l_i) = \theta_{L(j)}\}}{\sum z_i \, \mathbf{1}\{f(l_i) = \theta_{L(j)}\}}.$$

Given an initial assignment of (arbitrary) parameters, the EM algorithm computes the expected values of the numerator and denominator used for the maximum likelihood parameter estimation. It then updates the parameters of the model using the expected values computed. The algorithm is guaranteed to converge to a stationary point of the log-likelihood function, because the likelihood increases with every iteration [37].

We start with some (possibly random) parameter assignment, θ_L^1, and iterate between the following two steps:

Expectation (E-step). In this step the algorithm uses the parameters from the last M-step, θ_L^t, to compute the *expected* values for the sufficient statistics used in the maximum likelihood parameter estimation. The expected values are computed for sufficient statistics that incorporate "hidden" (unobserved) variables of the model. The expected sufficient statistic that needs to be computed for $\theta_{L(j)}$ is

$$\begin{aligned} E_{Z|X,L,\theta_L^t}[\sum_i z_i \, \mathbf{1}\{f(l_i)=\theta_{L(j)}\}] &= \sum_i E_{Z|X,L,\theta_L^t}[z_i \, \mathbf{1}\{f(l_i)=\theta_{L(j)}\}] \\ &= \sum_i z_i P(z_i|\mathbf{X},\mathbf{L},\theta_L) \, \mathbf{1}\{f(l_i)=\theta_{L(j)}\}, \end{aligned}$$

and can be computed from the posterior distributions obtained by the junction-tree algorithm on the moralized graph.

Maximization (M-step). In this step we use the expected sufficient statistics from the E-step to preform the maximum likelihood estimation:

$$\theta_{L(j)}^{t+1} = \frac{\sum x_i \, \mathbf{1}\{f(l_i)=\theta_{L(j)}\}}{E_{Z|X,\theta_L^t}[\sum z_i \, \mathbf{1}\{f(l_i)=\theta_{L(j)}\}]}.$$

The algorithm iterates between the E-step and the M-step until the increase in the likelihood function is smaller than a given threshold. After the algorithm has converged at θ_L^f, we can compute for each Y_i its posterior distribution using a final run of the junction tree algorithm.

The EM algorithm applied to the full model

In order to make inferences from the full model several additional parameters need to be estimated. Assuming we have a fully observed dataset, we can use the likelihood function to derive the maximum likelihood estimators for our model parameters:

$$\hat{\theta}_{L(i)} = \frac{\sum x_i \, \mathbf{1}\{f(l_i)=\theta_{L(j)}\}}{\sum z_i \, \mathbf{1}\{f(l_i)=\theta_{L(j)}\}},$$

$$\hat{\theta}_{y^a|w^b} = \frac{\sum_i \, \mathbf{1}\{y_i=y^a, w_i=w^b\}}{\sum_i \, \mathbf{1}\{w_i=w^b\}},$$

and

$$\begin{aligned} &\hat{\theta}_{w^a|w^b_{-1},c_{l_C,h_C},d_{l_D,h_D}} \\ &= \frac{\sum_i \, \mathbf{1}\{w_i=w^a, w_{i-1}=w^b_{-1}, f_1(c_i)=c_{(l_C,h_C)}, f_2(d_i)=d_{(l_D,h_D)}\}}{\sum_i \, \mathbf{1}\{w_{i-1}=w^b_{-1}, f_1(c_i)=c_{(l_C,h_C)}, f_2(d_i)=d_{(l_D,h_D)}\}}, \end{aligned}$$

where $f_1 : \{1,\dots c_{max}\} \to (l_C,h_C)$ takes in c_i and returns a pair, (l_C,h_C), such that $l_C \leq c_i \leq h_C$, and $f_2 : \{1,\dots,d_{max}\} \to (l_D,h_D)$ takes in d_i and returns a pair, (l_D,h_D), such that $l_D \leq d_i \leq h_D$.

As before, we start with some (possibly random) assignment of the parameters, and iterate between the following two steps until we determine convergence. For convenience, we denote by θ the complete parameter set for our model.

E-step. In this step the algorithm uses the parameters from the last M-step, θ^t, to compute expected sufficient statistics. For $\theta_{L(i)}$ this is done as previously described. Let $\mathbf{O} = (\mathbf{X}, \mathbf{L}, \mathbf{C}, \mathbf{D})$ be the set of observed variables, and $\mathbf{H} = (\mathbf{W}, \mathbf{Y}, \mathbf{Z})$ be the set of hidden variables. Using the junction-tree algorithm on the moralized graph, we can compute the probability of any assignment to some variable, Q, and its parents, $\mathbf{U}$, given the observed data and θ^t. In other words, after a run of the junction-tree algorithm, we know $P(q, \mathbf{u}|\mathbf{O}, \theta^t)$ for every assignment of Q and $\mathbf{U}$, $(q, \mathbf{u})$. We can then compute the needed expected sufficient statistics. For the enumerator and denominator of $\theta_{y^a|w^b}$ we compute

$$\begin{aligned}
&E_{H|O,\theta^t}[\sum_i \mathbf{1}\{y_i = y^a, w_i = w^b\}] \\
&= \sum_i E_{H|O,\theta^t}[\, \mathbf{1}\{y_i = y^a, w_i = w^b\}] \\
&= \sum_i P(y_i = y^a, w_i = w^b|\mathbf{O}, \theta^t)
\end{aligned}$$

and

$$E_{H|O,\theta^t}[\sum_i \mathbf{1}\{w_i = w^b\}] = \sum_i P(w_i = w^b|\mathbf{O}, \theta^t).$$

For $\theta_{w^a|w^b_{-1}, c_{(l_C,h_C)}, d_{(l_D,h_D)}}$ we compute

$$\begin{aligned}
&E_{H|O,\theta^t}[\sum_i \mathbf{1}\{w_i = w^a, w_{i-1} = w^b_{-1}, f_1(c_i) = c_{(l_C,h_C)}, f_2(d_i) = d_{(l_D,h_D)}\}] \\
&= \sum_i P(w_i = w^a, w_{i-1} = w^b_{-1}, f_1(c_i) = c_{(l_C,h_C)}, f_2(d_i) = d_{(l_D,h_D)}|\mathbf{O}, \theta^t)
\end{aligned}$$

and

$$\begin{aligned}
&E_{H|O,\theta^t}[\sum_i \mathbf{1}\{w_{i-1} = w^b_{-1}, f_1(c_i) = c_{(l_C,h_C)}, f_2(d_i) = d_{(l_D,h_D)}\}] \\
&= \sum_i P(w_{i-1} = w^b_{-1}, f_1(c_i) = c_{(l_C,h_C)}, f_2(d_i) = d_{(l_D,h_D)}|\mathbf{O}, \theta^t).
\end{aligned}$$

We recall that the running time complexity of the junction-tree algorithm is exponential in the size of the maximum sized clique of the moralized graph, and it is easy to see how the moralized network can be chorded such that, aside from the cliques generated in the previous model, only cliques of size three are introduced.

M-step. In this step we generate an updated set of parameters, θ^{t+1}, by using the expected sufficient statistics computed in the E-step. The computation of θ_L^{t+1} remains the same as before, and $\theta^{t+1}_{y^a|w^b}$ and $\theta^{t+1}_{w^a|w^b_{-1},d_{l_D,h_D},c_{l_C,h_C}}$ are updated as follows:

$$\theta^{t+1}_{y^a|w^b} = \frac{\sum_i P(y_i = y^a, w_i = w^b|\mathbf{O}, \theta^t)}{\sum_i P(w_i = w^b|\mathbf{O}, \theta^t)},$$

and

$$\begin{aligned}&\theta^{t+1}_{w^a|w^b_{-1},c_{(l_C,h_C)},d_{(l_D,h_D)}}\\ &= \frac{\sum_i P(w_i = w^a, w_{i-1} = w^b_{-1}, f_1(c_i) = c_{(l_C,h_C)}, f_2(d_i) = d_{(l_D,h_D)}|\mathbf{O}, \theta^t)}{\sum_i P(w_{i-1} = w^b_{-1}, f_1(c_i) = c_{(l_C,h_C)}, f_2(d_i) = d_{(l_D,h_D)}|\mathbf{O}, \theta^t)}.\end{aligned}$$

The algorithm iterates between the E-step and the M-step until the increase in the likelihood function is smaller than a given threshold. After the algorithm has converged at θ^f we can compute the posterior distribution for each Y_i by using a final run of the junction tree algorithm. We can now also compute for every W_i the probability that the *HpaII* site at index i is in an unmethylated cluster. We elaborate on this in the next section.

3.4 The MetMap package for methyl-Seq analysis

In the previous section we described a Bayesian network for the correction of bias introduced in the methyl-Seq experiment. We have implemented such a model in a free, open source software package called MetMap, available at:
`http://www.cs.berkeley.edu/~meromit/MetMap.html`.
In this section we describe several aspects of the MetMap implementation, and in the next section we present a study of the methyltypes of neutrophils from four human individuals that was used to validate MetMap's performance. Further details of our implementation of MetMap can be found at [148].

Unmethylated Clusters

Unmethylated sites of vertebrate genomes tend to cluster together, and it is therefore of interest to annotate the location and methylation states of such clusters.

Chapter 2 discussed the possibility of using the genome sequence to predict the whereabouts of such clusters, termed CpG islands. In short, due to a high mutation rate of methylated CpG sites into CpA or TpG sites, one can annotate regions that are constitutively unmethylated in the germline, or that are under selection, by finding segments of the genome that are rich in CpG sites. Chapter 2 discusses methods for the annotation of CpG islands from genomic sequences and highlights some of the challenges encountered in the process. The ability to make use of the link between mutation rates and methylation status to annotate potential unmethylated clusters is extremely powerful, but in presence of genome-wide methylation data there are several clear benefits to making use of it in the annotation process of unmethylated clusters. For example, while the genome sequence is uniform across tissues it has been shown that methylation states vary [71]. Given genome-wide methylation data we would like to know the actual methylation rates at regions that are suspected (due to their sequence) of being unmethylated. In addition, methylation data should enable finding regions that are unmethylated in the sample tested but do not have a strong genomic signature indicating this, and are therefore not annotated as CpG islands.

An important feature of MetMap is the annotation of unmethylated islands (called SUMIs - Strongly UnMethylated Islands) that are specific to the experiment at hand. MetMap specifies the coordinates of these islands and outputs for each island its mean methylation score. The SUMI regions annotated by MetMap are the union of two sets of regions. The first set consists of continuous regions in which each of the W variables received a probability of being in an unmethylated island that was larger than 0.1 and in which the methyl-Seq data directly supports the presence of at least two fragments. This set includes regions with relatively weak direct experimental evidence but with strong sequence evidence for being unmethylated. The second set was generated by setting a 600bp interval around each *HpaII* site for which the probability of being unmethylated was higher than the prior probability of being unmethylated outside of an unmethylated-island (0.1663) and for which the posterior probability of being in an unmethylated island was smaller than 0.1. All overlapping windows are concatenated and the regions selected are those in which at least 30% of the *HpaII* sites had a probability of being unmethylated larger than the prior-set threshold (0.1663) and in which the methyl-Seq data directly supports the presence of at least two fragments. This set includes regions with weaker sequence support for hypomethylation, but with extensive evidence that they are hypomethylated. Each SUMI receives a score, specifying the mean of the MetMap unmethylation scores at all sites within the SUMI. SUMIs can be used as "comparative units", facilitating the comparison of datasets.

Sites in the scope of the experiment

It is important to restrict analysis only to sites for which the methyl-Seq experiment holds some information regarding their methylation state. We reffer to such sites as being within the scope of the methyl-Seq experiment. MetMap identifies these sites from the structure of the graphical model.

A *HpaII* site is in the scope of an experiment if it lies on some fragment that has *HpaII* sites at its ends, and is of length l such that $l_{min} \leq l \leq l_{max}$, where l_{min} and l_{max} are the minimal and maximal fragment lengths for the methyl-Seq experiment. MetMap's graphical model identifies these sites; they are all *HpaII* site variables (Y variables) that have an edge to some fragment variable (Z variable). We note that this condition does not require a site to be at an end of a fragment that satisfies the length requirements; a site may be in the interior of such a fragment.

Single end sequencing

When given a single-end dataset as input, MetMap performs a transformation to an approximated paired-end dataset, and proceeds as if the initial data was paired-end. Conducting such a transformation is not trivial, since there is no bijection between single-end and paired-end datasets; one single-end dataset can be explained by different paired-end datasets. Theoretically, this issue may be addressed by having an inference procedure consider the different possible paired-end assignments for the given single-end dataset; however in our case doing so would result in an infeasible inference procedure due to the large sized cliques the single-end variables would introduce. We therefore chose to use a different approach.

Given a single-end dataset, each *HpaII* has a read count in both the upstream and the downstream directions, indicating the number of reads generated from sequencing starting at that site and proceeding to a specific direction. MetMap assumes that given a read count score at a site for a specific direction, it is most probable that most of the score originated from the shortest fragment for which there is an abundant signal at the corresponding site (the restriction site at the other end of the fragment). MetMap first generates a c_{max} value from the single-end read counts. c_{max} is set to be the value two standard deviations from the mean of the counts for sites within CpG islands (the CpG island track selected for this purpose was from the UCSC browser track [78]), assuming the read counts follow a poisson distribution. Dynamic scores are assigned to the different sites and different directions (meaning each site has two dynamic scores), starting out with each dynamic score being $min(c_i, c_{max})$, where c_i is the raw read count. Then, the method goes through the fragments of the genome from shortest to longest. When reaching a fragment, i, it is assigned a score $v_i = min(d_b, d_e)$ where d_b and d_e are the dynamic signals of the sites at the beginning and end of the fragment, in the corresponding directions. The method then updates each dynamic signal by subtracting v_i from it's dynamic score. After all assignments are through the leftover dynamic scores are distributed to the shortest fragment they are at an end of.

MetMap's output

The MetMap software takes as input: (1) the mapped reads of a methyl-Seq experiment, (2) the boundaries on the lengths of the fragments sequenced (determined by the size-selection step), and (3) a reference genome. The reference genome specifies the structure of the graphical model, and the methyl-Seq data is incorporated into this model by fixing the states of the appropriate variables. After running the inference procedure MetMap determines the methylation state of each *HapII* site, represented by Y_i, to be $E[Y_i|\mathscr{D}]$, where $\mathscr{D}$ is the observed fragment counts. Then, unmethylated islands (SUMIs) are determined.

MetMap outputs two files: (1) a list of the *HpaII* sites in the scope of the experiment with their inferred *unmethylated* states (assigning 0 to sites that are completely methylated and 1 to sites that are completely unmethylated), and (2) a list of SUMI regions with scores indicating the mean of the methylation states.

3.5 Evaluating MetMap's performance: methyltyping the human neutrophil

We carried out methyl-Seq on specimens of a single homogeneous and uncultured cell type, the neutrophil, from four male humans. HpaII fragments were size selected in the range 50-300bp and sequenced on a first generation Illumina Genome Analyzer yielding 23,731,359 32bp reads. Although longer reads are currently available, reads for our assay only need to be sufficiently long so that they can be mapped correctly to the reference genome. The reads were aligned to the reference human genome (hg18 [108]) with Bowtie [85] resulting in 18,218,420 alignments, and each of the four samples was analyzed with MetMap.

To infer methylation states from read depths, we first segmented the genome into 6,000 non-overlapping regions (of size 0.5Mbp) that could be analyzed separately. For each region, MetMap returned methylation probabilities for those CCGG sites for which information on site-specific methylation could be obtained from the methyl-Seq experiment, and annotated SUMIs. The CCGG subset contained 59% of the CCGG sites (4.8% of all CpG sites) in the human genome. Of the sites for which information could be obtained, 80% (1,035,243 sites) were outside CpG islands as annotated in the UCSC Genome Browser [78], and 20% (257,540 sites) were inside, resulting in a two-fold enrichment of the proportion of CCGG sites that are in such CpG islands.

To test whether MetMap was correcting bias in the raw counts, we directly determined the methylation status of 22 regions in the human genome using bisulfite sequencing [32]. Each CpG in the bisulfite experiment received a score from the set (0,0.25,0.5,0.75,1) based on the observed proportion of alleles in which that site was unmethylated in a sample [87].

We correlated the bisulfite scores (taken as being the true methylation status) with the read counts and with the MetMap predictions. Each of the 46 validated sites had three different scores for the extent to which it was unmethylated: a bisulfite score, a read count score, and a MetMap score. The Pearson correlation coefficient between the raw read counts and the bisulfite values was 0.67 while the Pearson correlation coefficient between the MetMap methylation score of those sites and the bisulfite values was improved to 0.90.

As the bisulfite scores may be an imprecise measure of the true extent of methylation (Methods) we tested the sensitivity of our results to the bisulfite scores. We "adjusted" bisulfite scores, assigning to each value of the two sets of scores, the read-count set and the MetMap predictions set, a separate "adjusted" bisulfite value, that is within a predetermined range. The range available for adjustment was determined by the initial bisulfite score. After this adjustment, the correlation coefficient of the read counts with the bisulfite scores was 0.73 and the correlation coefficient of the MetMap scores with the bisulfite scores was 0.95. While the correlation values increased as expected, the difference between the performance of MetMap and that of read counts remains similar. This indicates that the improvement in using MetMap instead of raw read counts was not due to the procedure by which bisulfite scores were assigned.

Examples of MetMap's ability to accurately detect partially and fully methylated sites are shown in Appendix B. Both the extent and variability of methylation in a region are better predicted by

MetMap than by the read counts.
In order to test MetMap's sensitivity, we considered a region in which the read counts varied considerably between two of the human samples, and observed that MetMap's inferences for this region are different for the two cases (Table 3.1). As an additional test, we artificially changed the read-counts at region chr19:4494679-4494763 and evaluated MetMap's inferences given the different input. The region was predicted by MetMap as unmethylated for the actual sample data, and was validated using bisulfite sequencing. We artificially fixed that region to be methylated, changing the read counts accordingly (assigning all to 0). MetMap's inferences indicated the region was methylated, as can be seen in Table 3.2.

	Sample 1		Sample 4	
Coordinate	Read Counts	MetMap Score	Read Counts	MetMap Score
chr16:66244444	5	0.069005	15	0.6234
chr16:66244465	2	0.128735	12	0.71821
chr16:66244537	1	0.21895	13	0.722445
chr16:66244541	2	0.24566	7	0.70159
chr16:66244599	5	0.592835	12	0.71721

Table 3.1: Different Read Counts in samples result in different MetMap predictions. The area of chr16:66244444-66244599 (156 bps) has different read counts across the region for two of the human samples. This difference is reflected in MetMap's output for this region.

Coordinate	Unmethylated (Original Input)	Methylated (Simulated Input)
chr19:4494679	0.999315	0.10772
chr19:4494683	0.76112	0.108185
chr19:4494716	0.97853	0.130815
chr19:4494763	0.9005	0.10113

Table 3.2: MetMap scores for bisulfite validated region, for different read-count inputs.

We next explored site-specific variability of methylation estimates among the four individuals, within the predicted SUMIs and outside of them. We noticed that methylation states tended to be more conserved at sites outside of SUMIs than within them (Pearson correlation was 0.955 and 0.81 for sites outside and inside of unmethylated islands, respectively), a finding also consistent with Bock et al. [18]. We note that similar read counts at orthologous restriction sites in two or more samples indicate that their methylation status is similar; however determination of their true extent of methylation requires a statistical method such as MetMap. Thus the degree of consistency observed among MetMap's site-specific inferences for different samples is supported by the high correlation of the corresponding raw read counts (e.g.: a correlation of 0.667 between sample 1 and sample 4).

Although the methylation status of individual sites within SUMIs was variable, the average probability of methylation for the whole SUMI was consistent across individuals (Figure 3.5). This observation suggests that the mean methylation state of a SUMI is more constrained than the methylation states of the individual sites within it, and thus a change in mean SUMI methylation is more likely to have functional consequences than a change at a specific site. Based on this, we propose that the mean SUMI methylation status is the more informative parameter for comparative or association studies.

Figure 3.5 illustrates the extent of consistency among samples, for sites inside SUMIs and for the mean MetMap scores of SUMIs. In all site-specific comparisons some sites are determined as fully unmethylated in one sample and unmethylated to some extent in the second sample. These can be seen as lines from the top right corner towards the bottom right corner, and from the top right corner towards the top left corner. These are sites that differ in methylation between the two samples, and the plots show that when a site differs in its methylation state between two samples, in the large majority of the cases the site is completely unmethylated in one of the two samples. At the lower right and upper left corners of the plots there are concentrations of sites that are determined to be highly, but not completely, unmethylated in one of the samples. We hypothesized that these concentrations occur because in those sites the prior (which is based on the genomic region in which the sites are located) has influenced the MetMap scores to be slightly lower than their true biological states. We confirmed this hypothesis by considering the percentage of sites that were in the UCSC CpG island set, and showing that the relevant corners of the plots are enriched for sites that are outside of these islands (see Methods). These findings show that MetMap's prior distribution does have an effect on the method's inferences, pulling scores down at times, but the extent to which the current prior distribution has changed the inferences for these sites is not large, and easily allows these sites to be detected as highly unmethylated.

MetMap identifies novel unmethylated islands associated with promoters and open chromatin regions

We mapped the 20,986 SUMIs present in at least one of the four individuals, and examined their relationship to purely sequence based definitions of CpG islands (Figure 3.6). Of the 20,986 SUMIs present in at least one of the four individuals, 4,652 do not overlap UCSC CpG islands, and 7,055 do not overlap the "bona fide" islands [17] with an epigenetic score larger than 0.5 (as recommended by Bock et al. [17], termed here BF islands). This result is consistent with the higher specificity, but lower sensitivity, of BF compared to UCSC island prediction. Details regarding the extent of overlap between SUMIs and the BF and UCSC islands can be seen in Table 3.3.

We compared the length distribution of our SUMIs with the length distributions of both the UCSC and BF islands (Figure 3.6.B). SUMIs were similar to BF islands, but the length distribution

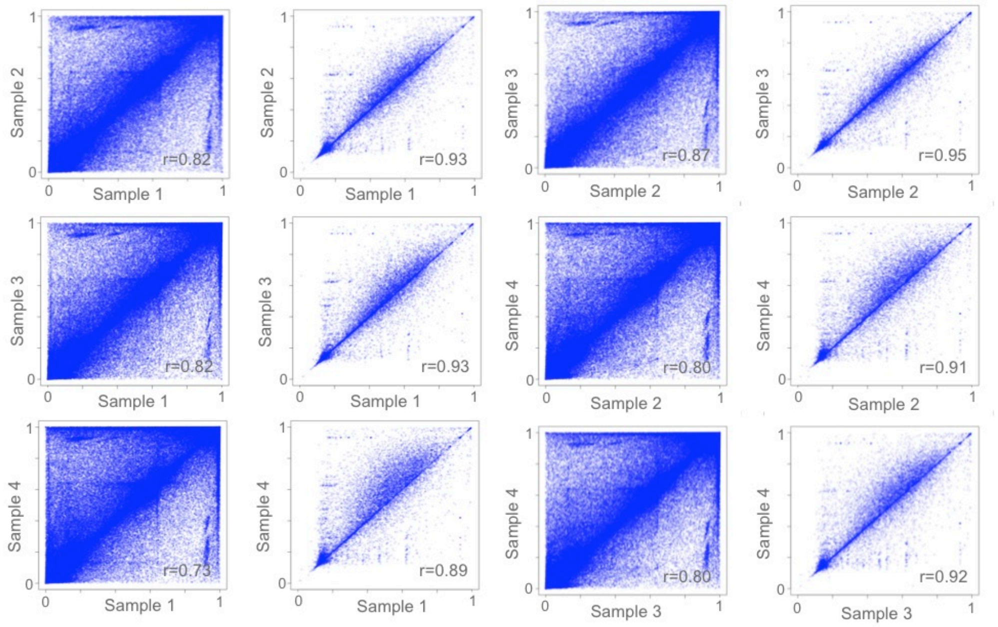


Figure 3.5: The average probability of methylation at SUMIs is highly stable across individuals of the same sex. All pairings among the four individuals tested are shown. On the left side of each pair the correlations between the site specific MetMap scores are presented for sites within SUMIs. On the right side of each pairing the correlations of the SUMI scores are presented.

	All Neutrophil SUMIs	Overlapping CGIs	Not Overlapping CGIs	Overlapping BFIs	Not Overlapping BFIs	Not Overlapping CGIs or BFIs
Sample 1	16,903	14,071	2,832	12,076	4,827	2,266
Sample 2	17,595	15,008	2,587	12,834	4,761	2,044
Sample 3	18,178	15,273	2,905	13,082	5,096	2,308
Sample 4	18,699	15,274	3,425	13,229	5,470	2,729
Union	20,985	16,334	4,651	13,931	7,054	3,797
Intersection	14,308	12,838	1,470	11,123	3,185	1,116

Table 3.3: Counts of the SUMIs annotated in the four human neutrophil samples. Union - The set of regions annotated as a SUMI in at least one of the four individuals. Intersection - The set of regions annotated as a SUMI in all four individuals. CGIs - UCSC CpG islands. BFIs - BF-islands.

of the UCSC CpG islands resembled a geometric distribution. The process by which UCSC CpG islands are annotated will produce false positives that follow a geometric length distribution, with

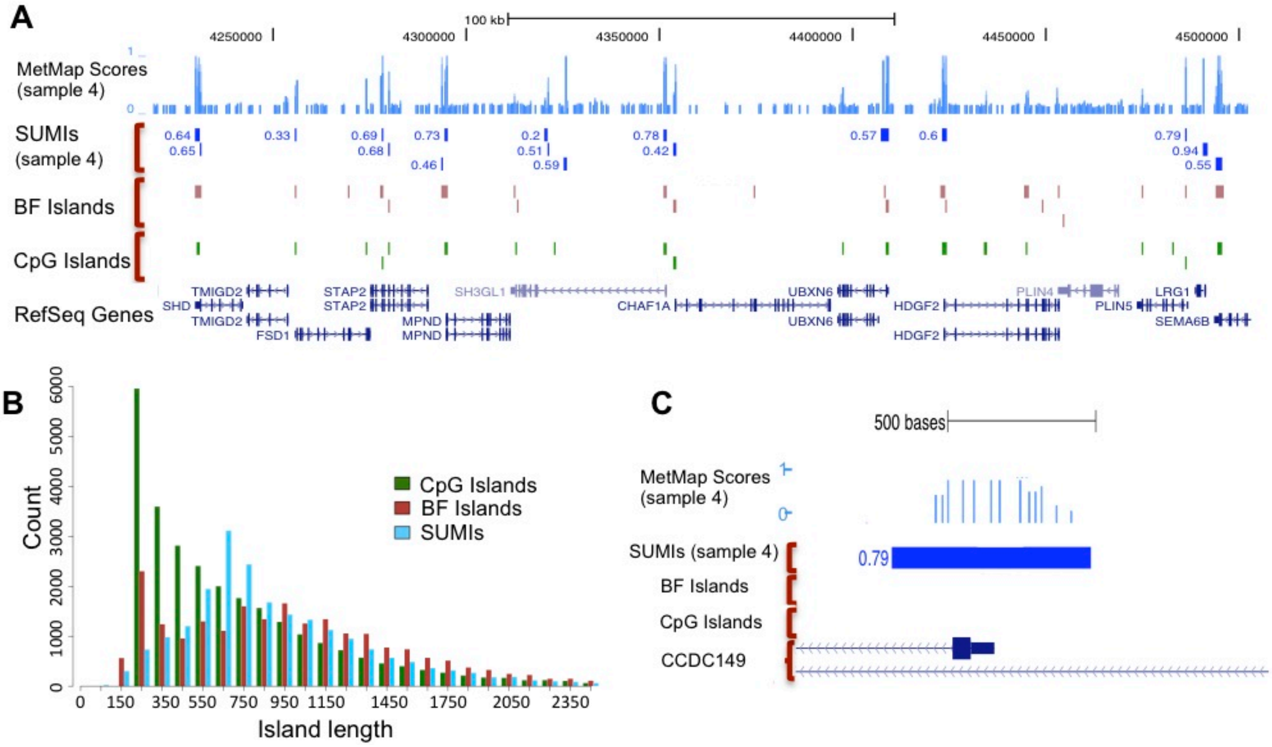


Figure 3.6: SUMIs in the neutrophil methylome. Genomewide SUMI predictions (a) reveal unmethylated islands that are proximal to genes and that do not always correspond to sequence based annotations of CpG islands shown in the tracks ‘BF islands’ and ‘CpG islands’ (e.g., the promoter of LRG1 and in an intron of SH3GL1). (b) SUMI and BF island length distributions have a different shape than the CpG island length distribution, suggesting numerous short false positives in the latter. (c) Some novel SUMIs appear 5’ of alternative promoter sites.

the number of false positive CpG islands increasing as a function of decreasing length (Methods). Since the length distributions of SUMIs and BF islands do not follow the same trend as the UCSC CpG island distribution, it is probable that at the shorter lengths the majority of predicted UCSC CpG islands are false positives. SUMIs did not overlap completely with BF islands: of the 21,626 BF islands, 13,899 were identified as SUMIs. BF islands are determined with a support vector machine that uses epigenetic data from multiple sources to train its prediction model. In contrast, MetMap’s SUMI predictions originate from an experimental signal for unmethylation in the cell type analyzed. The probable explanation for the MetMap/BF discrepancy is that the two methods have used epigenetic data from different tissues. More data from distinct cell types will shed light on this issue.

We validated with direct bisulfite sequencing five regions that are annotated as part of both a UCSC CpG island and a BF island, and did not overlap with SUMIs; we also sequenced three

	SUMIs	UCSC CpG islands	BF islands	Novel SUMIs
Open chromatin	70.0%	52.9%	65.3%	61.0%
Conservation	71.1%	68.5%	76.2%	49.6%
Gene regions	76.9%	77.7%	79.7%	59.9%
TSS regions	59.8%	52.2%	61.4%	22.0%

Table 3.4: Percentages of neutrophil SUMIs, UCSC islands and BF islands that overlap regions associated with functionality. For details on how the functional regions were determined see [148].

regions in BF islands that did not overlap with SUMIs or with UCSC CpG islands. In all cases those regions were validated as methylated in the neutrophil samples (see Appendix B). This is consistent with the notion that while these islands might be unmethylated in other cell types, they are methylated in the neutrophil. We analyzed four cases of SUMIs with scores higher than 0.5 that overlapped UCSC CpG islands but not BF islands. In each SUMI a region was picked and bisulfite sequenced. All four regions were determined as fully unmethylated (all CpG sites received a score of 1).

3,797 SUMIs do not overlap with BF islands or CpG islands, revealing new regions that are unmethylated in neutrophil cells. Of these novel SUMIs, 2,317 (61%) are within regions experimentally determined by the ENCODE project as open chromatin (Methods), 1,882 (50%) are within regions determined as conserved by the 17-way UCSC conservation track, 2,274 (60%) are within 2Kbp of RefSeq genes, and 837 (22%) are within 2Kbp of the 5' end these genes (Figure 3.6.C and Table 3.4).

Consistently with their similarity to conventional CpG islands, SUMIs are enriched near the transcription start sites (TSSs) of RefSeq genes, with a preference for the downstream side (Figure 3.7.A). We observe the same property also when we consider novel SUMIs alone (Figure 3.7.B), or when we consider only SUMIs that do not overlap BF islands (Figure 3.7.C) or UCSC CpG islands (Fig 3.7.D). This indicates that the distribution of novel SUMIs around the TSSs does not originate from a characteristic present in only one of these sets. We find that the proportion of SUMIs that maps at a distance from TSSs is larger for novel SUMIs than for all SUMIs, but that novel SUMIs have a degree of association with open chromatin similar to that observed for all SUMIs (Table 3.4); this suggests that novel SUMIs may often represent distal regulatory sequences.

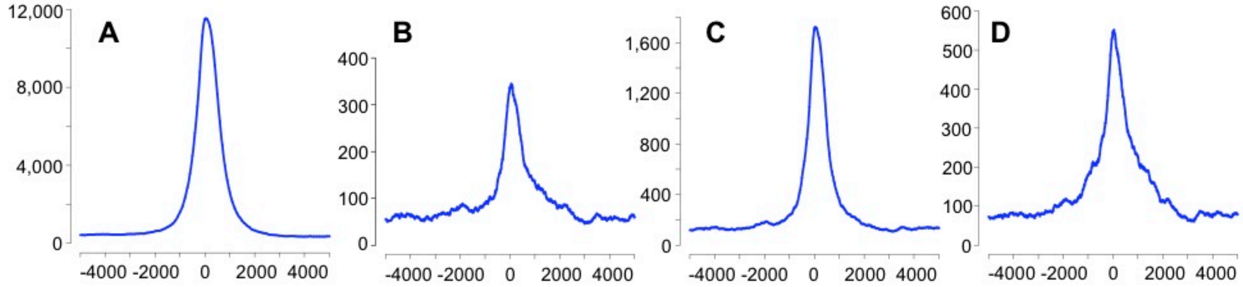


Figure 3.7: Transcription start sites and their close surroundings are enriched with novel SUMIs. The number of SUMIs that overlap each location within 5Kbp from RefSeq transcription start sites is shown for (a) all neutrophil SUMIs (b) Novel SUMIs (SUMIs that do not overlap UCSC CpG islands or BF islands) (c) SUMIs that do not overlap UCSC CpG islands (d) SUMIs that do not overlap BF islands.

3.6 Discussion

The possibilities and potential of DNA methylation analysis with new sequencing technologies have been described as a "revolution" [84]. The vast number of methods for methylation analysis, along with many papers describing exciting findings, support this statement. Methods that rely on the construction of a sequencing library produced by methyl-sensitive enzymes, followed by sequencing to measure methylation, are the practical approach for the analysis of large numbers of samples [84]. The efficient use of methyl-Seq data requires a computational method that can infer true methylation states by considering biases inherent in the technical method. We have developed a model based on a Bayesian network, implemented in the MetMap software, which makes it possible to use methyl-Seq for genome-scale methyltyping. MetMap facilitates the rapid calling of restriction-site-specific methylation, and of unmethylated regions, to produce methylation maps that are suitable for comparative analysis. Validation of MetMap calls with bisulfite sequencing shows that it compensates for bias present in the MethylSeq data. MetMap can combine experimental data and genome sequence to identify many unmethylated islands (SUMIs) that were previously unannotated, suggesting that it can identify novel functional regions.

The annotation of experiment-specific strongly unmethylated islands (SUMIs) reconciles the original definition of CpG islands, based on their sensitivity to methylation-sensitive restriction enzymes [13] with the sequence based definitions now used. The definition of SUMIs is functionally more exhaustive than the standard definition of CpG islands, since it couples sequence clues to methylation (abundance of CpGs) with experimental measurements of methylation. In our comparison of four humans, we noted that the average methylation states of SUMIs were more conserved among individuals than the methylation states of sites within them, suggesting that average methylation is more likely to be functionally important and so is a more informative parameter. SUMIs lie proximal to genes (77% are within 2Kbp of genes; 60% are within 2Kbp of the 5' end), and are likely to be directly involved in regulation of gene expression.

Overall, we predicted 3,797 SUMIs that do not overlap UCSC CpG islands or BF islands. Their sequence conservation and correlation with open chromatin suggests that they are functional, but they are less frequently associated with transcription start sites than the general set of SUMIs. We speculate that many novel SUMIs are enhancers. The discovery of these novel regions illustrates the utility of using experimental data to annotate CpG islands.

The main bias we have addressed here results from the experiment of interest (methyl-Seq) encompassing a non-random digestion of the genome followed by a size-selection step. Bayesian networks are suitable for such a case because they can be used to model the dependencies between neighboring sites. The model we have presented can be adjusted to account for similar biases in several other methyltyping methods, including the use of several different methylation sensitive restriction enzymes. It is also expected to prove useful in other high throughput sequencing experiments that are prone to such biases, for example some of the recent protocols developed for determining RNA secondary structure [166].

3.7 Methods

MethylSeq experiment

We obtained whole blood from four young adult male humans and obtained neutrophils by first isolating peripheral blood mononuclear cells by Ficoll separation, then purifying neutrophils with anti-CD16 antibodies conjugated to magnetic beads (Miltenyi); we verified that the purified samples contain $> 99\%$ neutrophils by Wright-Giemsa staining and visual inspection by a hematologist. Genomic DNA was isolated using the DNeasy Blood & Tissue isolation kit (Qiagen), quantified using a Nanodrop spectrophotometer, and quality-controlled for purity with an Agilent Bioanalyzer. Genomic DNA (2μg) was digested with HpaII under conditions that make it very likely that digestion is complete (overnight with enzyme boosting), fragments 50-300bp long were isolated from an agarose gel, and single-read sequencing libraries were prepared following the manufacturer's protocol (Illumina). Libraries were sequenced on a first-generation Illumina Genome Analyzer and 32 base reads were generated. Only reads beginning with "CGG" (the sequence of the ends produced by restriction with HpaII) were retained and analyzed with MetMap.

Validation with Bisulfite Sequencing

DNA was treated with the MethylEasy bisulfite conversion kit (Human Genetic Signatures), PCR-amplified with locus-specific primers that recognized human target sequences, and sequenced using standard Sanger chemistry. Since all epialleles from a single specimen were sequenced in bulk in the same mixture, we estimated the ratio of unmethylated/methylated alleles at each CpG in the sequence by examining the relative heights of the ‘C’ and ‘T’ traces in the sequencing output. Each CpG site received a score from the set (0,0.25,0.5,0.75,1), based on the relative C/T peak height [87]. A score of 1 indicates the site is fully unmethylated, meaning that only the ‘T’ trace was observed at the C position of a given CG, while a score of 0 indicates the site is fully methylated, meaning that only the ‘C’ trace was observed at the C position of a given CG.

“Adjusted” bisulfite values

We tested the extent to which our results may be affected by the representation of the bisulfite scores on a discrete five-point scale, since the true proportion of alleles that are unmethylated is a close to continuous measure. Each data point was assigned an ‘adjusted’ bisulfite score, within a tolerance window specified by the true bisulfite value of that data point. The ‘feasible ranges’ allowed for the ‘adjusted’ bisulfite scores were as follows: (0,0.15) for a 0 bisulfite score, (0.15,0.35) for a 0.25 score, (0.35,0.65) for a 0.5 score, (0.65,0.85) of a 0.75 score and (0.85,1) for a 1 score. For example, for a site with bisulfite score 0.25, read count score 0 and MetMap prediction 0.4 we would get two pairings (0.15,0) for (adjusted-bisulfite, read count score), and (0.35,0.4) for (adjusted-bisulfite, MetMap score). The score ranges were based on an assumption that assignments of “0.5” scores were the least precise. The adjustment of the bisulfite score to the read counts was done by generating a normalized read count value, in the 0-1 range, using the same “capping” value as MetMap.

Characterization of False-Positive UCSC CpG Islands

CpG islands in the UCSC track are defined in [50] as regions with a GC content of 50% or more, a length greater than 200bp, and a greater than 0.6 ratio of observed CpG dinucleotides to the expected number based on the GC content of the segment. The segments to consider are collected by scoring all dinucleotides (+17 for CpG and -1 for others) and identifying maximally scoring segments . Under this model, the probability that a region from the null model (sequence which is not an unmethylted region) fulfills these requirements increases as the length of the region decreases. This statement holds for models in which the probability of observing an A/T in the null model is larger than that of observing a C/G. This is indeed the case in humans. The likelihood of false positives in the UCSC CpG island set has been noted [17; 146].

Analysis of Figure 3.5

To test our hypothesis from figure 3.5, that the concentrations of sites at the corners of the site-specific plots are placed away from the plot borders due to the prior distribution, we took the sites within such a concentration and determined the proportion that are part of the UCSC CpG island set, and then compared this to the proportion in other subsets of sites. For our test we used the comparison between samples 3 and 4, and ran our test on the sites at the lower right corner of the plot. Specifically, we used sites that received a MetMap score for sample 3 that was between 0.1

and 0.4, and a MetMap score for sample 4 that was between 0.85 and 0.96. Of all 202,284 sites in the plot that compares samples 3 and 4, 80% are inside UCSC CpG islands. Of the sites that received a MetMap score lower than 0.4 for sample 3 (the sites in the lower half of the plot), 66.4% are inside UCSC CpG islands. Of sites with a MetMap score between 0.1 and 0.4 for sample 3, and a MetMap score higher than 0.96 for sample 4, 64.3% are within UCSC CpG islands. Finally, of sites with a MetMap score between 0.1 and 0.4 for sample 3, and a MetMap score between 0.85 and 0.96 for sample 4 (i.e. the sites that are part of the concentration), 31.2% are in UCSC CpG islands. This outcome supports our hypothesis, because the sites present in the concentration at the lower right corner are much less frequently part of a UCSC CpG island, and this characteristic affects the prior.

Open chromatin ENCODE files

One file of open chromatin was compiled from:
ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeChromatinMap/
using the union of the files:
wgEncodeUncFAIREseqPeaksH1hesc.narrowPeak
wgEncodeUncFAIREseqPeaksNhek.narrowPeak
wgEncodeUncFAIREseqPeaksGm12878V2.narrowPeak
wgEncodeUncFAIREseqPeaksHuvec.narrowPeak
wgEncodeUncFAIREseqPeaksPanislets.narrowPeak .

Chapter 4

A comparative study of DNA methylation in great apes

4.1 Introduction

Epigenetic modifications consist of a complex assortment of proteins and chemical modifications that are associated with DNA, control its transcription [10; 24] , and mediate stable phenotypic states. The genome and the epigenome are inherited together through cell divisions and generations, but the degree to which the epigenome is encoded by the genome is not known. Furthermore it is not clear to what extent the epigenome is maintained in the germline and transmitted between generations [45]. It might be reset in each generation using genetically encoded information, but persistence of epigenetic states in the germline creates the potential for semi-independent inheritance of epigenetic information [133; 139]. Finally, is not clear if epigenetic states that are present in the germline influence somatic epigenomes, or conversely if somatic epigenetic states are generated during cell differentiation using genetically encoded information.

We have explored the use of comparative epigenomic analysis ("phyloepigenomics") to obtain insights into changes in the epigenome in human evolution. Epigenetic differences provide a means to modulate the regulatory activity of noncoding regions. Functionally significant changes may be more readily identified in the epigenome than in the genome: sequence change is not always associated with functional change [19], but the epigenome mediates genome function by controlling transcription [10; 24], and so changes in it are more likely to reflect functional changes. Epigenomic comparison might thus complement the evidence of potentially adaptive genomic changes identified with multiple sequence-based approaches [59; 124; 127; 176].

This chapter explores these questions through the study of DNA methylation. We have compared the methylomes of human and chimpanzee in a homogeneous somatic cell type, the neutrophil, using the orangutan as an outgroup. Our comparison uses methyltypes generated by the coupling of methyl-Seq and the MetMap software (see Chapter 3 of this thesis).

We find that while the methyltypes of human and chimp are similar, a set of approximately 1500 stable differences in CpG island-like regions distinguishes human from chimp; these differences identify regions that may have diverged in gene regulatory function. The methylation states can be used to build a tree that recapitulates the phylogenetic relationship of the three species. Analysis of CG substitution patterns in CpG island-like regions that have conserved their methylated state in human, chimp, and orang indicates that methylation in the neutrophil reflects germline methylation. This raises the question of whether a germline epigenome is transmitted along with the genome, and if so, whether it is completely determined by genome sequence.

4.2 Results

A comparative epigenomic study should use cells of a single homogeneous type because different cell types have distinct epigenomes [94; 103], and cells should be uncultured because the epigenome can be distorted by cell culture [103]. Neutrophils are abundant circulating cells that are morphologically indistinguishable in humans and chimps, and can readily be isolated as a pure population without culturing. Since neutrophils in their circulating form are accessible and relatively homogenous, they are a suitable cell type for an interspecies comparison. We isolated neutrophils to $> 99\%$ purity from the peripheral blood of four young adult male humans (age 20-25 years old) and four age-equivalent male chimpanzees (age 12-16 years old, which is young adult, after accounting for differences in age of maturity [46]). To further attempt to control environmental variation, we selected individuals who were healthy, well nourished, afebrile, and not part of any study of infectious agents or other treatments.

In each sample, DNA was isolated from neutrophils and the methyl-Seq procedure was carried out: the DNA was digested with HpaII, 50-300 bp fragments isolated from an agarose gel, Illumina sequencing libraries constructed, and sequencing carried out on an Illumina sequencer. Single-end reads were quality filtered and aligned to their respective genomes with Bowtie [85], which produced stacks of reads at digested HpaII sites. Using methyl-Seq data and a reference genome sequence, MetMap assigns to each HpaII site within the scope of the experiment a probability of being unmethylated p(U) and a probability of being part of an unmethylated region p(I) (see Chapter 3 for a detailed description of MetMap and its output). MetMap produces a reduced representation survey of the methylome: of the 28,163,863 CGs in the human genome (the sites that can be methylated), 2,292,175 fall within a HpaII site, and of these 1,349,376 are within the scope of the experiment; similarly, there are 26,602,442 CGs in the chimpanzee genome; 2,122,178 fall within a HpaII site, of which 1,197,715 are within the scope of the experiment.

4.2.1 Characteristics of the human and chimp methylomes

MetMap annotates CpG island-like regions, called strongly unmethylated islands (SUMIs), based on experimental evidence of their unmethylated state. We annotated 20,986 SUMIs in the human neutrophil, that are present in at least one individual (Table 4.1). This set largely overlaps with the reference CpG island annotation, but 4,651 have not previously been annotated as CpG islands

	Human1	Human2	Human3	Human4	Human
SUMIs	16,904	17,596	18,179	18,700	20,986
not CGI	2,832	2,587	2,905	3,425	4,651

	Chimp1	Chimp2	Chimp3	Chimp4	Chimp
SUMIs	19,953	17,799	17,973	17,331	21,370
not CGI	4,298	3,015	3,163	2,836	5,228

Table 4.1: Summary of unmethylated islands identified in the human and chimp samples. The number of islands identified in each individual is shown in the row labeled "SUMI". The number of those SUMIs that do not overlap a CpG island (CGI) is shown in the row labeled "not CGI". The last column of each table shows the number of SUMIs and the number of SUMIs not overlapping a CpG island identified in at least one individual of the given species.

(Table 4.1). We obtained similar results for the chimp methylome (Table 4.1). Our comparison of the human and chimp methylomes used the 14,316 SUMIs that could be unambiguously mapped between the genomes of the two species.

Methylation probabilities calculated by MetMap were used to compare methylation states between human and chimp, revealing a high degree of similarity between the methylomes, but also significant differences that often involved the loss of a methylated state. Methylation states are more conserved at sites outside SUMIs than within them. At the 606,496 orthologous human and chimp HpaII sites that were not in SUMIs, methylation probabilities were highly correlated ($r^2 = 0.74$; Figure 4.1.A); 87% of these orthologous sites had $p(U) < 0.2$ (inferred probability of being unmethylated is smaller than 0.2), consistent with observations that CGs outside CpG islands are usually methylated [94; 103]. The 122,878 methylation probabilities at orthologous HpaII sites within SUMIs show a lower degree of correlation ($r^2 = 0.61$; Figure 4.1.B). However we observed a higher interspecies correlation when using the average p(U) calculated over all HpaII sites in each SUMI ($r^2 = 0.65$; Figure 4.1.C). The higher conservation of the methylation state of whole SUMIs, relative to the state of individual CGs within a SUMI, suggests that the average methylation of SUMIs is more informative in a comparative study.

4.2.2 Differentially methylated islands

We used permutation analysis to empirically define thresholds and identify SUMIs whose average methylation differs significantly between human and chimp (Methods). Among the 14,316 orthologous human and chimp SUMIs, we identified 1,525 that were significantly different at $p < 0.01$. The differentially methylated SUMIs have a length distribution and CG content very similar to SUMIs in general (Figure 4.2). Differential methylation of SUMIs between human and chimp is unlikely to be due to substitutions or polymorphisms at CGs in their genomes: analysis of CG dinucleotide content in these SUMIs indicates that 20% of differentially methylated SUMIs, in-

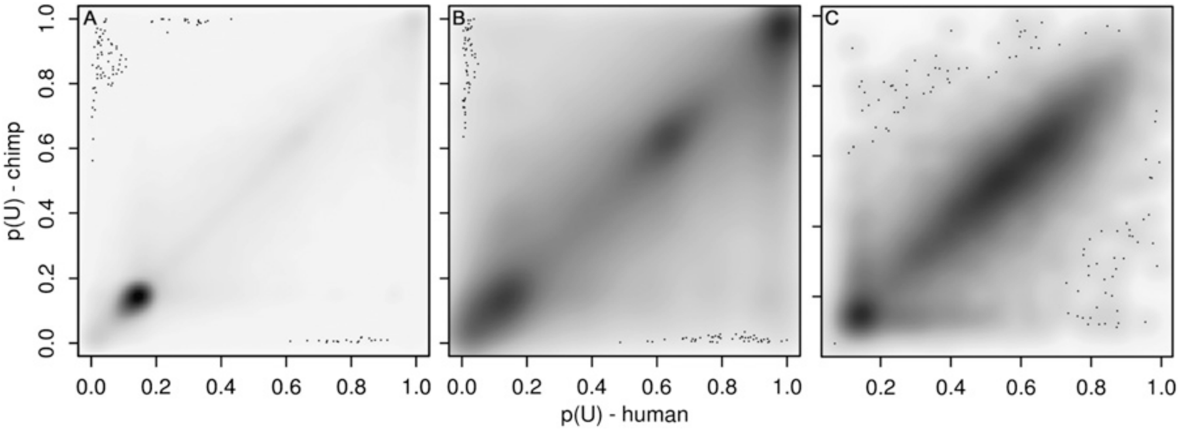


Figure 4.1: Comparison of the human and chimp neutrophil methylomes. Inferred probabilities of being unmethylated, $p(U)$, are plotted, with human on the x-axis and chimp on the y-axis. Low $p(U)$ indicates that a site is likely to be methylated, high $p(U)$ indicates that a site is likely to be unmethylated. Individual sites are plotted; grayscale intensity is proportional to the number of sites at each position. (A) Methylation probabilities for 606,496 orthologous human and chimp HpaII sites outside 14,316 orthologous human and chimp SUMIs. The sites are highly correlated ($r^2 = 0.74$, $p < 10^{-3}$ by permutation analysis) indicating conservation of methylation states between the species. (B) Methylation probabilities for 122,878 orthologous human and chimp HpaII sites within the 14,316 orthologous SUMIs. The sites are less correlated than sites outside SUMIs ($r^2 = 0.61$, $p < 10^{-3}$). (C) Mean methylation probabilities of the 14,316 orthologous human and chimp SUMIs ($r^2 = 0.65$, $p < 10^{-3}$). The distribution of methylation probabilities along the diagonal appears to be bimodal, with a cluster at $p(U) < 0.2$ and a cluster at $0.3 < p(U) < 0.9$; 15% of HpaII sites within SUMIs are methylated in at least one species.

cluding some of the most strongly differentially methylated, have no difference in CG content (Figure 4.3); restriction of the analysis to HpaII sites whose presence in both species could be confirmed by MspI digestion identifies essentially the same SUMIs as being differentially methylated; similarly, restriction of the analysis to HpaII sites that are not polymorphic in dbSNP produced the same results.

The differentially methylated SUMIs are associated with epigenetic features, including open chromatin (from FAIRE data) and histone tail modifications, that are consistent with these SUMIs' involvement in gene regulation (Table 4.2). Differentially methylated SUMIs are also modestly but significantly (p= .005) more likely than non-differentially methylated SUMIs to be associated with genes that were found by Blekhman et al. [15] to be differentially expressed in liver, heart, or kidney of human and chimp. Differentially methylated SUMIs are often found near transcription start sites, but not as frequently as SUMIs in general. Taken together, these data suggest that SUMIs in general, and differentially methylated SUMIs in particular, participate in transcriptional

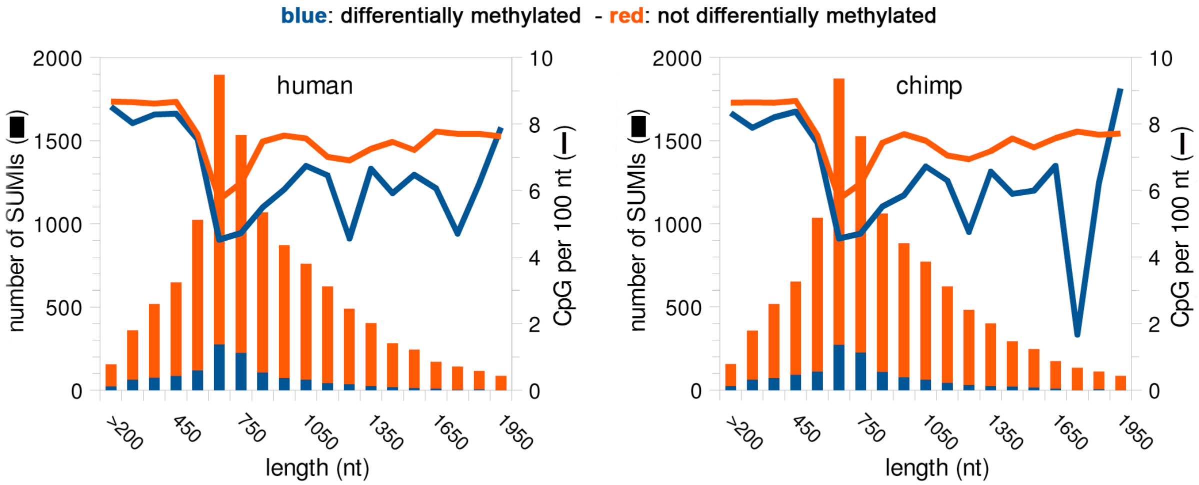


Figure 4.2: Length distribution and average CpG content of orthologous human-chimp SUMIs and differentially methylated SUMIs. The X axes show the lengths human (left) and chimp (right) SUMIs; the number of SUMIs of a given length is shown on the left-hand Y axes. The length distribution of SUMIs is similar between the two species, and between SUMIs that are differentially methylated (blue bars) and those that are not (red bars). This length distribution is also similar to that of computationally defined CG islands, as expected because these sets largely overlap [148]. CG content is reported as number of CG per hundred nucleotides (scale on right-hand Y axes). The CG content of differentially methylated SUMIs (continuous blue line) is slightly lower than in SUMIS that are not differentially methylated (red lines).

regulation.

4.2.3 The human, chimpanzee and orangutan methylomes recapitulate the phylogenetic tree

The orangutan methylome provides an outgroup to infer the ancestral state of the methylation differences noted between human and chimp SUMIs. Determination of the methylome of a single young adult male orangutan with the same procedure used for human and chimp identified 11,718 orthologous human-chimp-orang SUMIs. To determine if characteristic methylation states identify a species, we constructed a phylogenetic tree based on mean methylation probabilities of all 11,718 SUMIs that are orthologous in human, chimp, and orang, using each of the four humans, four chimps, and one orang as an independent operational taxonomic unit (Figure 4.4). To assign a SUMI to either a methylated or an unmethylated state, methylation probabilities were made binary using a stringent threshold of $p(U) = 0.2$. We calculated a distance matrix using Jukes Cantor-corrected Hamming distances and built a tree using Neighbor Joining. The tree recapitulates the

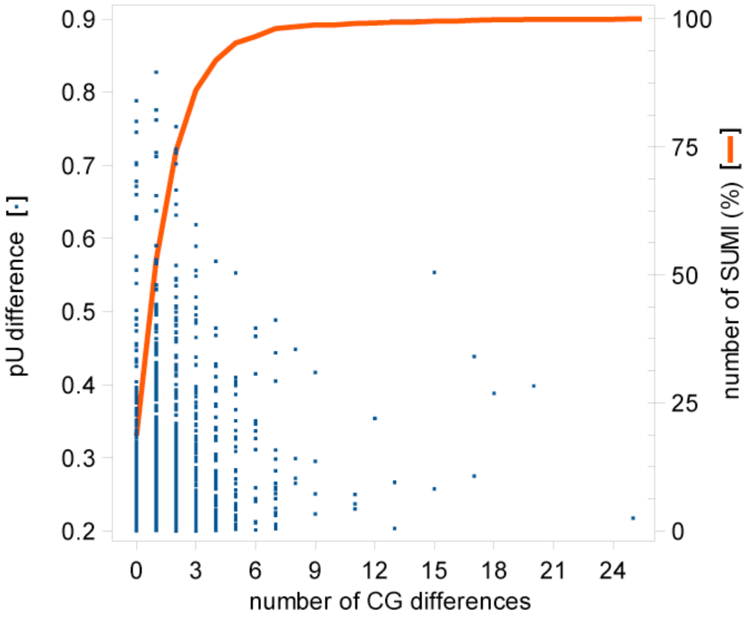


Figure 4.3: Differential methylation in human and chimp SUMIs is not related to gain or loss of CpGs in one of the two species. The blue dots indicate individual differentially methylated SUMIs (using the threshold of 0.2 to determine differential methylation), with the pU difference shown on the left-side Y axis and the number of CG differences on the X axis. The red line (scale on right-side Y axis) represents the running sum of the number of differentially methylated SUMIs with a number of CG differences between human and chimp smaller or equal to the value indicated on the X axis. Approximately 20% of the differentially methylated SUMIs have 0 CG differences, and approximately 50% have 0 or 1 CG differences. SUMIs with greater differential methylation do not have a larger number of CpG differences. Most SUMIs with the largest differential methylation have a small number of CG differences.

established phylogeny of the three species, with orang as the outgroup and all human individuals clustering together on a branch that is separate from the chimpanzee cluster (Figure 4.4). Bootstrap analysis indicates that this topology is highly significant. The tree also indicates, together with figure 4.5, that methylation of SUMIs has changed more frequently in human than in chimp, relative to the ancestral state. We obtained similar trees (with the same clusterings) using the subset of SUMIs that have the same numbers of HpaII sites in human and chimp, the subset of HpaII sites whose presence is confirmed by MspI digestion and deep sequencing, and the subset of HpaII sites that do not have sequence polymorphisms reported in dbSNP. The similar structures of these trees indicate that the tree structure in figure 4.4 is not an artifact due to sequence changes in one of the species. A tree built from the methylation states of all orthologous HpaII sites, irrespective of their location within a SUMI or not, also recapitulates the phylogeny of the three species. The mecha-

	FAIR	H3K27ac	H3K4me2	H3K4me3	H3K27me3
Orthologous HC	71%	68%	95%	87%	76%
Differentially methylated	67%	62%	92%	79%	79%

Table 4.2: Association between SUMIs and chromatin features. Percentage of overlap of SUMI sets with chromatin features associated with transcriptional regulation. All percentages of overlap were assigned $p < 0.02$ determined by 500 random permutations of the genomic locations of the different SUMI sets.

nisms leading to the methylation differences between species are unknown. The separate clustering of humans and chimps is consistent with the stable inheritance of methylation states within the two species; however, it does not demonstrate that those changes were driven by selection, or establish their functional significance, and it is also possible that at least some of the differences we observe are caused by factors in the separate environments of the humans and chimps in this study [29].

4.2.4 Neutrophil methylation is indicative of germline methylation

The apparent heritability of methylation states could simply reflect the determination of neutrophil methylation states by genome sequence. However it could also stem from stable maintenance of methylation states, or other epigenetic marks that determine methylation states, in the germline (i.e. pure epigenetic inheritance) [139]. Taken together with the tree structure in figure 4.4, evidence of a correspondence between somatic and germline epigenetic states would support epigenetic inheritance, although it could not prove it because of the possible dependence of epigenetic states on genome sequence or environmental factors.

The possibility that methylation states are heritable raises the question of whether the neutrophil methylation states are related to germline methylation states. If they are, then SUMIs identified as methylated in the neutrophil should have a higher rate of CG decay than SUMIs that are unmethylated in the neutrophil. Methylated CGs undergo mutation to TG (but not to AG or GG) much more frequently than unmethylated CGs [35]. The mutation is heritable if it occurs in the germline; this is the basis for the underrepresentation of the CG dinucleotide in vertebrate genomes [159]. We identified the subsets of orthologous SUMIs that had consistent methylation levels in human, chimp, and orangutan, varying from highly methylated to highly unmethylated, retrieved their sequences, and used Ambiore [70] to determine rates of the different substitution types involving the C in a CG dinucleotide. The rate for CG to TG transition was proportional to the probability that a SUMI is methylated (Figure 4.6). In contrast, rates of CG transversion to either AG or GG were independent of the neutrophil methylation state of a SUMI. These results are consistent with the hypothesis that neutrophil methylation states are related to germline methylation states. These inferences of germline methylation based on neutrophil methylation are in good correlation with the published methylome of an ES cell line determined by whole-genome bisulfite sequencing (correlation when considering all HpaII sites was 0.43). These results reveal a

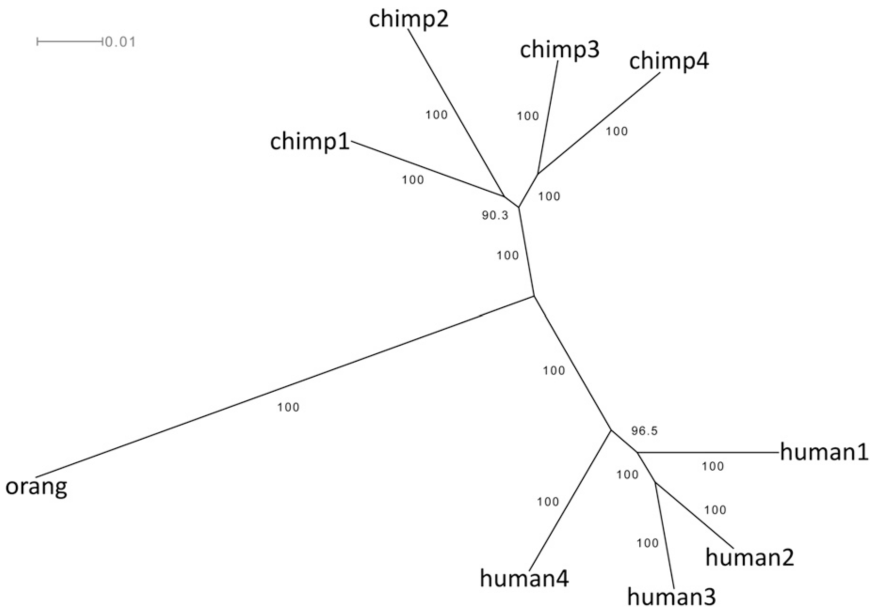


Figure 4.4: Phylogenetic tree built from mean methylation probabilities of the 11,718 orthologous human-chimp-orang SUMIs. The separate clustering of the human and chimp specimen indicates that certain methylation states are stably inherited within each species. The bootstrap values (1000 permutations) are shown next to each branch. The scale bar indicates the number of substitutions per site.

relationship between the neutrophil methylome and the methylome of germ cells, but do not mean that methylation is maintained at all stages of germ cell differentiation: germ cells undergo multiple stages of differentiation, and the pattern of CG decay indicates only that methylation is present in at least some of those stages.

We asked if the correlation between methylation states in the neutrophil and the germline is also valid for SUMIs whose methylation state has diverged in human, since these are responsible for the clustering of human and chimp on the tree in figure 4.4. If the correlation is valid, we should observe an excess of C/T polymorphism in human SUMIs that have become methylated. We compared the frequency of C/T polymorphism obtained from dbSNP130 in 312 human SUMIs that became methylated, and 438 SUMIs that became unmethylated, in human relative to the ancestral state inferred using the orang methylome. In regions that became methylated, we counted 71 CG to TG polymorphisms out of 10611 total CG sites; in regions that became unmethylated there were 42 CG to TG polymorphisms out of 13966 total CG sites. The difference in polymorphism counts is highly significant (p=3.6*10-05, binomial 2-sample proportion test). This data indicates that some human-specific changes in neutrophil methylation reflect changes in germline methylation states.

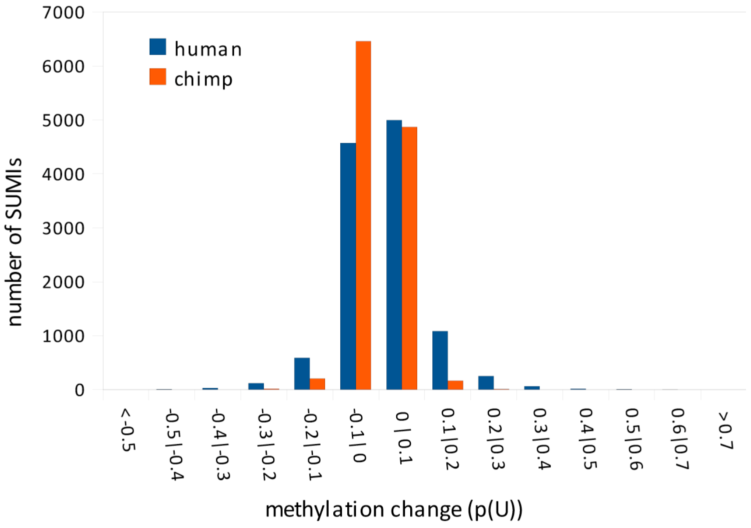


Figure 4.5: Distribution of methylation change, relative to the last common ancestor, in human and chimp SUMIs. The amount of change in methylation state of a SUMI was calculated as the difference between the p(U) values of the common ancestor, estimated from human, chimp and orang methylation states, and the extant species. The value is positive if the SUMI in the extant species is less methylated than the common ancestor, and negative if it is more methylated. The distributions of the amount of methylation change are shown for human (blue) and chimp (red). A larger number of human SUMIs than chimp SUMIs shows more extreme changes in methylation (SUMIs at the tails of the distribution).

4.3 Discussion

We have carried out a comparison of methylation states in multiple primates. Our comparison of neutrophils in humans and chimps reveals differences that occur more frequently within CpG island-like regions and often involve a loss of methylation relative to the ancestral state. Humans and chimps segregate on separate branches of a tree built from the neutrophil methylation data, and the CG decay and C/T polymorphism rates indicate that neutrophil methylation is related to germline methylation states. The CG decay analysis was made possible by the availability of our multispecies methylation data (to determine subsets of SUMIs with the same methylation state in the species analyzed) combined with the availability of genome sequences (to calculate substitution rates). It reveals regions whose methylation state is functionally constrained in human, chimp and orang.

The choice of cell type for our comparative study was dictated by our requirement for an accessible and homogeneous cell. Somatic tissues are composed of multiple differentiated cell types that have different epigenotypes; differences in the proportions of these cells can potentially

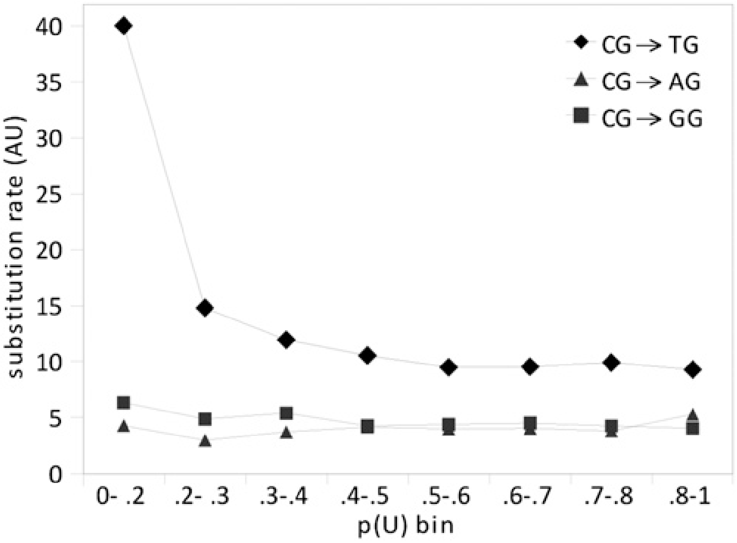


Figure 4.6: CG decay rates in orthologous human, chimp, and orang SUMIs. Substitution rates were calculated for all different types of substitutions involving the C in a CG dinucleotide in orthologous human, chimp, and orang SUMIs that have consistent methylation levels in the three species. Only the CG to TG substitution rate is expected to be affected by germline methylation. SUMIs in each $p(U)$ bin have neutrophil methylation probabilities within the indicated bin boundaries in human, chimp, and orang (i.e., SUMIs considered in this analysis did not change neutrophil methylation state during these species' evolution). Rates of CG to TG, but not to AG or GG, substitution vary as a function of methylation, in a manner consistent with neutrophil methylation states reflecting germline methylation states.

have a dramatic impact on the apparent epigenotype of a tissue. Blood is the most accessible tissue, but nucleated blood cells are made up of several cell types in proportions that can vary widely among individuals or even at different times in a single individual. Furthermore some types of blood cells, such as B and T lymphocytes, are made up of multiple subtypes. Neutrophils are among the most abundant nucleated blood cells, and they are the most homogeneous. They mature in the bone marrow and circulate briefly (∼12 hours) as mature cells. Immature forms, principally the band form, make up only a small proportion in healthy individuals (and were less than 5% in our subjects). Neutrophils form part of a system of non-specific immunity: they engulf and destroy microorganisms [112] in vertebrates and other animal species using mechanisms that do not require antigen-specific interactions [145]. This deeply conserved function is unlikely to have changed much in human evolution.

The SUMIs identified by MetMap are defined by criteria that include both CG content and methylation status, so that they are functional equivalents of CG islands. The number of SUMIs that are methylated in the germline may be larger than established by our study, for two reasons. First, because we wanted to compare CG decay rates in methylated vs. unmethylated SUMIs, we restricted our analysis to SUMIs that were consistently methylated or unmethylated in all three species. It is possible, or even likely, that some SUMIs are methylated in the germline of one of the species but not the others, but our analysis cannot address this. Second, there may be SUMIs that are methylated in the germline but unmethylated in neutrophils. A full definition of the set of SUMIs that are methylated in the germline would require analysis of many cell types.

The findings that somatic methylation states are related to germline methylation states, and that somatic methylation states recapitulate the phylogeny of human, chimp, and orang, raises some intriguing points. First, the phylogenetic trees built from methylation states show that epigenetic states can be maintained as characters that are predictably transmitted within a species. The ability to reconstruct phylogenies is not unique to CG methylation states: many phenotypic characters can be used for this purpose [43]. However these characters are not independently heritable, but reflect genomic sequences that encode the characters, i.e. the character merely acts as a surrogate for information encoded in the genome. This is not necessarily the case with CG methylation: while it may be determined by underlying genotype, it differs from other characters because it is a covalent modification of DNA itself. Thus it is intriguing to consider that the ability to reconstruct phylogenies using methylation states need not be a simple reflection of the inheritance of DNA sequence, but may instead reflect heritable epigenetic information carried by the methylation states themselves. It is nevertheless difficult to rule out genetic control of, or contribution to, the epigenetic states we observe. The relationship between methylation state and genotype is complex: examples exist in which a methylation state is completely determined by the DNA sequence on which it resides (obligate), is influenced but not determined by the DNA sequence (facilitated), or is purely epigenetic and can change without any change in DNA (Richards 2006) . Our data are consistent with any or all of these scenarios.

Furthermore the evidence for germline methylation states raises the possibility that epigenetic states are inherited directly, but does not demonstrate that methylation states per se are maintained and inherited in the germline. Germ cells go through a number of phases of differentiation, in some of which methylation is largely removed from the genome and subsequently replaced [126; 135]. The evidence for epigenetic resetting in germ cells implies that if methylation states are heritable, they must be so either because they are determined by DNA sequence, or because methylation is faithfully reestablished due to the retention of other components of the epigenome. piRNAs have been implicated in setting of germline methylation states and are good candidates for such a role [3].

Second, CG decay and C/T polymorphism analyses indicate that somatic methylation of SUMIs often reflects the presence of methylation at some stage in the germline. This observation is made possible by the availability of comparative methyltyping data, which has allowed us to obtain substitution rates in sequences with the same known methylation state. It raises the possibility that there are previously unsuspected constraints by germline methylation states on somatic states, i.e. that epigenetic information in the germline determines to some extent the epigenetic state of somatic cells. Phenotypic differences mediated by epigenetic inheritance would necessitate this type of control; however as discussed above, our finding does not demonstrate or require that germline epigenetic states be independent of genotype.

Regardless of the means (genetic or epigenetic) by which inheritance of the methylation state of a regulatory element is mediated, deviation from the inferred ancestral methylation state implies a change in its functional potential. King and Wilson suggested that mutations in transcriptional regulatory sequences would account for the phenotypic divergence of human and chimp [81]. A change in methylation state can accomplish the same thing as a regulatory mutation, without sequence change. In particular, loss of a methylated state provides a simple mode by which regulatory activity could be expanded without requiring gain-of-function through changes in DNA sequence. We speculate that a SUMI that is methylated in the germline remains methylated in most somatic cell types, but is active in a restricted set of cell types in which it is demethylated; germline transition to a demethylated state might broaden the spectrum of somatic cell types in which such a SUMI is active, thus in effect creating a regulatory sequence in the cell types that gain the new activity. Thus germline changes in methylation states could readily have functional and potentially adaptive consequences. This model of regulatory evolution is simpler than one requiring sequence change to create one or more transcription factor binding sites, but it remains to be seen if methylation states change without associated sequence change. Epigenetic differences, such as those we identified between the human and chimp methylomes, may be a novel source of variation to explain interspecies phenotypic divergence, and possibly phenotypic variation within a species.

4.4 Methods

Sample collection and isolation

Animal samples were collected with IACUC approval. Human samples were collected with IRB approval after obtaining informed consent. We obtained blood samples from four young adult male humans (age 20-25 years old) and four age-equivalent male chimpanzees (age 12-16 years old, which is young adult, after accounting for differences in age of maturity(Fleagle [46] p.257)). Young adults are fully developed but have not yet undergone age-related changes. To further attempt to control environmental variation, we selected individuals who were healthy, well nourished, afebrile, and not part of any study of infectious agents or other treatments. Immediately after phlebotomy, leukocytes were isolated by Ficoll centrifugation. Neutrophils were isolated from the leukocyte fraction with CD-16 microbeads (Miltenyi). An aliquot of each specimen was Wright-Giemsa stained and examined microscopically; all specimens contained >99% neutrophils.

Generation of methyl-Seq libraries

DNA was extracted by standard methods, and digested overnight with HpaII. HpaII cuts the sequence CCGG; methylation of the central cytosine on one or both strands protects the sequence from digestion with HpaII [63]. HpaII fragments 50-300 bp in length were isolated on an agarose gel. Single-read sequencing libraries were constructed from human and chimp samples using the standard Illumina kit, and sequenced on an Illumina GA to collect reads of 32 bases. A paired-end sequencing library was constructed from the orangutan sample and sequenced on an Illumina GAII to collect paired reads of 36 bases; only the first read of the paired-end sequencing reaction was analyzed in this study. As a control for the HpaII digestion, the DNA of human sample 1 and chimp sample 1 was digested with MspI, which cuts the sequence at CCGG and is methylation insensitive. Single read libraries were generated and sequenced as described for the HpaII libraries. Sequencing data were processed by the Illumina Pipeline for base calling and quality filtering. The first three sequencing cycles (corresponding to the 'CGG' sequence from the digested HpaII sites) of the orang sample were skipped to facilitate cluster calling. Only reads passing the Illumina quality filter (chastity filter = 0.6) were further processed. The sequence data have been deposited in the NCBI GEO database, with accession number GSE22376.

Generation of methylation maps from sequencing data

For a detailed description and explanation of the computational pipeline for analysis of methyl-Seq data, see the previous chapter of this thesis and [148]. Quality-filtered reads were aligned using Bowtie v0.9.9.2 [85] to their respective reference genomes, which were retrieved from the UCSC Genome Browser [137]: human genome (hg18, March 2006), chimpanzee genome (panTro2, March 2006), orangutan genome (ponAbe2, July 2007). We used an alignment policy that allows up to two mismatches in the first 28 bases and reports only reads that align with a single best match (parameters: -all -m 1; in Bowtie v 0.9.9.2, hits are "stratified" by default). Reads whose 5' end aligned to a CGG corresponding to a HpaII site were analyzed with MetMap to assign to each HpaII site within the scope of the experiment a probability of being unmethylated $p(U)$ and a probability of being part of an unmethylated region $p(I)$. For each of these sites, a species $p(U)$ was determined by averaging the $p(U)$ values of the four individuals (human and chimp) of that species at that site. For orang, we used the $p(U)$ values from the single individual analyzed.

Cross-genome mapping of SUMIs and HpaII sites

We define as orthologous a HpaII site or a SUMI that passes the following LiftOver procedure: a site or SUMI is mapped from the chimp or the orang genome to the human genome using LiftOver (hgdownload.cse.ucsc.edu/admin/exe/), and then mapped back to the first genome to ensure that it has a unique correspondent in both genomes. Furthermore, all SUMIs whose sequence contained a stretch of more than 10 contiguous 'N' in at least one species was excluded from the analysis.

Comparison of human and chimp methylomes

The human and chimp methylomes were compared using scatter plots of the $p(U)$ values of the 729,374 orthologous HpaII sites within the scope of the experiment or of the mean $p(U)$ values of the 14,316 orthologous human-chimp SUMIs. The significance of the correlation values of each scatter plot was determined by permutation analysis (1000 permutations). To determine the significance of differences in the methylation state of human and chimp SUMIs, we generated a null mean-$p(U)$ value distribution assuming that there is no significant difference in methylation between the four human and the four chimp samples. We generated groups for the null distribution by considering all divisions of the human and chimp individuals into two groups, such that in each group there were two humans and two chimps. The distribution of the absolute differences in mean-$p(U)$ values was used to set the thresholds for p=0.01 at absolute difference 0.19. Of the 14,316 SUMIs considered, 1,525 has an absolute difference in methylation that passed the threshold, and were therefore determined to differ in methylation between the species with a p-value < 0.01.

Overlap between SUMIs and chromatin features

Human genome (hg18) annotation of chromatin features was obtained from the UCSC genome browser. FAIRE data (determining open chromatin regions) was obtained from
ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeChromatinMap/ using the files:
wgEncodeUncFAIREseqPeaks{H1hesc,Nhek,Gm12878V2,Huvec,Panislets}.narrowPeak . Histone tail modification data were obtained from the EncodeBroadChipSeq dataset of the UCSC browser, using the union of the data for the cell lines: GM12878, H1hESC, HMEC, HSMM, HUVEC, NHEK, NHLF (H1hESC data was not available for H3K27ac). A SUMI and a chromatin feature were scored as overlapping if they shared at least one base. To evaluate the significance of the overlap between SUMIs and chromatin features, the location of the SUMIs was randomized 500 times using shuffleBed [129]. For each randomization, we computed the overlap between randomized SUMIs and chromatin features; p-values are reported as the frequency of the number of times a degree of overlap of a given chromatin feature with randomized SUMIs was equal or greater than that with the original SUMI data.

Association between differentially methylated SUMIs and differentially expressed genes.

We considered only SUMIs that have exactly one Transcription Start Site (TSS), as defined by the refGene table of the UCSC Genome Browser, within ±2000 bp from the SUMI boundary. Out of 14,316 orthologous human-chimp SUMIs (HC SUMIs), there are 6,457 such cases; out of the 1,525 differentially methylated SUMIs (diffSUMIs) there are 507 such cases. Each of these SUMIs was associated with the its proximal RefSeq gene. Blekhman et al generated human-chimp differential gene expression data for 17,231 genes from three tissues [15]. The intersection of the sets of SUMIs proximal to a TSS with the genes studied by Blekhman et al identified 5,277 HC SUMIs and 384 diffSUMIs for which we had gene expression data. Of the 384 diffSUMIs, 180 (46.9%) matched a gene that was differentially expressed, as determined by Blekhman, in at least one of the three tissues tested. In contrast, of the 5,277 general HC SUMIs, 2,139 (40.5%) matched a gene that was differentially expressed in at least one of the tissues they tested. To evaluated the significance of this difference, we randomized the association between SUMIs and genes and generated a p-value after 1,000 iterations - in each iteration 384 SUMIs were picked at random from the 5,277 SUMI set, and the number of those SUMIs for which the associated gene was determined as differentially expressed in one of the tissues was counted. In 5 of the 1,000 cases that number was larger or equal to 180, resulting in a p-value of 0.005 for the association of differentially methylated SUMIs with differentially expressed genes.

Inference of methylation state in the last common ancestor.

We used the orangutan methylome as the outgroup to infer the ancestral methylation state of human and chimp SUMIs. Out of the 14,316 orthologous human-chimp SUMIs, we identified 11,718 that had an orthologous orang SUMI meeting the same criteria described above in the section on cross-genome mapping. Only one unrooted tree topology is possible for any three species; for each orthologous human, chimp and orang SUMI, we calculated the branch length of the unrooted tree from mean $p(U)$ values of the three species using the REML method for continuous traits [43]. Given the branch lengths, the methylation state of the common ancestor for each SUMI was inferred using squared-change parsimony . We calculated the amount of change in methylation state as the difference for each SUMI between the $p(U)$ values of the common ancestor and the extant species. The value is positive if the extant species is less methylated than the common ancestor, and negative if the extant species is more methylated.

Construction of a phylogenetic tree based on methylation states

We built a phylogenetic tree based on the average methylation states of the 11,718 orthologous human-chimp-orang SUMIs. Each SUMI was assigned $p(U)$ value of 1 if its mean $p(U)$ score was larger than 0.2 and 0 otherwise (this conservative threshold of 0.2 for calling a SUMI 'methylated' is consistent with the CG decay analysis) The Jukes-Cantor distances (for binary characters) were calculated for each pair of individuals to obtain a distance matrix. We used the SplitsTreeprogram [69] to construct a phylogenetic tree using the Neighbor-Joining algorithm, and to bootstrap the resulting tree (1000 permutations) [43].

CG decay analysis

From the set of the 11,718 orthologous human, chimp and orang SUMIs, we determined the subsets of SUMIs in which all the three species had mean $p(U)$ within defined thresholds, computed multiple sequence alignments using ClustalW [86], and concatenated the alignments of all SUMIs within each subset. The concatenated multiple sequence alignments were submitted to Ambiore [70] to calculate the substitution rates of all possible substitution types involving the C of a CG dinucleotide.

To analyze CG decay in SUMIs that had changed methylation state in the human lineage, we identified human SUMIs whose methylation difference with the inferred last common ancestor was $p(U) < -0.19$ (identifying SUMIs that have become more methylated) or $p(U) > 0.20$ (identifying SUMIs that have become less methylated). For each SUMI, C/T polymorphisms mapping to a CG dinucleotide were retrieved from dbSNP build 130; only polymorphisms validated as "by-hapmap" and "1000genome" were used.

Part II

Error correction in HTS datasets

Chapter 5

Identification and correction of systematic error in high-throughput sequence data

5.1 Introduction

The technological advances that have enabled high throughput sequencing have opened the door to many fascinating fields. A substantial and crucial step in many of the studies that use such high-throughput data is determining the positions of single-nucleotide variants[1] . Calling of such variants is attempted by pinpointing places at which the sequencing data gives evidence of single-nucleotide resolution variance in the sample (for example - an individual that has a 'T' at some specific location on one chromosome and a 'C' at that same location on the second chromosome). Examples in which high-throughput sequence data is used for this purpose include studies that map rare variants across populations from the sequencing of individual genomes, RNA-editing events from RNA-Seq data, allele-specific DNA methylation from methyl-Seq datasets, and studies that determine differential allelic expression from RNA-Seq data, to name a few. Since high-throughput sequencing technologies allow sequencing at low cost at the expense of higher error rates [38; 67], improved statistical methods that accommodate these high error rates are needed in the calling of variants (heterozygous sites) [114]. The design of effective statistical methods requires precise characterization of error in high-throughput sequence data. Previous work has examined the behavior of individual base-call errors in sequence reads [38; 111; 161]. In this chapter we discuss a previously undescribed phenomenon in sequence data where these base-call errors aggregate at specific genomic locations across multiple sequence reads. We focus on Illumina technology, although we have observed systematic error on other platforms and return to this in the Discussion.

[1] A variant is a case in which at least two different nucleotides from {A,G,C,T} are present at some genomic coordinate, *x*, in the sample being sequenced. We then say that there is a variant at location *x*. Throughout this chapter we use variant and heterozygous site interchangeably.

We begin by describing the types of sequencing error that have been previously characterized, and their relationship to the type of error we have discovered. The likelihood of a base-call error occurring at any particular location in a sequence read is influenced by several parameters. It is known that base-call errors are more likely towards the ends of reads and that surrounding sequence motifs influence error frequencies [38; 111; 161]. For example, errors are more likely at positions preceded by GG or following a number of GGC motifs [111], but regardless of the preceding motif, errors become more likely towards the end of reads [38]. However, we have found that errors at some *genomic* positions appear with greater frequency than can be explained by these effects, and we refer to this as *systematic error*. Systematic error manifests as many individual base-call errors from separate sequence reads occurring at the same genomic position (Figure 5.1). Thus, a systematic error comprises many individual base-call errors (from different reads) that fall at the same genomic location.

These errors have the potential to be especially troublesome because they can confound methods that identify errors based on their sparsity among reads. For example, we show systematic errors affect current SNP (Single-Nucleotide Polymorphism) calling methods, where the first step involves computing the posterior probability for a SNP at every site based on relative nucleotide counts [91]. Although filters based on the depth of reads are frequently applied (mostly to screen for indels, copy number variants, or other structural variation) [1; 169], most existing approaches will not identify systematic errors, or distinguish them from true SNPs. Similarly, the detection of RNA editing sites in RNA-Seq data is complicated by systematic error, because an accumulation of errors at a transcriptome site can appear to be an edit event when compared with a reference genome that may have been sequenced using another technology [92]. In many studies that incorporate variant-calling from high-throughut sequence data, such as the examples noted in the first paragraph, the frequency at which a variant in the sequenced reads is observed is not expected to be 0.5. For example, in the case of a population study that aggregates low-coverage whole-genome sequencing data from many individuals, the observed frequency of a variant will be (in expectation) the frequency of that variant in the population being sequenced. We show that systematic errors are prevalent and are found in a large range of frequencies, requiring such studies to take special care to avoid falsely annotating systematic errors as variant calls.

In this paper we present a thorough characterization of systematic errors using Illumina short-read sequencing data. The data we use is optimized for the detection of errors because of high coverage and high numbers of paired-end reads in which the paired reads overlapped. We show that systematic errors must be accounted for when annotating variants from a dataset (heterozygous alleles), and that although improved base calling software can correct a small number of systematic errors, it is not sufficient by itself. We present an efficient statistical algorithm for the detection of systematic error and use it to show that systematic errors are present in other datasets, including an RNA-Seq dataset, a viral reference genome and Illumina HiSeq 2000 data from the 1000 genomes project.

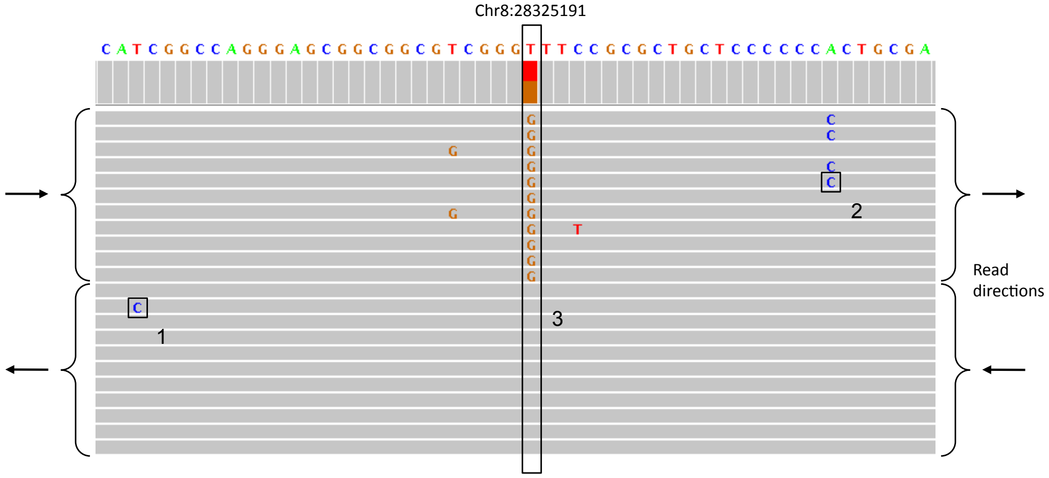


Figure 5.1: Types of errors. A screenshot from the IGV browser showing three types of error in reads from an Illumina sequencing experiment: (1) A random error likely due to the fact that the *position* is close to the end of the read. (2) Random error likely due to *sequence specific* error- in this case a sequence of Cs are probably inducing errors at the end of the low complexity repeat. (3) *Systematic error*: although it is likely that the GGT sequence motif and the GGC motifs before it created phasing problems leading to the errors, the extent of error is not explained by a random error model. In this case, all the base calls in one direction are wrong as revealed by the 11 overlapping mate-pairs. In particular, all differences from the reference genome are base-call errors, verified by the mate-pair reads, which do not differ from the reference. Given the background error rate, the probability of observing 11 *error-pairs* at a single location, given that 11 mate-pair reads overlap the location, is 1.5×10^{-26}. Moreover, given the presence of such errors at a single location, the probability that all of the errors occur on the same strand (i.e., on the forward mate pair) is $\frac{1}{1024} = 0.00098$. Note that the IGV browser made an incorrect SNP call at the systematic error site (colored bar in top panel).

5.2 Results

To investigate the types of errors present in whole-genome Illumina high throughput sequencing data, we conducted a paired-end methyl-Seq experiment on a male human individual with read length of 76 bp. A methyl-Seq experiment is ideal for investigating systematic error because the experiment results in high average coverage per covered location due to the fact that only sites cut by the restriction enzyme are assayed. The reads were mapped with Bowtie [85] allowing up to two mismatches.

Our experiment spanned 29,827,077 genomic locations at an average coverage of 35.4. Due to the small fragment size in methyl-Seq experiments many of the mate-pair reads overlapped, providing for each such location two base calls sequenced from the same DNA molecule (Figure 5.1) albeit from different directions. We made use of this to distinguish between base-call errors and true heterozygosity calls in the following manner: each pair of bases originating from a single mate-pair and sequencing the same position was denoted a *reference-pair* if both calls agreed with the reference genome, a *SNP-pair* if both calls disagreed with the reference genome and agreed among themselves, and an *error-pair* if one of the calls agreed with the reference genome but the other did not. A *SNP-pair* could consist of two base-call errors, in the case that both of the paired reads had an error at the same location, but the probability of such an event was low and we ignored such cases in this study.

Because we focused on overlapping mate-pairs, we report all results in terms of pairs. For example, when stating coverage we state the number of pairs overlapping a site (the coverage of the systematic error location in Figure 5.1 is 11), and when we state a location has 40% errors it means that of the pairs overlapping the location 40% were *error-pairs*. In our experiment 3,985,926 genomic locations were covered by both reads of some mate-pair but we restricted our analysis to the 2,226,445 of these locations with a coverage depth of at least 10. These 2,226,445 genomic locations where covered by a total of 85,782,923 base-call pairs, 223,957 of which were error-pairs.

5.2.1 Extent of systematic error

We found many locations at which there seemed to be an accumulation of errors. To test the extent of this phenomenon we computed the expected number of locations with each possible proportion of error. Let $c_{10},\ldots,c_j,\ldots,c_{565}$ be the number of locations with coverage j (in our data $\sum c_j = 2,226,445$), and $p := \frac{\#error-pairs}{\#pairs} = 0.002611$ be the probability of sequencing error. Let B_i be a random variable for the number of locations from $c_{10},\ldots,c_j,\ldots,c_{565}$ with proportion of errors i, and let B_{ij} be a random variable for the number of locations with coverage j and proportion of error i. We computed the expected number of locations to have each proportion of errors i as

$$E[B_i] = \sum_j E[B_{ij}] = \sum_j c_j \binom{j}{k_{ij}} p^{k_{ij}} (1-p)^{(j-k_{ij})},$$

where k_{ij} is the number of errors for coverage j that results in proportion of error i. Figure 5.2 shows a log-scale histogram of the expected and observed counts for these different error-proportions. The observed counts in the higher frequencies of errors are larger than the expected counts, indicating that there are more locations than expected that have a high frequency of base-call errors. We called such locations systematic errors, and set out to determine the characteristics of these locations, with the goal of lowering the false-positive rates in calling heterozygous sites.

For further characterization, we annotated a set of locations in which the number of *error-pairs* was significantly higher than expected, given the observed frequency of error. Setting $p = 0.002611$ as in the previous section, we compute a p-value for a given location with i errors and n coverage as $p(K \geq i|n) = \sum_{k=i}^{n} \binom{n}{k} p^k (1-p)^{(n-k)}$, where K is a random variable indicating the number of errors at a location. Of the 2,226,445 locations with coverage of at least 10, 2,116 locations were annotated as systematic errors, using a Bonferroni correction for a 0.05 signifi-

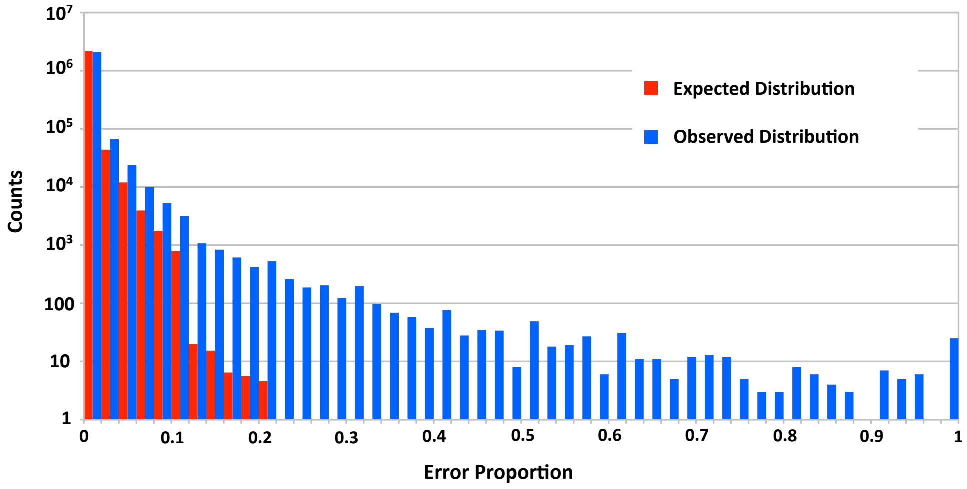


Figure 5.2: Proportion of base call errors across genomic sites. The observed (blue) number of locations with high base-call error frequencies significantly exceeds the expected amount (red).

cance level. We used a Bonferroni correction because it ensures that the probability of even one false-positive is ≤ 0.05, resulting in a set that is low in false-positives, and therefore suitable for characterizing the nature of systematic error. We note that this calculation yielded a lower bound on the frequency of systematic errors in our dataset of approximately 1 in 1000 bp.

5.2.2 Characterizing systematic errors

Having annotated the set of 2,116 systematic errors, we looked for characteristic features that could be identified in any high throughput sequencing experiment. Of the 2,116 sites we have determined as systematic errors, 953 had all base-call errors on the forward read and 1,062 had all base-call errors on the reverse read (an example is seen in Figure 5.1). We conclude from this that in systematic errors the base-call errors tend to appear on just one of the sequencing directions (forward or reverse). This tendency was noticed in [1], where the directionality on which errors occurred was used to filter false-positives from the set of heterozygous sites annotated. A possible explanation for this phenomenon is that the sequencing of some motifs, which are different on the opposite strands, have higher probability than others for base-call errors, resulting in systematic errors. This is consistent with the known overlap in absorption spectra of the G and T channels identified by a single laser in Illumina sequencing.

We therefore tested whether there are significant motifs surrounding systematic errors by generating a sequence logo [36; 144] for the reference sequences around the systematic errors (Figure 5.3). Interestingly, we found that the first base upstream of the systematic error has greater information regarding the presence of a systematic error than the base at which the error is present.

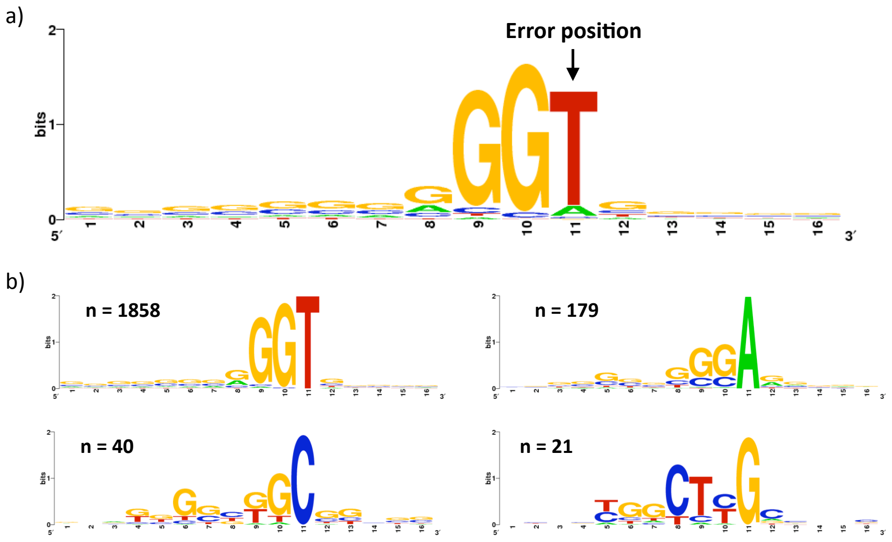


Figure 5.3: Sequence motifs at systematic error sites. (a) The motif around systematic errors reveals a strong enrichment for instances preceded by an occurrence of GG and for the error to occur at locations where the reference genome is T. (b) Categorized by the nucleotide at the error location. The number of systematic errors in each subset is denoted by n.

We found that the large majority of systematic errors are preceded by a G, and that two G bases followed by a T at the error site is by far the most common and characteristic sequence at systematic error locations. Although the GGT motif is a strong characteristic of systematic errors, an analysis restricted to GGT sites (estimating the expected error rate by that observed at GGTs, see Methods) showed that 660 sites, out of all 61,779 GGT sites, have a significant accumulation of errors . This shows that systematic errors are not accounted for by this motif alone.

To gain insight into the types of sequencing errors present at systematic errors we computed the frequencies of the different base substitutions in both systematic errors and throughout the entire dataset (Figure 5.4). We witnessed an extremely strong tendency for the $T > G$ error compared to all others. Our results show that there is a higher substitution rate to Gs than to the other nucleotides and that the substitution rate to A or T is considerably lower than the substitution rate to C. With respect to the reference bases at which systematic errors occur, there is a stronger tendency of error at A or T than at C or G. We divided the systematic error locations based on the reference base at which the error occurred, and tested for motifs in each of the fours sets (Figure 5.3.b). We concluded that the strongest motif at systematic errors is that of GGT where the error is at the T, resulting in an incorrect base call of G.

To test whether the quality scores at the locations of systematic errors account for the extent of base-call errors observed, we computed a p-value for each location given its specific quality scores: Given n (ordered) quality scores let K_i be a random variable for the number of errors at

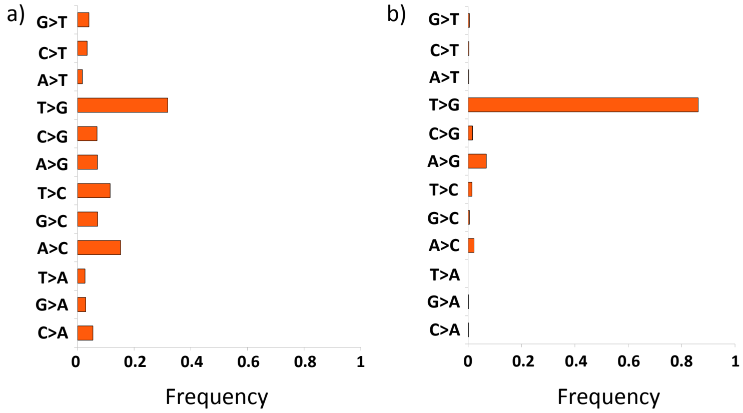


Figure 5.4: Base substitutions of systematic errors. Frequency of different base substitutions in (a) all errors (b) systematic errors.

locations 1 to i, and let X_i be an indicator variable for whether there was an error or not at location i. We then have that

$$p(K_n = k) = P(X_n = 1)P(K_{n-1} = k-1) + P(X_n = 0)P(K_{n-1} = k),$$

and can use dynamic programming to compute the p-value for each location in $O(n^2)$ time. Of the 2,226,445 positions with read count of at least 10, 268 had a significant accumulation of error under a Bonferroni correction for a significance level of 0.05 (the probability of even one false-positive is less than 0.05). It is interesting that significant positions were found, given that in general throughout the experiment the quality scores tend to predict a higher error rate than that observed ($\frac{\#error-pairs}{\#pairs} = 0.002611$ while the quality scores predict an error-pair frequency of 0.00416).

The characteristics of systematic errors, occurring mostly at GGT motifs where the error that occurs is a $T > G$ substitution, implies that the errors could be a result of the sequencing technology, which makes it hard to distinguish between a GGG and a GGT instance. It is the base-calling algorithm that makes such distinctions, given the images output from the Illumina machine. We asked whether systematic errors could be accounted for by base-callers that utilize sophisticated statistical techniques to reduce error. To test this we compared the systematic errors present in a dataset base-called by Bustard (Illumina's base-caller) to those present in the same dataset when base-called by naiveBayesCall [77], to our knowledge the most accurate base-calling algorithm available. We used for this the dataset that was used in [77] from the phiX174 virus. The genome of phuX174 is 5,386 bp long and has been extensively studied due to its use as a control for sequencing experiments. We found 59 systematic errors in the Bustard called dataset and 40 systematic errors in the naiveBayesCall dataset, amounting to a systematic error rate of 1 in 91 bp and 1 in 135 bp respectively. We believe the higher frequency of systematic errors is due to the phiX174 genome being richer than human in *GGT* motifs (data not shown) and to the high sequencing coverage (see Discussion section). These results show that while systematic error can be reduced with more sophisticated base calling, it is a persistent problem at a significant level even when using state of the art methods.

To test replicability of the locations at which systematic errors occur, we conducted a second methyl-Seq experiment on the same individual (see Methods section). The error frequency in this second experiment was determined as $p = \frac{\#error-pairs}{\#pairs} = 0.00162$ and of the 2,419,666 locations with coverage of at least 10 pair-calls, 3,272 locations were annotated as systematic errors using a Bonferroni correction of 0.05. From the 2,160,736 positions with at least 10 pair-calls in both of the experiments, 1,916 and 2,519 were annotated as systematic errors in the first and second experiments, respectively, and of those 1,279 locations were annotated as systematic errors in both experiments. This shows that while there is some variability in the locations determined as systematic errors, locations at which systematic errors occur are highly replicable (the expected number of systematic errors to be called at the same locations is 2). We tested whether the significant overlap of the locations at which systematic errors were detected was due to *GGT* motifs being more prone for systematic errors than other motifs. Of the 61,779 *GGT* sites that were overlapped by at least 10 pair-calls in each experiment, 1,596 and 2,080 locations were annotated as systematic errors in the first and second experiments, respectively, and of these 1,095 locations were annotated as systematic errors in both experiments (the expected number of systematic errors to be called at the same locations when restricting to *GGT* positions is 54). The lists of systematic errors for both experiments are available at: `http://bio.math.berkeley.edu/SysCall/systematic_error_lists/`.

5.2.3 Identification and correction of systematic errors

The main concern regarding systematic errors is that they may be incorrectly annotated as heterozygous sites in an individual or as rare variants in a population. Fortunately, in systematic error the extent of error at a location usually does not result in an equal ratio of reference to non-matching reference calls, making it easier for methods that expect such a ratio to identify these sites as non-SNPs. Nonetheless, SAMtools [91] identified 12 of the 2,116 systematic errors in our methyl-Seq dataset as SNPs (three of these are annotated as SNPs in dbSNP130), and in the SNP-calling procedure for the 1000 genomes project a filtering step based on directionality of sequencing was used to account for systematic errors (supplementary material of [1]). Systematic error may pose an even greater difficulty in studies where the ratio of reference to non-matching reference calls is not expected to be 1:1 for true heterozygous sites. This is the case in population studies that aim at characterizing rare variants, and in various functional genomic studies, such as RNA-Seq experiments in which variants are annotated alongside expression levels [178]. Systematic error may also affect RNA-Seq experiments in the bias it can introduce in coverage at systematic error sites. Such bias can in turn affect expression level estimates [163].

To account for this we have designed SysCall - a classifier which given a list of potential heterozygous sites and the reads from an Illumina experiment classifies each location as a system-

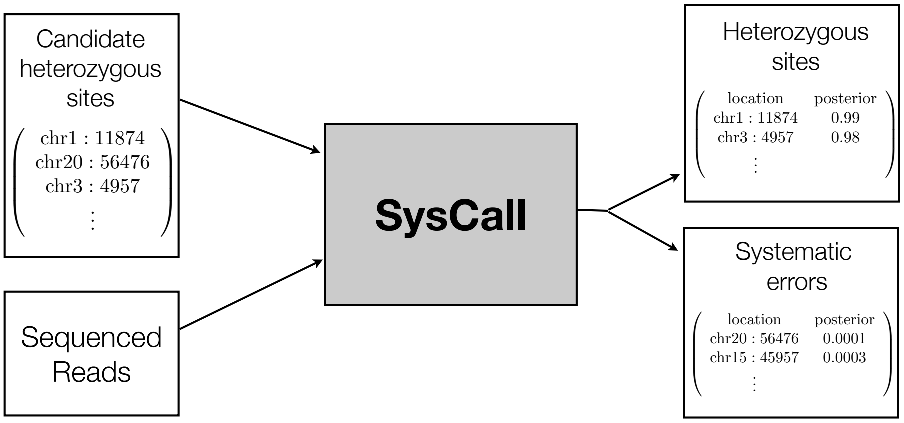


Figure 5.5: A flowchart of the SysCall classifier for distinguishing heterozygous sites from systematic errors.

atic error or a heterozygous site (Figure 5.5). Our classifier uses logistic regression to combine the different characteristics of systematic errors and make predictions. The SysCall algorithm is described in the Methods section. Importantly, SysCall does not assume that the experiment preformed is paired-end or that the expected frequency of variant observations is half, making it applicable to the different types of high throughput experiments discussed. SysCall is available at `http://bio.math.berkeley.edu/SysCall/`.

Assessing SysCall's performance

In order to test SysCall's performance we annotated a set of locations in our methyl-Seq dataset that would be candidates for heterozygous sites (where a significant amount of the base-calls differ from the reference) and for which using the overlap between paired reads we could call as systematic errors or heterozygous sites with high certainty. We used the same sets of locations that were annotated for training SysCall (see Methods section): a "SNPs" set consisting of 491 locations and a "Systematic errors" set consisting of 338 locations. From each mate-pair one of the reads was chosen at random to simulate a non-overlapping (and non paired-end) dataset.

As a first test of our classification algorithm we ran 100 iterations in which we generated training and test sets by randomly dividing the "SNPs" and "Systematic errors" sets into halves (from each of the "SNPs" halves 169 instances were randomly selected in order to have the same number of systematic errors and SNPs in the training and test sets). In each iteration we generated a feature matrix for the training and test sets, learned the coefficients of the logistic regression classifier from the training set, and classified the instances of the test set, recording the percentage of instances that were classified correctly (as either systematic errors or heterozygous sites). The distribution of the percentage of instances classified correctly from the 100 iterations had a mean

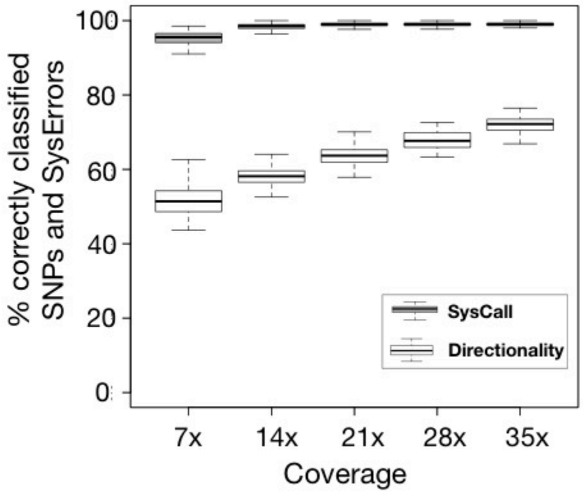


Figure 5.6: SysCall accurately distinguishes heterozygous sites from systematic errors. Proportion of correctly classified instances at different sequencing coverages for SysCall (grey) and for a logistic regression classifier that uses only the feature of directionality difference in error frequency (white).

of 99.0% and a standard deviation of 0.005 (Figure 5.6).

A strong characteristic of systematic errors is that the differences from the reference have a strong bias to occur on either the forward or reverse direction. We tested the ability to classify locations using the same logistic regression classifier but using only the directionality bias feature: $u_l = (q_{l1} - q_{l2})$. When running 100 iterations of training and testing as before using this classifier, the distribution of the percentage of instances classified correctly had a mean of 72.1% and a standard deviation of 0.021. Therefore, a significant amount of precision is gained when making use of all six features in the classification process.

A main purpose when designing SysCall was to be able to distinguish systematic errors from heterozygous sites in datasets of lower coverage than that available to us (35.4x). To evaluate SysCall's performance on different coverage depths, we simulated experiments of lower coverage by randomly sampling a given percentage from the initial set of reads. For each of 20%, 40%, 60% and 80% (resulting in coverage of 7x, 14x, 21x, and 28x respectively), we ran 100 iterations where in each iteration we randomly chose the given percentage from our reads, refined our set of locations to those with at least one base-call differing from the reference and proceed as in the previous test: divide the locations into a training and test set (the number of instances in each being half of the smaller sized set), compute features, train, classify, and record the percentage of instances classified correctly. The results for these tests, together with the results for the same tests when using only the directionality bias feature for classification are shown in Figure 5.6. SysCall's classifications are highly accurate at all of the coverage rates tested, and the improvement relative to

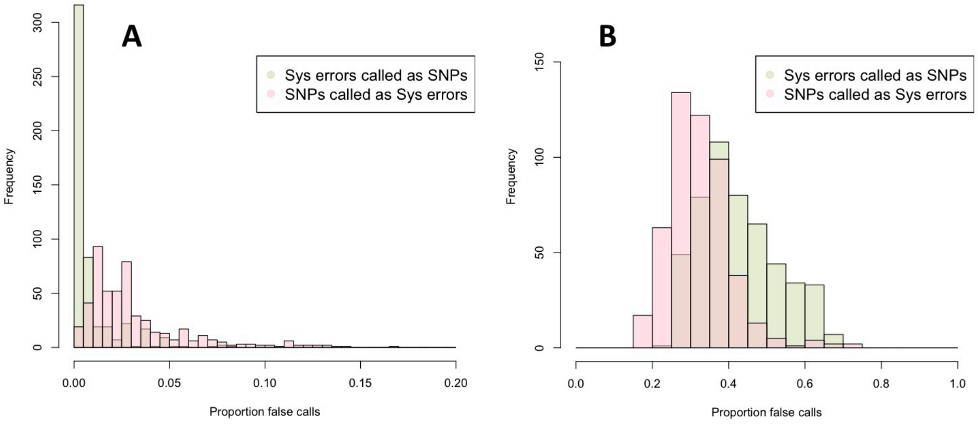


Figure 5.7: Histograms of the proportion of false-positive and false-negative classifications when considering all simulation runs and using (A) SysCall or (B) only the directionality feature. The proportions of Systematic errors called as SNPs (false-positives) and SNPs called as systematic errors (false-negatives) were determined for each individual simulation across the different coverage rates. The histograms show the frequency of each proportion of error across all simulations. When using all features in the prediction process the error rates of both types are significantly lower, and the miss-classifications are mostly false-negatives, whereas when using only the directionality feature for prediction the majority of wrong classifications are false-positives.

using only the directionality bias is negatively correlated with the mean coverage rate, as expected. The use of all features together results in less errors per simulation, and in more false-negatives (SNPs that are called as systematic errors) than false-positives (systematic errors that are called as heterozygous sites); an opposite trend than when classifying using only directionality (Figure 5.7).

To assess SysCall's ability to detect false SNP calls from Illumina datasets, we analyzed the GAII sequencing data available for NA18507, chromosome 21 (`http://www.illumina.com/truseq/tru_resources/datasets.ilmn`). SAMtools called 61,867 SNPs in the dataset and SysCall partitioned those locations into a set of 61,390 SNPs and 477 systematic errors. As a "gold standard" dataset we used the SNP calls for individual NA18507 available from the HapMap project (`http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/latest/`). From the set of SNPs called by SAMtools 11,984 (19.37%) were present in the "gold standard" dataset. Of the 61,390 SNPs called by SysCall 11,973 (19.50%) were in the "gold standard" set. Of the 477 systematic errors 11 (2.3%) were in the "gold standard" set. Our results show that SysCall helps clean the set of SNPs called by SAMtools from false-positiveSNP calls. We note that in this analysis half of the reads, in expectation, are expected to differ from the reference. When searching for variants in experiments where this is not the case (such as RNA-Seq, methyl-Seq, rare variant detection etc.) it is easier to mistake systematic errors for true variants and in such cases we expect SysCall's contribution will be even greater.

5.2.4 Presence of systematic errors in other datasets

In order to verify that systematic errors are not specific for the methyl-Seq procedure we looked for evidence of systematic errors in other high throughput datasets. We believe systematic error will be extremely important to correct for in RNA-Seq experiments, in which one attempts to annotate both heterozygous sites and expression levels to derive allele specific expression estimates. We therefore looked for systematic errors in the RNA-Seq dataset from Ambion Human Brain Reference by Illumina (accession SRA012427), on chromosome 1. Since this dataset did not contain overlapping paired reads we could not annotate *error-pairs*. Instead, we used directionality bias of the base-calls different from the reference to annotate systematic error. We could do so because the coverage in this dataset is high (at transcripts that are highly expressed). For each of the 857,570 locations covered by at least 10 forward and 10 reverse reads we conducted a chi-square test, testing for association between occurrence of mismatches and directionality of sequencing. Under a Bonferroni correction for a 0.05 significance level, we found 991 systematic errors. Thus we have approximately 1 in 1000 sites that are shown to be systematic errors. The method used here, using directionality bias, is statistically weaker than the method with which we identified systematic errors from the methyl-Seq experiment, where we used overlapping mate-pairs to identify base-call errors. The fact that the frequency of identified systematic errors in the RNA-Seq dataset is as high as in the methyl-Seq dataset implies that there are more systematic errors present in the RNA-Seq data than in the methyl-Seq data; this could be due to this dataset being produced by an older version of Illumina's GA.

We also looked at newer Illumina data generated by the HiSeq 2000 machines as part of the 1000 genomes project [1]. We analyzed exome data from chromosome 1 (accession ERX01220). We aligned reads to the reference genome with Bowtie and refined our analysis to the 848,742 sites that were covered by at least 10 reads in each direction. When conducting the same statistical test as for the RNA-Seq data, only 2 sites were determined as statistically significant with respect to the differences from the reference being present on one of the sequencing directions. However, testing for directionality bias of mismatches in this way has little power, and many strong systematic errors are missed by this method (Figure 5.8). This results in many locations that are not detected by this method as systematic errors but would be wrongly annotated as heterozygous sites due to their characteristics. We therefore annotated a set of candidate heterozygous sites as those locations with at least 10% of the base-calls being different from the reference sequence and with at least 5 differences from the reference, resulting in a set of 1,712 locations. Running SysCall on this set, 316 locations were classified as systematic errors. When annotating SNPs in the 1000 genomes project a filtering step was applied, detailed in sections 5.1.1 and 5.2.1 of the supplementary information of [1], designed specifically to filter out locations in which the base-calls different from the reference are not evenly distributed between the forward oriented and reverse oriented reads. The filtering step applied in [1] to avoid calling systematic errors as SNPs can decrease the number of false-positive SNP calls, but relies on having a sufficient number of reads from each

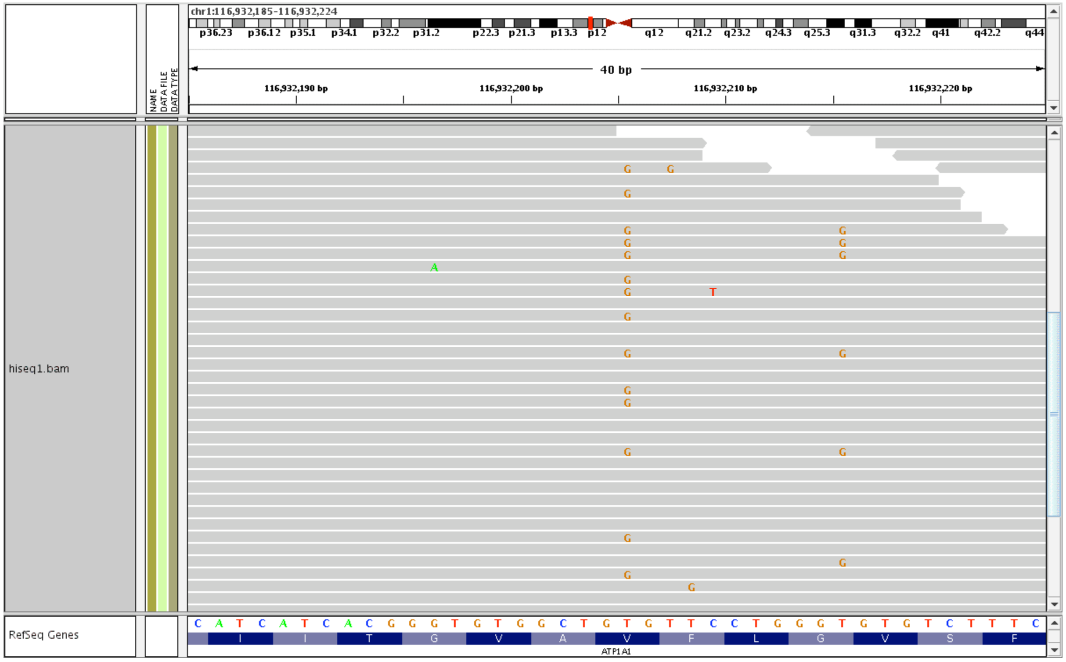

chr1:116,932,185-116,932,224
p36.23 p36.12 p35.1 p34.1 p32.2 p31.2 p22.3 p21.3 p13.3 p12 q12 q21.2 q23.2 q24.3 q25.3 q31.3 q32.2 q41 q42.2 q44
40 bp
116,932,190 bp
116,932,200 bp
116,932,210 bp
116,932,220 bp
NAME
DATA FILE
DATA TYPE
hiseq1.bam
RefSeq Genes
C A T C A T C A C G G G T G T G G C T G T G T T C C T G G G T G T G T C T T T C
I I T G V A V F L G V S F
ATP1A1

Figure 5.8: Systematic errors in HiSeq data. A screenshot from the IGV browser showing two systematic errors in the HiSeq dataset analyzed. These locations are not statistically significant under a chi-squared test for directionality bias (after correcting for multiple hypotheses), demonstrating the weakness of this test.

strand and makes use only of the strand-specific characteristic of systematic errors. As we have shown, distinguishing between systematic errors and heterozygous sites can be greatly improved by taking additional evidence into account.

5.3 Discussion

We have identified systematic error in Illumina sequence that is prevalent in different types of datasets, and that does not appear to be easily correctible during base-calling. This systematic error has significant implications for calling heterozygous sites, especially at low coverage [110]. Moreover, while increasing the extent of coverage enables the detection of rare variants in population studies and low expression rates in transcriptome studies, it also reveals locations of weaker systematic errors (locations at which there is a small accumulation of base-call errors). Thus, the problem of distinguishing systematic error from true heterozygous sites persists regardless of the extent of coverage. We detected this type of error, and could thoroughly characterize it, thanks to a dataset with overlapping paired-end reads and with very high coverage. Making use of our characterization we have designed and implemented a classifier to correct for systematic errors at much lower coverage depths and with no need for paired-end reads. We have shown that by using the different characteristics in the prediction process we gain a significant increase in performance over using directionality bias alone.

Although we have provided a preliminary characterization of systematic error, with further work and additional data it may be possible to better identify sequences associated with error. In particular, it should be possible to identify and characterize systematic error resulting from other sequencing technologies.

In this chapter we have presented how the methyl-Seq experimental method can be coupled with statistical tests to investigate the presence of systematic error in a dataset. We have also presented methods for annotating the characteristics of systematic errors and have demonstrated that logistic regression that uses a combination of different characteristics can be a powerful tool for distinguishing systematic errors from true variants. This workflow is generalizable, and can be used to test different sequencing technologies and work-flow chemistries for systematic errors at low cost (due to the low cost of the methyl-Seq procedure). Although such a comprehensive assessment is beyond the scope of this study, we have looked at RNA-Seq SOLiD data from [140] and have identified statistically significant systematic error. Additionally, a research group at UC Berkeley has recently discovered a new type of systematic error, that follows a different sequence logo than presented here, and at which base-calls different from the reference are found in both sequencing directions. This newly found error is currently being characterized and we anticipate that a collaboration to expand SysCall's framework for accommodating this type of error will be highly beneficial. We believe that as sequencing technology improves systematic errors should decrease, and we have observed this to be the case based on the Illumina samples we have investigated. Sequence from two years ago shows higher systematic error rates than recently sequenced data. Nevertheless, we believe that systematic error is a continuing characteristic of Illumina sequence.

5.4 Methods

methyl-Seq experiments

The human sample was collected with IRB approval from the Children's Hospital and Research Center, Oakland. The approval was granted for a single subject to draw blood for the purpose of examining his methylome and transcriptome, with the understanding that the subject is fully aware of the implications of collecting and analyzing personal genetic data. Immediately after phlebotomy, leukocytes were isolated by Ficoll centrifugation. B cells were isolated from the leukocyte fraction with an indirect magnetic labeling system for the isolation of untouched B cells which yields highly pure B cell preparations (Miltenyi). DNA was extracted by standard methods, and digested overnight with HpaII (NEB). HpaII cuts the sequence CCGG; methylation of the central cytosine on one or both strands protects the sequence from digestion with HpaII [63]. HpaII fragments 50-300 bp in length were isolated on an agarose gel. A paired-end sequencing library was constructed with the standard Illumina kit, and sequenced on an Illumina GAIIX to collect reads of 76 bases, resulting in 15,598,990 read pairs. Read pairs that did not terminate at CCGG restriction sites were removed, leaving 14,205,350 read pairs. The reads were mapped to the human reference genome (hg18) using Bowtie [85] as single end reads allowing 3 mismatches and requiring that the alignments be unique. Those that did not align were removed and the remaining reads were mapped again, this time as paired end reads with a mismatch limit of 2. The higher mismatch limit of 3 was used in the initial alignment step to avoid having reads with more base-call errors preferentially pass the uniqueness requirement. This produced 6,939,310 aligned read pairs mapped to 313,789 distinct locations. The same procedure was followed for the second methyl-Seq experiment from monocyte DNA. The experiment generated 14,432,723 read pairs, of which 7,265,035 were ultimately mapped to 274,230 distinct locations.

Annotating systematic errors at *GGT* sites

The error rate in our dataset at *GGT* sites was computed as $p_{GGT} := \frac{\#error-pairs\ at\ GGT}{\#all\ pairs\ at\ GGT} = 0.0194$. We tested whether there are specific *GGT* locations at which there is a significant excess of errors by computing a p-value for each *GGT* site, given the number of *error-pairs* and coverage at the location, using p_{GGT}, and using a Bonferroni correction of 0.05. The number of significant locations remained substantial at 660, out of 61,779 *GGT* sites considered.

Annotating systematic errors in the phiX174

To test the influence different base callers have on the extent to which systematic errors are present in a dataset we looked for systematic errors in the non-paired reads reported in [77]. In [77], several sets of base-called reads were obtained from one run of sequencing of the phiX174 genome, each using a different base calling method to process the images generated by the sequencing machine. In this work we compared two base calling methods: Bustard, which is Illumina's base-caller, and naiveBayesCall, presented in [77]. The sequencing run generated 74,686 non-paired reads, resulting in an extremely high coverage dataset for the 5,386 bp long genome.

We mapped the reads from each method to the virus genome using Bowtie, obtaining 382.2x coverage for the Bustard called reads and 394.2x coverage for the naiveBayesCall called reads. Since phiX174 is only 5,386 bp long and has been thoroughly studied for heterozygous sites due to its use as a sequencing control, we excluded the five known SNP sites from our analysis, and at the remaining sites called all base-calls that were different from the reference as base-call errors. We computed the probability of a base-call error for each dataset of mapped reads by $p = \frac{\#\ base-call\ errors}{\#\ base\ calls}$, and identified locations with a significant accumulation of errors by computing a p-value for every given location with i errors and coverage n as previously described in the text, using a Bonferroni correction for a 0.05 significance level. We used the frequency of base-call errors in the Bustard called reads of 0.0029 as the error probability for both datasets, since this was the higher of the two frequencies.

We found 59 systematic errors in the Bustard called dataset and 40 systematic errors in the naiveBayesCall dataset, amounting to a systematic error rate of 1 in 91 bp and 1 in 135 bp respectively. When restricting to cases in which more than 10% of the base-calls had errors we found 15 systematic errors for Bustard and 10 systematic errors for naiveBayesCall, 7 of which were at the same sites.

SysCall's design and implementation

In this section we describe SysCall, a logistic regression classifier designed to distinguish heterozygous sites from systematic errors, based on the characteristics of systematic errors we have discussed. We will begin with describing the features used in SysCall's model, continue with how the model parameters were learned, and end with a description of the prediction procedure given a new dataset. Importantly, the special features of the methyl-Seq dataset (overlap of paired-reads and deep coverage) were used only for the first two stages. There is no need for the dataset on which SysCall is used to have such features. As we show in Figure 5.6, SysCall preforms well on a single-end dataset of 7x.

Model features

We have chosen features to be used in SysCall based on our findings regarding the characteristics of systematic errors. Given a dataset and a location, l, SysCall annotates a vector of features, x_l, as follows: First a sequencing direction is chosen (forward or reverse) as the direction with the larger proportion of base-calls that differ from the reference. SysCall only considers sites at which there is at least one base-call that differs from the reference. Let q_{l1} and q_{l2} be that proportion for the chosen and not chosen directions respectively. For example, for the location annotated as a SNP in Figure 5.1, we would choose the forward direction and have $q_1 = 1$ and $q_2 = 0$. Let b_i be the nucleotide that is i places from l in the chosen direction and let w_i be the vector of quality scores at the location i places from l, attained from the reads overlapping that location. A feature vector is then annotated for l as:

$$x_l = (b_{-2}, b_{-1}, b_0, q_{l1} - q_{l2}, q_{l1}, \mathrm{PT}(w_0, w_1)),$$

where $\mathrm{PT}(w_0, w_1)$ is the paired t-test result on the two vectors w_0 and w_1. This paired t-test feature is computed due to our observation that the quality scores at systematic error locations tend to be lower relative to the quality scores at their neighboring sites (Figure 5.9), and this can help distinguish them from true heterozygous sites. As an example, for the location annotated as a SNP in Figure 5.1 the feature vector is $(G, G, T, 1, 1, -5.56)$.

Parameter estimation

We learned parameters for SysCall using training sets constructed from our methyl-Seq dataset. In that dataset, due to both overlap of paired-reads and high coverage, it was possible to determine many sites with high certainty as either heterozygous sites or systematic errors. We annotated a

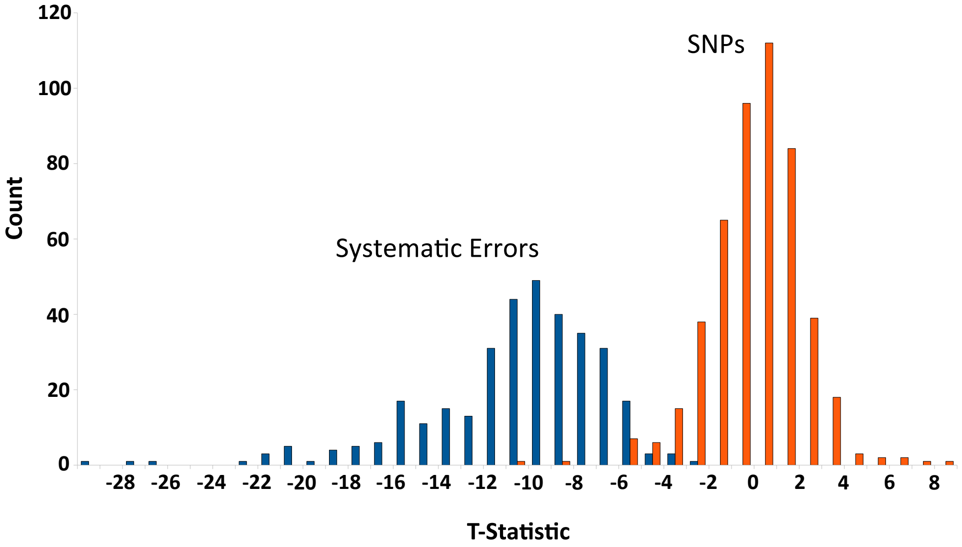


Figure 5.9: The paired t-test statistic helps distinguish true SNPs from systematic errors. The paired t-test ($PT(w_0, w_1)$) was computed for the "SNPs" and "Systematic errors" sets used for training SysCall. The histogram of paired t-test for the "SNPs" set (red) is centered around 0 (mean: 0.0024, std: 2.035), indicating that the quality scores at those locations were similar to their neighboring quality scores. The histogram of the "Systematic errors" set (blue) formed an almost disjoint distribution (mean: -10.505, std: 3.919). We note that at lower coverage rates the separation is not expected to be as strong.

list of locations that would be candidates for heterozygous sites (where a significant amount of the base-calls differ from the reference) and which we could call as systematic errors or heterozygous sites with high certainty. Of the 905 locations in our dataset with coverage of at least 40 (paired-calls) and at which 10-90% of the base-calls on the forward strand differed from the reference we annotated two sets: (1) "SNPs" - the 491 locations at which all differences from the reference were *SNP-pairs*. (2) "Systematic errors" - the 338 locations at which all differences from the reference were *error-pairs*. From each mate-pair one of the reads was chosen at random to simulate a non-overlapping (and non paired-end) dataset. Also, 338 locations were chosen at random for the "SNPs" set to ensure the predictions were feature-based only. A feature matrix was built for these 676 locations (the training set), and the parameters for a logistic regression model were computed by maximum likelihood estimation using the software package R. Note that when assessing SysCall's performance the data on which the classifier was trained was different from that used to asses its performance (in each iteration only half of this dataset was used for training).

At different depths of coverage the different features may be indicative to different extents. For example, at high sequencing depths the paired t-test statistic and the frequency of error on each direction may have a more significant effect than at lower sequencing depths, where the sequence motif is more informative. To account for this we simulated experiments of lower coverage by randomly sampling a given percentage from the initial set of reads. For each of 20%, 40%, 60% and 80% (resulting in coverage of 7x, 14x, 21x, and 28x respectively), we randomly chose the given percentage from our reads, refined our set of locations to those with at least one base-call differing from the reference and proceeded as before to construct a different training set for every coverage.

Prediction procedure

SysCall takes as input a list of genomic locations and a sequencing dataset. For n given locations, SysCall constructs an $n \times 7$ feature matrix, M, where $M_{i,*} = (1, x_i)$, x_i being the feature vector for location i. Then, SysCall computes the mean coverage for the given dataset and uses the model parameters learned from the training set with coverage closest to that observed, β, to compute the vector of posterior probabilities as

$$p_i = \frac{1}{1 + e^{-[\beta^T M^T]_i}}$$

for $i = 1, \ldots, n$. Using a threshold of 0.5 on the posterior probability, SysCall partitions the locations into "true heterozygous sites" ($p_i \geq 0.5$) and "systematic errors" ($p_i < 0.5$) and prints out two files accordingly, along with the posterior probability assigned to each location. In the case of multiple mappings of reads, each mapping of a read is considered by SysCall, independently of other mappings.

SysCall is implemented in R. The running time for classification is instantaneous, and the running time for feature assembly depends on the number of sequenced reads in the experiment and the number of locations considered, currently taking 10 seconds per 100,000 reads when classifying 900 locations, and is trivially parallelizable. SysCall is available at `http://bio.math.berkeley.edu/SysCall/`.

Appendix A

Biological terms

Note: Many terms in this table do not have a single concrete definition due to ongoing discoveries that call for reassessment. The descriptions given here do not serve as concrete definitions, but rather as sufficient background for readers with little or no prior biological knowledge.

Array hybridization A technique to conduct high-throughput genomic studies that preceded high-throughput sequencing. This technique uses arrays - chips to which known segments of DNA are attached. Methodologies using array hybridization can approximate the extent to which each of the DNA segments on the chip is present in a sample.

Case-control studies Studies that aim at finding regions of the DNA that are causal of disease (or that are associated with causal regions). Such studies are done by analyzing the DNA of individuals with the disease (cases) and individuals that do not have the disease (controls).

Chromatin The combination of DNA and "packing" proteins that are bound to it. The packing proteins (histones) can be arranged in either a dense (closed chromatin / heterochromatin) or a sparse (open chromatin) formation.

Coding region A region of the genome that contains sequence that is translated (with an mRNA intermediate) to a protein sequence.

Cultured cells Cells that have been adapted to be grown in the laboratory, and were initially plant or animal cells.

Enhancer A short region of the DNA that can be bound with proteins and enhance gene transcription.

Exon A region on the DNA that codes for mRNA that remains part of the protein-generating mRNA after the editing process (known as splicing).

Fibroblasts The principle active cells of connective tissue in animals.

Gene A continuous segment of DNA that provides the coded instructions for the synthesis of either an enzyme or a functional RNA molecule.

Germline Cells from which gametes (reproductive cells) are derived.

Heterochromatin A tightly packed form of the DNA. Genes present on heterochromatin are not transcribed (are silenced).

Heterozygous site A site of the DNA in which different instances compared (e.g. individuals or species) can display different nucleotides.

Imprinting A phenomenon in which the parent of origin determines the expression rate of a gene. Imprinting has been shown to involve DNA methylation.

Neutrophil The most abundant type of white blood cell, also known as polymorphonuclear leukocyte.

Open chromatin see Chromatin

Ortholog regions Regions of the DNA in two species that originated from the same region in the common ancestral species.

PCR amplification An efficient method for the (near exponential) amplification of DNA molecules.

Phenotype The observable physical or biochemical characteristics of an organism.

piRNA Non-coding RNA molecules that form RNA-protein complexes and can be involved in gene silencing.

Promoter region A sequence of nucleotides, associated with a gene, that must bind with mRNA polymerase before transcription can proceed. The promotor region is where many transcription factors bind and have an affect on transcription.

Protein coding region (see Coding region)

Restriction enzymes DNA cutting enzymes, found naturally in bacteria.

SNP (Single Nucleotide Polymorphism) A single-nucleotide location that can differ between individuals of a species. The complete definition of SNP requires that the variant be found in at least 5% of the population.

Somatic cells The cells that form the body of the organism, that are not part of the germline. For example, skin, liver or heart cells.

Stem cells Unspecialized cells with the potential to develop into different cell types.

Transcription The process in which RNA is made from the DNA code (one of the two main steps in the expression of proteins in the cell, see Translation).

Translation The process in which the sequence of an mRNA is read to make an amino acid sequence (a protein).

Transcription factor A protein that aids in activation and regulation of transcription. There are many different known transcription factors.

Uncultured cells see Cultured cells.

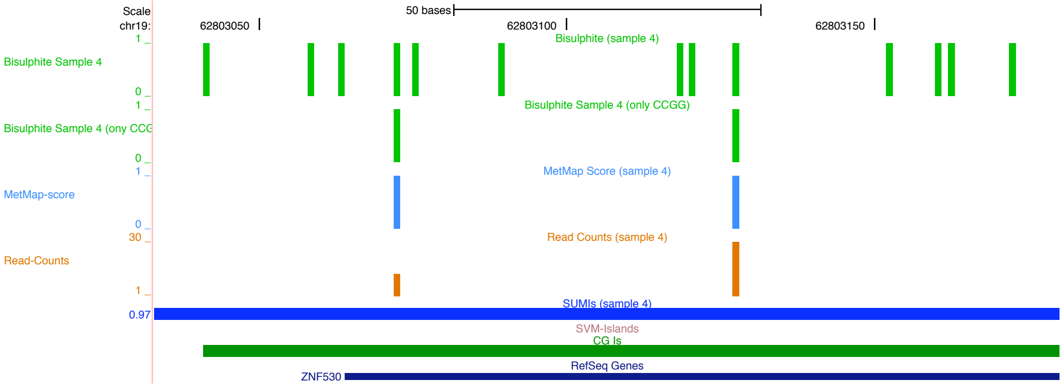
Appendix B

Further examples of bisulfite validation for MetMap

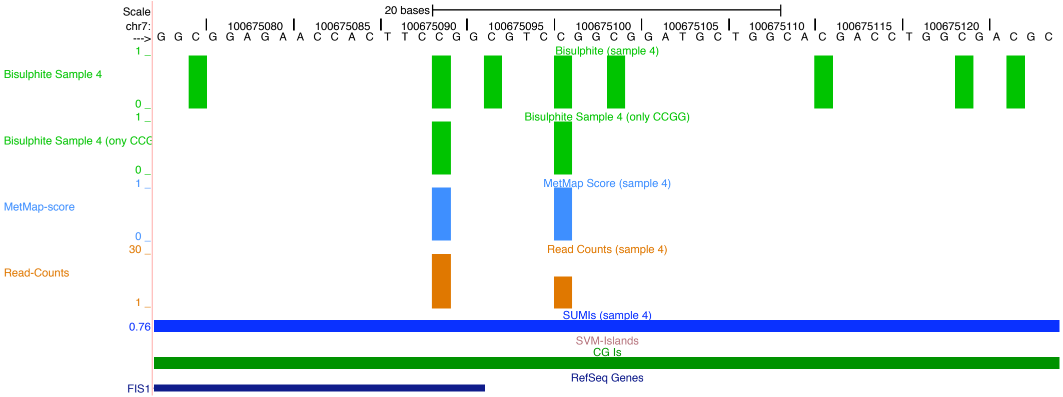
In this appendix the different regions of the genome for which direct bisulfite sequencing was used to evaluate MetMap are presented. In each subfigure the following are shown (from top to bottom):

- Bisulfite values assigned to each CG site in the scope of sequencing, normalized to a 0-1 scale.
- Bisulfite values only for the HpaII sites (CCGG sites), normalized to a 0-1 scale (this is a subset of the previous item).
- MetMap site-specific scores.
- Read counts. For this representation, any HpaII site with read counts larger than 30 was set to 30, the "capping" value for sample 4.
- SUMI regions, along with their scores
- "bona fide" islands (denoted as the SVM-island track, because these islands are inferred with the use of a support-vector machine).
- UCSC CpG islands
- RefSeq genes

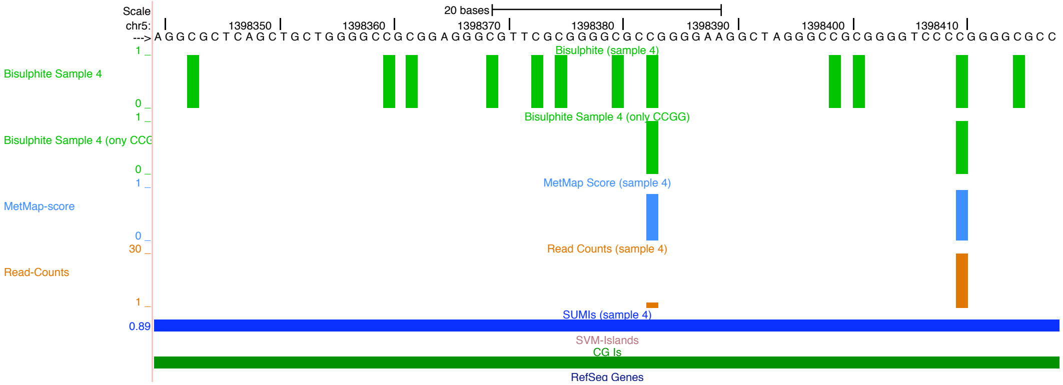
The bisulfite scores of the HpaII sites are better aligned with the MetMap scores than with the MethylSeq read counts. Moreover, the overall methylation state of the region (seen in the uppermost track) is well correlated with the presence or absence of SUMIs, and with the SUMI scores. The SUMI scores are not the mean of the MetMap scores in the validated regions because the regions validated were only a small portion of the SUMIs, which may have regions that are methylated to different extents.



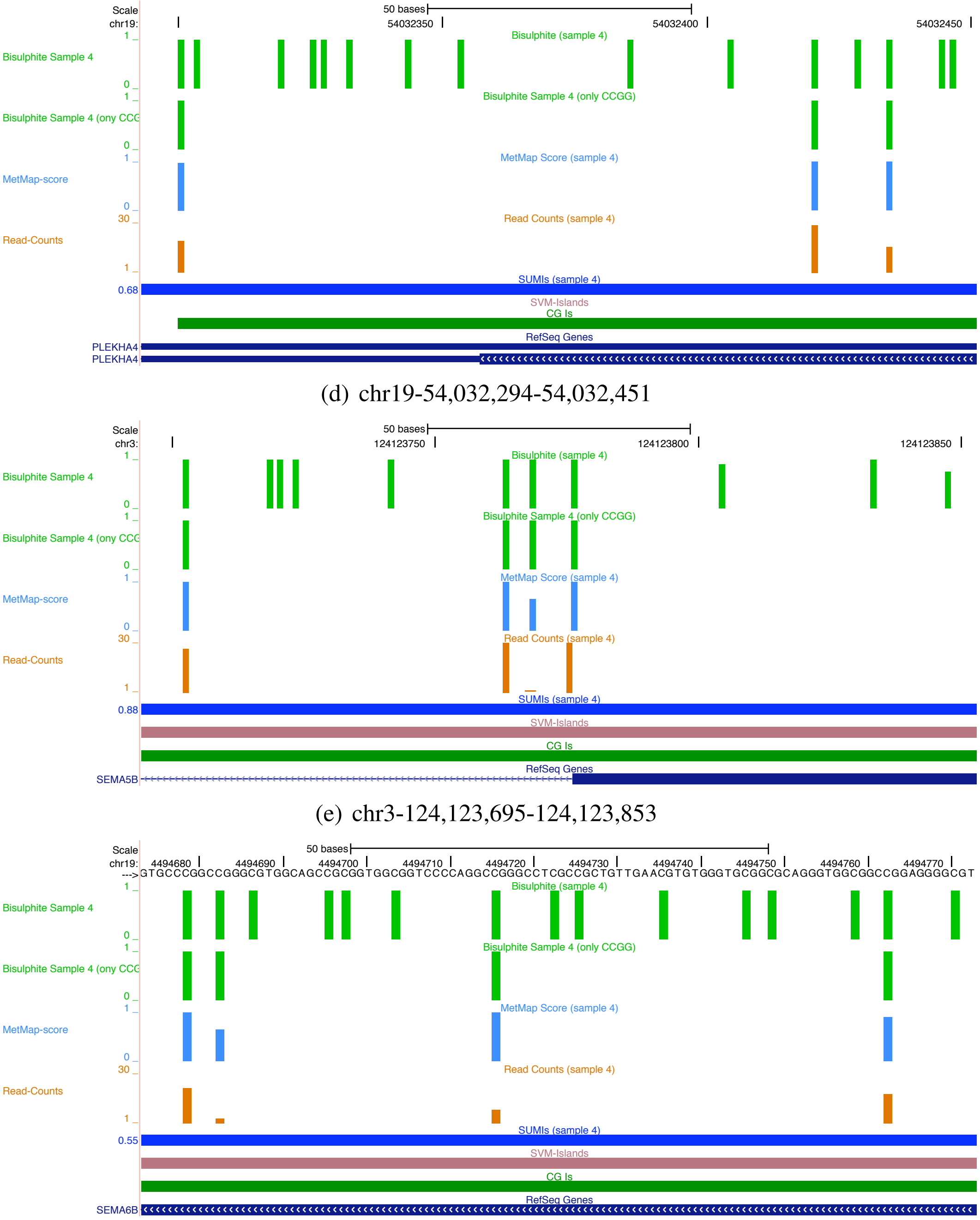
(a) chr19-62,803,034-62,803,180



(b) chr7-100,675,073-100,675,124



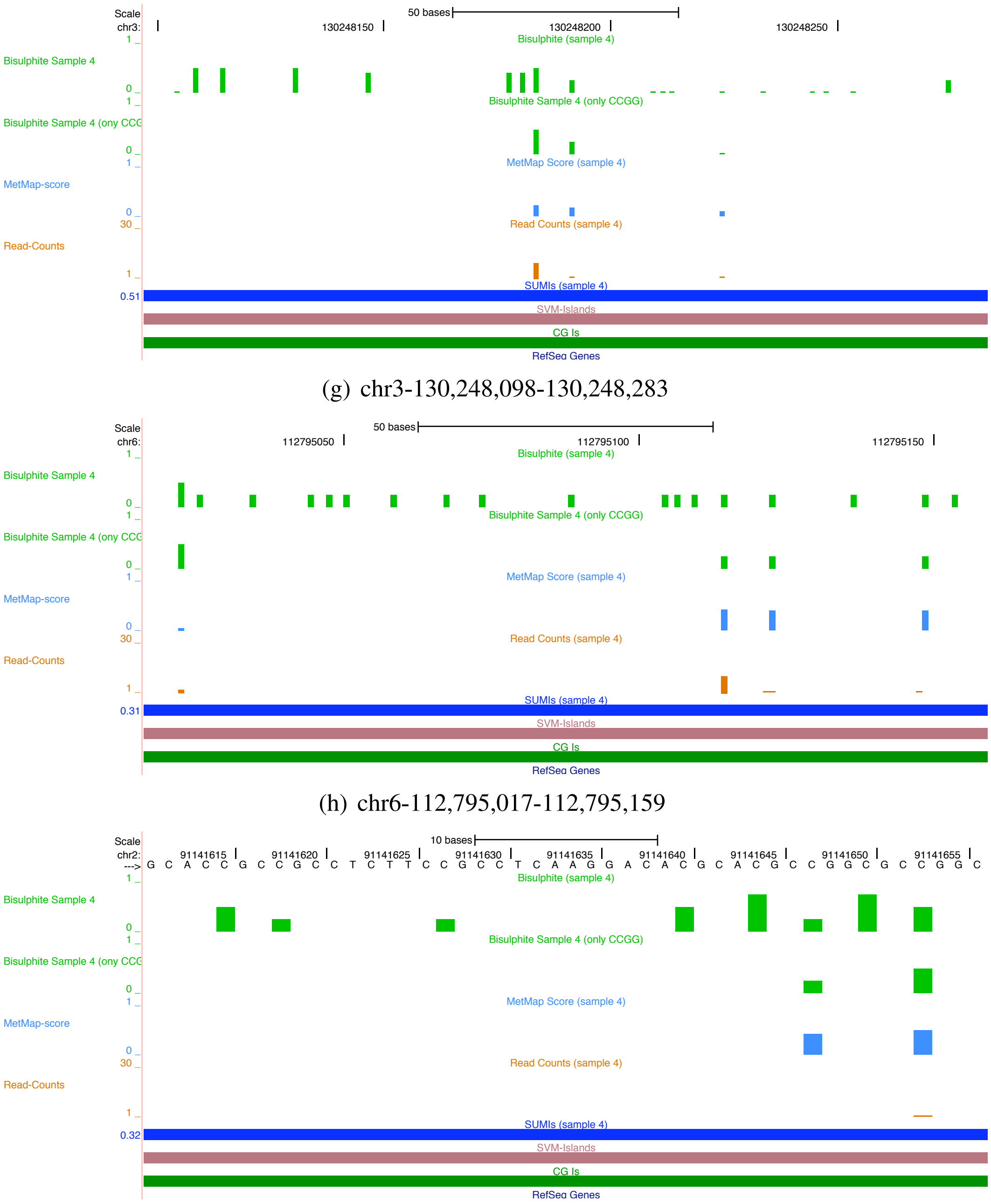
(c) chr5-1,398,340-1,398,418



(d) chr19-54,032,294-54,032,451

(e) chr3-124,123,695-124,123,853

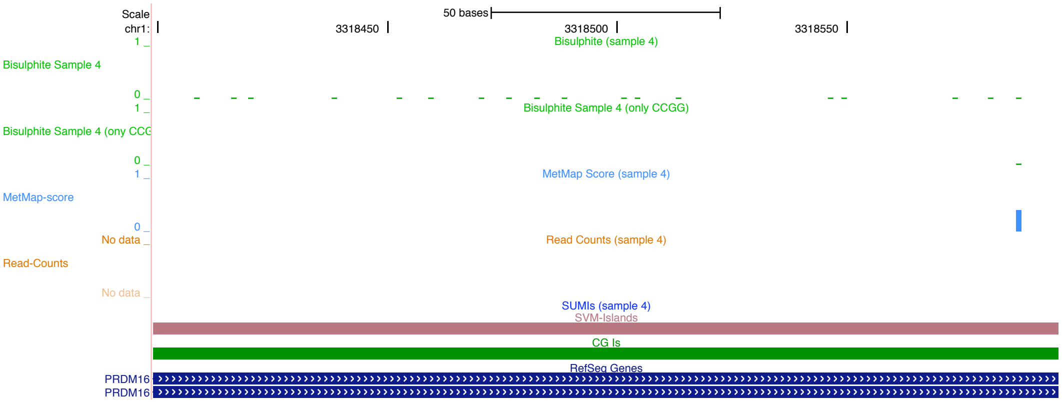
(f) chr19-4,494,674-4,494,773



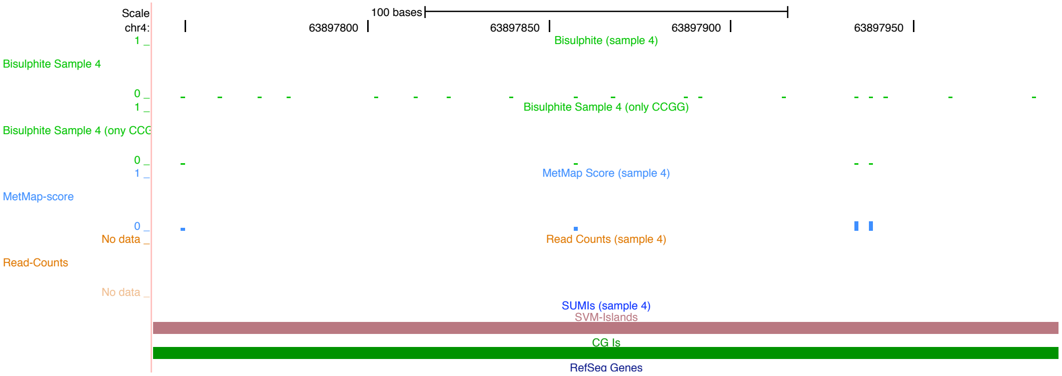
(g) chr3-130,248,098-130,248,283

(h) chr6-112,795,017-112,795,159

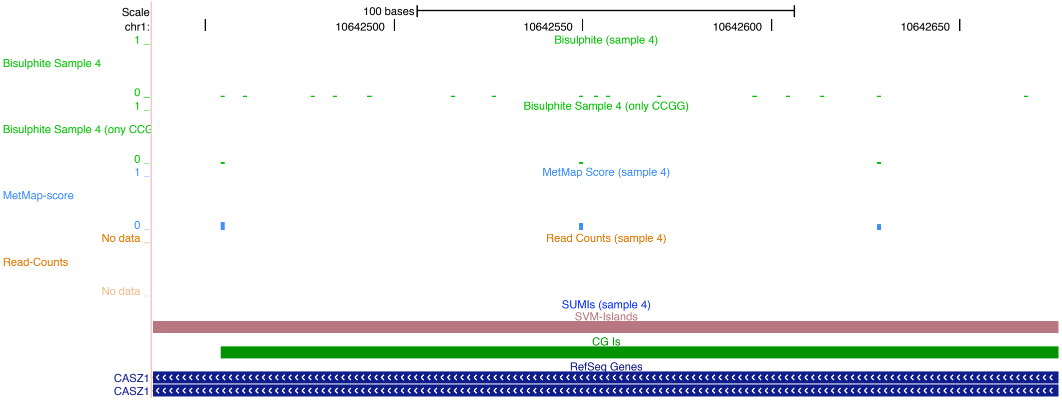
(i) chr2-91,141,611-91,141,656



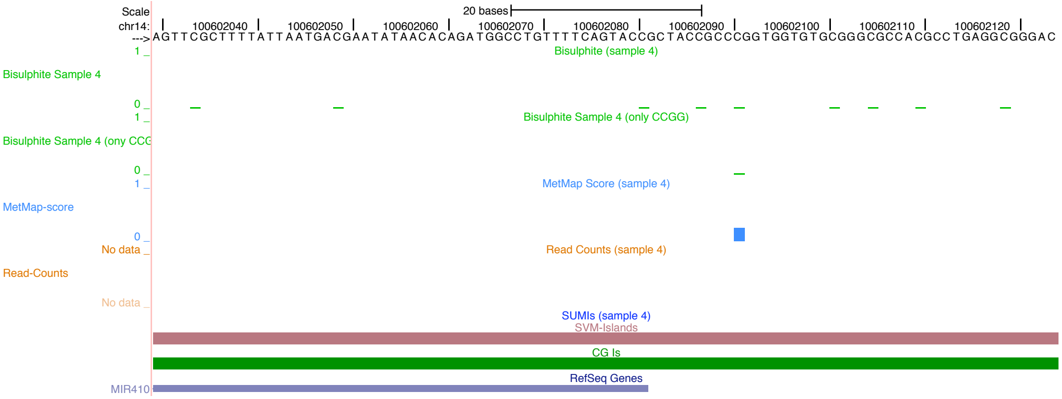
(j) chr1:3,318,400-3,318,596



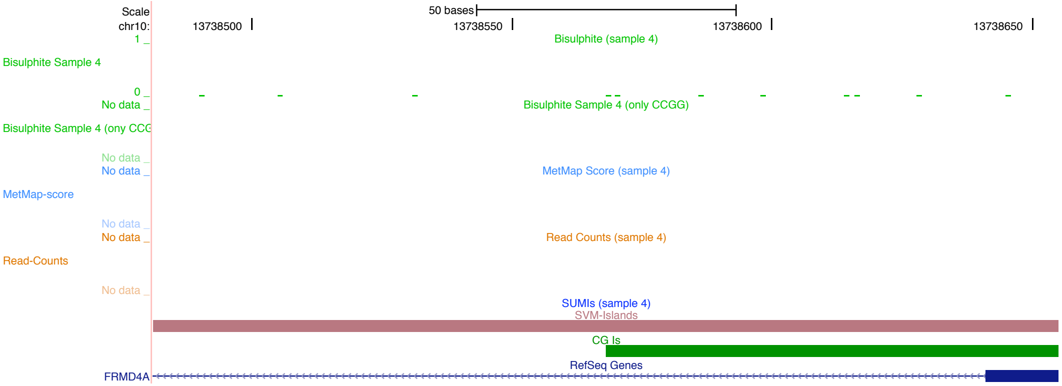
(k) chr4:63,897,742-63,897,990



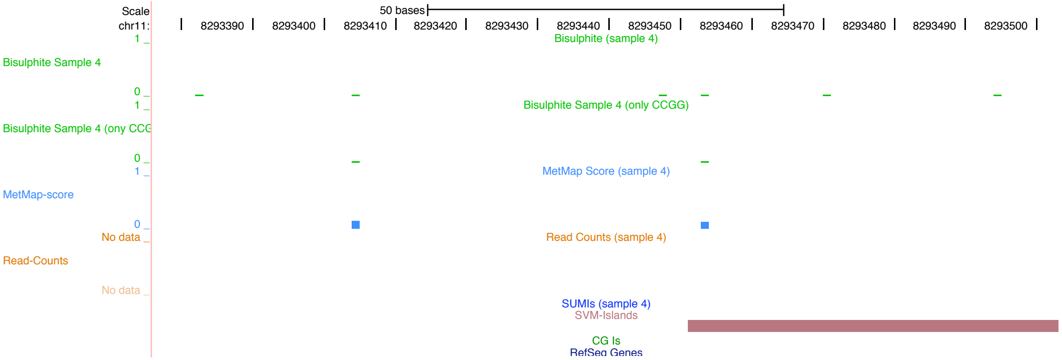
(l) chr1:10,642,437-10,642,676



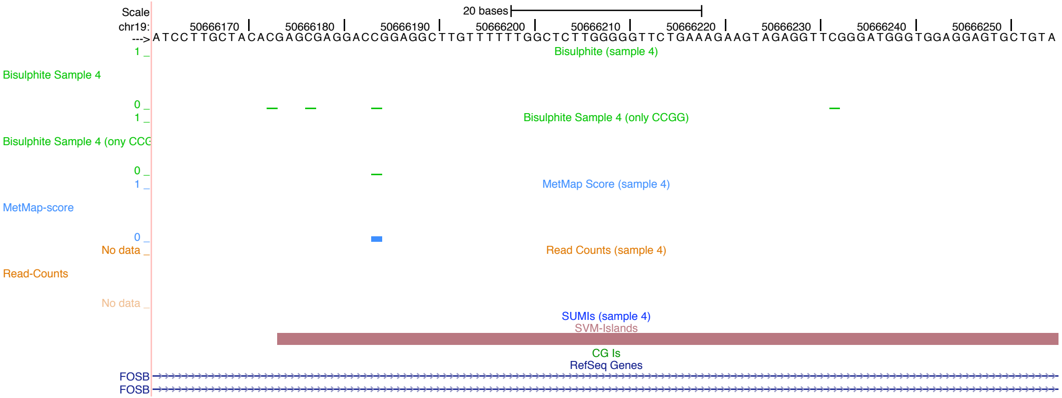
(m) chr14-100,602,030-100,602,124



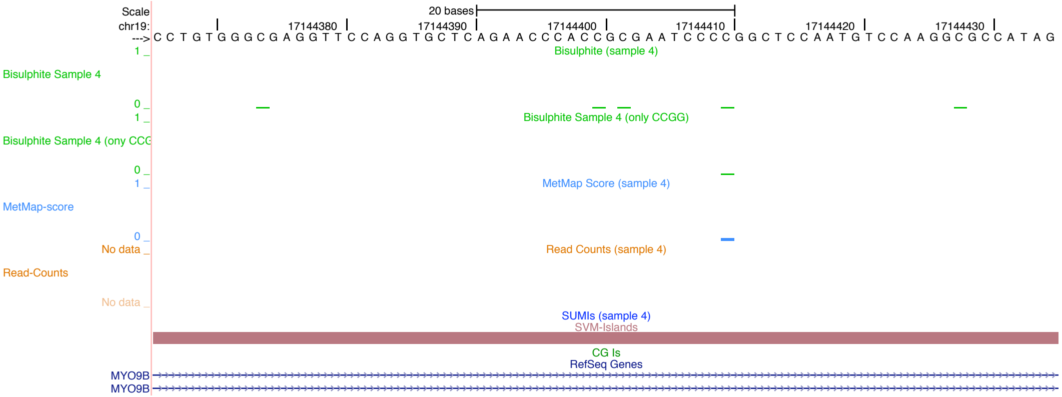
(n) chr10:13,738,482-13,738,655



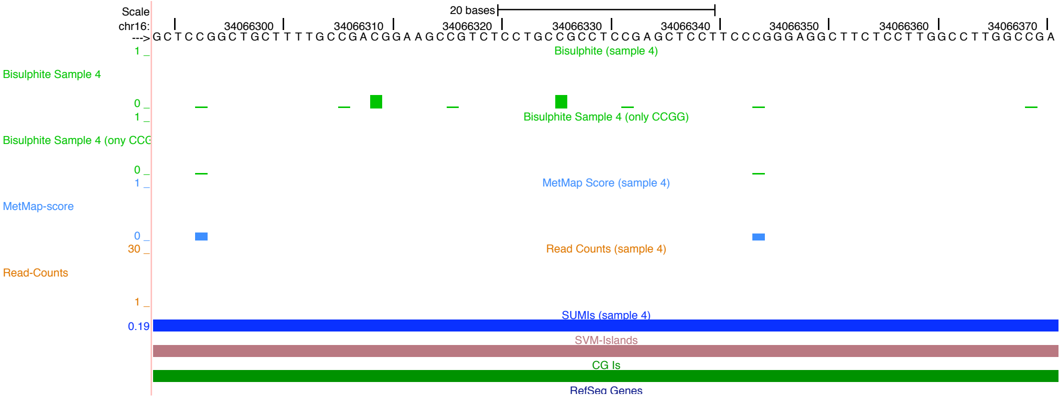
(o) chr11:8,293,377-8,293,503



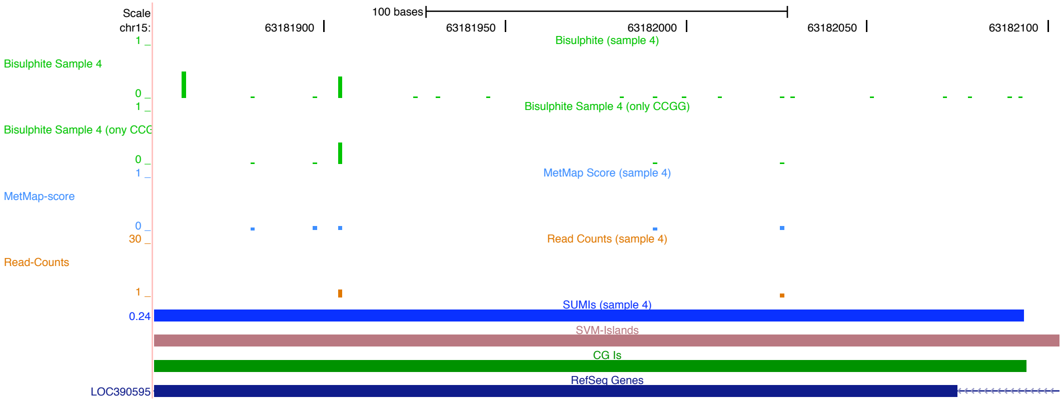
(p) chr19-50,666,161-50,666,255



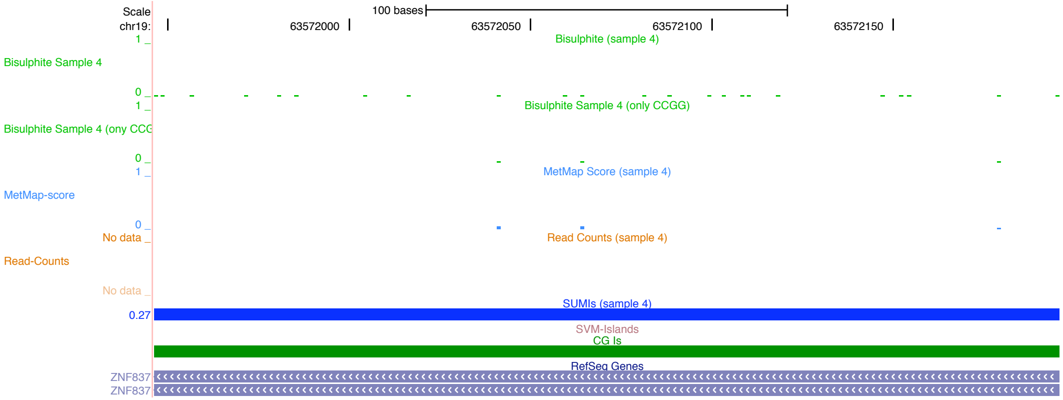
(q) chr19-17,144,366-17,144,435



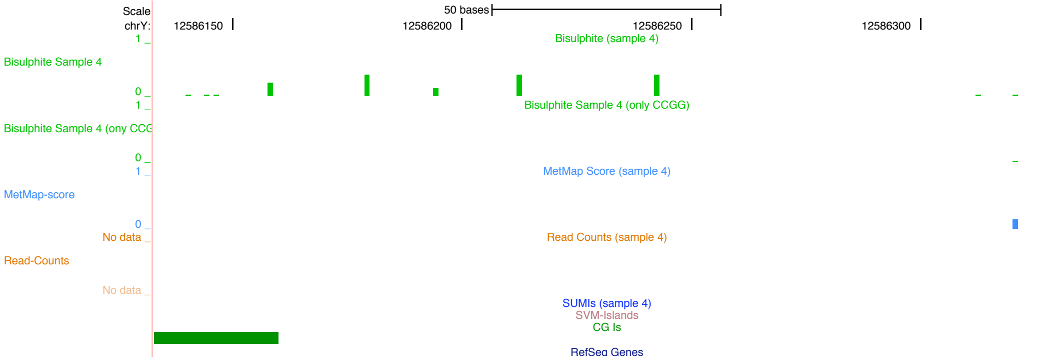
(r) chr16-34,066,289-34,066,371



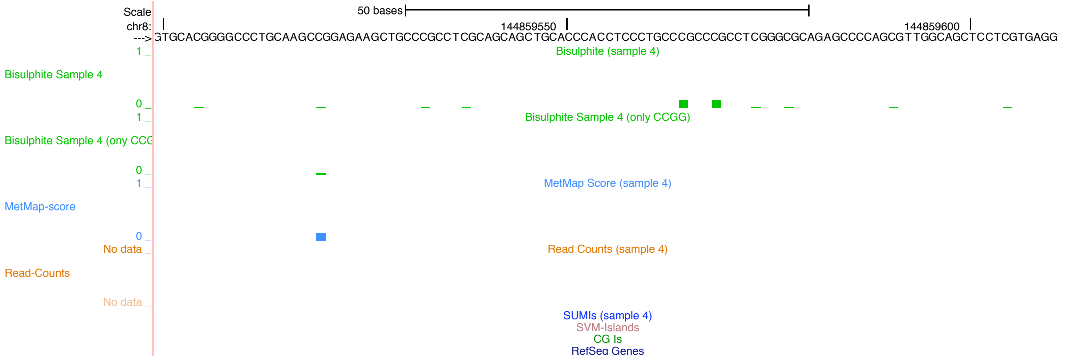
(s) chr15-63,181,854-63,182,103



(t) chr19-63,571,947-63,572,196



(u) chrY-12,586,134-12,586,330



(v) chr8:144,859,500-144,859,611

Bibliography

[1] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, October 2010.

[2] H. Almagor. A Markov analysis of DNA sequences. *J. Theor. Biol.*, 104:633–645, 1983.

[3] A. A. Aravin, R. Sachidanandam, D. Bourc'his, C. Schaefer, D. Pezic, K.F. Toth, T. Bestor, and G. J. Hannon. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Molecular cell*, 31(6):785–799, 2008.

[4] K. Atteson. Calculating the exact probability of language-like patterns in biomolecular sequences. In *Proceedings of the ISMB 98 Conference*, pages 17–24, 1998.

[5] M. P. Ball, J. B. Li, Y. Gao, J. Lee, E. M. Leproust, I. Park, B. Xie, G. Q. Daley, and G. M. Church. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature Biotechnology*, 27(4):361–368, 2009.

[6] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1):289–300, 1995.

[7] Y. Benjamini and T. P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research, doi:10.1093/nar/gks001*, 2012.

[8] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[9] B. P. Berman, D. J. Weisenberger, J. F. Aman, T. Hinoue, Z. Ramjan, Y. Liu, et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina associated domains. *Nature Genetics*, 44(1):40–46, 2011.

[10] B. E. Bernstein, A. Meissner, and E. S. Lander. The Mammalian Epigenome. *Cell*, 128(4):669–681, February 2007.

[11] A. Bird. DNA methylation patterns and epigenetic memory. *Genes Dev*, 16(1):6–21, January 2002.

[12] A. Bird, M. Taggart, M. Frommer, O.J. Miller, and D. Macleod. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*, 40(1):91, 1985.

[13] A. P. Bird. CpG-rich islands and the function of DNA methylation. *Nature*, 321(6067):209, 1986.

[14] B. E. Blaisdell. Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucaryotic nuclear DNA sequences both protein-coding and noncoding. *J. Mol. Evol.*, 21:278–288, 1985.

[15] R. Blekhman, A. Oshlack, A. E. Chabot, G. K. Smyth, and Y. Gilad. Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS genetics*, 4(11):e1000271, 2008.

[16] C. Bock. Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, 13(10):705–719, 2012.

[17] C. Bock, J. Walter, M. Paulsen, and T. Lengauer. CpG island mapping by epigenome prediction. *PLoS Computational Biology*, 3(6):e110, 2007.

[18] C. Bock, J. Walter, M. Paulsen, and T. Lengauer. Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. *Nucleic Acids Research*, 36(10):e55, 2008.

[19] D. Boffelli, J. McAuliffe, D. Ovcharenko, K. D. Lewis, I. Ovcharenko, L. Pachter, and E. M. Rubin. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611):1391–1394, 2003.

[20] J. Borgel, S. Guibert, Y. Li, H. Chiba, D. Schübeler, H. Sasaki, T. Forné, and M. Weber. Targets and dynamics of promoter DNA methylation during early mouse development. *Nature Genetics*, 42(12):1093–1100, 2010.

[21] M. Borodovsky, J.D. McIninch, E. V. Koonin, K. E. Rudd, C. Medigue, and A. Danchin. Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic Acids Research*, 23(17):3554–3562, 1995.

[22] S. Branciamore, Z. X. Chen, A. D. Riggs, and S. N. Rodin. CpG island clusters and pro-epigenetic selection for CpGs in protein-coding exons of hox genes and other transcription factors. *Proc. Nat. Acad. Sci. USA*, 107(35):15485–15490, 2010.

[23] C. V. Breton, H. M. Byun, M. Wenten, F. Pan, Al. Yang, and F. D. Gilliland. Prenatal tobacco smoke exposure affects global and gene-specific DNA methylation. *American journal of respiratory and critical care medicine*, 180(5):462–467, 2009.

[24] R. A. Brink. Paramutation and chromosome organization. *Quarterly Review of Biology*, pages 120–137, 1960.

[25] A. L. Brunner, D. S. Johnson, S. W. Kim, A. Valouev, T. E. Reddy, et al. Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Research*, 19:1044–1056, January 2009.

[26] P. Bühlmann and A. J. Wyner. Variable length Markov chains. *Annals of Statistics*, 27(2):480–513, 1999.

[27] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268:78–94, 1997.

[28] E. I. Campos and D. Reinberg. Histones: Annotating Chromatin. *Annual Review of Genetics*, 43(1):559–599, 2009.

[29] B.R. Carone, L. Fauquier, N. Habib, J.M. Shea, C.E. Hart, R. Li, et al. Paternally induced transgenerational environmental reprogramming of metabolic gene expression in mammals. *Cell*, 143(7):1084–1096, 2010.

[30] H. Cedar and Y. Bergman. Epigenetics of haematopoietic cell development. *Nature Reviews Immunology*, 11(7):478–488, 2011.

[31] G. A. Churchill. Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology*, 51(1):79–94, 1989.

[32] S. J. Clark, J. Harrison, C. L. Paul, and M. Frommer. High sensitivity mapping of methylated cytosines. *Nucl. Acids Res.*, 22(15):2990–2997, August 1994.

[33] N. M. Cohen, V. Dighe, G. Landan, S. Reynisdóttir, A. Palsson, S. Mitalipov, and A. Tanay. DNA methylation programming and reprogramming in primate embryonic stem cells. *Genome Research*, 19(12):2193–2201, 2009.

[34] S. J. Cokus, S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, 452:215–219, 2008.

[35] C. Coulondre, J. H. Miller, P. J. Farabaugh, and W. Gilbert. Molecular basis of base substitution hotspots in Escherichia coli. *Nature*, 274(5673):775, 1978.

[36] G. E. Crooks, G. Hon, J.-M. M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome Research*, 14(6):1188–1190, June 2004.

[37] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[38] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer. Substantial biases in ultrashort read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16):e105, 2008.

[39] J. R. Edwards, A. H. O'Donnell, R. A. Rollins, H. E. Peckham, C. Lee, et al. Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Research*, 20:972–980, 2010.

[40] M. Ehrlich. DNA methylation in cancer: too much, but also too little. *Oncogene*, 21(35):5400–5413, 2002.

[41] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011.

[42] M. Esteller. The necessity of a human epigenome project. *Carcinogenesis*, 27(6):1121–1125, June 2006.

[43] J. Felsenstein. *Inferring phylogenies*, volume 2. Sinauer Associates Sunderland, 2004.

[44] S. Feng, S. J. Cokus, X. Zhang, P. Y. Y. Chen, M. Bostick, M. G. Goll, et al. Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences of the United States of America*, 107(19):8689–8694, May 2010.

[45] S. Feng, S. E. Jacobsen, and W. Reik. Epigenetic reprogramming in plant and animal development. *Science Signalling*, 330(6004):622, 2010.

[46] J. G. Fleagle. *Primate adaptation and evolution*. Academic Press, 1999.

[47] M. F. Fraga, E. Ballestar, M. F. Paz, S. Ropero, F. Setien, M. L. Ballestar, et al. Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30):10604–9, 2005.

[48] T. S. Furey. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics*, 2012.

[49] F. Gao, X. Liu, X.-P. Wu, X.-L. Wang, D. Gong, et al. Differential DNA methylation in discrete developmental stages of the parasitic nematode Trichinella spiralis. *Genome Biology*, 13(10):R100, 2012.

[50] M. Gardiner-Garden and M. Frommer. CpG islands in vertebrate genomes. *Journal of Molecular Biology*, 196(2):261–282, 1987.

[51] M. S. Gelfand. Prediction of function in DNA sequence analysis. *Journal of Computational Biology*, 2(1):87–115, 1995.

[52] P. G. Giresi, J. Kim, R. M. McDaniell, V. R. Iyer, and J. D. Lieb. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*, 17:877–885, 2007.

[53] J. L. Glass, R. F. Thompson, B. Khulan, M. E. Figueroa, E. N. Olivier, et al. CG dinucleotide clustering is a species-specific property of the genome. *Nucleic acids research*, 35(20):6798–6807, 2007.

[54] M. G. Goll and T. H. Bestor. EUKARYOTIC CYTOSINE METHYLTRANSFERASES. *Annual Review of Biochemistry*, 74(1):481–514, 2005.

[55] A. Goren and H. Cedar. Replicating by the clock. *Nature Reviews Molecular Cell Biology*, 4(1):25–32, 2003.

[56] H. Gowher, O. Leismann, and A. Jeltsch. DNA of Drosophila melanogaster contains 5-methylcytosine. *EMBO J*, 19(24):6918–6923, December 2000.

[57] R. T. Grant-Downton and H. G. Dickinson. Epigenetics and its implications for plant biology 2. The 'Epigenetic Epiphany': Epigenetics, Evolution and Beyond. *Annals of Botany*, 97(1):11–27, 2006.

[58] S. I. S. Grewal and D. Moazed. Heterochromatin and epigenetic control of gene expression. *Science*, 301(5634):798–802, 2003.

[59] S. R. Grossman, I. Shylakhter, E.K. Karlsson, E. H. Byrne, S. Morales, et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, 327(5967):883–886, 2010.

[60] H. Gu, C. Bock, T. S. Mikkelsen, N. Jager, Z. D. Smith, E. Tomazou, A. Gnirke, E. S Lander, and A. Meissner. Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nature Methods*, 7(2):133–136, 2010.

[61] F. Hach, F. Hormozdiari, C. Alkan, F. Hormozdiari, I. Birol, E. E. Eichler, and S. C. Sahinalp. mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nature methods*, 7(8):576–577, 2010.

[62] M. Hackenberg, C. Previti, P. L. Luque-Escamilla, P. Carpena, J. Martínez-Aroza, and J. L. Oliver. CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics*, 7:446, 2006.

[63] R. M. Harland. Inheritance of DNA methylation in microinjected eggs of Xenopus laevis. . *Proc Natl Acad Sci USA*, 79(7):2323–2327, 1982.

[64] R. A. Harris, T. Wang, C. Coarfa, R. P. Nagarajan, C. Hong, et al. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature Biotechnology*, advance online publication doi:10.1038/nbt.1682, September 2010.

[65] I. R. Henderson and S. E. Jacobsen. Epigenetic inheritance in plants. *Nature*, 447(7143):418–424, 2007.

[66] H. Hikoya. Discovery of bisulfite-mediated cytosine conversion to uracil, the key reaction for DNA methylation analysis — A personal account. *Proceedings of the Japan Academy, Series B*, 84(8):321–330, 2008.

[67] K. Hoff. The effect of sequencing errors on metagenomic gene prediction. *BMC Genomics*, 10(1):520+, November 2009.

[68] F. Hsieh, S. C. Chen, and K. Pollard. A nearly exhaustive search for CpG islands on whole chromosomes. *The International Journal of Biostatistics*, 5(1):14, 2009.

[69] D. H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution*, 23(2):254–267, 2006.

[70] D. G. Hwang and P. Green. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 101(39):13994–14001, 2004.

[71] R. Illingworth, A. Kerr, D. Desousa, H. Jørgensen, P. Ellis, J. Stalker, et al. A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biology*, 6(1):e22, 2008.

[72] R. S. Illingworth and A. P. Bird. CpG islands–'a rough guide'. *FEBS letters*, 583(11):1713–1720, June 2009.

[73] R. A. Irizarry, C. Ladd-Acosta, B. Wen, Z. Wu, C. Montano, P. Onyango, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genetics*, 41(2):178–186, January 2009.

[74] R. A. Irizarry, H. Wu, and A. P. Feinberg. A species-generalized probabilistic model-based definition of CpG islands. *Mammalian Genome*, 20:674–680, 2009.

[75] H. Ji, L. I. R. Ehrlich, J. Seita, P. Murakami, A. Doi, P. Lindau, and others. Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature*, 467(7313):338–342, 2010.

[76] P. A. Jones and S. B. Baylin. The epigenomics of cancer. *Cell*, 128(4):683–692, 2007.

[77] W.-C. Kao and Y. Song. naiveBayesCall: An Efficient Model-Based Base-Calling Algorithm for High-Throughput Sequencing. *Journal of Computational Biology*, 18:365–377, 2011.

[78] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*, 32(Database issue):D493–D496, January 2004.

[79] P. V. Kharchenko, M. Y. Tolstorukov, and P. J. Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature biotechnology*, 26(12):1351–1359, 2008.

[80] K. B. Kim. CpG islands detector: a window-based CpG island search tool. *Genomics and Informatics*, 8:58–61, 2010.

[81] M. C. King and A. C. Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116, 1975.

[82] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

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

[84] P. W. Laird. Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, 11(3):191–203, 2010.

[85] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[86] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, 2007.

[87] L. Laurent, E. Wong, G. Li, T. Huynh, A. Tsirigos, et al. Dynamic changes in the human methylome during differentiation. *Genome Research*, 20(3):320–331, 2010.

[88] T. F. Lee, J. Zhai, and B. C. Meyers. Conservation and divergence in eukaryotic DNA methylation. *Proceedings of the National Academy of Sciences*, 107(20):9027–9028, 2010.

[89] E. Li, T. H. Bestor, and R. Jaenisch. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, 69(6):915–26, June 1992.

[90] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[91] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[92] M. Li, I. X. Wang, Y. Li, A. Bruzel, A. L. Richards, J.M. Toung, and V. G. Cheung. Widespread RNA and DNA Sequence Differences in the Human Transcriptome. *Science*, 333(6038):53–58, July 2011.

[93] R. Lister, R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, H. A. Millar, and J. R. Ecker. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133(3):523–536, 2008.

[94] R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, 2009.

[95] E. R. Mardis. A decade's perspective on DNA sequencing technology. *Nature*, 470(7333):198–203, 2011.

[96] T. Marschall and S. Rahmann. Probabilistic arithmetic automata and their application to pattern matching statistics. *Proceedings of the 19th Annual Symposium on Combinatorial Pattern Matching*, pages 225–232, 2008.

[97] D. I. Martin, M. Singer, J. Dhahbi, G. Mao, L. Zhang, G. P. Schroth, L. Pachter, and D. Boffelli. Phyloepigenomic comparison of great apes reveals a correlation between somatic and germline methylation states. *Genome Research*, 21(12):2049–2057, 2011.

[98] A. K. Maunakea, R. P. Nagarajan, M. Bilenky, T. J. Ballinger, C. D'Souza, S. D. Fouse, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, 466(7303):253–257, 2010.

[99] R. McDaniell, B.-K. Lee, L. Song, Z. Liu, A. P. Boyle, et al. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science. doi: 10.1126/science.1184655*, 2010.

[100] F. Meacham, D. Boffelli, J. Dhahbi, D. Martin, M. Singer, and L. Pachter. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, 12(1):451+, November 2011.

[101] T. A. Medvedeva, M. V. Fridman, N. J. Oparina, D. B. Malko, E. O. Ermakova, et al. Intergenic, gene terminal, and intragenic CpG islands in the human genome. *BMC Genomics*, 11:48, 2010.

[102] A. Meissner, A. Gnirke, G. W. Bell, B. Ramsahoye, E. S. Lander, and R. Jaenisch. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, 33(18):5868–5877, 2005.

[103] A. Meissner, T. S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454:766–770, July 2008.

[104] M. L. Metzker. Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1):31–46, January 2010.

[105] F. Mohn, M. Weber, D. Schübeler, and T. C. Roloff. Methylated DNA Immunoprecipitation (MeDIP). In John M. Walker and Jörg Tost, editors, *DNA methylation*, volume 507 of *Methods in Molecular Biology*, chapter 5, pages 55–64. Humana Press, Totowa, NJ, 2009.

[106] A. Molaro, E. Hodges, F. Fang, Q. Song, W. R. McCombie, G. J. Hannon, and A. D. Smith. Sperm Methylation Profiles Reveal Features of Epigenetic Inheritance and Evolution in Primates. *Cell*, 146(6):1029–1041, September 2011.

[107] H. D. Morgan, H. G. Sutherland, D. I. Martin, and E. Whitelaw. Epigenetic inheritance at the agouti locus in the mouse. *Nature Genetics*, 23(3):314–318, 1999.

[108] M. J. Morgan. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[109] A. Murrell, V. K. Rakyan, and S. Beck. From genome to epigenome. *Human Molecular Genetics*, 14(suppl 1):R3–R10, 15 April 2005.

[110] Malhis. N. and S. J. Jones. High quality SNP calling using Illumina data at shallow coverage. *Bioinformatics*, 26:1029–1035, 2010.

[111] K. Nakamura, T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, et al. Sequence-specific error profile of Illumina sequencers. *Nucleic acids research*, 39(13):e90, July 2011.

[112] C. Nathan. Neutrophils and immunity: challenges and opportunities. *Nature Reviews Immunology*, 6(3):173–182, 2006.

[113] P. Nicodeme, B. Salvy, and P. Flajolet. Motif statistics. *Theoretical Computer Science*, 287:593–617, 2002.

[114] R. Nielsen. Genomics: In search of rare human variants. *Nature*, 467(7319):1050–1051, October 2010.

[115] G. Nuel. Effective p-value computations using Finite Markov Chain Imbedding (FMCI): application to local score and pattern statistics. *Algorithms in Molecular Biology*, 1:5, 2006.

[116] G. Nuel. Numerical solutions for pattern statistics on Markov chains. *Statistical Applications in Genetics and Molecular Biology*, 5(1):26, 2006.

[117] J. E. Ohm, K. M. McGarvey, X. Yu, L. Cheng, K. E. Schuebel, et al. A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat Genet*, 39(2):237–242, 2007.

[118] L. Pachter. Models for transcript quantification from RNA-Seq. *arXiv:1104.3889v2 [q-bio.GN]*, 2011.

[119] A. A. Pai, J. T. Bell, J. C. Marioni, J. K. Pritchard, and Y. Gilad. A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS genetics*, 7(2):e1001316, 2011.

[120] P. J. Park. ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009.

[121] C. A. Perry, C. D. Allis, and A. T. Annunziato. Parental nucleosomes segregated to newly replicated chromatin are underacetylated relative to those assembled de novo. *Biochemistry*, 32(49):13615–13623, 1993. PMID: 8257695.

[122] G. J. Phillips, J. Arnold, and R. Ivarie. The effect of codon usage on the oligonucleotide composition of the E.coli genome and identification of over- and underrepresented sequences by markov chain analysis. *Nucleic Acids Research*, 15(6):2627–2638, 1987.

[123] J. K. Pickrell and J. K. Gilad, Y.and Pritchard. Comment on “Widespread RNA and DNA Sequence Differences in the Human Transcriptome”. *Science*, 335(6074):1302, 2012.

[124] K.S. Pollard, S.R. Salama, B. King, A.D. Kern, T. Dreszer, et al. Forces shaping the fastest evolving regions in the human genome. *PloS genetics*, 2(10):e168, 2006.

[125] L. Ponger and D. Mouchiroud. CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics*, 18(4):631–633, 2002.

[126] C. Popp, W. Dean, S. Feng, S.J. Cokus, S. Andrews, M. Pellegrini, S.E. Jacobsen, and W. Reik. Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature*, 463(7284):1101–1105, 2010.

[127] S. Prabhakar, J.P. Noonan, S. Pääbo, and E.M. Rubin. Accelerated evolution of conserved noncoding sequences in humans. *Science*, 314(5800):786–786, 2006.

[128] K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 35(Database issue), January 2007.

[129] A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[130] V. Rakyan, T. Down, N. Thorne, P. Flicek, E. Kulesha, et al. An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res.*, 18:1518–1529, June 2008.

[131] V. K. Rakyan, S. Chong, M. E. Champ, P. C. Cuthbert, H. D. Morgan, K. V. K. Luu, and E. Whitelaw. Transgenerational inheritance of epigenetic states at the murine AxinFu allele occurs after maternal and paternal transmission. *Proceedings of the National Academy of Sciences of the United States of America*, 100(5):2538–2543, 2003.

[132] V. K. Rakyan, T. A. Down, D. J. Balding, and S. Beck. Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics*, 12(8):529–541, 2011.

[133] V.K. Rakyan and S. Beck. Epigenetic variation and inheritance in mammals. *Current opinion in genetics & development*, 16(6):573–577, 2006.

[134] M. Regnier and W. Szpankowski. On pattern frequency occurrences in a Markovian sequence. *Algorithmica*, 22(4):631–649, 1998.

[135] W. Reik. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143):425–432, 2007.

[136] G. Reinert, S. Schbath, and M. S. Waterman. Probabilistic and statistical properties of words: an overview. *Journal of Computational Biology*, 7(1/2):1–46, 2000.

[137] B. Rhead, D. Karolchik, R. M. Kuhn, A. S. Hinrichs, A. S. Zweig, et al. The UCSC genome browser database: update 2010. *Nucleic acids research*, 38(suppl 1):D613–D619, 2010.

[138] P. Ribeca and E. Raineri. Faster exact Markovian probability functions for motif occurrences: a DFA-only approach. *Bioinformatics*, 24(24):2839–2848, 2008.

[139] E. J. Richards. Inherited epigenetic variationÑrevisiting soft inheritance. *Nature Reviews Genetics*, 7(5):395–401, 2006.

[140] A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, 12:R22, 2011.

[141] K. D. Robertson. DNA methylation and human disease. *Nature Reviews Genetics*, 6(8):597–610, 2005.

[142] S. Saxonov, P. Berg, and D. L. Brutlag. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Nat. Acad. Sci. USA*, 103(5):1412–1417, 2006.

[143] R. J. Schmitz, M. D. Schultz, M. G. Lewsey, R. C. O'Malley, M. A. Urich, O. Libiger, N. J. Schork, and J. R. Ecker. Transgenerational Epigenetic Instability Is a Source of Novel Methylation Variants. *Science*, 334(6054):369–373, 2011.

[144] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100, 1990.

[145] A. W. Segal. How neutrophils kill microbes. *Annual review of immunology*, 23:197, 2005.

[146] L. Shen, Y. Kondo, Y. Guo, J. Zhang, L. Zhang, et al. Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *PLoS Genet*, 3(10):2023–2036, October 2007.

[147] Y. Shufaro, O. Lacham-Kaplan, B. Z. Tzuberi, J. McLaughlin, A. Trounson, H. Cedar, and B. E. Reubinoff. Reprogramming of DNA replication timing. *Stem Cells*, 28(3):443–449, 2010.

[148] M. Singer, D. Boffelli, J. Dhabhi, A. Schönhuth, G. P. Schroth, D. I. Martin, and L. Pachter. MetMap enables genome-scale methyltyping for determining methylation states in populations . *PLoS Computational Biology*, 6(8):e1000888, 2010.

[149] M. Singer, A. Engström, A. Schönhuth, and L. Pachter. Determining coding CpG islands by identifying regions significant for pattern statistics on Markov chains. *Statistical Applications in Genetics and Molecular Biology*, 10(1):43, 2011.

[150] M. Singer and L. Pachter. *Book Chapter: Bayesian networks in the study of genomewide DNA methylation*. Oxford University Press, accepted for publication.

[151] M. Slatkin. Epigenetic inheritance and the missing heritability problem. *Genetics*, 182(3):845–850, 2009.

[152] S. A. Smallwood and G. Kelsey. De novo DNA methylation: a germ cell perspective. *Trends Genet*, 28(1):33–42, January 2012.

[153] S. A. Smallwood, S. Tomizawa, F. Krueger, N. Ruf, N. Carli, et al. Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nature Genetics*, 43(8):811–814, June 2011.

[154] D. J. Smiraglia, L. J. Rush, M. l. C. Fruhwald, Z. Dai, W. A. Held, et al. Excessive CpG island hypermethylation in cancer cell lines versus primary human malignancies. *Hum. Mol. Genet.*, 10:1413–1419, 2001.

[155] B. D. Strahl and C. D. Allis. The language of covalent histone modifications. *Nature*, 403(6765):41–45, 2000.

[156] R. Straussman, D. Nejman, D. Roberts, I. Steinfeld, B. Blum, N. Benvenisty, I. Simon, Z. Yakhini, and H. Cedar. Developmental programming of CpG island methylation profiles in the human genome. *Nature Structural and Molecular Biology*, 16(5):564–571, 2009.

[157] Y. Sujuan, A. Asaithambi, and Y. Liu. CpGIF: an algorithm for the identification of CpG islands. *Bioinformation*, 2(8):335–8, 2008.

[158] M. M. M. Suzuki and A. Bird. DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews. Genetics*, 9:465–476, May 2008.

[159] J. Sved and A. Bird. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proceedings of the National Academy of Sciences*, 87(12):4692–4696, 1990.

[160] D. Takai and P. A. Jones. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences of the United States of America*, 99(6):3740–3745, 2002.

[161] M. Taub, H. C. Bravo, and R. A. Irizarry. Overcoming bias and systematic errors in next generation sequencing data. *Genome Medicine*, 2:87, 2010.

[162] G. Toperoff, D. Aran, J.D. Kark, M. Rosenberg, T. Dubnikov, et al. Genome-wide survey reveals predisposing diabetes type 2-related DNA methylation variations in human peripheral blood. *Human molecular genetics*, 21(2):371–383, 2012.

[163] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28:511–515, May 2010.

[164] B. M. Turner. Defining an epigenetic code. *Nature Cell Biology*, 9(1):2–6, 2007.

[165] B. Tycko. Mapping allelespecific DNA methylation: A new tool for maximizing information from GWAS. *The American Journal of Human Genetics*, 86(2):109–112, 2010.

[166] J. G. Underwood, A. V. Uzilov, S. Katzman, C. S. Onodera, J. E. Mainzer, D. H. Mathews, et al. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nature Methods*, 7(12):995–1001, 2010.

[167] N. Vergne. Drifting Markov models with polynomial drift and applications to DNA sequence. *Statistical Applications in Genetics and Molecular Biology*, 7(1):6, 2008.

[168] H. Wakaguri, R. Yamashita, Y. Suzuki, S. Sugano, and K. Nakai. DBTSS: database of transcription start sites, progress report 2008. *Nucleic Acids Research*, 36(Database issue):D97–D101, 2008.

[169] J. Wang, W. Wang, R. Li, Y. Li, G. Tian, et al. The diploid genome sequence of an Asian individual. *Nature*, 456(7218):60–65, November 2008.

[170] R. Y. Wang, C. W. Gehrke, and M. Ehrlich. Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues. *Nucleic Acids Research*, 8(20):4777–4790, 1980.

[171] Y. Wang and F. C. C. Leung. An evaluation of new criteria for CpG islands in the human genome as gene markers. *Bioinformatics*, 20(7):1170–1177, 2004.

[172] M. Weber, I. Hellmann, M. B. Stadler, L. Ramos, S. Pääbo, M. Rebhan, and D. Schübeler. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics*, 39(4):457–466, 2007.

[173] D. Weigel and V. Colot. Epialleles in plant evolution. *Genome Biology*, 13(10):249, 2012.

[174] M. Widschwendter, H. Fiegl, D. Egle, E. Mueller-Holzner, G. Spizzo, et al. Epigenetic stem cell signature in cancer. *Nat Genet*, 39(2):157–158, 2007.

[175] H. Wu, B. Caffo, H. A. Jaffee, R. A. Irizarry, and A. P. Feinberg. Redefining CpG islands using hidden Markov models. *Biostatistics*, 11(3):499–514, 2010.

[176] X. Yi, Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science Signalling*, 329(5987):75, 2010.

[177] A. Zemach, I. E. McDaniel, P. Silva, and D. Zilberman. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, 328(5980):916–919, 2010.

[178] K. Zhang, J. B. Li, Y. Gao, D. Egli, B. Xie, et al. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nature Methods*, 6(8):613–618, July 2009.

[179] X. Zhang, J. Yazaki, A. Sundaresan, S. Cokus, S. W. L. Chan, H. Chen, I. R. Henderson, P. Shinn, M. Pellegrini, S. E. Jacobsen, and J. R. Ecker. Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell*, 126(6):1189–1201, 2006.

[180] Y. Zhang, C. Rohde, S. Tierling, T. P. Jurkowski, C. Bock, D. Santacruz, S. Ragozin, R. Reinhardt, M. Groth, J. Walter, and A. Jeltsch. DNA methylation analysis of chromosome 21 gene promoters at single base pair and single allele resolution. *PLoS Genetics*, 5(3):e1000438, 2009.

[181] M. J. Ziller, F. Müller, J. Liao, Y. Zhang, H. Gu, C. Bock, et al. Genomic Distribution and Inter-Sample Variation of Non-CpG Methylation across Human Cell Types. *PLoS Genet*, 7(12):e1002389+, December 2011.