Summer 9-1-2014

On the Analysis of DNA Methylation

Michael Stevens
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

School of Engineering and Applied Science

Department of Computer Science and Engineering

Dissertation Examination Committee:
Ting Wang, Chair
Jeremy Buhler
Joe Costello
Gary Stormo
Kilian Weinberger

On the Analysis of DNA Methylation

by

Michael William Stevens

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2014

Saint Louis, Missouri

Contents

[1] Non-biologists readers are referred to Chap 2 for biology background

Acknowledgments

I'd like to acknowledge my committee for their support. In particular, to Gary for giving me a way back in the game when I didn't think anyone would. To Jeremy, actually, the very first faculty I ever talked to when I came asking around about being graduate student, and for introducing me to Ting when trouble brewed again. To Kilian for inspiring me to dig in and just learn the math. To Joe for getting Ting going in methylation. And especially for Ting for taking on the unpleasant task of pushing me when I needed it while also letting me follow my interests.

And thanks goes to the many graduate students and distinguished faculty in both computer science and genetics who make this a cool place to work and the staff that make it all work.

Michael William Stevens

Washington University in Saint Louis
August 2014

Dedicated to all the people who populated the Internet with so much damn good information.

ABSTRACT OF THE DISSERTATION

On the Analysis of DNA Methylation

by

Michael William Stevens

Doctor of Philosophy in Computer Science

Washington University in St. Louis, 2014

Professor Ting Wang, Chair

Recent genome-wide studies lend support to the idea that the patterns of DNA methylation are in some way related either causally or as a readout of cell-type specific protein binding. We lay the groundwork for a framework to test whether the pattern of DNA methylation levels in a cell combined with protein binding models is sufficient to completely describe the location of the component of proteins binding to its genome in an assayed context. There is only one method, whole-genome bisulfite sequencing, *WGBS*, available to study DNA methylation genome-wide at such high resolution, however its accuracy has not been determined on the scale of individual binding locations. We address this with a two-fold approach. First, we developed an alternative high-resolution, whole-genome assay using a combination of an enrichment-based and a restriction-enzyme-based assay of methylation, *methylCRF*. While both assays are considered inferior to WGBS, by using two distinct assays, this method has the advantage that each assay in part cancels out the biases of the other. Additionally, this method is up to 15 times lower in cost than WGBS. By formulating the estimation of methylation from the two methods as a structured prediction problem using a

conditional random field, this work will also address the general problem of incorporating data of varying qualities -a common characteristic of biological data- for the purpose of prediction. We show that methylCRF is concordant with WGBS within the range of two WGBS *methylomes*. Due to the lower cost, we were able to analyze at high-resolution, methylation across more cell-types than previously possible and estimate that 28% of CpGs, in regions comprising 11% of the genome, show variable methylation and are enriched in regulatory regions. Secondly, we show that WGBS has inherent resulution limitations in a read count dependent manner and that the identification of unmethylated regions is highly affected by GC-bias in the underlying protocol suggesting simple estimate procedures may not be sufficient for high-resolution analysis. To address this, we propose a novel approach to DNA methylation analysis using change point detection instead of estimating methylation level directly. However, we show that current change-point detection methods are not robust to methylation signal, we therefore explore how to extend current non-parametric methods to simultaneously find change-points as well as characteristic methylation levels. We believe this framework may have the power to examine the connection between changes in methylation and transcription factor binding in the context of cell-type specific behaviors.

Chapter 1

Introduction

> Button-holes! there is something lively in the very idea of 'em - and trust me, when I get amongst 'em - you gentry with great beards - look as grave as you will - I'll make merry work with my button-holes - I shall have 'em all to myself - 'tis a maiden subject - I shall run foul of no man's wisdom or fine sayings in it.
>
> - Laurence Sterne, The Life and Opinions of Tristram Shandy, Gentleman

1.1 Motivation

The mathematician, John von Neumann, approached the problem of designing self-replicating machines that can evolve. His solution -five years before Watson and Crick [86] proposed a structure for a DNA molecule with complementary pairs of nucleic acids into which could be encoded instructions, as the mechanism for inheritance- was to use a tape to encode information necessary for replicating itself as well as the construction and the execution of the means of replication [61]. Encoding the self-replication ability as information on a tape (as opposed to some property of the material or components the machine is made of -such as

how crystals grow in a lattice) is critical to the design, because it endowed the machine with an unbounded capacity to increase in complexity, not just changes in property, simply by errors in copying of the tape during replication.

We all have a single genome with 3 billion base pairs encoding the information that unfolded during our development to create who we are [44]. However, we each have over 200 distinct cell types that respond differently depending on its identity. Each of our cells respond to some set of stimuli differently. Many of the responses require building proteins and this requires activating the instructions in our genome. The stimulus, in general, comes to the genome via a limited set of DNA-binding proteins. Where these proteins bind in the genome, determines which set of proteins are created in response to which stimulus.

If you believe Darwin [17], then the implication is that our genomes descend from that of a single-celled organisms. A question then is how does a single genome, evolved for a single cell, multiplex to support 100's of different identities? One solution could be to just copy the 'tape' for every new cell type. However, while copying does appear to take place, this doesn't seem to be an explanation for multi-cellularity. The single-celled Amoeba, dubia, has 670 billion bases in its genome [28]. Instead, it appears that the map itself changes -where a particular stimulus evokes one gene to express in one cell type, another gene will express in another.

1.2 Gaps, Directions, and Needs [2]

1.2.1 Increasingly changes in DNA methylation are shown to be related to cell-type specific regulatory regions.

While not required for stem cell survival [82], methylation is required for differentiation [47]. Since methylation of DNA can block some transcription factors from binding [37] and since methylation is not easily modified, it raises the tantalizing possibility that DNA methylation could function, at least in part, as a mechanism that allows one genome to provide a variable set of responses to environmental changes depending on which cell-type it is part of -as

[2]Non-biologists readers are referred to Chap 2 for biology background

such, the axiomatization (possibly in terms of a statistical model) of the patterns of DNA methylation and how it varies could be used to inform a framework that addresses the definition of cell-type itself by defining cell-type as a mapping of stimulus to response as apposed to by morphology, apparent function, or lineage. This quantitative definition of cell-type is both measurable and more robust. It retains coherency in face of growing evidence of cell-type plasticity. It is also better suited to quantitative models required for synthetic biology.

However, the role is likely not simple and very little in known. As mentioned, mouse stem cells without methylation can proliferate and maintain stem cell characteristics [82]. Also, while some transcription factor binding is prevented by methylation, it is specifically bound by others [37]. Additionally, for example the transcription factor, CTCF, has been shown to instead instruct methylation in that it is both necessary and sufficient to de-methylate transgenic-ally inserted promoters and its binding was not altered in embryonic stem cells lacking methylation [76]. However, this is not universal as CTCF binding was absent at a subset of CpG islands (specifically at regions that are known to control genes that are repressed in a parent-specific manner). Nonetheless, 20% of the variation in methylation of low-methylation regions between mouse ESCs and neural progenitors was explained by binding of 126 JASPAR matrices using a linear model taking methylation and sequence-based prediction of factor binding as input. The predicted activities of the individual transcription factors were in agreement with their changes in transcription level.

Additionally, even with whole-genome bisulfite-sequencing (*WGBS*) coverage was low as 8x's, 300k tissue-specific differently methylated regions (*tsDMRs*) were found amoung 17 adult mouse tissues [31]. Known heart-specific factor motifs predicted in p300 binding sites in heart were more specifically marked by tsDMRs than by chromatin-predicted enhancers, Fig 1.1, and so the author suggest using tsDMRs as a method to define putative regulatory regions at high resolution.

Since most cognate sequences of DNA-binding proteins are small, typically 8-20 bp, examining the potential influence of DNA methylation requires requires high-resolution assays. Since CpGs can be up to 500 bp apart, accurate CpG methylation estimates are required. The only methylation assay with this potential is WGBS. However, analysis has only been performed on the region level, and it is not known whether accurate estimates are possible.

1.2.2 Current approaches of assaying DNA methylation are limited to either high-cost, low resolution, or low coverage.

Current whole genome assays of methylation fall into three groups: enrichment-based, restriction-enzyme-based, and bisulfite-based. Enrichment based methods, such as *MeDIP-seq* [87, 53], have low resolution -since the CpG in a read responsible for binding is unknown- and measure enrichment rather than methylation directly. Additionally, MeDIP-seq enrichment is globally sensitive to experimental conditions and locally sensitive to CpG density. Restriction-enzyme methods, like *MRE-seq*, usually only sample less than a third of the CpGs in the genome. Of the three common ways to readout bisulfite treated libraries, both reduced-representation bisulfite-sequencing *RRBS*, and micro-arrays also sample only a fraction of the CpGs in the genome [77]. The third way, whole genome bisulfite treatment followed by next generation sequencing (*NGS*), (*WGBS*) [49, 14, 46], is often considered the gold standard as it both estimates methylation directly and potentially covers all CpGs in the genome. However, because every experiment essentially re-sequences the genome, its application is limited by its high cost. It's has been estimated the 70-80% of the information produced by WGBS is wasted as most CpGs do not change their methylation levels across cell-types. [90]

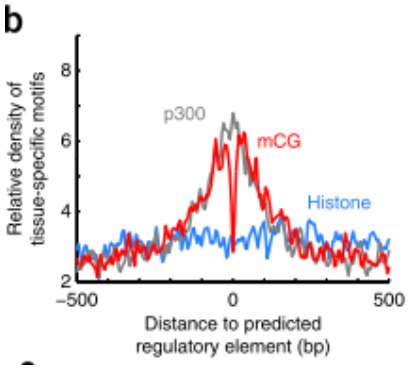


Figure 1.1: Both p33 and tsDMR have higher specifity than histone marks to predicted regulatory elements. *Reproduced from [31]*

1.2.3 The leading assay of DNA methylation has not been extensively studied for quality or accuracy in methylation estimates.

WGBS is assumed to be accurate although it has several unique characteristics among NGS assays that deserve consideration. It is a consensus that as the cost of sequencing goes down, WGBS will be the preferable and dominant assay of methylation. However, there is a lack

of evidence that the consensus is founded on an evaluation of the accuracy of WGBS and is in contrast to some known biases in its methylation estimates. WGBS is a relatively new method with the breakthrough paper in 2009 [50]. In this method, DNA is treated with bisulfite which converts unmethylated C's to T's (and G's to A's for reverse strand reads) which requires non-standard alignment to the reference genome. This results in reads with a fundamentally different base distribution creating an alignment problem qualitatively different than other NGS assays.

Two strategies based on modifications to standard alignment have emerged, wild-card matching and three letter alignment [6], Fig 1.2. In the former, a C in the reference matches either a T or C in the read, while in the later, all T's are converted to C's in both the read and the reference before alignment (for reverse strand reads the conversion is from G to A). Because of the wild-card asymmetry between C and T (read T's match reference C's, while read C's only match C's), reads with less methylation can align more places potentially leading to over-estimates of methylation. While three letter alignment addresses this bias, it does so at the cost of lower coverage as up to 50% of reads are thrown out due ambiguous mapping or low quality. See [6] for a fuller treatment of these biases. The effect of either these methods in terms of alignment methodology has not been characterized.

1.2.4 Analysis of genome-wide DNA methylation has focused on the change in methylation, either between genomic regions or between samples, for example, between cell-types or biological replicates

Although several approaches have been applied, there is no consensus on the best use of the high-resolution, genome-wide methylation data. The first type of analysis focuses on finding specific patterns of methylation organized around a characteristic methylation level. For example, a hidden markov model (*HMM*) was used to segment the genome into methylated, unmethylated, and lowly methylated regions (HMR,UMR,LMR) [76]. LMRs represent about 30% of CpGs in the genome and are located in relatively CpG-poor regions far from protein coding loci but show histone marks for enhancers and are occupied by cell-type-specific transcription factors. The second type of analysis defines and identifies windows of differential

methylation between two experiments or samples (*DMR*'s). For example, using 17 adult mouse tissue methylomes [31], an HMM was able to segment the genome using a χ^2 statistic generated across the samples. The segments generated by the state with the highest χ^2 where considered tissue-specific DMRs representing around 7% of the mouse genome. Using 42 WGBS data sets, in another approach the authors first identified dynamic CpGs using a pair-wise comparison using a beta-difference distribution [90]. In a second step, closely located dynamic CpGs were merged resulting in DMRs which account for around 19% of the human genome.

1.3 Contribution

1.3.1 Develop an alternative high-resolution method to WGBS by integrating MeDIP-seq and MRE-seq assays using a statistical model.

We combined methylation predictions from two relatively cheap measures of DNA methylation in a statistical model to approximate the high-resolution estimates of the more expensive method, WGBS. In addition to significant cost-savings, this method will provide an alternative to compare the accuracy of WGBS, will demonstrate how using two assays provides higher accuracy, and will demonstrate a general method for combining assays of varying quality into predictive models.

MeDIP-seq and MRE-seq are independent, yet complimentary, assays of methylation. MeDIP-seq is a precipitation-based method and so enriches for methylated CpGs, while MRE-seq is a restriction enzyme-based method and so isolates fragments between pairs of unmethylated CpGs. MeDIP-seq is proportional to methylation while MRE-seq is inversely proportional, Fig 1.3. While MeDIP-seq is enrichment-based and MRE-seq is rather sparse (requiring multiple unmethylated CpGs within the limit of the fragment size), it was shown that the combination of the two could provide CpG methylation estimates genome-wide. [30]

Figure 1.2: Approaches to WGBS alignment: wildcard versus 3-letter. *Reproduced from [6]*

We used a conditional random field (*CRF*) [41] for structured prediction of methylation by using both MeDIP-seq and MRE-seq as input to predict single-CpG methylation genome-wide. CRFs are discriminately trained graphical models that model the probability of the output, in this case methylation, conditioned on the global observations, in this case MeDIP-seq and MRE-seq. In contrast to HMMs which model the joint probability of the observation and

output, CRFs take the observation as a given and so don't waste effort modeling dependencies amongst the observations. As such, they should be able to train and predict efficiently even as the size of the model increases. Additionally as the number of parameters grow, CRFs are less prone to over-fitting than traditional HMMs. As HMM parameters are learned with maximum likelihood estimates using a training data set, with enough random variables an HMM could effectively memorize its training data. CRFs, on the other hand, are typically trained by iteratively ascending their gradients and are stopped when performance on a second, test, data set decreases. In this way, a CRF is in a sense optimized on its training data until the point where further optimization would result in learning the sample distribution instead of the desired true distribution.

Most work on CRFs has been done in the field of natural language processing and so model words -which are a discrete type data. MeDIP-seq and MRE-seq are pseudo-continuous in that even though they take on discrete values, the number of unique values is too large for even a moderate sized CRF. While a continuous extension for CRFs does exist, [67], it is well known that discretized data can do better than continuous data in Bayesian networks.

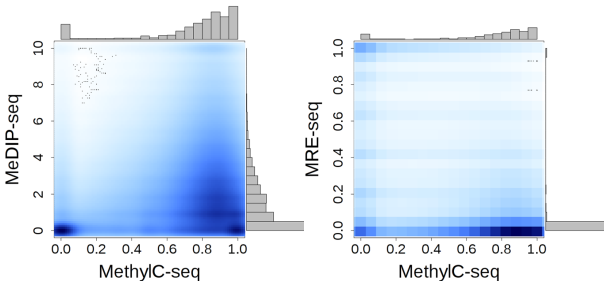


Figure 1.3: MeDIP/MRE are related to WGBS

CRFs are already known to handle millions of features [41] and good implementations already exist, so it was not necessary to build our own. Since CRFs condition on the observations and so don't need to examine their interaction, we included additional data likely to effect methylation as observations in the model: CpG density, the local GC%, and different kinds of genomic features, such as CpG-islands versus exons. We allowed ℓ_1 normalization to perform feature selection for us.

Concordance between methylCRF and WGBS CpGs in a 25% methylation window is within the range of concordance between two WGBS data sets. The concordance did not appreciably decrease on a separate cell type. methylCRF is provided as a turn-key application for a

target audience of bioinformaticians or biologists familar with the standard GNU tool-chain. methylCRF is up to 15 times cheaper than WGBS.

1.3.2 Show methylCRF methylomes are generally comparable to WGBS methylomes for biological insight

Using data generated from the Epigenetic Roadmap Consortium [5], we compared 54 methylCRF methylomes and 33 WGBS methylomes -more than doubling the size of previous analyses. We found that high resolution analysis provides novel biological insights including that an estimated 28% of CpGs show significant differences between cell types and when clustered account for 11% of the genome. These regions are suggestive of regulatory potential as they are enriched for enhancer marks and transcription factor binding in a tissue-dependent manner.

1.3.3 Joint analysis of methylCRF and WGBS identifies systemic WGBS biases suggesting limited resolution in both methylation levels and individual CpGs.

On top of a reduced alphabet for alignment, WGBS puts greater demands on NGS than do enrichment-based assays. Other than genotyping, NGS is mostly used to measure enrichment. However, the final readout of WGBS is a ratio of converted and unconverted Cs. Seen as a Bernoulli trial (where the true rate of methylation is the probability of success), several issues become apparent. The number of converted Cs is subject to random fluctuations due to among other things sampling of a small number of alleles from a library, PCR-biases, and sequencing-biases. Therefore one can expect that the computed methylation is a combination of true methylation and error. Combined with the fact that WGBS non-specifically sequences the whole genome, this highlights the fundamental trade-off of WGBS cost versus accuracy due to small-sample size. As an example, even with 100x's coverage 6% of CpGs will vary more than 10% from true methylation due solely to error induced by sampling, while at 10x's coverage 80% of CpGs will vary more than this and more than 10% of CpGs will vary more than 25% from true methylation.

Even on closely related data sets WGBS, shows discordance at levels beyond expectation. We looked at WGBS generated from embryonic stem cell (*ESCs*) lines, H1, H9, and HSF1 and only 45-55% of the CpGs genome-wide were within 10% methylation of each other (75-90% were within 25%). When we looked at two H1 datasets, these numbers did not increase. Using two different processing pipelines on the same library, still only 83% of CpGs were within 10% of each other (97% were within 25%).

These results suggests that WGBS assays require a statistical model for accurate estimation as well as better comparability with methylCRF. Currently methylation levels are characterized in coarse resolution: methylated, intermediately methylated, lowly methylated, and unmethylated. Since on a single gemone, a CpG is either methylated or not, there doesn't seem to be biological justification for a resolution on the order of, say, 100 distinct levels of methylation. Additionally as mentioned, there is an inherent small-sampling issues when calculating methylation from a small sample of reads. Therefore, it should be more accurate to include an estimate of confidence in a methylation call. Additionally, since methylation levels are locally correlated [77], we also used an HMM, (*Twiposn*), to model methylation estimates. However, we used binomial emission probabilities to vary probabilities for different read count. Using expectation maximization to learn up to five distinct states, on several data sets, this model could only distinguish between a highly and one or two lowly methylated states -which does not even recover known biology. This result suggest that even with standard statistical models, WGBS can not both simultaneously recover known biology as well as accurately estimate methylatio.n Additionally, although read-count-based HMM approaches have the potential to correct for the small-sample issue, it does so as a compromise in two ways. One, since it borrows information from neighboring CpGs, it effectively reduces the ability to detect real changes in methylation by mistaking it for noise, and secondly, it does impose a limitation in the resolution of the dynamic range of methylation.

1.3.4 $TV_{reductio}$: a non-parametric statistical method for filtering WGBS methylation to detect features suitable for DNA-binding events.

Due to the limitations of both methylCRF and WGBS detailed above, we present a new representation of a methylome as a signal and focus directly on the problem of detecting changes in the signal that are potentially associated with factor binding. This is a work in progress. I propose to do this by co-optimizing 1) the number of distinct changes in methylation using step-detection via total variance de-noising (*TVD*) [72] and 2) the likelihood of the fewest number of known transcription factors binding to either the boundaries or the interiors of the windows of low-methylation defined by the changes. This method will only require pre-computed binding potentials of known transcription factors and a genome-wide, single-CpG resolution assay of methylation -either WGBS or methylCRF. I can use existing genome-wide assays of transcription factor binding (*ChIP-seq*) to determine both the specificity and sensitivity of the method.

I hypothesize that a significant subset of changes in methylation are a result of transcription factor binding. In this sense, a change in methylation refers to any CpGs that are not methylated in an experiment and are, as well, not constitutively unmethylated. Specifically, I will look for changes in methylation best explained by the binding potential of the least number of transcription factors. Further, since this method only requires one assay, it could potentially generalize both transcription factor and chromatin mark ChIP-seq thus providing a single method to determine both the transcription factors binding in a cellular context as well as their combinatorial relation providing hypotheses as to the causes of changes in transcription. This could be a general method for the community in order economically study this relationship in any biological context.

Additionally, this method will extend the use of TVD to a new field and extend the methodology of TVD with a new class of constraints.

1.3.5 Subtext

As a computer scientist working exclusively in biology, I have struggled, seemingly endlessly, with the question: what am I doing? Is my work data analysis, is it determining which problems are computable, is it data mining, is it biology, is it statistics or machine learning? Biology, like many fields, is becoming a *big data* field in part due to the success of computer science and engineering and venture capital. This data presents all kinds of new challenges and there is a gap in the ability and capacity to absorb and cull results of the data it is generating. Mathematicians, chemists, physicists, statisticians, economists, engineers, computer scientists, and a new breed of biologists are all converging in these gaps. From the perspective of biology each one is useful in terms of what they can contribute to biological understanding. Each of these fields have their own tools, traditions, and communities, though, requiring biologists to learn all of the in-and-outs, the caveats, and the gotchas for each one individually in order to incorporate their methods responsibly into their research -which, of course, would leave no time to study biology. There is a need, then, of some sort of general methodology for big data biology. It is natural to look to these technical and mathematical fields for guidance. However, each field has its own focus and concerns. As nominally domain agnostic, I would root a methodology in machine learning and data mining both of which are a mixture of fields including math and statistics. It is necessary, then, to first understand what are the distinct concerns and issues with big data biology to determine what parts of these fields to take, what to leave, and what to modify. This is the first step in establishing a sub-domain of these fields as work in it can dually extend biological knowledge as well as insight to the nature of data itself.

Biology is stupendously complex and different in many ways from the rest of the physical world and offers many distinctive challenges that have historically driven advances in math and statistics. For example, in order to study the inheritance of seed size, Galton invented the regression line and described the concept of regression toward the mean [12]. I propose these as the core concerns for big data biology:

1. Representation: While machine learning is concerned with representation, it refers to the form of the data used as input to a method, such as representing a photograph as a vector of pixel intensities. Here I refer to representation as the form of the *result* of the model, for example, a cluster of photographs containing a similar object. This includes

statistical concepts like point estimates, confidence intervals, and hypothesis tests. It also includes less well defined things as clusters, change points, graphs, or a smoothed signals. For example, a kernel smoother applied to a methylation signal would be represented by a smoothly changing signal with noise filtered out. Trivial information that could be extracted from this representation are things like point estimates of a region or a list regions with the highest or lowest values. However, this representation in itself, can not trivially produce a set of methylated regions in the signal or whether a dip in the signal is significant or not.

2. Encoded intuition. This is prior knowledge about an object of study that is somehow encoded in the model. For example, in methylCRF we used the knowledge that close CpGs are correlated as a first-order state transition distribution in a CRF and an HMM in Twiposn. We used the same intuition as a regularizer in $TV_{reductio}$.

3. Generalizability. This is somewhat central to all science. In machine learning it refers to accuracy on unseen data. In learning theory it is related to justifiable induction. It refers to whether the accuracy of prediction on future data as well as the likelihood of results being spurious random events. This is further elaborated in Chapter 3.

4. Interpretability. This is the ultimate goal of any model in biology. This is measure of how useful a representation and a model is in context to what is known. This involves the representation of the result as well as the assumptions of the model, the way the model uses the data, and the confidence in the result. This can be a major portion of an analysis and often the hardest part. Methods that are simple in these terms are of great value -even at the expense of accuracy. For example, it was argued that since all models are approximations using samples (an approximation for the object of interest), exact learning is not necessary [8]. While this is true, there is a cost in interpretability as this additional approximation adds another factor to consider in the meaning or the reliability of results. Additionally, interpretation includes convincing skeptical experts in the domain.

Since this is very much just a start and beyond the scope of this thesis, I will instead, through this work, explore and place in context through working examples, these concepts. These ideas are not new and in some form or another are currently used in practice by most people in all technical fields but not explicitly focused on as essentials of the quality of biological data

and for the use in biology. It is my hope that in this era of the emergence of big data, as a field, by focusing on these core issues, big data biology, will grow into a symbiotic relationship with math, statistics, and other technical fields where the flow goes in both directions.

Chapter 2

Biology Background

> I regard as quite useless the reading of large treatises of pure analysis: too large a number of methods pass at once before the eyes. It is in the works of applications that one must study them; one judges their ability there and one apprises the manner of making use of them.
>
> Joseph Louis Lagrange

Since the completion of the draft of the human reference genome in 2001 [44], genomics has revolutionized the study of biology including allowing the easy association of genomic loci to disease, the study and effect of genomic diversity, the location of where DNA interacting proteins bind, a representative catalog of protein coding genes, global analysis of transcript expression levels, the study of chromosomal modifications in cancer, and the global study of chemical modifications to chromosomal constituent molecules whose dynamic patterns help define cell-type specific responses to stimuli [57]. All of these have been facilitated by the development of the post-reference genome technology of next generation sequencing (*NGS*). Where the reference human genome project cost $2.7 billion, NGS-based genomes are now, for example, publicly available from DNADTC for $7000.

As exemplified by Illumina sequencing, NGS involves first fragmenting large sequences into smaller fragments, via sonification for example, which are then read out producing 100-150

nucleotides (or *base pairs*) long sequences (*100-150 bp reads*). According the HudsonAlpha's latest statistics, a single experiment using Illumina's newest HiSeq 2500, can produce up to 300 million 100 bp reads passing a basic quality filter. Given the reference genome size of 3 billion bp, an experiment can produce then theoretically on average up to 10 reads covering every base, referred to as *10x's coverage*. The genome from DNADTC mentioned above provides 30x's coverage.

In addition to sequencing genomes, NGS is combined with a host of other molecular biology techniques to provide a variety of information. In enrichment-based techniques a 'probe' molecule that has an affinity for another molecule of interest is used to bind to that molecule and is designed in a such way that it can be isolated from the rest of the milieu of the cell. If the molecule of interest is associated with DNA, then the DNA isolated with the molecule of interest can be sequenced to show where in the genome that molecule was associating with DNA. Versions of this technique allow the determination of the location of *DNA-binding proteins*. These are often proteins that are correlated with and widely considered to causally regulate variously the promotion, enhancement, or inhibition of gene transcription and are along with their binding partners known as *transcription factors or TFs*. Additionally, the molecule of interest can be as specific as a particular chemical modification of a molecule, commonly referred to as a *mark*. Individual marks of interest include a large class that histones are subject to -*histones* are proteins that DNA is packaged with. These marks, among other things, can effect which genomic locations transcription factors can bind. And so, since these marks are reversible, changes in their patterns are potentially one source of cell-type specific behavior. In enrichment-based techniques, the more reads that align to a genomic region, the more consistently your molecule of interest occurs at that location. However, due to variability of many layered experimental factors (in tissue preparation, in enzyme specificity, in sequencing process, in the analysis process, etc), enrichment values are not comparable across experiments.

Another NGS-based method utilizes *restriction endonucleases* -useful because these enzymes are able to cut DNA in a sequence-specific manner. The specific sequences recognized by the enzymes are typically around 4-8bp long and so may occur at up to 12 million sites in the genome. However, only the fragments that are of a certain length can be sequenced, so the amount of the genome accessible to this technique is limited. DNA itself is subject to one known chemical modification, the addition of a methyl group to a base, referred to as

DNA methylation. There are *methylation sensitive enzymes* that will only cut the DNA if it is unmethylated. This technique, *MRE-seq* [53], allows the sequencing of fragments whose both ends are unmethylated and so gives an estimate of where DNA is not methylated.

DNA exists as a double stranded molecule made of complementary base pairs. Most enzymatic reactions, like DNA synthesis, are directional, and so the two strands are anti-parallel. In vertebrates, DNA methylation primarily occurs on a C which is followed by a G, and is termed a *CpG*. The human reference genome contains around 28 million CpGs. Whether CpGs are methylated or not has effects on whether proteins can bind to DNA, gene expression, the local and global structure of chromosomes, the ability of the cell to defend against parasitic sequences from duplicating themselves, embryonic development, transcription, sex-specific chromosome modifications, as well as the ability to control genes based on their parent-of-origin ([69, 79, 42, 34]).

Transcription refers the process of creating RNA from DNA. According to ideas put forth in the 'central dogma' [15], this is the intermediate step in the creation of proteins from instructions encoded in DNA. Simply put, the initiation of gene transcription is controlled via the binding of transcription factors in a gene's *promoter* which is a region on the genome preceding the transcription start site. Additionally, transcription factor binding to sites non-local to the transcription start sites, *enhancers*, can also effect the decisions for transcription as well as effect the amount of transcription as measured in a population of cells. It is widely thought that the identification of transcription factors binding sites, (*TFBS's*), the factors that bind them, and the genes they effect is an essential step in unraveling gene regulatory networks and thus gene regulation itself. The sequence features that a particular DNA-binding binds to can be represented as a position-weight-matrix, (*PWM*), [78] where the binding potential of a TFBS to a transcription factor is factorized by the contributions of the individual bases in the sequence. Databases such as JASPER [66] and TRANSFAC [52] provide pre-computed estimate PWM's for a number of transcription factors.

Chapter 3

Justified Induction

> ... And then came the grandest idea of all! We actually made a map of the country, on the scale of a mile to the mile!"
> "Have you used it much?" I enquired.
> "It has never been spread out, yet," said Mein Herr: "the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well.
>
> -Lewis Carroll, Sylvie and Bruno Concluded

3.1 A Story

First let me draw a caricature and gross simplification of the use of large data sets in biology labs. A graduate student, let's call him or her (*'G' could be either, of course*), G Iwa Napayper, working in Gottap Ublish's lab. Both G and Gottap came away from Francy Slide's presentation a little slack-jawed. Francy had an unbiased, data driven result about a pathway related to theirs. There were heatmaps, bar charts with lots of bars, something called a kernel density plot, a connectivity cloud, and lists of important genes. Interestingly,

the list of genes showed enrichment in pathways related to things no one ever considered would be connected. Gottap was thinking, wow, 1) *un-biased* and *data driven* are powerful arguments, maybe this is the future of biology, and 2) if I don't get some results like this soon, my competition will. G was thinking, dang, that looks a lot easier way and a lot faster way to publish something. So, they agree to broaden G's project to include an RNA-seq library. G asks people in the labs around who've done it, they order their reagents, set up their model system, and go for it. After 6 months G has finally submitted and gotten back data and again checks the tribal knowledge for what others have used -RNA-seq isn't so rare anymore, so there are a few options. G picks the easiest. He (*shorter to write*) tries it but only gets a couple differentially expressed genes. He runs the program a few more times with the same results, fear grips him as he wonders whether he screwed up the expensive experiment. He calms down after re-checking over the weekend, the QC he did. He asks around and gets told to try this other program. This time it worked he got a many, many more differentially expressed genes. Some of them are the canonical members of their pathway, so it worked. Gottap's smells a headline opportunity and asks G to look for enrichment for the genes they found. It turns out the genes are enriched in eye development, dna metabolism, and viral response. Considering they are studying the effect of alcohol on liver enzyme production, they are blown away by the depth of the possible connections! The connections are solid, because the enrichment for this gene set has a p-value of 0.0000005, so there's no way it's just spurious. G starts looking in the literature for possible connections and they start formulating a story on why liver function is connected to eye development. G shows his results to Gottap, and they plan out experiments to test if one of the eye development genes is necessary for their phenotype in liver as well as another just as a backup. Three years later G and Gottap publish on the backup gene. In their discussion they talk about the connections to eye development and sketch out a new theory incorporating their liver phenotype to some part of eye development. The reviewers were also microbiologists, so they focused on the controls used for the experiments on the backup gene.

You may agree or disagree with G's and Gottap's approach (*it is a caricature after-all*). To some degree this is happening in labs across the world as biologists try to incorporate the large amount of data they are now generating. A central question in biology today is how to learn from this data to gain insights about the machinery, processes, and dynamics of living things. Given limited resources, it would be folly to fund biologists to test all possible inferences in a trial-and-error random walk. The generation of model systems and

the design and execution of controlled experiments can take years, many man-hours and many materials. It is of central important to the field, then, in order to maximize its return to society, to maximize the likelihood of insights drawn from large data sets. This is the problem of *induction*. Similar to the theory of science, *learning theory* provides a framework for reasoning about what is *justifiable induction*.

3.2 Generalizability

Most people are familiar with *Ockham's razor* (also referred to as *the principle of parsimony*) and most scientists either explicitly or implicitly follow it as a maxim in their approach to science. Among the many versions and phrasings, the principle can be summarized as: a simpler explanation that explains the data is preferable to a more complicated one. In biology, where experiments can be long and costly and where so little is known about the degree of connection between entities, the idea of simplicity, is in many ways used as a proxy for repeatability. That is to say that if the inferences from an experiment require little additional theory or modifications to existing models, then it has some credibility even without being repeated. In this way, simplicity is used as a heuristic to distinguish the quality of a model or hypothesis, that is, it is a direct property of the model object.

When applied to specifically to mathematical models, Ockham's razor is usually interpreted as the number of variables in the model. A model of one variable is simpler than one of two. A linear model is simpler than a second-degree polynomial. A simpler model that describes some set of data as accurately as a more complicated one is more likely to capture the underlying structure of the data and, so, *generalize* better in terms of giving similar accuracy on future data.

The conjecture of a learnable underlying structure is turned inside out by Popper [65], however, where he equates simplicity to *falsafiability*. In the extreme, a model or theory that is not falsifiable is not even scientific. In the example above, then, a linear model is better than a non-linear model specifically because, given the same domain, the linear model can accurately represent quantifiably fewer mappings from the domain to the range. He changes the focus of justifiable induction from a heuristic about a quality of the model to a lower bound on its applicability in some domain at a given level of accuracy. This change in focus from

accurately representing the object of induction, to a performance-based criteria, may have opened the doors to use the parsimony heuristic in new ways to improve generalizability.

In the 1960's some statisticians started to examine completely different forms of simplicity, ie other than the number of parameters, and learning theorists provided tools to think about and evaluate both forms in a unified way. Using a similar lower-bound perspect, Vapnik formulated the problem of justifiable induction or learning [83], in general, as a problem of function estimation of some data distribution -presumably representing something of interest. His estimation model consist of three components:

- G: some generator of random vectors $x \in \mathbb{R}^n$ from some fixed, but unknown distribution $P(X)$.
- S: a supervisor which returns a mapping from $x \rightarrow y$, according to a, also fixed, but unknown, conditional probability distribution $P(y|x)$.
- LM: a learning machine capable of implementing a set of functions $f(x, \alpha)$, $\alpha \in \Lambda$ where Λ is a set of vectors or more generally, a set of functions.

The learning problem, then is to choose a function f from the set of functions that best predicts S in the best possible way. One notion of the quality of a prediction is the expected discrepancy between S and f for a given x. The discrepancy function, L is referred to as the *loss* whose expectation or *risk*, R is is given by:

$$R(\alpha) = \int L(y, f(x, \alpha)) dP(x, y) \tag{3.1}$$

Using R, the task then is to learn a function, $f(x, \alpha_0)$ that minimizes $R(\alpha)$. However, we don't know $P(X, Y)$ or $P(Y|X)$ and, further, in most cases only a small sample of data from $F(x, y)$ is available to train with. We can instead, as a proxy for R, use the principle of induction using *empirical risk minimization* over the training data we have:

$$R_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \alpha) \tag{3.2}$$

where Q is a generalized loss function -in a form needed to include density estimation- and $z_i = (x_i, y_i)$. Note that this form of induction does not require any knowledge or explicit forms of the distributions. How does this work then? -it's so general Q could just be an explicit map from every y to an x.

The *Key Theorem* of ERM learning specifies that to be consistent it is necessary and sufficient conditions for the proxy, R_{emp}, to *uniformly* converge in probability to R in the limit as $\ell \to \infty$. Here converge in probability is used in a PAC sense, that is, there exists a sample size $m(\delta, \epsilon)$ for which $P\{R_{emp} - R \leq \epsilon\} < \delta$ for any ϵ, $\delta > 0$. The condition is essentially that even in the worst case, R_{emp} converges to R. Specifically:

$$\lim_{\ell \to \infty} Prob \left\{ \sup_{\alpha \in \Lambda} (R(\alpha) - R_{emp}(\alpha)) > \epsilon \right\} = 0 \ , \forall \epsilon \tag{3.3}$$

To describe a constructive formulation of when this equality holds, the notion of capacity is required. For simplicity, we'll discuss the case for where Q is an indicator function giving 0 if $f(x) = y$ and 1 otherwise. Consider a pseudo-boolean function of a particular class of functions, Λ, evaluated on a particular data set, $(z_1, ...z_\ell)$:

$$N^{\Lambda}(z_1, ...z_i) : Q(z, \alpha)^{\ell} \to \mathbb{Z} \tag{3.4}$$

$N(\Lambda)$ gives the number of binary vectors possible across all $\alpha \in \Lambda$. From this we can define the *growth function*:

$$G^{\Lambda}(\ell) : \ \ln \left\{ \sup_{z_1, ...z_\ell} N^{\Lambda}(z_1, ..., z_\ell) \right\} \tag{3.5}$$

Note that if $G^{\Lambda}(\ell) = \ln 2^{\ell}$ there exists an $\alpha \in \Lambda$ for every possible combination of $(x, y)^{\ell}$ and so Λ has the capacity to learn any configuration, that is, Λ is unfalsifiable for a set of ℓ data points. If however:

$$\lim_{\ell \to \infty} \frac{G^{\Lambda}(\ell)}{\ell} = 0 \tag{3.6}$$

then Eq. (3.3) holds, and so Eq. (3.6) is necessary and sufficient for justifiable induction using ERM induction. It turns out that the growth function either grows linearly in ℓ or is bounded by a logarithmic function. Further, if Eq. (3.6) holds, then the rate of convergence is also *fast*. A rate of convergence is fast if for any $\ell > \ell_0$:

$$P\{R(\alpha_\ell) - R(\alpha_0) > \epsilon\} < \exp^{-c\epsilon^2\ell} \tag{3.7}$$

where $c > 0$ is a constant.

Further, the implication from all this is a bound on the real risk, R, (which uniformly decreases with ℓ):

$$R(\alpha) \leq R_{emp}(\alpha) + (B - A)\sqrt{\frac{G^{\Lambda,B}(2,\ell)}{\ell} - \frac{ln(\eta/4)}{\ell}} \tag{3.8}$$

where A and B are the smallest and largest value Q takes over all $\alpha \in \Lambda$ and $1 - \eta$ is the probability that the inequality holds. Note that this is a distribution free bound meaning it holds for any admissible Q. One important implication from the inequality is that when you have enough data -'enough' defined by the growth function, the empirical risk is almost the same as true risk. The further value of this inequality, then, depends on situations where data is limited.

In this case, you can see that in order to reduce the risk, $R(\alpha)$, there are only a few options available. One is to reduce the empirical risk. Since we always minimize, this is only an issue for classes of functions for which there is no method to find the global optima. Otherwise, when there are methods to find global optima, there is nothing more to be done method-wise -except for computational concerns such as memory or time constraints. The other way is to choose a Λ with a smaller growth function, ie, a smaller VC-dimension -explained below. The final thing you can do is to get more data. This also shows the trade-off between the capacity of the class of functions and the amount of data. A final approach is to reduce your expectation being correct -that is, increasing η also decreases $R(\alpha)$ which indeed seems like a weird thing to do. However, if, for example, as a meta-approach, multiple models on different data lead to the same underlying conclusion, your confidence of the results may be high even if individually each model has a lower probability of generalization. While all of these factors are intuitively important, this line of learning theory shows that these are the only things that are important in determining justifiable induction.

Since the growth function is often difficult to derive for a given Λ, the *VC dimension* is used instead. The VC dimension for a set of indicator functions, Λ, applied to a data set is the largest ℓ for which $G^\Lambda(\ell) = 2^\ell$. The VC-dimension is infinite if there is not such ℓ, otherwise,

the VC-dimension is finite and is also necessary and sufficient for Eq. (3.6). It turns out that the VC-dimension is the point where the growth function switches from linear to logarithm, so reducing the VC-dimension provides a way to reduce the growth function.

One might ask why this is all necessary. As long as you cross-validate, the model should generalize well completely independent of any sense of parsimony or capacity. The assumption of cross-validation, however, is that the training data is representative of G. An almost universal character of biological data is that 1) it is derived from some technique that is a proxy for G, and 2) the data has multiple layers of systemic biases due to any number of issues in biological variance, preparation, sensing, or in analysis. WGBS, for example, originally had only one sample, and has many biases (detailed in our later chapters). Also, note that G in this case is methylation on chromosomes, but WGBS measures C-to-T conversion rate across a population of cells -the two are not the same thing.

I propose, then, that the concept of training error plus capacity is useful -even in the case of learning a representation of the data.

Chapter 4

Estimating absolute methylation levels at single CpG resolution from methylation enrichment and restriction enzyme sequencing methods.

Content appeared in [77].

> The art of discovering the causes of phenomena, or true hypothesis, is like the art of decyphering, in which an ingenious conjecture greatly shortens the road.
>
> Gottfried Wilhelm Leibniz

4.1 Abstract

Recent advancements in sequencing-based DNA methylation profiling methods provide an unprecedented opportunity to map complete DNA methylomes. These include whole genome

bisulfite sequencing (WGBS, MethylC-seq, or BS-seq), Reduced-Representation Bisulfite-Sequencing (RRBS), and enrichment-based methods such as MeDIP-seq, MBD-seq, and MRE-seq. These methods yield largely comparable results, but differ significantly in extent of genomic CpG coverage, resolution, quantitative accuracy, and cost, at least while using current algorithms to interrogate the data. None of these existing methods provides single-CpG resolution, comprehensive genome-wide coverage, and cost feasibility for a typical laboratory. We introduce methylCRF, a novel Conditional Random Fields-based algorithm that integrates methylated DNA immunoprecipitation (MeDIP-seq) and methylation-sensitive restriction enzyme (MRE-seq) sequencing data to predict DNA methylation levels at single CpG resolution. Our method is a combined computational and experimental strategy to produce DNA methylomes of all 28 million CpGs in the human genome for a fraction (<10%) of the cost of whole genome bisulfite sequencing methods. methylCRF was benchmarked for accuracy against Infinium arrays, RRBS, WGBS sequencing, and locus specific-bisulfite sequencing performed on the same human embryonic stem cell line. methylCRF transformation of MeDIP-seq/MRE-seq was equivalent to a biological replicate of WGBS in quantification, coverage and resolution. We used conventional bisulfite conversion, PCR, cloning and sequencing to validate loci where our predictions do not agree with whole genome bisulfite data, and in 11 out of 12 cases methylCRF predictions of methylation level agree better with validated results than does whole genome bisulfite sequencing. Therefore, methylCRF transformation of MeDIP-seq/MRE-seq data provides an accurate, inexpensive and widely accessible strategy to create full DNA methylomes.

4.2 Introduction

The haploid human genome contains approximately 28 million CpGs that exist in methylated, hydroxymethylated, or unmethylated states. The methylation status of cytosines in CpGs influences protein-DNA interactions, gene expression, and chromatin structure and stability, and plays a vital role in the regulation of cellular processes including host defense against endogenous parasitic sequences, embryonic development, transcription, X chromosome inactivation, and genomic imprinting, as well as possibly playing a role in learning and memory [34, 42, 69, 79]. Understanding the role of DNA methylation in development and disease requires accurate assessment of the genomic distribution of these modifications [42]. Recent

advancements in sequencing-based DNA methylation profiling methods provide an unprecedented opportunity to map complete DNA methylomes. Techniques for high throughput detection of cytosine methylation include bisulfite conversion of unmethylated cytosines to uracil, immunoprecipitation with antibodies specific for methylated DNA, and cleavage of CpG containing restriction sites by methylation sensitive or methylation-dependent restriction endonucleases followed by sequencing or microarray hybridization [6].

The most comprehensive method, bisulfite treatment followed by sequencing, (whole genome bisulfite sequencing, or WGBS, including MethylC-seq [50] and BS-seq [14, 46], measures single cytosine methylation levels genome- wide and directly estimates the ratio of molecules methylated rather than enrichment levels. However this method requires essentially re-sequencing the entire genome multiple times for every experiment (with up to half the reads not even covering CpG sites). To obtain a complete DNA methylome, the total sequencing depth required for adequate coverage of each strand is equivalent to 30X of the human genome (90Gb) which remains an expensive experiment. In addition to its high cost, bisulfite converted genomes have lower sequence complexity and reduced GC content. Therefore, performance of WGBS based methods is also influenced by potential differences in the efficiency of amplification of methylated and unmethylated DNA copies of the same locus, and the ability to accurately align bisulfite converted sequencing reads to the genome, which is more challenging than alignment of conventional reads [39]. As noted, 10% of CpGs in the mammalian genome remain refractory to alignment of bisulfite-converted reads [42].

Reduced Representation Bisulfite-Sequencing (RRBS) [54] addresses the cost issue by measuring single CpG methylation only in CpG dense regions. For the human genome, it requires only around 3Gb of sequencing to achieve the same degree of sequencing depth for most regions of interest. However, RRBS's ability to interrogate a locus is dependent on its MspI cut-site (CCGG) density and consequently measures 10-15% of the CpGs in the human genome [7, 30].

Restriction enzyme methods (e.g. MRE-seq [53]), on the other hand, typically incorporate parallel digestions with 3-5 restriction endonucleases. Utilizing multiple cut- sites, MRE-seq can cover close to 30% of the genome and saturates at 3 Gb of sequencing [58]. These methylation-sensitive enzymes cut only restriction sites with unmethylated CpGs and each read indicates the status of a single CpG. While methylated CpG's could be inferred by the

absence of reads at cutting sites, this would require assuming perfect digestion that is not typically done in practice.

In contrast to MRE-seq, methods utilizing monoclonal antibodies against 5-methylcytosine (MeDIP-seq) ([87, 53]) or methylated DNA-binding proteins (MBD-seq, domains of MBD2 alone, or in combination with MBD3L) [74] rich for methylated DNA independent of DNA sequence have been estimated to saturate at 5 Gb of sequencing [58]. An important advantage of MeDIP over restriction enzyme methods is a lack of bias for a specific nucleotide sequence, other than CpGs. However, the relationship of enrichment to absolute methylation levels is confounded by variables such as CpG density [63]. Another inherent limitation of MeDIP-seq in its current form is its lower resolution (≈150bp) compared to MRE-seq or bisulfite-based methods in that one or more of the CpGs in the immunoprecipitated DNA fragment could be responsible for the antibody binding.

Finally, array-based platforms are widely utilized. Approaches which couple bisulfite conversions with hybridization-based arrays (e.g. Illumina's HumanMethylation450 BeadChip arrays), while having single base pair resolution, are limited to a priori targeted regions. For example, the Illumina BeadChip array assesses methylation at 485K targeted CpGs, averaging 17 CpGs per gene spread across CpG islands and gene loci. (For this analysis we also utilized the previous BeadChip version which contains 27K CpGs.)

As sequencing costs drop, the number of complete single nucleotide DNA methylome maps is increasing, however, still only a few are publicly available for human. Barring a disruptive technological advance, the need for DNA methylome maps to address fundamental biological questions will likely continue to far outpace the production of new maps for years. In contrast, many more lower-cost DNA methylomes of either lower resolution or lower coverage have been generated across diverse biological and disease states. For example, the NIH Roadmap Epigenomics Project's current release of the Human Epigenome Atlas [5] contains 8 WGBS datasets, 119 RRBS datasets, 49 MeDIP-seq and 45 MRE-seq datasets. These methods yield largely comparable results, but differ significantly in extent of genomic CpG coverage, resolution, quantitative accuracy, and cost, at least using current algorithms to interrogate the data (Bock et al. 2010; [30]). None of these existing methods provides single-CpG resolution, comprehensive genome-wide coverage, and a cost that is affordable for a typical laboratory, particularly when many samples are assayed. To address these

needs we describe here methylCRF, a novel Conditional Random Fields-based algorithm that integrates methylated DNA immunoprecipitation (MeDIP-seq) and methylation-sensitive restriction enzyme (MRE-seq) sequencing data to predict DNA methylation levels at single CpG resolution.

4.3 Results

4.3.1 Motivation for integrating MeDIP-seq and MRE-seq data

MeDIP-seq and MRE-seq provide complementary readouts of DNA methylation [53]. Their protocols are simple and differ in important ways from bisulfite treatment methods [30][53]. By using simple heuristics, the combination of these two methods gave promising results in identifying differentially methylated regions (DMRs) and intermediate or mono-allelic methylation [30]. Here we further explore the complementary nature of MeDIP-seq and MRE-seq. All genome-wide DNA methylation profiling methods have their own unique biases which can lead to errors in assessing methylation states. Observed genome-wide measurements (MeDIP-seq, MRE-seq, WGBS, etc) are derived from the actual methylation states of the sample, which is unknown or "hidden" from the investigators. These hidden methylation states are often inferred from the observed data, which are usually sequencing reads aligned to the reference genome. Because MeDIP-seq and MRE-seq are independent, complementary measurements of the same methylation state, our confidence in inferring the methylation state can be significantly increased when results from these two methods are integrated. For example, a decrease in MeDIP-seq signal could reflect a biological event (we infer that this region is unmethylated) or could be a methodological artifact; but if the inferred unmethylated state is corroborated by an increase of MRE-seq signal, then the inference of unmethylation is stronger. Thus, integrating MeDIP-seq and MRE-seq is expected to significantly improve our ability to predict methylation levels accurately.

While MeDIP-seq and MRE-seq data correlate to a certain degree with WGBS measurements (Fig. 1A, B), their relationship is not well represented by a simple linear translation. It has also remained technically challenging to infer absolute methylation levels from enrichment measurements alone [21, 63, 9]. Importantly, existing high-resolution methylomes and prior

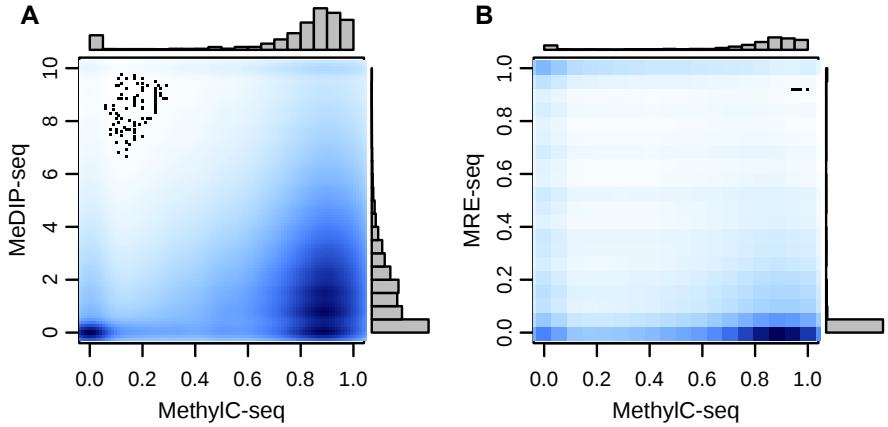


Figure 4.1: Relationship between MeDIP-seq and MRE-seq and MethylC-seq. (A) A kernelized density plot of per CpG MeDIP-seq normalized read count values as a function of MethylC-seq methylation levels shows a complex, approximately proportional relationship. (B) MRE-seq normalized read counts as a function of MethylC-seq methylation levels shows a complex, approximately inversely proportional relationship.

regional analyses reveal that CpG methylation levels are highly non-random throughout the genome [50]. The levels vary strongly with local CpG density, display distinct genomic feature-specific characteristics, and show strong correlation between neighboring CpG sites (Fig. 2A, B, C). These properties motivated us to use a formal statistical model to explore these complex relationships with the goal of making an accurate, comprehensive, high-resolution prediction of DNA methylation levels from MeDIP-seq and MRE-seq data.

4.3.2 Summary of the methylCRF algorithm

We chose a Conditional Random Field (CRF) model to integrate MeDIP-seq and MRE-seq data to predict genome-wide single CpG methylation levels. Like the Hidden Markov Model (HMM), which has been extensively adopted by the computational biology field[22], CRFs were initially developed for natural language processing [41] but their application in biological research has been limited [85]. However, CRFs have several distinct advantages when modeling data with complex inter-dependencies which is a common feature of biological data.

The primary advantage is due to that CRFs model the conditional probability of the variable of interest (CpG methylation states) given observed variables (e.g. MRE scores and MeDIP

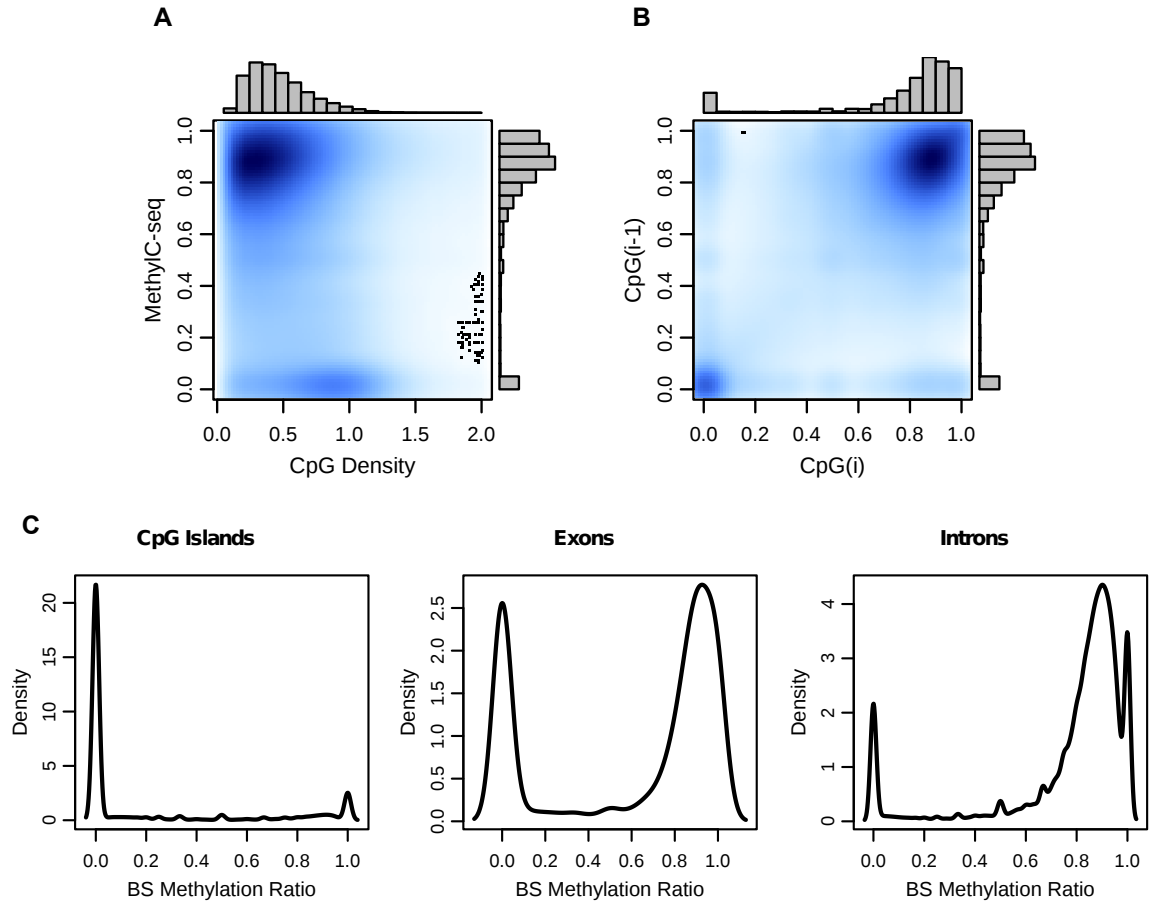
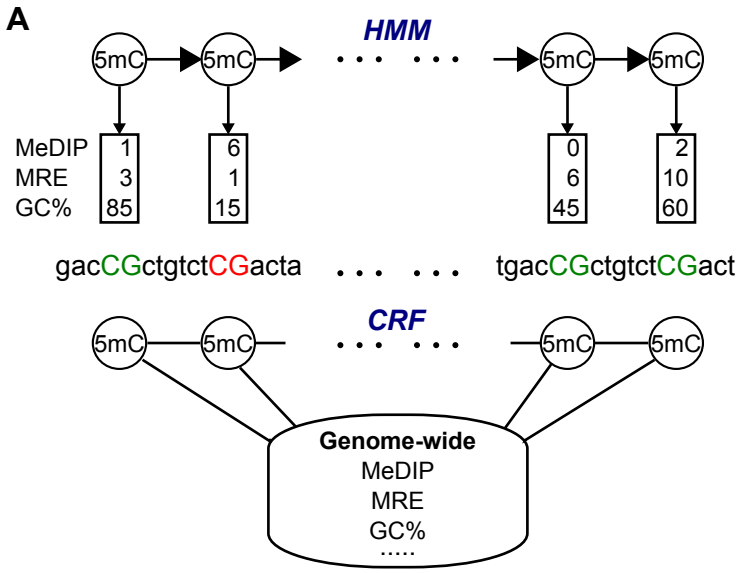


Figure 4.2: CpG methylation levels are non-random throughout the genome. (A) A kernelized density plot of MethylC-seq methylation levels as a function of CpG density. Methylation varies in a CpG density dependent manner with the majority of CpGs at 0-0.75 density with 75%-100% methylation and a smaller group at 0.75-1.25 density with almost 0% methylation. (B) CpG methylation levels as a function of their immediately 5' CpG methylation level (up to 750bp). (C) Distribution of MethylC-seq methylation levels at CpG islands, exons, and introns.

scores), whereas HMMs model the joint probability of all model variables (Fig. 3A). By conditioning on the observed variables, CRFs are not confounded by correlation between them. This then allows CRFs to efficiently model more complex relationships between CpG methylation levels and larger numbers of potentially correlated and distant observations. In contrast, modeling the full joint probability either requires modeling the inter-dependencies between the observations (for example, between MeDIP-seq read count and CpG density) or making the assumption that they are conditionally independent (Fig. 3B) âĂŞ which in this case, they are not [63]. Since very little is actually known about possible additional confounding factors in these assays, this construction gives us significant freedom in choosing what data to use in predicting methylation levels. Considering the number of features that may influence CpG methylation, we believe this is a critical benefit. This also allows the dependency of the methylation ratio of a CpG on any single observation or groups of observations anywhere in the genome to be accounted for with the model complexity growing only by feature number and not by distance of the dependency as would happen in a HMM. Long range-interactions are a common problem [19, 89]. Of note, a CpG's methylation ratio can depend on observations in both directions which would introduce loops in a HMM. Additionally, by not having to consider dependencies between observations, the addition of agglomerative and derivative features is trivial. In our case, this was extremely useful as we could define features incorporating large windows of MeDIP-seq and MRE-seq scores at each CpG without increasing the complexity of the inference and without considering their complicated dependence on individual MeDIP-seq and MRE-seq scores.

An additional, practical benefit of using CRFs is that the specification of the model is created by defining functions in a way that offers great flexibility. In a typical implementation, these functions are then instantiated with every combination of assignments of values for its parameter variables. However, one could instantiate with only one or a subset of the values a variable can take without requiring the addition of a full probability distribution over all of the values. In a large model, this can provide a significant reduction in model complexity. Feature functions can also overlap or use subset of variables of another. While subsetting does not add any expressiveness to the model, it provides an elegant and automated way to handle missing observations as well as to take advantage of a more detailed feature for some configurations of the variables while having a simpler representation for other configurations.



B

$$
\begin{aligned}
HMM &:= P(5mC)P(MeDIP, MRE, ... \mid 5mC) \\
&= P(5mC, MeDIP, MRE, ...) \\
&= P(5mC \mid MeDIP, MRE, ...)P(MeDIP, MRE, ...)
\end{aligned}
$$

$$
CRF := P(5mC \mid MeDIP, MRE, ...)
$$

Figure 4.3: CRF versus HMM. (A) In a HMM, the labels generate the observations, while in a CRF co-occurrences of the label and observations are associated. (B) HMMs model the joint probability and must consequently model the dependencies in the observations, while CRFs only model the dependencies of the label on the data.

Lastly, since we are only interested in predicting correct methylation levels given our experimental data and the experimental data is fixed at test time, we do not believe there is any sacrifice in power by not modeling the full joint probability distribution. Also note that methylation ratios can be interpreted as a maximum likelihood estimate of the probability of a particular CpG being methylated and as our results indicate this probabilistic interpretation appears effective.

Our complete CRF model, methylCRF, is described in detail in the Methods section. Features include MeDIP-seq and MRE-seq measurements covering individual CpGs, distance between neighboring CpGs, distribution of MRE sites, and genomic annotations including CpG islands, genes, repeats, and evolutionary conservation of DNA sequences. We also generated a variety of derived scores representing averaged experimental measurements in genomic windows of different sizes. We trained a separate CRF for each genomic feature, and for the final methylation estimates, we averaged the predictions for any CRF whose predictions overlapped. Training was performed using MethylC-seq [50] measured methylation levels in randomly chosen regions representing 20% of the genome Table 6.1. Methylation levels were predicted genome-wide, and performance was evaluated using CpGs that were not used for training.

4.3.3 High concordance between methylCRF and WGBS predictions

Using methylCRF, we predicted individual methylation levels of 28 million CpGs for Human H1 Embryonic Stem Cells (ESC) with combined MeDIP-seq and MRE-seq data. Our predictions are in high concordance with MethylC-seq predictions on the same H1 cells, with a genome- wide correlation of 0.77 (Fig. 4A). methylCRF recapitulates the bimodal distribution of methylation levels identified by MethylC-seq (Fig. 4A). Using a previously developed concordance measurement (defined as the percent of CpGs with a methylation proportion difference less than 0.1 or 0.25)[30], methylCRF and MethylC-seq are 91% concordant within a 0.25 difference (Fig. 4B). This high concordance is illustrated by a genome- browser comparison between methylCRF and MethylC-seq of a representative genomic locus (Fig. 4C).

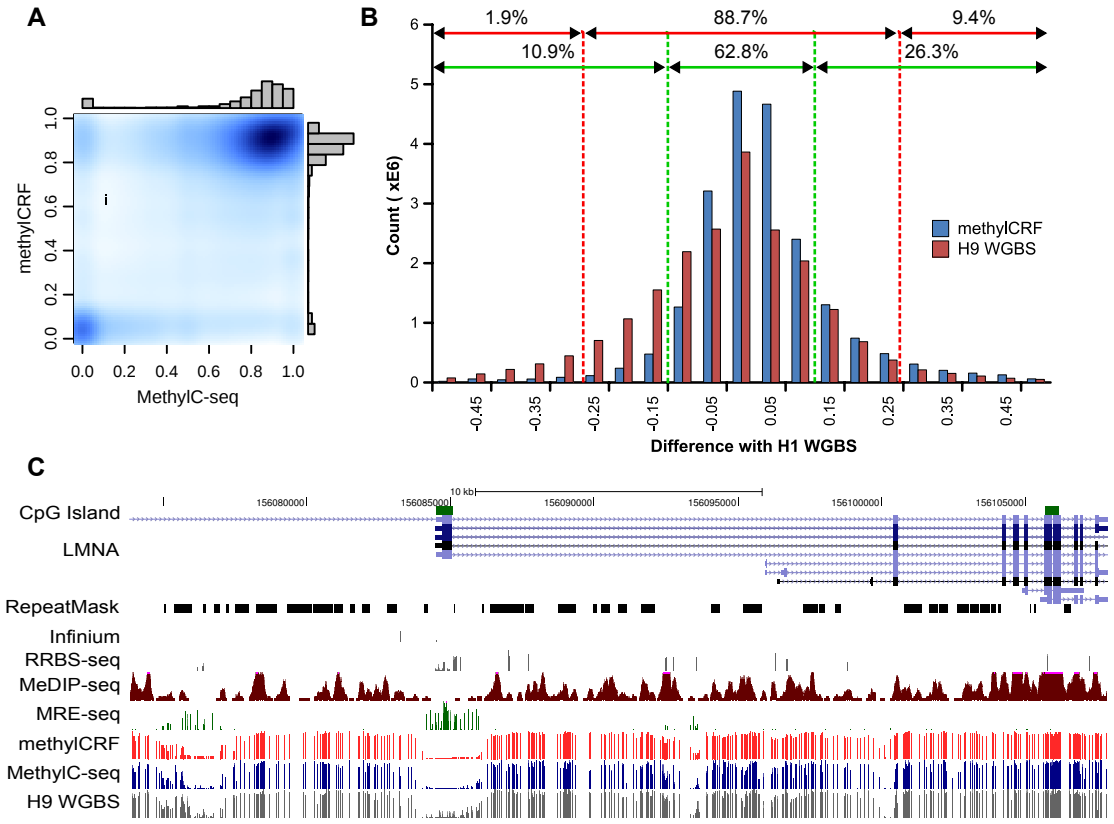


Figure 4.4: Concordance Between MethylC-seq and methylCRF. (A) Kernelized density plot comparing H1 ESC (male) MethylC-seq and methylCRF methylation levels at each CpG with at least 1 MethylC-seq read. MethylCRF recapitulates the bi-modal distribution of MethylC-seq. (B) The number of CpGs as a function of the difference between MethylC-seq and methylCRF methylation levels -the two agree within 25% for 91% of the CpGs and within 10% for 70% of the CpGs. The difference between BS-seq (H9 ESC, female) and MethylC-seq (H1 ESC) on common CpGs is also plotted for comparison. (C) Genome Browser view of per CpG methylation levels across a representative test region on chromosome 1.

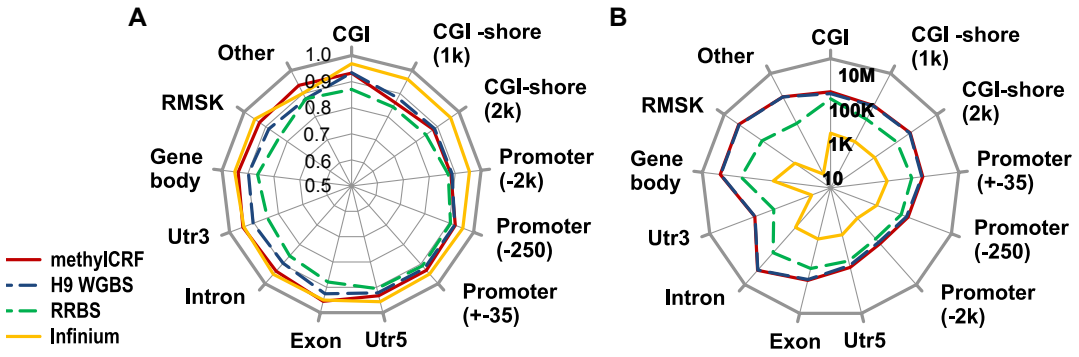


Figure 4.5: Comparison between MethylC-seq and methylCRF and other methylation assays. (A) Concordance of methylCRF, BS-seq, RRBS, and Infinium array with MethylC-seq within a 25% window broken out by annotated genomic features. Note that BS-seq (H9) is a female sample, while MethylC-seq (H1) is a male sample. Only CpG's in common were compared. (B) The number of CpGs used for each comparison on a log10 scale.

We next compared methylCRF and MethylC-seq on various genomic features (Fig. 5A). methylCRF and MethylC-seq agreed at an exceptionally high level for CpGs within CpG islands, promoters, 5' UTRs, and exons with 93%, 93%, 93%, and 96% respective concordances. The concordance decreased in Repeat-Masker annotated regions and regions with no annotation (Fig. 5A), possibly reflecting higher mapping errors in these regions, particularly for the reduced complexity reads from bisulfite conversion.

4.3.4 Benchmarking against other experimental methods

Several additional DNA methylation datasets exist for the H1 ESC line, including data obtained with RRBS and Infinium methylation array. In addition, a WGBS dataset was generated for the H9 human embryonic stem cell line (BS-seq) [46]. Data from this closely related ESC line provides the closest "biological replicate" of the MethylC-seq ESC H1 WGBS dataset.

When compared to MethylC-seq, methylCRF's performance is almost indistinguishable the comparison between MethylC-seq and BS-seq on these ESC cell lines (Fig. 4B, 4C, 5A). Specifically, within a 28% difference window, per-CpG methylation levels between the MethylC- seq (H1) and BS-seq (H9) are 90% concordant while methylCRF (H1) predictions reach the same concordance with a window of 23%. These windows decrease to 26% for H9

and 18% for methylCRF when we limit the comparison to CpGs with high MethylC-seq (H1) read coverage (e.g. >10 reads) and not in repetitive regions.

RRBS has comparable concordance levels to methylCRF when compared to MethylC-seq. The Infinium array data appears to have slightly higher concordance, which might be a result of having many fewer (28,000) CpGs for comparison and/or non-random selection of CpGs on the Infinium platform (Fig. 4C and 5A). The high concordance among these popular methods is consistent with previous comparisons ([7][30]. However, these methods clearly interrogate very different fractions of the DNA methylomes, as evidenced by the Genome Browser view (Fig. 4C) and CpG coverage comparison (Fig. 5B).

4.3.5 Robust performance across a variety of measurements

The strength of WGBS predictions is significantly influenced by sequencing coverage. Previous analyses suggest that the methylation level of individual CpGs can only be confidently estimated when sequencing depth is at least 10 [30]. Therefore, typically the minimum requirement for a WGBS experiment is to sequence the bisulfite converted genome to a depth of 30X [39]. However, even at this sequencing depth, a significant number of CpGs still are not covered by enough reads (Fig. 6A). Indeed, we observe increased concordance with increasing MethylC-seq coverage. For example, with a minimum 10-read coverage level the concordance within a 0.25 threshold window between methylCRF and MethylC-seq increased to 93% (from 91%, minimum 1-read coverage)(Fig. 6A).

CpG density is a major confounding factor in analyzing methyl-cytosine enrichment based methods [21, 63]. For example, inferring methylation levels in CpG-poor regions is thought to be highly inaccurate or impossible using MeDIP-seq [63]. Therefore, we examined methylCRF's performance across regions with differing CpG density and found the concordance between methylCRF and MethylC-seq does not vary significantly based on CpG density (Fig. 6B).

We also compared methylCRF with BATMAN [21], a popular method for analyzing MeDIP-seq data. Since BATMAN predicts methylation levels in windows of fixed-size and not

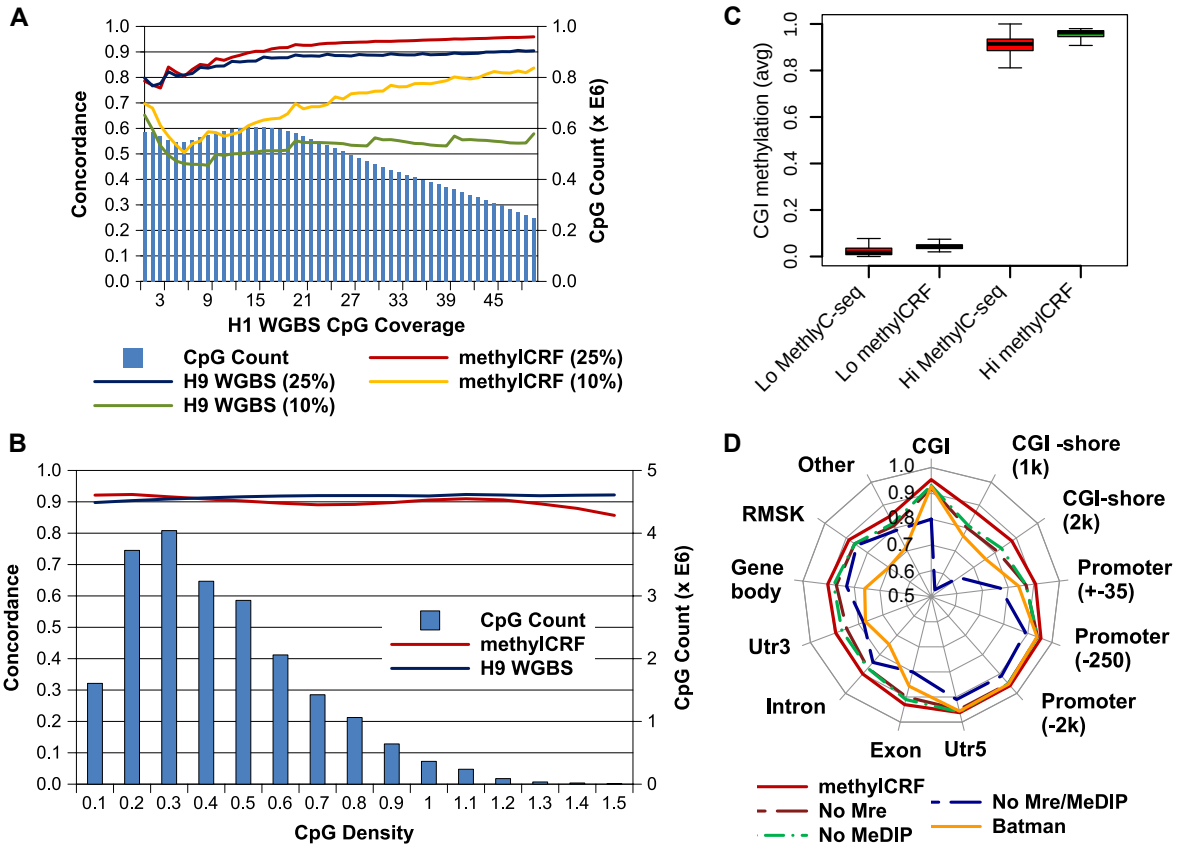


Figure 4.6: Factors Affecting Concordance Between MethylC-seq and methylCRF. (A) Concordance with MethylC-seq as a function of MethylC-seq read count (CpG coverage) at 10% and 25% windows for both methylCRF and BS-seq. The right y-axis (blue bars) indicates the number of CpGs with that coverage. (B) Concordance with MethylC-seq as a function of CpG density at 25% windows for both methylCRF and BS-seq. (C) Concordance of methylCRF within a 25% window broken out by annotated genomic features when only MeDIP-seq, MRE-seq, or genomic features are used. Concordance of BATMAN using MeDIP-seq is also plotted for comparison. (D) methylCRF accuracy on CGIs with high or low methylation (as defined by MethylC-seq). The Lo set of CGIs are those with an average CpG methylation ≤ 0.2, while the Hi set are those with an average methylation ≥ 0.8.

of single CpGs, we assigned each CpG the methylation level of its window. methylCRF consistently outperforms BATMAN in all categories (Fig. 6D).

Since our model learns separate CRFs for each genomic feature, we asked if it is possible that the high correlations between methylCRF and MethylC-seq could be explained by each CRF capturing the a priori methylation distributions of genomic features instead of using the experimental data. To examine this relationship, we applied methylCRF: 1) without MeDIP-seq data, 2) without MRE-seq data, and 3) with neither MeDIP-seq nor MRE-seq data, i.e. with only genomic features (Fig. 6C). The experimental data does indeed make a large difference in our predictions. Interestingly, MRE-seq alone performs slightly better than MeDIP-seq alone. This may be due to the ability of a CRF to incorporate a priori knowledge that most CpGs are methylated, thus making some MeDIP-seq information redundant. However, it is important to note that the combination of MeDIP-seq and MRE-seq improves performance significantly.

To further demonstrate that experimental data, but not the a priori methylation status of genomic features drives our prediction, we compared the rates of concordance for methylated and unmethylated CpG islands. Using MethylC-seq scores for H1 ES cells, we defined 17,189 unmethylated and 6,728 methylated CpG islands with an average methylation level of <=0.2 and >= 0.8 respectively. We compared these with the average methylCRF scores for H1 ES cells for each of these CpG islands (Fig. 6D). Clearly, methylCRF predicts similar sets of methylated and unmethylated CpG islands as MethylC-seq and it does both equally well. Furthermore, on a per-CpG level, unmethylated CpG islands concordance is 0.98, while methylated CpG islands concordance is 0.96. This analysis strongly suggests that while we take advantage of a priori information, like genomic features, the algorithm clearly integrates experimental data, relationships within the data, and between data and genomic features to make accurate predictions. We performed a similar analysis on the subset of CpG islands located in promoters - that is, a partitioning of CpG islands independent of the model of the CpG island-specific CRF - and obtained similar results (Supplementary Fig. S1). Similar results were also obtained when we restricted our analysis to intergenic CpG islands (Supplementary Fig. S2).

4.3.6 methylCRF accuracy is robust when applied to a second sample

Having demonstrated that methylCRF can accurately predict differential DNA methylation of CpGs independent of the characteristic methylation status of their genomic feature, we tested whether our model, trained on data from H1 embryonic stem cells, would generalize to data of other samples. This includes testing whether methylCRF can predict differential DNA methylation at a genomic locus between different samples. We reason that if methylCRF is completely dependent on genomic features, or is over-trained with ESC data, we would not be able to distinguish between datasets generated from other cell or tissue types.

We generated WGBS, Infinium HumanMethylation450 BeadChip, MeDIP-seq and MRE-seq data profiles of a human fetal neural stem cells (NSCs) culture (Hu-F-NSC-02, neurosphere cultured cells, ganglionic eminence derived, fetal age of 21 weeks) (Supplementary Table 1). We generated a single CpG resolution DNA methylome of this sample using methylCRF. We performed similar concordance analysis between predictions of methylCRF and that of WGBS, and between methylCRF and Infinium array. The overall concordance is consistently high for comparison of methylCRF against either WGBS or Infinium arrays (Fig. 7A). Specifically, methylCRF and WGBS were 88% concordance within a 0.25 difference window, and 65% concordance within a 0.10 difference window. Additionally, we defined differentially methylated CpG islands between the H1 ESC and the fetal NSCs using the WGBS data. Out of 26,845 CpG islands, we identified 233 that have significantly different methylation status between H1 ESC and fetal NSCs, such that their average methylation levels are less than 0.2 in one sample but greater than 0.8 in the other. These WGBS defined, cell type-specific differences in CpG island methylation were mirrored by similar differences between H1 ESC and fetal NSCs estimated by methylCRF (Fig. 7B), suggesting methylCRF can faithfully predict differential DNA methylation between two samples.

Finally, we evaluated the ability of methylCRF to predict intermediate methylation levels. We did not include imprinted control regions (ICRs) as a genomic feature in training. However, when we examined methylation status of known ICRs (obtained from `https://atlas.genetics.kcl.ac.uk`, and summarized in Supplementary Table 2), we found that majority of the ICRs exhibited intermediate methylation levels based on methylCRF prediction in both H1 ESC and fetal NSCs (Fig. 8A, B), and the levels were consistent with those

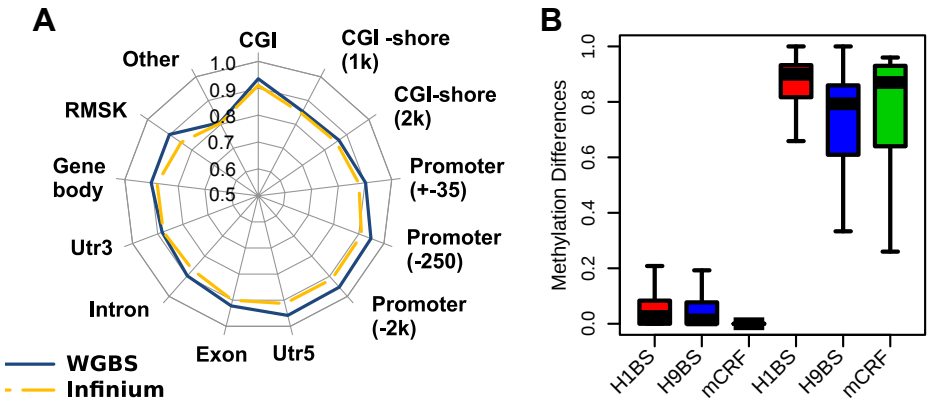


Figure 4.7: Applying methylCRF to fetal NSCs. (A) Concordances between methylCRF and WGBS data and between methylCRF and Infinium array broken out by annotated genomic features. (B) CpG islands were grouped as "indifferent" and "different" based on their methylation levels in H1 ESC and fetal NSC (Hu-F-NSC-02) assessed by WGBS data. Actual difference distributions were plotted between H1 ESC (WGBS, red), or H9 ESC (WGBS, blue), or H1 ESC (methylCRF, green) and fetal NSCs (WGBS).

determined by WGBS based methods (Fig. 8A, B). Genome Browser views of the data were provided for two exemplar ICRs (Fig. 8C, D).

4.3.7 Experimental validation

For regions where methylCRF and MethylC-seq results were discordant in H1 ESC, we experimentally validated methylation status by performing PCR amplification of bisulfite converted DNA, followed by Sanger sequencing of cloned amplicon DNA. Out of 12 regions that show disagreement, bisulfite validation agreed better with methylCRF in 11 cases, and agreed better with MethylC-seq in only 1 case. Two of the tested loci are shown in Figure 9, while the remaining sites are summarized in Table 6.1 and Supplementary Figure S3.

4.4 DISCUSSION

DNA methylation is an epigenetic mark that has important regulatory roles in a broad range of biological processes and diseases [34]. Understanding the role of DNA methylation in

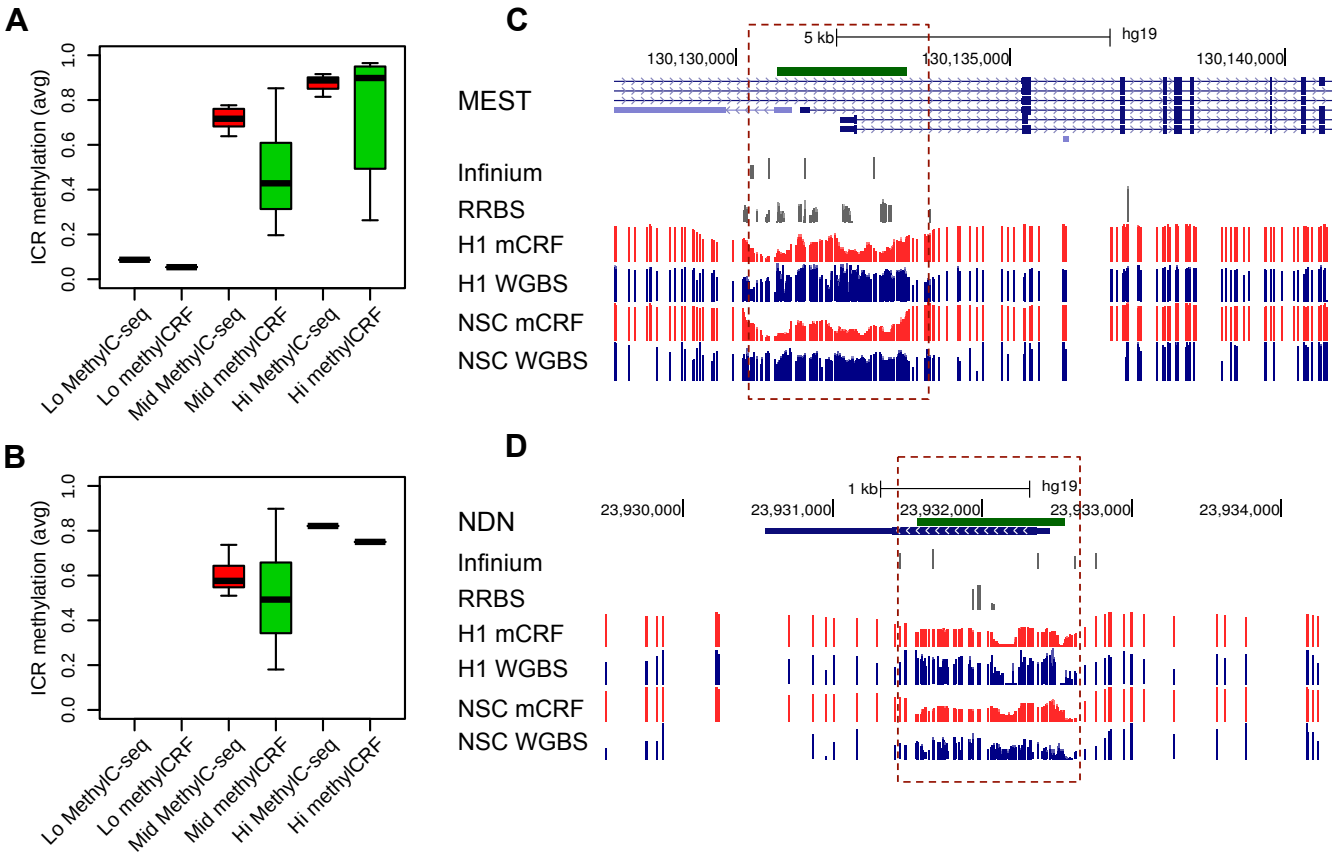


Figure 4.8: Comparing methylCRF and WGBS predictions in imprinted control regions. (A) Known imprinted control regions (ICRs, `https://atlas.genetics.kcl.ac.uk`, Supplementary Table 2) were grouped based on WGBS data (H1 ESC, MethylC-seq), as "Lo" (average methylation <=0.2), "Mid" (average methylation between 0.2 and 0.8), and "Hi" (average methylation >=0.8). Boxplots represent average methylation levels of these ICRs based on MethylC-seq and methylCRF. (B) Same as (A), except for fetal NSCs. (C) A genome browser view of ICR near gene MEST (mesoderm specific transcript, chr7). (D) A genome browser view of ICR near gene NDN (necdin, chr15).

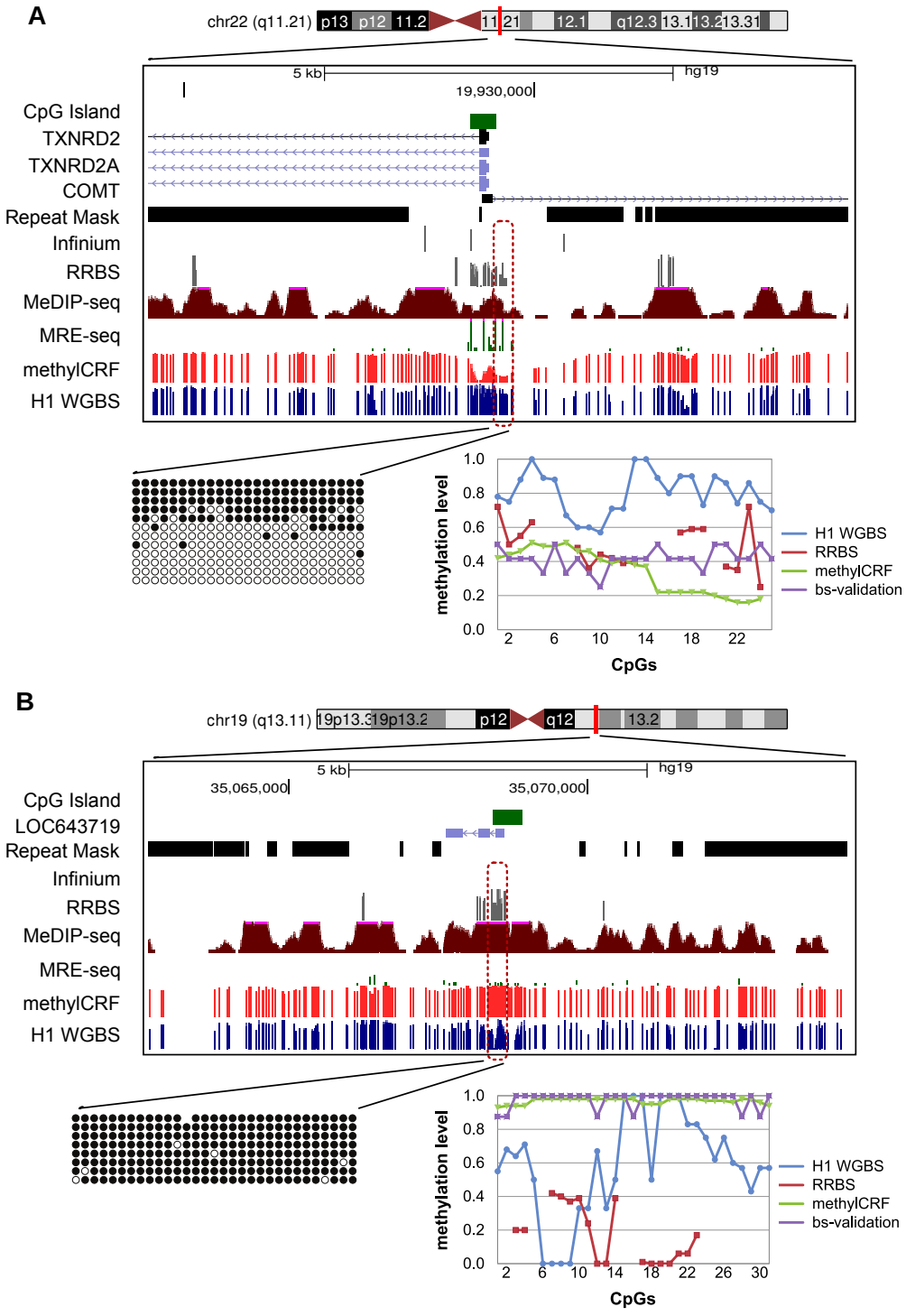


Figure 4.8: Experimental validation of regions where there is discordance between MethylC-seq and methylCRF. Genome browser view and site-specific bisulfite sequencing validation for each region (open circle: unmethylated CpG; filled circle: methylated CpG). The line graph shows the methylation levels estimated by MethylC-seq, RRBS, bisulfite validation, and methylCRF. (A) chr22:19929336-19929659; MethylC-seq predicted on average a methylation level of 80% methylated, while methylCRF and bisulfite validation agreed on a level of 40% methylated. (B) chr19: 35068305-35068683; MethylC-seq predicted on average a methylation level of 60% methylated, while methylCRF and bisulfite validation agreed that the region is more than 90% to completely methylated.

Fig. Panel	Chr	Strand	Tested Coordinates (hg19)	Fwd Primer	Rev Primer	MethylC-seq RMSE	methylCRF RMSE
S2A.	chr2	+	233216826-233217069	TTTTTTTAGAATTTAAATTTGGGTGAA	ATCCTACCTTAAATAAACACCTACC	0.28	0.12
S2B.	chr1	+	146551336-146551644	TTTTTTTTGGTTGAGGTTAGTTTAT	CCCAAACTCTAAATCAAAACTTTTT	0.16	0.13
S2C.	chr2	+	37571975-37572244	TAGTTTGGTTAGAGGAGAAGGTGAG	AACCCAAAAAAAAACCAATAACATC	0.52	0.06
S2D.	chr10	+	94820761-94821132	TTAGGAGTTAGGAAAAAGTTTTGAG	ACTAAACCAAACTAAACAACAAACC	0.44	0.11
S2E.	chr15	+	57025677-57025990	TGATTGGAGTTTTGAGGAGGA	CCCACATAAAAACAAAACCCTAAC	0.46	0.19
S2F.	chr1	+	200343036-200343274	GGAGGGGAAGAATATAAGAAATAATTAGT	TCTAAATCCCAATCCCTAACTACAA	0.06	0.82
S2G.	chr2	+	228736110-228736500	ATGTAGTTTAGGTTGTGGTTTAGGT	CAATCTAAAAACCCAAAATCCC	0.31	0.05
S2H.	chr4	+	103940626-103940995	TTAAGAATTTTATTGAATTGAGGGG	TAAACAAAAAACACACCAAACAATC	0.2	0.04
8A.	chr22	+	19929336-19929659	GTTTTTGGGGTAGTTAGGGTTGT	CTCAACTTTCCACAAAAAATCTAAAA	0.4	0.12
S2I.	chrX	+	48929750-48930067	GTAGGTAGGTTAATGGAGTGGTGAG	ACCAAAAAAAACAACCAAAACATACT	0.77	0.03
S2J.	chr13	+	58208240-58208505	TTTATATTTATGTGTTTGTGAATTTTA	ATACTCACCAAATAACCCAAACC	0.53	0.32
8B.	chr19	+	35068305-35068683	TTTTGGTTAGAAATTGGTTAATGAT	CTAAAATACCACAAACCCCACTAC	0.5	0.05
NA	chr16	+	2666554-2666783	GGAGGATTTAGTGTTATTTGTTT	TCTTAATTATTACTTAAAATTTTAAATACC	NA (no value)	0.24

Table 4.1: Summary of validation results. Validation of 12 targets chosen within windows where there is discordance between MethylC-seq and methylCRF for H1 ESC. The root-mean-squared error (RMSE) is shown between MethylC-seq or methylCRF and bisulfite validation using the listed primers.

development and disease requires knowledge of the distribution of these modifications in the genome. The technology is now available for studying DNA methylation genome-wide, at high resolution and in a large number of samples [6]. Previous comparisons suggest that many popular methods yield largely comparable results, but they differ significantly in extent of genomic CpG coverage, resolution, quantitative accuracy, and cost [7, 30], at least using current algorithms to interrogate the data.

We introduce a combined computational and experimental strategy to produce single CpG resolution DNA methylomes of all 28 million CpGs in the human genome at a fraction of the cost of whole genome bisulfite sequencing methods. Our computational model, methylCRF, is based on Conditional Random Fields, a model similar to the well-known Hidden Markov Model, initially developed for natural language processing but less applied in biomedicine. Using this model, we integrated data from two complementary DNA methylation assays (MeDIP-seq and MRE-seq) to predict methylation at single CpG resolution that were similar to the results from WGBS on the DNA of the same cell line. However, the cost of our two assays combined is less than 10% of a whole genome bisulfite sequencing methylome. We showed that methylation levels assessed by methylCRF from MeDIP-seq/MRE-seq data are indistinguishable from a biological replicate of whole genome bisulfite sequencing.

A complete genome-wide DNA methylome of a given sample will describe methylation levels of every CpG in the genome, approximately 28 million in humans. WGBS based method is considered the only approach capable of producing such single CpG resolution DNA methylomes. It is perhaps the most celebrated method in DNA methylomics to date and generally considered superior to enrichment-based methods [39]. One important reason that WGBS appears conceptually superior to enrichment-based methods is that transformation of sequencing results to direct estimates of methylation levels of individual CpGs is straightforward -once data is aligned to the reference genome, one can simply count converted and unconverted Cs to infer methylation levels. Although WGBS does not directly measure single CpG methylation levels of a sample, investigators can easily infer methylation levels based on experimental data (by sequencing alleles from multiple cells) derived from the true methylation states, i.e., observed counts of converted and unconverted Cs. Such intuitive heuristics makes WGBS seem straightforward.

Similarly, enrichment-based data are also derived from true methylation states. However, current analytical methods for enrichment-based data usually calculate enrichment scores that are indicative of regional DNA methylation levels corrected by local CpG distribution [21], [9], but do not predict single CpG methylation levels. Our novel algorithm closes this gap - we can predict single CpG methylation levels based on MeDIP-seq and MRE-seq data, two fundamentally different methods. The algorithm represents a fundamental advancement in statistical modeling over methods currently applied to enrichment-based methods.

There are still significant barriers to individual laboratories adopting WGBS as a routine assay, mainly the high production cost. Our method costs only a fraction of that of WGBS, yet can achieve comparable results. Importantly, the cost saving is scalable; any anticipated reduction in sequencing cost will reduce the cost of WGBS and our method in equal proportions. We performed saturation analysis of MeDIP and MRE and concluded that 30M MRE reads and 50M MeDIP reads are required to reach saturation for measuring a human DNA methylome (Supplementary Notes, Supplementary Fig. S4). This translates to 1-1.5X coverage of the human genome. For WGBS, the requirement is at least 20-30X coverage. This striking 20-fold difference in required coverage will remain unchanged across different next-generation sequencing platforms. To generate 30X coverage for a human sample is still expensive or even prohibitive for most labs. Very often investigators need to assay multiple samples to identify biologically interesting differences with reasonable statistical significance.

Additionally, bisulfite converted genomes have lower sequence complexity. This not only causes problems in library construction and cluster formation on a sequencing plate, but also more profoundly affects alignment of bisulfite reads to the genome, i.e. mapping. Several algorithms have been developed to improve mapping of WGBS data [13, 27, 38, 62, 88], but the problem remains not entirely solved. The confidence of mapping WGBS reads is generally lower than mapping standard, non-bisulfite-converted reads. Since CpG methylation calls are predicted based on aligned reads, how accuracy of methylation calls relates to mapping quality remains uncharacterized. For example, while the effect of biased (C doesn't match T) versus non-biased (C matches T) alignment has been analyzed [39], no one, to our knowledge, has examined the possibly more critical, alignment biases based on the number of Cs in a read in either type of alignment.

In contrast, MeDIP and MRE protocols produce standard sequencing reads for which existing statistics were designed. Although two libraries are constructed, the total cost (reagents and labor) is comparable to constructing one WGBS, according to published protocols `http://www.roadmapepigenomics.org/protocols`. Mapping of MeDIP and MRE reads uses standard mapping tools and is more accurate than mapping of reads from WGBS.

In our experimental validation, we examined 13 loci where MethylC-seq and methylCRF predictions do not agree in H1 ESC. We used bisulfite conversion, PCR, cloning and sequencing as our validation method because it is considered the gold-standard for targeted DNA methylation prediction, and we could exclude the possibility that differences are caused by bisulfite conversion. Nevertheless, in 11 out of 12 loci the gold-standard approach gave methylation levels that were closer to those from methylCRF predictions than from WGBS. Our interpretation of this result is that most errors made by WGBS might be due to misalignment, however a comprehensive analysis of WGBS mapping is needed to be certain. Alternatively, these differences may reflect true biological variation. This raises the profound question again âĂŞ how much of the WGBS predicted DNA methylome is actually incorrect due to challenges in mapping bisulfite converted reads? We are eager to explore this question in future studies.

When compared to RRBS and Illumina arrays, our method is obviously much more comprehensive. Our method provides 10 to 20 fold more coverage than RRBS or available methylation arrays. Investigators may want to use array-based assays when their target CpG sites are directly interrogated by the array. However, many regions of interest, for example, repeats or cryptic promoters, will not be assessed. There are also a number of examples where the specific CpG site interrogated by an array does not reflect the true methylation status of the genomic feature (e.g. a promoter) and may lead to false conclusions. Our method not only provides a comprehensive method for exploratory studies, but as the cost of WGBS drops sufficiently for exploratory analysis, the concomitant drop in cost of methylCRF application on MeDIP-seq and MRE-seq will provide investigators a platform to comprehensively address biological questions by comparing multiple samples, conditions, or variances genome-wide.

The accuracy of methylCRF was benchmarked against WGBS, RRBS, Infinium array, and locus specific-bisufite sequencing on H1 human ESCs. In addition, we determined that the high concordance between methylCRF and WGBS is consistent across most genomic feature sets

and across all CpG density levels. The power of the method stems from its integrative nature âĂŞ methylCRF is able to integrate a priori information about the expected methylation states of various types of genomic features, two complementary and independent experimental measurements of methylation states, and hidden relationships among neighboring CpG sites. Genomic sequences and features provide a default expectation of their methylation status. Indeed, CpG content of the genome reflects germ cell methylation states during the course of evolution [48] and has been used to estimate methylation levels directly [18]. This is reflected by the overall concordance of 0.66 when methylCRF makes predictions based on genomic features alone, which represents an expectation of methylation of a majority of CpGs in a normal somatic cell. The concordance is significantly improved when either MeDIP-seq or MRE-seq data are integrated, and the highest concordance is obtained when the datasets are combined. Importantly, methylation predictions made by methylCRF are conditioned on both genomic features and experimental data, and not driven by genomic features alone. This is supported by the accurate separation of methylated CpG islands from unmethylated ones (Fig. 6D), even when focusing on promoter regions and/or intergenic CpG islands (Supplementary Fig. 1, 2).

The accuracy of methylCRF was further benchmarked on WGBS and Infinium array data from a second sample. Here methylCRF trained on H1 ESC data was applied to MeDIP-seq and MRE-seq of a fetal brain NSCs sample. The concordances between methylCRF and WGBS, and between methylCRF and Infinium array were at similarly high levels as those obtained on analyzing H1 ESC data. Moreover, methylCRF can reliably identify differentially methylated regions between the two samples. This strongly suggests that our model trained on ESC data can be applied to data of other samples.

In the current implementation of methylCRF we only consider CpG methylation and assume all signals obtained from MeDIP-seq and MRE-seq are results of CpG methylation. We also assume WGBS produced methylation signal and ignore complications caused by hydroxymethylation. We note that methylations of cytosines in the context of other than CpG (i.e., CHG and CHH) are rare in somatic cells but are indeed present in embryonic stem cells, usually in low levels and are associated with highly methylated CpGs [50]. Biological significance of CHG and CHH methylation in mammalian cells is yet to be determined. Our statistical model is general enough to incorporate non-CpG cytosine methylation, but we

focused on CpG methylation in this study. Our statistical model is also general enough to incorporate data on hydroxymethylation when they become more and more available.

Our study has several limitations. Because the cell line we used to train, H1 ESC, is male it is possible that the additional X chromosomes in female samples may not be as accurate. This is because males will only have one allele aligning the reference X chromosome whereas a female will have two, resulting in twice as many reads. In fact, this may also effect WGBS accuracy. The concordance within 0.25 difference window between H1 and H9 on chrX alone drops to 81% whereas excluding chrX raises the concordance 92%. The concordance of methylCRF with H9 on chrX is 79% whereas without it is 90%. Note that this also provides a natural experiment that suggests how methylCRF will perform in cases of large scale genomic aberrations such as segmental or even chromosomal duplication or deletion, which is frequently found in cancer. Both WGBS and methylCRF seem to be proportionally less accurate when alleles and possibly segments are added or deleted. Nevertheless, once more WBGS data becomes available, we can trivially extend methylCRF to include a field for structural variations which could be estimated by standard means. Additionally, since MeDIP-seq and MRE-seq are sequencing based, we can make use of existing tools to add SNP-based features and to include input (un-enriched sequences). We note that WGBS is at a disadvantage when considering SNPs, in particular C->T SNP's will either be reported as an unmethylated C unless strand- specific alignment is available or even worse will align to cause false-positive alignments in other locations.

Another potential limitation is that the H1 WGBS (MethylC-seq) and MeDIP-seq and MRE-seq were performed on separate passages of the H1 ESC line and assayed in separate labs. This may explain why methylCRF was consistently validated by the bisulfate cloning method over WGBS. However, if this were true, two important inferences can be drawn. One, notable changes in methylation can be seen even between passage numbers. Two, these validations show methylCRF's sensitivity in detecting DMRs based on experimental data - even in very similar biological contexts. Additionally, the accuracy on fetal NSCs is slightly lower than on H1 ESC's. This may suggest that H1 ESC's may have differences in their global methylation than other cell types. While this does not stop methylCRF from detecting tissue-specific DMR's, it suggests that the accuracy may improve further if we re-train methylCRF simultaneously on WGBS from multiple cell-types.

Finally, because of our use of genomic feature-specific predictions, methylCRF accuracy may suffer when some part of the methylation machinery breaks down or behaves differently, for instance as in some cancers. We have indirectly tested this in a sample case, however. Fig. 2C shows CpG islands to have an extremely biased distribution of low methylation. However, Fig. 6D shows equivalent accuracy in CpG islands that represent a genomic feature with, in the statistical view of methylCRF, aberrant methylation.

Despite the promise of WGBS based methods, the number of publicly available, complete, single CpG resolution DNA methylomes is still small in contrast to the number of lower resolution and/or lower coverage DNA methylomes generated by less expensive methods (e.g., MeDIP-seq and MRE-seq generated by individual labs and by the Roadmap Epigenomics project). Our method can convert these data into single base resolution, complete DNA methylomes, thus significantly increase the value of such existing datasets.

In summary, our results suggest that methylCRF is an effective statistical framework capable of integrating two fundamentally different sequencing-based DNA methylation assays, MeDIP-seq and MRE-seq, to predict genome-wide, single CpG resolution methylome maps. The concordance of our methylCRF predictions with WGBS falls within the range of concordance between two WGBS experiments on similar cells. methylCRF will thus significantly increase the value of high-coverage DNA methylomes produced using much less expensive methods, and provide a general statistical framework for integrating contributions from various types of DNA methylation data regardless of their coverage, resolution, and nature of their readout.

4.5 Methods

4.5.1 methylCRF implementation

methylCRF is implemented using the theoretical framework of conditional random fields [41]. This general framework expresses the conditional probability $Pr(Y|X)$ of a series of hidden states, the random variables, Y, given observed data X:

$$P(Y|X) = \frac{1}{Z} \prod_{c=1}^{C} e^{\sum_{k=1}^{K} w_k * f_k(c, y_{c-1}, y_c, X)}$$

Y is the methylation level of every CpG; X is the observations (MeDIP-seq, MRE-seq, genomic context). The C CpGs are indexed by c, and the K feature functions, f, are indexed by k. The weights, $w1..wk$, are learned via gradient ascent of the log likelihood. Z is the partition function which provides the global normalization and is the sum of all sequences of methylation levels given X:

$$Z = \sum_{y \in Y} \prod_{c=1}^{C} e^{\sum_{k=1}^{K} w_k * f_k(c, y_{c-1}, y_c, X)}$$

Our approach to data features was to initially include anything that we thought might in some way effect methylation. We then let an L1 normalization term during training determine which features were not important by pushing their weights to 0. Therefore, the choice of important features was learned from the data. We split the data into different ranges of effect. We included MeDIP and MRE scores both at the CpG (D0 and M0) and within windows of 20bp, 200bp, 2kbp, and 20kbp windows. We included whether a CpG was at an MRE restriction site (ER) and the distance in bp to the previous CpG (PC). From UCSC genome browser tracks [36] we included a 46-way mammalian phastCons conservation score. We included GC% in 20bp, 150bp, and 500bp windows, and CpG density in a 150bp window.

We defined one CRF feature for each one of these data features combined with the methylation level of the current and previous, 5‘, CpG. We also added CRF features for the next two MeDIP and MRE scores on both the 5‘ and 3‘ sides. We then defined compound features. We included a feature combining D0 and PC to possibly address the nonlinear relationship between MeDIP and CpG density. We also included one large feature including factors that appeared to be interacting (data not shown) including both the current and previous CpG methylation as well as D0, M0, and PC for the current CpG as well as the two CpGs to the upstream and downstream of the current. This feature also included MeDIP and MRE in 20bp and 2kbp windows, ER, and GC in 20bp and 150bp windows for the current CpG as well as ER for the surrounding two CpG's. We additionally included four more features

consisting of subsets of these as a fallback for rare combinations of values. A diagram of the complete model is illustrated in Supplementary Fig. S5.

The distributions of methylation levels are genomic feature-specific (Fig. 2C), so we reasoned that the methylation level transitions between neighboring CpGs are also genomic feature-specific. To address this, we trained a separate CRF for each genomic feature: one for each of the genomic annotations in RefGene (5‘ UTR, gene body, exon, intron, 3‘ UTR, CGI), for the derived types (distal promoter, TSS-2kb; proximal promoter, TSS-250bp; core promoter, TSS +/- 35bp; 1kbp flanking each CpG Island; and 2kbp flanking each CpG Island), one for each Repeat class (DNA, LINE, LTR, RNA, SINE, low complexity sequence and simple repeats, and other), and one for the remainder of the genome not covered by any of the previous CRFs.

Training was performed using MethylC-seq [50] measured methylation levels in randomly chosen regions representing 20% of the genome. We used only CpGs with at least 10-read coverage. We performed separate discretization for each CRF. Each of the CRFs were trained using crfsgd [8] using default settings.

CRFs are typically discriminately trained by iteratively ascending their gradient. While the function is convex and so converges to a global maximum, the whole CRF must be evaluated once for every iteration in the ascent which poses performance issues. However, CRFs have been shown to handily model millions of features [75]. Additionally, the ascent can be performed online providing two benefits: 1) potentially less over-fitting due to the less optimal solution, and 2) speed of analysis [8]. Being discriminately trained, though, CRFs need to be handled carefully so as to ensure their generalizability to future data.

For CpG's that are annotated with multiple features, we combine the methylation predictions by averaging the predictions of the corresponding CRF's and giving each CRF an equal vote. Performance was evaluated using CpGs that were not used for training.

4.5.2 Discretization Heuristic

CRFs are rooted in the Natural Language Processing (NLP) community and so model discrete rather than continuous variables. There has been at least one paper extending CRFs to rankings which requires developing a continuous CRF [67]. Additionally, in the derivation of CRFs there is no restriction on the form of the random variables, so continuous predictors are also an option. However, the theoretical and practical work on modeling and training of discrete CRFs is very extensive. Additionally, the relationship between the predictors and the methylation ratios are complex. Finally, Naive Bayes is known to perform better with discretized variables [20]. We, therefore, decided to follow in this line of work by representing the relationship between the predictors and methylation ratios as piece- wise constant. The trade-off in avoiding the choice of the correct family of continuous distributions for the methylation ratios as well as the predictors in the continuous case, is that we must determine where to cut the range of a predictor into pieces. This is equivalent to discretization for which considerable work has been done.

While equal range or equal size discretization is straightforward, they did not perform very well (data not shown). We instead chose to use supervised discretization using the methylation ratios to guide the discretization. While good entropy-based methods do exist, they would require the methylation ratios to already be discretized posing a chicken and the egg problem for which we could not find an existing solution. This lead us to develop a 2-step heuristic consisting of clustering to discretize the methylation ratios followed by supervised discretization of each predictor. In the first step, which we term "order-preserving clustering", we cluster the predictors and the methylation ratios together. We iteratively up-weight the methylation ratios and re-cluster until there is an order between clusters such that all of the methylation values in one cluster are larger than the previous (Supplementary Fig. 6). Given this as a data model- driven discretization of the methylation ratios, we then use supervised discretization on each predictor individually. Note that this heuristic can use any pair-wise distance metric, clustering method, or supervised discretization method. We used k-means [51] and Euclidean distance for clustering and CAIM [40] for discretization.

4.5.3 MeDIP-seq, MRE-seq, and WGBS data

All data were obtained from the NIH Roadmap Epigenomics Mapping Centers' repository for human reference epigenome atlas [5]. Experiments were performed under the guidelines of the Roadmap Epigenomics project `http://www.roadmapepigenomics.org/protocols`. Specifically, MeDIP-seq and MRE-seq experiments were performed as described previously [53]. All data have been previously submitted to NCBI, and are listed in Supplementary Table 1.

The reads were aligned with bowtie [45] to HG19. The MRE reads were normalized to account for differences in enzyme efficiency and scoring consisted of tabulating reads with ends at each CpG [53]. To allow for comparison between experiments, the CpG read counts for MeDIP were scaled so that the 75th percentile of CpGs with at least one read is 10. Since for each MeDIP read, the CpG that was bound by the antibody cannot be determined, a fractional count was added to each CpG for the read. The final MeDIP score is the sum of CpG scores within the specified window.

4.5.4 Genomic features

RepeatMasker annotations, CpG islands, genomic super duplications, 46-way phastCons, and refGene coding loci features were all downloaded from UCSC Genome Browser [36]. The GC percent, CpG density, and MRE sites were calculated using HG19.

4.5.5 Training and prediction

For training, we randomly selected both the location and size of genomic fragments of 75kb to 750kb in length comprising roughly 20% of the genome. We used only CpGs with at least 10 BS reads. We performed separate discretization for each CRF. For the k-means step in the discretization of BS methylation values, we arbitrarily chose 10 clusters as this seemed a reasonable cutoff for the granularity for the measurement of CpG methylation that we would be interested in. We used Euclidean distance as our metric. Further details on our discretization method are discussed above. Each of the CRFs were trained using crfsgd

[8] using default settings. The data for each CRF was split on gaps of greater than 750bp between consecutive CpG's.

For prediction, the data was created as for the training data. For the final methylCRF predictions, we combined the predicted methylation levels of all the CRFs by averaging the predictions for CpGs that were shared by multiple CRFs. Genome Browser tracks are available as part of Roadmap Epigenomics Project's data visualization hub: `http://VizHub.wustl.edu`

4.5.6 Bisulfite treatment and library construction for WGBS

1 to 5ug gDNA was sonicated to an approximate size range of 200-400bp. Size selection is performed on a PAGE gel to obtain DNA fragments of 200-300bp. DNA are quantified by fluorescent incorporation (Qubit, Invitrogen). The library preparation includes end-repair and phosphorylation with NEBNextTM or Illumina Sample Prep Kit reagents, and addition of an âĂŸA' base to the 3' end of the DNA fragments. Methylated adapters are ligated and size selection is performed to remove excess free adaptors. The ligated DNA is quantified by Qubit, and approximately 100ng DNA is used for bisulfite conversion. Methylated-adaptor ligated to unmethylated lambda-phage DNA (NEB) is used as internal control for assessing rate of bisulfite conversion. The ratio of target library to Lambda is 1600:1. Bisulfite conversion of the methylated adapter-ligated DNA fragments follows the FFPE Tissue Samples Protocol from Qiagen's Epitect Bisulfite Kit. Cleanup of the bisulfite converted DNA is performed, and a 2nd round of conversion is applied. Enrichment of adaptor-ligated DNA fragments is accomplished by dividing the template into 5 aliquots followed by 8 cycles PCR with adaptor primers. Post PCR size-selection of the PCR products from the 5 reactions is performed on a PAGE gel. Following 100bp paired-end sequencing on a HiSeq2000, sequence reads were aligned and processed through the Bismark pipeline.

4.5.7 Infinium assay

Bisulfite conversion was performed on 1 ug of genomic DNA using the EZ DNA methylation kit (Zymo research) as per the manufacturer's alternative incubation conditions protocol.

The bisulfite converted DNA was amplified and hybridized to an Infinium HumanMethylation450 beadchip (Illumina) following the Infinium HD methylation assay protocol at the UCSF Genomics Core facility. Methylation levels (beta values) were determined using the Methylation Module of the Illumina GenomeStudio software.

4.5.8 Bisulfite validation

Total genomic DNA underwent bisulfite conversion following an established protocol10 with the following modifications: incubation at 95 °C for 1 min, 50 °C for 59 min for a total of 16 cycles. Regions of interest were amplified with PCR primers Table 6.1 and subsequently cloned using pCR2.1/TOPO (Invitrogen). Individual bacterial colonies were subjected to PCR using vector-specific primers and sequenced (Quintara biosciences). The data were analyzed with online software BISMA [70]. Results are summarized in Table 6.1 and Supplementary Fig. 3.

4.6 SOFTWARE AVAILABILITY

methylCRF is complete open source software. The source code, parameter sets, genomic data sets, as well as instructions are available at: `http://methylcrf.wustl.edu`

4.7 ACKNOWLEDGEMENT

We thank collaborators in the Reference Epigenome Mapping Centers (REMC), Epigenome Data Analysis and Coordination Center and NCBI who have generated and processed data used in this project. We acknowledge support from the NIH Roadmap Epigenomics Program, sponsored by the National Institute on Drug Abuse (NIDA) and the National Institute of Environmental Health Sciences (NIEHS). J.B.C is supported by a Career Development Award from the Dermatology Foundation. J.F.C., M.H. and T.W. are supported by NIH grant 5U01ES017154. T.W. is supported in part by the March of Dimes Foundation, the Edward Jr. Mallinckrodt Foundation, P50CA134254 and a generous start up package from

Department of Genetics, Washington University. K.L.L. and C.L.M. are supported by NIH Grants P01CA095616 and P01CA142536 and a grant from the Sontag Foundation. T.W. thanks David Haussler for insightful discussions during the early development of the project.

4.8 Supplemental

1. Additional comparison of MethylC-seq and MethylCRF predictions of differentially methylated CpGs with the same genomic feature.

   We reason that if methylCRF relies more on a priori information of genomic features than on experimental data, then it would predict that CpGs with the same genomic features have the same methylation level. In the main text, we examined methylated and unmethylated CpG islands. Here, we examined only CpG islands that are located in promoter regions, using both the general CRF model and the CRF model trained on CpG islands alone. These promoter CGIs, 421 methylated and 11,418 unmethylated, represent a subset of CGIs with an a priori distribution distinct from the set of all CGIs (i.e. that of 6,728 methylated and 17,189 unmethylated), Fig. S1. However, methylCRFs accuracy on this set is equal to that of all CGIs (Fig. 6D). Furthermore, the predictions from the CGI CRF (i.e. the component CRF that predicts just CGI methylation - which has no information to distinguish promoter CGIs from the set of all CGIs), is equally accurate on methylated CGIs, Fig. S2

2. Genome browser views, per-CpG methylation, and line graph of locations validated by location-specific bisulfite sequencing, Fig. S3 A-J. Two additional locations are presented in the main text.

3. MeDIP-seq/MRE-seq Saturation Analysis

   Both MeDIP and MRE assays approach saturation with less than 50M reads. We plotted number of CpGs measured by MeDIP or MRE against number of reads. Specifically, assuming 40 million mappable reads as the total reads for a MeDIP library, 90% of the total reads can interrogate 98% total CpG sites; similarly, assuming 30 million mappable reads as the total reads for a MRE library, 90% of the total reads can interrogate 97% total CpG sites (Fig. S4.A). When combined, a saturated MeDIP library and a saturated MRE library are equivalent to 1 1.5X coverage of the human genome. For

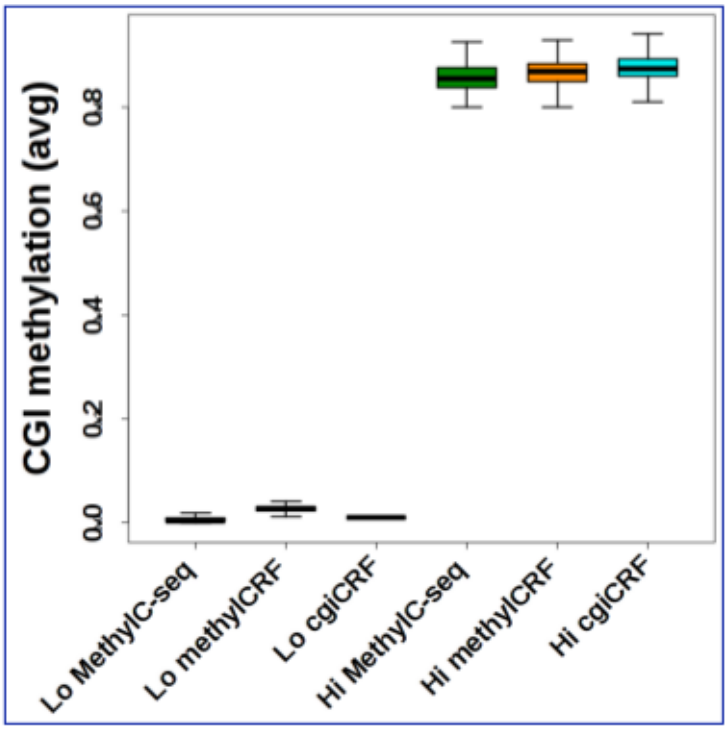


Figure S1: methylCRF accuracy on methylated (avg MethylC-seq score >= 0.8 and unmethylated (avg MethylC-seq <= 0.2) promoter CpG Islands (CGIs).

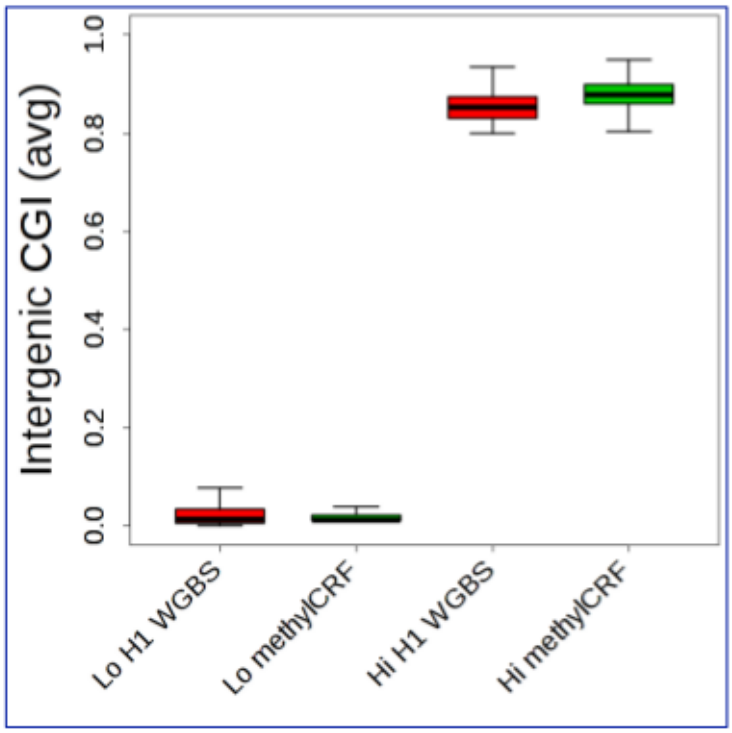
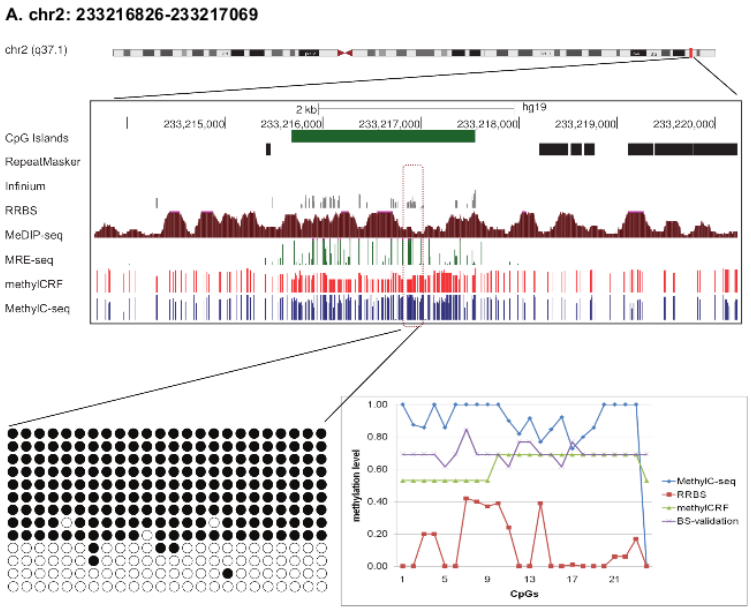
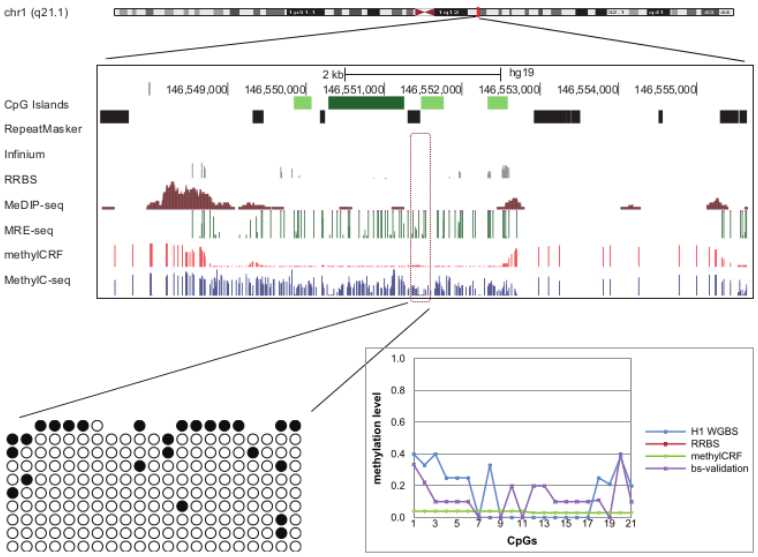


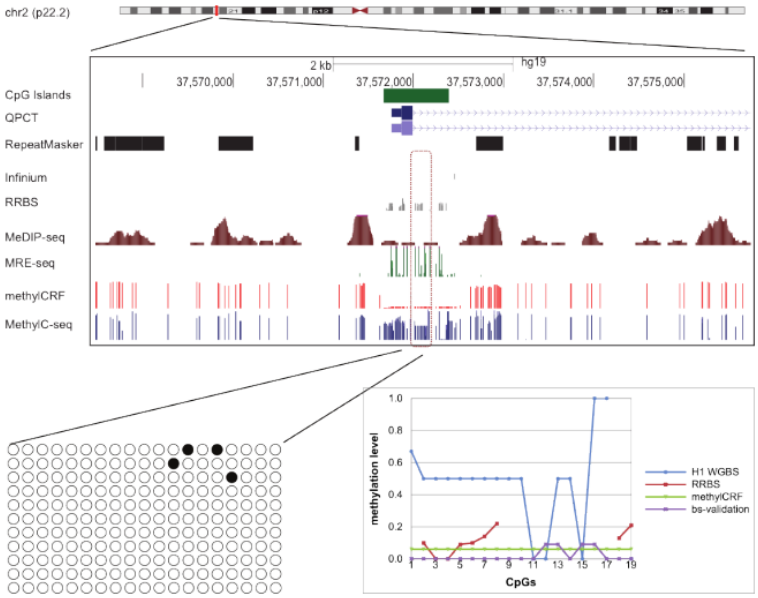
Figure S2: methylCRF accuracy on methylated (avg MethylC-seq score >= 0.8 and unmethylated (avg MethylC-seq <=0.2) intergenic CpG islands (CGIs)



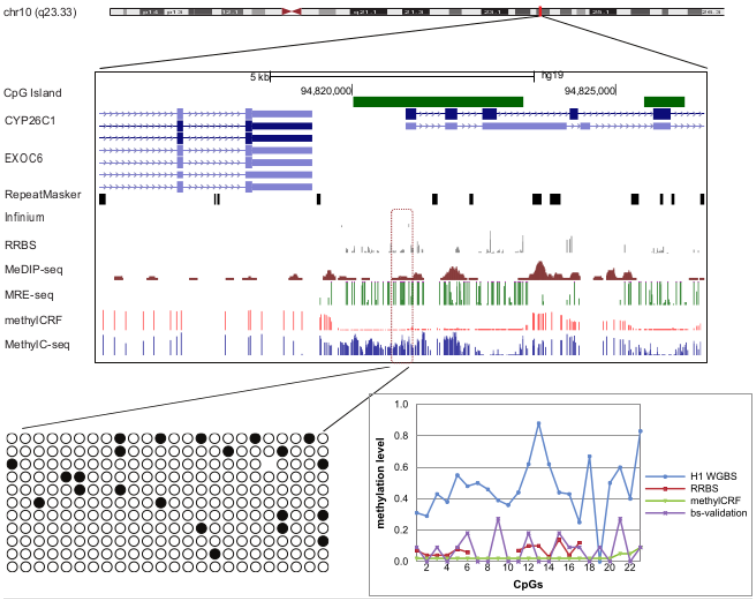
B. chr1: 146551336-146551644



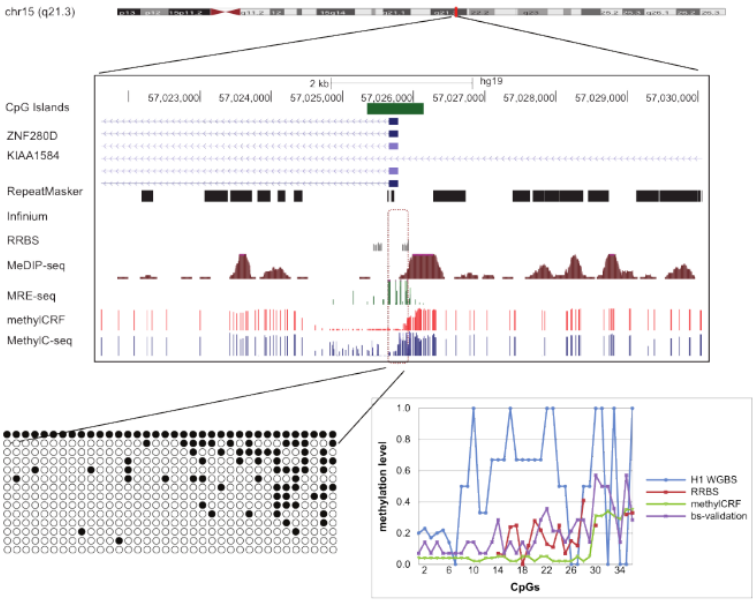
C. chr2: 37571975-37572244



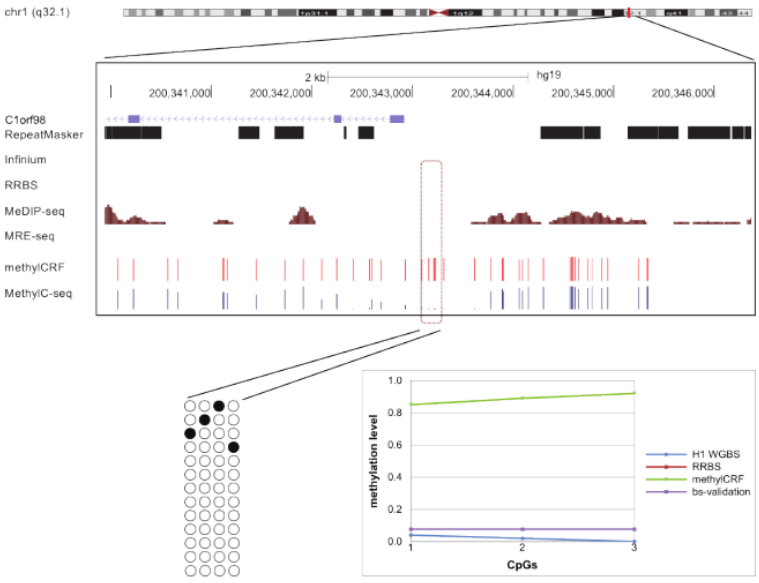
D. chr10: 94820761-94821132



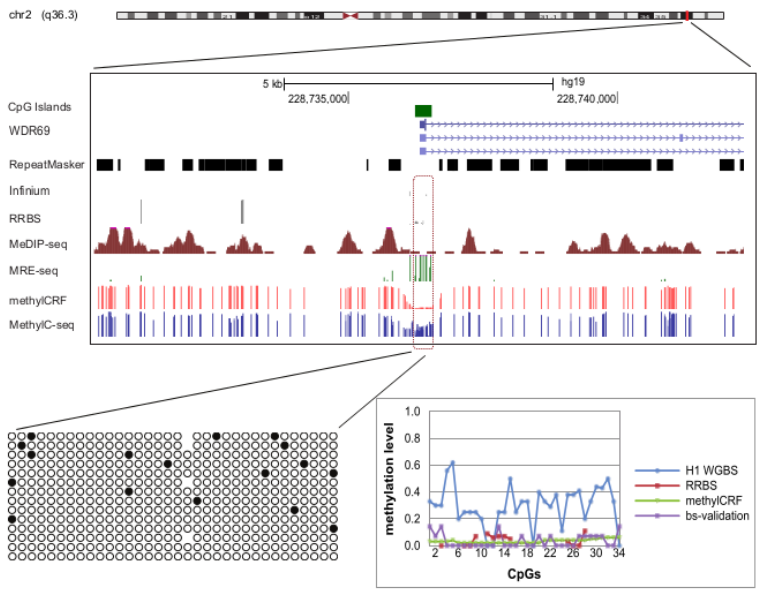
E. chr15: 57025677-57025990



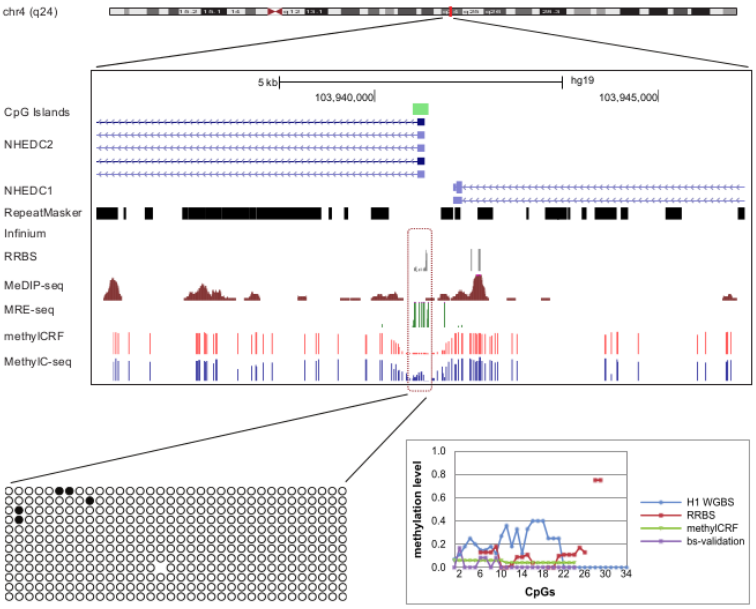
F. chr1: 200343036-200343274



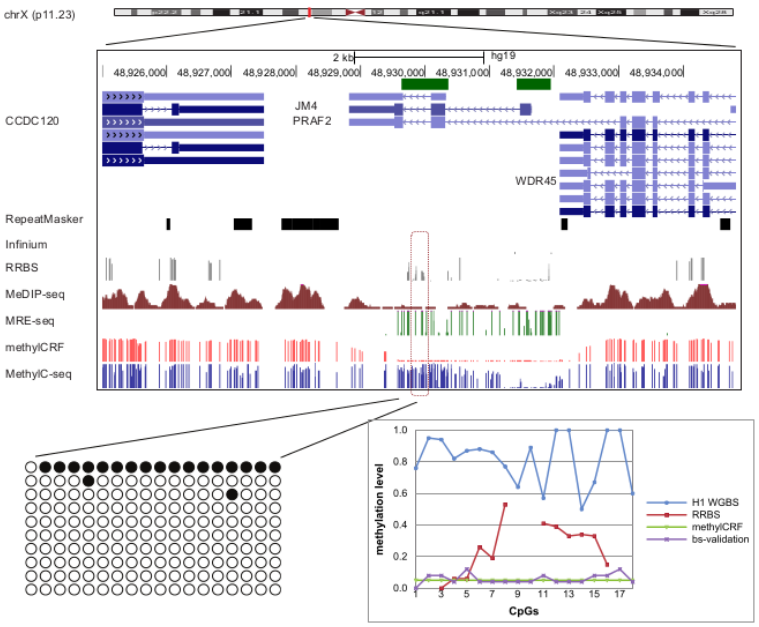
G. chr2: 228736110-228736500



H. chr4: 103940626-103940995



I. chrX: 48929750-48930067



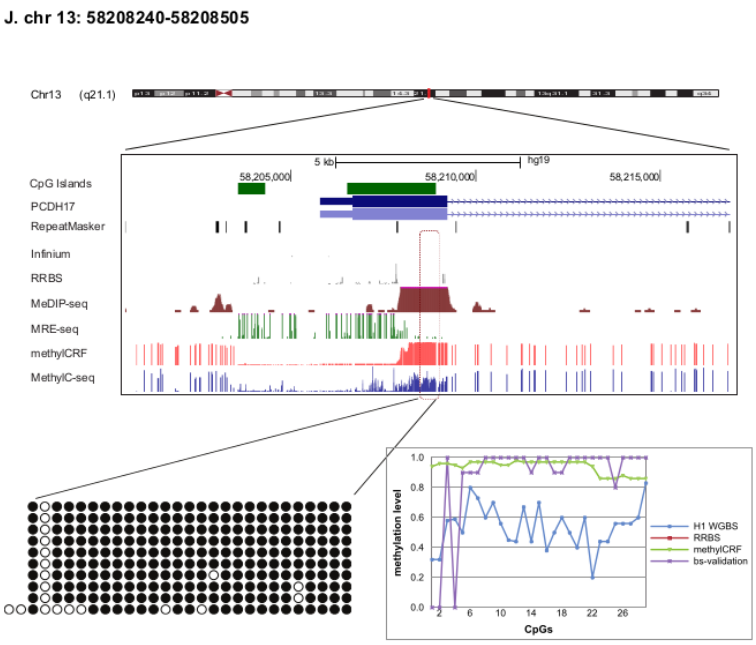


Figure S3: Additional validation results, A-J

WGBS the standard requirement is 30X coverage. The striking ratio of more than 20 fold difference will remain unchanged across different next-gen sequencing platforms.

In order to determine the depth of MeDIP and MRE required for given accuracies we ran methylCRF on randomly chosen subsets of reads (Fig. S4.B). We ran three types of tests, 1) reducing both MeDIP and MRE, 2) reducing MeDIP only, and 3) reducing MRE only. The model is relatively robust and works well with a range of coverages. Using only 40% of both MeDIP and MRE reads still gives an overall correlation close to using 100%, 0.82 vs. 0.85, while reducing either MeDIP or MRE to 40% gives a correlation around 0.83. Interestingly, using only 60% of the MeDIP reads gave better results than using over 80%.

4. Complete data model.

5. Order Preserving Clustering Heuristic.

   For multivariate clustering, the order preserving heuristic iteratively increases the scale of a subset of variables, in this case the CpG methylation estimate (labeled Y), until the clustering algorithm outputs a clustering in which the CpG methylation estimates within each cluster are contiguous. (A) a two- dimensional example, where there are

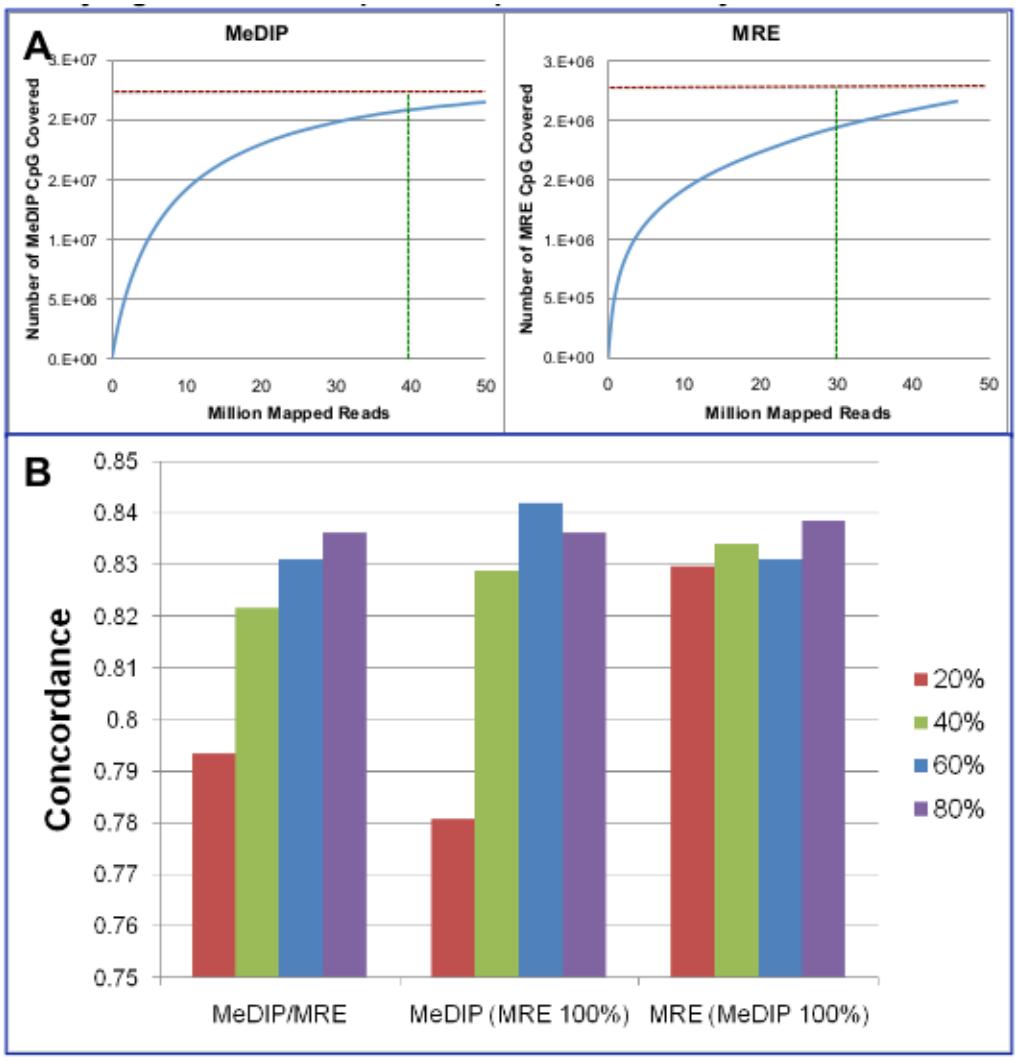


Figure S4: MeDIP-seq/MRE-seq saturation analysis

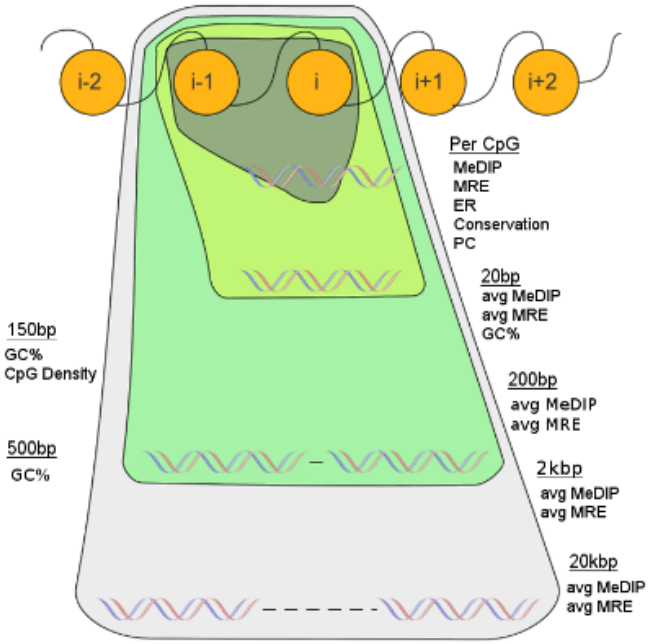


Figure S5: Diagram of the complete methylCRF model.

three obvious clusters. Grouping these into two clusters using a distance-based metric results in two clusters whose projection onto Y, does not suggest a discretization of Y. In contrast, multiplying Y by some sufficiently large value, w, produces a clustering whose projection onto Y, partitions Y into single interval clusters. (B) shows psuedo-code for the heuristic.

6. High methylCRF Concordance with WGBS is Independent of Accuracy Threshold

   The concordance of methylCRF methylation levels when compared to MethylC-seq is similar to that of BS-seq to MethylC-seq (Fig. 1, 2). This similarity is robust across the whole range of accuracies (Fig S7.A). Interestingly, the performance is actually rather good even on shuffled methylation values of either BS-seq or methylCRF, which underscores the critical importance of proper null model choice in methylome analysis. 80% of methylCRF methylation estimates are within $\pm$ 18% of the MethylC-seq levels (19% for BS-seq). To determine if our estimates also have a similar correlation pattern between neighboring CpGs, we looked at the accuracy of the change in methylation between neighboring CpGs (Fig. S7.B). That is, for every threshold and every CpG we calculate:

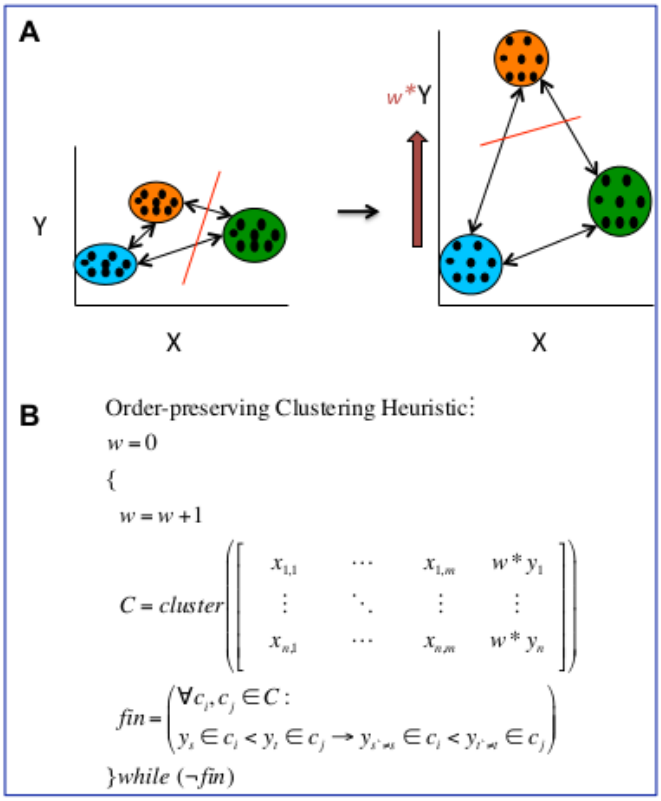


Figure S6: Order preserving heuristic. (a) the scale of Y is increased until a distance measure partitions X correspond to consistent ranges of Y. (b) psuedo-code

$$|methylCRF(CpG_i - CpG_{i-1}) - MethylCseq(CpG_i - CpG_{i-1})|$$

Similar to the methylation estimates in (Fig. S7.A), 80% of the differences are within ± 17% of differences of MethylC-seq (26% for BS-seq). In fact, the only significant difference between methylCRF and BS-seq estimates that we noted occur when we analyze the change in methylation of neighboring CpGs as a function of MethylC-seq methylation level (instead of by difference threshold). Even at a difference threshold of Âś 40%, less than 20% of the differences between methylCRF estimates of neighboring CpGs are concordant with that of MethylC-seq differences greater than 50%, while on the other hand, BS-seq estimates are more than 50% percent concordant. (Fig. S7.C) This suggests that methylCRF does not match large changes in methylation between neighboring CpGs well. However, since the methylation level correlation is, nevertheless, high between methylCRF and MethylC-seq as well as between neighboring CpGs in general, this suggests that methylCRF disagrees with the exact location of boundaries between regions of high and low methylation of MethylC-seq by a few CpGs rather than completely missing changes in methylation.

7. methylCRF can identify developmental enhancers with lower methylation.

   The concordance at human Vista enhancers [64, 84] with positive expression in mouse embryos is lower than the global concordance, 0.85. However, methylCRF is nevertheless able to identify which set of enhancers had reduced methylation. To show this we split the Vista enhancers into two groups: methylated (>0.66) and reduced methylation ($\leq$ 0.66) based on their average WGBS methylation. We choose 0.66 to be the cut-off for the large node in the distribution of methylation values, Fig 4A.

4.8.1 SUPPLEMENTARY TABLES

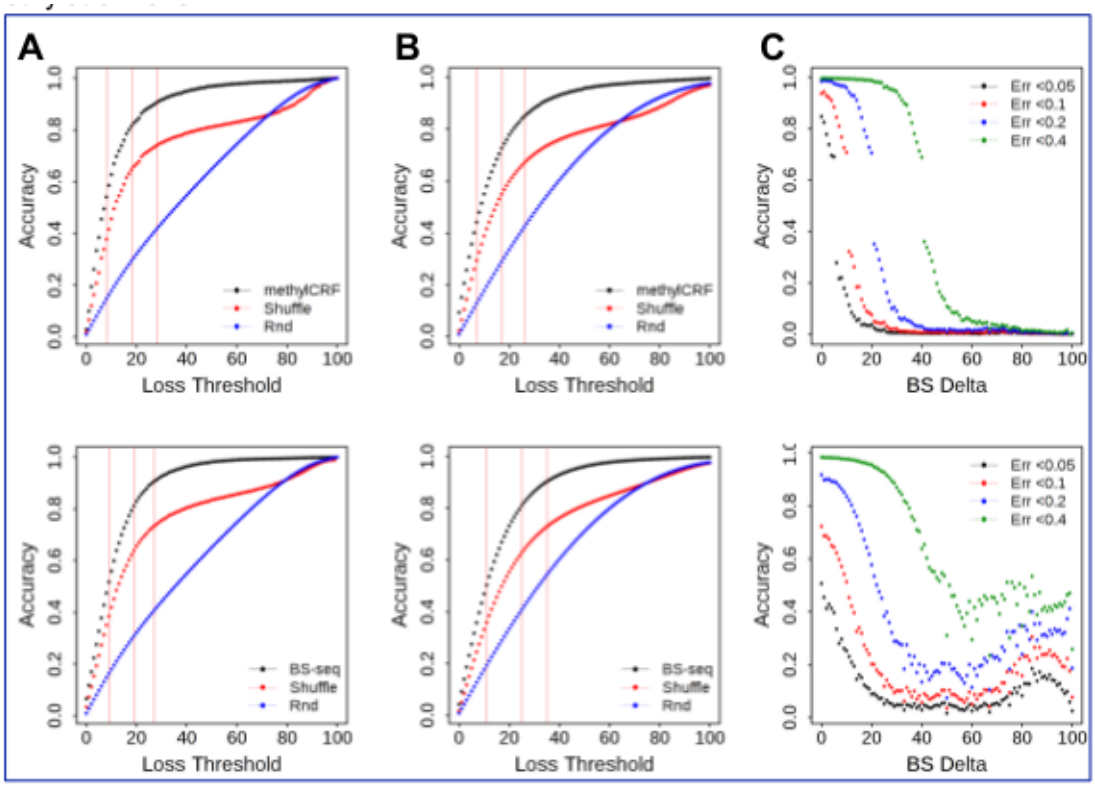


Figure S7: methylCRF [top] and BS-seq [bottom] accuracy as compared to MethylC-seq. Loss refers to the threshold of difference between methylCRF and MethylC-seq methylation levels to call them concordant. Accuracy refers to the ratio of CpGs within that threshold. (a) Accuracy by difference between methylation estimates, shuffled methylCRF estimates, and random methylation values (red lines show 50%, 80%, and 90% accuracy); (b) accuracy across similarity thresholds of the difference between neighboring CpG's methylation for methylCRF, shuffled methylCRF values, and random methylation values; (c) accuracy of 4 concordance thresholds of the difference between neighboring CpG's methylation by MethylC-seq methylation level.

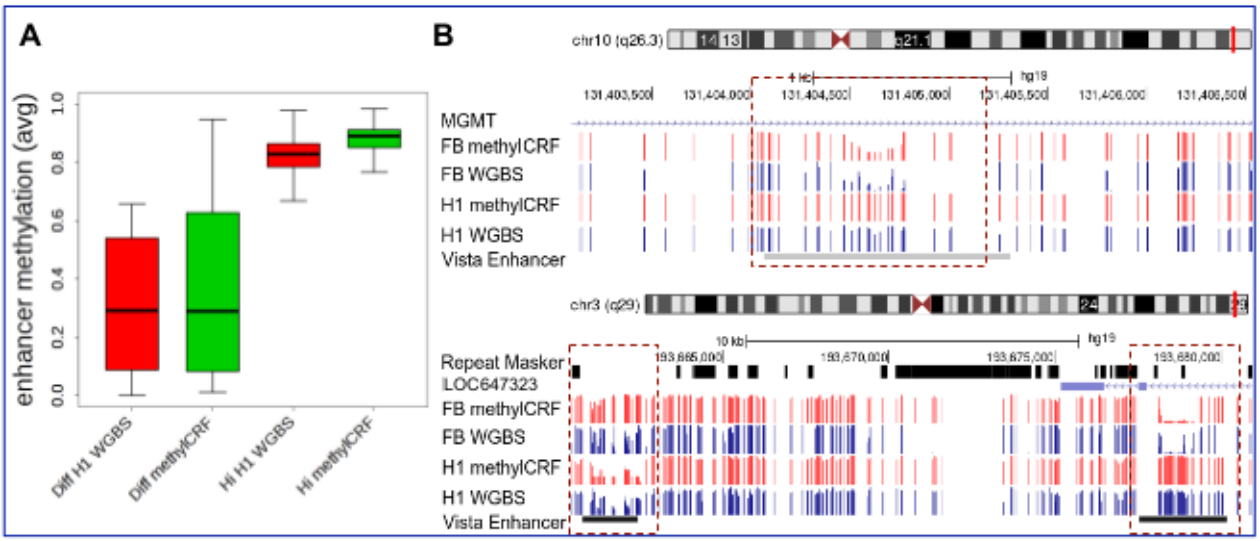


Figure S8: DMR detected at Vista enhancers. (A) methylCRF can identify enhancers with reduced methylation. (B) Example browser shots of DMR's at differentially methylated enhancers identified by the UCSC Genome Browser's Vista enhancer track. Upper panel: MGMT loci, associated with gliomas showing demethylation in NSC. Lower panel: LOC647323 loci showing a far upstream enhancer DMR demethylated in ES and an intronic enchancer DMR demethylated in NSC.

Experiment	Sample	GEO ID
MeDIP-seq	H1 ESC Batch1	GSM543016
	H1 ESC Batch2	GSM456941
	Fetal Brain HuFNSC02	GSM669613*
MRE-seq	H1 ESC Batch1	GSM428286
	H1 ESC Batch2	GSM450236
	Fetal Brain HuFNSC02	GSM669603*
WGBS	H1 ESC	GSM432686 GSM429321 GSM429322 GSM429323
	H9 ESC	GSM491349
	Fetal Brain HuFNSC02	GSM941746*

Table 4.2: Complete datasets used in this study
(* denotes unpublished data)

Chrom	Coordinate (hg19)	.	Imprinted gene
chr4	89618367	89620597	NAP1L5
chr6	144329408	144329947	PLAG1
chr7	50849753	50850871	MEG1
chr7	94286182	94286557	PEG10
chr7	130132066	130132356	MEST
chr8	141107838	141110984	PEG13
chr10	121577530	121578385	INPP5F_V2
chr11	2019368	2023499	HG19/IGF2
chr12	2806850	2808502	KCNQ1OT1
chr13	48892636	48893857	RB1
chr14	101275673	101277556	GTL2
chr15	23931560	23932547	NDN
chr15	25199662	25200343	SNURF
chr18	44554880	44556671	TCEB3C
chr19	57351728	57352173	PEG3
chr19	57630348	57630725	USP29
chr20	30135077	30135292	MCTS2
chr20	36147118	36151058	NNAT
chr20	57464743	57464960	GNAS

Table 4.3: Imprinted control regions used in this study. Adopted from `https://atlas.genetics.kcl.ac.uk`.

Chapter 5

Multiple cell-type DNA methylation dynamics at single CpG resolution captured by combinatorial methylCRF prediction

5.1 Abstract

DNA methylation is an important epigenetic modification involved in many fundamental biological processes and diseases. Many studies have shown methylation changes associated with embryogenesis, cell differentiation and cancer at a genome-wide scale. Our understandings of genome-wide methylation changes in a developmental or disease-related context have been steadily growing. However, the overarching view and understanding of methylation patterns in many different normal cell or tissue types are still lacking. Here we present an in-depth analysis of single-CpG resolution methylomes predicted using methylCRF on 58 cells. We found that methylCRF can accurately predict dynamic DNA methylation patterns across cell types. We categorized the 26 million human autosomal CpG based on their methylation level across all the cells and focused on variably methylated CpGs for further analysis. Among all the autosomal CpGs, only 28% show significant differences among cell types. We then grouped these CpGs into variably methylated regions (VMRs) to explore their functional importances. Overall, there are more than 400 thousand VMRs occupying 11% of the genome. We found that VMRs enrich enhancer histone modification marks, suggesting their role as regulatory

enhancers likely during cell differentiation. VMRs enrich transcription factor binding sites in a tissue-dependent manner; furthermore they enrich SNPs and GWAS variants, suggesting VMRs could potentially be implicated in disease progression. Taken together, these analyses demonstrated the power of methylCRF in characterizing CpG methylation and variably methylated regions, many of which harbor regulatory potentials. Our results highlighted the link among CpG variation, genetic variation and disease risk in a tissue-specific manner.

5.2 Introduction

DNA methylation refers to the chemical modification of the addition of a methyl group at the C5 position of cytosine on DNA sequences. Methylation on cytosine can occur in different genomic contexts but largely in a CpG dinucleotide context(Fazzari and Greally 2004). Proper establishment of DNA methylation early in embryogenesis is vital for normal development in many organisms(Law and Jacobsen 2010). DNA methylation clearly plays a role in genomic imprinting and X-chromosome inactivation where methylation of one parental allele suppresses its expression and leads to monoallelic gene expression(Reik and Lewis 2005). In addition, epigenetic modifications of the chromatin, including DNA methylation and histone modifications, orchestrate heritable, cell type- and developmental stage-specific gene expressions in vertebrates(Robertson 2005; Portela and Esteller 2010).

Since the advent and wide adaptation of next-generation sequencing in epigenomic studies(Harris et al. 2010), insights have been gained over different aspects of the functions of DNA methylation at a genome-wide scale in various contexts. We now have a catalogue of the DNA methylomes of many cell types in different organisms(Lister et al. 2010; Kobayashi et al. 2013; Lister et al. 2013; Shen et al. 2012; Zhang et al. 2013) and discovered while the majority of the DNA methylation remain stable once the cell is fully differentiated, dynamic DNA changes occur during embryogenesis, cell differentiation, tissue development, aging and disease progression and many more(Jiang et al. 2013; Meissner et al. 2008; Laurent et al. 2010). While the majority of the efforts have been focused on identifying specific DNA methylation changes induced by specific treatment or environmental stimuli and by disease progression, little knowledge is known about the systematic pattern of DNA methylation across many cells types from different tissues under physiologically normal conditions. Among

the 28 million CpG in the human genome, we still have little idea of what proportion of them show dynamic difference and could be functionally important.

Through the effort of the Roadmap Epigenomics Project(Bernstein et al. 2010), we now have a large collection of genome-wide DNA methylation profile spanning multiple tissue and cell types, in the form of complementary methylated DNA immunoprecipitation (MeDIP-seq) and methylation-sensitive restriction enzyme (MRE-seq) data(Maunakea et al. 2010). We recently have introduced a novel conditional random fields-based algorithm, methylCRF, which combines both MeDIP-seq and MRE-seq data and predicts single-CpG resolution DNA methylomes(Stevens et al. 2013). Here, we have leveraged the new algorithm and analyzed a large number of full DNA methylomes across multiple cell and tissue types. We have found a relatively small percentage of the autosomal CpGs that show dynamic changes among the cell types we investigated. Merging the dynamic CpGs into dynamic regions, we have characterized the features of these regions and their co-localization with various regulatory elements such as enhancer associated histone marks, transcription factor binding sites and disease associated GWAS hits and uncovered many important functions these dynamic regions might possess.

5.3 Results

5.3.1 Characterization of autosomal CpG methylation patterns

Through Roadmap Epigenomics Project, we have collected 43 MeDIP-seq and MRE-seq methylation datasets and together with another 15 normal human DNA methylomes in-house, we are interrogating the DNA methylomes of 58 normal human primary cell samples including fetal brain, cortex derived and ganglionic eminence derived neurosphere cells, fibroblast, keratinocyte, melanocyte of skin, luminal epithelial, myoepithelial and stem breast cells, CD4 memory, CD4 naive, CD14, whole blood, granulocyte and peripheral blood mononuclear cell (PBMC), and endometrium. For a complete list of cell and tissue types, refer to Supplemental Table 1.

We have applied methylCRF to these MeDIP-seq and MRE-seq datasets and generated 58 single CpG resolution complete DNA methylomes. For all the following analysis, only autosomal CpGs were considered. To test the similarity between the methylation levels of our methylCRF prediction and those of Whole Genome Bisulfite Sequencing (WGBS)(Ziller et al. 2013), we calculated their genome-wide concordance (defined as the percent of CpGs with a methylation level difference less than 0.25) between samples profiled by both methods. Although the samples were from different sources, the concordance between tested pairs is as high as 85.3% (Supplemental Table 2).

We first looked at the global pattern of DNA methylation in different cell types. The average DNA methylation level of each cell type ranges from 75.4% to 81.7% and displays small variation among tissue types and fairly consistent level of methylation among different cell types from the same tissue type (Supplemental Table 1). For each cell type, the distribution of the overall DNA methylation level follows a pattern similar to other known methylomes where the majority (more than 78% of the total CpGs for every sample) are either highly methylated or unmethylated (about 11% of the total CpGs for every sample) and a small percentage (less than 7% for every sample) of CpGs are intermediately methylated (Supplemental Table 1, Supplemental Figure 1a,b).

We next explored the relationships among 58 DNA methylomes and see if there are some apparent patterns from a global scale. To this end, we binned CpGs into 1kb windows and used the average methylation of each window for Principle Component Analysis (PCA) and hierarchical clustering analysis (Figure 1a and Supplemental Figure 1c). Both analyses clearly showed the separation and clustering of samples by tissue types and loosely by cell types. In particular, CD4 memory cell and CD4 naive cell methylomes are particularly similar among each other whereas cells such as breast luminal epithelial cells are not as closely clustered, reflecting cell-type dependent variability in global DNA methylome patterns. This variability could be due to inherent epigenomic variability in certain cell types or cell type mixture during sample harvesting steps.

With regard to the average methylation pattern at different genomic features (promoters, exons, introns etc.), we found that all the cell types examined shared the known pattern of relatively low average methylation level in promoter regions compared to gene bodies (exons and introns) or intergenic regions (Figure 1b).

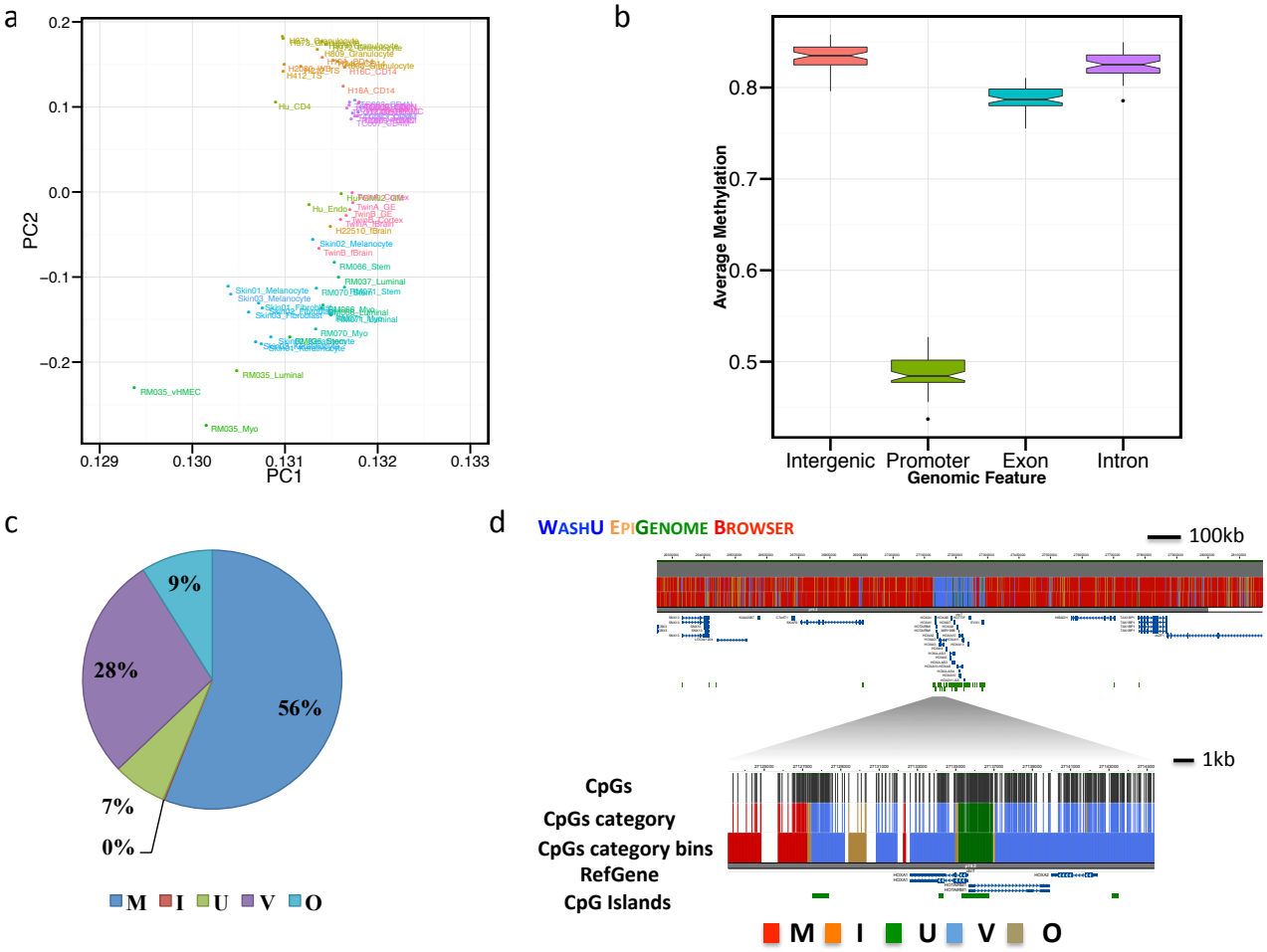


Figure 5.1: Characterization of autosomal CpG methylation patterns across 58 sample datasets. A. Principal component analysis of CpG methylation levels for 1kb genomic bins across 58 methylCRF datasets. B. Average methylation level of genomic features over 58 samples. C. Categorization of autosomal CpGs based on methylation level across 58 datasets. M: constitutively methylated CpGs; I: constitutively intermediately methylated CpGs; U: constitutively unmethylated CpGs; V: variably methylated CpGs; O: others. D. A browser shot of a genomic region showing different categories of CpGs and the corresponding merged regions

5.3.2 Categorization of CpGs

To date, most of the studies on DNA methylation were only focused on a few samples at a genome-wide scale and there is a lack of understanding of which CpGs are always methylated or unmethylated, regardless of the cell type being studied. With the availability of samples in our collection, we thought to categorize all the autosomal CpGs based on their methylation patterns across a large number of samples. Specifically, we picked 70% and 30% methylation as cutoffs to categorize CpGs into five categories: constitutively methylated CpGs (M), constitutively intermediately methylated CpGs (I), constitutively unmethylated CpGs (U), variably methylated CpGs (V) and the rest (O) (See Methods for detailed explanation on CpG categorization). Based on this classification scheme, we found that 56% of autosomal CpGs are constitutively methylated in all the samples we examined, 7% constitutively unmethylated and less than 1% constitutively intermediately methylated (Figure 1c and Supplemental Table 3). This result suggests that more than half of the autosomal CpGs are stably methylated (>= 70% methylation) in all the samples we looked, in line with the notion that DNA methylation is a stable epigenetic mark. In addition, nearly 60% of constitutively methylated CpGs are located in repeats and 64.2% of CpGs in repeats are constitutively methylated (Supplemental Figure 2a,b). On the other hand, majority (76.1%) of constitutively unmethylated CpGs are located in CpG islands and furthermore, 68.1% of all the CpGs in CpG islands are constitutively unmethylated CpGs (Supplemental Figure 2c,d).

5.3.3 Identification and characterization of variably methylated CpGs and regions

Given the nature of variably methylated CpGs and their potential of playing a role in regulating gene transcription, we then focused our analysis on the set of variably methylate CpGs and determined what portion of CpGs shows a dramatic DNA methylation level difference among cell types. To this end, we calculated the difference between the highest and lowest methylation score of all samples for each CpG and use 40% difference as an empirical cutoff (See Methods, Supplemental Figure 3) to define these variably methylated CpGs (VMCs). After applying the cutoff, 7.5 million CpGs were identified as variably methylated CpGs that account for 28.2% of all autosomal CpGs in the genome (Supplemental

Figure 4a). We then merged neighboring CpGs into windows to generate variably methylated regions (VMRs) based on the merging criteria (described in Methods). This merging leads to the identification of 560765 VMRs whose base coverage account for 16.9% of the genome (Supplemental Table 3, Figure 1d).

Next, we used several metrics to characterize these VMRs. The sizes of VMRs vary from 100bp up to 5.3kb (Figure 2, Supplemental Figure 4b). However, majority of the VMRs are small as 70.7% and 89.3% of VMRs are smaller than 1kb and 2kb, respectively. The number of variable CpGs covered in VMRs mostly falls within 50 (Supplementary Fig. 4c). Over 65% of CpGs covered in VMRs are methylated (>= 70% methylation) and less than 10% are unmethylated (<= 30% methylation) on average across samples (Supplementary Fig. 4d). We found that majority of the VMRs are located in intron and intergenic regions and only a small percentage overlapping promoter or exons (Figure2b). In addition, majority of VMRs are located far away from annotated transcription start sites (TSSs, Figure 2b), suggesting they might play a role as distal regulatory elements, specifically, a VMR with hypomethylation in certain cell types might give access to other cis- or trans- regulatory elements such as transcription factors or other DNA-binding proteins to modulate nearby gene expressions. To test this hypothesis, we first determined hypomethylation in each cell type (see Methods, Figure 2d) and then employed publicly available datasets from ENCODE(Consortium et al. 2013) on various histone marks and transcription factor binding sites and gene expression profiles and integrated them to explore the potential functions of these tissue specific hypomethylated VMRs.

5.3.4 VMRs enrich transcription factor binding sites

We first examined co-localization between VMRs and over 160 transcription factor binding site (TFBS) ChIP-seq peaks. We found that 40% of the total VMRs overlap with at least one TFBS peak and 24% with three or more TFBS peaks (Figure 3a). This result suggests wide spread co-localization between TFBS and identified VMRs even given the currently limited amount of ChIP-seq data and it's tempting to postulate that those VMRs that didn't overlap with current TFBS peaks might also harbor regulatory potentials through modulating transcription factor binding. To further pinpoint tissue-specific contribution to these observed co-localization, we calculated enrichment of TFBS signals over size-matched randomly selected

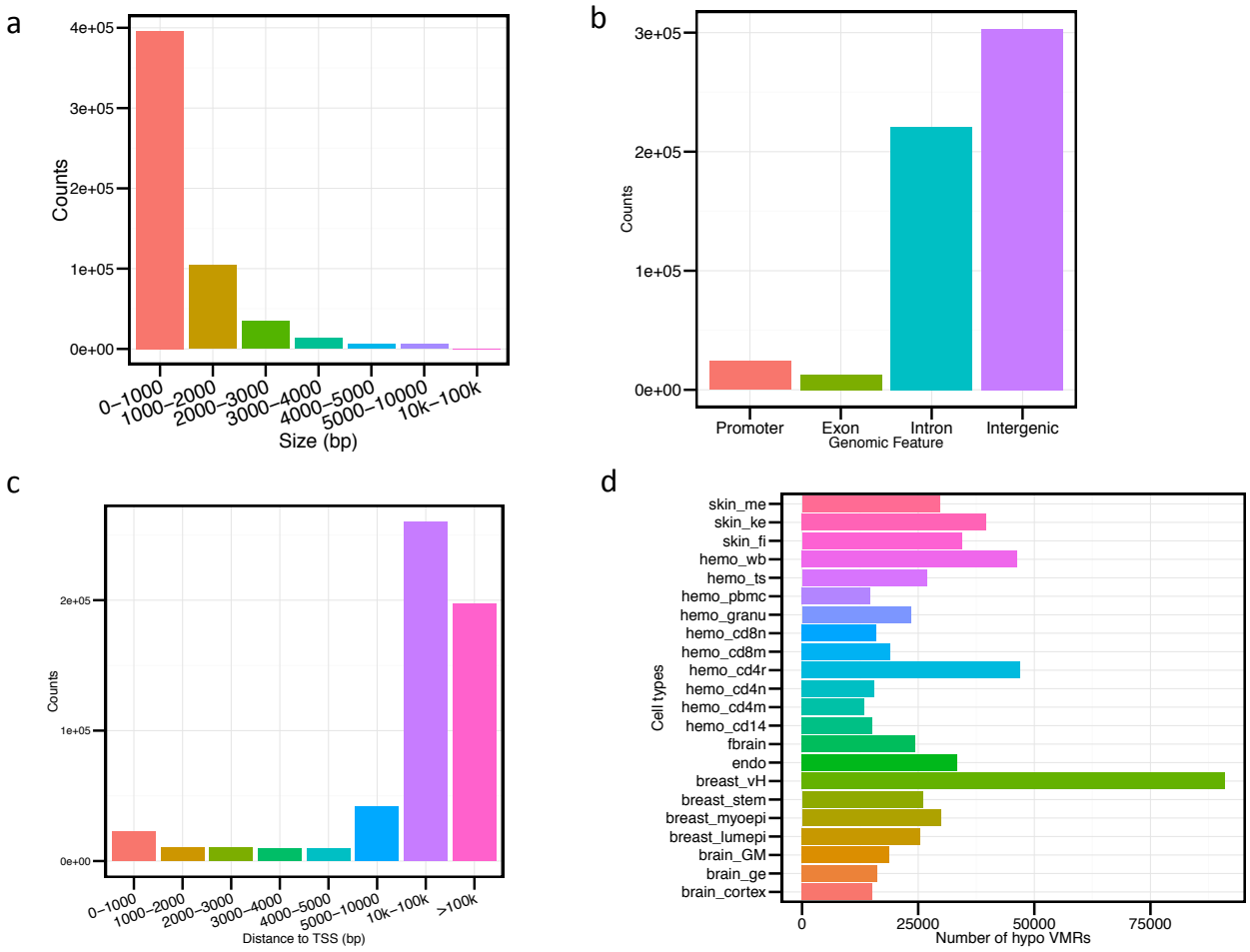


Figure 5.2: Characterization of variably methylated regions. A. Distribution of VMR sizes at each size ranges. B. Distribution of VMRs over different genomic features. C. Distance of VMRs to their nearest transcription start sites (TSSs). D. Number of hypomethylated VMRs in each cell type

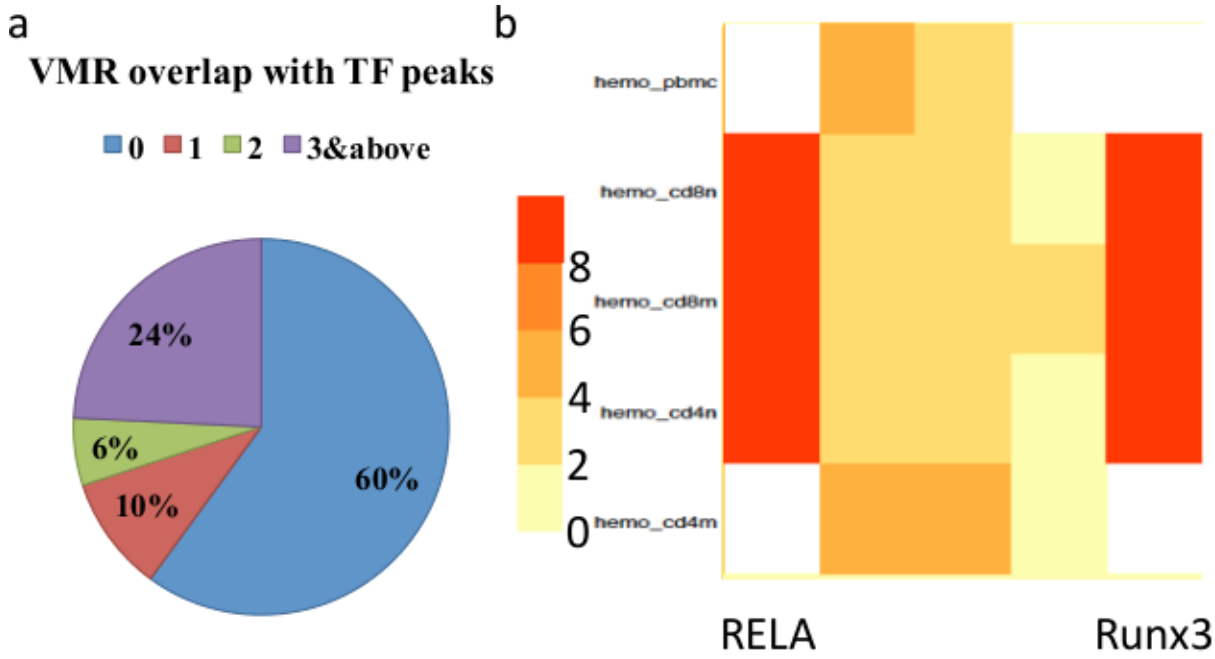


Figure 5.3: VMRs enrich transcription factor binding sites. A. Co-localization between hypomethylated VMRs and transcription factor binding sites. B. Enrichment of specific transcription factor binding sites in each cell type specific hypomethylated regions

genomic regions (See Methods). We found specific enrichment of certain transcription factor peaks in VMR that include several interesting examples of known functionally important transcription factors in particular cell types (Figure 3b). For example, RelA is an important transcription factor that plays a role in immune responses and its deletion cause defects in hematopoietic stem cell function(Grossmann et al. 1999; Stein and Baldwin 2013). Through our enrichment analysis, we found high enrichment over CD8M, CD8N, and CD4N specific hypomethylated VMRs. The idea is that RelA bind to tissue-specific hypomethylated regions and participates in the regulation of various RelA target genes.

5.3.5 VMRs co-localize enhancer histone marks and many possess enhancer potentials and validated enhancer activities

Given the location of the identified VMRs are largely far away from TSSs, we sought to explore their regulatory potentials by examining their relationships with various histone

marks assayed on the same cells. We called ChIP-seq peaks of histone mark for each cell type and found that up to 72% of the cell type specific hypomethylated VMRs overlap with enhancer histone mark ChIP-seq peaks represented by H3K4me1 (Supplemental Figure 6a) in the same cell type. Indeed, when we calculated the enrichment of various histone peaks at VMRs, we found a strong enrichment of peaks for marks that possess enhancer or active transcription activities, such as H3K4me1, H3K4me3 and H3K9ac (Figure 4a). In contrast, there is also a depletion of repressive histone marks (H3K36me3 and H3K9me3) at these VMRs. We also calculated the average signal density of ChIP-seq data over 10 kb regions centered around VMRs and found in general higher level of enhancer or active transcription mark in the VMRs compared to its flanking sequences (Figure 4b). These results thus highlighted that many of the VMRs we identified could have potential enhancer activities in different cell types.

Next, we analyzed the functional enrichment for genes near the VMRs in each tissue type and indeed found in cells enrichment of genes whose functions are relevant for that particular cell types. For example, we found many genes near brain specific hypoVMRs that encode functions such as axon extension, nerve cell development etc (Figure 4c). This evidence further supports the hypothesis that hypomethyled VMRs act as enhancers in specific cells to regulate nearby functionally related genes, possibly through transcriptional regulation mechanism.

To validate the enhancer hypothesis, we utilized the VISTA enhancer project resources and determined whether our VMRs could include some validated enhancer regions(Visel et al. 2007). Indeed, we found that 78% of positively validated human VISTA enhancers overlap with at least one of the VMRs and this overlap is statistically significant (Figure 4d, Fisher's exact test). For example, VISTA enhancer hs1546 was validated to have enhancer activity in mouse forebrain and it overlaps with a VMR hypomethylated in our fetal brain samples (Figure 4d). Because VISTA enhancer list only have more than 2000 loci validated, we predict that many more VMR could possess enhancer activities and they could potentially function during cell differentiation.

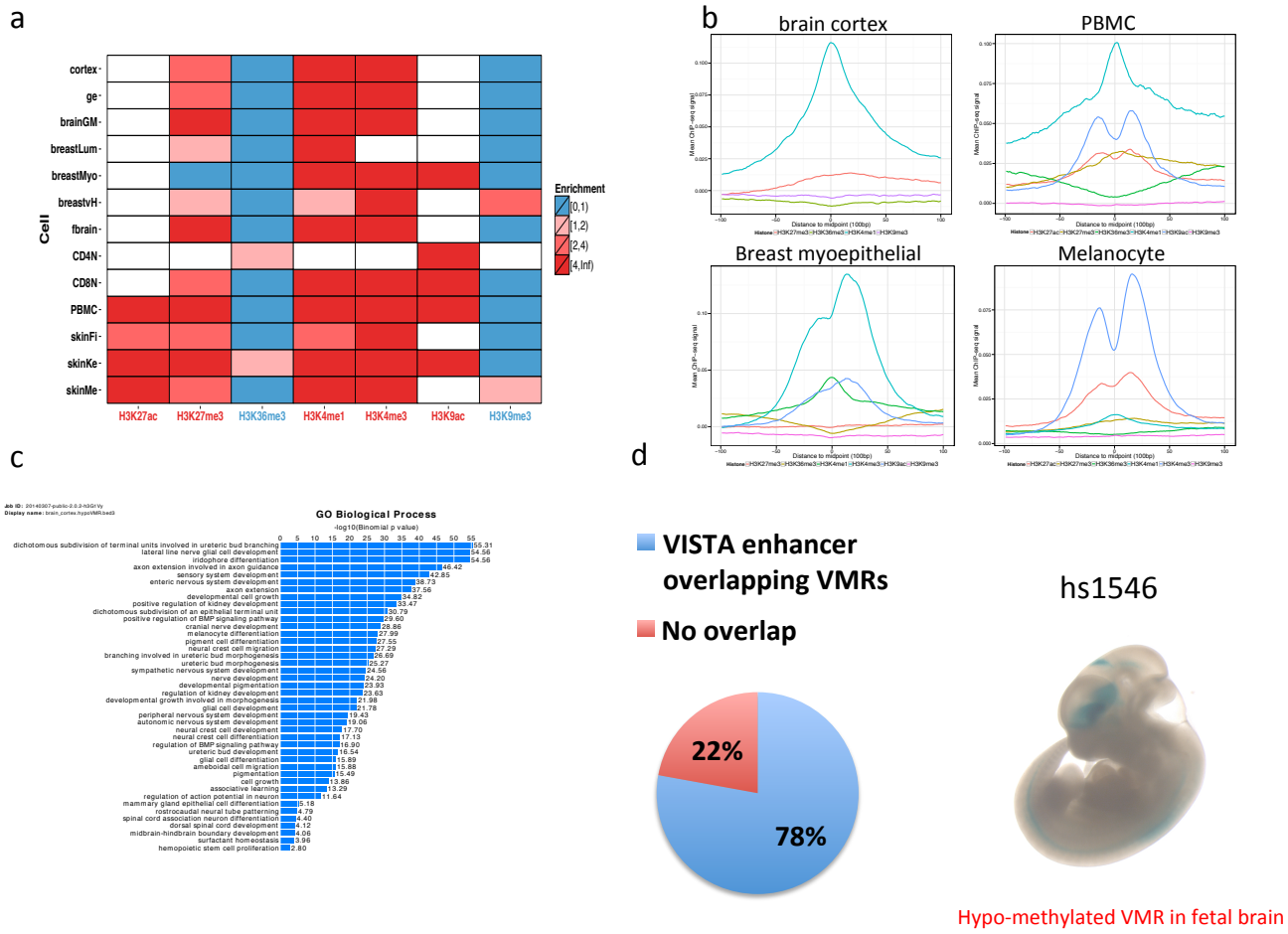


Figure 5.4: VMRs co-localize enhancer histone marks and possess enhancer potentials. A. Enrichment of histone marks ChIP-seq peaks in VMRs. B. ChIP-seq signal density over 10 kb region centered on VMRs. C. GO term enrichment of genes near VMRs. D. Overlap of VISTA enhancers with VMRs

5.3.6 VMRs enrich SNPs and GWAS variants

It was recently shown that genomic variations often co-localize with presumed regulatory regions (Consortium et al. 2013; Manolio 2013). We hypothesized that VMRs might contain many functional genomic variations and methylation might play a role in regulating these genomic variations. Thus, we examined the co-localization and enrichment of dbSNPs and variants from disease related Genome-wide Association Studies within our VMRs. We found that VMRs in total encompass 3323 published GWAS variants and there is slightly enrichment of GWAS variants in each cell type specific hypomethylated VMR (Figure 5b).

To further pinpoint whether methylation could play a role in regulating these genomic variations, we overlap variants with variably methylated CpGs and indeed we found many co-localizations. One very interesting example lies at SNP rs1805007. This SNP was located in the exon of the gene MC1R, the melanocortin 1 receptor. The common allele C resides in a CpG context and its risk allele is T. Studies have known that risk allele T induces a nonfunctional MC1R variant that is strongly associated with red hair phenotype and increased risk to melanoma(Frandberg et al. 1998; Flanagan et al. 2000). And this CpG was identified to have low methylation specifically in melanocyte and high methylation in other cell types including keratinocyte and fibroblast in skin. We speculate that this particular CpG's unmethylation signature confers a protective function to melanocyte because methylated Cs have a much higher rate to deaminate into T, especially under the exposure of UV(Ikehata and Ono 2006; Fryxell 2004). On the other hand, the high methylation level of this CpG in keratinocyte confers susceptibility of the locus to mutate into T and this discovery corroborates the fact that the majority of skin cancers are of keratinocyte origin(Albert and Weinstock 2003). And it'd be reasonable to suspect that methylation of this CpG in melanocyte might contribute to the progression towards melanoma under excessive sunlight exposure.

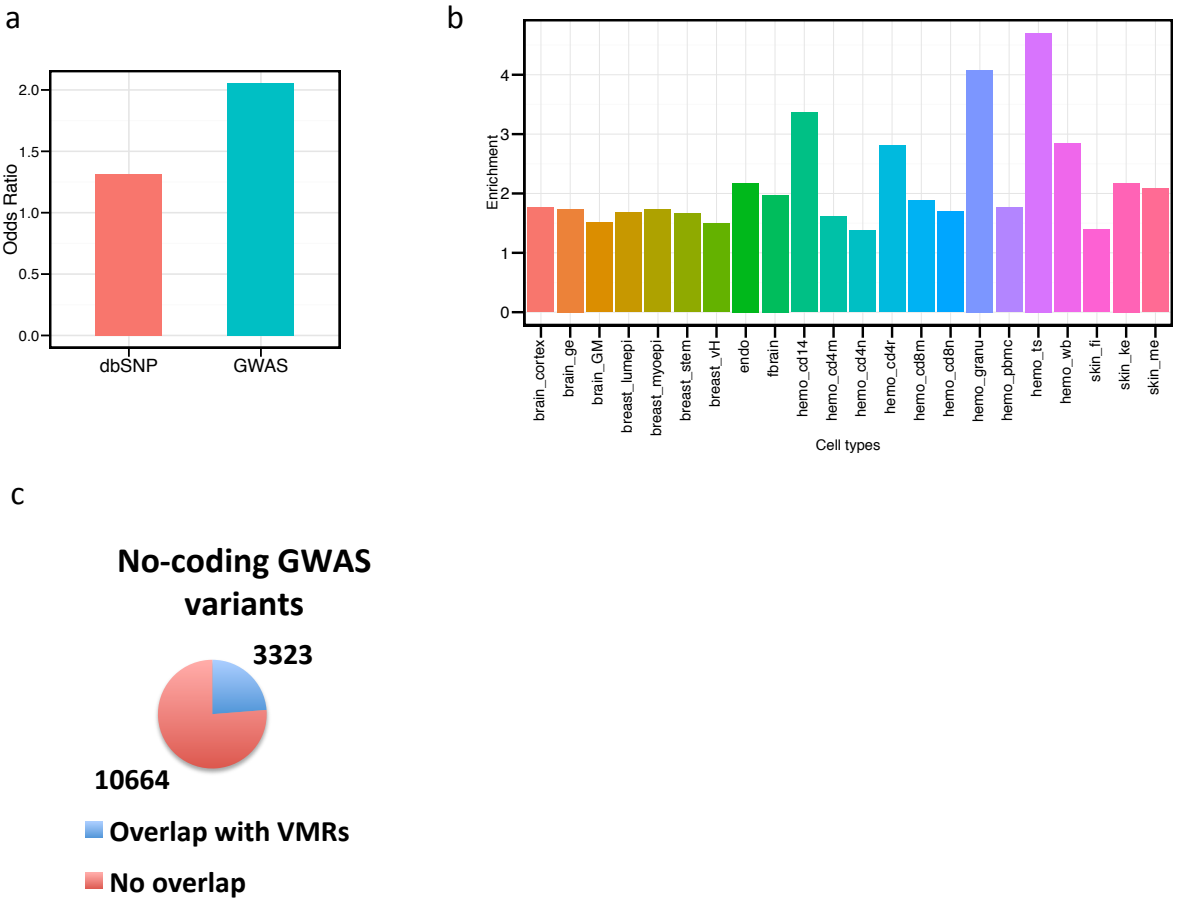


Figure 5.5: CpG variation is linked to genetic variation and disease risk in a tissue-specific manner. A. Odds ratio of VMRs overlapping with dbSNP and GWAS variants. B. Enrichment of GWAS variants in tissue specific hypomethylated VMRs. C. Overlap between non-coding GWAS variants and VMRs

5.3.7 Hypomethylated VMRs correlate with nearby gene expressions

To support that cell type specific hypomethylated VMRs could act as enhancers to mediate gene expression, we analyzed gene expression data from the same cell type and see whether we can observe expression differences correlated with methylation differences in VMRs. To this end, we identified genes near these VMRs and calculated their expression level from RNA-seq data and compared their levels between two cell types. We found that in general between two cell types, genes near those hypomethylated VMRs tend to have statistically significantly higher expression level (Supplemental Figure 8). This supports the hypothesis that hypomethylated VMRs act as tissue-specific enhancers to drive nearby gene expressions.

5.3.8 Average methylations on VMRs cluster samples by tissue type

To show the utility of the VMR set, we calculated average methylation level of each sample at these VMRs and cluster them with hierarchical clustering. Indeed, methylation level at VMRs can largely cluster the samples by their tissue types (Supplemental Figure 9).

5.3.9 Characterization of other categorical regions

Besides identifying variably methylated CpGs in the genome, we were also interested CpGs that show consistent methylation levels across all the samples. With the availability of 58 methylomes spanning many cell types, we were able to identify CpGs that are either constitutively highly methylated ($> 70\%$ methylation) or constitutively lowly methylated ($< 30\%$ methylation) in every sample. In total, we identified almost 14.9 million constitutively highly methylated CpGs that account for 56.01% of the total autosomal CpGs. In contrast, only 1.8 million CpGs (6.74% of all autosomal CpGs) shows constitutively low methylation in every samples assayed in this study. And the constitutively intermediated CpGs (methylation levels are between 30% and 70%) only account a tiny fraction (0.16%) of the total CpGs. There are about 2.4 million CpGs that don't fall into any of the above categories. Following

the same rule used to merge variably methylated CpGs, we merged CpGs in other categories and applied post-merging trimming so the resulting categorical regions are non-overlapping (See Methods).

To characterize these regions, we also looked at their size distribution, distance to nearest TSS and genomic distributions. We found that constitutively methylated (MMRs) and unmethylated regions (UMRs) have a larger median size than VMRs whereas constitutively intermediately methylated (IMR) and other regions (OR) are smaller in median size yet the overall distribution agree with that of VMRs (Supplemental Figure 10 a, b). On the distance of region to nearest TSS, the MMRs and ORs have similar distribution to VMRs and UMRs have a higher percentage located near TSS and promoters in agreement with the fact that many of the gene promoter proximal regions are unmethylated (Supplemental Figure 10 c, d).

5.3.10 Comparison with WGBS-based dynamic CpGs and DMRs

Ziller et al. recently published a paper detailing the first dozens of DNA methylomes using whole genome bisulfite sequencing (WGBS) [90] in normal developmental tissues and cultured cell lines. They defined a list of differentially methylated regions (DMRs) and we set out to determine the extent to which their DMRs overlap with our VMRs. First of all, the numbers of total covered bases are similar (487 million bases for VMRs vs. 492 million bases for DMRs). We looked at the intersection between bases covered by VMRs and bases covered by DMRs and we found that 36.4% of VMRs covered bases overlap with 36% of DMR covered bases. VMRs and DMRs share the similar properties such as small in size yet the number of DMRs is much larger than that of VMRs (716087 vs. 560765) and further 44.2% of VMRs overlap with 32.9% of DMRs (Supplemental Figure 11 a, b). To understand where the discrepancy lies, we looked at the size distribution of DMR-specific and VMR-specific regions. It appears that VMR-specific regions have slightly larger size than DMR-specific regions (Supplemental Figure 11 c). We checked how many of DMRs or VMRs are derived from a single CpG and found that 36.6% of DMRs but only 5.7% of VMRs are single CpG regions (Supplemental Figure 11 d). This difference could be due to the fact that WGBS relies on sufficient number of reads to accurately call methylation and some of the single CpG DMRs could have low coverage and thus inaccurate methylation calls. Indeed we found evidence suggesting that single CpG DMRs have lower read coverage than the rest of CpGs covered by

DMRs (Supplemental Figure 11 e). If we only consider the non-single-CpG regions in both sets, we observed almost the same percentage of covered base overlap and a slightly higher overlap of regions between VMR and DMRs (Supplemental Figure 11 f). The discrepancy not explained by single-CpG regions are probably due to either the inherent difference between WGBS and TFBS or different methods used to define them or both.

5.4 Disscussion

Here we present the first multiple cell type DNA methylomes analysis using methylCRF predictions. We have previously shown that methylCRF is highly robust in predicting single CpG methylation level genome-wide with high accuracy in comparison to direct whole genome bisulfite sequencing but at a much lower cost(Stevens et al. 2013). And in this study, we applied methylCRF to a total of 58 samples spanning multiple tissue and cell types and we were able to categorize CpGs based on their patterns across all the samples examined. The field has known for decades that the majority of the CpGs in the genome are methylated and this is confirmed in our analysis that about 56% of the autosomal CpGs are constitutively methylated (>= 70% CpG methylation) regardless of the cell type. In contrast, 7% of the autosomal CpGs are constitutively unmethylated (<= 30% CpG methylation) and as we have shown, most of these unmethylated regions located in CpG islands and gene promoter proximal regions. Focusing on the variably methylated CpGs and subsequently merged regions, we have identified 28% of the genomic CpGs that show dynamic CpG methylation across 58 samples. These CpGs could be functioning in different cell types by varying their methylation level and influence the transcriptional network in a particular cell types.

5.5 Methods

5.5.1 Data processing and methylCRF prediction

MeDIP-seq and MRE-seq data were aligned using BWA (Li and Durbin 2009) against human hg19 and then processed with methylCRF pipeline to generate single CpG methylation

predictions for each of the 58 samples(Stevens et al. 2013). For the rest of the analysis, only autosomal CpGs are considered.

5.5.2 CpG categorization

Each autosomal CpGs was scanned and assigned into one of the five categories: constitutively methylated CpGs (M), constitutively unmethylated CpGs (U), constitutively intermediately methylated CpGs (I), variably methylated CpGs (V), and others (O). A CpG is called to be constitutively methylated if every methylCRF prediction is equal to or greater than 0.7. Likewise, a CpG is called constitutively unmethylated if every methylCRF prediction is equal to or smaller than 0.3. And if every methylCRF prediction falls between 0.3 and 0.7, a CpG is intermediated methylated. To call variably methylated CpGs, we calculated the different between the maximum and the minimum CpG methylation levels for all the CpGs. If the difference for a CpG is equal to or greater than 0.4, it is considered a variably methylated CpG. We estimated the likelihood of identifying false positive variably methylated CpGs with our method by applying the same calculation to three CD4 memory and three CD4 naÃŕve datasets. Based on the concordance and Pearson correlation among these datasets (Supplemental Figure 2), we made the assumption that these datasets are similar enough to be considered as controls for our variably methylated CpG determination and any CpGs identified from within these datasets would approximate variably methylated CpGs detected by chance. With 40% methylation difference, the number of variably methylated CpGs determined from CD4 memory and CD4 naÃŕve data are below 10% of that from 58 samples.

5.5.3 Merging CpGs into regions

We merged CpGs into regions for CpGs in each category by adopting the merging roles from Ziller et al [90]. For CpGs within 500bp of each other, merge them into one window. For CpGs more than 500bp away from its closest neighboring CpG, extend the coordinate by 50bp in both directions to get 100bp windows. After the above merging step, if the size of a region is not greater than 100bp, extend it to 100bp from the center of the region. After CpGs were merged into regions for each category, we filter out overlapping regions in the following order. For variably methylated regions (VMRs), keep them as they are and remove those

overlapping with VMRs from constitutively unmethylated regions (CUMRs). Then filter out those overlapping with the union of VMRs and UMRs from constitutively intermediately methylated regions (CIMRs). Next, filter out those overlapping with the union of VMRs, CUMRs and CIMRs from constitutively methylated regions (CMRs). Lastly, filter out those overlapping with the union of VMRs, CUMRs, CIMRs and CMRs from other regions. In this way, all the resulting categorical regions are non-overlapping.

5.5.4 Determining hypomethylation of VMRs

Since we are averaging methylation level of VMRs for samples of the same cell type, we loosed the cutoff and called a VMR hypomethylated in a cell type if the average methylation of the region across samples in the cell type is below 40% and hypermethylated if above 70

5.5.5 Browser tracks

All the methylCRF data and custom tracks are displayed using the WashU Epigenome Browser (Zhou et al. 2011).

5.5.6 Genomic features

Transcription start sites (TSSs) and other genomic feature information were downloaded from UCSC genome browser (hg19)(Kent et al. 2002).

5.5.7 Histone ChIP-seq peak calling and enrichment calcuation

REMC ChIP-seq data were downloaded from GEO (GEO ID: GSE16368). Processed BED files were used for each ChIP-seq dataset. Histone peaks were called using SICER using the default parameters against hg19 genome(Zang et al. 2009). The enrichment of histone peaks in VMRs were calculated using the number of VMRs overlapping called histone peaks divided

by the number of size-matched randomly selected regions from the genome overlapping the same histone peaks.

5.5.8 TFBS ChIP-seq enrichment

TFBS ChIP-seq peak data were downloaded from ENCODE consortium. The enrichment of TFBS ChIP-seq peaks in VMRs were calculated using the number of VMRs overlapping peaks divided by the number of size-matched randomly selected regions overlapping the same TFBS ChIP-seq peaks.

5.5.9 GWAS variants

GWAS variants data were downloaded and processed from GWAS Catalog of National Human Genome Research Institute (Hindorff et al.).

5.6 Supplemental Figures

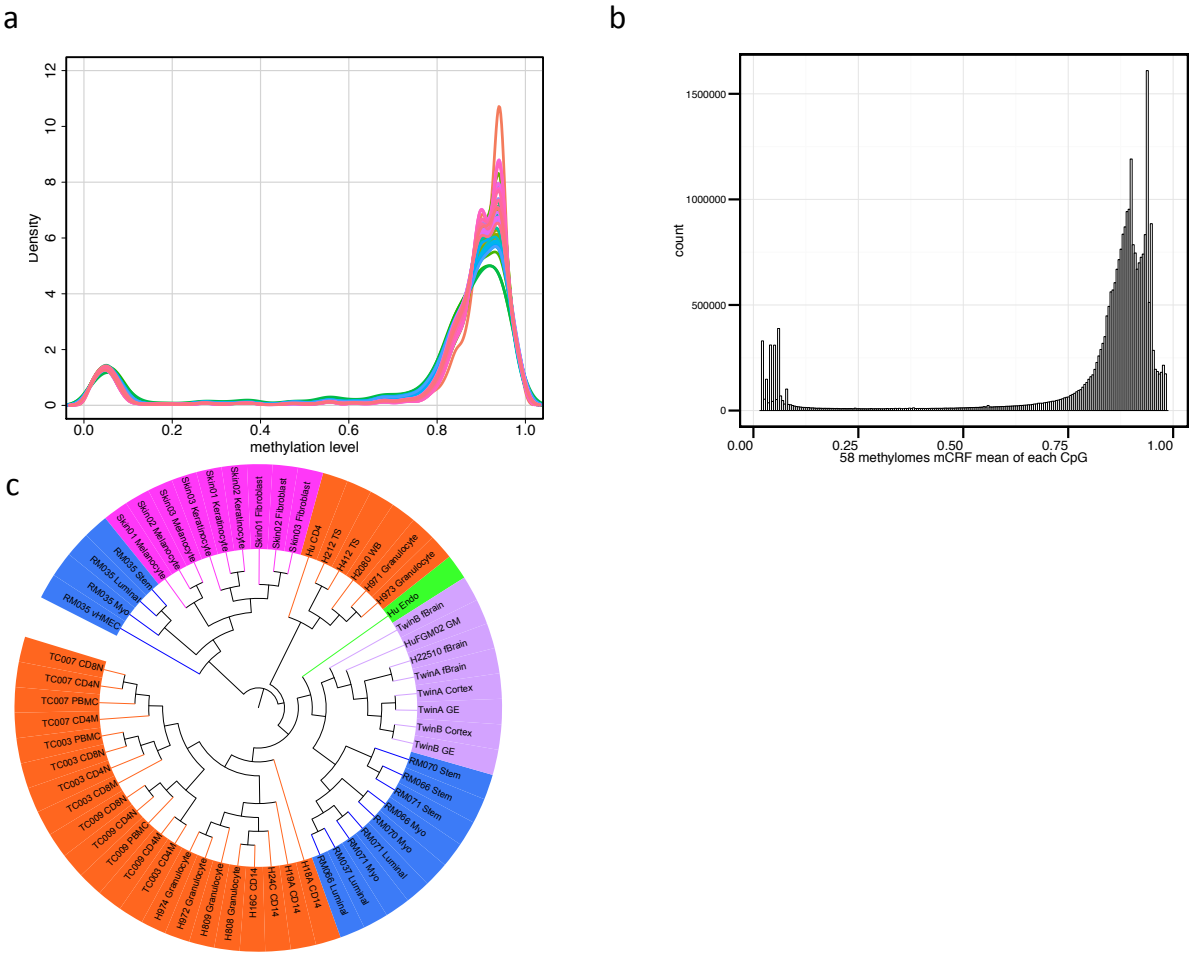


Figure S1: Characterization of CpG methylation across examined samples. A. Density plot of methylCRF prediction of all 58 samples. B. Distribution of individual CpG methylation level across 58 samples. C.Clustering of 58 samples based on average methylation level at 1kb bins

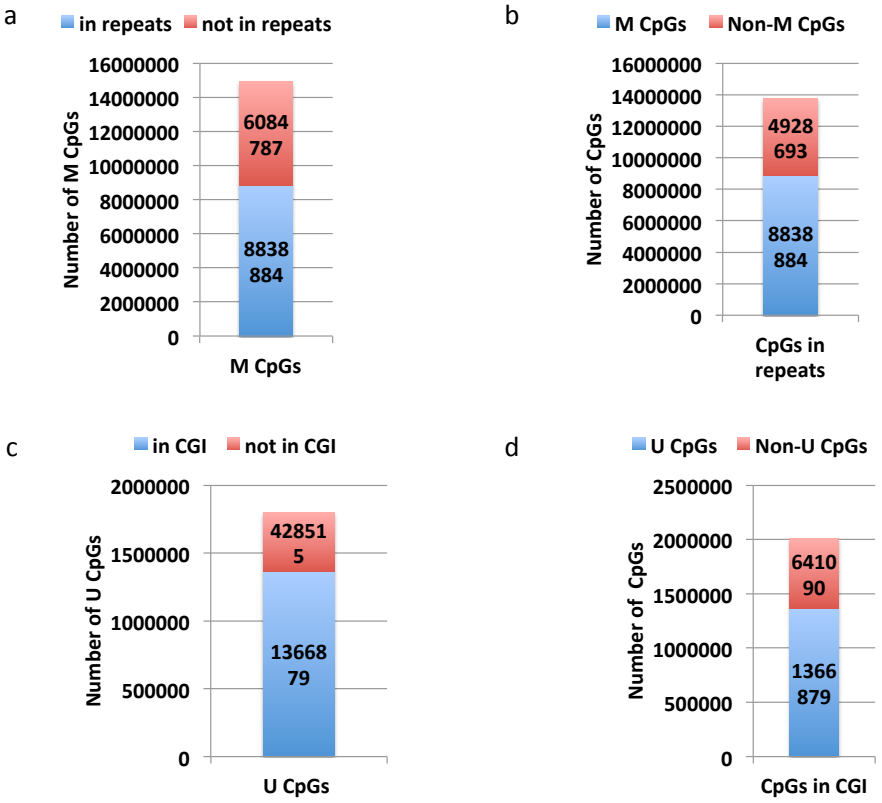


Figure S2: Overlap of constitutively methylated regions with repeats, constitutively unmethylated regions with CpG islands. A. Number of constitutively methylated CpGs in repeats vs not in repeats. B. Number of constitutively methylated CpGs and non-constitutively methylated CpGs in repeats. C. Number of constitutively unmethylated CpGs in CpG islands vs not in CpG islands. D. Number of constitutively unmethylated CpGs and non-constitutively methylated CpGs in CpG islands

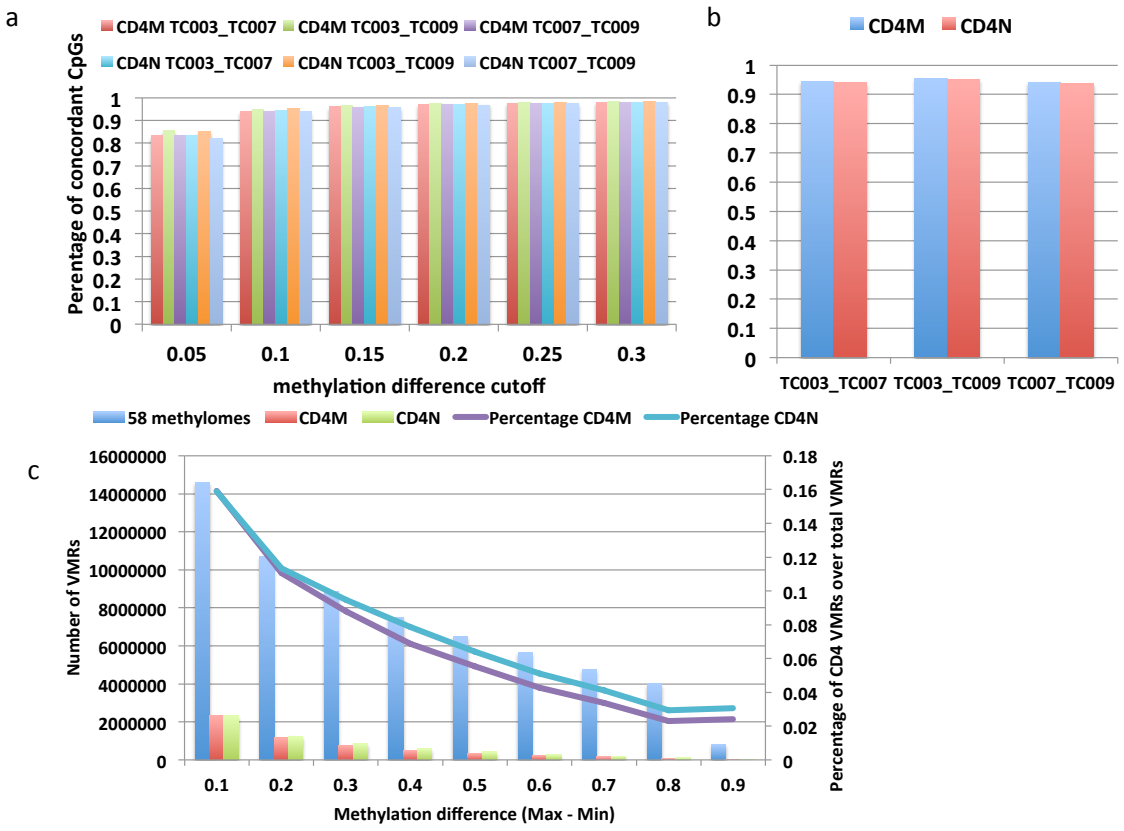


Figure S3: Concordance and Pearson correlation and estimation of falsely identified variably methylated CpGs. A. Percentage of CpGs that are concordant among three CD4M methylomes and three CD4N methylomes at different cutoffs. B. Pearson correlation among three CD4M methylomes and three CD4N methylomes. C. VMR number at different methylation difference (maximum methylation minus minimum methylation) and the percentage of VMRs from CD4M and CD4N comparison in the total number of VMRs detected from 58 methylomes.

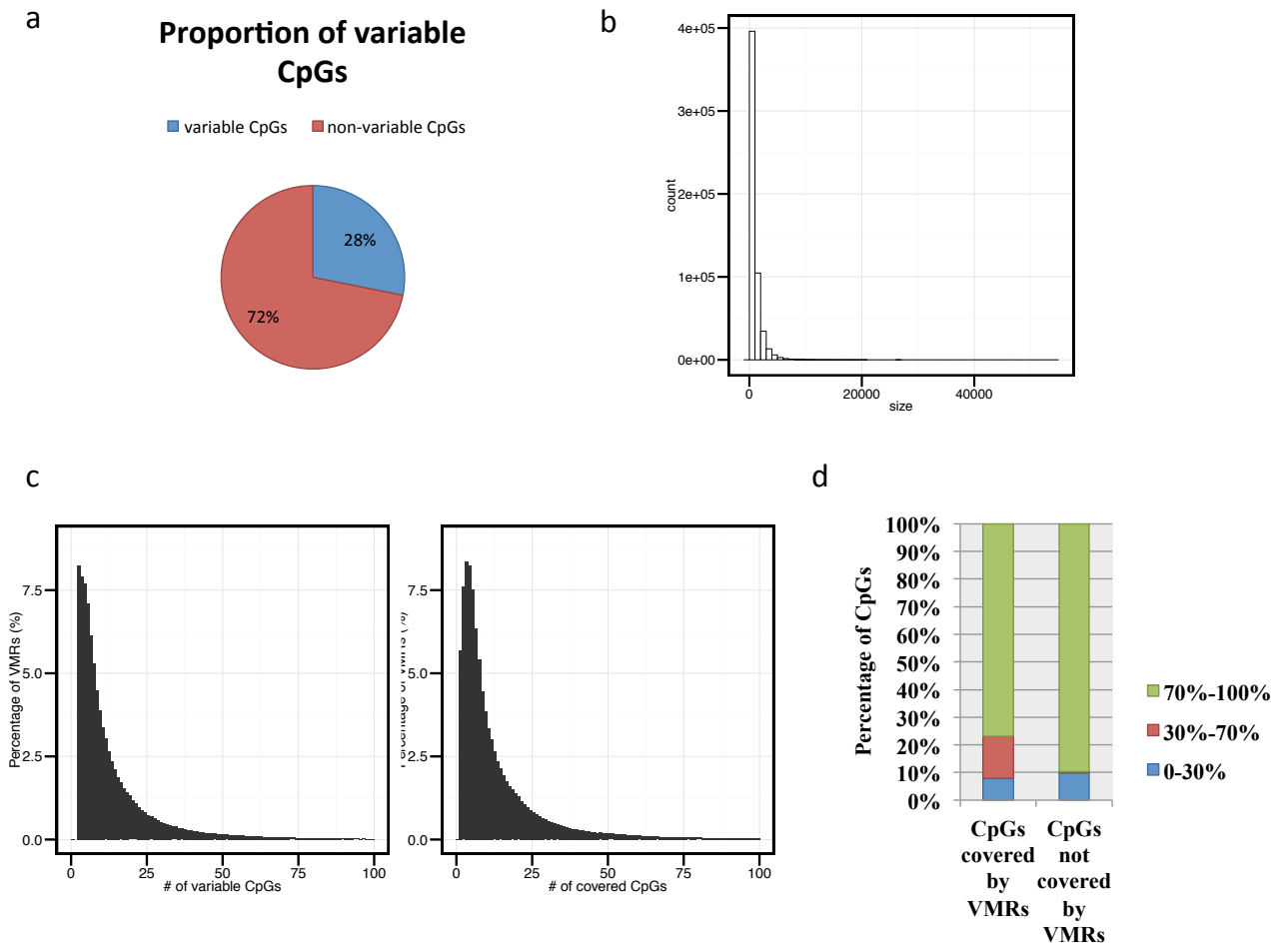


Figure S4: Characterization of variably methylated regions. A. Histogram of sizes of variably methylated regions (VMRs). B. Histogram of the number of CpGs covered and merged VMRs. C.Average methylation level for each CpG in VMRs or non-VMRs.

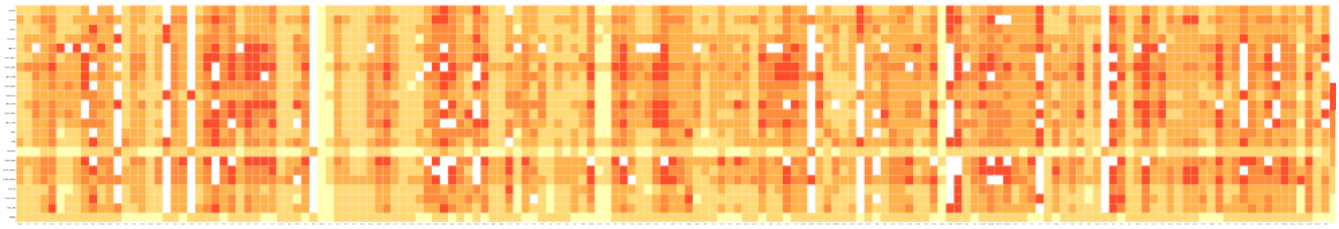


Figure S5: Enrichment of TFBS signal over hypomethylated VMRs.

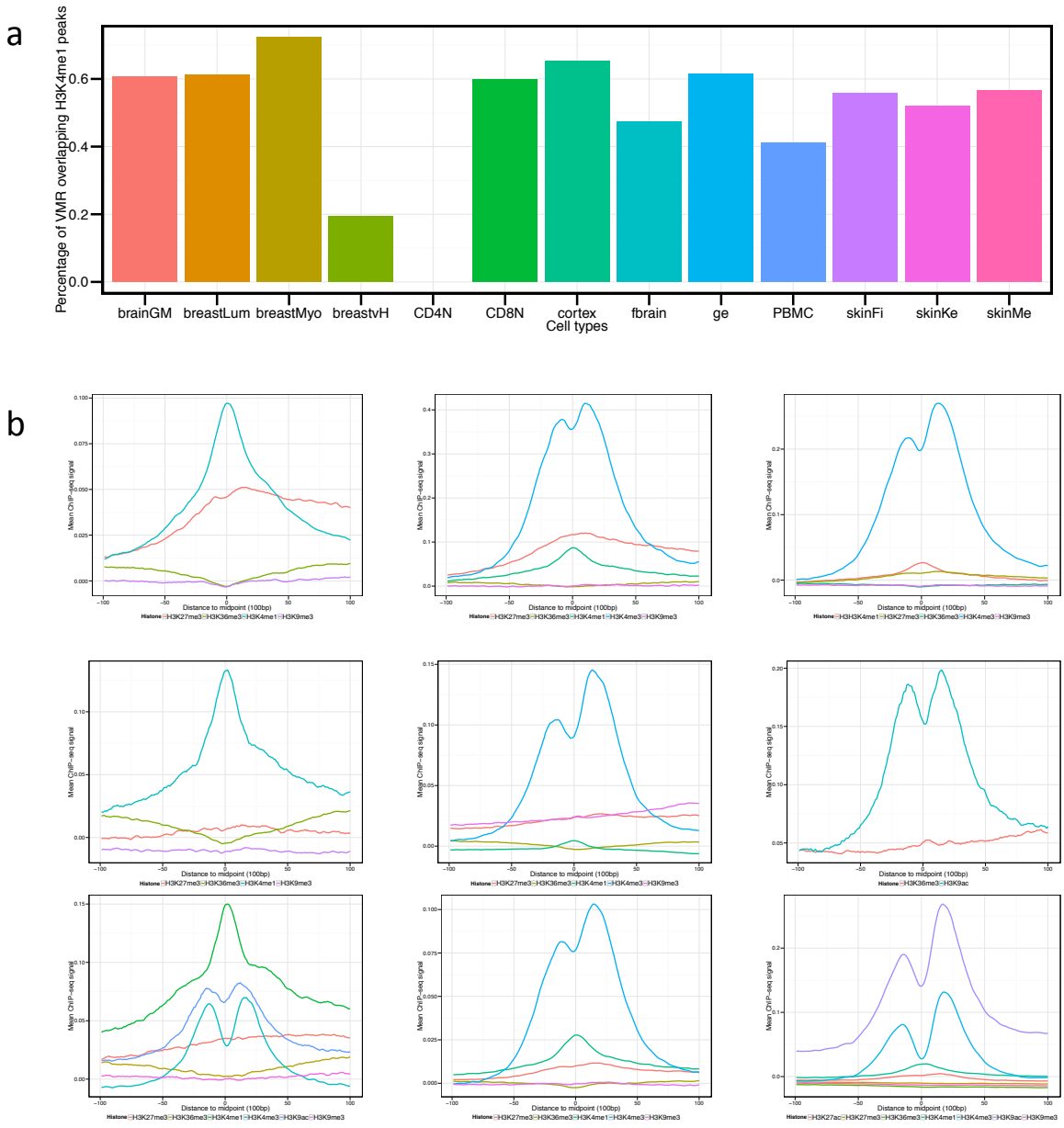


Figure S6: Tissue specific hypoVMRs enrich enhancer or active transcription marks. A. Percentage of tissue specific hypoVMRs overlapping H3K4me1 peaks. B. Mean ChIP-seq signal over 10kb regions centered on the middle point of VMRs (Row 1: brain_ge, brain_GM, fbrain; Row 2: breast_lumepi, breast_vH, hemo_cd4n; Row 3: hemo_cd8n, skin_fibroblast, skin_keratinocyte.

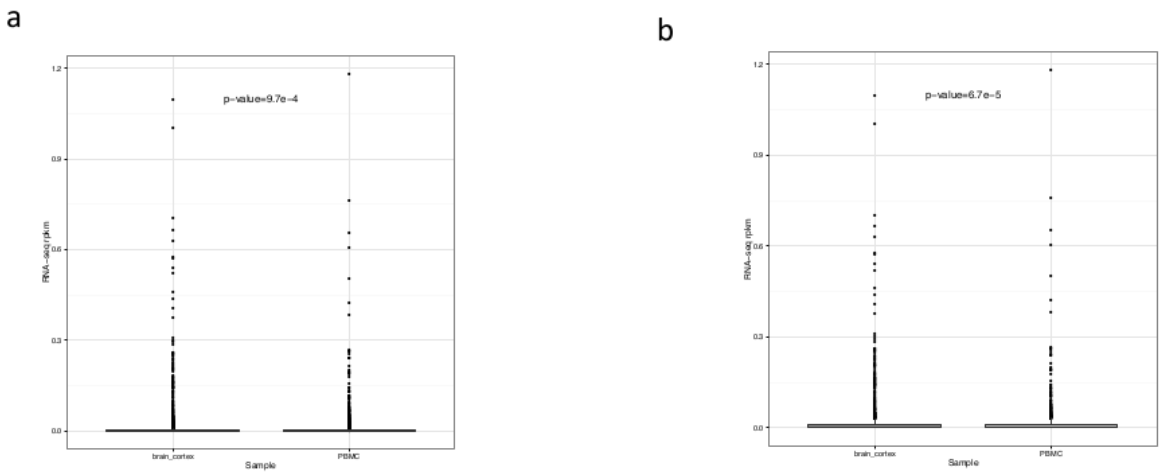


Figure S7: Hypomethylated VMRs correlates with elevated nearby gene expressions. A. RNA-seq RPKM for genes located in 10 kb regions surrounding VMRs B. Same as a. except genes with no expression in both cell types were excluded

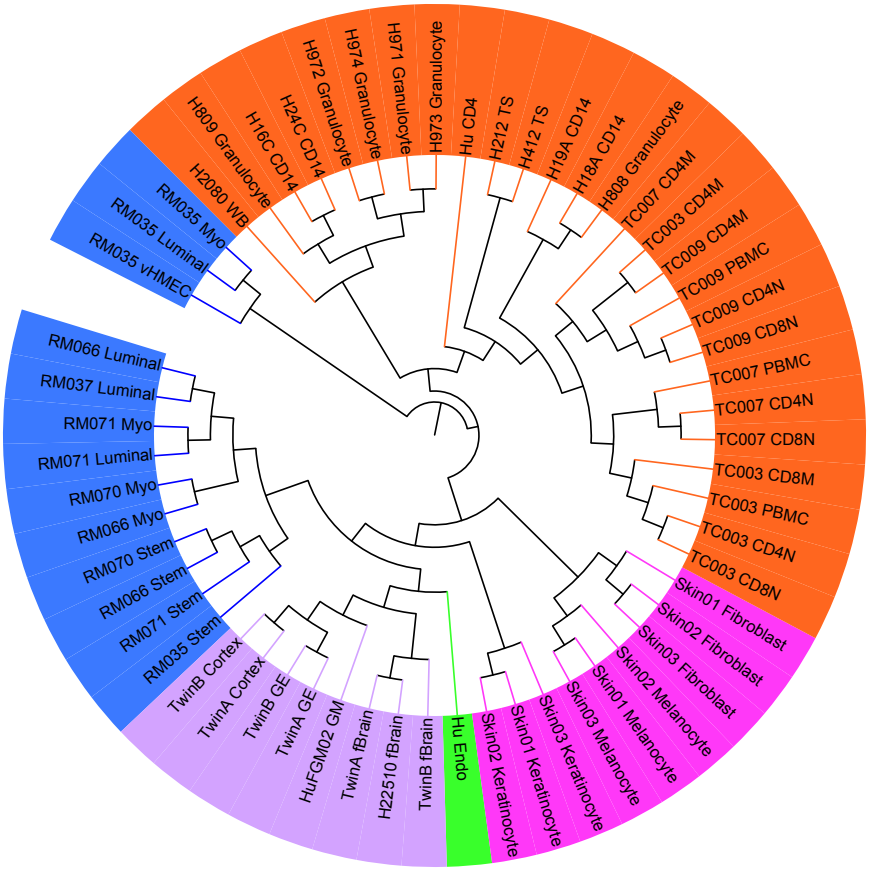


Figure S8: Clustering of samples based on average methylation of VMRs

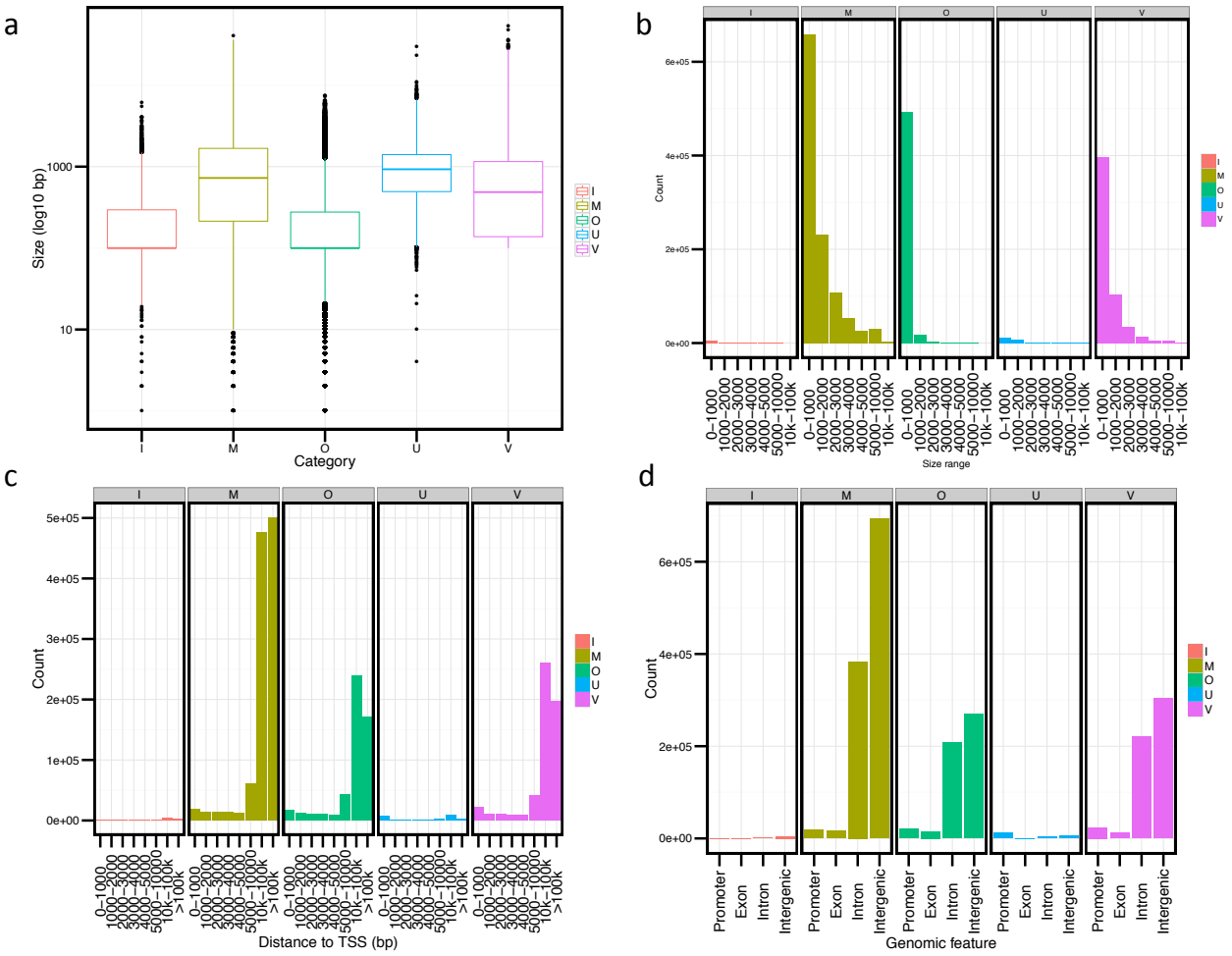


Figure S9: Characterization of each categorical regions. A. Size distribution of different categories of regions. B. Same as a but shown in size ranges. C. Distribution of distances of region to nearest TSS. D. Distribution of each categorical region with respect to genomic features

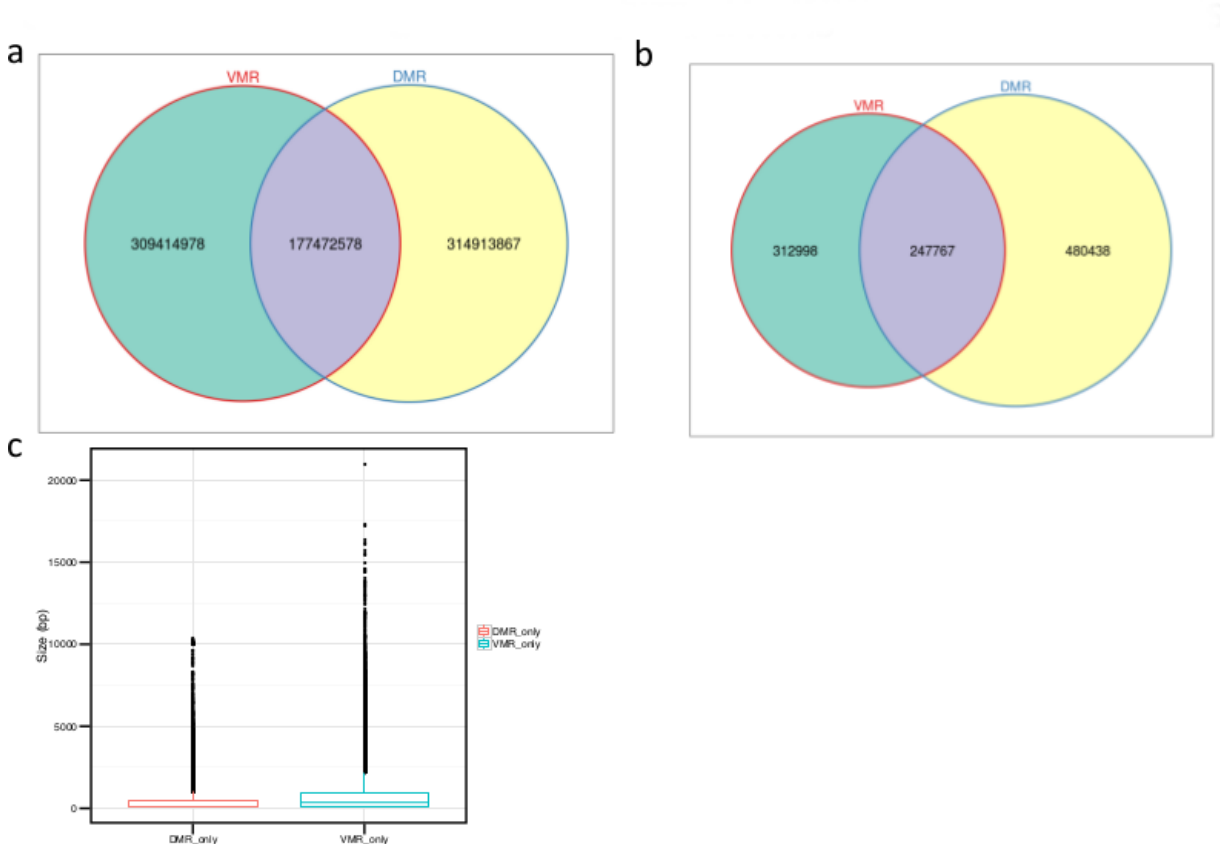


Figure S10: Comparison between methylCRF predicted VMRs and WGBS predicted DMRs. A. Overlap between bases covered by VMRs and those by DMRs. B. Overlap between regions of VMRs and DMRs. C. Size distribution of VMR-specific and DMR-specific regions. D. Number of DMR and VMR with single CpG. E. Median read coverage for each WGBS libraries on CpGs in single CpG DMRs or non-single CpG DMRs. F. Overlap between bases covered by non-single CpG VMRs and DMRs and overlap between regions of non-single CpG VMRs and DMRs

Chapter 6

Detailed analysis of methylCRF and WGBS concordance and resolution

(he) was serious; – he was all uniformity; – he was systematical, and, like all systemstick reasoners, he would move both heaven and earth, and twist and torture every thing in nature to support his hypothesis. In a word, I repeat it over again; – he was serious.

-Laurence Sterne, The Life and Opinions of Tristram Shandy, Gentleman

6.1 Introduction

methylCRF [77] and WGBS are the only two whole genome, single-CpG assays of methylation currently available. WGBS is considered the standard for assaying methylation. This was conveyed consistently in reviewers comments to the methylCRF manuscript. Since methylCRF is up to 15 times cheaper than WGBS, its widespread use has potential to much more quickly expand the number of assayed methylomes than WGBS given the same field-wide expenditure. Interest in DNA methylation continues to increasingly move toward center stage in a variety of areas including central societal concerns such as epigenetic inheritance, cancer, and many

environmental factors in disease. As epigenetic editing is starting to become possible, clinical and pharmaceutical interest in DNA methylation is creating excitement. Therefore it is critical to asses whether methylCRF and WGBS can be used interchangeable to further our understanding of methylation.

6.1.1 Review of WGBS

From the reviewer's responses to the methylCRF manuscript, it appeared to us that our perspective of WGBS reliability and accuracy was not shared in general. So, we review some reasons for our concerns. We start by briefly reviewing a few details about the generation of WGBS libraries and then analyze the possible effects.

Generating WGBS Libraries [6] provides an in-depth analysis of issues involved with WGBS libraries, we summarize a few of them here:

- A large percentage of high-ranking WGBS DMRs overlap with repetitive regions. This could indeed reveal a very interesting unexpected aspect of biology. However repetitive regions are difficult to confidently align next generation sequencing short reads to as seen by the high percent of reads mapping to multiple locations.
- Spike-in controls show evidence of incomplete bisulphite conversion of unmethylated C's to T's ($< 99\%$)and, surprisingly, over-conversion of methylated C's ($> 1\%$) as well. This results in a complicated scenerio resulting in reads where some methylated C's become T while some unmethylated C's remain C. Additionally, bisulphite treatment degrades DNA preventing their amplification which may provide some unknown selection bias for some sequence feature.
- Sequencing of constitutively methylated adapter and constitutively unmethylated C's added to reads during end-repair.

The authors mention that the second two can be somewhat reduced by "aggressive" adapter trimming. However, trimming reduces the length of reads which may induce more multi-mapped reads and thus reduce coverage. It has not yet been examined whether the first point

is reflective of biological dynamics of methylation or whether it reflects some experimental artifact of the protocol, the sequencing, or the analysis pipeline.

Variable GC-bias in Illumina Reads The existence of biases due to GC-content (the ratio of G's and C's to A's and T's) in sequencing protocols and how it has changed over time has been well document [1, 29, 71, 2, 10, 16, 11, 59]. Because of the bisulfite induced C to T conversion prior to PCR amplification of fragments, estimated WGBS methylation values might be uniquely sensitive to this bias Fig. 6.1(left). Unlike most short read alignment applications where only the number of reads is used, it is not clear how to correct for this as WGBS relies on the C to T ratio, and so the GC content, to estimate methylation levels. However correction methods have been proposed for targeted sequencing [56] suggesting that it may be possible. The bias appears due to GC content of the fragments rather than the parts of the fragments that are sequenced, suggesting PCR is the source of bias and both AT-rich and GC-rich are under-represented [4].

Additionally, after library preparation, reads with high GC content, regardless of their fragment GC content, may also be depleted due to differential cluster formation on 'sequencing-by-synthesis' based sequencing [1]. Illumina's new TruSeq cBot kit appears to better balance AT and GC rich regions [32] mediate this bias for newer libraries. However as recent as 2011 using Illuminas TruSeq WGBS kit, GC content maybe be positively correlation with coverage [55].

6.2 Results

6.2.1 Comparing methylCRF and WGBS

methylCRF has more CpGs with extreme values than WGBS where concordance is higher

First, we compare our original methylCRF and WGBS data sets[77, 50], we used for the original manuscript. The concordance across methylation levels shows a pattern of discordance

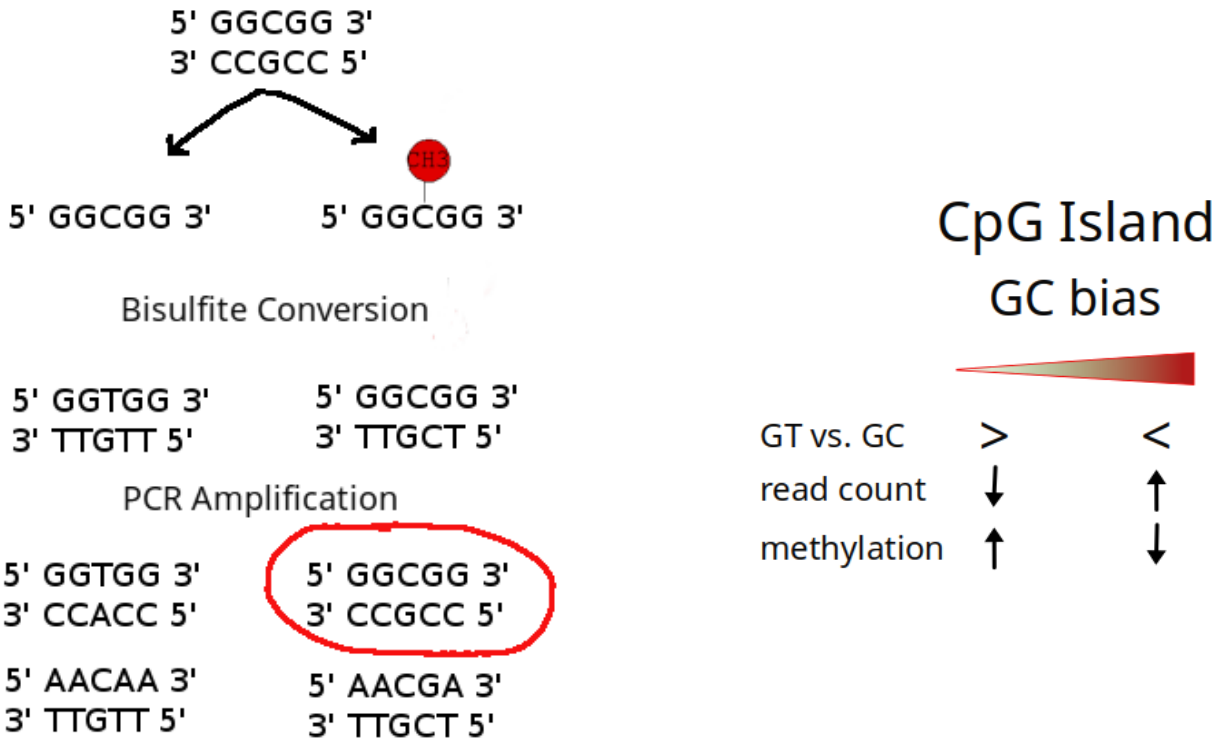


Figure 6.1: Effect of GC bias. Left: In the WGBS protocol, fragments are first denatured. Bisulfite is then applied, converting unmethylated C's to T's. The fragments are then amplified with PCR. If there is GC-bias in amplification, then this may introduce a bias such that the ratio of methylated and unmethylated chromatids differ from what is sequenced. Right: GC dense regions, like CpG Islands should show read counts that are proportional to GC bias and methylation that is inversely proportional to GC bias.

BS	CpG	=	0.1	+-0.25
0.0	382k	0.30	0.67	0.68
0.1	138k	0.46	0.73	0.76
0.2	084k	0.04	0.12	0.64
0.3	086k	0.05	0.08	0.38
0.4	146k	0.02	0.05	0.13
0.5	263k	0.02	0.04	0.13
0.6	458k	0.02	0.05	0.35
0.7	1.1M	0.03	0.07	0.93
0.8	4.4M	0.22	0.59	0.99
0.9	8.0M	0.77	0.97	0.99
1.0	2.3M	0.12	0.73	0.99

Table 6.1: methylCRF/WGBS concordance across rounded methylation deciles.

for CpGs with WGBS values, 0.2-0.75, Fig. 6.2. However, the majority of discordance, by CpG count, is in the WGBS range 0.6-0.8 where methylCRF tends to be 0.8-1.0, Fig. 6.3.C. For the purpose of this manuscript we refer to these regions as *intermediate*, *shifted*, and *high* for 0.2-0.75, 0.6-0.8, and 0.8-1.0 methylation respectively as well as another region, *low*, for 0.0-0.1.

WGBS has 2.8M CpGs in the shifted region, 94.5% of these have methylCRF values outside of this range. These CpGs tend to sit in regions of high WGBS methylation Fig. 6.4.A. Note that if this discrepancy was due to methylCRF and WGBS disagreeing on the exact location of boundaries between regions of high and low methylation -as was suggested in [77] Supplemental- you would expect the average window methylation to instead be close or lower than the CpG.

Additionally, 17% of the low methylCRF values have intermediate WGBS values Fig. 6.3.C. These results show that WGBS has more intermediate methylation than methylCRF. Given the bi-model distribution of WGBS methylation values and methylCRFs statistical nature, we were concerned that methylCRF was suffering from a type of class imbalance type issue in training.

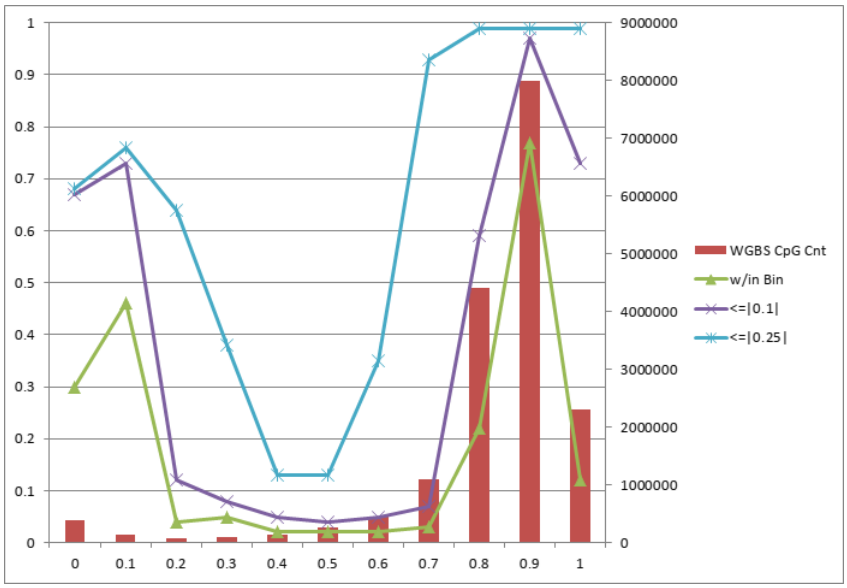


Figure 6.2: Concordance between methylCRF and WGBS methylation values for H1ES. The barplot shows the number of WGBS CpGs in each decile. The green line shows the percent of those with methylCRF in that decile. Purple and blue show the percent with methylCRF values within a window of 0.10 and 0.25, respectively. Less then 15% of the WGBS CpGs from 0.2-0.7 have methylCRF values within 0.1, and within a 0.25 window, the concordance for WGBS CpGs between 0.3-0.6 is below 40%. Interestingly, 25% WGBS CpGs from 0.0-0.2, have methylCRF values more than 25% greater.

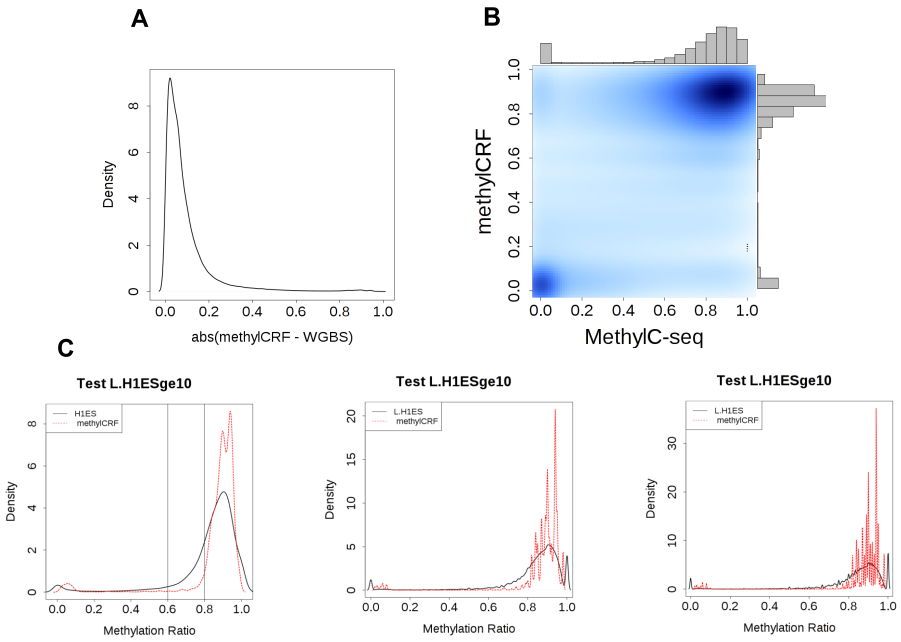


Figure 6.3: A: per CpG difference between methylCRF and WGBS values shows that the vast majority of discrepency within 0.2 methylation. B: Kernalized scatterplot of the same data showing that the discrepancy is tends to be CpGs WGBS calls 0.6-0.8 versus 0.8-1.0 in methylCRF. C: Left: Density plot of WGBS and methylCRF methylation values across the genome. Above 0.6, methylCRF tends to concentrate values in the 0.8-1.0 region more than WGBS. Additionally, this reveals a tendency of methylCRF to concentrate values from 0.0-0.1 closer to 0 than WGBS. Middle/Right: This pattern is robust to choice of kernel width.

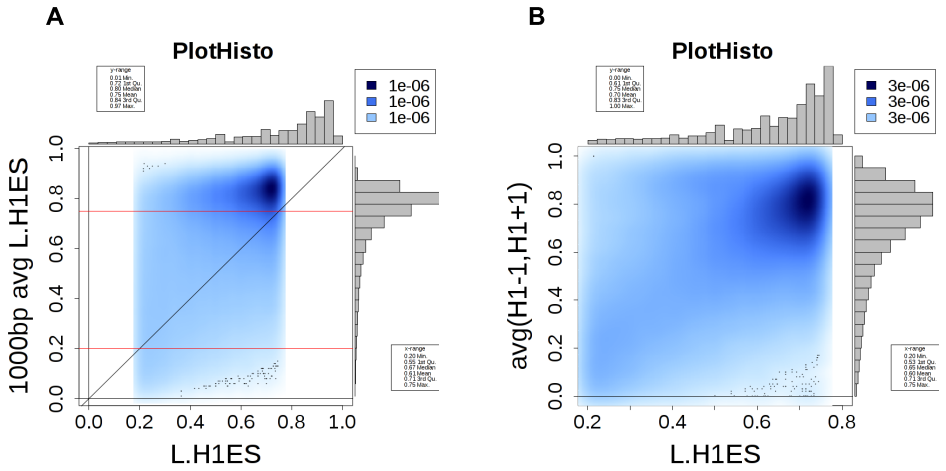


Figure 6.4: A: Density plot of WGBS values between 0.2 to 0.75 on the x-axis and the average methylation of 1kbp windows centered on each CpG. The majority of these CpGs sit in 1000bp windows of average 0.8 to 1.0 methylation. The two red lines highlight 0.2 and 0.75 methylation and the black line is has slope 1 representing equal values between CpGs and their 1kbp window. B: Similar plot for CpGs with 2 neighbors within 50bp (24%).

Low concordance in intermediate methylation between methylCRF and WGBS

A detailed look at the concordance rates with 10% and 25% windows, Table 6.1 supports that the concordance between methylCRF and WGBS is uniformly the lowest between 0.2 and 0.75 methylation. We chose 0.75 because the bins are rounded.

Majority of intermediate WGBS CpGs are in windows of high WGBS methylation. In order to examine this region, we make the assumption that intermediate WGBS CpGs within a 1kbp window centered around them with an intermediate average methylation are *likely* intermediately methylated, otherwise, we term them *questionable*. . There are 2.1M intermediate CpGs with WGBS values and at least 10 reads [77], of which only 6% are likely. Note that 98% of the CpGs have 1kbp windows with at least 3 CpGs. 69% of intermediate CpGs are in windows with high methylation and 1% in windows with ≤ 0.2, Fig. 6.4.A. This pattern holds when considering CpGs and their neighbors within 50bp, 24%, Fig. 6.4.B

Of the 6% likely intermediate CpGs, 22%, 135k CpGs, have methylCRF values in this range, we call these *MM*. We can conclude that according to this definition, the concordance on determining true intermediate methylation between WGBS and methylCRF is over 20%. However, note that the positive predictive power of WGBS to predict intermediate methylation at single-CpG resolution is only 30%.

Low methylCRF CpGs in windows of intermediate WGBS suggest methylCRF may lack fine resolution. Of the 519K CpGs low methylCRF CpGs (L), 17% are in windows of intermediate WGBS methylation,(LM). LM CpGs are more likely to occur in CGI's and repeatmasker RNAs and less likely to occur in SINEs, LINES, LTRs, and exons than MM CpGs, (1.1'x, 1.7'x, 0.2'x, and 0.2'x, 0.5'x, and 0.5'x respectively). The MeDIP-seq and MRE-seq values suggest that the LM CpGs are a distinct group within the L group, Fig. 6.5 with consistently higher methylation. Nonetheless, LM CpGs are much closer in methylation to the L group than the MM group.

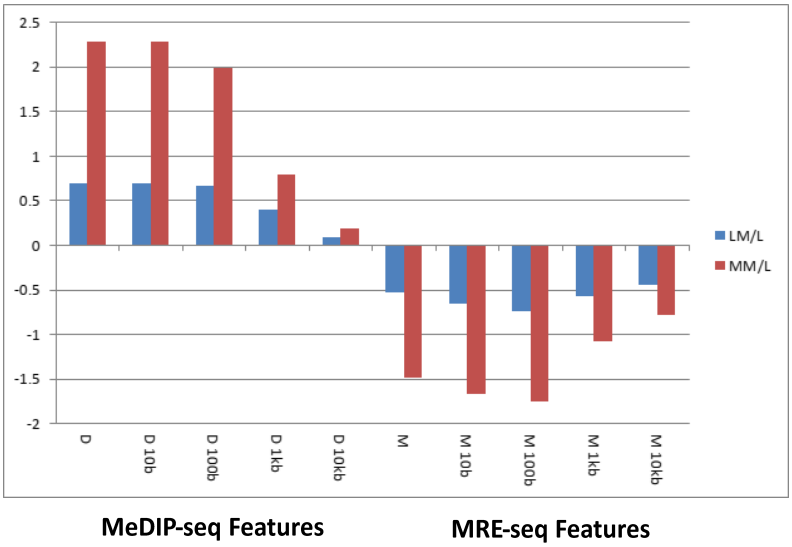


Figure 6.5: Ratio of LM/L and MM/L methylCRF values for methylcRF MeDIP-seq (5 left) and MRE-seq (5 right). The experimental data supports that the LM are indeed a separate group from the L group with higher MeDIP values and lower MRE values suggesting these are CpGs with higher methylation. However, they suggests significantly lower methylation than CpGs called intermediate by methylCRF.

Key	Sample	methylCRF	SBS	RRBS
E03	Fetal Brain	x		x
E10	H1 (Cell Line)	x	x	x
E11	H9 (Cell Line)		x	x
E17	Mobilized CD34		x	x
E38	CD184+ (hESC Derived)		x	x
E41	Neurosphere Ganglion Emenence	x	x	

Table 6.2: Epigenome Consortium Methylome Platform Replicates

At biologically motived features, RRBS, WGBS, and methylCRF show widespread but minor discordance.

The incredible resource generated by the NIH Roadmap Epigenome Consortium [5] provided us an opportunity to look at whole genome methylation assays more comprehensively. We were able to compile 58 methylCRF, WGBS, and RRBS methylomes across a variety of tissues and cell-types. There are six cases where multiple assays were performed on the same samples Table 6.2 -one with all three. To focus on critically on regions with potential for the most biological insight, we used ChromHMM [25] histone mark-based enhancers. ChromHMM defines enhancers as a characteristic combinatorial histone code by combining any number of histone ChIP-seq assays. Of the CpGs with values in all of the libraries, we found 50k CpGs that overlapped a ChromHMM enhancer from any of the libraries with ChromHMM annotation.

Using this enhancer CpG set, we performed hierarchical clustering across all 58 data sets. While some of the pairs cluster together (E17, E38), we expected all of them to Fig. 6.6(left). We next checked the clustering using instead annotated CpG Islands -which show some of the highest concordance between WGBS and methylCRF in [77]. However, the results were similar, Fig. 6.6(right).

methylCRF libraries are more self-similar than WGBS. Despite the platform base clustering of leaf nodes, the branch distance at leaves appeared relatively small suggesting the result may not be robust to small differences in the distance statistic. In fact, the Neurosphere (E41) methylCRF and WGBS pair's closest pair-wise libraries reveal that methylCRF libraries

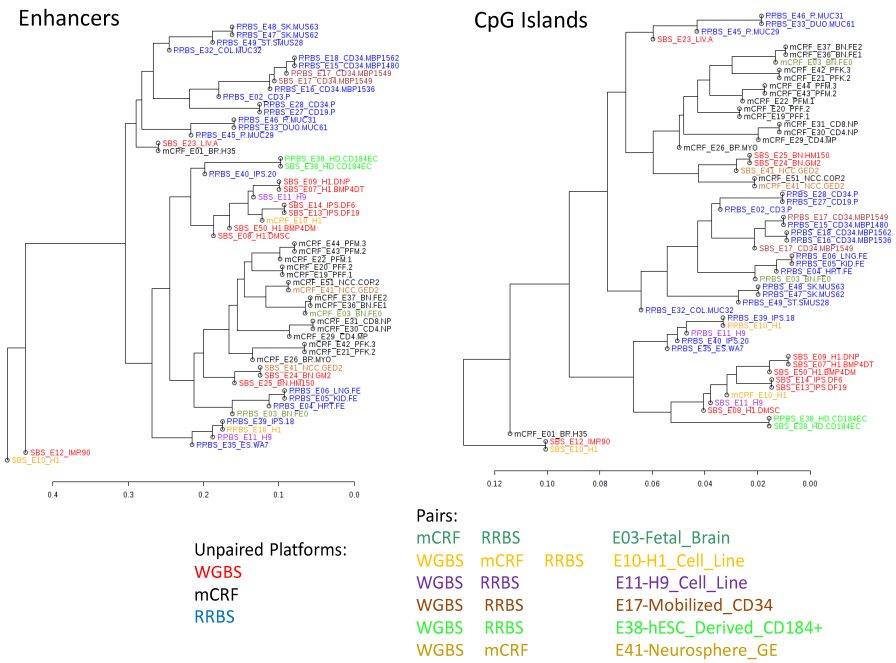


Figure 6.6: Hierarchical clustering of 58 methylomes (methylCRF, WGBS, RRBS) reveals platform bias. Additionally, libraries on the same cell-type using different platforms are not guaranteed to cluster together, although paired WGBS and RRBS are more likely to cluster together than either with methylCRF. Left are CpGs in ChromHMM-defined enhancers , and Right are CpGs within CpG Islands.

	SBS_E41_NCC.GED2
SBS_E41_NCC.GED2	1.00
SBS_E24_BN.GM2	0.96
SBS_E25_BN.HM150	0.94
mCRF_E41_NCC.GED2	0.94

	mCRF_E41_NCC.GED2
mCRF_E41_NCC.GED2	1.00
mCRF_E51_NCC.COR2	0.97
mCRF_E36_BN.FE1	0.95
mCRF_E03_BN.FE0	0.94
mCRF_E37_BN.FE2	0.94
SBS_E41_NCC.GED2	0.94

Figure 6.7: Left: the WGBS and methylCRF libraries closest to the E41 WGBS library using concordance within a 25% window. Right: similarly for the methylCRF E41 library. Interlibrary distance between methylCRF is in general smaller than between WGBS libraries which prevents WGBS and methylCRF technical replicates from clustering together.

are in general slightly more self-similar than WGBS libraries are Fig. 6.7. While WGBS is methylCRFs fifth closest library (third closest for WGBS), they are nonetheless 94% concordant. A similar pattern is seen between RRBS and methylCRF fetal brain (E03) pair. The clustering was performed using 25% concordance. Other distant metrics, like correlation and euclidean distance we tried gave similar results. The clustering result is due to rather small differences in concordance and may not be the best way to test the quality of these assays globally.

Highly discordant enhancer CpGs have proportionally higher WGBS values and lower methylCRF methylation At our enhancer set in the E41 methylCRF/WGBS pair, the vast majority of discordant CpGs are within 0.1 methylation Fig. 6.8. However, for CpGs greater than 0.1 discordant, the WGBS methylation values are proportional to the amount of discordance, while for methylCRF the methylation values are inversely proportional. Note that as the concordance increases, methylCRF values approach the distribution for all enhancers while at the highest concordance WGBS values are distinctly methylated. Also note that CpGs below 0.10 discordance actually have intermediate methylation in both WGBS and methylCRF. In the case of WGBS, it is even higher than CpGs that are up to 0.25 discordant. The inversely proportional relationship between discordance and methylation for methylCRF is supported by lower MeDIP-seq values and high MRE-seq values suggesting that these might be CpGs which are problematic for WGBS.

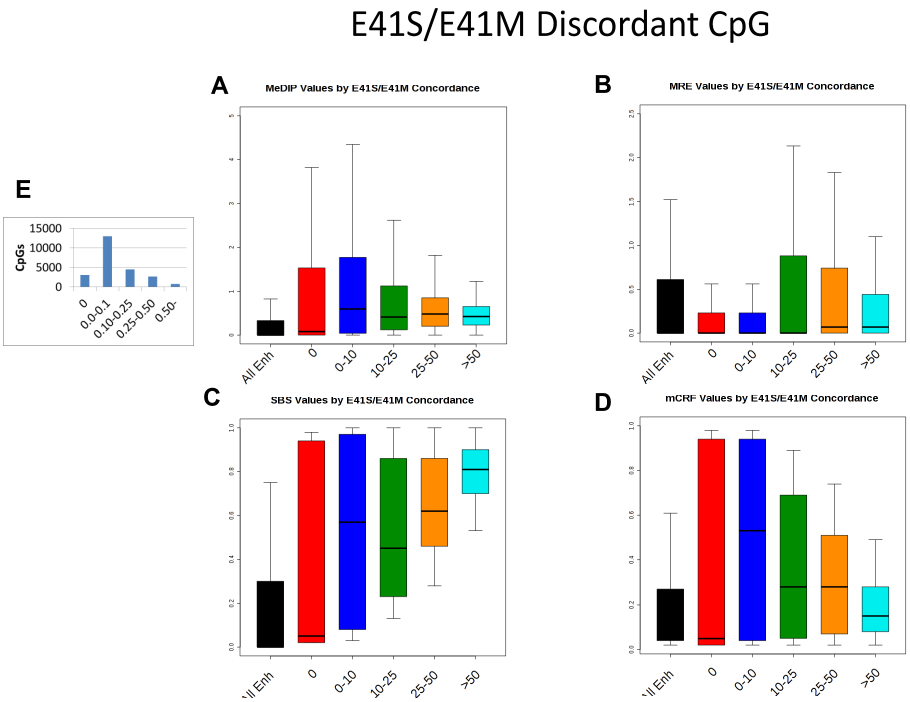


Figure 6.8: Comparison of enhancer CpGs assay scores between E41 WGBS and methylCRF by amount of discordance. A: MeDIP-seq values, discordant CpGs tend to have higher average values. B: MRE-seq value, CpGs more 25% discordant have slightly higher average values. C: WGBS methylation values, concordant and slightly discordant have a wide variation of values, while discordance greater than 10% is proportional to methylation values. D: methylCRF methylation values, CpG 0-10% discordant are similar to WGBS, however discordance greater than 10% is inversely proportional to methylation. Note that these CpGs have both higher MeDIP and MRE value on average. E: the number of CpGs in each discordance bin showing most CpGs are 0-10% discordant.

Highly discordant enhancer CpGs are enriched in UTR3, CGI-shores, and un-annotated regions At these CpGs, discordance is proportional to enrichment in CGI-shores. Above 0.1 discordance, un-annotated regions are high. While UTR3's show the same trend, there is a large drop above 0.5 discordance. SINEs are enriched for less than 0.1 discordance, but depleted in large discordance -suggesting that SINEs are 'noisy' in some sense, but large discrepancies between WGBS and methylCRF are relatively few. Intergenic CpG Islands are enriched for 0.0 methylation in both WGBS and methylCRF and depleted in large discordance. Note that all enhancer CpGs are enriched in SNPs and are most enriched in largely discordant CpGs. Also of note, 20% of CpGs in this enhancers set fall in 2kb regions flanking CpG Islands. This opens a question of whether CpG Island shores are in fact enhancers or whether it might be advantageous for ChromHMM states to include these as a sub-type. Also note that due to the large number, these regions might play a large part in the platform first hierarchical clustering.

WGBS has higher coverage at CpGs where WGBS and methylCRF disagree. On average, WGBS has 58x's coverage at enhancers. However, at enhancer CpGs where methylCRF and its closest four libraries agree on methylation Fig. 6.7, the E41 WGBS has 70x's coverage at CpG-Islands and 46x's coverage otherwise. This raises the possibility that of a relationship between coverage and WGBS methylation estimates.

Higher variance in WGBS than methylCRF at CpG Island shores [33]. Browser shots of CpG Island shores seem to indicate that neighboring CpGs are less correlated in WGBS than methylCRF in these regions, Fig. 6.9. In all four cases, methylCRF tends to call these CpGs uniformly lowly methylated while WGBS shows both higher methylation and much greater variation in methylation. This is entirely, consistent with the divergent relationship between methylation and discordance in highly discordant CpGs in Fig. 6.8.C and D.

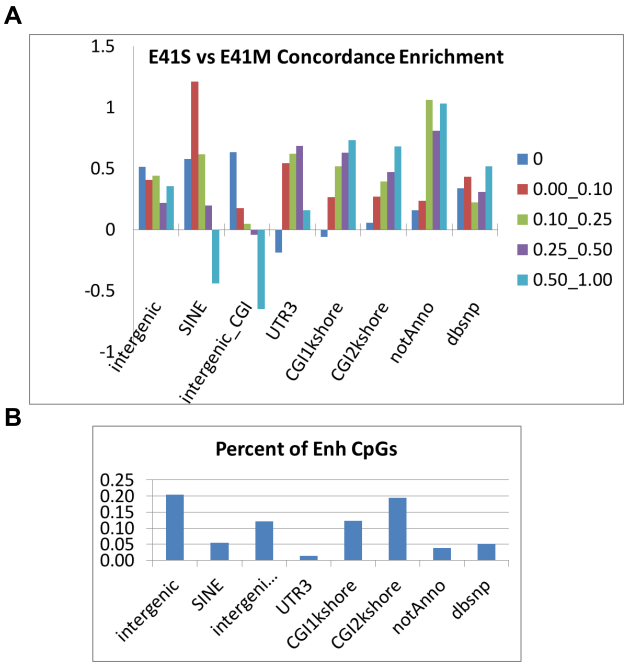


Figure 6.9: Discordant CpGs buy genomic annotation. A: enrichment by annotation feature. B: percent of discordant CpGs annotated with each feature. SINEs show enrichment for low discordant CpGs but depletion for highly discordant CpGs suggesting that SINEs are a source of low discordance. High discordance is also depleted in intergenic CGIs. Highly discordant CpGs seem to be concentrated in CGI-shores, UTR3, and not annotated CpGs. Since shores make up a large number of enhancer CpGs, this suggests CGI-shores are the most problematic in terms of WGBS/methylCRF concordance.

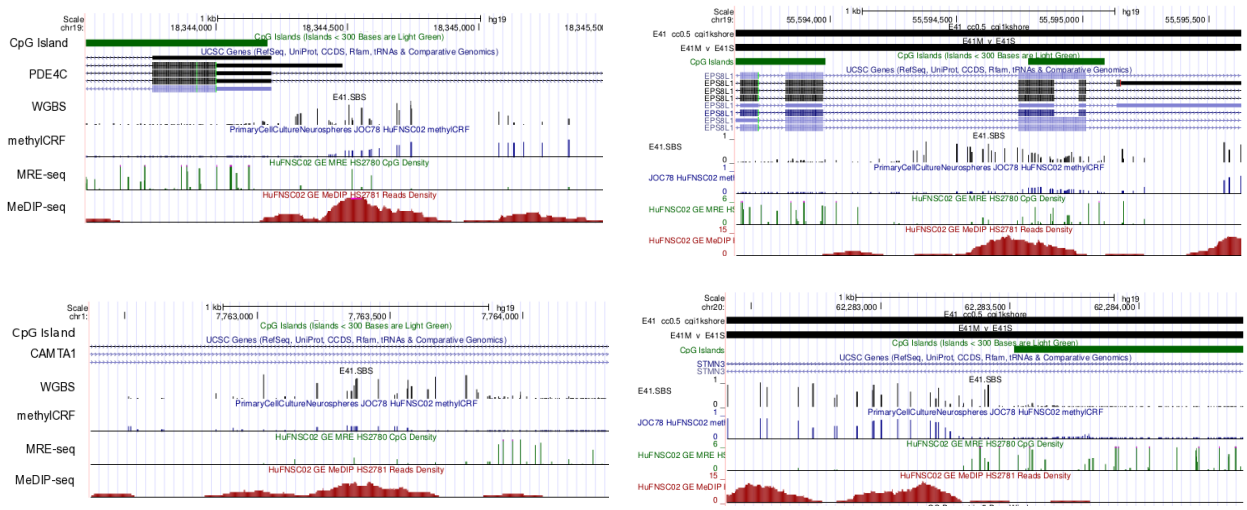


Figure 6.10: Four CpG Islands showing highly discordant CpGs between WGBS and methylCRF. Generally, WGBS shows less correlation between neighboring CpGs than methylCRF in these regions. Additionally, methylCRF tends to have lower methylation values than WGBS which is consistent with divergent methylation patterns in highly discordant CpGs in Figure 7 C and D.

6.2.2 Analyzing WGBS

Extensive analysis of methylCRF is provided in [77]. The above analysis of the differences between methylCRF and WGBS prompted us to examine WGBS in more detail.

Coverage and GC-bias affect determination of unmethylated regions

For this analysis, we used 33 of the WGBS library set of normal tissue used in [90] from which the authors estimate that 22% of the CpGs in the human genome have variable methylation. In order to examine the effect of coverage and GC-bias, we chose to look at unmethylated regions (UMRs). These represent a small fraction of CpGs Fig. 6.3 and Table 6.1 and tend to appear in groups [77] and have been the main subject of interest in terms of methylation's regulatory potential since DNA methylation was discovered.

3ML conservative estimation of UMRs To identify UMRs, we sought to define a set that would be robust to modeling choice so as to make the results as generally applicable as

possible. We therefore looked for a conservative set of UMRs. We developed a method, we termed *Moving Mean/Median Locater* (3ML). This is an exceedingly simple and intuitive model applicable to a wide-range of sequence or time-series-based problems. Conceptually, we combine benefits of 1) a moving mean filter, as a robust indicator of low-regions of the signal, and 2) a moving median filter, as a way to compensate for the smoothing effect of the moving mean to more accurately find the boundaries of regions of low signal. Specifically, we find all windows of the moving median filtered signal below a given threshold and filter out any regions which do not contain a CpG from the moving mean filtered signal below the threshold. For this analysis, we used a threshold of 0.2 methylation and a moving window size of 1kbp -which corresponds to an average sized *differentially methylated region* (DMR) [6]

To give some intuition into the suitability of this approach, we demonstrate the effect of each component on chromosome 1 of one of the above WGBS sets. Using mean windows only found 3495 UMRs (3.4Mb), while median windows alone found 5153 UMRs (4.9Mb). Requiring both to be true, resulted in 3492 UMRs (3.4Mb), while 3ML resulted in 3373 UMRs (4.9Mb). Thus 3ML was able to find fewer UMRs than mean alone but with 35% larger average UMR size. Note also that the use of the mean criteria significantly reduces the number of UMRs determined by median alone, suggesting that the median filter only set includes many short UMRs. Since these are relatively large windows of CpGs with median values below 0.2 and at least one CpG with a mean below 0.2, we conjecture that any method used to determine UMRs would include these regions. In this way, we claim that this is a conservative and so universal set of UMRs.

UMR count is confounded by coverage and GC-bias We ran 3ML on the 33 WGBS libraries using 1kb moving windows and a threshold of 0.2 to find UMRs in each library. The number of UMRs ranged from 20-60 thousand showing a 3-fold difference from lowest to highest UMR count. These libraries range from 3-51x's coverage. There is strong negative correlation between UMR count and coverage (r^2 0.39), Fig. 6.11 A. Additionally, there is a correlation between ratio of coverage at UMR's and over-all coverage (r^2 of 0.08), when corrected for over-all coverage, the r^2 rises to 0.09, Fig. 6.11 B. The is a similar correlation between UMR and GC% within 75bp of each CpG and its read count (r^2 0.08), Fig. 6.11 C. The UMR coverage ratio is very strongly correlated (r^2 0.85) with the correlation between

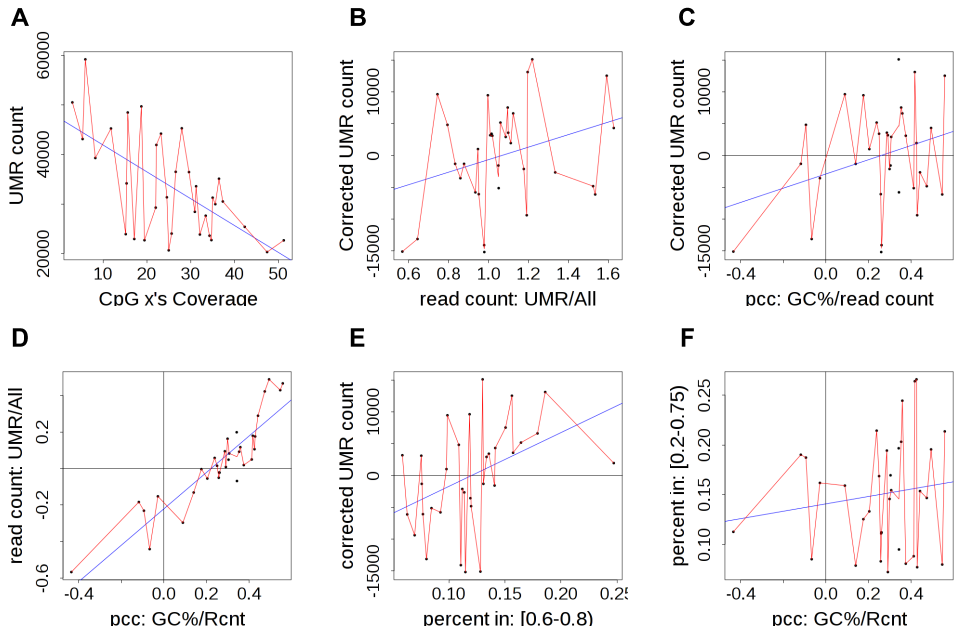


Figure 6.11: Bias in WGBS libraries. A) UMR count is negatively correlated with library coverage. B) UMR count corrected for library coverage is correlated with the ratio of coverage at UMRs and whole coverage. C) UMR count corrected for library coverage is also correlated with relation between GC% and read coverage. D) UMR read count ratios are highly correlated with the ratio of GC% and read count. E) UMR count corrected for library coverage is also the percentage of CpGs with values from 0.6 to 0.8 -although the percentage of CpGs in this range is not highly associated the ratio of GC% over read count (r^2 0.02). F) The percent of CpGs with methylation 0.2 to 0.75 slightly correlated with the ratio of GC% over read count.

GC% and read count, Fig. 6.11 D. This suggest that up to 9% of the variation in UMR count is due to GC% alone, even after coverage affects are accounted for.

Interestingly, the corrected UMR count is additionally associated (r^2 0.18) with the percentage of CpGs in the shift range between methylCRF and WGBS, 0.6-0.8, Fig. 6.11 E. However, correlation between GC read count bias and percent of CpGs with values in this region is lower (r^2 0.02), suggesting that there are additional factors effecting both UMR count and the distribution of CpG values. Intermediate regions, 0.2-0.75, show a similar relationship Fig. 6.11 F.

6.2.3 Theoretical and Empirical Single CpG, WGBS Variability

The high number shift region WGBS CpGs in windows of high methylation Fig. 6.4 prompted us to speculate whether we could determine whether these were likely noise or reflective of methylation in the cell. We were able assemble a group of WGBS datasets performed on human ES cells providing an interesting framework within which to study sources of variability at multiple levels. This set includes one methylome from HSFI (F1), one H9 (H9) [46], and three H1 (H1.1a,H1.1b,H1.2) [49] . For simplicity, we consider all five as methylomes as *biological replicates*, Fig. 6.12.A The three H1 methylomes include libraries from two different labs which we consider *technical replicates*. H1.1a and H1.1b are from two different analysis pipelines on the same library which we consider *computational replicates*. H1.1a is the methylome from [49] and H1.1b is from the NIH Roadmap Epigenome Consortium [5] used in the above analysis.

With a 10% threshold of difference in methylation, CpGs across biological replicates, pair-wise, are discordant on average 52% of the time. Technical replicate CpGs are discordant on average 44% of the time, while CpGs in the computational replicates are discordant 17% of the time. With a 25% threshold, the percent of discordant CpGs were 18%, 10%, and 3%, respectively, Fig. 6.12.B.

To help understand the source of the high discordance, we modeled the estimated methylation at each CpGs as draws from a binomial distribution. To focus on the effect of sampling, we make the simplifying assumption that each read has one CpG. Since we remove identical reads during alignment (likely to be due to over amplification), we can further assume that each read is from a different allele. As an example, consider a CpG for which 50% of the alleles in a sample are methylated. The distribution of methylation values if we sample 100 reads has low variance, Fig. 6.12. In fact, only 6% of such CpGs will have estimated methylation more than 10% different from the true methylation. However, if we sample only 10 reads, then 75% of CpGs will differ more than 10%. In contrast, short read-based methods that use the count of reads as their statistic (as opposed to the ratio), such as ChIP-seq or RNA-seq, have distributions that could be modeled as an expected value with normally distributed error. For example, then, for bases for which 10 reads corresponds to the true strength of some signal, in order have similar chances to have a value of 9 or 11 (here equating a 10% change

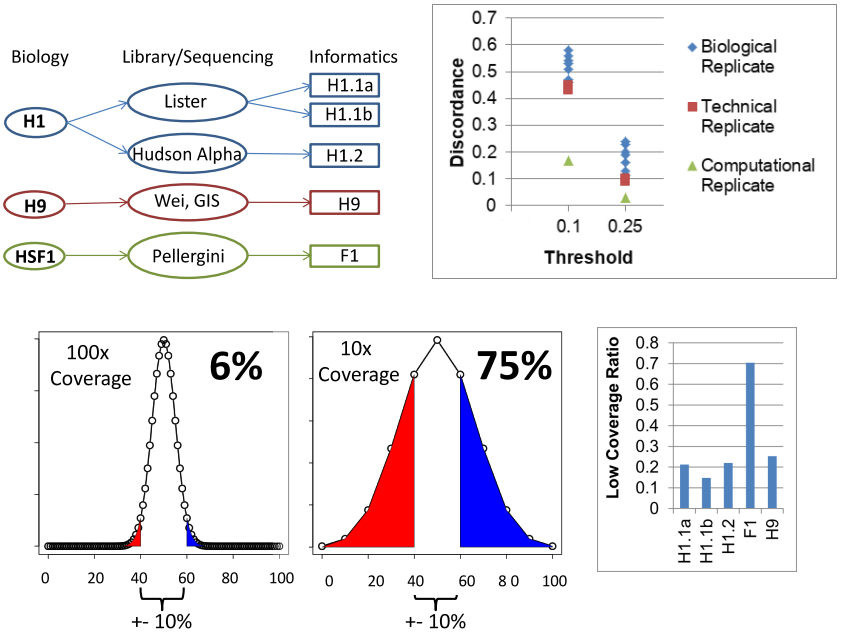


Figure 6.12: A) Samples used for this analysis represent three kinds of replicates: 1) biological: across different ESC cell lines (H1,H9,F1), 2) technical: one cell line across different labs (Lister, Hudson Alpha), and 3) informatics: across different processing pipelines using the same library (H1.1a, H1.1b). B) CpG discordance at 10% and 25% threshold for concordance for all three types of replicates. C) Binomial distribution of the ratio of methylated alleles from hypothetical CpGs that are 50% methylated, sampled 100 at a time. 6% of the CpGs are more than 10% discordant from their true value. D) Similar distribution where alleles are sampled 10 at a time showing that 80% of CpGs are greater than 10% discordant. E) The ratio of CpGs that have 10 or fewer reads in each library.

in methylation to a change of one read for a count-based readout), the standard deviation would have to be close to 4 reads.

On average, these ESC libraries average around 20% CpGs with 10 or fewer reads (except for F1 which is 70%), Fig. 6.12.E. Additionally, these libraries average roughly 25x's coverage -for CpGs with 25 reads, one would expect over 42% would be more than 10% difference. Note that this is quite close to the 44% average number of discordant CpGs for technical replicates Fig. 6.12. However, it does not fully explain the 52% for biological replicates and can not explain the 17% for computational replicates.

Empirical orphan analysis suggests coverage as a potential source of variance. In order to further examine the effect of rare values in a binomial distribution we focused on methylation *orphans*. Since CpG methylation is locally correlated [77], CpGs with large variation from their true methylation will tend to have neighbors with the same true methylation, however, they will tend to have smaller variation. That is, the CpGs with large variation will tend to look like outliers, or orphans, in windows of different average methylation Fig. 6.13.A. If there are many orphans, than either the local correlation structure itself varies in some way or these CpGs are bad estimates of their true methylation in the cell and may help explain the discordance between methylCRF and WGBS in the shifted range, Fig. 6.3.C.

We define orphans as being at least some threshold of methylation either above or below both its 5' and 3' neighbors as well as the average methylation in 400bp flanking the orphan. We additionally require the flanking regions to have at least 2 CpGs each and we call all CpGs with 2 CpGs in their flanking regions, orphan candidates. On average, 73% of CpGs are candidates. The prevalence of 10% orphans ranges from 10-20% of candidate CpGs, while 25% orphans range from 2-7% of candidates, Fig. 6.13.B In comparison, for methylCRF, only 1% of candidates are 10% orphans. Additionally, in every library 25% orphans had lower read count than average, Fig. 6.13 -only 56% of read count on average of candidates overall. Furthermore, the ratio of orphans detected in a library is strongly negatively correlated with their average read count over the read count of all candidates, ($r^2 = 0.57$). This shows a strong relationship between CpGs with depletion in read count, relative to average read count, and their variance in methylation with their neighbors.

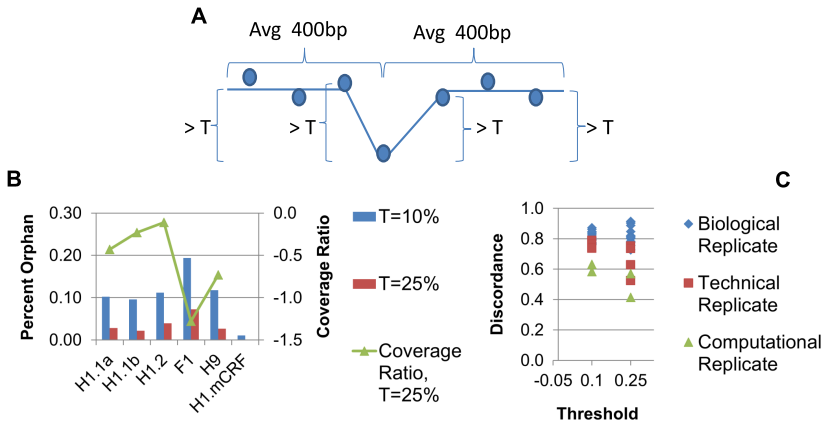


Figure 6.13: A) Criteria for calling orphan CpGs The CpG has 'threshold' greater methylation than or less methylation than both its neighbors as well as the average methylation of CpGs within 400bp on either side. Each 400bp flank must at least 2 CpGs. B) The percentage of orphan CpGs in each WGBS library, as well as H1 methylCRF, at a difference threshold of 10% and 25%. The right axis shows the ratio of orphan read count over all potential orphan's average read count and is for the green line. C) The percentage of discordant orphans, that is, orphans that are unique to reach library in pair-wise comparisons.

In pair-wise comparisons, on average 81% of 10% orphans are unique to a biological replicate (76% and 61% for technical and computational replicates respectively), Fig. 6.13.C. For 25% orphans the numbers change to 83%, 66%, and 49%. The concordant subset of 25% orphans have, on average, much higher read count than orphans as a whole, 138%. However, notably, concordant orphans still have lower read count than candidates overall -only 88% the coverage on average.

6.2.4 Poisson-based HMM suggests 2 states of methylation

Formulating methylation as a state model Given the discordance between WGBS and methylCRF, Fig. 6.3, the lower concordance between WGBS libraries at CpGs with fewer than 10 reads [77], Fig: 6.A., that CpGs with high intermediate methylation sit in windows of higher methylation, Fig. 6.4, potential discordance between replicates due to sampling error, Fig. 6.12, the correlation between read count and UMR count as well as with ratio of read count at UMRs to CpGs overall, and the correlation between orphan read count and concordance, we sought a method to incorporate read count into the estimation of methylation using WGBS. The mechanism of the deposition of methylation is known and it is traditionally thought demethylation occurs passively, ie, the lack of deposition of methylation on the new strand of forming chromatid during cell replication. However, there is evidence of non-replication based demethylation and active demethylation mechanisms have been hypothesized.

Each CpG in a chromatin is either methylated or not. However over a population of chromatin, CpGs have a distribution of methylation values. These results suggest that, at least in part, the variation of methylation values across CpGs in a methylome may be an artifact -whether due to the regulatory mechanisms themselves or to artifacts of measuring methylation. The shape of the distribution of methylation values across the genome, Fig. 6.4, in combination with the local correlation between neighboring CpGs, [77], supports the conjecture of an underlying principle of DNA methylation: the genome, in general, switches between multi-CpG methylated and unmethylated regions. Further, the above empirical and theoretical analysis suggests that many of the CpGs with intermediate methylation may be due to random error and that identified fine structure may be inadequately controlled for artifacts. This then suggests a model where the methylation values for CpGs in either state are imperfectly

enforced (and/or assayed), and so when observed across all CpGs in those states, indeed appear to follow a binomial distribution. That is to say, across a population of cells, at each 'methylated' CpG the expected number of unmethylated chromatin follow a binomial distribution and vice verse for 'unmethylated' CpGs. We assume chromatin sampling is i.i.d and that the regulatory mechanisms providing protection from and enforcement of methylation (and/or artifacts of measuring methylation) have independent error rates. Note that this is a distinct (although related) approach from assuming, for example, that the distribution of methylation values in a methylated region follow a particular distribution.

Accordingly, we developed an HMM, *Twiposn* , with two kinds of states: 1) the methylated state where the number of T's gives the number of errors, and 2) the unmethylated state where instead the number of C's gives the number of errors. Since errors are somewhat rare, for convenience, we use Poisson emission probabilities and instead interpret the number of C's and T's as the number of events occurring in an interval of size of the total read count -and so, both states share the same interval. In order to include CpGs with varying read counts and to simplify our understanding, we define each state over an interval of 1 read. Since the sum of C Poisson distributions is Poisson distributed with interval $\sum_{i=1}^{C} \lambda\_i$ where λ_i is the interval for distribution i, we model each state as a family of Poisson distributions, one for each total read count, that share the same C-to-T ratio.

Incorporating read count in methylation prediction However, given a state with an expected C-to-T ratio -as the total read count increases- the emission probability (of even the most likely C-to-T ratio) decreases. This is the opposite of our intuition that higher read counts leads to more confident estimation! In this way, CpGs with lower read count, ie those with which we have lower confidence, contribute more probability mass to each potential state path than those with high read count. However, MAP sequence estimation for HMM's uses the Viterbi, dynamic-programming, algorithm which performs an $argmax$ at each step. While the probability of each state emitting particular C-to-T ratio decreases with increasing read count, the log odds ratio between methylated and unmethylated state grows (decays) exponentially in favor of the more likely state thus increasing the odds for that state to be taken by Viterbi, Fig. 6.14 -thereby correctly encoding our intuition. For example at a particular CpG with u unmethylated reads, m methylated reads, and $u + m = n$, the odds

ratio is:

$$\frac{P(\text{unmethylated state})}{P(\text{methylated state})} = \frac{f_m(u; n, \lambda_m)}{f_u(m; n, \lambda_u)} \sim \frac{Binom_m(u)}{Binom_u(m)}$$

where λ_m and λ_u are the expected methylated ratio for each state. Then for a given number of methylated and unmethylated reads, m and u, the odds ratio is:

$$\frac{Binom_m(u)}{Binom_u(m)} = \frac{\binom{n}{u}\lambda_m^u(1-\lambda_m)^{u-1}}{\binom{n}{m}\lambda_u^m(1-\lambda_u)^{m-1}}$$

which, since $n = c + m$, reduces to:

$$\frac{\lambda_m^u(1-\lambda_m)^{u-1}}{\lambda_u^m(1-\lambda_u)^{m-1}}$$

Now, say, for a particular c and m, $\frac{P_m}{P_u} > 1$, and so,

$$\lambda_m^u(1-\lambda_m)^{u-1} > \lambda_u^m(1-\lambda_u)^{m-1}$$

Then for any combinations of read counts ur and mr with $r \in \mathbb{R}$, and assuming we can use the gamma function to calculate the factorial for non-integers, this inequality holds, ie

$$\left(\lambda_m^u(1-\lambda_m)^{u-1}\right)^r > \left(\lambda_u^m(1-\lambda_u)^{m-1}\right)^r$$

and the odds ratio as a function of r increases (decays) exponentially for all combinations of counts:

$$f(r) = \left(\frac{Binom_m(u)}{Binom_u(m)}\right)^r$$

Twiposn defined methylation states. Using Baum-Welch (an adaptation of Expectation Maximization that uses the forward-backward algorithm to estimate posterior marginals), we trained the 2-state model on the H1 methylome [49] on which we trained and tested methylCRF, including all CpG's with at least one read. Unlike a simple thresholding model of methylated versus unmethylated, the distribution of methylation in each state can and does overlap, Fig. 6.14. The overlap ranges from 0.43 - 0.50 methylation which has 478k CpG's. Note that this range is much more restricted from the whole intermediate range where methylCRF and WGBS have low concordance.

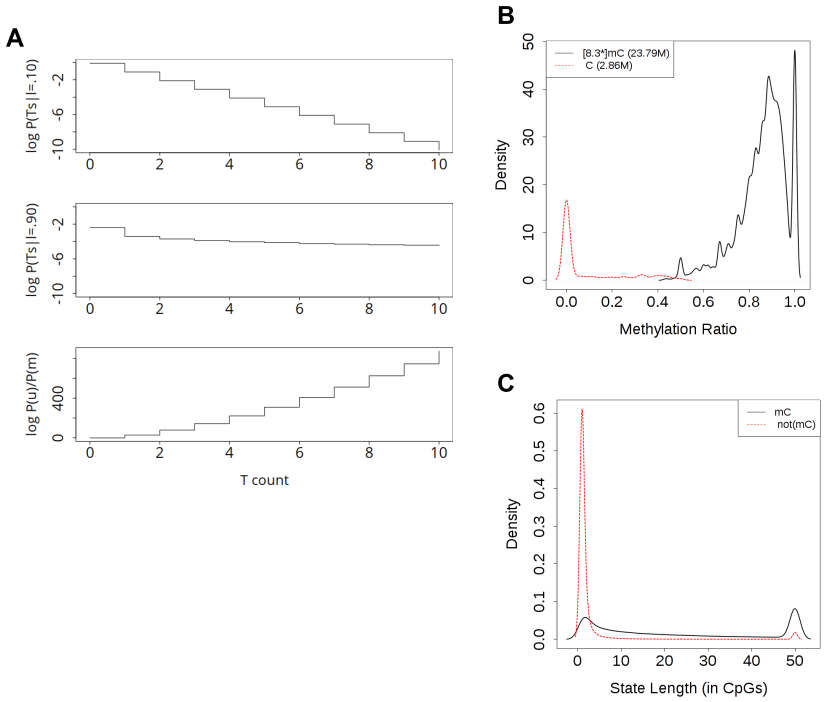


Figure 6.14: For a CpG with only T's aligned to it, the log odds emission ratio for the unmethylated over methylated state increases as the read count increases -even though the probability of emitting all ratios of C-to-T's decreases as the read count increases. A) Top: The log probability of emitting all T's in an unmethylated state (10% C-to-T's). X-axis is the number of reads. Middle: The log probability of emitting all T's in a methylated state (90% C-to-T's). Bottom: The log odds ratio for the unmethylated state over the methylated state. B) CpG methylation distribution for the 2-state Twiposn model. C) State length distribution for the same model. Lengths are truncated at 50 CpGs.

There were 23.8M and 2.9M methylated and unmethylated states. The model suggests that methylated CpGs are 86% methylated in expectation, while unmethylated CpGs are 9%. A model trained on H9 [46] showed similar levels: 85% and 10%. Unmethylated states have much fewer CpGs than methylated states Fig. 6.14.C

For Twiposn states we define orphans as states consisting of only one CpG. 10% of the methylated states were orphans, while 77% unmethylated states were orphans This is consistent with H9, 11% and 77%. 25% of unmethylated orphans had values less than 0.4 and had less than 10 reads, while 40% had less than 0.4 mC and had at least 10 reads. This is reversed in H9, with 40% low and 25% high read, respectively.

We next sought to model more states, such as intermediate methylation at imprinted regions. Accordingly we added a binomial state, since Poisson distributions become worse approximations to binomial distributions as the λ grows given a constant n. We used Baum-Welch again on four ES libraries (H1.1a,H1.1b,H2,H9) allowing the methylated state, the unmethylated state and up to three binomial states with randomly initialized means. Surprisingly, H1.1 and H1.2 resulted in one binomial state having no CpGs, while H1.1b and H9 resulted in both the binomial states being empty. Suggesting that either, methylation in a population of cells consists of two states or the Twiposn model is not powerful enough to detect some states. Interestingly the binomial state in both H1.1a and H2 was a super-low methylation state with expected 0.6% and 0.002% methylation. This allowed their unmethylated state to have higher methylation, 18% and 20% versus 14% and 12% for H1.1b and H9. We re-ran with different initialization 10 times each with similar results.

6.3 Discussion

While we set out to show that methylCRF-based methylomes are as good as WGBS methylomes, we find that it is not a case of whether methylCRF is as good as WGBS. We believe these results raise a serious question about the quality of WGBS estimates. By characterizing where methylCRF and WGBS do agree, we believe we can present high confidence results about methylation, genome-wide. We suggest where they disagree should be considered with lower confidence.

Chapter 7

$TV_{reductio}$

I think that it is a relatively good approximation to truth – which is much too complicated to allow anything but approximations – that mathematical ideas originate in empirics. But, once they are conceived, the subject begins to live a peculiar life of its own and is ... governed by almost entirely aesthetical motivations. In other words, at a great distance from its empirical source, or after much "abstract" inbreeding, a mathematical subject is in danger of degeneration. Whenever this stage is reached the only remedy seems to me to be the rejuvenating return to the source: the reinjection of more or less directly empirical ideas.

- John von Neumann, "The Mathematician", in The Works of the Mind

7.1 Introduction

Since the majority of the CpGs in the genome are methylated [77], we are mostly interested in CpG's where it is not. However, since CpG methylation is estimated over a population, we do not know how much of a change in methylation is meaningful. In its simplest form, then, finding these locations of sufficiently lower methylation is an instance change-point detection. That is, given a series of data points ordered along one or more dimensions, ie, often either in time or in space (such as pixels in an image), assume there are abrupt changes in the value when seen along the ordered dimension that are distinct from random fluctuation. The task is to find whether a change has occurred in the signal and if so where. This is a general problem with applications in many domains such as manufacturing quality control, intrusion detection, spam filtering, website tracking, and medical diagnostics. A methylome therefore could be represented as a set of locations where the methylation value changes.

However, since neighboring CpG's are highly correlated and global methylation levels are bi-model [77], DNA methylation can be characterized as a piece-wise constant signal with the pieces bounded by the change-points. This constraint leads to a special case of change detection with the additional hypothesis that the change points occur relatively infrequently. Used in many domains, this has been referred to as step filtering, step smoothing, and shift, jump, or edge detection.

HMM-based Twiposn is not well suited for WGBS methylation estimation. Our HMM-based methylation state estimator for WGBS libraries, Twiposn, did not work well for segmenting the methylome. Twiposn models a methylated and unmethylated with Poisson distributed emissions and any number of intermediately methylated states with binomial distributions. However, even when we did Baum-Welch training with the two Poisson states and three intermediate states, the learned model only used up to one binomial state. The results were similar in several ES cell methylomes. Further, the binomial state was not even intermediate, it was a super-lowly methylated stated with an expected methylation below the unmethylated Poisson state. However, it is well known that DNA at imprinted regions show intermediate methylation. It is also widely thought methylation might have other such regions.

It is possible that each of our hypothesized states are actually a mixture of states. We could model each state as a mixture of Poisson distributions. While relatively trivial to implement, this adds a fair amount of complexity to the model, which just defers the cost of any potential utility to the interpretation of the results and its integration to existing knowledge about methylation. Without which, all we are doing is adding parameters to the model and thus potentially fitting the data instead of the unknown dynamics we are trying to model. What is the relationship between the constituent components between a region of roughly 86% average methylation versus 40%? If we don't plan to interpret this, then this starts to suggest that this modeling framework may not be well suited for what we are looking for out of this data.

Additionally, HMMs are imbued with an implicit exponential state duration. The probability of staying in a given state with a self-transition probability, p, state for d is $p^{d-1}(1-p)$ [68]. Although the authors describe how to generalize an HMM to explicit state durations, these models require up to D times more storage and $O(D^2)$ more computation, where D is the maximum allowed state length. Additionally, unless parametric densities are used for state length, the number additional parameters needed to train makes over-fitting more of an issue. Since some of the methylated states have over 5000 CpGs this could become prohibitive to model with an HMM.

While there is a possibility to further develop this method, a more fundamental problem exists. We don't actually know if methylation states indeed exist, how long they are, or what distribution they have. Our only estimate is through Twiposn -which, if anything, imposes an exponential (more specifically, geometric) length distribution on states. This leaves us with the possibility of model fishing, guessing, and distribution assumptions. For these reasons -the lack of distinct states and the inherent limitations in modeling state length, we concluded that this type of framework was not well suited for our purposes. We sought a fundamentally different approach to estimating methylation.

Our intuition about local CpG correlation is better encoded as a non-parametric constriant. We used an HMM to incorporate two intuitions about the methylation data 1) by observation and analysis, the values don't change much and small changes could be noise, and 2) the relative large changes are not likely explained by noise. However, this requires us

to first hypothesize a small number of 'states' that stochastically generate the observation. We then use the data to estimate the states, their transition frequencies, and their emission frequencies. Non-parametric regression methods provide a more direct way to incorporate our intuitions without invoking additional hypotheses.

Parametric approaches (such as HMMs), produce simplified representations of data in a sort of top down fashion by controlling the number of parameters used to describe the data globally. If we consider an indicator function that returns 1 only if the likelihood is within some bound, we can frame this in terms of VC-theory [83]. A benefit of controlling the number of parameters is that it controls -for many types of models- the space of possible signals the model can convincingly represent, ie within some probability bound. Consider a type of data of n data points, $\{y|y \in (0,1]^N\}$ and a model with one state with a constant integer emission probability. The model is only powerful enough to have non-zero posterior probability for 10 signals, $\{x|x \in (0,1]^N, x \in Z^n\}$. Roughly, abusing VC-theory, the idea is that the lower the ratio between the number of possible representations a model can take over the number of possible signals, the *capacity*, the better the generalizability of a method is to that class of signals. From this perspective it is easy to see that controlling the capacity of a model can be done through intuitions defining relationships such as between neighboring data points. That is, restricting the type and form of relationships restricts the set of representations.

Using a constrained optimization formalization, the observations about methylation could be encoded directly as a property of the representation x of methylation $y \in \mathbb{R}^n$. Here we seek a representation that is as simple as possible in terms of the number of times it changes value but is still arbitrarily close to y:

$$\begin{aligned} \underset{x^n}{\text{minimize}} \quad & \sum_{i=1}^{n-1} \delta(y_{i-1} \neq y_i) \\ \text{subject to} \quad & f(y,x) \leq S \end{aligned} \tag{7.1}$$

where f is some measure of the distance between y and x and δ is the indicator function. Clearly as S is increased, the fewer the number of changes can exist in x -which reduces the combinatorial number of vectors x can take on. While easy to formulate, this is hard to solve. In fact, it is a variant of the *sparse approximate solution* which is known to be NP-hard. [60].

One mathematical approach to characterizing the amount of change of a function of one variable is to simply add up the total amount of change in the co-domain across the domain. This sum, the *total variation* [35] is defined then for function f over the interval $[a, b] \subset \mathbb{R}$ as the supremum over the sum of differences across all partitions of $[a, b]$, $\mathcal{P} = \{P = \{x_1 \leq x_2 \leq ... \leq x_{nP}\}\}$:

$$TV_a^b(f) = \sup_P \sum_{i=1}^{nP-1} |f(x_{i+1}) - f(x_i)| \tag{7.2}$$

It is easy to see that for functions with discrete domains, that the solution set of partitions always includes the partition of all data points. If a solution, P, didn't contain a data point, x_i, x_i is either 1) between its neighbors, in which case, you could add it without changing the total variation, or 2) it is either higher or lower than its neighbors, in which case, you could add it and increase the total variation contradicting the assumption on P. x_0 or x_n is always in P. The total variation across its whole domain is the sum of its first derivative across all consecutive points. Unless specified, in this manuscript, TV will refer to the discrete case over its entire domain.

Note that for a constant function, $TV(f) = 0$, while for a monotonic function, $TV(f) = |f(n) - f(1)|$ (we will revisit this fact later). In the case of representation, $f(x) = x$, so for notational convenience, we borrow from the notation for norms and specify it as the ℓ_1-norm of the discrete first-derivative :

$$\|x\|_{TV} = \sum_{i=1}^{n-1} |x_{i+1} - x_i|$$

In the fields of image de-noising and signal processing, a now classic, paper [72] using partial differential equations as a means, incorporated the TV statistic as a constraint in the Lagrangian form of an optimization problem in order to separate noise from signal of an image. The intuition is that most objects in images are covered by multiple pixels, and so variation in neighboring pixels tends to be 'spurious oscillations' while large, sharp, and enduring changes are not. Controlling this variation then provides a tractable approach to globally remove noise from an image. The problem then is to find values for all pixels in the image that trades-off fidelity to the image with the total variation between neighboring pixel values. Termed *total variation denoising (regularization)*, the authors use a quadratic loss for

the fidelity, here defined as a discrete 1D version:

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^n}{argmin} \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_i)^2 + \lambda \left\| \beta \right\|_{TV} \tag{7.3}$$

where y has length n, β is the optimal representation, and λ is a tunable parameter varying the trade-off between the two terms. This approach turned out to be a watershed advance -over 6500 citations (`http://scholar.google.com`)- over smoothing methods, because the constraint tends to reduce the transition between objects to sharp borders as apposed to smooth transitions. In fact, edge detection is the topic of Rudin's PhD thesis where he advocated for the importance of singularities of functions in numerical analysis [73].

The ℓ_1 norm has been used extensively in statistics and machine learning to induce simplicity in terms of the number of explanatory variables needed predict an output variable. Among other properties it is the closest convex approximation for cardinality-based statistics, for example the ℓ_0 norm, found in combinatorial problems. As a constraint for ordinary least squares regression, *lasso regression*, the ℓ_1 norm has a tendency to push feature weights to zero [80]. Within the lasso regression framework, this method of inducing simplicity was applied to *signal regression*, where, for example, images or sound files are used as explanatory variables in regression [43]. The intuition was that for regression, representing the signals by higher order features -such as curves in an image or phonemes in a spoken language track, would improve prediction. One could think of this as *representation regression* to distinguish it from variable selection-type regression. In the context of regression, the authors refer to the TV norm regularization as *first order variable fusion*. This idea was extended and formalized as the *fused lasso* which used a combination of both the lasso and TV norms [81]. While their formulation retains the regression context, they applied it to representation problems. A method for the full regularization path of the fused lasso, using a modified piece-wise coordinate descent, was shown to be faster than either convex optimization [26] or LARS [23] for large problems. This was generalized to arbitrary graph structures with piece-wise constant signal across adjacent nodes.

7.2 Results

7.2.1 Fused Lasso, TV Regularization

We applied the fused lasso to a sample of 1000 CpGs of our methylation data. While it does indeed provide a simplified representation, some limitations became immediately apparent Fig. 7.1. The orange bars indicate where we would like to have one segment. One the left, $\lambda = 0.1$ gives a reasonable representation here for a fine resolution. However, this same representation, creates a finely graduated stair-step in the middle of the bar on the right. This kind is detail hard to interpret. On the other hand, overall, $\lambda = 1.0$ gives a much simpler representation by matching some intuition about the gross segments. However, it too shows stair-step patterning. The stair-step on the left makes little sense. A seemingly simple interpretation of this window is that there is a segment with an outlier. Additionally, the window under the right bar could be considered either 1) one long segment where the righthand side has an unusual, but arguably random nevertheless, cluster of lower tending values, or 2) there is in intermediate valued segment starting 2 CpGs right of the segmentation. Additionally, the one CpG with its own level doesn't seem justified at all. Note that all reasonable values of λ fail in this region.

This example highlights one of the limitations of TV normalization: it is myopic. While λ itself is a global parameter -that controls the weight of the total sum of TV terms, each TV term is the relationship between only two variables. As such, there is no consideration of the length of a segment -which certainly is important in understanding the pattern of a signal. For example, consider whether any of the segmentations are correct for the left orange bar Fig. 7.1. It would seem the simplest explanation is that 12 data points with similar values and one outlier -especially when considered in context of all the data. The myopic restriction leads to several notable limitations.

The TV norm can not enforce any kind of simplicity between monotonic regions. That is, in these regions, all segmentations have identical TV penalties. Consider, for example, the regularization path of any set of data $\{x_1, ..., x_n | x_1 \leq x_2 \leq ... \leq x_n\}$. With $\lambda = 0$, the representation will fit the data exactly. As λ increases, the only change to the representation it will cause is to gradually grow the two ends by fusing variables -which consequentially

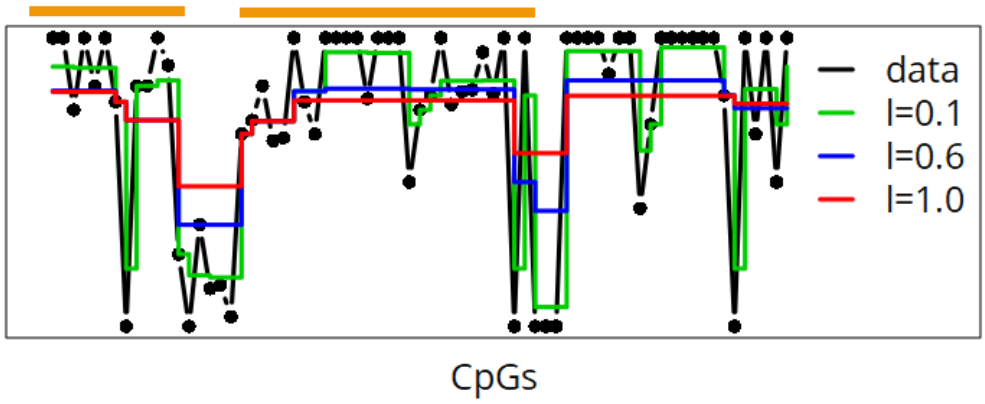


Figure 7.1: Fused lasso run on CpG methylation. The original data for 75 CpGs is in black. The three lines representations using a λ of 0.2, 0.7, and 1.0. The orange bars mark regions we would like to be one segment. A λ lower than 0.1 generally fits the data, while higher than 1.0 approaches the mean of the data set.

reduces the contrast as the end segments (which are the global max and min segments) approach the mean. This then suggests a scenario where for a given signal, the regularization path smooths a signal one fusion at a time [26], gradually forming a representation as a series of gross level minimums and maximums. The segments between these extrema, however are not regularized. Any further smoothing comes by extending and regressing them to the global mean until finally, there is one segment at the mean.

Another fundamental limitation of TV normalization is that it can not take into account local context in its choice of change points. There are limitations to this for the 2D case.

The myopic nature of TV also leads to a problem of interpretation. In general, as $n \to \infty$, or more likely in this case, as the number of analyzed methylomes increases, there will be segments of all values in the range. To see this, consider that with even a random distribution of values, there are finite probabilities of long consecutive segments of equal points everywhere across the range. This then suggests that our conclusions about characteristic values and the number of values methylation has is a function of the number of methylomes we analyze -that is, an artifact of the experiment.

One way to compensate for the myopic limitation of the TV-norm, could be to post-process the segmentation by globally clustering the values of the segments. However, the histogram of values suggests that the clustering would not be justifiable, Fig. 7.2. Additionally, the bottom histogram demonstrates, the contrast reduction mentioned above. Whereas the data ranges from 0.0 to 1.0, the TV representation, only ranges from 0.4 to 0.9. Fig. 7.1 shows that regardless of λ, the TV representation misses the most striking feature -the segments of 0.0 and 1.0 methylation.

7.2.2 Segmental/Structured K-means

Despite its wide success in many fields, the TV norm does not appear robust enough to represent our data well. An alternative approach to simple representation is through feature extraction or learning via dimensionality reduction. In this case, we refer to range of methylation values as the dimensionality of the data. The idea is to create a representation that is close to the date, but only uses a relatively few number of values. As a central task in machine learning, methods span many approaches from simple heuristic clustering, like k-means, to binning and locality-sensitive hashing, to global variance-based methods like principle component analysis, to local similarity-based manifold learning methods. We chose to simply try K-means clustering. This simple method allowed a simple to implement extension to incorporate the observation that methylation values are locally correlated. Standard k-means is an iterative method. Each iteration consists of two steps, 1) data points are assigned to their closest mean, ie cluster, and 2) re-estimate the means based on their membership. We modified the first step by performing Viterbi decoding to find the optimal global assignments of CpGs to means. The cost to change from one mean to another is tunable via a hyper-parameter, η. Note that as in HMMs, this not MAP assignment, but maximum path assignment. Also of interest is connection to the TV norm, in that segmental k-means regularizes the $\eta * \ell_0$ TV norm on the number of changes in value. It thus incorporates our intuition directly without resorting to a convex relaxation.

Despite the methodological and conceptual simplicity of this method, it performed quite well on the same 1000 CpGs tested above Fig. 7.3 with five means and η of 0.0, 0.3, and 0.8. Note that unlike Twiposn, Segmental K-means used all five possible states. Surprisingly, standard K-means, ie, with η of 0.0, follows the data quite well. Similar to the TV norm with a low λ,

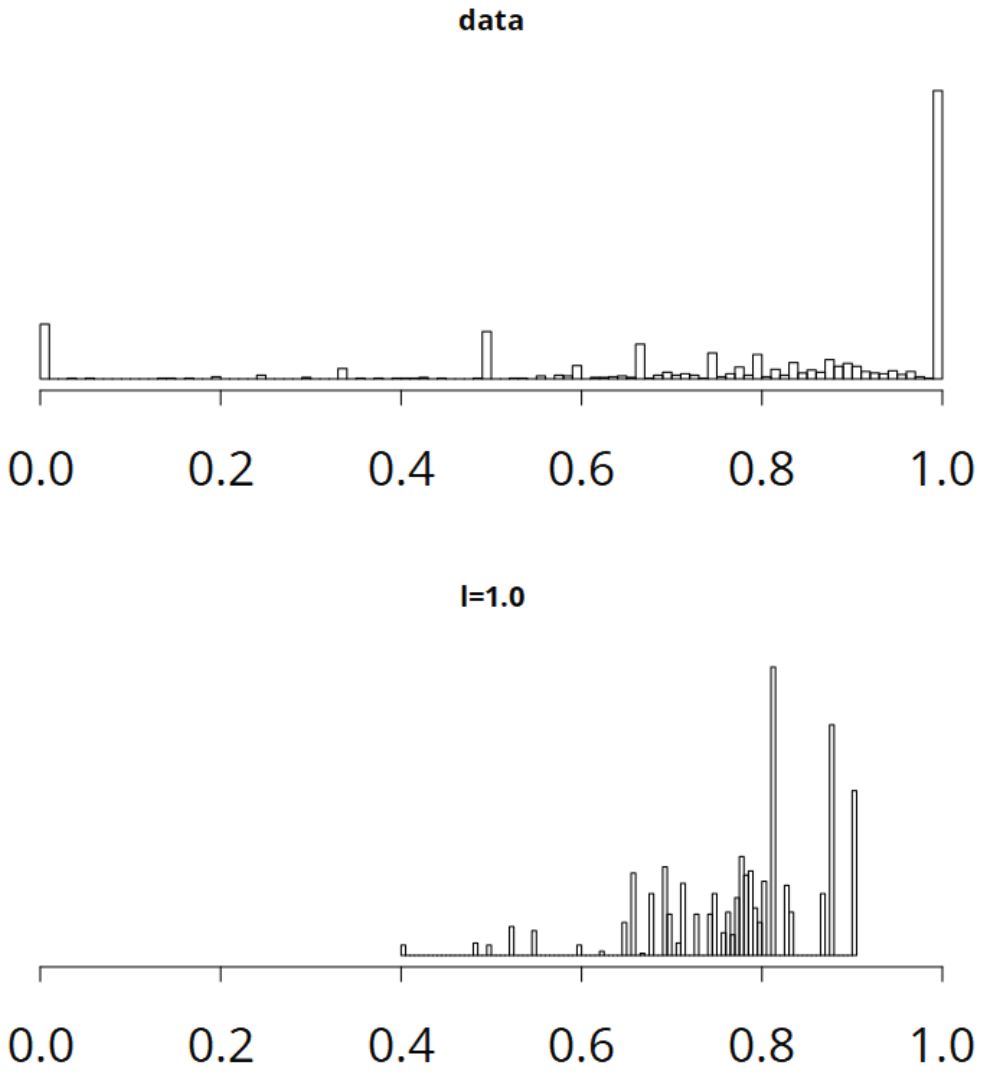


Figure 7.2: Histogram of CpG methylation values for the 1000 CpGs (top) and their fused lass segmentation with a $\lambda = 1$ (bottom).

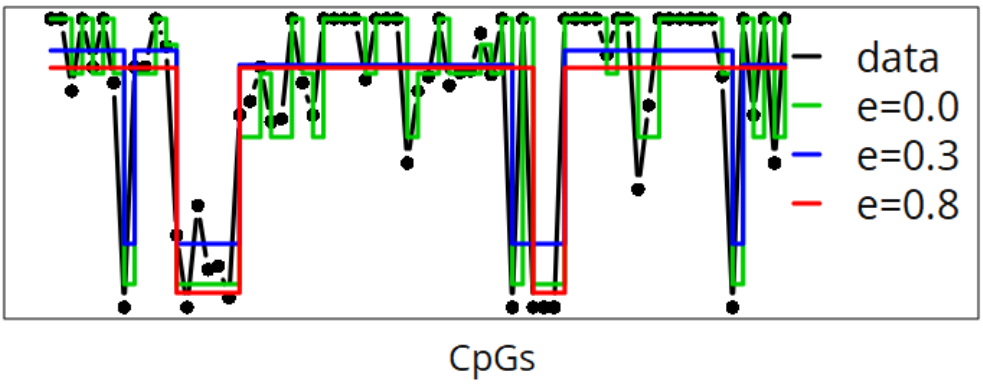


Figure 7.3: Segmental K-mean run on CpG methylation with five means. The original data for 75 CpGs is in black. The line representations using a η of 0.0, 0.3 and 0,8.

segmental k-means with η of 0.3 adds break points for outliers. However, it has many fewer break points. As can be seen with in the histograms, Fig. 7.4, areas with very many CpGs will tend to be split up into much finer clusters. Since the means are randomly initialized, at each iteration, they will tend to move in the direction of the most CpGs in the re-estimate step. If most CpGs are within the 0.7-1.0 range, then this region will tend to have more means. Note that in the assignment step segmental k-means is similar to using the TV norm.

7.2.3 $TV_{reductio}$

On the far left of Fig. 7.3 is a CpG with a much value in a window of noisy, but higher level. What information is required to give us some confidence to determine whether this is a window containing a CpG with one quite unusual value or three windows, two high valued ones with a short low valued window in between. Information such as 1) the frequency of short windows, 2) the frequency of large jumps between windows, 3) the frequency of windows with values similar to the three windows, as well as the 4) variation of values within the windows. While the TV norm contains 2 and 4 to some degree in its objective, it does not have the memory required for 3 or 4. The objective for segmental k-means on the other hand,

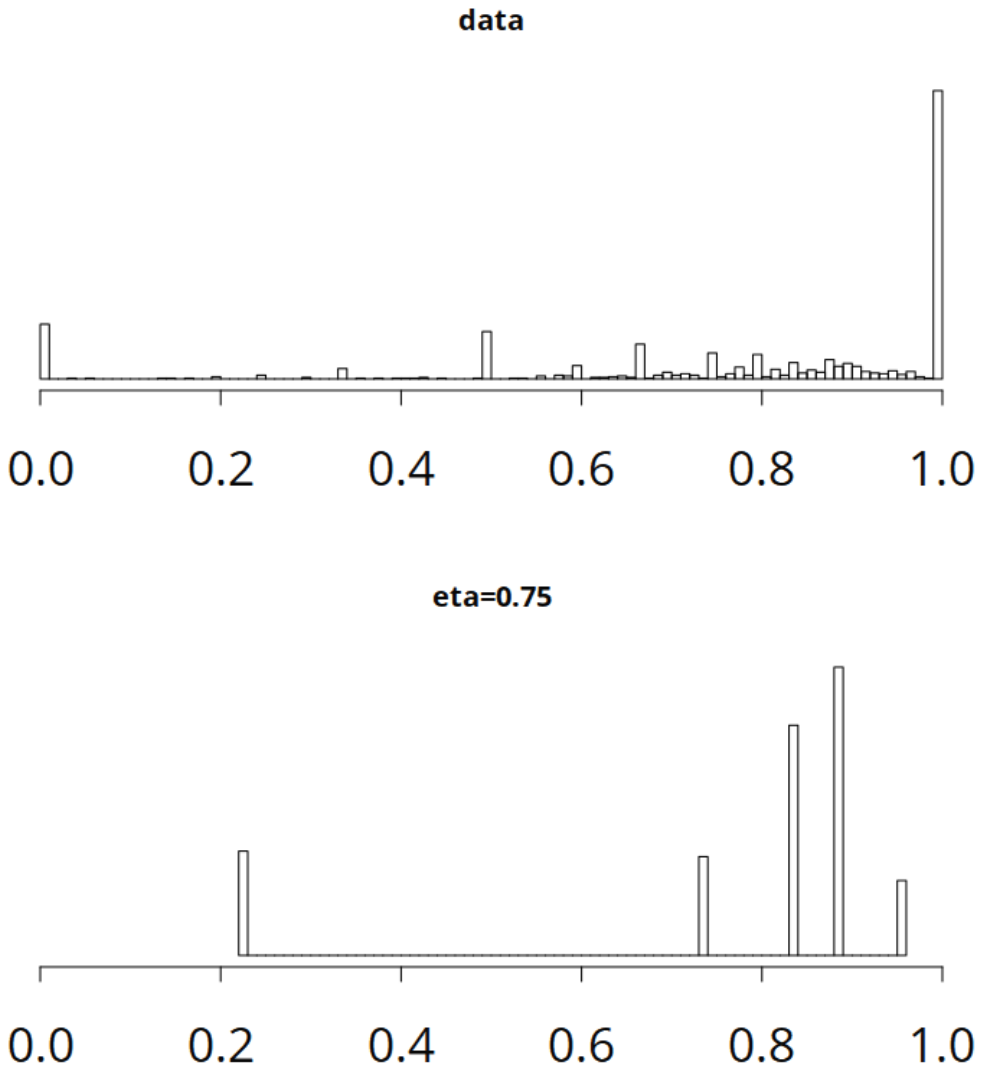


Figure 7.4: Histogram of CpG methylation values for the 1000 CpGs (top) and their segmental k-means segmentation with an $\eta = 0.8$ (bottom).

in some sense contains 3 and 4, however, 2 is not. We propose to combine these two objectives (ie small number of values and small number of changes) in the hope incorporate as much of our intuition about this data as we can. We hope to generalize both the parametric HMM framework to include non-parametric intuitions as well as to extend the non-parametric TV framework, to include parametric intuitions.

Jointly constraining change points and unique value cardinality

Our modified problem we set out to approach for $|y| = n$ is:

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^n}{argmin} \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_i)^2 + \lambda_1 \|\beta\|_{TV} + \lambda_2 \|\{v|v \in \beta\}\|_1 \tag{7.4}$$

That is, we want to find a representation of y that is both segmented into flat segments that come from constrained number of values.

Total variation of sorted values Possibly the simplest approach would be to sort y and add an ℓ_1 constraint between consecutive variable. In sense, this is the TV norm on the sorted data. This results in only two additional terms for each parameter during gradient ascent and so if very efficient. However, this becomes a form of isotonic regression, which as discussed above, for which the ℓ_1 norm (or any pairwise norm) can not sparsify. In fact, testing this method revealed very little difference, regardless of the regularization parameter.

Total smoothness A possibly better approach would be additionally regularize on the *complete total variation*, that is the pair-wise variation among *all* parameters.

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^n}{argmin} \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_i)^2 + \lambda_1 \|\beta\|_{TV} + \lambda_2 \sum_{i=2}^{n} \sum_{j=2}^{n} |\beta_i - \beta_j| \tag{7.5}$$

Similar to TV, increasing the number of an nxn matrix of all pair-wise distances. Note that this sum indirectly counts the number of unique values in the representation. However, it adds a penalty between unique value is not what we want. Hopefully, the number of zeros combined with the other terms would offset this limitation. For the coordinate-wise gradient

descent [26], this formulation would add $2n$ additional terms to each partial derivative as well as additionally checking up to $2n$ regions between discontinuity points for each of the terms when the derivative is not minimized. Additionally, [26], noted that the coordinate-wise ascent sometimes fails where further descent require two parameters to jointly move. They argue that by updating λ in small enough increments, only one pair of parameters would require this at any given step. It is not clear whether this could be addressed with the complete total variation.

However, one way to speed this approach is based on the insight that most of the complete total variation comparisons are not necessary. In fact, they don't make sense. A y with value 1 another with value 0, should not be penalized. To account for this one could instead use a *truncated lasso* for which values greater than α have no penalty. However, this is not a convex penalty and becomes combinatorial in nature. An alternative to this is to instead give values greater than α infinite penalty -which is convex. In order to use this, though, we have to change the minimization to be over $\sum_{i=1}^{C} C^n$ functions where C is the maximum number of unique values allowed in the representation. Where each function is a unique subset of n^2 terms complete total variation. This is, of course, intractable to represent explicitly. However, that is not necessary. Because of the truncated ℓ_1 norm, the majority of functions could not possibly decrease the function value during coordinate-wise descent. However, the number of value regions needed to check during coordinate-wise descent could grow substantially. However, the may be additional heuristics to limit. Regardless, since the problem is convex, convergence is possible. Still, the problem of variable dependence described above may not be addressable. Although, the existence of a dual would allow a certificate to know when convergence is reached.

Reduced resolution Another way to control the number of unique values is to instead use a discretized representation. For many representations of real-world phenomena, there is a limit to the resolution that is useful. Often continuous representation are nonetheless discretized before use. This is the case for methylation -the difference between 3.4% and 3.5% is not interpretable. For a discrete set of values A, Eq. (7.3) becomes The problem becomes:

$$\hat{\beta} \in \underset{\beta \in A^n}{argmin} \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_i)^2 + \lambda \left\| \beta \right\|_{TV} \tag{7.6}$$

This then is the form of a Markov random field for which graph-cuts can solve for graph structures of 2 or 3 node cliques. Since Eq. (7.6) contains only 2 nodes cliques, however, we can use the Viterbi algorithm to it. Then we need to wrap this in another minimization to regularize the size of A:

$$\hat{\beta} \in \min_{a \subseteq A} \left(\underset{\beta \in a^n}{argmin} \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_i)^2 + \lambda_1 \left\| \beta \right\|_{TV} \right) + \lambda_2 \left\| a \right\|_1 \tag{7.7}$$

Since the minimum of a convex functions is itself convex, Eq. (7.7) consists minimizing a convex plus a modular function. It turns out the convex functions are upper bounded by a modular function [24] for which the authors develop a majorization-minimization algorithm that is guaranteed to converge. However convergence is local. Using submodular minimization, [3], showed how to solve functions of this form taking the Lovasz relaxing of submodular constraint. However, as [24] pointed out the consistency proofs for Lovasz extensions are typically given with respect to the relaxation. This leads non-optimal solutions of the original, non relaxed, objective.

7.3 Discussion

We are left with a choice to solve Eq. (7.7), 1) find a local solution, or 2) find a global solution to a relaxed form.

Chapter 8

Conclusion

8.1 Summary

> It is the nature of an hypothesis, when once a man has conceived it, that it assimulates every thing to itself as proper nourishment; and, from the first moment of your begetting it, it generally grows the stronger by every thing you see, hear, read, or understand. This is of great use.
>
> -Laurence Sterne, The Life and Opinions of Tristram Shandy, Gentleman

In this thesis, we set out to use DNA methylation as a means of discovering the component of DNA-binding proteins binding in a cell in the service of understanding how one genome might provide a polymorphic interface allowing for multiple cell-type identities. Since current methods of assaying DNA methylation are limited in either resolution or comprehensiveness or are prohibitively expensive, we first developed an alternative, *methylCRF*, that is comprehensive, and high-resolution. We achieved this by combining two complimentary assays of methylation using a statistical model. It is concordant within the range of replicates of WGBS while being 15 times cheaper. A benefit of this price differential is that we were able to examine methylation across more cell-types than was previously possible. We examined 58

methylomes to reveal the extent of methylation variation at single-CpG resolution. Results include that only 28% CpGs vary across these cell-types and that only 11% of the genome has variably methylated regions and these regions are indeed enriched in potential regulatory regions. Detailed analysis of methylCRF and WGBS revealed systemic differences between the two bringing into question whether either are accurate enough to use in helping identify transcription factor binding and suggests that a statistical model is required WGBS. Additionally, WGBS seems to strongly suffer from GC-bias in the underlying protocol. Given these limitations, we re-formulated the representation of DNA methylation, from un-methylated and methylated regions to that of change-point detection. While change-point detection is a core problem with methods used across wide cross-section of technical fields, through detailed analysis, we found that current methods are lacking in the ability to represent the methylation signal as a few change-points between a limited number of methylation levels. We propose to extend existing total variation method to simultaneously learn the methylation levels present in a population of cells as well as the change-points of those levels.

As ongoing work, we will use either submodular minimization or a maximization/majorization approximation framework to implement our proposal. The result of this will be a radical extension of change-point detection of possible wide applicability across the fields using it. Specifically, this method can be combined with transcription factor analysis to our goals of finding a representation for DNA methylation that could lead to understanding of multi-cellular capability of our genome. Additional ongoing work is the conceptual development and promotion of a unifying set of objectives, values, and concerns for big data biology to ensure its vitality, ability to advance biological knowledge and to place it in reciprocal relationships with other technical fields.

References

[1] Daniel Aird, Michael G. Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David B. Jaffe, Chad Nusbaum, and Andreas Gnirke. Analyzing and minimizing PCR amplification bias in illumina sequencing libraries. *Genome Biology*, 12(2):R18, February 2011. 00177 PMID: 21338519.

[2] Manuel Allhoff, Alexander Schonhuth, Marcel Martin, Ivan G Costa, Sven Rahmann, and Tobias Marschall. Discovering motifs that induce sequencing errors. *BMC Bioinformatics*, 14(Suppl 5):S1, April 2013.

[3] Francis Bach. Structured sparsity-inducing norms through submodular functions. 2010. 00077.

[4] Yuval Benjamini and Terence P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10):e72–e72, May 2012.

[5] Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, Peggy J Farnham, Martin Hirst, Eric S Lander, Tarjei S Mikkelsen, and James A Thomson. The NIH roadmap epigenomics mapping consortium. *Nature Biotechnology*, 28(10):1045–1048, October 2010. 00197 PMID: 20944595.

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

[7] Christoph Bock, Eleni M Tomazou, Arie B Brinkman, Fabian Müller, Femke Simmer, Hongcang Gu, Natalie Jäger, Andreas Gnirke, Hendrik G Stunnenberg, and Alexander Meissner. Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nature biotechnology*, 28(10):1106–1114, October 2010. 00206 PMID: 20852634.

[8] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. *Advances in neural information processing systems*, 20:161–168, 2008.

[9] Lukas Chavez, Justyna Jozefczuk, Christina Grimm, Jörn Dietrich, Bernd Timmermann, Hans Lehrach, Ralf Herwig, and James Adjaye. Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. *Genome Research*, 20(10):1441–1450, October 2010. 00064 PMID: 20802089.

[10] Yen-Chun Chen, Tsunglin Liu, Chun-Hui Yu, Tzen-Yuh Chiang, and Chi-Chuan Hwang. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS ONE*, 8(4):e62856, April 2013. 00005.

[11] Bastien Chevreux. mira's cheveux GGC.g, March 2009.

[12] Brian E. Clauser. The life and labors of francis galton: A review of four recent books about the father of behavioral statistics. *Journal of Educational and Behavioral Statistics*, 32(4):440–444, December 2007. 00003.

[13] Cristian Coarfa, Fuli Yu, Christopher A Miller, Zuozhou Chen, R Alan Harris, and Aleksandar Milosavljevic. Pash 3.0: A versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel DNA sequencing. *BMC bioinformatics*, 11:572, 2010. 00023 PMID: 21092284.

[14] Shawn J Cokus, Suhua Feng, Xiaoyu Zhang, Zugen Chen, Barry Merriman, Christian D Haudenschild, Sriharsa Pradhan, Stanley F Nelson, Matteo Pellegrini, and Steven E Jacobsen. Shotgun bisulphite sequencing of the arabidopsis genome reveals DNA methylation patterning. *Nature*, 452(7184):215–219, March 2008.

[15] F Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, August 1970. Cited by 1277.

[16] Jesse Dabney and Matthias Meyer. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *BioTechniques*, 52(2):87–94, February 2012.

[17] Charles Darwin. *The Descent of Man, and Selection in Relation to Sex*. D. Appleton, 1872. 00604.

[18] Rajdeep Das, Nevenka Dimitrova, Zhenyu Xuan, Robert A Rollins, Fatemah Haghighi, John R Edwards, Jingyue Ju, Timothy H Bestor, and Michael Q Zhang. Computational prediction of methylation status in human genomic sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 103(28):10713–10716, July 2006. 00107 PMID: 16818882.

[19] D. DeCaprio, J. P. Vinson, M. D. Pearson, P. Montgomery, M. Doherty, and J. E. Galagan. Conrad: Gene prediction using conditional random fields. *Genome Research*, 17(9):1389–1398, July 2007. 00062.

[20] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 194–202, 1995. 01652.

[21] Thomas A Down, Vardhman K Rakyan, Daniel J Turner, Paul Flicek, Heng Li, Eugene Kulesha, Stefan Graf, Nathan Johnson, Javier Herrero, Eleni M Tomazou, Natalie P Thorne, Liselotte Backdahl, Marlis Herberth, Kevin L Howe, David K Jackson, Marcos M Miretti, John C Marioni, Ewan Birney, Tim J P Hubbard, Richard Durbin, Simon Tavare, and Stephan Beck. A bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotech*, 26(7):779–785, July 2008.

[22] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, May 1998.

[23] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499. 03618.

[24] M. El Halabi, L. Baldassarre, and V. Cevher. To convexify or not? regression with clustering penalties on graphs. In *2013 IEEE 5th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 21–24, December 2013. 00000.

[25] Jason Ernst and Manolis Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotech*, 28(8):817–825, 2010.

[26] Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, December 2007. 00653.

[27] Martin C Frith, Ryota Mori, and Kiyoshi Asai. A mostly traditional approach improves alignment of bisulfite-converted DNA. *Nucleic acids research*, 40(13):e100, July 2012. 00014 PMID: 22457070.

[28] C T Friz. The biochemical composition of the free-living amoebae chaos chaos, amoeba dubia and amoeba proteus. *Comparative biochemistry and physiology*, 26(1):81–90, July 1968. 00022 PMID: 4249055.

[29] Kasper D Hansen, Steven E Brenner, and Sandrine Dudoit. Biases in illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research*, 38(12):e131, July 2010.

[30] R Alan Harris, Ting Wang, Cristian Coarfa, Raman P Nagarajan, Chibo Hong, Sara L Downey, Brett E Johnson, Shaun D Fouse, Allen Delaney, Yongjun Zhao, Adam Olshen, Tracy Ballinger, Xin Zhou, Kevin J Forsberg, Junchen Gu, Lorigail Echipare, Henriette O'Geen, Ryan Lister, Mattia Pelizzola, Yuanxin Xi, Charles B Epstein, Bradley E Bernstein, R David Hawkins, Bing Ren, Wen-Yu Chung, Hongcang Gu, Christoph Bock, Andreas Gnirke, Michael Q Zhang, David Haussler, Joseph R Ecker, Wei Li,

Peggy J Farnham, Robert A Waterland, Alexander Meissner, Marco A Marra, Martin Hirst, Aleksandar Milosavljevic, and Joseph F Costello. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotech*, 28(10):1097–1105, October 2010.

[31] Gary C. Hon, Nisha Rajagopal, Yin Shen, David F. McCleary, Feng Yue, My D. Dang, and Bing Ren. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nature Genetics*, advance online publication, September 2013. Cited by 0000.

[32] Illumina Inc. TruSeq SR cluster kit v3-cBot-HS support. 00000.

[33] Rafael A Irizarry, Christine Ladd-Acosta, Bo Wen, Zhijin Wu, Carolina Montano, Patrick Onyango, Hengmi Cui, Kevin Gabo, Michael Rongione, Maree Webster, Hong Ji, James B Potash, Sarven Sabunciyan, and Andrew P Feinberg. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genetics*, 41(2):178–186, February 2009.

[34] Peter A Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics*, 13(7):484–492, July 2012. Cited by 0194.

[35] C. Jordan. Sur la série de fourier. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences, Paris*, 92:228–230, 1881. 00121 MSC2010: 26A45 = Functions of bounded variation (one real variable).

[36] W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and and David Haussler. The human genome browser at UCSC. *Genome Research*, 12(6):996 –1006, June 2002. 03033.

[37] Robert J. Klose and Adrian P. Bird. Genomic DNA methylation: the mark and its mediators. *Trends in Biochemical Sciences*, 31(2):89–97, February 2006. Cited by 1115.

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

[39] Felix Krueger, Benjamin Kreck, Andre Franke, and Simon R Andrews. DNA methylome analysis using short bisulfite sequencing data. *Nat Meth*, 9(2):145–151, February 2012. 00061.

[40] Lukasz A. Kurgan and Krzysztof J. Cios. CAIM discretization algorithm. *IEEE Trans. on Knowl. and Data Eng.*, 16(2):145–153, February 2004. 00273.

[41] John Lafferty. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. pages 282–289, 2001.

[42] Peter W. Laird. Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet*, 11(3):191–203, February 2010. 00509.

[43] Stephanie R. Land and Jerome H. Friedman. Variable fusion: A new adaptive signal regression method. Technical report, Technical Report 656, Department of Statistics, Carnegie Mellon University Pittsburgh, 1997. 00018.

[44] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H. Waterston, Richard K. Wilson, LaDeana W. Hillier, John D. McPherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kymberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Minx, Sandra W. Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan-Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J. Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L. Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R. Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Hong Mei Lee, JoAnn Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W. Davis, Nancy A. Federspiel, A. Pia Abola, Michael J. Proctor, Bruce A. Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W. Richard McCombie, Melissa de la Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu-Chuan

Chen, Deanna Church, Michele Clamp, Richard R. Copley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G. R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V. Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P. Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F. A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I. Wolf, Kenneth H. Wolfe, Shiaw-Pyng Yang, Ru-Fang Yeh, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A. Evans, Maria Athanasiou, Roger Schultz, Aristides Patrinos, and Michael J. Morgan. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001. 00002.

[45] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, March 2009.

[46] Louise Laurent, Eleanor Wong, Guoliang Li, Tien Huynh, Aristotelis Tsirigos, Chin Thing Ong, Hwee Meng Low, Kin Sung, Ken Wing, Isidore Rigoutsos, Jeanne Loring, and Chia-Lin Wei. Dynamic changes in the human methylome during differentiation. *Genome Research*, February 2010. 00355.

[47] E Li, T H Bestor, and R Jaenisch. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, 69(6):915–926, June 1992. 02812 PMID: 1606615.

[48] Jian Li, R Alan Harris, Sau Wai Cheung, Cristian Coarfa, Mira Jeong, Margaret A Goodell, Lisa D White, Ankita Patel, Sung-Hae Kang, Chad Shaw, A Craig Chinault, Tomasz Gambin, Anna Gambin, James R Lupski, and Aleksandar Milosavljevic. Genomic hypomethylation in the human germline associates with selective structural mutability in the human genome. *PLoS genetics*, 8(5):e1002692, 2012. 00029 PMID: 22615578.

[49] Ryan Lister, Ronan C O'Malley, Julian Tonti-Filippini, Brian D Gregory, Charles C Berry, A Harvey Millar, and Joseph R Ecker. Highly integrated single-base resolution maps of the epigenome in arabidopsis. *Cell*, 133(3):523–536, May 2008.

[50] Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, Lee Edsall, Jessica Antosiewicz-Bourget, Ron Stewart, Victor Ruotti, A Harvey Millar, James A Thomson, Bing Ren, and Joseph R Ecker. Human DNA methylomes at base resolution

show widespread epigenomic differences. *Nature*, 462(7271):315–322, November 2009. 01438 PMID: 19829295.

[51] JB Macqueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[52] V Matys, E Fricke, R Geffers, E Gössling, M Haubrock, R Hehl, K Hornischer, D Karas, A E Kel, O V Kel-Margoulis, D-U Kloos, S Land, B Lewicki-Potapov, H Michael, R Münch, I Reuter, S Rotert, H Saxel, M Scheer, S Thiele, and E Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1):374–378, January 2003.

[53] Alika K Maunakea, Raman P Nagarajan, Mikhail Bilenky, Tracy J Ballinger, Cletus D'Souza, Shaun D Fouse, Brett E Johnson, Chibo Hong, Cydney Nielsen, Yongjun Zhao, Gustavo Turecki, Allen Delaney, Richard Varhol, Nina Thiessen, Ksenya Shchors, Vivi M Heine, David H Rowitch, Xiaoyun Xing, Chris Fiore, Maximiliaan Schillebeeckx, Steven J M Jones, David Haussler, Marco A Marra, Martin Hirst, Ting Wang, and Joseph F Costello. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, 466(7303):253–257, July 2010.

[54] Alexander Meissner, Tarjei S Mikkelsen, Hongcang Gu, Marius Wernig, Jacob Hanna, Andrey Sivachenko, Xiaolan Zhang, Bradley E Bernstein, Chad Nusbaum, David B Jaffe, Andreas Gnirke, Rudolf Jaenisch, and Eric S Lander. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770, August 2008. 01168 PMID: 18600261.

[55] André E. Minoche, Juliane C. Dohm, and Heinz Himmelbauer. Evaluation of genomic high-throughput sequencing data generated on illumina HiSeq and genome analyzer systems. *Genome Biology*, 12(11):R112, November 2011. 00101 PMID: 22067484.

[56] Evgeny A Moskalev, Mikhail G Zavgorodnij, Svetlana P Majorova, Ivan A Vorobjev, Pouria Jandaghi, Irina V Bure, and Jörg D Hoheisel. Correction of PCR-bias in quantitative DNA methylation studies by means of cubic polynomial regression. *Nucleic Acids Research*, 39(11):e77–e77, June 2011.

[57] Nasheen Naidoo, Yudi Pawitan, Richie Soong, David N Cooper, and Chee-Seng Ku. Human genetics and genomics a decade after the release of the draft sequence of the human genome. *Human Genomics*, 5(6):577–622, October 2011. Cited by 0007.

[58] Shalima S Nair, Marcel W Coolen, Clare Stirzaker, Jenny Z Song, Aaron L Statham, Dario Strbenac, Mark W Robinson, and Susan J Clark. Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias. *Epigenetics: Official Journal of the DNA Methylation Society*, 6(1):34–44, January 2011.

[59] Kensuke Nakamura, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, Margaret C. Linak, Aki Hirai, Hiroki Takahashi, Md Altaf-Ul-Amin, Naotake Ogasawara, and Shigehiko Kanaya. Sequence-specific error profile of illumina sequencers. *Nucleic Acids Research*, 39(13):e90–e90, July 2011. 00134 PMID: 21576222.

[60] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, April 1995. 00911.

[61] John Von Neumann. *Theory of Self-Reproducing Automata*. University of Illinois Press, Champaign, IL, USA, 1966. 04994.

[62] Christian Otto, Peter F Stadler, and Steve Hoffmann. Fast and sensitive mapping of bisulfite-treated sequencing data. *Bioinformatics (Oxford, England)*, 28(13):1698–1704, July 2012. 00006 PMID: 22581174.

[63] Mattia Pelizzola, Yasuo Koga, Alexander Eckehart Urban, Michael Krauthammer, Sherman Weissman, Ruth Halaban, and Annette M. Molinaro. MEDME: An experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. *Genome Research*, 18(10):1652 –1659, October 2008.

[64] Len A Pennacchio, Nadav Ahituv, Alan M Moses, Shyam Prabhakar, Marcelo A Nobrega, Malak Shoukry, Simon Minovitsky, Inna Dubchak, Amy Holt, Keith D Lewis, Ingrid Plajzer-Frick, Jennifer Akiyama, Sarah De Val, Veena Afzal, Brian L Black, Olivier Couronne, Michael B Eisen, Axel Visel, and Edward M Rubin. In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502, November 2006. 00615 PMID: 17086198.

[65] Karl Popper. *The Logic of Scientific Discovery*. Routledge, April 2014. 00000.

[66] Elodie Portales-Casamar, Supat Thongjuea, Andrew T Kwon, David Arenillas, Xiaobei Zhao, Eivind Valen, Dimas Yusuf, Boris Lenhard, Wyeth W Wasserman, and Albin Sandelin. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 38(Database issue):D105–110, January 2010.

[67] Tao Qin, Tie Liu, Xu Zhang, De Wang, and Hang Li. Global ranking using continuous conditional random fields. In *NIPS*, pages 1281–1288. MIT Press, 2008.

[68] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989. 18826.

[69] Keith D. Robertson. DNA methylation and human disease. *Nat Rev Genet*, 6(8):597–610, 2005.

[70] Christian Rohde, Yingying Zhang, Richard Reinhardt, and Albert Jeltsch. BISMA–fast and accurate bisulfite sequencing data analysis of individual clones from unique and repetitive sequences. *BMC Bioinformatics*, 11:230, 2010.

[71] Michael G. Ross, Carsten Russ, Maura Costello, Andrew Hollinger, Niall J. Lennon, Ryan Hegarty, Chad Nusbaum, and David B. Jaffe. Characterizing and measuring bias in sequence data. *Genome Biology*, 14(5):R51, May 2013. 00018 PMID: 23718773.

[72] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1–4):259–268, November 1992. 06659.

[73] Leonid Iakov Rudin. *Images, numerical analysis of singularities and shock filters.* phd, California Institute of Technology, 1987. 00092.

[74] David Serre, Byron H Lee, and Angela H Ting. MBD-isolated genome sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Research*, 38(2):391–399, January 2010.

[75] Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. pages 213–220, 2003.

[76] Michael B. Stadler, Rabih Murr, Lukas Burger, Robert Ivanek, Florian Lienert, Anne Scholer, Christiane Wirbelauer, Edward J. Oakeley, Dimos Gaidatzis, Vijay K. Tiwari, and Dirk Schubeler. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, advance online publication, December 2011.

[77] Michael Stevens, Jeffrey B Cheng, Daofeng Li, Mingchao Xie, Chibo Hong, Cécile L Maire, Keith L Ligon, Martin Hirst, Marco A Marra, Joseph F Costello, and Ting Wang. Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome research*, 23(9):1541–1553, September 2013. Cited by 0001.

[78] G D Stormo. DNA binding sites: representation and discovery. *Bioinformatics (Oxford, England)*, 16(1):16–23, January 2000. Cited by 1033.

[79] Miho M Suzuki and Adrian Bird. DNA methylation landscapes: provocative insights from epigenomics. *Nature reviews. Genetics*, 9(6):465–476, June 2008. Cited by 0851.

[80] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 10061.

[81] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, pages 91–108, 2005. 00561.

[82] Akiko Tsumura, Tomohiro Hayakawa, Yuichi Kumaki, Shin-ichiro Takebayashi, Morito Sakaue, Chisa Matsuoka, Kunitada Shimotohno, Fuyuki Ishikawa, En Li, Hiroki R Ueda, Jun-ichi Nakayama, and Masaki Okano. Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases dnmt1, dnmt3a and dnmt3b. *Genes to Cells: Devoted to Molecular & Cellular Mechanisms*, 11(7):805–814, July 2006.

[83] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000. 00285.

[84] Axel Visel, Simon Minovitsky, Inna Dubchak, and Len A Pennacchio. VISTA enhancer browser–a database of tissue-specific human enhancers. *Nucleic acids research*, 35(Database issue):D88–92, January 2007. 00203 PMID: 17130149.

[85] Hanna M. Wallach. Conditional random fields: An introduction. *Technical Reports (CIS)*, 2004. 00268.

[86] J D WATSON and F H CRICK. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, April 1953. 00014 PMID: 13054692.

[87] Michael Weber, Jonathan J Davies, David Wittig, Edward J Oakeley, Michael Haase, Wan L Lam, and Dirk Schübeler. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature genetics*, 37(8):853–862, August 2005. Cited by 0969.

[88] Yuanxin Xi and Wei Li. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, 10(1):232, July 2009.

[89] Xiao-Lin Yin and Jing Li. Detecting copy number variations from array CGH data based on a conditional random field model. *Journal of bioinformatics and computational biology*, 8(2):295–314, April 2010. 00008 PMID: 20401947.

[90] Michael J. Ziller, Hongcang Gu, Fabian Müller, Julie Donaghey, Linus T.-Y. Tsai, Oliver Kohlbacher, Philip L. De Jager, Evan D. Rosen, David A. Bennett, Bradley E. Bernstein, Andreas Gnirke, and Alexander Meissner. Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500(7463):477–481, August 2013. Cited by 0000.