MACHINE LEARNING APPROACHES TO UNRAVELLING THE ROLE OF MAMMALIAN DNA METHYLATION IN GENE REGULATION

by

Milos Pavlovic

APPROVED BY SUPERVISORY COMMITTEE:

Dr. Michael Q. Zhang, Chair

Dr. Min Chen, Co-Chair

Dr. Pradipta Ray

Dr. Jeff DeJong

Dr. Tae Hoon Kim

Dr. Zhenyu Xuan

Copyright 2017

Milos Pavlovic

All Rights Reserved

Dedicated to my wife Kristina Pavlovic.

MACHINE LEARNING APPROACHES TO UNRAVELLING THE ROLE OF MAMMALIAN DNA METHYLATION IN GENE REGULATION

by

MILOS PAVLOVIC, MS

DISSERTATION
Presented to the Faculty of
The University of Texas at Dallas
in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY IN
MOLECULAR AND CELL BIOLOGY

THE UNIVERSITY OF TEXAS AT DALLAS
December 2017

ACKNOWLEDGMENTS

Firstly, I would like to thank my family: mostly my wife Kristina, my mother Milanka and my father Milisa, who have always been supportive of me throughout this long and challenging PhD journey. Next, I would like to thank my supervisors and all committee members for their valuable research input. To Michael, thank you for admitting me to your lab, and for providing support and guiding my research in the right direction. Special thanks goes to Pradipta for spending extensive hours training me in Computer Science and Machine Learning and for being my mentor. I would also like to acknowledge Min for advising me and teaching me how to perform statistical modeling. Finally, I would like to thank all my colleagues from the Zhang and Xuan labs for sharing research ideas and being good friends, especially my best man Peng Xie.

October 2017

MACHINE LEARNING APPROACHES TO UNRAVELLING THE ROLE OF MAMMALIAN DNA METHYLATION IN GENE REGULATION

Milos Pavlovic, PhD
The University of Texas at Dallas, 2017

Supervising Professor: Michael Q. Zhang

Transcriptional regulation is a highly complicated and dynamic process established by regulatory pathways involving cascades, feedbacks and other sophisticated control mechanisms. Epigenetic mechanisms are key regulatory processes involving heritable modifications to the genome that do not require the substitution of constituent nucleotides in the DNA sequence but which may be suitably reprogrammed in germ cells. 5-Methylcytosine and 5-Hydroxymethylcytosine in DNA are major epigenetic modifications known to be implicated in mammalian gene regulation. The literature suggests that DNA methylation in a promoter or enhancer region causes transcription repression, while hydroxymethylation abundance in enhancers coincides with elevated expression of proximal genes. Accordingly, obtaining, analyzing, and interpreting Next Generation Sequencing methylation data could give us a deeper insight into the trancriptome, as well as modes of epigenetic gene regulation. However, performing whole-genome methylation assays is expensive and unfeasible to conduct for every physiological or perturbation condition, and often generates incomplete genome-wide methylation profile. For that purpose we created a novel, supervised, ensemble-learning classification framework to perform whole-genome methylation and hydroxymethylation status predictions in CpG dinucleotides. Additionally, we developed a platform to perform *in silico*, high-throughput hypotheses testing based on such predictions. For the purpose of performing *de novo* methylome reconstruction, we adopted the concept of invariant methylation across mammalian reference methylomes, and incorporated it into our framework by creating the *consensus reference methylome*. Our toolkit performs fast and accurate prediction and imputation on large amounts (~Terabytes) of data in existing sequencing datasets. Since we do not use cell type specific features such as Transcription Factor Binding Sites, models trained on one cell type can be used to predict the epigenetic profile of a related cell type, thereby showing great promise for transfer learning scenarios. We test our approach on H1 human embryonic stem cells and H1-derived neural progenitor cells. Our predictive model is comparable in accuracy to other state-of-the-art DNA methylation prediction algorithms, and is the first *in silico* predictor of hydroxymethylation achieving high whole-genome accuracy, paving the way for large-scale reconstruction of hydroxymethylation maps in mammalian model systems. We designed a novel, beam-search driven feature selection algorithm to identify the most discriminative predictor variables, and developed a platform for performing integrative analysis and reconstruction of the epigenome. Our toolkit DIRECTION provides predictions at single nucleotide resolution and identifies relevant features based on resource availability. This offers enhanced biological interpretability of results potentially leading to a better understanding of epigenetic gene regulation. Our tool is publicly available and can be downloaded from: utdallas.edu/~mxp114330/direction

TABLE OF CONTENTS

Acknowledgments v

Abstract vi

List of Figures x

List of Tables xiii

CHAPTER 1 AN INTRODUCTION TO EPIGENETICS AND COMPUTATIONAL APPROACHES FOR CHARACTERIZING DNA METHYLATION 1

- 1.1 Prior publication 2
- 1.2 Contribution 2
- 1.3 Abstract 3
- 1.4 Epigenetics and epigenetic modifications 3
- 1.5 DNA methylation and its role in gene regulation 4
- 1.6 Detection and high-throughput quantification of DNA methylation 6
- 1.7 Computational approaches to modeling of BS-seq and TAB-seq data and functional analysis of DNA methylation 9
- 1.8 DNA methylation prediction: Context in literature 11
- 1.9 Evaluation of prediction quality 13
- 1.10 Importance of predicting DNA methylation and hydroxymethylation 16

CHAPTER 2 INTER-METHYLOME SIMILARITIES: CONSENSUS REFERENCE METHYLOME AND ITS USAGE IN DNA METHYLATION PREDICTION 18

- 2.1 Prior Publication 19
- 2.2 Abstract 19
- 2.3 Introduction 20
- 2.4 Methods 20
- 2.5 Results 23
- 2.6 Summary 24

CHAPTER 3 DNA METHYLATION PREDICTION 26

- 3.1 Prior Publication 27
- 3.2 Abstract 27

3.3 Introduction 28
3.4 Methods 29
3.5 Results 53
CHAPTER 4 DNA HYDROXYMETHYLATION PREDICTION 70
4.1 Prior Publication 71
4.2 Abstract 71
4.3 Introduction 72
4.4 Methods 73
4.5 Results 78
CHAPTER 5 DNA METHYLATION AND HYDROXYMETHYLATION TRANSFER LEARNING…... 95
5.1 Prior Publication 96
5.2 Abstract 96
5.3 Introduction 97
5.4 DNA methylation transfer learning 98
5.5 5-hmC transfer learning 100
5.6 Limits of transfer learning using DIRECTION 101
CHAPTER 6 SUMMARY 102
REFERENCES 107
BIOGRAPHICAL SKETCH…....118
CURRICULUM VITAE

LIST OF FIGURES

Fig. 2.1: DNA methylation reconstruction framework 21

Fig. 2.2: DNA methylation predictions harnessing intra- and inter-methylome similarities 22

Fig. 3.1: Whole methylome empirical distributions of BS-seq CCRs in NPC and liver tissue 29

Fig. 3.2: Dependence of the performance metric based (F-score) based on training and test set size 36

Fig. 3.3: Schema of our prediction framework outlining beam search (feature selection), training, testing, and cross-validation modes 45

Fig. 3.4: An example path traversed by beam search through Precision-Recall space, while optimizing F-score (in parentheses) for H1 non-CGI SVM model. 46

Fig. 3.5: Beam search algorithm feature set exploration shown for beam width = 2, for two levels of the search tree 48

Fig. 3.6: Schematic representation of optimal feature set finding algorithm embedded within DIRECTION 49

Fig. 3.7: Precision/Recall plot for balanced sets prediction of DNA methylation status using the methylation status of the nearest neighbors 51

Fig. 3.8: Precision/Recall plot for balanced sets methylation status imputation using methylation status of nearest neighboring CpG site as function of distance to nearest neighbor 52

Fig. 3.9: Cumulative Distribution Function of the fraction of low coverage CpG sites w.r.t distance to the nearest high coverage site in BS-seq datasets 52

Fig. 3.10: Empirical distributions of high coverage methylome such as NPC and low coverage methylome such as Fetal Small Intestine 53

Fig. 3.11: CGI and non-CGI SVM model DNA methylation status predictions using balanced sets in NPC 54

Fig. 3.12: Prediction metrics for DNA methylation balanced sets status prediction in H1 cells CGI and non-CGI regions. 55

Fig. 3.13: Whole genome methylation status prediction accuracy obtained by binning the whole genome based on the BS-seq level in H1, and NPC 60

Fig. 3.14: Balanced sets predictions in H1 and NPC based on exclusion and inclusion of intermediate methylation in CGI, and non-CGI 61

Fig. 3.15: Comparison of DIRECTION and classification tree for NPC methylation prediction. 62

Fig. 3.16: Assorted feature sets generated by beam search by using various priority evaluation metrics for DNA methylation status prediction in NPC 63

Fig. 3.17: AUC curves for methylation status prediction in NPC CGI and non-CGI regions.......63

Fig. 3.18: Hierarchical clustering of features in OFS for predicting methylation status in NPC CGI and non-CGI regions, and changes in precision and recall with respect to OFS. 64

Fig. 4.1: Empirical distributions of 5-hmC levels in NPC and H1 cell lines................................73

Fig. 4.2: Distribution of 5-hmC levels in the set of CpG sites identified as significantly hydroxymethylated in Yu et al. 2012 (114) 74

Fig. 4.3: Concordance rate between TAB-seq NPC replicates as a function of minimum sequencing depth of mapping at either replicate...75

Fig. 4.4: Inferred 5-mC level distribution in NPC by the tool MLML (31) by jointly analyzing BS-seq and TAB- seq CCRs in NPC. 76

Fig. 4.5: 5-hmC status prediction on balanced sets using SVM in NPC and H1..........................78

Fig. 4.6: Precision/Recall plot for 5-hmC status predictions in NPC using various 5-hmC level thresholds for SVMs. 80

Fig. 4.7: ROC curves for 5-hmC status predictions in NPC RF model.. 81

Fig. 4.8: OFS feature clustering 84

Fig. 4.9: Visualization of 5-hmC status prediction and discriminative input features in PCDH17… 85

Fig. 4.10: Visualization of 5-hmC status prediction and discriminative input features in MEIS2… 85

Fig. 4.11: 5-hmC status prediction in enhancer regions using SVM and RF models...................86

Fig. 4.12: Precision/Recall plot for 5-hmC status predictions in NPC in enhancer regions using balanced sets for SVM. 87

Fig. 4.13: Sequencing depth in NPC enhancers........87

Fig. 4.14: Enhancers with high 5-hmC enrichment and their proximal genes 90

Fig. 5.1: Hierarchical clustering of reference methylomes from the Roadmap Epigenome Consortium........99

LIST OF TABLES

Table 1.1: Literature survey of methylation prediction 14

Table 2.1: List of cell and tissue types that were used to create the reference methylome 23

Table 2.2: Relative size of the consensus reference methylome w.r.t to disagreement thresholds. 23

Table 3.1: Summary on different modes of our methylation analysis toolkit 32

Table 3.2: Fisher's Exact Test p-values for various BS-seq and TAB-seq sequencing depths. ...38

Table 3.3: List of features used for predicting DNA methylation and hydroxymethylation 43

Table 3.4: Initial feature sets for NPC and H1 methylation status predictions 44

Table 3.5: F-score trajectory as beam search algorithm searches through feature space 47

Table 3.6: Similarity of OFSs across different beam width values 50

Table 3.7: Beam search algorithm shown as a pseudocode 50

Table 3.8: Whole genome BS-seq status prediction evaluation in NPC dataset using SVM 55

Table 3.9: Balanced set evaluations for DNA methylation prediction 56

Table 3.10: Whole genome evaluations for DNA methylation prediction 58

Table 3.11: Feature sets for methylation status prediction. 65

Table 4.1: Feature sets for 5-hmC status prediction in NPC, NPC enhancers, and H1 79

Table 4.2: Whole genome 5-hmC status prediction evaluation metrics in NPC and H1 81

Table 4.3: Balanced set and whole genome evaluations for 5-hmC status predictions 83

Table 4.4: Prediction metrics for 5-hmC status prediction based on nearest neighbor's 5-hmC status 91

CHAPTER 1

AN INTRODUCTION TO EPIGENETICS AND COMPUTATIONAL APPROACHES FOR CHARACTERIZING DNA METHYLATION

Authors: Milos Pavlovic, Pradipta Ray, Kristina Pavlovic, Aaron Kotamarti,

Min Chen and Michael Q. Zhang

Department of Biological Sciences

The University of Texas at Dallas

800 W. Campbell Road

Richardson, TX 75080-3021

1.1 Prior publication

Milos Pavlovic (M.P.) and Pradipta Ray (P.R.) wrote the manuscript. Min Chen (M.C.) advised M.P. and Michael Zhang (M.Q.Z.) supervised the project. This chapter provides a broad introduction to the field of Epigenetics, primarily focusing on DNA methylation and computational methods for prediction and downstream analysis of this epigenetic modification. Per the policy of OUP Bioinformatics, the publication of material in a PhD thesis is permitted with the publication of a peer-reviewed manuscript in their journal. The original manuscript (1) "DIRECTION: A machine learning framework for predicting and charactering DNA methylation and hydroxymethylation in mammalian genomes" by Milos Pavlovic, Pradipta Ray, Kristina Pavlovic, Aaron Kotamarti, Min Chen and Michael Q. Zhang, published in 2017, is reproduced by permission of Oxford University Press and appears online at the following web address: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btx316. Supplementary information is available online at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btx316. The main text is partially altered compared to the online version of the manuscript, and figures and tables do not chronologically correspond to the online manuscript numbering.

1.2 Contribution

Milos Pavlovic wrote and Pradipta Ray edited this chapter. This chapter provides a broad literature review of the field of Epigenetics, with special focus on DNA methylation and related computational methods, which have been developed for the purpose of predicting and characterizing this epigenetic modification.

1.3 Abstract

Epigenetics is an emerging discipline which domain of study encompasses a set of stable and heritable traits that are unrelated to the underlying DNA sequence. Major epigenetic modifications involve DNA methylation and histone modifications, which role proved to be essential in gene regulation. Therefore, the undergoing research in the field of epigenetics has great promise for improving personalized medicine. The following literature review will mostly cover biological mechanisms behind establishing, maintaining and detecting DNA methylation in mammals, as well as assorted computational methods, which have been developed for prediction and functional analysis of DNA methylation.

1.4 Epigenetics and epigenetic modifications

Epigenetics is a rapidly evolving discipline of molecular biology whose domain of study involves a set of stable and heritable traits which are implicated in gene regulation, and which cannot be explained by the underlying DNA sequence (2). Epigenetic mechanisms are key regulatory processes involving heritable modifications to the genome that do not require the substitution of constituent nucleotides in the DNA sequence but which may be suitably reprogrammed in germ cells (3). In mammalian model systems, the assembly including DNA, histone proteins and certain RNAs gives rise to a complex structure called chromatin, whose primary role is to package the genome in a space efficient manner and control vital cellular processes such as cell division and DNA replication. The top-level chromatin structure largely depends on the underlying DNA methylation and histone modifications, which predominantly determine epigenetic and gene expression profiles of a cell (4). The published literature strongly indicates that aberrant epigenetic patterns are hallmarks of many diseases such as cancer and various brain disorders. Additionally, epigenetic marks play an important role in various developmental processes; can be faithfully transmitted from parents to progeny, and therefore give rise to various distinct phenotypes (epi-phenotypes) where the underlying genotype remains the same. Consequently, deep understanding of the fundamental epigenetic mechanisms and its role in gene regulation have been subject to extensive biomedical research as they hold a great promise for improving personalized medicine (3, 5-8).

The next chapter will cover the main biological mechanisms that are responsible for establishment and maintenance of DNA methylation in mammals, as well as its role in gene regulation.

1.5 DNA methylation and its role in gene regulation

DNA methylation represents a sequence of processes in which methyl groups are added onto the DNA molecule by enzymatic activity. The most prevalent form of DNA methylation in mammals is 5-mC, in which a methyl group from S-adenosylmethionine is transferred onto the 5th position of the carbon ring in the Cytosine base, and is frequently associated with gene silencing in mammals (9, 10). 5-mC presence has also been evidenced in plants, fungi and bacteria, involving processes such as bacterial immune response (restriction-modification system) to phage infection (11). In mammalian genomes the vast majority (approximately 98%) of methylated cytosines are present in a form of a CpG dinucleotide, with non-CpG methylation being significantly more common in plants (12).

The enzymes that are involved in initial establishment of 5-mC modification (DNMT3A, DNMT3B) belong to the group of *de novo* methyl transferases (13). Subsequently, methylation patterns are faithfully maintained throughout the course of cell divisions as a result of the hemi-methylation mechanism, governed by the DNMT1 enzyme which recognizes a methyl group on the mother DNA strand and copies it onto the daughter strand (14). In cancer, these processes are often altered, leading to the aberrant methylation landscape (15), suggesting an important functional role of DNA methylation in disease vs. healthy state.

Depending on its localization in the genome, 5-mC can have different effects on gene expression. In the gene promoters, enhancer regions, or in the proximity to the transcription start sites (TSS), 5-mC is known to cause gene silencing, whereas its presence in the gene body can boost transcription elongation. Other functions of 5-mC include centromere stability and recruitment of the enzymes involved in chromatin remodeling and formation of the transcription initiation complex (15).

It is noteworthy that in the demethylation process, 5-mC is oxidized by Ten-eleven Translocation (TET) enzymes, thereby giving rise to its oxidative derivatives 5-hydroxymethylcytosine (5-hmC), 5-formylcytosine (5-fC), and 5-carboxylcytosine (5-caC) (10).

The mostly studied and characterized oxidative derivative of 5-mC is 5-hmC, which presence was first discovered in bacteriophages in 1952 (16). However, the elucidation of the role of 5-hmC in methylation dynamics, gene regulation and cell development has only recently been characterized (17-19).

5-hmC is generated in TET1 and TET2 mediated oxidation reaction of 5-mC, and represents an intermediate step in the cascade of events that ultimately lead to passive demethylation of cytosine in mouse primordial germ cells (20). The proposed mechanism suggests spontaneous replication dilution of 5-hmC, since there is no described mechanism for 5-hmC maintenance in mammalian genomes. However, in the recent study Bachman et al. (21) found that 5-hmC levels do not significantly change during mammalian cell cycle progression, implying that 5-hmC is a predominantly stable epigenetic modification whose functional role is tightly coupled with cell proliferation. From previous work it is known that 5-hmC closely associates with enhancers (22), exon-intron boundaries (23), elevated C-to-G conversion rates (24), labile nucleosomes and CTCF binding. Additionally, latest literature suggests that 5-hmC abundance across different tissues varies significantly, with brain tissue being 5-hmC enriched and certain cancer tissues (breast, blood) exhibiting 5-hmC depletion, suggesting an important role of 5-hmC in determining healthy vs. disease state (25, 26).

1.6 Detection and high-throughput quantification of DNA methylation

Methylation detection techniques can be broadly divided into following categories:

a) **Methylation-specific enzyme digestion**: Technologies like HELP (HpaII tiny fragment Enrichment by Ligation-mediated PCR) (27, 28) relies on restriction enzymes HALPII and its isoschizomer MSPL to selectively cut the unmethylated cytosines, followed by PCR amplification and sequencing. The biggest shortcoming of such approach is that only 4 out of 100 off all non-repeat CpG dinucleotides are being recognized and subsequently cleaved by HPAII, leaving out a huge portion of CpG sites undetected (29).

b**) Antibody based**: Technologies like MEDIP (Methylated DNA Immunoprecipitation) involve the treatment of DNA with the anti-5mC monoclonal antibody, followed by PCR amplification of the pulled down DNA fragments. The amplified PCR products containing input DNA and methylated DNA can be differentially labeled by cyanine dyes Cy5 and Cy3 and further cohybridized on oligonucleotide arrays or high-throughput sequenced by MEDIP (29). Such methods are overly dependent on the quality and cross reactivity of anti 5-mC antibody, and are additionally biased in their immunoprecipitation step since DNA methylation is not uniformly distributed across the genome. For this purpose, additional statistical modeling is required, performed by tolls like BATMAN (30).

c) **Whole genome or reduced representation amplification:** Methylation-specific PCR (14) has been used to selectively amplify methylated regions, followed by sequencing of amplified fragments. However, the most comprehensive and the *de facto* standard technique for whole genome methylation quantification is Sodium Bisulfite treatment of DNA (31), which causes methylated cytosines to remain intact while unmethylated cytosines are deaminated to uracils (C-to-U conversion) (32). Traditionally, PCR amplification is the next step followed by sequencing. However, other options involve restriction enzyme digestion using Combined Bisulfite Restriction Analysis (COBRA) or Methylation-sensitive Single Nucleotide Primer Extension (Ms-SNuPE) followed by sequencing (33). Variations on this theme are used by employing methylation-specific primers (34). Whole-genome shotgun Bisulfite Sequencing (BS-seq or WGBS) involves all PCR fragments genome-wide, while the Reduced Representation Bisulfite-sequencing (RRBS-seq) protocol leads to a small fraction of the fragments being selected (35). RRBS-seq involves digestion by a restriction enzyme constraining DNA fragments to have CpG sites at both ends, in the process leading to approximately 1% sampling of the whole genome (36).

BS-seq experiments allow us estimate C-to-U conversion rate (CCR) or methylation level for each cytosine in the genome, which serves as an estimator of the degree of methylation. This is widely regarded to be the most faithful quantification of DNA methylation levels, but sources of noise that need to be modeled in BS-seq data include low CCRs (~1%) which can lead to false positive methylation calls (37), DNA depurination due to bisulfite treatment which can cause breaks among DNA strands (38), and imputation of missing data in RRBS sequencing (39). However, the biggest confounding factor is that BS-seq cannot distinguish between 5-mC and 5-hmC, hence the estimated methylation level is due to both 5-mC and 5-hmC.

In order to quantify the degree of hydroxymethylation, alternate protocols like TET-Assisted BS-seq (TAB-seq) (22) and Oxidative BS-seq (oxBS-seq) (40) were developed.

TAB-seq is a bisulfite-based technique that distinguishes between 5-mC and 5-hmC. DNA is first treated with an enzyme 5-hmC glucosyltransferase, which will selectively attach a glucose molecule onto 5-hmC modified cytosines only. Next step involves treating already glycosylated DNA with TET1 and TET2 enzymes, which will induce oxidation of 5-mC and its subsequent conversion to unmethylated cytosine, whereas already glycosylated 5-hmC modified cytosines will remain intact since TET cannot oxidize glycosylated 5-hmC (22). The following steps include bisulfite treatment and sequencing. Consequently, every cytosine “coming out of the sequencing machine” is a 5-hmC site. In addition to BS-seq based, hydroxymethylation identification and quantification assays can be also be restriction enzyme (41) and antibody (42) based.

In this dissertation, detectable modifications from BS-seq experiments (yielding summation of 5-mC and 5-hmC levels) are referred to as methylation, and genome-wide characterization of methylation as methylome. Detectable modifications from TAB-seq (yielding solely 5-hmC driven CCRs or 5-hmC levels) are referred to as hydroxymethylation, and corresponding genome-wide maps as hydroxymethylome.

1.7 Computational approaches to modeling of BS-seq and TAB-seq data and functional analysis of DNA methylation

Traditionally, Next Generation Sequencing (NGS) techniques require either a *de novo* assembly of the sequenced reads or mapping the reads to a known reference genome. For BS-seq and TAB-seq data, C-to-U conversion represents a challenge, since uracils are recognized by sequencing machines as thymidines, which causes a prevalence of single nucleotide variations, leading to complications in indexing-based mapping schemes. Typically, indices of bisulfite-treated genomes are generated *in silico* to aid the mapping process (43). Various tools like BSMAP, RMAP, BS-Seeker and BISMARK (43-46) perform end-to-end mapping analysis or build wrappers around state-of-the-art generic NGS read mapping tolls like Bowtie (47). Based on variant calling at every cytosine, individual cytosines or CpG dinucleotides can be called “methylated” or “unmethylated”, or categorized as one of 5-mC, 5-hmC or unmethylated based on BS-seq and TAB-seq data. Typically, most such methylation calling strategies use a filtering scheme to count with high quality sequencing and alignment scores, followed by a simple binomial probability test (22). However, it is noteworthy that in the population of cells BS-seq and TAB-seq protocols provide CCRs for cytosines ranging from 0 (unmethylated) to 1 (fully methylated). Additionally, BS-seq and TAB-seq datasets often disagree for the portion of the data, due to experimental or sampling error. Model based approaches like MLML (31) are routinely used to ensure consistency between the datasets in question, by systematically discarding overshoot indices in which the sum of 5-mC and 5-hmC levels is greater than 1. Multiple primary sources of methylation data have been integratively modeled to derive a single methylation level consistent with all data sources: MethylCRF (48) uses MEDIP-seq and restriction enzyme based methylation data to reconstruct the whole methylome using Conditional Random Fields (CRF).

Based on quantification of methylation levels at individual cytosines or CpG sites, various downstream functional analyses of methylomes have been successfully performed in the last decade. Markov chain based models have been developed for contrasting different methylomes to identify Differentially Methylated Regions (DMRs), as well as identification and characterization of contiguous domains of Fully Methylated, Lowly Methylated and Unmethylated Regions (FMRs, LMRs, and UMRs respectively) in the genome (49).

Methylation data has been successfully used to build models for predicting active regulatory regions (MethylSeekR (50)) and for predicting cancer drug sensitivities (51).

Recently, multiple association studies largely predicated upon classical Genome-Wide Association Studies (GWAS) (52) and Expression Quantitative Trait Loci (eQTL) models (53) have found that alterations in certain CpG sites methylation levels strongly correlate with proximal genetic variation and expression of a proximal gene (54-56). Such CpG sites known as meQTLs often occur in contiguous genomic regions, exhibit significant correlation with local chromatin and Transcription Factor (TF) binding profiles (54) suggesting that changes in DNA methylation levels occur in harmony with other major gene regulatory processes in healthy cells. However, no such study has considered confounding factors in DNA methylation signal coming from 5-hmC modification, which is known to be correlated with enhancer-like histone modifications (H3K27ac, H3K4me1) and which overabundance in the gene bodies often coincides with elevated expression levels of a proximal gene in question (1). Therefore, when in possession of both BS-seq and TAB-seq data, individual eQTL models corresponding to 5-mC and 5-hmC would provide a solution to deconvoluting inevitably mixed DNA methylation signals and would help decrease discovery of potentially false associations.

1.8 DNA methylation prediction: Context in literature

Over the past decade, high-throughput assays and corresponding computational models have been actively pursued to annotate and predict the epigenome (57, 58) including several approaches for predicting methylation as either a binary or continuous variable in CpG dinucleotides. The earliest methods for DNA methylation prediction used sequence-based and structure-based information to train Support Vector Machines (SVMs) and decision tree classification models. All of these algorithms predict binary DNA methylation status (discrete approximation of a methylation level, using CCR of 0.5 as a threshold) of entire CpG islands (CGI) or contiguous genomic fragments of approximately 100bp size, and achieve accuracy in the range between 85% and 94% (59-63). Concretely, HDMFinder (61) exclusively uses DNA sequence derived features, such as TF Binding Sites (TFBS) and Alu repeats, to predict methylation status of 800bp long CpG-centered regions across the genome, and achieves 86% accuracy. Analysis of discriminative DNA sequence-based features for methylation prediction was performed which lead to identification of DNA motifs corresponding to aberrant methylation patterns in cancer (64). Semi-supervised learning approaches for methylation prediction (harnessing clustering of unlabeled data) have been used for predicting hyper-methylated genes in cancer (65). The underlying structure based on the density of the input features has been utilized for predicting DNA methylation in CGIs using k nearest neighbor (k-NN) algorithm (66). More generic tools have been developed to use DNA sequence derived features to predict both histone modification and DNA methylation states across the genome, as in EPIGRAM, which uses Random Forest (RF) and achieves 91% accuracy (67). In their study *Flores et al.* were the first to use evolutionary data such as genome-wide CG depletion signatures to predict methylation status (68). However, sequence-based prediction of methylation is limited in its ability to identify cell type, tissue, or condition-specific methylation patterns across datasets as underlying sequence features remain unchanged. Since such methylation patterns are of specific interest to biologists, several studies analyzed correlation between methylation and various assays profiling TF ChIP-seq, DNAse-seq or chromatin landscape. CPGIMethPred (69) uses epigenomic and sequence derived features, while *McCabe et al.* (70) uses polycomb binding and genomic composition features to predict methylation states of CGIs. *Wrzodek et al.* (71)*,* *Kondo et al.* (72)*, and Luu at al.* (73) analyzed correlational patterns between methylation states and various epigenome and ChIP-seq derived input features. However, these are all correlative studies or predictive algorithms which predict genome-wide methylation levels of CpG islands only, which count for a small fraction of all CpG sites in the genome (74).

Such knowledge has been leveraged to build explicit predictive models of DNA methylation based on histone modification, nucleosome positioning, chromatin accessibility and TFBS, including several at single nucleotide and dinucleotide resolution. *Whitaker et al.* (67) uses discriminative sequence motifs for individual datasets to predict CpG methylation. *Ma et al.* (75) uses SVM regression to predict methylation as a continuous-valued response variable in CpG sites across tissues, and *Zhang et al.* (76) uses RF on genome, epigenome and ChIP-seq derived traits and neighboring CpG methylation levels for imputing methylation arrays. *Yan et al.* (77) uses RFs on sequence and epigenome-derived features by training on BS-seq data, while *Wang et al.* (78) uses SVMs and deep neural networks on topological domains and other features by training in RRBS-seq data. *Fan et al.* (79) predict stem cell CpG methylation for methylation arrays and BS-seq data, while *Angermueller et al.* (80) are the first to predict methylation status of CpG sites in single cells, across five tissues, for single cell BS-seq (sc-BS-seq) and single cell RRBS-seq data. Finally, *Pavlovic et al.* used SVMs and RF trained models for methylation status prediction, were the first to predict hydroxymethylation status using TAB-seq data to achieve 82% genome-wide accuracy, and implemented the concept of invariant methylation in methylation status prediction to obtain 97% genome-wide accuracy (1) (Table 1.1 for comprehensive survey updated from (76)).

1.9 Evaluation of prediction quality

In machine learning assorted evaluation metrics are routinely employed for evaluating performance of various predictive models. In this dissertation, for evaluating predictions the following metrics were used to evaluate prediction quality on the whole-genome scale and balanced sets (same number of examples from positive and negative class used for evaluation):

1. $Precision = \frac{TP}{TP+FP}$
2. $Recall\ or\ Sensitivity\ or\ True\ Positive\ Rate = \frac{TP}{TP+FN}$
3. $Specificity\ or\ True\ Negative\ Rate = \frac{TN}{TN+FP}$
4. $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
5. $F - score\ or\ F1 - score = \frac{2 * Sensitivity * Specificity}{Sensitivity + Specificty}$
6. AUC: Which is created by plotting the False Positive Rate (1-True Negative Rate) on the x-axis and the True Positive Rate on the y-axis, and the resulting area under the curve is calculated.

Where TP=number of true positives, TN=number of true negatives, FP=number of false positives, and FN=number of false negatives.

Table 1.1: Literature survey of methylation prediction (Methods: NB: Naive Bayes, LR: Logistic Regression, k-NN: k Nearest Neighbor, RF: Random Forest, SVM Support Vector Machine, LDA: Linear Discriminant Analysis, ANN: Artificial Neural Network) (Metrics: ACC: Accuracy, MCC: Matthews Correlation Coefficient, CC: Correlation Coefficient, R: Regression Coefficient, RMSE: Root Mean Square Error) [1] (81) [2] (82) [3] (83) [4] (84) [5] (85)

Citations	Samples	Model	Features	Response variable	Performance metric
(86)	Restriction Landmark Genome Scanning for control fibroblast and DNMT1-overexpressed fibroblast cell lines	LDA	k-mer and consensus motifs in CGI	Methylation prone CGIs vs Methylation resistant CGIs for DNMT1 overexpression among unmethylated CGIs in controls	ACC: 0.82
(59)	MethDB (curated database of ~5,000 experimentally determined methylation of DNA fragments in species from plants to humans)[1]	SVM (best), ANN, NB, LR, k-NN, decision tree	Genomic features (binary sparse encoding of sequence)	Methylation status of DNA fragments of 39bp	SVM (polynomial kernel degree 6) metrics: ACC: 0.7506, MCC: 0.504, AUC: 0.82
(64)	Restriction Landmark Genome Scanning for control fibroblast and DNMT1-overexpressed fibroblast cell lines	LDA	Discriminative motifs in CGI obtained using MAST	Methylation prone CGIs vs Methylation resistant CGIs for DNMT1 overexpression among unmethylated CGIs in controls	ACC: 0.84
(60)	Methylation status of CGI in the non-repetitive parts of human Chromosome 21 (HpaII-McrBC PCR method)- 149 CGIs[2]	SVM linear kernel (best), RBF SVM, Decision tree, AdaBoost	k-mer and nucleotide content, predicted DNA structure, repeat regions, TFBS, evolutionary conservation, SNP frequency	CGI methylation status for whole CGI	Linear SVM metrics: CC:0.74, ACC:0.915
(61)	Human brain data[3] with methylation status of ~5,500 genomic domains	SVM RBF kernel (best), K-means, LDA, LR	k-mer content and repeat regions	Methylation status of 800bp regions	RBF SVM metrics: ACC: Overall: 0.86, CGIs: 0.965, non-CGIs: 0.84
(62)	Human brain data[3] with methylation status of ~5,500 genomic domains	SVM (linear kernel)	Nucleotide and dinucleotide content, Alu element, TFBSs	Methylation status of CpG-rich 200-500bp regions (CGI fragments)	ACC: 0.8303-0.8499, CC: 0.567-0.686
(25)	Bisulfite treated tumor and normal human samples followed by targeted 454 sequencing of 25 gene-related CGIs	NB (best), SVM (SMO), ANN, kNN (k=3)	30bp flanking sequence of each CpG site	Methylation status of randomly selected 41 CpG sites from sequenced dataset (methylation level ≥0.5 or ≤ 0.01)	NB metrics: ACC:>0.75
(87)	Methylation status of CGI in the non-repetitive parts of human Chromosome 21 (HpaII-McrBC PCR method)[2]	SVM (linear kernel)	DNA sequence patterns, repeat distribution, predicted DNA helix structure, predicted TFBS, genetic variation, and CGI attributes	Methylation status of CGI	CC: 0.698, ACC: 0.868
(88)	Human Epigenome Project[4] data for chromosomes 6, 20, and 22, using methylation status in human CD4+ T lymphocytes	SVM (linear kernel)	Nucleotide content, Alu annotation, TFBS, and histone methylation (H3K4me1, H3K4me2, H3K4me3, and H3K9me1)	CGI methylation status	ACC: 0.8994
(63)	Methylation status of CGI in the non-repetitive parts of human Chromosome 21 (HpaII-McrBC PCR method)[2]	Alternative decision tree (best), decision tree, AdaBoost, SVM	4-mer frequencies in CGI	Methylation status of CGIs on chromosome 21	Alternating decision tree metrics: ACC: 0.9063, AUC: 0.8906, MCC: 0.742
(85)	Various vertebrate epigenomic datasets[5]	AdaStump, Decision Tree, RF, NB, LR, SVM (linear, RBF kernels)	DNA sequence content, predicted DNA structure, evolutionary history and population variation, annotation of repeats, genes, regulatory regions, chromosomal bands and isochores, histone modification	Prediction of various epigenetic features (including DNA methylation)	AdaStump metrics: for all epigenome predictions: CC: 0.498, ACC: 0.749

(89)	Human Epigenome Project[4] data for chromosomes 6, 20, and 22, using methylation status for all samples, and Epigraph datasets[5]	Decision tree (best), SVM	Nucleotide content, evolutionary conservation, DNA structure prediction	CGI methylation status (2-way: methylated/unmethylated, or 4-way: methylation patterns across tissues)	Decision tree metrics: 2-way: CC:0.775, ACC: 0.9167; 4-way: CC: 0.707, ACC: 0.8939
(90)	Human Epigenome Project[4] data for chromosomes 6, 20, and 22, using methylation status in human CD4+ T lymphocytes	k-NN	5-mer frequency in 499bp upstream and downstream of CpG site	Methylation status of CpG sites	ACC: 0.7745
(91)	Human Epigenome Project[4] data for chromosomes 6, 20, and 22 across 1.9 million CpG sites, using methylation status in human CD4+ T lymphocytes	SVM (linear kernel)	DNA sequence derived features: GC content, GC observed/expected ratio, Alu repeats, and repeat masker. 214 TFBS and 38 histone marks.	CGI methylation status in chromosomes 6, 20, and 22	ACC: 0.94, CC: 0.81
(92)	Human Epigenome Project[4] data for chromosomes 6, 20, and 22, using methylation status in human CD4+ T lymphocytes	SVM	Sequence length, nucleotide and dinucleotide content, promoter and TFBS annotation, nucleosome positioning	Methylation status of CGI in chromosome 22	ACC: 0.9059, CC: 0.65
(93)	MethDB (curated database of ~5,000 experimentally determined methylation of DNA fragments in species from plants to humans)[1]	SVM (RBF kernel)	3-mer composition of DNA fragments	Methylation status and level for 400 human DNA fragments in MethDB	Methylation status prediction: ACC: 0.8207, MCC: 0.6411 Methylation level prediction: R: 0.8223, RMSE: 0.2042
(69)	Human Epigenome Project[4] data for chromosomes 6, 20, and 22, using methylation status in several human tissue or cell types	SVM	Gardiner-Garden criteria, 4-mer composition, conserved TFBSs and conserved elements, predicted DNA structure, functional annotation of proximal genes, nucleosome positioning, histone methylation and acetylation	Methylation status of CGI	Metric in human CD4+ lymphocyte: ACC: 0.9313, CC: 0.8302
(94)	BS-seq for H1 and IMR90 cell lines	Linear regression	Dinucleotide sequence derived features created using the sequence environment of 78bp. Each nucleotide interpreted as a categorical variable with 16 states.	DNA methylation levels at CpG nucleotides within partially methylated domains	R=0.86 (for the sequence context of 140bp)
(75)	Methylation array data of multiple human tissues	Support vector regression (RBF kernel) (best), linear regression	Methylation beta values in surrogate tissue	Methylation beta values for different tissues	Methylation level prediction: For probes in beta-value range 0.2 to 0.8: R^2: 0.89-0.98
(77)	BS-seq for H1, NPC, IMR90 cell lines	RF (best), SVM (RBF kernel), LR, Decision Tree, NB	Nucleotide composition, 16 histone marks, RNA-seq	Methylation status of genomic segments (based on CpG_MPs tool)	RF metrics: H1: AUC: 0.99, NPC: AUC: 0.99, IMR90: AUC: 0.92
(76)	100 blood samples for 450K arrays	RF	Sequence composition, evolutionary rate, copy number variation, haplotype score, recombination rate, SNP presence, annotation of gene body, promoters, CGIs, repeats, DNase, Pol2 and TF ChIP-seq, histone marks, neighboring CpG site methylation level and distance, chromatin states	Methylation status and levels at single CpG sites	Classification: CGI: ACC: 0.98, Whole genome: ACC: 0.92, Regression: R=0.9, RMSE=0.19
(78)	GM12878 and K562 cell lines (RRBS-seq)	Deep Nets (ANN) and SVM	Genomic features, neighboring CpG sites, and Hi-C	Methylation status at CpG dinucleotides across 1kb windows	ACC: 0.721-0.897
(79)	BS-seq and methylation arrays for H1 and H9 cell lines	RF (best), LR, SVM	Nucleotide, dinucleotide frequencies and NpN ratios for 500bp flanks, methylation data for 1000bp flanks, histone marks, chromosome organization, chromatin structure, evolutionary features, repeats, TFBS	Methylation status and levels at CpG sites	Metrics for RF: Classification: ACC: 0.93, MCC: 0.86, Regression: Spearman correlation coefficient: 0.7602

1.10 Importance of predicting DNA methylation and hydroxymethylation

Prediction of DNA methylation and hydroxymethylation remains important for several reasons. Despite the availability of high-throughput assays for querying DNA hydroxymethylation, there only exists a handful of publicly available TAB-seq and oxBS-seq datasets, and performing whole-genome BS-seq, TAB-seq, and oxBS-seq requires significant expenditure and skilled labor. Sequencing (or hybridization) based assays are also invasive and destructive procedures that may be unfeasible in certain experimental setups. It is also impossible to set up high-throughput assays for all cell or tissue types and every developmental stage, physiological condition or perturbation, necessitating *in silico* prediction. In such situations, reconstruction of the whole epigenome predicated upon available data for correlated traits and a predictive model trained on a similar cell type is a practical, economical and efficient way to query methylation or hydroxymethylation. Additionally, DNA sequencing based protocols have amplification and fragment selection steps, effectively creating a biased sampling procedure that nay cause a fraction of cytosines in the genome to be unrepresented or underrepresented in the survey. This is especially evident for protocols like RRBS-seq where only a small fraction of cytosines have reliable coverage for querying methylation (35). Such missing or low quality data can be imputed using predictive models, which can be trained using available high quality data. Also, inherent stochasticity of the sampling process makes it inevitable that some estimations of methylation levels using high coverage sequencing data can be potentially erroneous. However, *in silico* predictive models, trained using high-quality data with multiple input predictor variables, would be able to robustly predict DNA methylation status.

Aberrant genome-wide methylation patterns (not restricted to promoters or gene bodies) serve as early detection markers of multiple pathological disorders including cancer (95). Such studies cannot be successfully completed in missing data scenarios, suggesting an immense importance of *in silico* predictive models. Finally, the advent of new single cell sequencing technologies lead to a development of protocols such as single cell BS-seq (sc-BS-seq) (96), which are often unable to provide complete and sufficient genome-wide CpG coverage, therefore necessitating *in silico* prediction. Model based predictors remain relevant to date, as neural networks were recently used to successfully predict CpG methylation status of the 3000 bp windows in single cells (80), by employing epigenetic and sequence-derived features.

CHAPTER 2

INTER-METHYLOME SIMILARITIES: CONSENSUS REFERENCE METHYLOME AND ITS USAGE IN DNA METHYLATION PREDICTION

Authors: Milos Pavlovic, Pradipta Ray, Kristina Pavlovic, Aaron Kotamarti,
Min Chen and Michael Q. Zhang
Department of Biological Sciences
The University of Texas at Dallas
800 W. Campbell Road
Richardson, TX 75080-3021

2.1 Prior Publication

Milos Pavlovic (M.P.) performed the majority of experiments, and Pradipta Ray (P.R.) designed the majority of experiments. M.P. and P.R. wrote the manuscript. Min Chen (M.C.) advised M.P. and Michael Zhang (M.Q.Z.) supervised the project. This chapter introduces the concept of inter-methylome similarities and relates it to methylation prediction. Per the policy of OUP Bioinformatics, the publication of material in a PhD thesis is permitted with the publication of a peer-reviewed manuscript in their journal. The original manuscript (1) "DIRECTION: A machine learning framework for predicting and charactering DNA methylation and hydroxymethylation in mammalian genomes" by Milos Pavlovic, Pradipta Ray, Kristina Pavlovic, Aaron Kotamarti, Min Chen and Michael Q. Zhang, published in 2017, is reproduced by permission of Oxford University Press and appears online at the following web address: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btx316. Supplementary information is available online at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btx316. The main text is partially altered compared to the online version of the manuscript, and figures and tables do not chronologically correspond to the online manuscript numbering.

2.2 Abstract

Here we analyzed 25 reference methylomes from the NIH Roadmap Epigenome consortium, and discovered a portion of CpG sites whose methylation status was invariant across all analyzed methylomes. We utilized this information to perform methylation status prediction in H1 human embryonic stem cells (H1) and H1 derived neural progenitor cells (NPC) across CpG sites that

exhibited an invariant methylation status in the reference. We implemented an optional dictionary-based approach to perform such predictions, and further improved overall model-based prediction accuracy by creating an additional feature, which partitions a target methylome into its invariant and variant portion.

2.3 Introduction

Binding of DNMT1 to DNA is extremely selective and requires a linker DNA sequence of a minimum 20 bp in length (97), whereas the underlying sequence composition of a genomic region has been documented to shape DNA methylation patterns locally (22). The binding of DNMT1 to DNA results in a 6000 bp random walk of an enzyme and subsequent methylation of 50 CpG sites on average, resulting in spatially contiguous stretches of hypermethylated CpG sites, which are seldom interrupted by unmethylated CpGs. Accordingly, the accurate methylome predictions using sequence composition-derived features were performed in its own right (64). This suggests that a proportion of CpG sites have invariant methylation status across cell or tissue types and conditions, and therefore hold a great potential for the imputation of missing data, and provide potentially valuable insights into underlying aberrant methylation patterns.

2.4 Methods

The high predictive ability of DNA methylation predictive models which use only sequence derived features (in multiple datasets) suggests that a portion of DNA methylation status in CpG sites is governed by the underlying sequence, and should be unchanged across cell and tissue types and conditions. Therefore, we obtained 25 publicly available WGBS (Table 2.1) from the NIH Roadmap Epigenome consortium (98) excluding H1 and H1-derived cells, and estimated

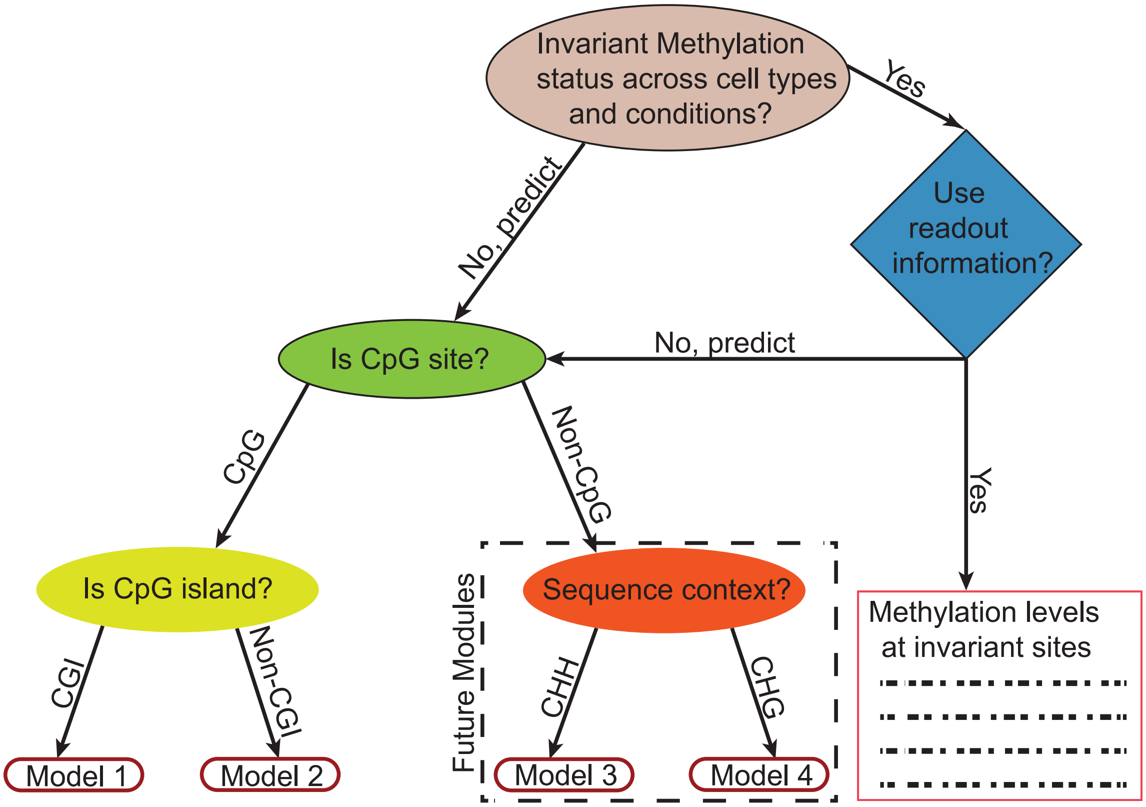


Fig. 2.1: DNA methylation reconstruction framework: Decision Tree for partitioning methylome based on different prediction paradigms such as model based predictions and dictionary-based driven predictions using invariant methylation readout.

its methylation status by thresholding the CCR at 0.5 for each cytosine, and compared the respective binary methylation statuses (high and low methylation) across all the CpG sites with coverage ≥ 5 across the 25 datasets. We identified a portion of CpG sites, which exhibited invariant methylation statuses across all analyzed methylomes, and we optionally used their methylation as an additional feature for performing whole methylome reconstruction or imputation in other datasets (Fig. 2.1).

Based on 25 high-quality reference human methylomes from the NIH Roadmap Epigenome consortium (98), we identified the majority methylation status for each CpG site with reliable sequencing depth across the 25 datasets.

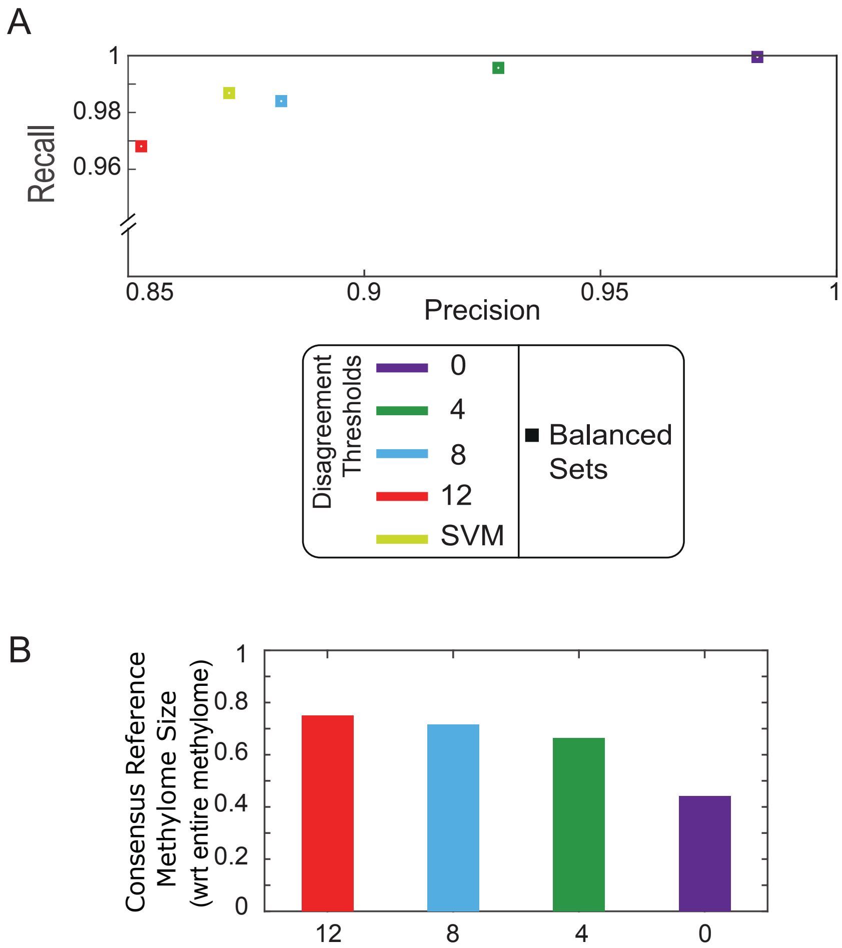


Fig. 2.2: DNA methylation predictions harnessing intra- and inter-methylome similarities A) Balanced sets predictions on methylation-invariant CpG sites using consensus reference methylome and SVM. B) Consensus Reference Methylome size as fraction of total methylome for disagreement thresholds 0, 4, 8, and 12.

We refer to the set of cytosines and their corresponding majority methylation status as the *consensus reference methylome.* We systematically decrease the set of cytosines by additionally constraining that no more than 8, 4, or none out of the 25 reference methylomes could be different from the methylation status of the majority of methylomes, referring to these variations as “consensus reference methylome with disagreement threshold n”. While determining methylation status in NPC using such consensus-based predictors, we identified a trade-off between accuracy and applicability. As we increase stringency of the disagreement criterion

Table 2.1: List of cell and tissue types that were used to create the reference methylome

Cell lines and tissues used to create reference methylome	
H9 Cell Line	Gastric
HUES64 Cell Line	Left Ventricle
iPS DF 6.9 Cell Line	Lung
iPS DF 19.11 Cell Line	Ovary
4star	Pancreas
IMR90 Cell Line	Psoas Muscle
Mobilzied CD34 Primary Cells Female	Right Atrium
Neurosphere Cultured Cells Cortex Derived	Right Ventricle
Penis Foreskin Keratinocycte Primary Cells skin03	Sigmoid Colon
Aorta	Small Intestine
Adult Liver	Thymus
Brain Hippocampus Middle	Spleen
Esophagus	

from 12 to 0, the prediction accuracy improves from 0.85 to 0.99 (on balanced sets) (Fig. 2.2 (A)), while the fraction of CpG sites in the genome that can be used to perform this prediction drops from 75% to 44% (Fig. 2.2 (B), Table 2.2). Given high predictive ability of the consensus reference methylome with zero disagreement, we optionally use this dictionary driven approach as a predictor to reconstruct a portion of the methylome.

Table 2.2: Relative size of the consensus reference methylome w.r.t to disagreement thresholds. Size of the consensus reference methylome with disagreement thresholds 0, 4, 8 and 12 as a fraction of the entire methylome (in terms of CpG cytosines).

Reference methylome sizes	Disagreement threshold			
Size of the reference methylome relative to the whole methylome	0	4	8	12
	0.44236051	0.664482316	0.716133503	0.751079765

2.5 Results

Based on the 44% of CpG sites that are methylation invariant in our reference we compared our SVM prediction model (detailed explanation behind creation and evaluation of the SVM model can be found in the Methods section Chapter 3) to the prediction based on the consensus reference methylome. Both predictors were highly accurate and comparable on the set of cytosines underlying the consensus reference methylome with zero mismatches, and on balanced subsets, the precision of the SVM was 0.87, compared to 0.99 of the most stringent consensus based predictor (Fig. 2.2 (A)). On whole genome datasets we noticed incremental improvement in NPC methylation prediction accuracy (0.97) as opposed to solely SVM or RF models that govern (0.96) accuracy (Table 3.9).

In addition to using a dictionary-based approach to perform predictions on the invariant portion of the methylome, we also created an additional feature that splits a target methylome into its invariant and variant portion, and used it in our model-based methylation status prediction framework. Further details on the predictive ability of the aforementioned feature will be discussed in the chapter 3.

2.6 Summary

In this chapter we introduced the concept of invariant DNA methylation and for that purpose we created the consensus reference methylome using the set of 25 publicly available methylomes from Roadmap Epigenome consortium (98). In order to improve methylation prediction accuracy in NPC, we employed the consensus reference methylome as an optionally driven standalone predictor, as well as an additional feature in our model-based methylation prediction framework (details in the results section chapter 3). The primary purpose of creating the consensus reference methylome was to use it as an additional predictive variable, while its usage in a context of a standalone predictive model was performed mostly as a feasibility study. The use of this feature is optional, and can be removed when required, i.e., when the primary goal is to identify differentially methylated regions for a cell type or tissue in question which is substantially different from the reference methylomes constituting the consensus reference methylome. The use of such predictor is probably most useful when only a limited number of input variables are present, such as resource-scarce scenarios. By creating the consensus reference methylome we paved the way for a new generation methylome reconstructions, which involve a synergy of model-based and dictionary-based prediction approaches to achieve high accuracy. Depending on the reconstructed methylome, the consensus reference methylome can be created using a different set of relevant reference methylomes, and can potentially provide insight into aberrant CpG methylation in perturbation or disease studies (such as cancer) known to affect DNA methylation (99).

CHAPTER 3

DNA METHYLATION PREDICTION

Authors: Milos Pavlovic, Pradipta Ray, Kristina Pavlovic, Aaron Kotamarti,
Min Chen and Michael Q. Zhang
Department of Biological Sciences
The University of Texas at Dallas
800 W. Campbell Road
Richardson, TX 75080-3021

3.1 Prior Publication

Milos Pavlovic (M.P.) performed the majority of experiments, and Pradipta Ray (P.R.) designed the majority of experiments. M.P. and P.R. wrote the manuscript. Aaron Kotamarti (A.K.) performed feature engineering. Min Chen (M.C.) advised M.P and Michael Zhang (M.Q.Z.) supervised the project. This chapter covers various aspects of DNA methylation prediction such as the choice of machine learning methods to perform such predictions, the rationale behind choosing the predictive methods and ultimately describes results generated employing such predictive methods. Per the policy of OUP Bioinformatics, the publication of material in a PhD thesis is permitted with the publication of a peer-reviewed manuscript in their journal. The original manuscript (1) "DIRECTION: A machine learning framework for predicting and charactering DNA methylation and hydroxymethylation in mammalian genomes" by Milos Pavlovic, Pradipta Ray, Kristina Pavlovic, Aaron Kotamarti, Min Chen and Michael Q. Zhang, published in 2017, is reproduced by permission of Oxford University Press and appears online at the following web address: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btx316. Supplementary information is available online at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btx316. The main text is partially altered compared to the online version of the manuscript, and figures and tables do not chronologically correspond to the online manuscript numbering.

3.2 Abstract

Here we performed DNA methylation status predictions at single nucleotide resolution in H1 and NPC cell lines on the whole-genome scale and balanced sets using supervised machine learning methods to achieve high accuracy. For that purpose we devised model-based learning algorithms, predicated upon SVMs and RFs. We implemented a sophisticated feature selection procedure to identify most important predictive variables. In addition to the model based approaches we utilize neighboring CpG site methylation information to predict methylation status of the nearby CpGs. Based on the consensus reference methylome we predicted DNA methylation at invariant CpG sites, and created an additional feature that splits the whole methylome into its invariant and variant portion to improve predictions on balanced sets. We extensively describe the overall architecture of our prediction framework; we justify the importance of conducting DNA methylation predictions and relate our findings to the published literature. Finally, we demonstrate significant biological interpretability of results generated by our prediction framework. The consensus reference methylome prediction and transfer learning DNA methylation status prediction are covered in chapter 2 and chapter 5 of this dissertation respectively.

3.3 Introduction

High-throughput assays to detect DNA methylation are expensive, unfeasible in certain contexts and often leave a portion of the methylome unqueried. For that purpose we developed a novel, supervised integrative learning framework to predict whole-genome methylation predictions in CpG dinucleotides. Our machine learning framework yields high-accuracy single nucleotide resolution predictions of DNA methylation (either 5-mC or 5-hmC) and solely 5-hmC modifications in mammalian model systems. Our publicly available tool DIRECTION (Discriminative Integrative whole Epigenome Classification at single nucleotide resoluTION) can be trained on shotgun sequencing-based mammalian methylation and hydroxymethylation datasets, by identifying and using available, correlated, high-throughput assays and genomic sequence-based traits as predictor variables. DIRECTION can be downloaded from: http://www.utdallas.edu/~prr105020/direction http://www.utdallas.edu/~mxp114330/direction

3.4 Methods

Even though DNA methylation is innately heritable (maintained through cell divisions) (15), a portion of it will be reprogrammed due to factors like genomic imprinting, progression of developmental stage or may be altered by “external noise” resulting from various environmental factors. We have built a supervised machine learning framework for methylation prediction using 3 basic assumptions: a) Homogeneity of the methylation states across all cells in the cell-type or tissue-type under question. b) Temporal stability of the 5-mC modification in a particular cell type. c) Imprinting loci are ignored: loci with one methylated and one unmethylated cytosine due to imprinting have so far been identified in approximately 100 mammalian genes (100), and are being ignored for the purposes of our model.

Bisulfite treatment protocols followed by short-read sequencing (BS-seq or TAB-seq) provide CCRs at single nucleotide resolution for cytosines ranging form 0 (unmethylated) to 1 (fully

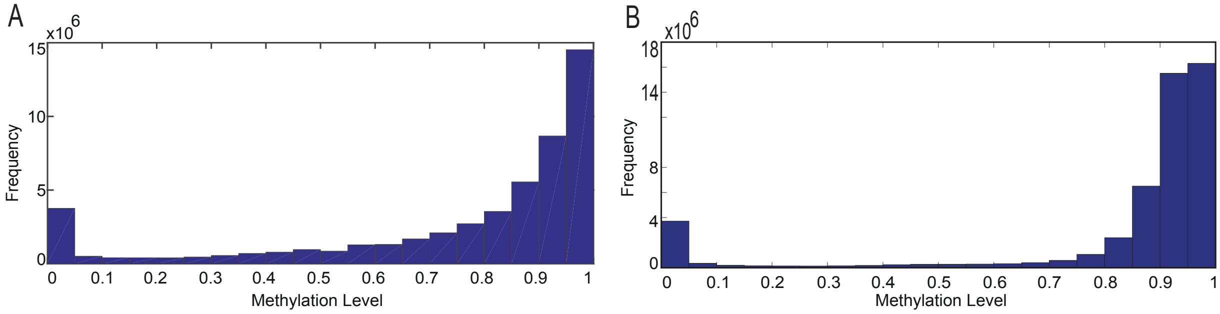


Fig. 3.1: Whole methylome empirical distributions of BS-seq CCRs in NPC (A) and liver tissue (B). Both cell types exhibiting bimodal genome-wide CCR distributions

methylated). Typically, prediction of an epigenetic mark can be appropriately performed using either classification or regression models depending on whether one is predicting a discrete or continuous-valued epigenetic trait. We formulate prediction of DNA methylation as a binary classification problem due to the bimodal nature of the distribution of CCRs in BS-seq experiments. Genome-wide empirical distributions of CCRs in mammalian reference methylomes (98) from inbred cell lines and sourced whole tissue (with low and high cellular heterogeneity respectively) show clear evidence of a bimodal distribution of CCRs (Fig. 3.1), with peaks close near CCRs of 0 and 1. Accordingly, we used a well-established CCR of 0.5 to distinguish between low and high methylation classes (76).

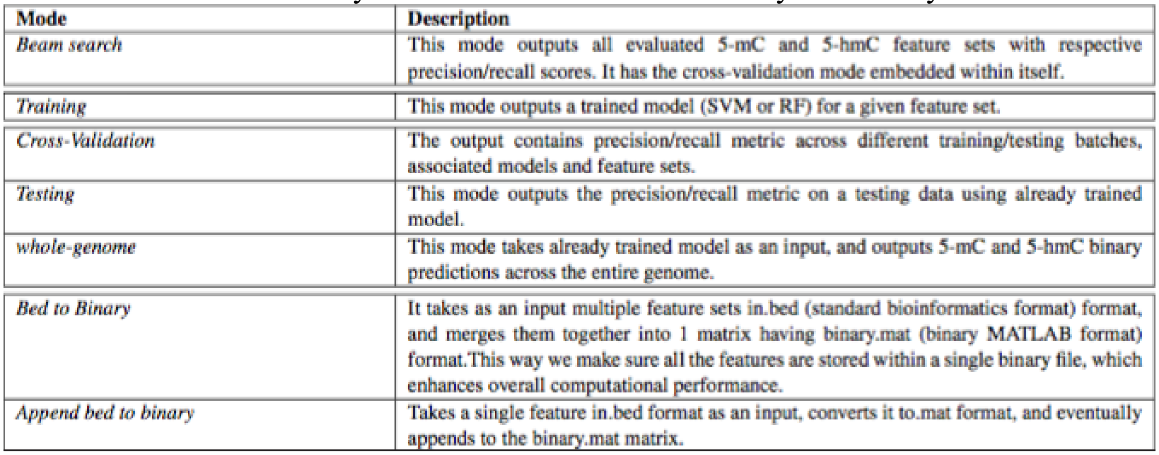
3.4.1 Tradeoffs underlying classification frameworks for methylation prediction

DNA methylation is essentially a discrete phenomenon at the level of individual alleles as there are only 3 possible methylation states that are distinguishable by bisulfite treatment followed by NGS: both alleles unmethylated (CCR of 0), one allele methylated one unmethylated (CCR of 0.5), and both alleles methylated (CCR of 1). Our premise of a two state methylation level is a simpler model, which captures the basic nature of DNA methylation (Fig. 3.1) while retaining relative simplicity over regression models. These evidences lend weight to the tractability of predicting 5-mC modifications. Therefore methylation prediction lends itself naturally to a classification framework. We aim to learn a function that will map a set of input features $\{x1, x2, \dots x_n\}$ to binary class labels $\{low, high\}$ for the purpose of reconstructing a discretized approximation of the BS-seq CCRs at individual cytosines, by using a CCR of 0.5 to threshold between low and high methylation classes. However, bisulfite based sequencing assays typically agglomerate signal form both alleles across millions of cells, thereby giving rise to CCRs that may be closer to 0.5 (intermediate methylation) than 0 or 1. Classification algorithms excel at predicting methylation status in cytosines with CCRs that have extremal values, where CCRs are near 0 or 1. However, their performance degrades when predicting methylation status in cytosines with CCRs commensurate with intermediate methylation levels due to the near-arbitrariness of class label assignments for such intermediate CCRs. DIRECTION is designed for predicting methylation at CpG cytosines in mammalian model systems, which are well known to exhibit a bimodal distribution of CCRs even in highly divergent and heterogeneous mammalian tissues such as muscle (101) and brain (102). Despite the fact that the majority of mammalian methylomes exhibit bimodal distribution, certain mammalian datasets can posses significant amount of intermediate methylation (103, 104). Cancer datasets, with their underlying mixture of cell-types and genome heterogeneity, can be a source of such abundant intermediate methylation in mammalian genomes (105). It is noteworthy that in invertebrates, the degree of intermediate methylation is known to be higher (103). In such situations, a regression-based approach is possibly more suitable (76). Given the flexibility of our prediction framework, such regression-based models can be conveniently incorporated when needed.

3.4.2 Direction toolkit and its uniqueness

Firstly, DIRECTION is able to deconfound effects of 5-mC and 5-hmC modifications, as it can be separately trained on BS-seq and TAB-seq datasets for a given cell type. This is the first time 5-hmC modifications have been predicted *in silico* (with a whole genome accuracy of 0.82), allowing us to systematically reconstruct 5-hmC modifications maps in different cell and tissue types. Secondly, DIRECTION provides different usage modes (Table 3.1) including imputation and whole methylome reconstruction (based on training a model in a related cell or tissue type).

Table 3.1: Summary on different modes of our methylation analysis toolkit



Mode	**Description**
Beam search	This mode outputs all evaluated 5-mC and 5-hmC feature sets with respective precision/recall scores. It has the cross-validation mode embedded within itself.
Training	This mode outputs a trained model (SVM or RF) for a given feature set.
Cross-Validation	The output contains precision/recall metric across different training/testing batches, associated models and feature sets.
Testing	This mode outputs the precision/recall metric on a testing data using already trained model.
whole-genome	This mode takes already trained model as an input, and outputs 5-mC and 5-hmC binary predictions across the entire genome.
Bed to Binary	It takes as an input multiple feature sets in.bed (standard bioinformatics format) format, and merges them together into 1 matrix having binary.mat (binary MATLAB format) format.This way we make sure all the features are stored within a single binary file, which enhances overall computational performance.
Append bed to binary	Takes a single feature in.bed format as an input, converts it to.mat format, and eventually appends to the binary.mat matrix.

This is possible because we do not use predictor variables likely to be relevant only in specific cell-types such as DNA-binding motifs of cell type restricted TFs, enabling transfer learning. Thirdly, DIRECTION is equipped with a sophisticated feature selection algorithm and is able to heuristically identify an optimal feature set (OFS) for predictions based on the set of available predictor variables (optionally using regional methylation patterns and methylation information from other cell types), allowing use in resource-poor scenarios and providing biologically interpretable results. DIRECTION performs methylation predictions at single nucleotide resolution, allowing us to collate predictions to any biologically relevant resolution such as CpG dinucleotide, CGI, or gene for purposes of downstream functional analysis. Direction implementation is based upon a novel decision tree based topology (Fig. 2.1), in which different classifiers correspond to each leaf of the tree. This tree partitions the methylome by selecting the most appropriate classifier given the availability of predictor variables and their efficacy on the basis of biologically relevant methylation paradigms. Additionally, we identified CpG sites with invariant methylation (see methods section chapter 2) by contrasting available reference methylomes, and implemented it as an optional feature for methylation prediction (see Results).

3.4.3 DIRECTION architecture

DIRECTION offers three primary modes of usage: for existing datasets, it can identify an OFS for predicting methylation or hydroxymethylation status based on available input feature sets, or impute low quality or missing data. Additionally, our toolkit allows us to perform whole methylome and hydroxymethylome reconstruction based on a user-provided feature set and SVM or RF model trained on a similar cell or tissue type. For other modes see (Table 3.1).

Machine learning based approaches, most prominently SVM and RF models have been successfully used to predict DNA methylation in the past (61, 76). Since we aim to perform genome-wide prediction, we chose not to use a single predictive model, but instead designed a scalable ensemble-learning framework that would be able to deconvolve multiple methylation paradigms that are at work in in different regions of the genome. For this purpose, a decision tree with a biologically motivated topology is used (Fig. 2.1) which partitions the methylome for methylation status prediction, based on available predictor variables and methylation paradigms. At each partition, we train separate predictive models predicated upon an SVM and RF, which exhibit comparable predictive accuracy. We also identified CpG sites with invariant methylation status across a set of high-quality reference methylomes, which can optionally be used as an additional feature to predict methylation status. With research on 5-hmC functionality still underway, and due to a lack of reference hydroxymethylomes, we used a single predictive model (SVM or RF) to perform 5-hmC status prediction.

3.4.4 Model-based classification: SVM and RF

As previously mentioned, we approach DNA methylation prediction as a binary classification problem, and for that purpose we devised an ensemble supervised learning framework predicated upon SVM and RF. The classification problem can be broken up into separate stages: the estimation or training stage: in which we use training data to learn a model, and the subsequent inference or testing stage: in which we use a trained model to make optimal class assignments (106). Thus we aim to learn a function $f(x): R^n \Rightarrow \{0,1\}$ that will map a set of input features $\{x1, x2 \ldots x_n\}$ to the binary class labels $\{low, high\}$. While some previous works have performed dimensionality reduction (69) we have decided to perform iterative feature selection in order to identify the OFS for methylation and hydroxymethylation prediction.

SVM classification: The SVM framework seeks to maximize the distance of training instances from the decision boundary in input space. It trains a hyperplane (or a set of hyperplanes for multi-class classification) based on the support vectors (a subset of data points closest to the decision boundary) of the training data. SVMs are a non-probabilistic, maximum margin classifiers. The optimization problem is thus to maximize the distance between the support vectors and the decision boundary (known as functional margin). Since a linear hyperplane may not necessarily suffice for good classification performance, most SVMs map input data to a higher dimensional space for separation by hyperplane, and use the kernel trick for calculating necessary pairwise inner products rather than performing all computations in the higher dimensional space. We chose to use the popular Radial Basis Function (RBF) kernel (106), previously used to predict methylation status (61). The RBF kernel between two input feature vectors x and x' is defined as: $f(x, x') = \exp\,(-\lambda \left|\left|x - x'\right|\right|^2$, and is proven to be robust with respect to other kernels such as linear and polynomial for classification purposes (106).

RF classification: RF is an ensemble-learning algorithm comprised of numerous decision trees (weak learners), well known for its high classification performance and resistance to overfitting. It averages predictions and feature weights across multiple decision trees and randomly samples subsets of features, subsequently separating class labels by splitting input features to optimize Gini Impurity or entropy (107).

We include both models into our framework since they have differing strengths: for example SVMs work well even with small training sets, while RFs are naturally resistant to outliers, thereby letting the user choose the model depending on the dataset and training data availability. It is noteworthy that these two models with comparable efficacy for our data.

3.4.5 Training and testing strategies for SVM and RF and prediction quality evaluation

Training and test set sizes were decided based on evaluation metric stability (Fig. 3.2). In order to evaluate the performance of our SVM and RF based predictive models we perform 5 fold cross-validation using balanced sets having 10,000 data points. The balanced sets are comprised of 5,000 positive and 5,000 negative examples, where 4,000 of each class are used for training and the remaining 1,000 of each class for testing. We discovered that the aforementioned design decisions govern the best trade-off between stably and accurately estimating prediction metrics versus computational time (Fig. 3.2). We thus chose k=5 for k-fold cross-validation on 10,000 sampled training examples (5,000 of each class) to balance out the trade-off between the training and testing set size. Namely, if k is too large the testing set size will be too small, and conversely if k is too small the training set size is too small and the number of experiments may not be enough to estimate the prediction performance.

It is worth noting that using more than 20,000 data points to train the SVM may cause the MATLAB built-in function svmtrain to be very slow, which may effectively result in non-convergence from a practical point of view.

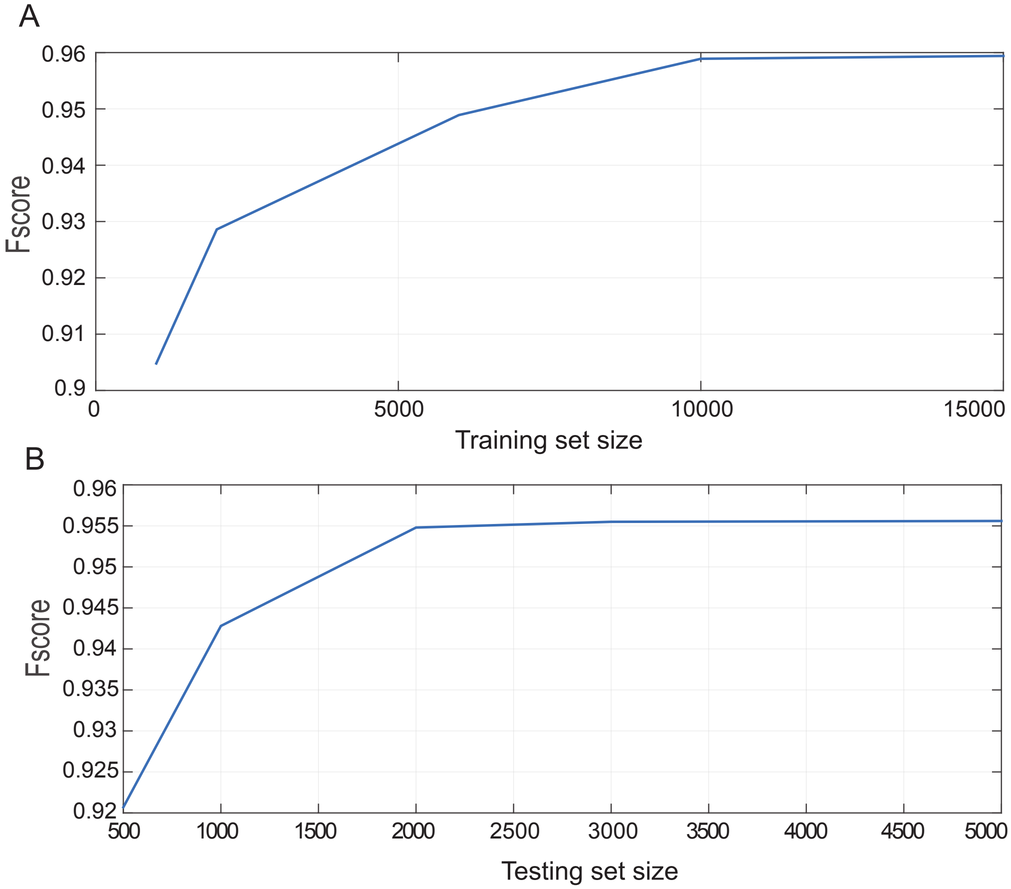


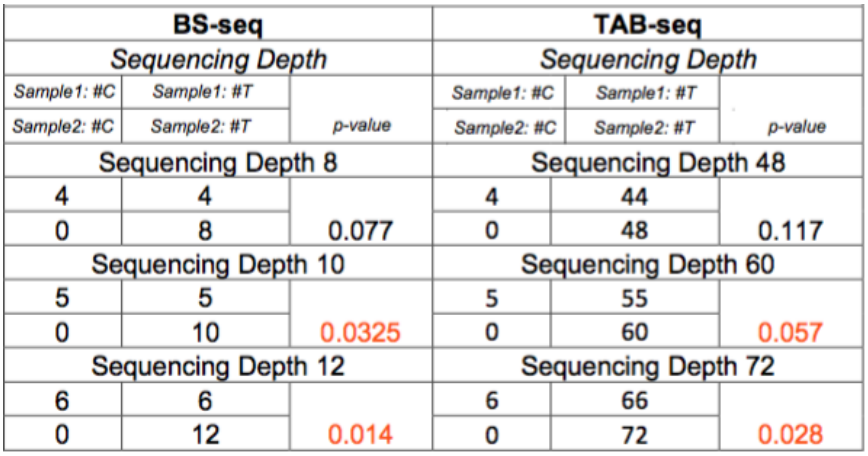
Fig. 3.2: Dependence of the performance metric based (F-score) based on: A) Training set size B) Test set size

Increasing label fidelity for training and testing samples: We identified the sequencing depth required for cytosines used for training (inclusion in the training set) and evaluating (inclusion in the testing set) our models based on the minimum sequencing depth that would always distinguish unmethylated (or non-hydroxymethylated) cytosine from marginally methylated or hydroxymethylated (CCR of 0.5 or 0.09 for BS- seq and TAB-seq respectively) given representative sampling. Due to sampling variance at low sample sizes causing small sample sizes to often not be representative, we performed a non-parametric categorical test (Fisher's Exact Test) between categorical distributions where for one sample the CCR is zero, versus another sample where CCR of the marginally methylated or hydroxymethylated sample is faithfully represented in the sample. We perform this over a range of sequencing depths fixed for both samples to identify when Fisher's Exact Test is able to identity a statistically significant difference between the two samples. This was performed to ensure label fidelity of training and testing samples. For BS-seq datasets, we need to minimally differentiate between completely unmethylated cytosines with a CCR of 0 with respect to marginally methylated cytosines with a CCR of 0.5. Given representative sampling, the minimum sequencing depth at a cytosine required to differentiate between the cases is two. However, we find that for the Fisher's Exact Test, we get a statistically significant p-value ($p \leq 0.05$) when sequencing depth for both samples is 10. In practice, for the SVM and RF models, both balanced set predictions and whole genome predictions were performed with cytosines where coverage ≥ 20. We find that out of 56,434,896 annotated CpG cytosines, 50,379,832 have coverage ≥ 20 in H1, and 49,134,499 have coverage ≥ 20 in NPC, suggesting that even in datasets with high sequencing depth, between 11% and 13% of cytosines do not have satisfactory coverage depth and can be imputed using DIRECTION. For the Reference Methylome predictor variable based predictions, and SVM model is compared with the Reference Methylome predictor, since sequencing depth ≥ 20 across all reference methylomes causes a large drop in the number of cytosines eligible for training and testing, a more modest sequencing depth constraint of ≥ 5 was used. Similarly, when Nearest

Neighbor evaluations were performed, the more modest sequencing depth constraint of ≥ 5 was used in order to capture more cytosines in the evaluation process. Additionally, Consensus Reference Methylome and Nearest Neighbor were introduced as input predictor variables into our toolkit, and cytosines with sequencing depth ≥ 5 were chosen for this purpose.

For TAB-seq datasets, we need to minimally differentiate between completely non-hydroxymethylated cytosines with a CCR of 0 with respect to marginally hydroxymethylated cytosines with a CCR of 0.09. Given representative sampling, the minimum sequencing depth at a cytosine required to differentiate between these cases is 20. We find that for the Fisher's Exact Test, we get a statistically significant p-value (p< or ~0.05) when sequencing depth for both samples is 60. In practice, for the SVM model, both balanced set predictions and whole genome predictions were performed with cytosines where coverage ≥ 60. See Table 3.2 for p-values obtained by Fisher exact test.

Table 3.2: Fisher's Exact Test p-values for various BS-seq and TAB-seq sequencing depths. Fisher's Exact Test shows statistical significance (p-value < or ~0.05) for distinguishing between a sample that is unmethylated (or non-hydroxymethylated) versus a sample that is marginally methylated (or hydroxymethylated) at sequencing depths of 10 for BS-seq data and 60 for TAB-seq data.



BS-seq			**TAB-seq**		
Sequencing Depth			*Sequencing Depth*		
Sample1: #C	*Sample1: #T*		*Sample1: #C*	*Sample1: #T*	
Sample2: #C	*Sample2: #T*	*p-value*	*Sample2: #C*	*Sample2: #T*	*p-value*
Sequencing Depth 8			Sequencing Depth 48		
4	4		4	44	
0	8	0.077	0	48	0.117
Sequencing Depth 10			Sequencing Depth 60		
5	5		5	55	
0	10	0.0325	0	60	0.057
Sequencing Depth 12			Sequencing Depth 72		
6	6		6	66	
0	12	0.014	0	72	0.028

SVM model decisions: The parameters used to train the SVM are as follows: kkt violation fraction =0.05, maximum number of sampled training sets used for training in order to achieve SVM convergence=3, maximum number of iterations in each training for SVM convergence $=10^7$. The average number of support vectors per 8000 training examples within different BS-seq optimal feature sets varied between 1200-1300, suggesting an upper bound of the experimental error rate range of 0.15-0.1625.

RF model decisions: When training the RF, we randomly sample one third of all available features in the training set, and perform sampling of training data-points with replacement. Splitting on input features is performed in a way that minimizes Gini Impurity score. Depending on the prediction paradigm we grow between 50 and 150 decision trees in the forest (for example CGI methylation status predictions can be successfully performed using 50 decision trees: when classification error reaches its minimum). Additional information about different modes implemented in our toolkit can be found in Table 3.1.

For evaluating predictions on balanced sets, we used Precision and Recall, F-score, and Area Under Curve (AUC). True Positive and True Negative Rates were used to evaluate whole genome predictions (see chapter 1.9 for details about evaluation metrics used in this dissertation). The metrics commonly used to assess the performance of a supervised learning algorithm belong to one of the following three categories: threshold metrics, rank metrics, or probability metrics (108). Since we perform classification using non-likelihood based approaches (SVM and RF), we use appropriate metrics in the “threshold- based” metrics category. The decision of which one to chose mostly depends on the nature of the problem that needs to be addressed. For prediction of skewed classes, special care needs to be taken such that the metric does not get inflated by simply predicting one class more often than the other. Concretely, we perform both methylation and 5-hmC predictions using balanced sets (avoiding skewed classes) and report the performance using Precision, Recall, F-Score (harmonic mean of Precision and Recall), and AUC while whole-genome prediction performance (where the frequency of the two classes are skewed for both methylation and 5-hmC status prediction) is evaluated using True Positive Rate (Sensitivity or Recall), True Negative Rate (Specificity) and Accuracy.

3.4.6 Feature engineering and feature selection

Feature engineering: We use a variety of genomic and epigenomic traits as input to train our classifier (Table 3.3). Features we do not model include gene annotation because histone modification data implicitly contain this information and enable us to discern between active, poised, and repressed cis-regulatory (57) and transcribed regions. Such annotation-based features may be incorporated when histone modification datasets are not available. Additionally, we do not model spatial contiguity explicitly into our predictive model. Since DNA methylation response variable (thresholded BS-seq CCRs) and various input features (e.g. histone modifications) are very well correlated spatially, our predictions are able to identify stretches of similar methylation without a need for explicit spatial auto-correlative models like Hidden Markov Model (HMM) or explicit spatial input features. TAB-seq CCRs are not spatially auto-correlated as well as BS-seq CCRs, but 5-hmC enriched regions and large stretches of 5-hmC depletion can be identified. Finally, features such as discriminative k-mers and motifs or ChIP-seq datasets of TF binding that can predict the methylation status were not used since the expression of such TFs are likely to be cell-type specific and accordingly not suitable for transfer learning purposes in the context of whole methylome reconstruction. Only the near ubiquitously expressed CTCF and p300 TF ChIP-seq data were used in the Initial Feature Set for predicting H1 methylation status, and these features were not used for NPC methylation status prediction, transfer learning for methylation status prediction, or 5-hmC status prediction.

All genomic features (tracks) such as Alu repeats, CGI as well as the genomic positions of CpG sites in the human genome (hg19 assembly) were obtained from the UCSC genome browser (109), or calculated based on the downloaded sequence and annotation. Histone mark ChIP-seq, DNase-seq and Transcription Factor binding ChIP-seq data (CTCF, p300) were obtained from the Roadmap Epigenome consortium (98) under the NCBI GEO GSE16256 accession (http://egg2.wustl.edu/roadmap/web_portal/processed_data.html). Genome-wide signal coverage tracks (negative log10 transform of the p-value) based on the uniformly processed Roadmap Epigenome Consortium datasets were used for ChIP-seq and DNase-seq features (98). All the raw features were matched against the list of available CpG sites using the IntersectBed tool from the Bedtools toolkit (110). After initial processing all the features were stored into a single matrix. The features were normalized to zero mean and variance one before training the model.

BS-seq and TAB-seq data sourcing and processing: BS-seq and TAB-seq datasets from the NIH Roadmap Epigenome consortium (98) were used for training and testing our predictive model. Read counts for estimating CCRs in H1 human embryonic stem cell (ESC) line and H1-derived NPC neural progenitor BS-seq datasets (GEO GSE16256) were obtained from the uniformly processed data published by the Roadmap Epigenome consortium (98) while the BISMARK tool (43) was used for mapping and obtaining the CCRs for H1 (GEO GSE36173) and NPC (GEO

GSM882245, GSM1463129) TAB-seq datasets. These cell types were chosen due to availability of BS-seq and TAB-seq data, and since previous studies performing functional enrichment and analysis of 5-hmC in human and mouse ESCs (17, 114, 116, 117) and neural progenitors (118-120), especially in neural development.

We have devised the pipeline for end-to-end mapping and variant calling of raw BS-seq and TAB-seq reads using the BISMARK BS-seq read mapper (43). Scripts that were used to calculate the reads sequencing depth and hydroxymethylation levels were coverage2cytosine and bismark methylation extractor. The final output to the .bed format was performed by the bismark2bedGraph. This was performed to generate H1 and NPC TAB-seq CCRs. H1, NPC, MSC, and IMR90 BS-seq CCRs were obtained from the uniformly processed datasets of the NIH Roadmap Consortium (http://egg2.wustl.edu/roadmap/web_portal/processed_data.html) processed from GEO series GSE16256 datasets by the Consortium as fractional methylation value and read coverage for each CpG cytosine.

For 5-hmC status prediction, BS-seq CCR (and not the predicted methylated status) was used as an input feature. An additional feature was created for methylation status imputation based on the methylation status of the CpG cytosine nearest to the cytosine in question (nearest neighbor feature, see section 3.4.7).

However, a similar feature was not used for 5-hmC status imputation since 5-hmC modifications do not occur in long stretches even though they can be somewhat locally enriched (chapter 4). Finally, based on the invariance of methylation statuses across reference methylome datasets

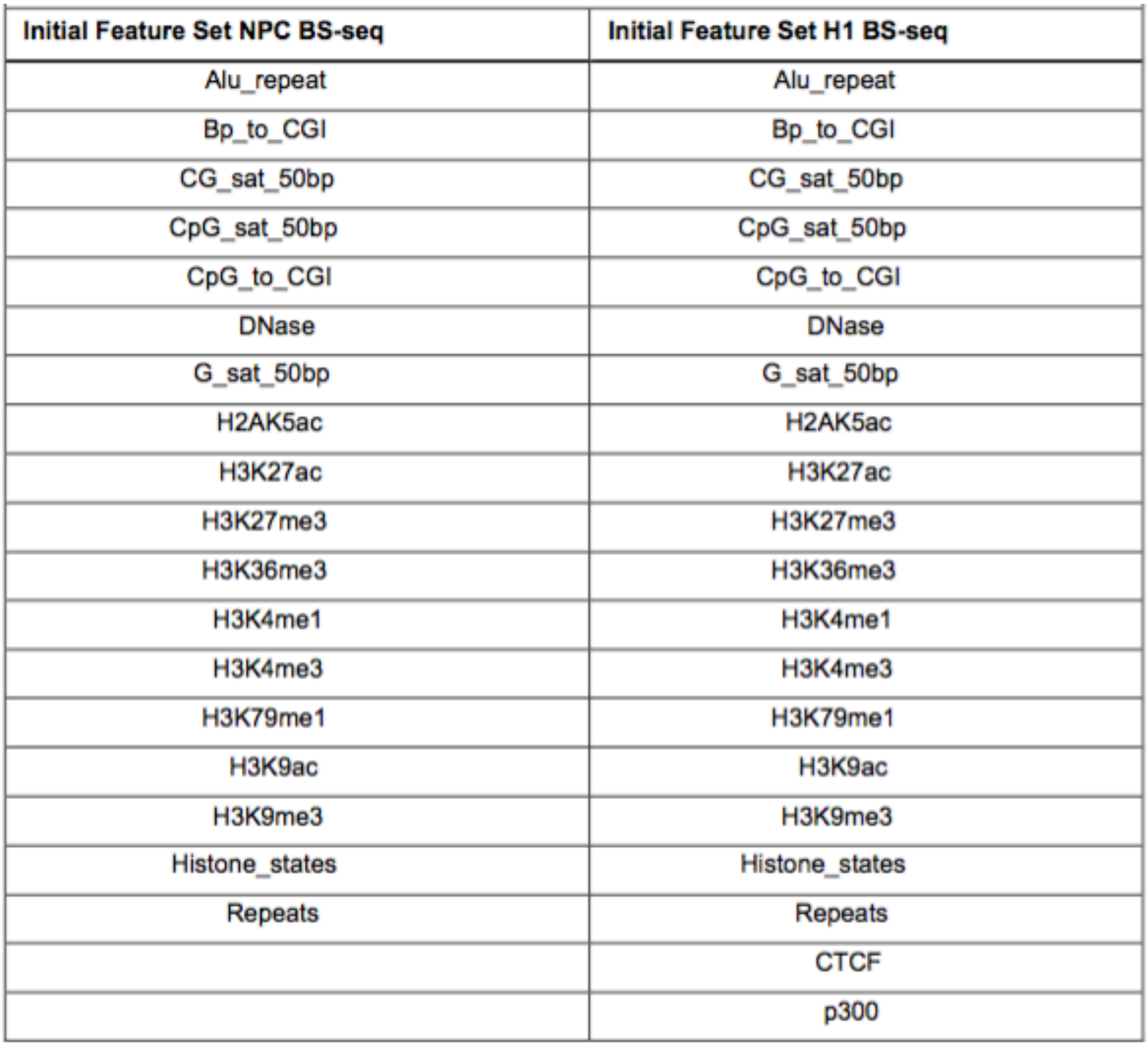
Table 3.2: List of features used for predicting DNA methylation and hydroxymethylation. All features for methylation prediction were used for 5-hmC predictions as well, since 5-hmC is on the demethylation pathway. [1](111), [2](112), [3](57)

Feature	**Type**	**Description**	**Motivation**
		Genome-derived features (processed from UCSC Genome Browser datasets (113)**)**	
CpG island (CGI)	Binary	Presence/absence of CGI annotation at CpG site	CGIs tend to be significantly unmethylated in comparison to non-CGI regions of the genome
Distance to nearest CGI (in bps)	Non-negative integer	Helps distinguish CpGs in CGI, CGI “shores” and non-CGI	Cytosines on the CGI shores (near CGIs) tend to be highly methylated and govern most of methylation within non-CGI regions)
Distance to nearest CGI (in CpGs)	Non-negative integer	Alternative feature for distance to CGI, measured in number of intervening CpGs, rather than genomic coordinates	As above
GC content	Continuous $\in [0,1]$	Percentage of nucleotides which are G/Cs in centered window around CpG site (window sizes: 50, 100, 200, 400, 800bp used)	Higher GC content empirically shows lower methylation levels: fact corroborated in CGIs
CpG density	Continuous $\in [0,1]$	Percentage of dinucleotides which are CpGs in centered window around CpG site (window sizes: 50, 100, 200, 400, 800bp used)	As above
Strand-specific guanine density	Continuous $\in [0,1]$	Percentage of guanines in centered window around CpG site (window sizes: 50, 100, 200, 400, 800bp used)	5-hmC levels can be asymmetrically distributed in a CpG site between strands (114)
Repeats (SINEs, LTRs)	Binary	Presence/absence of SINE or LTR annotation at the CpG site	Higher methylation suppresses transcription in repeat regions (115)
Alu	Binary	Presence/absence of Alu annotation at the CpG site	As above
		Epigenome-derived features	
Enhancers	Binary	Created using a cutoff value of the ChIP-seq H3K27ac and H3K4me3 signal generated using MACS tool[1]	5-hmC is known to be overrepresented in enhancers (49)
Core histone modification ChIP-seq signal	Continuous	$-\log_{10}$ transformed ChIP-seq p-values based on ChIP binding and input control, as calculated by the MACS tool[1]. (H3K9me3, H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K27ac: available for 109 epigenomes)[2]	Repressive marks like H3K9me3 and H3K27me3 are often mutually exclusive with DNA methylation
Auxiliary histone modification ChIP-seq signal	Continuous	Similarly processed data for additional histone modifications available for a limited number of epigenomes (H2AK5ac, H2AZ, H2BK120ac, H2BK12ac, H2BK15ac, H2BK20ac, H2BK5ac, H3K14ac, H3K18ac, H3K23ac, H3K23me2, H3K4ac, H3K4me1, H3K4me2, H3K56ac, H3K79me1, H3K79me2, H4K20me1, H4K5ac, H4K8ac, H4K91ac)[2]	As above
Histone states	Discrete: 1-15	Using core histone modification signal for core marks to segment data into posterior decoded 15-state HMM annotation tool ChromHMM[3], based on (Chadwick, 2012)	Histone states have been shown to be well correlated with DNA methylation (98)
BS-seq CCR	Continuous $\in [0,1]$	Percentage of cytosines remaining unchanged based on the Roadmap Epigenome consortium datasets[2]	Used only for predicting 5-hmC status, since 5-hmC modifications show up as part of the BS-seq CCRs
		ChIP-seq TF binding-derived features	
DNase-seq signal	Continuous	Regions of open chromatin characterized by DNase digestion and sequencing: coverage signal contrasted with uniformly distributed read set simulation, and $-\log_{10}$ transform of p-value used[2]	DNase hypersensitive regions positively correlated to active regulatory regions, negatively correlated to 5-mC
CTCF ChIP-seq signal	Continuous	$-\log_{10}$ transformed ChIP-seq p-values based on ChIP binding and input control for CTCF binding[2]	Well-known insulator. Used only for H1 methylation and 5-hmC status prediction.
p300 ChIP-seq signal	Continuous	$-\log_{10}$ transformed ChIP-seq p-values based on ChIP binding and input control for p300 binding[2]	p300 marks active transcription sites. Used only for H1 methylation and 5-hmC status prediction

we created an additional feature that splits the methylome into its invariant and variant portion, in order to improve balanced sets predictions in NPC (see results).

Initial feature elimination: We identified and eliminated redundant features based on feature clustering and reduced the size of the full feature set (listed in Table 3.3), and ultimately created

Table 3.3: Initial feature sets for NPC and H1 methylation status predictions



Initial Feature Set NPC BS-seq	**Initial Feature Set H1 BS-seq**
Alu_repeat	Alu_repeat
Bp_to_CGI	Bp_to_CGI
CG_sat_50bp	CG_sat_50bp
CpG_sat_50bp	CpG_sat_50bp
CpG_to_CGI	CpG_to_CGI
DNase	DNase
G_sat_50bp	G_sat_50bp
H2AK5ac	H2AK5ac
H3K27ac	H3K27ac
H3K27me3	H3K27me3
H3K36me3	H3K36me3
H3K4me1	H3K4me1
H3K4me3	H3K4me3
H3K79me1	H3K79me1
H3K9ac	H3K9ac
H3K9me3	H3K9me3
Histone_states	Histone_states
Repeats	Repeats
	CTCF
	p300

the “Initial Feature Sets” (IFS) (Table 3.4). We identify clusters of highly correlated features and keep only one representative feature for each cluster and eliminate the others.

The total set of predictor variables include several features that were engineered at multiple genomic resolutions (in bins of 50bp, 100bp, 200bp, 400bp, and 800bp) to predict DNA methylation and hydroxymethylation in genomic regions of corresponding size, and these naturally cluster in redundant groups. Since DIRECTION is trained to classify methylation and 5-hmC status at single nucleotide resolution, engineered features at the smallest resolution (50bp) were kept for the IFS, and the lower resolution features were discarded. These decisions

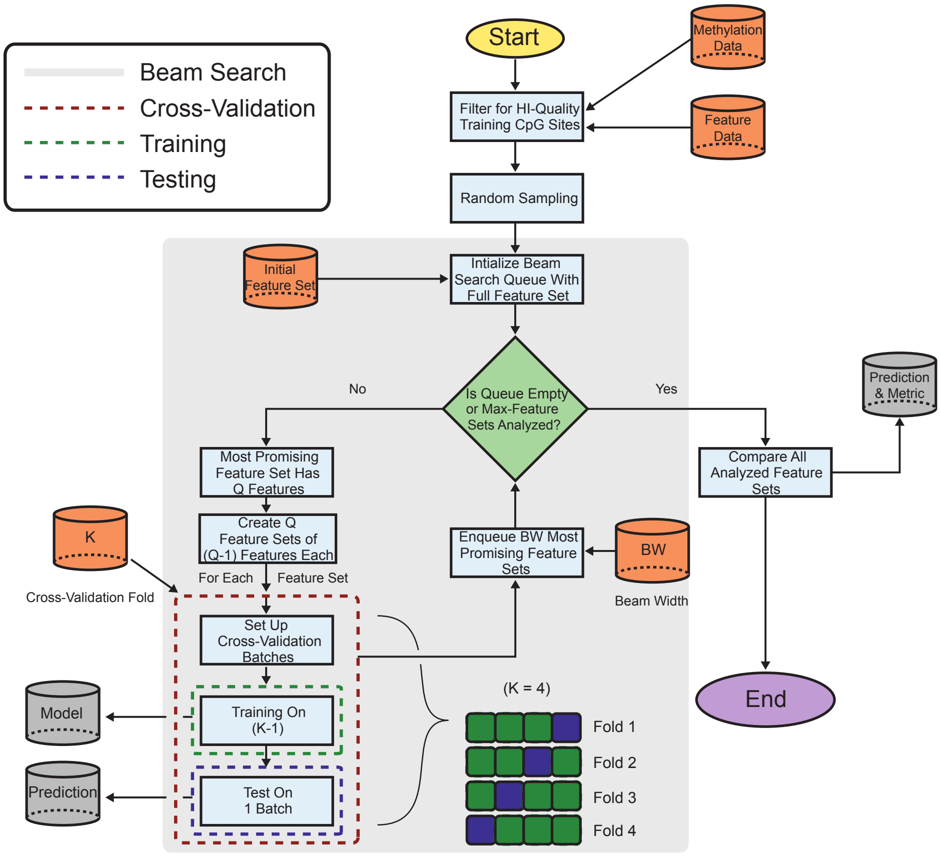

Beam Search
Cross-Validation
Training
Testing
Start
Methylation Data
Filter for HI-Quality Training CpG Sites
Feature Data
Random Sampling
Initial Feature Set
Intialize Beam Search Queue With Full Feature Set
No
Is Queue Empty or Max-Feature Sets Analyzed?
Yes
Prediction & Metric
Most Promising Feature Set Has Q Features
Compare All Analyzed Feature Sets
Create Q Feature Sets of (Q-1) Features Each
K
Cross-Validation Fold
For Each
Feature Set
Enqueue BW Most Promising Feature Sets
BW
Beam Width
Set Up Cross-Validation Batches
(K = 4)
Model
Training On (K-1)
Prediction
Test On 1 Batch
Fold 1
Fold 2
Fold 3
Fold 4
End

Fig. 3.3: Schema of our prediction framework outlining beam search (feature selection), training, testing, and cross-validation modes

resulted in saving a reasonable amount of computational time, and significantly reduced the possibility of overfitting our model.

Feature selection (beam search algorithm)**:** Typically, machine learning models with more input parameters tend to fit the response variable better, occasionally resulting in overfitting (121). This leads to a trade-off between predictive power and feature sparsity. Some previous approaches to perform optimal feature selection include dimensionality reduction (69) and removal of individual features from the full feature set to create the Gini index (76, 77), which will rank the features according to their contributions to the prediction metric. In order to gain additional insight about features and their additive effects we implemented a modified version of the recursive feature elimination algorithm that provides information about the discriminative nature of individual features and features subsets (Fig. 3.3).

Recursive feature elimination is a well-established strategy that was successfully used to determine the most predictive features and feature sets for methylation prediction (61). However, performing a top-down exhaustive search given a high number of input features (N) can be

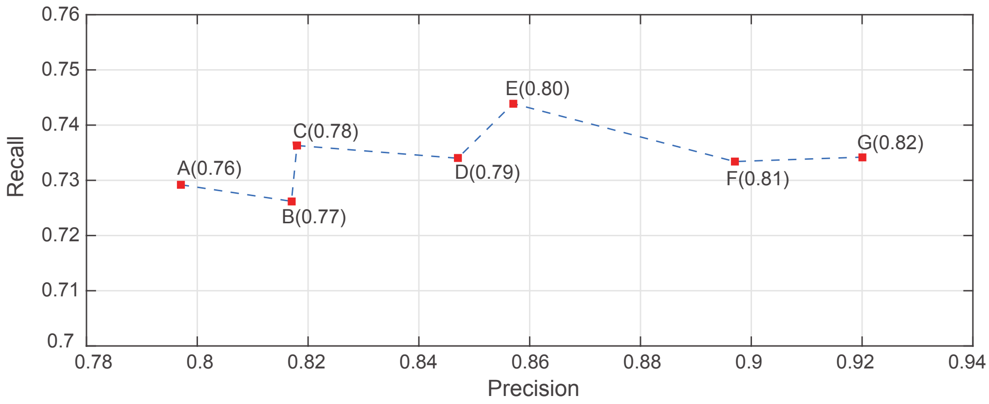
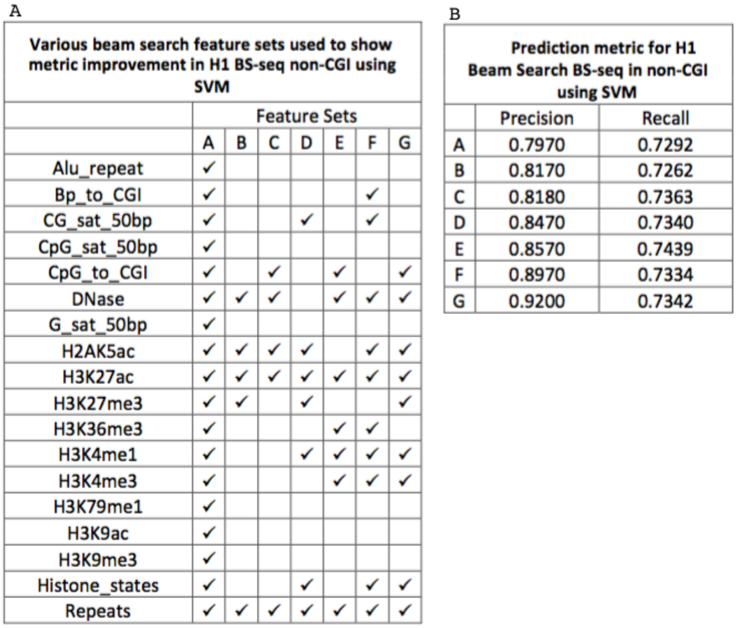


Fig. 3.4: An example path traversed by beam search through Precision-Recall space, while optimizing F-score (in parentheses) for H1 non-CGI SVM model.

extremely time consuming and computationally demanding since the number of explored feature sets may reach $2^N - 1$, leading us to consider heuristic approaches in determining the OFS.

All input features (listed in Table 3.3) were first preprocessed for use in our predictive framework. Identifying OFSs for classification is computationally intractable for a large number of input features (122) due to the curse of dimensionality. The problem is additionally complicated by the presence of noise in input features, label infidelity in the response variable, missing or low quality data for certain features, and high inter-feature correlation. While OFS selection and model training can be jointly performed (123) we heuristically identified an OFS using a recursive feature elimination strategy (Fig. 3.3) not limited to a specific learning algorithm, providing flexibility to choose a predictive model. Recursive feature elimination allows us to pick feature sets with fewer features that fit the data better in an iterative fashion,

Table 3.4: Underlying data for Figure 3.4, showing the F-score (for H1 non-CGI methylation status prediction) trajectory as the beam search algorithm searches through feature space. (A) For incrementally improved F-scores, the feature sets are shown. (B) For each improved F-score, the corresponding precision and recall values are shown.



A

Various beam search feature sets used to show metric improvement in H1 BS-seq non-CGI using SVM

	Feature Sets						
	A	B	C	D	E	F	G
Alu_repeat	✓						
Bp_to_CGI	✓					✓	
CG_sat_50bp	✓			✓		✓	
CpG_sat_50bp	✓						
CpG_to_CGI	✓		✓		✓		✓
DNase	✓	✓	✓		✓	✓	✓
G_sat_50bp	✓						
H2AK5ac	✓	✓	✓	✓		✓	✓
H3K27ac	✓	✓	✓	✓	✓	✓	✓
H3K27me3	✓	✓		✓			✓
H3K36me3	✓				✓	✓	
H3K4me1	✓			✓	✓	✓	✓
H3K4me3	✓				✓	✓	✓
H3K79me1	✓						
H3K9ac	✓						
H3K9me3	✓						
Histone_states	✓			✓		✓	✓
Repeats	✓	✓	✓	✓	✓	✓	✓

B

Prediction metric for H1 Beam Search BS-seq in non-CGI using SVM

	Precision	Recall
A	0.7970	0.7292
B	0.8170	0.7262
C	0.8180	0.7363
D	0.8470	0.7340
E	0.8570	0.7439
F	0.8970	0.7334
G	0.9200	0.7342

implicitly enforcing sparsity. We performed an initial feature elimination step based on inter-feature correlational redundancies (see *Initial Feature Elimination*).

We then conducted recursive feature elimination on the remaining features (Table 3.3) by implementing the beam search algorithm (Fig. 3.3): a classical artificial intelligence search procedure, utilizing heuristic pruning rules to explore a graph with nodes corresponding to all possible feature sets (124). Nodes (feature sets) are sorted in a queue according to classification evaluation metrics evaluated by 5-fold cross-validation, and the queued node having the highest metric is explored further by the algorithm until all nodes are evaluated or a maximum number of iterations are reached while simultaneously recoding the feature set with the optimum metric (Fig. 3.4, Table 3.4).

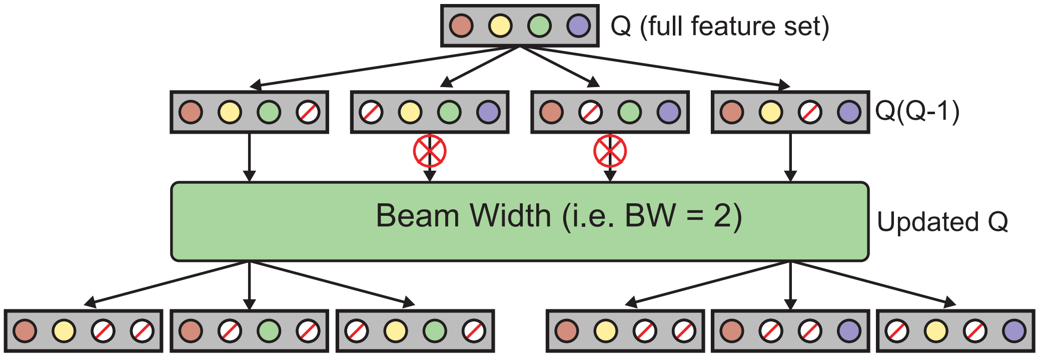


Fig. 3.5: Beam search algorithm feature set exploration shown for beam width = 2, for two levels of the search tree

The beam width parameter controls the number of nodes subject to further exploration and subsequent evaluation (Fig. 3.5). Across different beam width values, we find that beam search exhibits stability since it generates similar results (Table 3.5).

The algorithm for identifying optimal feature sets is shown as pseudocode (Table 3.6 and a flowchart Fig. 3.6). While OFSs can be optimized for multiple classification evaluation metrics (i.e., precision, recall) in our framework, in this paper "OFS" typically refers to the feature set corresponding to highest F-score metric, unless otherwise mentioned explicitly.

OFSs for NPC methylation status prediction that were obtained by optimizing different evaluation metrics are discussed in the results section. Finally, we examined contributions of individual features to the predictive ability of the OFS (see results Fig. 3.11).

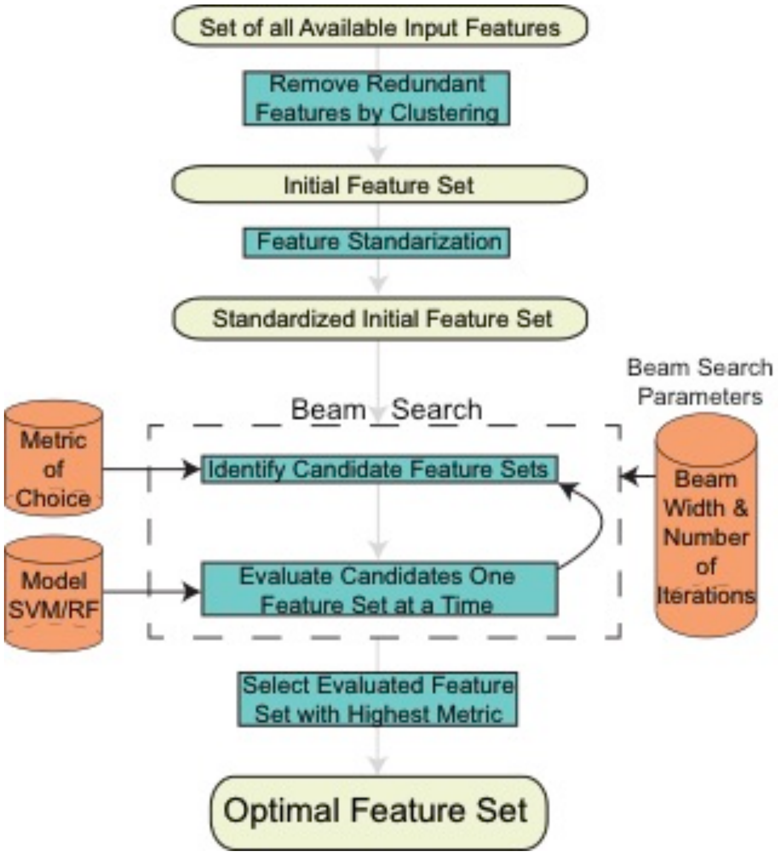
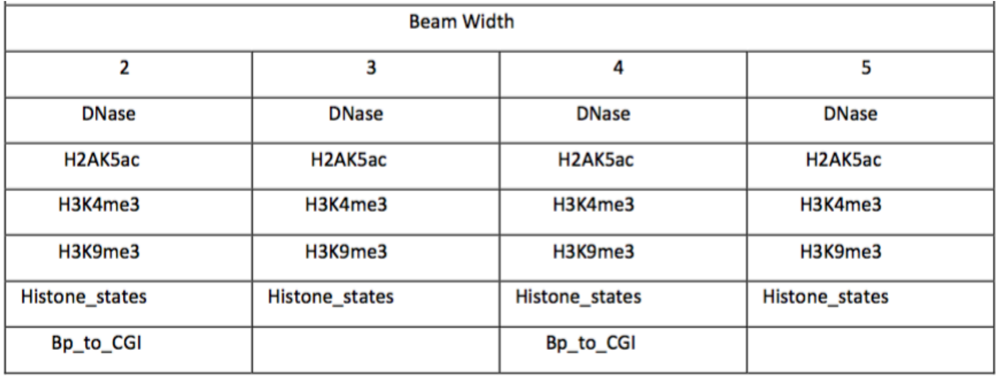

Set of all Available Input Features
Remove Redundant Features by Clustering
Initial Feature Set
Feature Standarization
Standardized Initial Feature Set
Beam Search Parameters
Beam Search
Metric of Choice
Identify Candidate Feature Sets
Beam Width & Number of Iterations
Model SVM/RF
Evaluate Candidates One Feature Set at a Time
Select Evaluated Feature Set with Highest Metric
Optimal Feature Set

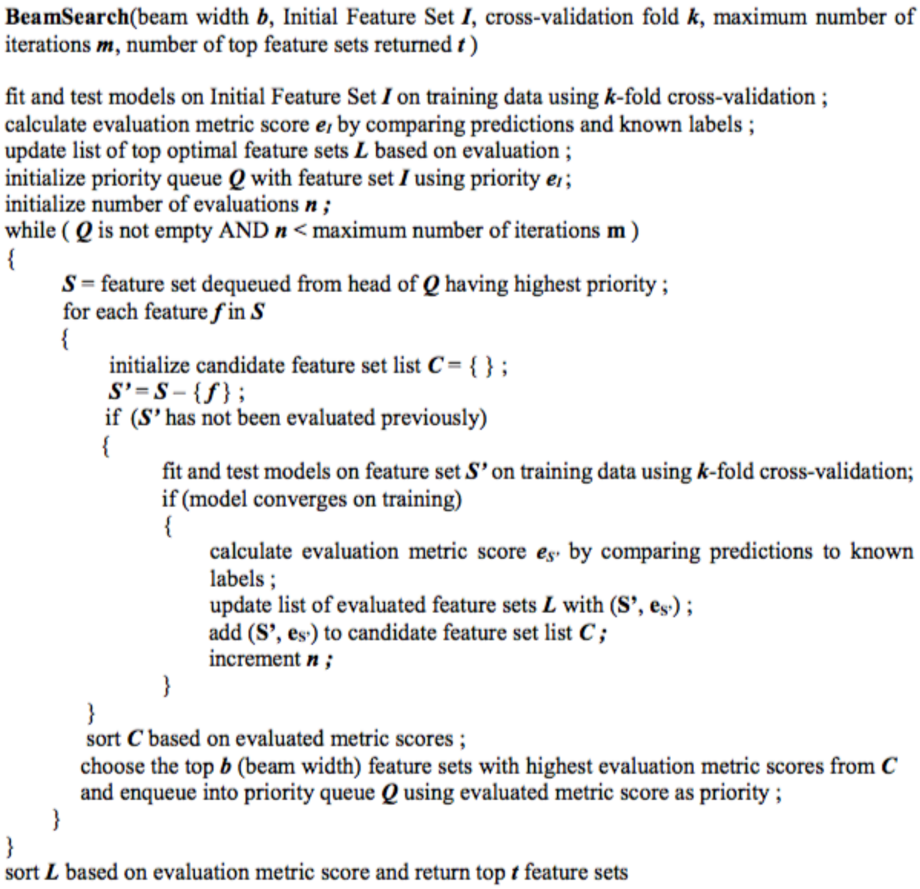
Fig. 3.6: Schematic representation of optimal feature set finding algorithm embedded within DIRECTION

Table 3.5: Similarity of OFSs across different beam width values for beam search using SVM model for methylation status prediction in NPC CGI dataset



Beam Width			
2	3	4	5
DNase	DNase	DNase	DNase
H2AK5ac	H2AK5ac	H2AK5ac	H2AK5ac
H3K4me3	H3K4me3	H3K4me3	H3K4me3
H3K9me3	H3K9me3	H3K9me3	H3K9me3
Histone_states	Histone_states	Histone_states	Histone_states
Bp_to_CGI		Bp_to_CGI	

Table 3.6: Beam search algorithm shown as a pseudocode



```
BeamSearch(beam width $b$, Initial Feature Set $I$, cross-validation fold $k$, maximum number of iterations $m$, number of top feature sets returned $t$ )

fit and test models on Initial Feature Set $I$ on training data using $k$-fold cross-validation ;
calculate evaluation metric score $e_I$ by comparing predictions and known labels ;
update list of top optimal feature sets $L$ based on evaluation ;
initialize priority queue $Q$ with feature set $I$ using priority $e_I$ ;
initialize number of evaluations $n$ ;
while ( $Q$ is not empty AND $n$ < maximum number of iterations $m$ )
{
        $S$ = feature set dequeued from head of $Q$ having highest priority ;
        for each feature $f$ in $S$
        {
              initialize candidate feature set list $C = \{ \}$ ;
              $S' = S - \{f\}$ ;
              if ($S'$ has not been evaluated previously)
              {
                     fit and test models on feature set $S'$ on training data using $k$-fold cross-validation;
                     if (model converges on training)
                     {
                           calculate evaluation metric score $e_{S'}$ by comparing predictions to known labels ;
                           update list of evaluated feature sets $L$ with ($S'$, $e_{S'}$) ;
                           add ($S'$, $e_{S'}$) to candidate feature set list $C$ ;
                           increment $n$ ;
                     }
              }
              sort $C$ based on evaluated metric scores ;
              choose the top $b$ (beam width) feature sets with highest evaluation metric scores from $C$ and enqueue into priority queue $Q$ using evaluated metric score as priority ;
        }
}
sort $L$ based on evaluation metric score and return top $t$ feature sets
```

3.4.7 Exploiting correlation between datasets

We engineered several predictor variables based on methylation status of neighboring CpG sites, previously used to impute methylation data (76). Cytosines in CpG sites were divided into “high-coverage” and “low-coverage” sets (sequencing depth at CpG site in the dataset was ≥ or < 5) in NPC. To predict methylation status at each low-coverage cytosine, we compared predictive abilities of the methylation status of the three nearest high-coverage CpG sites to the CpG site in question. We additionally contrasted another predictor constructed by using the most common methylation status (performing a majority vote) across the three nearest high-coverage sites.

We find that the precision of prediction drops from the nearest to furthest neighbor, and methylation status of the nearest neighbor’s predictive performance is comparable to the majority methylation status of the three nearest neighbors (Fig. 3.7).

We analyzed the predictive quality of the nearest neighbor based on distance between the predicted CpG site and the nearest neighbor. As distance increases from contiguous up to

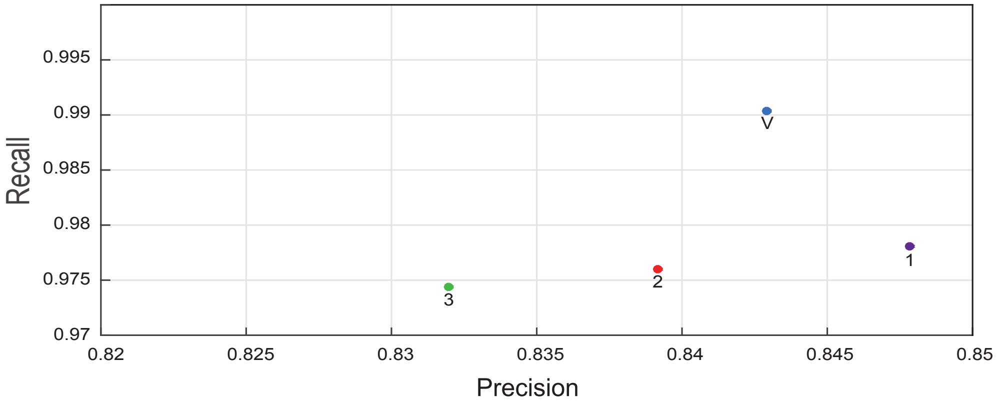


Fig. 3.7: Precision/Recall plot for balanced sets prediction of DNA methylation status using the methylation status of the first, second and third nearest neighbor, and a vote amongst all three.

2500bp, both precision and recall decrease (Fig. 3.8), with a significant drop after 500bp. Thus, methylation status of the nearest neighboring high-coverage CpG site within 500bp was used as a discriminative predictor variable.

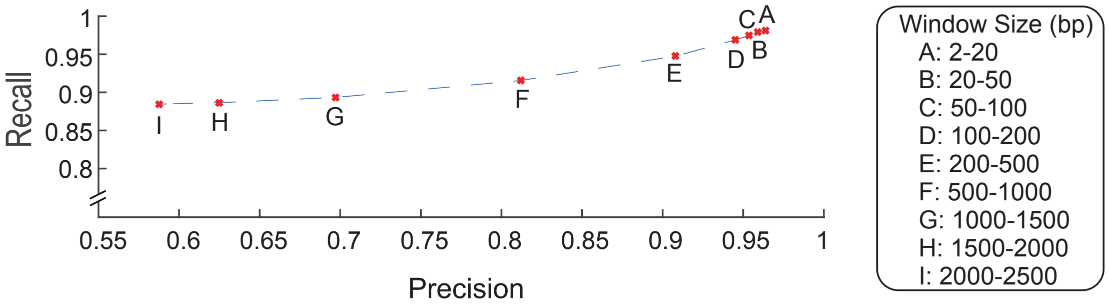


Fig. 3.8: Precision/Recall plot for balanced sets methylation status imputation using methylation status of nearest neighboring CpG site as function of distance to nearest neighbor (Table 3.9 for results)

Since the predictive power of neighboring CpG sites drops with distance, we wanted to determine what fraction of CpG sites with low coverage (<5) have high coverage (≥5) neighboring CpG sites within 500bp, making them a good candidate for imputation. Therefore we computed the Cumulative Distributive Function (CDF) of the fraction of low coverage sites with respect to distance to the nearest high coverage neighboring site (Fig. 3.9), in high coverage

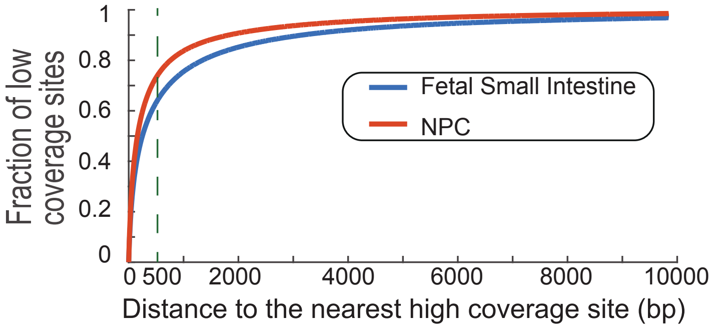


Fig 3.9: Cumulative Distribution Function of the fraction of low coverage CpG sites w.r.t distance to the nearest high coverage site in a typical high-coverage and low-coverage BS-seq dataset (NPC and fetal small intestine respectively)

(NPC) and low coverage (Fetal Small Intestine) Roadmap Epigenome consortium datasets.

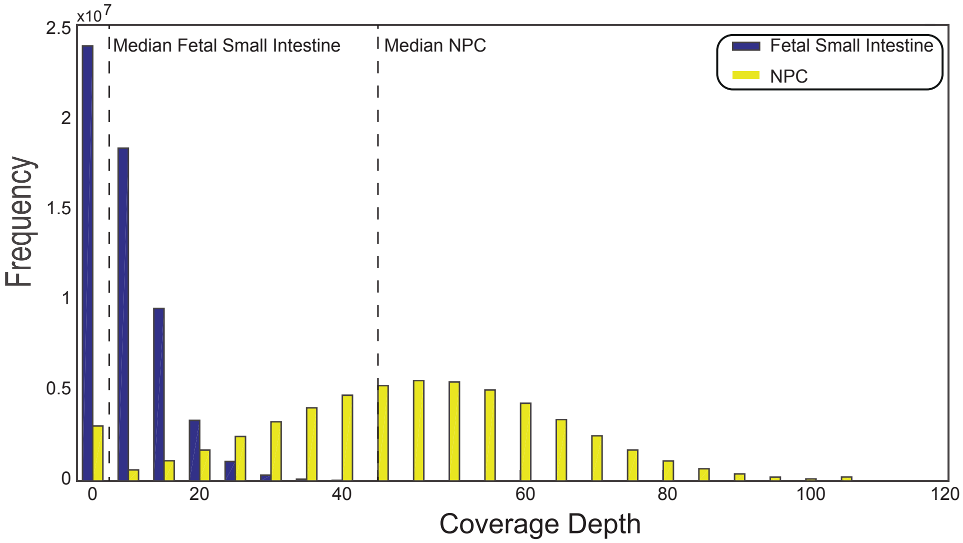


Fig. 3.10: Empirical distributions of high coverage methylome such as NPC (yellow) and low coverage methylome such as Fetal Small Intestine (blue)

Even in a low coverage methylome such as Fetal Small Intestine (Fig. 3.10), more than 60% of low coverage CpG sites had a corresponding high coverage neighbor within 500bp, suggesting high probability of them being correctly imputed (Fig. 3.8). Since a large fraction of CpG sites have a high coverage neighbor within 500bp even for moderately sized BS-seq datasets (Fig. 3.9), this feature was added to the beam search-identified OFS and the model was retrained for imputation.

3.5 Results

BS-seq datasets from the NIH Roadmap Epigenome consortium were used for training and testing our predictive model. Read counts for estimating CCRs in H1 human embryonic stem cell line and H1-derived NPC neural progenitor BS-seq datasets (GEO GSE16256) were obtained from the uniformly processed data published by the Roadmap Epigenome consortium (98).

3.5.1 DNA methylation prediction

Since there is no precedent for *in silico* prediction of the 5-hmC modification, we first built a framework for conventional two-state classification of DNA methylation in CpG sites, supervised using BS-seq data. Since distributions and spatial contiguity patterns of highly and lowly methylated CpG sites vary between CGI and non-CGI regions, we trained two classifiers with separately inferred OFSs (Fig. 2.1, Model 1, Model 2). Significant differences in prediction quality were observed among different feature sets (agreeing with previous studies (61, 76) suggesting the importance of feature set selection.

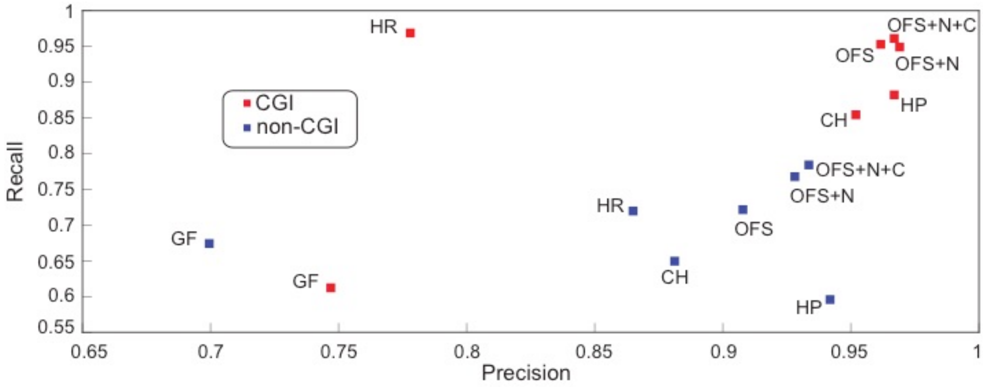


Fig. 3.11: CGI and non-CGI SVM model DNA methylation status predictions using balanced sets in NPC: GF (Genomic Features), CH (Chromatin Features), HR (Highest Recall Features), HP (Highest Precision Features), OFS (Highest F-Score Features), OFS+N (OFS+nearest neighbor), OFS+N+C (OFS+N+consensus reference methylome)

We performed optimal feature selection using our beam search algorithm, and identified feature sets with the best precision, recall, and harmonic mean of the two (F-score) for training and testing balanced sets of both classes in H1 and NPC with minor performance differences (NPC: Fig. 3.11, H1: Fig. 3.12 (A, B)). Whole genome predictions (Table 3.7) were performed subsequently (Table 3.9 for detailed results). The whole genome predictions were also used to assess the performance of DIRECTION across varying values of BS-seq CCRs (see next section).

DIRECTION more accurately predicts positive than negative methylation class (Table 3.8). This can be attributed to the fact the majority of mammalian methylomes are highly methylated (Fig. 3.1), and that highly methylated CpGs often occur in long stretches that are seldom interrupted by lowly methylated CpGs. Since we use histone marks (which often occur in large domains) as input, our model will inevitably misclassify some of aforementioned lowly methylated CpGs.

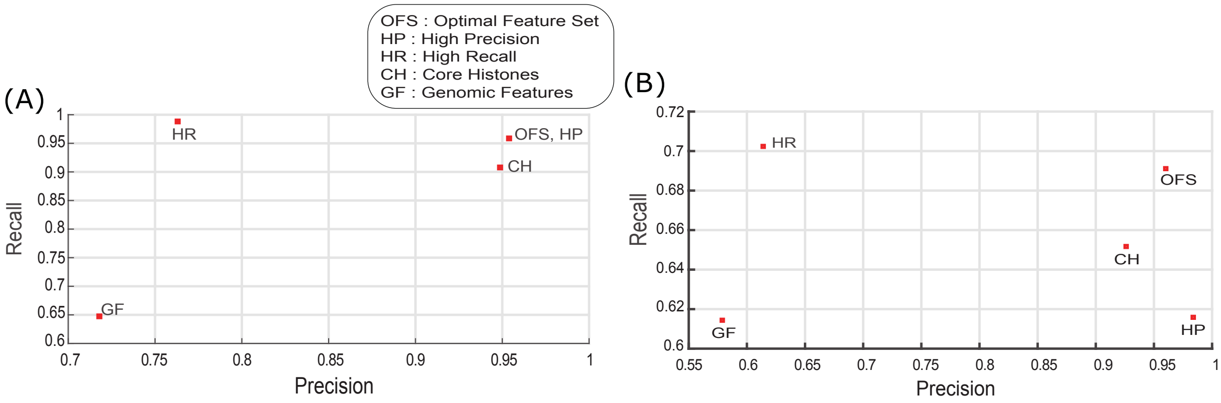


Fig. 3.12: Prediction metrics for DNA methylation balanced sets status prediction in H1 cells CGI (A) and non-CGI (B) regions.

Table 3.7: Whole genome BS-seq status prediction evaluation in NPC dataset using SVM

	CGI	**non-CGI**	**Total**
True Positive Rate	0.99	0.99	0.99
True Negative Rate	0.81	0.68	0.7
Accuracy	0.96	0.96	0.96

Table 3.8: Balanced set evaluations for DNA methylation prediction

Evaluation on genomic loci subsets by sampling balanced sets

Comparison of different predictive models in NPC dataset

				Datasets		Evaluation metrics		
Data type	**Sampling loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**Precision**	**Recall**	**Fscore**
BS-seq	CGI cytosines	SVM	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 20)	0.96	0.95	0.95
BS-seq	CGI cytosines	RF	(RF OFS)	NPC (depth >= 20)	NPC (depth >= 20)	0.95	0.96	0.95
BS-seq	CGI cytosines	Classification Tree	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 20)	0.94	0.95	0.94
BS-seq	CGI cytosines	Ensemble model with SVM and	SVM + N + C = (SVM features: SVM OFS + nearest neighbor	NPC (depth >= 20)	NPC (depth >= 20)	0.97	0.96	0.96
BS-seq	non-CGI cytosines	SVM	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 20)	0.91	0.72	0.80
BS-seq	non-CGI cytosines	RF	(RF OFS)	NPC (depth >= 20)	NPC (depth >= 20)	0.89	0.74	0.81
BS-seq	non-CGI cytosines	Classification Tree	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 20)	0.71	0.71	0.71
BS-seq	non-CGI cytosines	with SVM and consensus	SVM OFS + nearest neighbor feature) + Consensus	NPC (depth >= 20)	NPC (depth >= 20)	0.93	0.78	0.85

Comparison of predictive abilities for different feature sets in NPC dataset

				Datasets		Evaluation metrics		
Data type	**Sampling loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**Precision**	**Recall**	**Fscore**
BS-seq	CGI cytosines	SVM	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 20)	0.96	0.95	0.95
BS-seq	CGI cytosines	SVM	GF	NPC (depth >= 20)	NPC (depth >= 20)	0.75	0.61	0.67
BS-seq	CGI cytosines	SVM	CH	NPC (depth >= 20)	NPC (depth >= 20)	0.95	0.85	0.90
BS-seq	CGI cytosines	SVM	HP	NPC (depth >= 20)	NPC (depth >= 20)	0.97	0.88	0.92
BS-seq	CGI cytosines	SVM	HR	NPC (depth >= 20)	NPC (depth >= 20)	0.78	0.97	0.86
BS-seq	CGI cytosines	SVM	SVM + N = (SVM features: SVM OFS + nearest neighbor feature)	NPC (depth >= 20)	NPC (depth >= 20)	0.97	0.95	0.96
BS-seq	CGI cytosines	with SVM and consensus reference methylome based	SVM + N + C = (SVM features: SVM OFS + nearest neighbor feature) + Consensus Reference Methylome	NPC (depth >= 20)	NPC (depth >= 20)	0.97	0.96	0.96
BS-seq	non-CGI cytosines	SVM	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 20)	0.91	0.72	0.80
BS-seq	non-CGI cytosines	SVM	GF	NPC (depth >= 20)	NPC (depth >= 20)	0.70	0.67	0.68
BS-seq	non-CGI cytosines	SVM	CH	NPC (depth >= 20)	NPC (depth >= 20)	0.88	0.65	0.75
BS-seq	non-CGI cytosines	SVM	HP	NPC (depth >= 20)	NPC (depth >= 20)	0.94	0.60	0.73
BS-seq	non-CGI cytosines	SVM	HR	NPC (depth >= 20)	NPC (depth >= 20)	0.87	0.72	0.79
BS-seq	non-CGI cytosines	SVM	SVM + N = (SVM features: SVM OFS + nearest neighbor feature)	NPC (depth >= 20)	NPC (depth >= 20)	0.93	0.77	0.84
BS-seq	non-CGI cytosines	with SVM and consensus reference methylome based predictor	SVM + N + C = (SVM features: SVM OFS + nearest neighbor feature) + Consensus Reference Methylome	NPC (depth >= 20)	NPC (depth >= 20)	0.93	0.78	0.85

Comparison of predictive abilities for different feature sets in H1 dataset

				Datasets		Evaluation metrics		
Data type	**Sampling loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**Precision**	**Recall**	**Fscore**
BS-seq	CGI cytosines	SVM	OFS	H1 (depth >=20)	H1 (depth >=20)	0.96	0.96	0.96
BS-seq	CGI cytosines	SVM	GF	H1 (depth >=20)	H1 (depth >=20)	0.72	0.65	0.68
BS-seq	CGI cytosines	SVM	CH	H1 (depth >=20)	H1 (depth >=20)	0.95	0.91	0.93
BS-seq	CGI cytosines	SVM	HP	H1 (depth >=20)	H1 (depth >=20)	0.96	0.96	0.96
BS-seq	CGI cytosines	SVM	HR	H1 (depth >=20)	H1 (depth >=20)	0.76	0.99	0.86
BS-seq	non-CGI cytosines	SVM	OFS	H1 (depth >=20)	H1 (depth >=20)	0.96	0.69	0.80
BS-seq	non-CGI cytosines	SVM	GF	H1 (depth >=20)	H1 (depth >=20)	0.56	0.62	0.59
BS-seq	non-CGI cytosines	SVM	CH	H1 (depth >=20)	H1 (depth >=20)	0.93	0.65	0.77
BS-seq	non-CGI cytosines	SVM	HP	H1 (depth >=20)	H1 (depth >=20)	0.98	0.62	0.76
BS-seq	non-CGI cytosines	SVM	HR	H1 (depth >=20)	H1 (depth >=20)	0.61	0.70	0.65

Comparisons of predictions involving the Consensus Reference Methylome

				Datasets		Evaluation metrics		
Data type	Sampling loci constraints	Predictive Model	Input Features Used	Trained on	Tested on	Precision	Recall	Fscore
BS-seq	Cytosines with disagreement threshold = 0	SVM	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 5)	0.87	0.99	0.93
BS-seq	Cytosines with disagreement threshold = 0	Single predictor variable	Consensus Reference Methylome	NPC (depth >= 5)	NPC (depth >= 5)	0.98	0.99	0.98
BS-seq	Cytosines with disagreement threshold <= 4	Single predictor variable	Consensus Reference Methylome	NPC (depth >= 5)	NPC (depth >= 5)	0.93	0.99	0.96
BS-seq	Cytosines with disagreement threshold <= 8	Single predictor variable	Consensus Reference Methylome	NPC (depth >= 5)	NPC (depth >= 5)	0.88	0.98	0.93
BS-seq	Cytosines with disagreement threshold <= 12	Single predictor variable	Consensus Reference Methylome	NPC (depth >= 5)	NPC (depth >= 5)	0.85	0.97	0.91
BS-seq	CGI cytosines	with SVM and consensus reference methylome based predictor	SVM + N + C = (SVM features: SVM OFS + nearest neighbor feature) + Consensus Reference Methylome	NPC (depth >= 20)	NPC (depth >= 20)	0.97	0.96	0.96
BS-seq	non-CGI cytosines	with SVM and consensus reference methylome based predictor	SVM + N + C = (SVM features: SVM OFS + nearest neighbor feature) + Consensus Reference Methylome	NPC (depth >= 20)	NPC (depth >= 20)	0.93	0.78	0.85

Comparisons of predictions involving the Nearest Neighbor Methylation Status predictor

				Datasets		Evaluation metrics		
Data type	Sampling loci constraints	Predictive Model	Input Features Used	Trained on	Tested on	Precision	Recall	Fscore
BS-seq	All cytosines with nearest neighbor distance within 2 - 20 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	0.96	0.98	0.97
BS-seq	All cytosines with nearest neighbor within distance 20 - 50 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	0.96	0.98	0.97
BS-seq	All cytosines with nearest neighbor within distance 50 - 100 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	0.95	0.97	0.96
BS-seq	All cytosines with nearest neighbor within distance 100 - 200 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	0.95	0.97	0.96
BS-seq	All cytosines with nearest neighbor within distance 200 - 500 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	0.91	0.95	0.93
BS-seq	All cytosines with nearest neighbor within distance 500 - 1000 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	0.81	0.92	0.86
BS-seq	All cytosines with nearest neighbor within distance 1000 - 1500 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	0.70	0.89	0.78
BS-seq	All cytosines with nearest neighbor within distance 1500 - 2000 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	0.62	0.89	0.73
BS-seq	All cytosines with nearest neighbor within distance 2000 - 2500 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	0.59	0.88	0.71

Prediction metrics with intermediate methylation removed

				Datasets		Evaluation metrics		
Data type	Sampling loci constraints	Predictive Model	Input Features Used	Trained on	Tested on	Precision	Recall	Fscore
BS-seq	CGI cytosines	SVM	(SVM OFS)	H1 (depth >= 20, no intermediate methylation sites)	H1 (depth >= 20, no intermediate methylation sites)	0.97	0.97	0.97
BS-seq	non-CGI cytosines	SVM	(SVM OFS)	H1 (depth >= 20, no intermediate methylation sites)	H1 (depth >= 20, no intermediate methylation sites)	0.96	0.72	0.82
BS-seq	CGI cytosines	SVM	(SVM OFS)	NPC (depth >= 20, no intermediate methylation sites)	NPC (depth >= 20, no intermediate methylation sites)	0.97	0.97	0.97
BS-seq	non-CGI cytosines	SVM	(SVM OFS)	NPC (depth >= 20, no intermediate methylation sites)	NPC (depth >= 20, no intermediate methylation sites)	0.94	0.77	0.85

Table 3.9: Whole genome evaluations for DNA methylation prediction

Evaluation on genomic loci subsets

Comparison of SVM predictive model in NPC and H1 datasets

				Datasets		Evaluation metrics						
Data type	**Evaluation loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**TP**	**TN**	**FP**	**FN**	**TPR**	**TNR**	**Accuracy**
BS-seq	All cytosines	SVM	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 20)	43477143	3808315	1628764	220277	0.99	0.70	0.96
BS-seq	All cytosines	SVM	(SVM OFS)	H1 (depth >=20)	H1 (depth >=20)	43625145	3253354	1440859	2060474	0.95	0.69	0.93

Transfer learning between datasets using SVM predictive model

				Datasets		Evaluation metrics						
Data type	**Evaluation loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**TP**	**TN**	**FP**	**FN**	**TPR**	**TNR**	**Accuracy**
BS-seq	All cytosines	SVM	(SVM NPC OFS)	NPC (depth >= 20)	H1 (depth >=20)	41317737	3727930	966283	4367882	0.90	0.79	0.89
BS-seq	All cytosines	SVM	(SVM H1 OFS)	H1 (depth >=20)	NPC (depth >= 20)	42286042	2999939	2437140	1411378	0.97	0.55	0.92
BS-seq	All cytosines	SVM	(SVM NPC OFS)	NPC (depth >= 20)	MSC (depth >=20)	23703248	2779019	1132119	3528772	0.87	0.71	0.85
BS-seq	All cytosines	SVM	(SVM NPC OFS)	NPC (depth >= 20)	IMR90 (depth >=20, no sex chromosomes)	24457244	2292688	7853852	3990412	0.86	0.23	0.69

Comparison of different predictive models in NPC dataset

				Datasets		Evaluation metrics						
Data type	**Evaluation loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**TP**	**TN**	**FP**	**FN**	**TPR**	**TNR**	**Accuracy**
BS-seq	All cytosines	SVM	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 20)	43477143	3808315	1628764	220277	0.99	0.70	0.96
BS-seq	All cytosines	with SVM and	SVM OFS) + Consensus	NPC (depth >= 20)	NPC (depth >= 20)	43516018	4072279	1364800	181402	0.99	0.75	0.97
BS-seq	All cytosines	RF	(RF OFS)	NPC (depth >= 20)	NPC (depth >= 20)	43409038	3928194	1508885	288382	0.99	0.72	0.96

Comparisons of predictions involving the Consensus Reference Methylome

				Datasets		Evaluation metrics						
Data type	**Evaluation loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**TP**	**TN**	**FP**	**FN**	**TPR**	**TNR**	**Accuracy**
BS-seq	Cytosines with disagreement threshold = 0	SVM	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 5)	22892862	1724692	298403	48224	0.99	0.85	0.99
BS-seq	Cytosines with disagreement threshold = 0	Single predictor variable	Consensus Reference Methylome	NPC (depth >= 5)	NPC (depth >= 5)	22931737	1988656	34439	9349	0.99	0.98	0.99

Comparisons of predictions involving the Nearest Neighbor Methylation Status predictor

				Datasets		Evaluation metrics						
Data type	**Evaluation loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**TP**	**TN**	**FP**	**FN**	**TPR**	**TNR**	**Accuracy**
BS-seq	All cytosines with nearest neighbor distance within 2 - 20 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	29550709	14763379	564837	564881	0.98	0.96	0.98
BS-seq	All cytosines with nearest neighbor within distance 20 - 50 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	49875774	24255756	1059991	1063485	0.98	0.96	0.97
BS-seq	All cytosines with nearest neighbor within distance 50 - 100 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	74673162	38853553	1929159	1939160	0.97	0.95	0.97
BS-seq	All cytosines with nearest neighbor within distance 100 - 200 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	134860662	72343025	4301602	4328057	0.97	0.94	0.96
BS-seq	All cytosines with nearest neighbor within distance 200 - 500 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	343604957	178749445	18992938	19013419	0.95	0.90	0.93
BS-seq	All cytosines with nearest neighbor within distance 500 - 1000 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	535404825	184117716	49462992	49353293	0.92	0.79	0.88
BS-seq	All cytosines with nearest neighbor within distance 1000 - 1500 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	516462743	97126961	61567274	61614285	0.89	0.61	0.83
BS-seq	All cytosines with nearest neighbor within distance 1500 - 2000 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	506944740	57313501	65247592	64967145	0.89	0.47	0.81
BS-seq	All cytosines with nearest neighbor within distance 2000 - 2500 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	502844979	40255457	65773735	65603667	0.88	0.38	0.81

Comparisons of predictions involving the Nearest Neighbor Methylation Status predictor

				Datasets		Evaluation metrics						
Data type	Evaluation loci constraints	Predictive Model	Input Features Used	Trained on	Tested on	TP	TN	FP	FN	TPR	TNR	Accuracy
BS-seq	All cytosines	Nearest neighbor status (N1)	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	92474170	9903311	2108509	2075216	0.98	0.82	0.96
BS-seq	All cytosines	2nd Nearest neighbor status (N2)	2nd Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	92279738	9765000	2246820	2269648	0.98	0.81	0.96
BS-seq	All cytosines	3rd Nearest neighbor status (N3)	3rd Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	92127642	9647770	2364050	2421744	0.97	0.80	0.96
BS-seq	All cytosines	Vote among 3 Nearest neighbor status (V)	Vote among 3 Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	46819091	4897408	1108502	455602	0.99	0.82	0.97

Comparisons of predictions in H1 and NPC for different ranges of BS-seq CCRs

				Datasets		Evaluation metrics
Data type	Evaluation loci constraints	Predictive Model	Input Features Used	Trained on	Tested on	Accuracy
BS-seq	All cytosines	RF	(RF OFS)	NPC (depth >= 20, methylation CCR range [0,0.2)	NPC (depth >= 20, methylation CCR range [0,0.2)	0.81
BS-seq	All cytosines	RF	(RF OFS)	NPC (depth >= 20, methylation CCR range [0.2,0.4)	NPC (depth >= 20, methylation CCR range [0.2,0.4)	0.61
BS-seq	All cytosines	RF	(RF OFS)	NPC (depth >= 20, methylation CCR range [0.4,0.6)	NPC (depth >= 20, methylation CCR range [0.4,0.6)	0.54
BS-seq	All cytosines	RF	(RF OFS)	NPC (depth >= 20, methylation CCR range [0.6,0.8)	NPC (depth >= 20, methylation CCR range [0.6,0.8)	0.92
BS-seq	All cytosines	RF	(RF OFS)	NPC (depth >= 20, methylation CCR range [0.8,1.0]	NPC (depth >= 20, methylation CCR range [0.8,1.0]	0.99
BS-seq	All cytosines	SVM	(SVM OFS)	H1 (depth >= 20, methylation CCR range [0,0.2)	H1 (depth >= 20, methylation CCR range [0,0.2)	0.72
BS-seq	All cytosines	SVM	(SVM OFS)	H1 (depth >= 20, methylation CCR range [0.2,0.4)	H1 (depth >= 20, methylation CCR range [0.2,0.4)	0.69
BS-seq	All cytosines	SVM	(SVM OFS)	H1 (depth >= 20, methylation CCR range [0.4,0.6)	H1 (depth >= 20, methylation CCR range [0.4,0.6)	0.57
BS-seq	All cytosines	SVM	(SVM OFS)	H1 (depth >= 20, methylation CCR range [0.6,0.8)	H1 (depth >= 20, methylation CCR range [0.6,0.8)	0.96
BS-seq	All cytosines	SVM	(SVM OFS)	H1 (depth >= 20, methylation CCR range [0.8,1.0]	H1 (depth >= 20, methylation CCR range [0.8,1.0]	0.97

Performance of Direction for different BS-seq CCR values: We analyzed the results for whole methylome predictions in H1 and NPC by binning all high-coverage cytosines (sequencing depth ≥20) in the BS-seq datasets based on their CCRs. We created 5 bins, based on intervals of 0.2 from 0 (completely unmethylated) to 1 (completely methylated) based on the CCR. We find that for the SVM model (tested on the H1 dataset) and the RF model (tested on the NPC dataset), our accuracy for the extremal values of BS-seq CCRs are accurate (Fig. 3.13 (A, B) respectively, Table 3.9), while the performance is limited in the interval [0.4, 0.6) corresponding to intermediate methylation. This suggests that cytosines in these regions correspond to data points near the classification boundary, and are prone to be misclassified due to their proximity to the boundary.

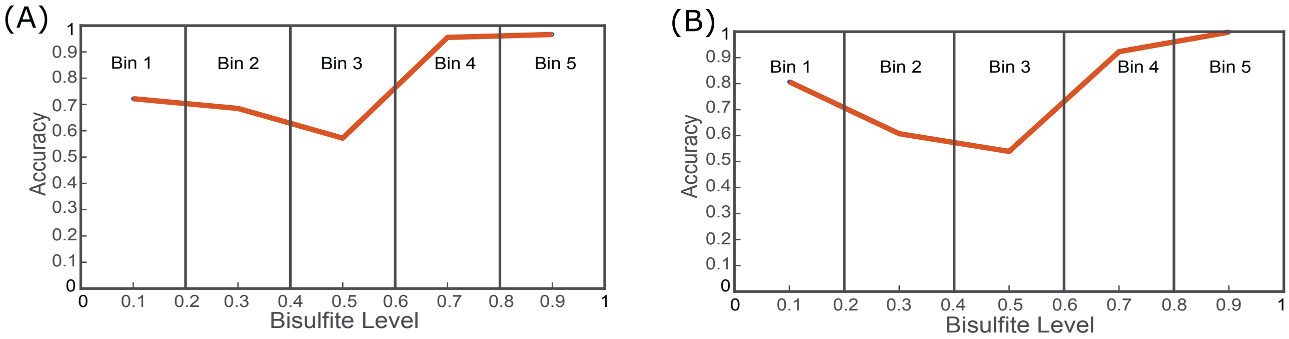


Fig. 3.13: Whole genome methylation status prediction accuracy obtained by binning the whole genome based on the BS-seq level in H1 (A), and NPC (B)

However, intermediate methylation is relatively uncommon in *in vitro* cell lines due to their homogeneity, and in mammalian systems (H1 and NPC: <3%, (102)). Thus, we find that the lower predictive ability of DIRECTION in cytosines with intermediate methylation only has a modest effect on the overall prediction metric by contrasting precision and recall in balanced sets sampled from the methylome by including or withholding cytosines with intermediate methylation (Fig. 3.14 (A, B)). It is important to note that such intermediate methylation is scarce in mammalian model systems (125), even in heterogeneous tissues like brain (101). For datasets with significantly higher amounts of intermediate methylation, we recommend using regression-based approaches (76).

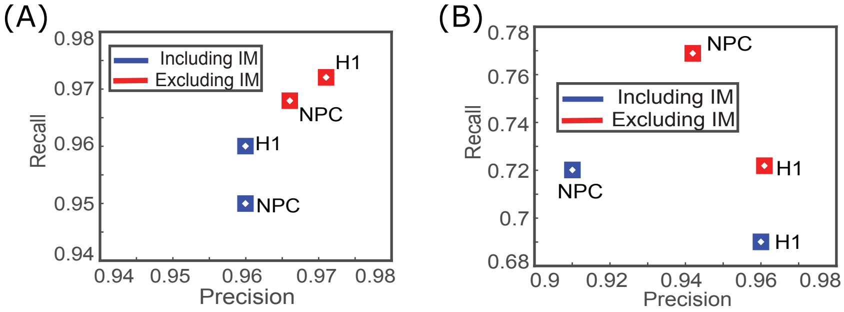


Fig. 3.14: Balanced sets predictions in H1 and NPC based on exclusion and inclusion of intermediate methylation [0.4,0.6) in CGI (A), and non-CGI (B).

3.5.2 Comparison with other DNA methylation prediction tools

Different methylation prediction algorithms work at differing genomic resolutions, on different datasets, using different predictor variables, to predict different response variables; making it challenging to set up unbiased comparisons between models. However, based on reported performances, DIRECTION is comparable to state-of-the-art high-resolution methylation prediction algorithms (Whole-genome accuracy: DIRECTION: 0.96 versus (76): 0.91, Table 1.1). Also, under the constraint of the same predictor variable set, DIRECTION outperformed the well-established inbuilt MATLAB classification tree function (Fig. 3.15).

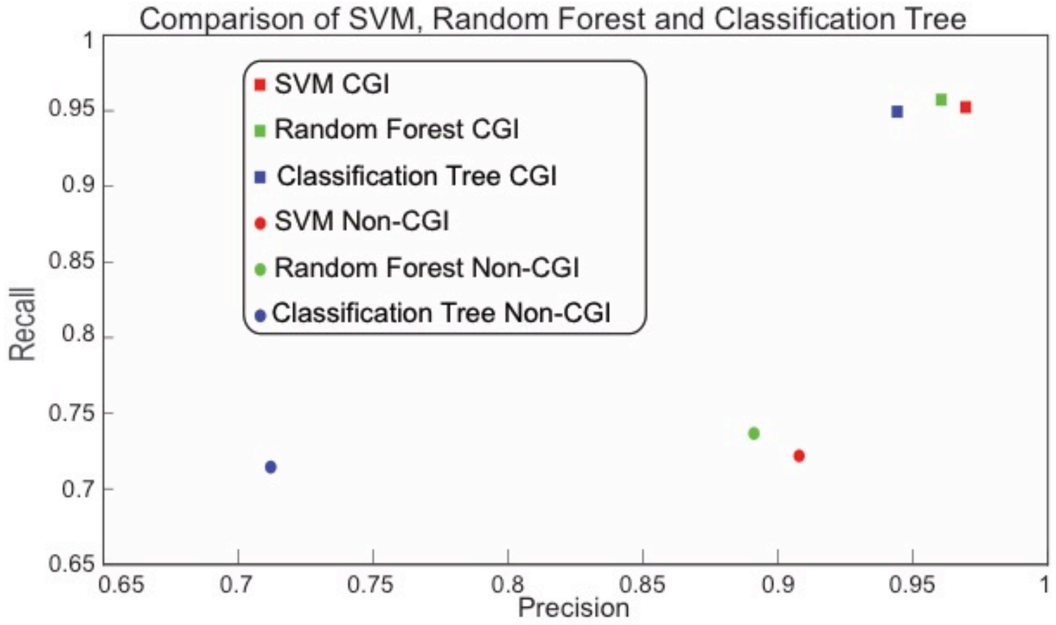
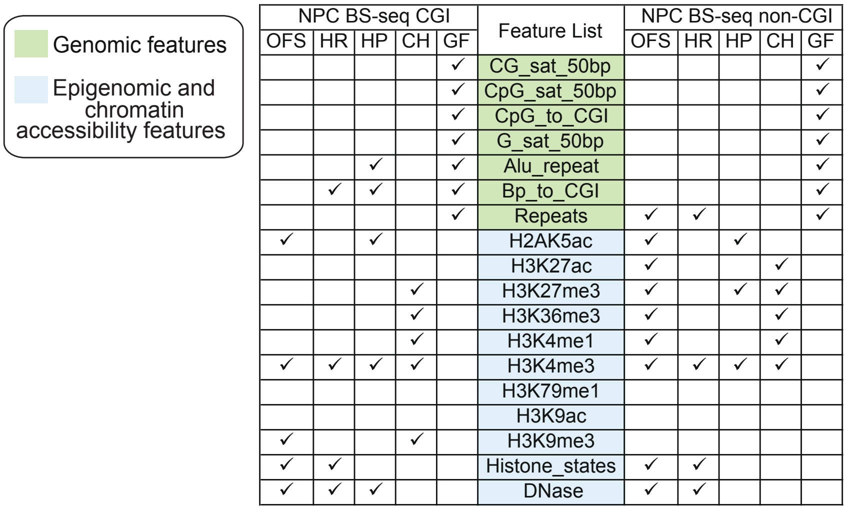


Fig. 3.15: Comparison of DIRECTION and classification tree for NPC methylation prediction.

3.5.3 OFS for DNA methylation prediction

The most discriminative features, contributing to high recall and precision, in DNA methylation predictions in NPC CGI regions were chromatin “states” inferred by the ChromHMM model (57) and H2AK5ac histone modification (Fig. 3.16 and Table 3.10). The underlying biological interpretation of our findings is supported by published literature as H2AK5ac histone modification was shown to be enriched in regions of euchromatin and low methylation (126). Also, the OFS for predicting DNA methylation in NPC CGI regions has only 5 features (Fig. 3.16), including transcription activation (H3K4me3, H2AK5ac) and repression (H3K9me3) associated histone marks, and DNase hypersensitivity which is known to be discriminative with respect to the underlying DNA methylation (127).



🟩 Genomic features

🟦 Epigenomic and chromatin accessibility features

NPC BS-seq CGI					Feature List	NPC BS-seq non-CGI				
OFS	HR	HP	CH	GF		OFS	HR	HP	CH	GF
				✓	CG_sat_50bp					✓
				✓	CpG_sat_50bp					✓
				✓	CpG_to_CGI					✓
				✓	G_sat_50bp					✓
		✓		✓	Alu_repeat					✓
	✓	✓		✓	Bp_to_CGI					✓
				✓	Repeats	✓	✓			✓
✓		✓			H2AK5ac	✓		✓		
					H3K27ac	✓			✓	
			✓		H3K27me3	✓		✓	✓	
			✓		H3K36me3	✓			✓	
			✓		H3K4me1	✓			✓	
✓	✓	✓	✓		H3K4me3	✓	✓	✓	✓	
					H3K79me1					
					H3K9ac					
✓			✓		H3K9me3					
✓	✓				Histone_states	✓	✓			
✓	✓	✓			DNase	✓	✓			

Fig. 3.16: Assorted feature sets generated by beam search by using various priority evaluation metrics (default F-score for OFS) for DNA methylation status prediction in NPC

Contrasting CGI to non-CGI OFSs, we find several histone features (H3K27ac, H3K27me3, H3K36me3, and H3K4me1) in the non-CGI, as opposed to H3K9me3 in the CGI. The non-CGI OFS also contains the Repeat feature, which is expected since repeat-containing retrotransposons in the human genome are silenced by methylation (128). The major changes in predictive ability are depicted by significantly different recall (Fig. 3.11) and AUC (Fig. 3.17 (A, B).

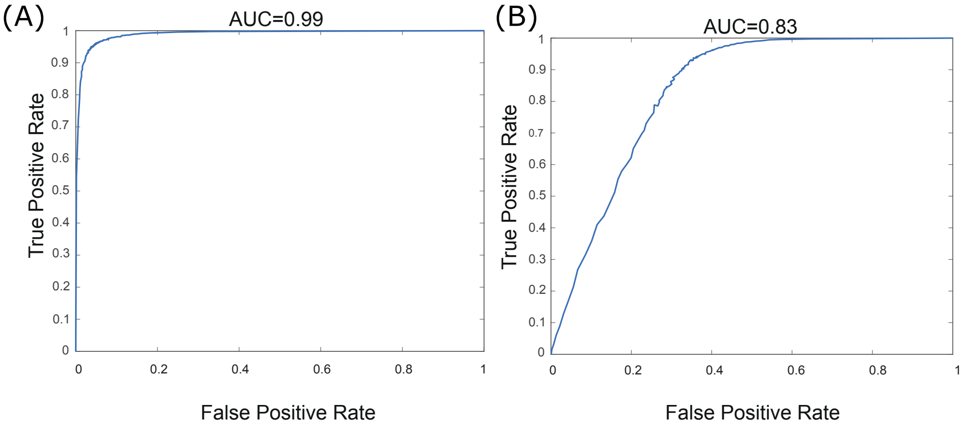


Fig. 3.17: AUC curves for methylation status prediction in NPC CGI (A) and non-CGI

We discovered that DNase and histone state features impacted the recall in CGI regions significantly, whereas high precision values were predominantly governed by H2AK5ac histone modification, known to be associated with regions of active chromatin and insulator region shores (120). Similarly, if any of the aforementioned features or the clusters they belong to are removed, the DNA methylation prediction in non-CGI regions drops (Fig. 3.18 (A, B)), suggesting similar informational content of predictors in CGI and non-CGI OFSs. In summary, a small set of features (H3K4me3; either DNase or Histone states; and H2AK5ac, along with Repeats for Non-CGI regions) can near optimally predict methylation status at single nucleotide resolution. Many aspects of our learned models are consistent with previous findings: a significant gain in prediction accuracy when highly discriminative epigenomic features are included (Fig. 3.17 and Table 3.10) (77), and significantly improved prediction performance in CGI regions with respect to non-CGI regions in both NPC and H1 cell line.

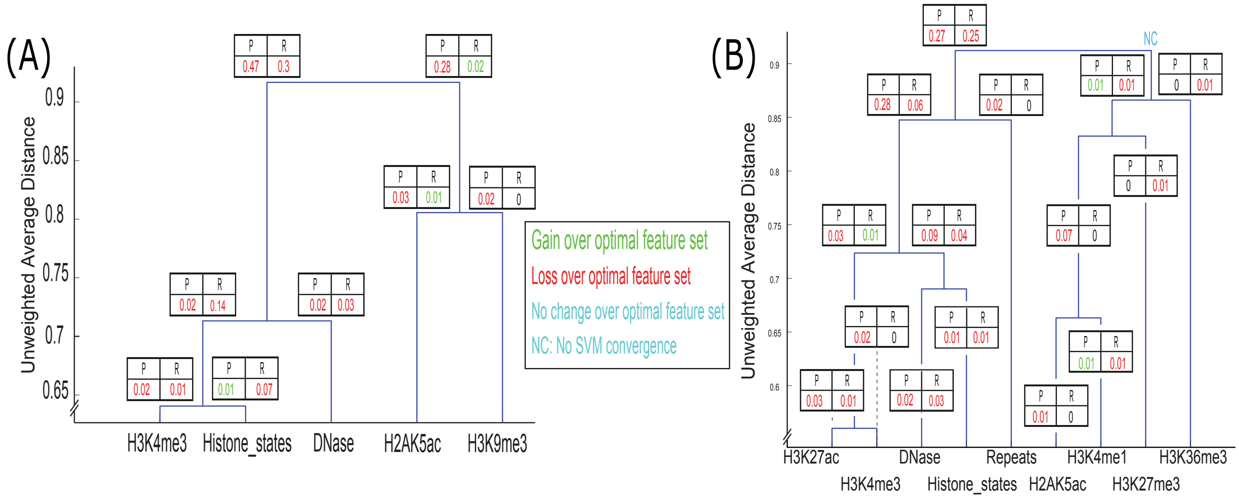
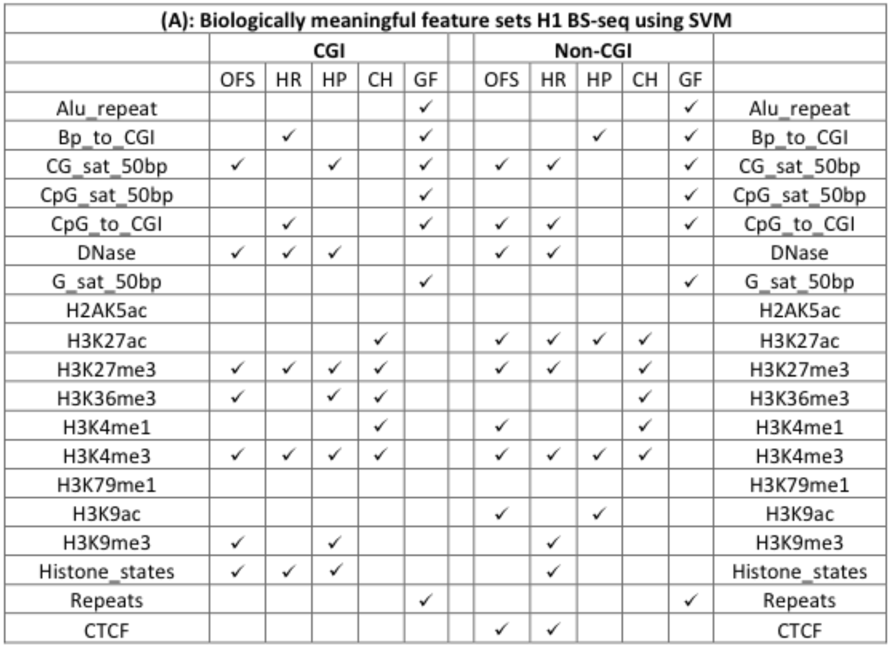
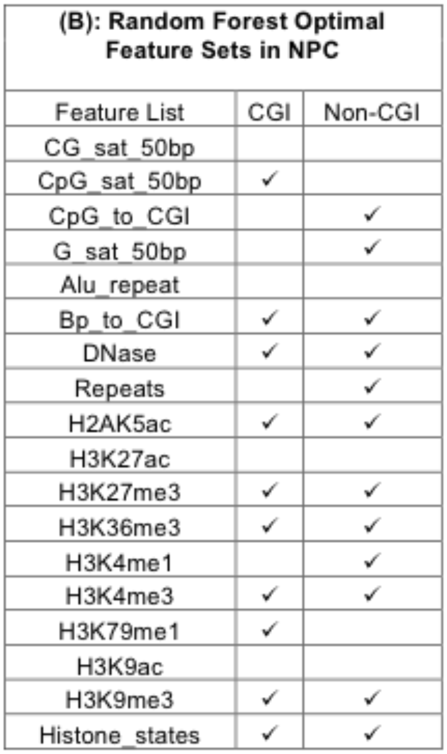


Fig. 3.18: Hierarchical clustering of features in OFS for predicting methylation status in NPC CGI (A) and non-CGI (B) regions, and corresponding changes in precision and recall with respect to OFS.

Table 3.10: Feature sets for methylation status prediction. (A) Feature sets for methylation status prediction using SVM in H1 CGI and non-CGI datasets. (B) OFS for methylation status prediction using RF in NPC CGI and non-CGI datasets.



(A): Biologically meaningful feature sets H1 BS-seq using SVM												
	CGI						**Non-CGI**					
	OFS	HR	HP	CH	GF		OFS	HR	HP	CH	GF	
Alu_repeat					✓						✓	Alu_repeat
Bp_to_CGI		✓			✓				✓		✓	Bp_to_CGI
CG_sat_50bp	✓		✓		✓		✓	✓			✓	CG_sat_50bp
CpG_sat_50bp					✓						✓	CpG_sat_50bp
CpG_to_CGI		✓			✓		✓	✓			✓	CpG_to_CGI
DNase	✓	✓	✓				✓	✓				DNase
G_sat_50bp					✓						✓	G_sat_50bp
H2AK5ac												H2AK5ac
H3K27ac				✓			✓	✓	✓	✓		H3K27ac
H3K27me3	✓	✓	✓	✓			✓	✓		✓		H3K27me3
H3K36me3	✓		✓	✓						✓		H3K36me3
H3K4me1				✓			✓			✓		H3K4me1
H3K4me3	✓	✓	✓	✓			✓	✓	✓	✓		H3K4me3
H3K79me1												H3K79me1
H3K9ac							✓		✓			H3K9ac
H3K9me3	✓		✓					✓				H3K9me3
Histone_states	✓	✓	✓					✓				Histone_states
Repeats					✓						✓	Repeats
CTCF							✓	✓				CTCF



(B): Random Forest Optimal Feature Sets in NPC		
Feature List	CGI	Non-CGI
CG_sat_50bp		
CpG_sat_50bp	✓	
CpG_to_CGI		✓
G_sat_50bp		✓
Alu_repeat		
Bp_to_CGI	✓	✓
DNase	✓	✓
Repeats		✓
H2AK5ac	✓	✓
H3K27ac		
H3K27me3	✓	✓
H3K36me3	✓	✓
H3K4me1		✓
H3K4me3	✓	✓
H3K79me1	✓	
H3K9ac		
H3K9me3	✓	✓
Histone_states	✓	✓

Characterization of feature subset contributions to predictive ability of the OFS: While creating the IFSs eliminated highly correlated features, OFSs identified by the beam search algorithm can still contain somewhat correlated, partially redundant features. For performance issues, we want to have some degree of redundancy in the OFS to make the prediction robust, but on the other hand we want to also assess the contribution to the predictive ability by subsets of features in the OFS. We thus performed the following assessment. We performed a standardization (Z-transformation (121)) across all features and hierarchically clustered them to identify similarity across features. Based on the feature clustering in the OFS, we left out individual features and feature subsets according to the nodes of the dendrogram, and retrained our classifier. The difference in performance metrics with respect to the OFS provides a clear indication of both feature redundancy and contributions of subsets of features to the OFS prediction metric.

While max-margin models do not explicitly posses a likelihood-based inferential framework to directly apply information theoretic approaches to sparse model selection like the Aikake Information Criterion (121), our approach provides an intuitive platform to identify smaller subsets of the OFS having comparable predictive power, and also identifies subsets of features that have major contributions to the precision and recall (Fig. 3.18).

The notion behind identifying a "minimal" feature set was based on the notion of several correlated input features potentially being part of the OFS, each only contributing a limited amount of predictive power to the overall OFS. By clustering the individual features in the OFS and eliminating them one at a time, we identified the effect each (or a subset) possesses on the predictive power, in a manner agnostic to the classification algorithm. The tradeoff between obtaining a smaller feature set versus improving classification performance metrics can thus be clearly identified, allowing the user to decide on a choice of the input feature set for related experiments.

3.5.4 Using neighboring CpG sites as predictor variables

For improving imputation, the methylation status of the nearest neighboring CpG site within 500bp was used to create an input feature. Our feature engineering analyses (See Methods section, Fig. 3.8) suggests that the predictive quality of the feature significantly decreases after 500bp (a distance corresponding to the average size of CGIs (129)), in agreement with findings that CGIs are typically consistently methylated or demethylated. We tested the ability of this feature to contribute to predictions in CGI and non-CGI SVM models by adding it to the beam search-identified OFS, followed by retraining the SVMs on balanced sets. It makes insignificant impact on the CGI SVM (where precision and recall are > 0.95) but strikingly improves recall of the non-CGI SVM from 0.72 to 0.77 (Fig 3.11), suggesting that even in non CpG-rich regions, spatial contiguity of methylation status is commonplace.

3.5.5 Summary and Discussion

Here we introduced DIRECTION, a state-of-the-art machine learning toolkit that performs DNA methylation and hydroxymethylation predictions at single nucleotide resolution in mammalian genomes. DIRECTION implementation is predicated upon 2 learning algorithms: SVM and RF, which characteristics are detailed in the methods section of this chapter. We provided an extensive discussion why classification based predictive models are better suited than regression based models to perform such predictions. We introduced all input variables we used to perform such predictions, as well as the beam search: a stochastic feature selection procedure that identifies most important predictor variables. We created an additional predictor variable based on the methylation status of neighboring CpG sites, and demonstrated how its addition to the input feature space affects prediction rate. Also, we utilized the *consensus reference methylome* to predict DNA methylation status of invariant CpGs (see chapter 2). Based on the *consensus reference methylome* we created an additional feature that splits the methylome into its invariant and variant portion to improve predictions on balanced sets in NPC. Finally, in order to show that obtained feature sets are somewhat conserved across different cell lines, we performed transfer learning between H1, NPC and mesenchymal stem cells (MSC) (See chapter 5 Transfer learning between H1 and NPC cell lines).

Both BS-seq (36) and TAB-seq (130) protocols have reduced representation versions where assays query a limited set of CpGs. DIRECTION is ideally suited to impute methylation or hydroxymethylation status in such reduced representation datasets (as well as existing low coverage whole genome datasets), being able to make use of relevant genome-wide traits (based on genomic annotation, DNA sequence and relevant publicly available genome-wide assays) to create whole-genome scale datasets.

Widespread use of epigenome-querying assays like BS-seq naturally leads to a discussion of relevance of *in silico* epigenome prediction. However, for an *in vivo* sourced sample with a limited DNA yield (like clinical samples), only a few assays can be performed, necessitating the *in silico* prediction of some assays based on the outcome of others. Secondly, paralleling the rise of whole genome assays, are reduced representation BS-seq (36) and TAB-seq (130) assays, for which *in silico* prediction is especially relevant. Recent developments in single cell technologies allow BS-seq assays to be performed on individual cells (131) with some studies contemplating single-cell TAB-seq as future work (132). Given the destructive nature of next generation sequencing, *in silico* prediction tools can be potentially useful for using single-cell methylation data and underlying genomic sequence for imputing methylation status or to make a model-based prediction for 5-hmC status (for 5-hmC status prediction see Chapter 4).

CHAPTER 4

DNA HYDROXYMETHYLATION PREDICTION

Authors: Milos Pavlovic, Pradipta Ray, Kristina Pavlovic, Aaron Kotamarti,
Min Chen and Michael Q. Zhang
Department of Biological Sciences
The University of Texas at Dallas
800 W. Campbell Road
Richardson, TX 75080-3021

4.1 Prior Publication

Milos Pavlovic (M.P.) performed the majority of experiments, and Pradipta Ray (P.R.) designed the majority of experiments. M.P. and P.R. wrote the manuscript. Min Chen (M.C.) advised M.P. and Michael Zhang (M.Q.Z.) supervised the project. This chapter covers various aspects of 5-hmC status prediction, such as tractability of 5-hmC modifications in mammalian genomes, whole-genome 5-hmC status imputation, whole-genome and balanced set model-based prediction results in NPC and H1 cell lines, as well as the OFS selection using DIRECTION. Per the policy of OUP Bioinformatics, the publication of material in a PhD thesis is permitted with the publication of a peer-reviewed manuscript in their journal. The original manuscript (1) "DIRECTION: A machine learning framework for predicting and charactering DNA methylation and hydroxymethylation in mammalian genomes" by Milos Pavlovic, Pradipta Ray, Kristina Pavlovic, Aaron Kotamarti, Min Chen and Michael Q. Zhang, published in 2017, is reproduced by permission of Oxford University Press and appears online at the following web address: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btx316. Supplementary information is available online at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btx316. The main text is partially altered compared to the online version of the manuscript, and figures and tables do not chronologically correspond to the online manuscript numbering.

4.2 Abstract

Here we performed a pioneering work of predicting 5-hmC modification status *in silico* using DIRECTION. We obtained high whole-genome accuracy, identified most important predictor

variables, and paved the way for large-scale reconstruction of hydroxymethylation maps in mammalian model systems. We found that the OFS harboring accurate 5-hmC predictions is comprised of enhancer-like features, most notably H3K27ac and H3K4me1 histone modifications, and therefore 5-hmC in enhancer regions was considered as a separate prediction paradigm. Based on such accurate predictions we identified enhancer regions, which exhibit differential hydroxymethylation and potentially serve as cis-regulatory regions of proximal protein coding genes. We performed whole hydroxymethylome reconstruction in small TAB-seq datasets and built an *in silico* platform for high-throughput hypothesis testing based on such predictions. Finally, we predicted 5-hmC status in the regions containing high BS-seq coverage, to show that 5-hmC status can be accurately *de novo* predicted based on BS-seq data and a few other available features.

4.3 Introduction

BS-seq is not able to differentiate between 5-mC and 5-hmC modifications, and therefore the overall degree of methylation represents the summation of the two. Despite the availability of high-throughput assays for querying DNA hydroxymethylation, there only exists a handful of publicly available TAB-seq or oxBS-seq datasets, and performing whole-genome BS-seq, oxBS-seq or TAB-seq requires significant expenditure and skilled labor. These arguments lend weight to predicting genome-wide 5-hmC modification *in silico.* Additionally, CpG dinucleotides may be asymmetrically modified for 5-hmC (114), therefore we used DIRECTION to solely predict 5-hmC status at single nucleotide resolution, as opposed to CpG dinucleotide.

Since 5-hmC is known to closely associate with enhancers and thereby affect gene expression of proximal genes (116), we considered 5-hmC status predictions in enhancer regions of the genome as a separate prediction paradigm (see Results).

4.4 Methods

5-hmC is an intermediate molecular state in the demethylation pathway, and TAB-seq CCRs tend to be significantly lower than BS-seq CCRs (see Chapter 3 Methods for BS-seq CCR). For analyzing 5-hmC levels, a naive analysis yields a distribution with maximal frequency at the CCR of 0, and no other observable secondary modes in the distribution (Fig. 4.1). However, it is well characterized that while most CpG sites have a CCR of 0, statistically significant hydroxymethylated CpG sites have a CCR frequency distribution with a mode of 0.18 (Fig. 4.2) (114).

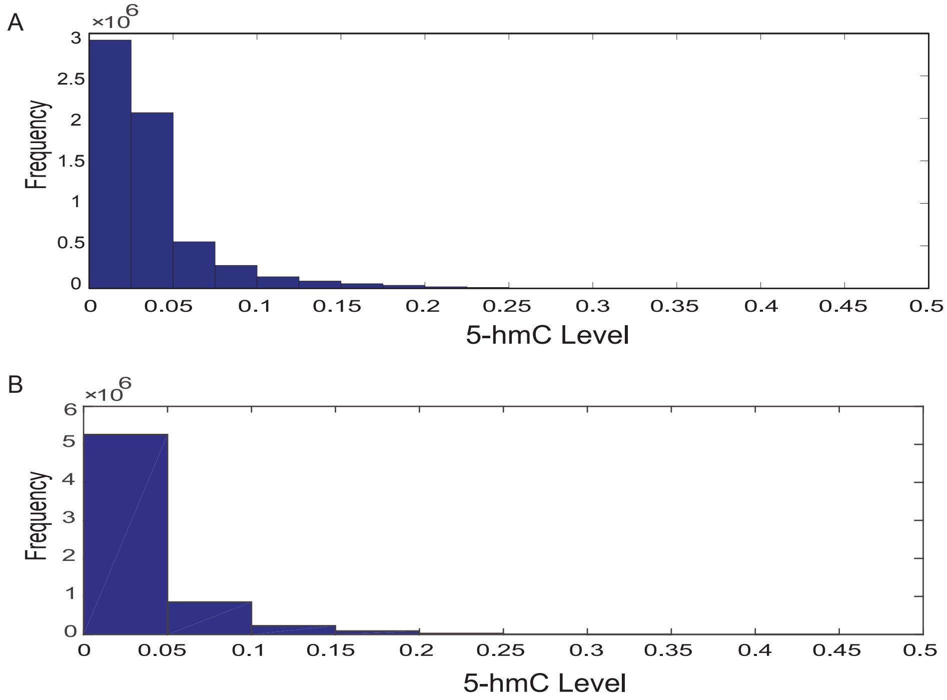


Fig. 4.1: Empirical distributions of 5-hmC levels in NPC (A) and H1 (B) cell lines.

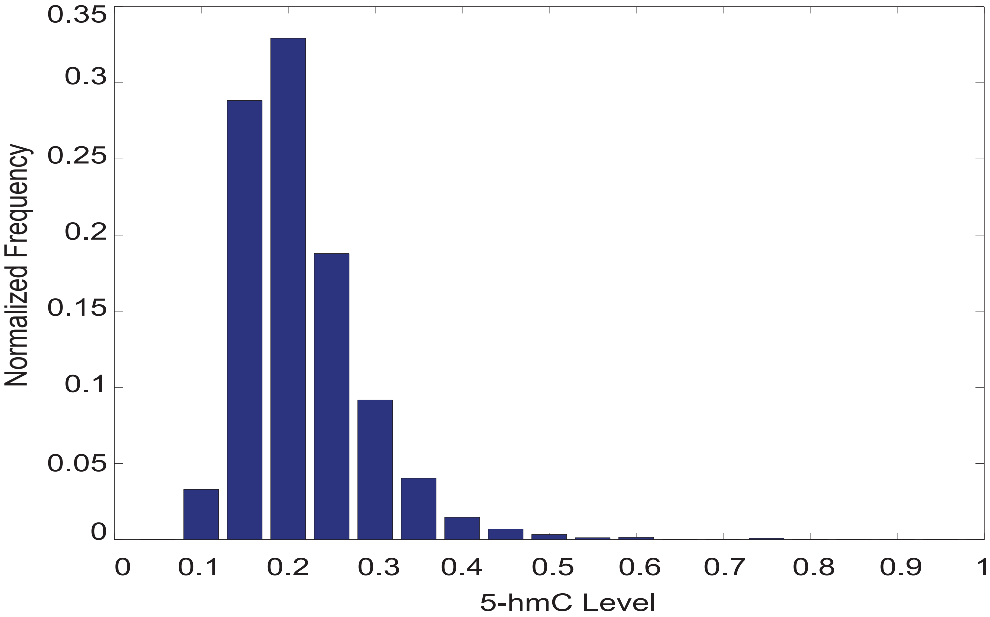


Fig. 4.2: Distribution of 5-hmC levels in the set of CpG sites identified as significantly hydroxymethylated in Yu et al. 2012 (114)

Hence, given the unimodal distribution of CCRs generated by TAB-seq, in order to label individual cytosines into significantly hydroxymethylated versus non-hydroxymethylated classes, we choose a threshold of 0.09, equidistant from 0 and 0.18. Accordingly, as with 5-mC we also model 5-hmC prediction as a binary classification problem.

4.4.1 Tractability of 5-hmC modification and feasibility of 5-hmC status prediction

Strong preference of 5-hmC for open chromatin regions, as well as its positive correlation with gene expression and bias towards exon inclusion were previously documented in literature (133, 134), suggesting a functional role and consistency of 5-hmC modifications across biological replicates. 5-hmC modifications have also been shown to be temporally stable (21) further suggesting a strong signal to noise ratio in hydroxymethylation assays.

However, since there is no *in silico* precedent performing 5-hmC predictions we first needed to verify that 5-hmC is sufficiently tractable in mammalian genomes for the purpose of performing such predictions. Therefore, at the outset, we performed pairwise comparison of hydroxymethylation levels across biological replicates in NPC, using binary discretization of hydroxymethylation levels. TAB-seq CCR correlation across biological replicates is less faithful than BS-seq by exhibiting some stochasticity in the signal. However, we obtain a concordance rate (fraction of cytosines where 5-hmC status between replicates agree) of 82% (in CpG sites with coverage >60) for 5-hmC status between biological replicates in NPC (Fig. 4.3).

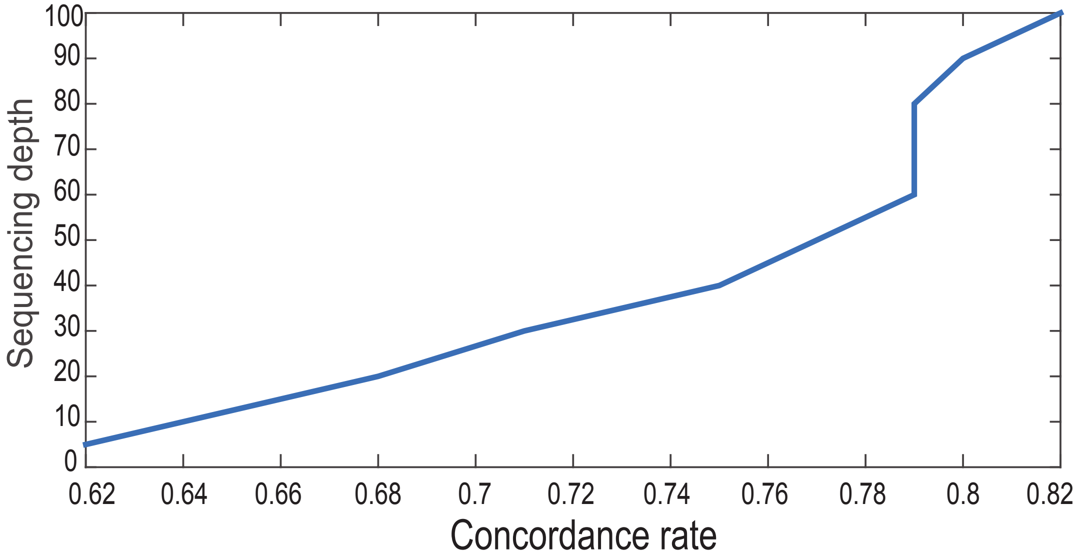


Fig. 4.3: Concordance rate between TAB-seq NPC replicates as a function of minimum sequencing depth of mapping at either replicate

For practical purposes, this may be considered as an approximate upper bound of possible predictive accuracy when evaluating 5-hmC status predictions. Thus, we approach hydroxymethylation prediction as a binary classification problem for the purpose of reconstructing a discretized approximation of TAB-seq CCRs at individual cytosines.

Further we looked at consistency of our BS-seq and TAB-seq datasets in NPC. MLML (31) is a method that uses read counts from data obtained by TAB-seq (or oxBS- seq), and BS-seq to estimate CCRs for the 5-mC and 5-hmC modifications jointly. It identifies indices exhibiting "overshoot" where the sum of estimated CCRs for 5-mC and 5-hmC sum to greater than 1. Upon running MLML on our BS-seq and TAB-seq datasets in NPC we obtained the maximum likelihood distribution of 5-mC levels (Fig. 4.4) that strongly resembled the one of BS-seq levels (Fig. 3.1). Additionally, out of the 52,531,101 CpG sites being analyzed (sites without coverage in either of the experiments are discarded) the number of overshoot indices was only 3,186 or 0.006% in NPC, suggesting that our BS-seq and TAB-seq datasets show good consistency between experiments. Most of the overshoot indices contained very low coverage (2,654 CpG sites) in both BS-seq and TAB- seq experiments and were systematically discarded prior to training our model. These evidences lend weight to the tractability of predicting 5-hmC modifications.

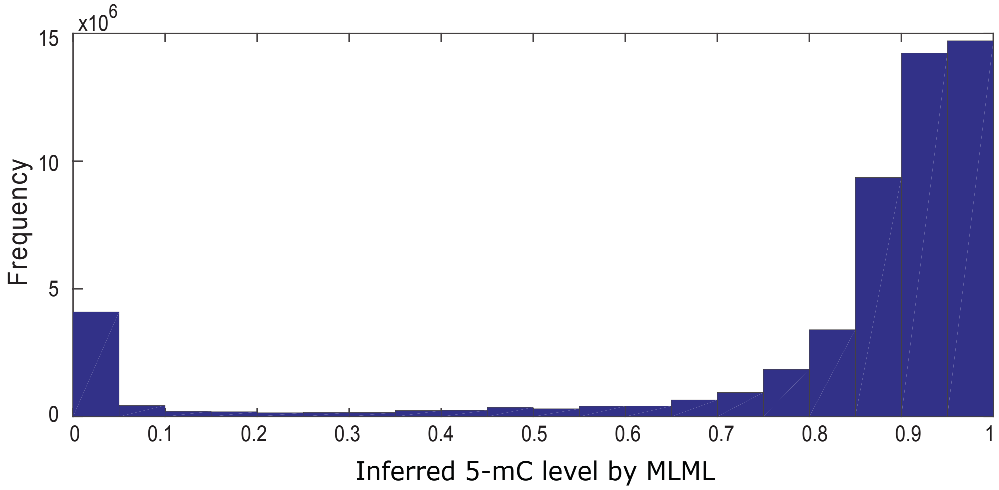


Fig. 4.4: Inferred 5-mC level distribution in NPC by the tool MLML (31) by jointly analyzing BS-seq and TAB- seq CCRs in NPC.

4.4.2 Feature engineering and feature selection

The same set of features that was used to predict DNA methylation status in H1 and NPC cell lines (listed in Table 3.3) was employed to predict 5-hmC status, including the following exceptions: a) Bisulfite level feature was added for the purpose of training and testing the predictive model. b) CpG island feature was ignored (5-hmC does not exhibit preference between CpG and non-CpG island regions). c) Enhancer feature (binary feature) was added to the model, as 5-hmC is known to closely associate with enhancers (116). The feature selection procedure to identify OFS for 5-hmC status predictions was conducted in the same fashion as with DNA methylation status predictions (see Methods Chapter 3).

Due to a scarcity of publicly available TAB-seq datasets we were not in position to collect a sufficiently large set of reference hydroxymethylomes. Conversely, we obtained a plethora of publicly available BS–seq datasets from Roadmap Epigenome Consortium (98), which we used to create the consensus reference methylome (for *consensus reference methylome* see Chapter 2). Therefore, this dictionary lookup based feature was not optionally used for 5-hmC status prediction.

Analogous to using neighboring CpG sites methylation status as predictor variables (see Chapter 3 Methods) neighboring 5-hmC status information was used to predict 5-hmC status of a CpG site in question. The default hydroxymethylation level value of 0.09 was used as threshold for separating classes of highly and lowly hydroxymethylated CpGs (for details see Results).

4.5 Results

Tab-seq datasets for H1 and NPC cell lines were obtained under the following accession numbers: H1: (GEO GSE36173) and NPC (GEO GSM882245, GSM1463129), while the BISMARK (43) was used for mapping and obtaining the CCRs. Scripts that were used to calculate the reads sequencing depth and hydroxymethylation levels were coverage2cytosine and bismark methylation extractor. The final output to the .bed format was performed by the bismark2bedGraph. This was performed to generate H1 and NPC TAB-seq CCRs. These cell types were chosen due to availability of BS-seq and TAB-seq data, and since previous studies performing functional enrichment and analysis of 5-hmC in human and mouse ESCs (17, 114, 116, 117) and neural progenitors (118-120) especially in neural development.

4.5.1 5-hmC status prediction

We performed 5-hmC status prediction using features from the initial feature set for methylation status prediction model, using methylation level as an additional feature (Table 4.3). In order to identify the most discriminative features for 5-hmC status prediction, we ran our beam search algorithm and obtained discriminative feature sets.

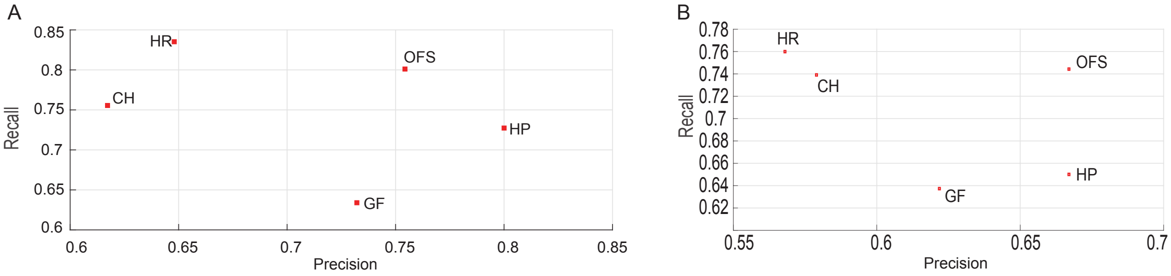


Fig. 4.5: 5-hmC status prediction on balanced sets using SVM in NPC (A) and H1 (B) (Table 4.3 for results).

Table 4.1: Feature sets for 5-hmC status prediction in NPC, NPC enhancers, and H1 using SVMmodel (A) and in NPC dataset using RF model (B).

(A) Biologically meaningful feature sets for 5-hmC status prediction on balanced sets using SVM

	NPC enhancers					**NPC**					**H1**				
	OFS	HR	HP	CH	GF	OFS	HR	HP	CH	GF	OFS	HR	HP	CH	GF
Alu_repeat	✓	✓	✓		✓	✓				✓	✓				✓
BS-seq_CCR	✓	✓				✓	✓				✓	✓	✓		
Bp_to_CGI					✓					✓					✓
CG_sat_50bp	✓				✓	✓	✓	✓		✓	✓		✓		✓
CpG_sat_50bp					✓			✓		✓					✓
CpG_to_CGI					✓			✓		✓					✓
DNase	✓	✓				✓					✓				
G_sat_50bp					✓			✓		✓	✓				✓
H2AK5ac															
H3K27ac	✓								✓		✓			✓	
H3K27me3				✓					✓					✓	
H3K36me3				✓					✓			✓	✓	✓	
H3K4me1	✓			✓		✓	✓		✓		✓	✓		✓	
H3K4me3				✓					✓					✓	
H3K79me1															
H3K9ac															
H3K9me3	✓			✓		✓					✓				
Histone_states		✓						✓			✓				
Repeats		✓			✓		✓	✓		✓	✓				✓
CpG_Island								✓							
CTCF											✓	✓	✓		

(B) Random Forest OFS for NPC 5-hmC status prediction

	Features
Alu_repeat	
Bp_to_CGI	
CG_sat_50bp	✓
CpG_sat_50bp	✓
CpG_to_CGI	✓
DNase	✓
G_sat_50bp	
H2AK5ac	
H3K27ac	✓
H3K27me3	✓
H3K36me3	
H3K4me1	✓
H3K4me3	✓
H3K79me1	✓
H3K9ac	
H3K9me3	✓
Histone_states	✓
Repeats	
BS-seq_CCR	✓
CpG_Island	

Based on the experimental design previously outlined, the performance of the OFS was compared against other biologically and statistically meaningful feature sets (NPC: Fig. 4.5 (A), F- score 0.78; H1: Fig. 4.5 (B), F-score 0.7). The most distinguishing characteristic of assorted 5-hmC feature sets in both cell types was the profound presence of active enhancer histone modifications H3K4me1 and H3K27ac (135) DNase and other genomic derived features including CpG content, and Alu repeats (Table 4.1).

Insightfully, a single addition to the OFS when our predictor was constrained to the enhancer regions was H3K27ac, suggesting biological interpretability of our results. The absence of H3K27ac from the 5-hmC OFS (when the predictor is not constrained to enhancer regions) can be explained by the presence of another enhancer chromatin mark (H3K4me1) in the OFS, and the relatively small size of enhancer regions compared to the non-enhancer portion of the genome.

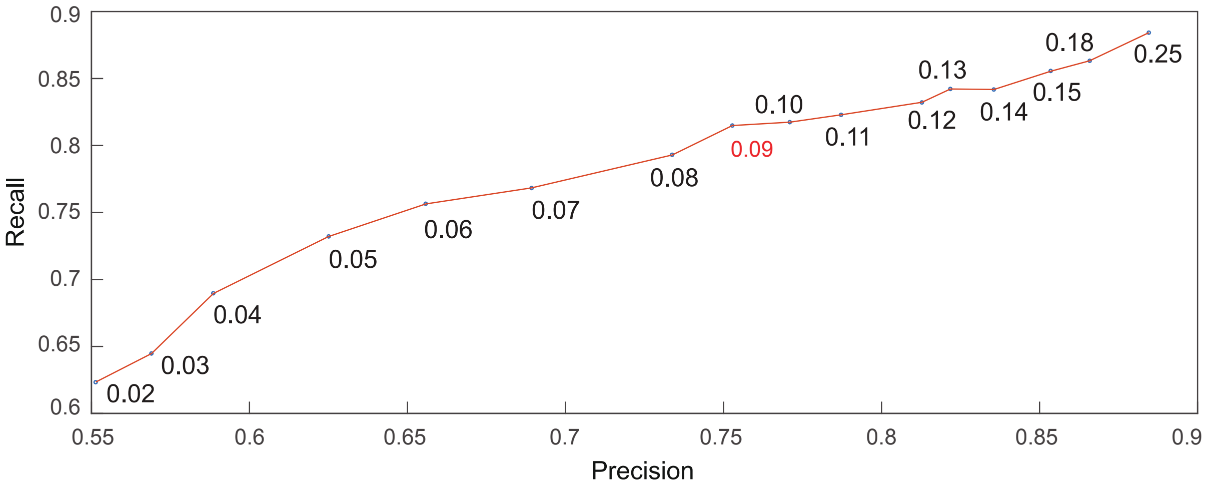


Fig. 4.6: Precision/Recall plot for 5-hmC status predictions in NPC using various 5-hmC level thresholds for SVMs. Threshold of 0.09 is marked red to symbolize the default value that was used in this paper.

Unsurprisingly, we find the H3K4me1 enhancer mark being one of the most promising predictive features due to its presence in both the high recall and optimal feature sets. Significant depletion of 5-hmC in H3K9me3 rich heterochromatin regions, and its positive correlation with H3K4me3 active histone modification (136), clearly designates these chromatin marks as suitable candidates for the OFS.

In order to show that the obtained OFS is discriminative towards 5-hmC signal, we predicted 5-hmC status across various TAB-seq level thresholds and noticed that the prediction metric grows slowly with the increase in threshold value (Fig. 4.6), and shows consistent AUC for a range of thresholds (Fig. 4.7 (A, B)).

We performed whole-genome 5-hmC predictions in NPC and H1 and obtained 0.82 and 0.75 accuracy respectively (NPC: Table 4.2 (A); H1: Table 4.2 (B)). These results together suggest that 5-hmC status can be fairly accurately reconstructed in our datasets. Lower prediction accuracy in H1 can putatively be attributed to a lower coverage depth in the training data.

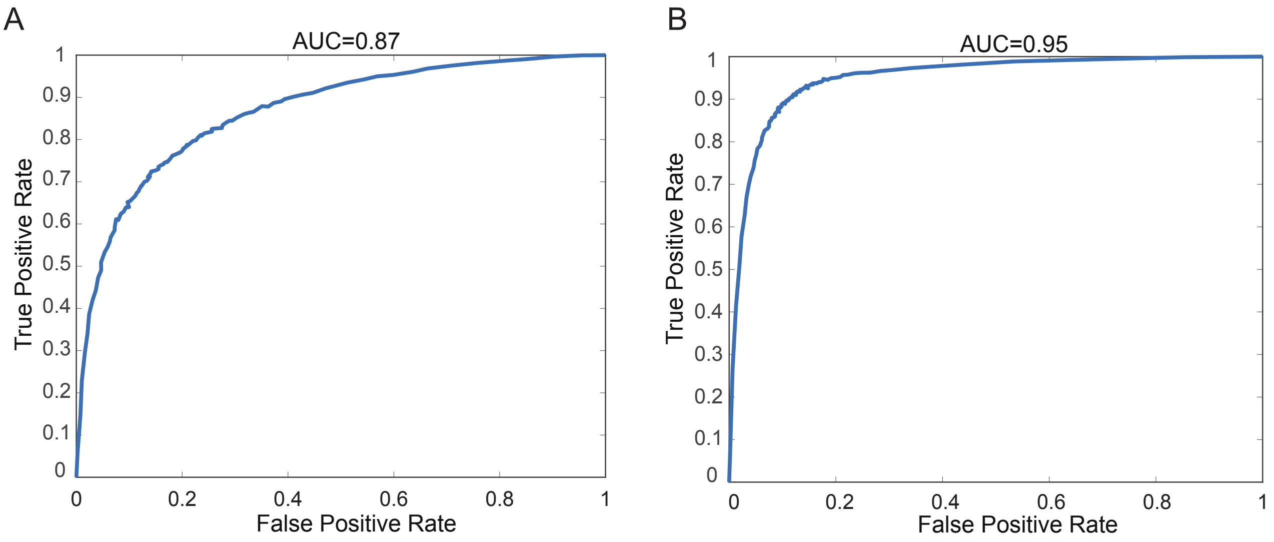


Fig. 4.7: ROC curves for 5-hmC status predictions in NPC RF model. (A) ROC curve for 5-hmC status predictions in NPC RF model. (B) ROC curve for 5-hmC status predictions using threshold of 0.25 in NPC RF model.

Table 4.2: Whole genome 5-hmC status prediction evaluation metrics in NPC (A) and H1 (B).

A

	Enhancer	**Non-enhancer**	**Total**
True Positive Rate	0.85	0.73	0.75
True Negative Rate	0.72	0.83	0.82
Accuracy	0.74	0.83	0.82

B

Prediction Metric	H1 5-hmC Whole Genome
True Positive Rate	0.67
True Negative Rate	0.75
Accuracy	0.75

We performed 5-hmC predictions restricted to cytosines with high BS-seq CCRs, yielding comparable results to our previous analyses, implying that the numerous public BS-seq datasets together with additional input features can be used to predict 5-hmC maps (see *BS-seq driven 5-hmC status identification)*. For 5-hmC transfer learning in H1 and NPC see Chapter 5.

BS-seq driven 5-hmC status identification: There is a vast number of BS-seq datasets which are publicly accessible, and only a handful of these have an accompanying TAB-seq counterpart. We used the NPC BS-seq and TAB-seq datasets to train and test a 5-hmC status prediction classifier using only CpG sites where BS-seq CCR could be reliably estimated (coverage ≥ 20). We trained our model using the 5-hmC OFS, where the BS-seq level feature was excluded. Such a classifier performs comparably to our previously reported classifiers, achieving a precision of 0.74, recall of 0.8 and an F-score of 0.77 (Table 4.3). Hence, we show that our method has the capability of performing *de novo* 5-hmC modifications map reconstruction based on the BS-seq dataset and a handful of other features. Such an approach trades off the size and diversity of the training data for a smaller, higher quality training set, and can likely be useful in reconstructing 5-hmC maps of experimental conditions with published BS-seq data.

Table 4.3: Balanced set and whole genome evaluations for 5-hmC status predictions.

Evaluation on genomic loci subsets by sampling balanced sets

Comparison of predictive abilities for different feature sets in NPC dataset

				Datasets		Evaluation metrics		
Data type	**Sampling loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**Precision**	**Recall**	**Fscore**
TAB-seq	All cytosines	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	0.75	0.80	0.77
TAB-seq	All cytosines	SVM	GF	NPC (depth >= 60)	NPC (depth >= 60)	0.73	0.63	0.68
TAB-seq	All cytosines	SVM	CH	NPC (depth >= 60)	NPC (depth >= 60)	0.62	0.76	0.68
TAB-seq	All cytosines	SVM	HP	NPC (depth >= 60)	NPC (depth >= 60)	0.80	0.73	0.76
TAB-seq	All cytosines	SVM	HR	NPC (depth >= 60)	NPC (depth >= 60)	0.65	0.84	0.73

Comparison of predictive abilities for different feature sets in H1 dataset

				Datasets		Evaluation metrics		
Data type	**Sampling loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**Precision**	**Recall**	**Fscore**
TAB-seq	All cytosines	SVM	(SVM OFS)	H1 (depth >= 60)	H1 (depth >= 60)	0.67	0.74	0.70
TAB-seq	All cytosines	SVM	GF	H1 (depth >= 60)	H1 (depth >= 60)	0.62	0.64	0.63
TAB-seq	All cytosines	SVM	CH	H1 (depth >= 60)	H1 (depth >= 60)	0.58	0.74	0.65
TAB-seq	All cytosines	SVM	HP	H1 (depth >= 60)	H1 (depth >= 60)	0.67	0.65	0.66
TAB-seq	All cytosines	SVM	HR	H1 (depth >= 60)	H1 (depth >= 60)	0.57	0.76	0.65

Comparison of predictive abilities for different feature sets in NPC enhancer dataset

				Datasets		Evaluation metrics		
Data type	**Sampling loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**Precision**	**Recall**	**Fscore**
TAB-seq	Enhancer cytosines	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	0.77	0.82	0.79
TAB-seq	Enhancer cytosines	SVM	GF	NPC (depth >= 60)	NPC (depth >= 60)	0.73	0.63	0.68
TAB-seq	Enhancer cytosines	SVM	CH	NPC (depth >= 60)	NPC (depth >= 60)	0.63	0.75	0.68
TAB-seq	Enhancer cytosines	SVM	HP	NPC (depth >= 60)	NPC (depth >= 60)	0.93	0.53	0.68
TAB-seq	Enhancer cytosines	SVM	HR	NPC (depth >= 60)	NPC (depth >= 60)	0.63	0.82	0.71

Comparison of different predictive models in NPC dataset

				Datasets		Evaluation metrics		
Data type	**Sampling loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**Precision**	**Recall**	**Fscore**
TAB-seq	All cytosines	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	0.75	0.80	0.77
TAB-seq	All cytosines	RF	(RF OFS)	NPC (depth >= 60)	NPC (depth >= 60)	0.78	0.82	0.80

Comparison of different predictive models in NPC dataset

				Datasets		Evaluation metrics		
Data type	**Sampling loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**Precision**	**Recall**	**Fscore**
TAB-seq	Cytosines with high BS-seq levels	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	0.74	0.80	0.77
TAB-seq	All cytosines	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	0.75	0.80	0.77

Evaluation on genomic loci subsets

Comparison of SVM models in NPC and H1 datasets

				Datasets		Evaluation metrics						
Data type	**Evaluation loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**TP**	**TN**	**FP**	**FN**	**TPR**	**TNR**	**Accuracy**
TAB-seq	All cytosines	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	339376	7269832	1578026	114373	0.75	0.82	0.82
TAB-seq	All cytosines	SVM	(SVM OFS)	H1 (depth >= 60)	H1 (depth >= 60)	84682	1664105	552163	40920	0.67	0.75	0.75

Comparison of SVM models in NPC enhancer regions, non-enhancer regions, & whole genome

				Datasets		Evaluation metrics						
Data type	**Evaluation loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**TP**	**TN**	**FP**	**FN**	**TPR**	**TNR**	**Accuracy**
TAB-seq	Enhancer cytosines	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	67008	311652	123148	12030	0.85	0.72	0.74
TAB-seq	Non-enhancer cytosines	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	272368	6958180	1454878	102343	0.73	0.83	0.82
TAB-seq	All cytosines	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	339376	7269832	1578026	114373	0.75	0.82	0.82

Comparison of different predictive models in NPC dataset

				Datasets		Evaluation metrics						
Data type	**Evaluation loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**TP**	**TN**	**FP**	**FN**	**TPR**	**TNR**	**Accuracy**
TAB-seq	All cytosines	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	339376	7269832	1578026	114373	0.75	0.82	0.82
TAB-seq	All cytosines	RF	(RF OFS)	NPC (depth >= 60)	NPC (depth >= 60)	348195	7260210	1587648	105554	0.77	0.82	0.82

Transfer learning between H1 and NPC datasets using SVM predictive model

Data type	**Evaluation loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**TP**	**TN**	**FP**	**FN**	**TPR**	**TNR**	**Accuracy**
TAB-seq	All cytosines	SVM	(SVM NPC OFS)	NPC (depth >= 60)	H1 (depth >=60)	80328	1510583	705685	45364	0.64	0.68	0.68
TAB-seq	All cytosines	SVM	(SVM H1 OFS)	H1 (depth >=60)	NPC (depth >= 60)	295697	6559183	2288675	158052	0.65	0.74	0.74

4.5.2 OFS feature contributions: We constructed a dendrogram (Fig. 4.8 for the 5-hmC status prediction OFS, and eliminated subsets of features (see Methods Chapter 3)).

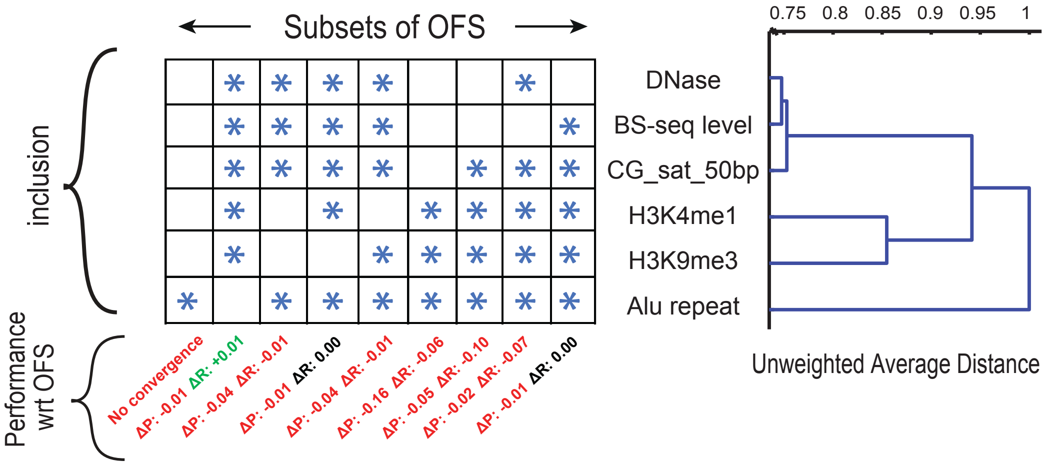


Fig. 4.8: OFS feature clustering. At each node, leaves (features) under it were removed from OFS to create new feature sets. For these, feature inclusion (starred) and resultant change in precision/recall w.r.t. OFS (by reclassifying dataset) characterize features' contribution to classification quality

The most notable changes to recall were observed upon elimination of the BS-seq CCR feature, while precision was affected by H3K4me1 and GC saturation removal, signifying the importance of these features to the prediction rate. Only four features (BS-seq CCR, GC saturation, DNase, and Alu) are sufficient to capture the majority of TAB-seq signal by garnering 0.75 F-score in NPC (Fig. 4.8). Several of these were identified in the literature to be enriched in regions of high hydroxymethylation (114). We show our 5-hmC prediction at work in two genomic regions proximal to PCDH17 and MEIS2 genes (Fig. 4.9, Fig. 4.10), previously implicated in synapse formation and interneuron development (137, 138).

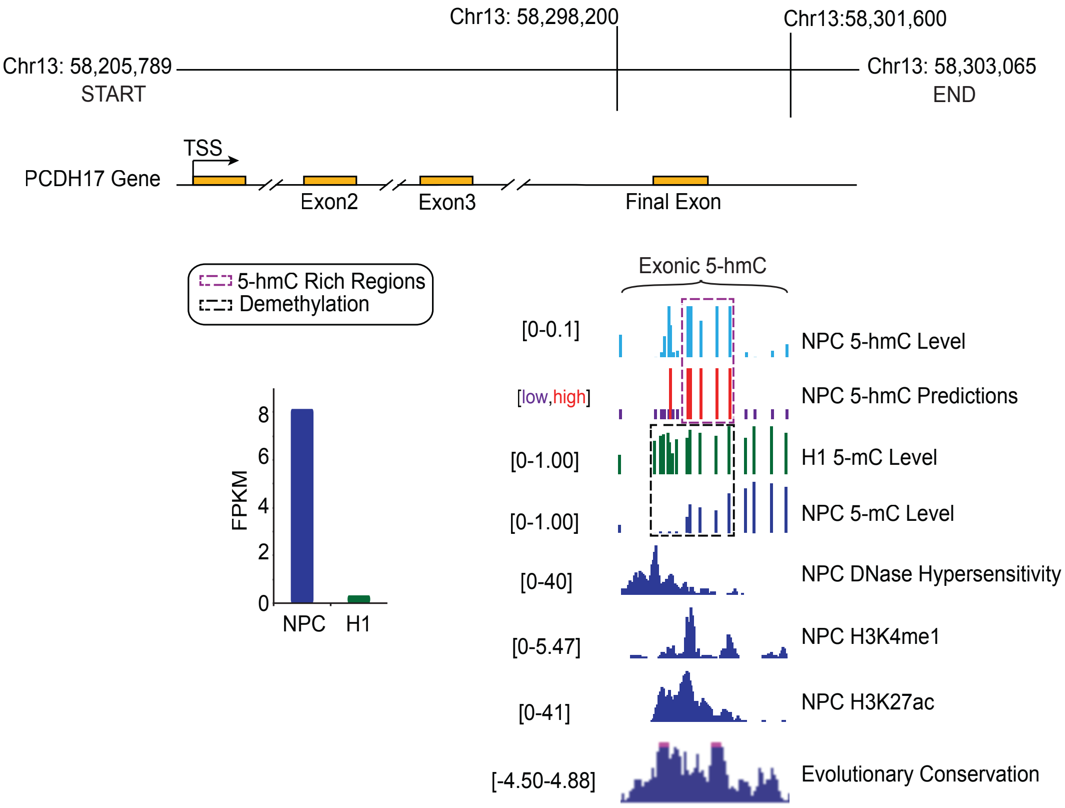


Fig. 4.9: Visualization of 5-hmC status prediction and discriminative input features in PCDH17

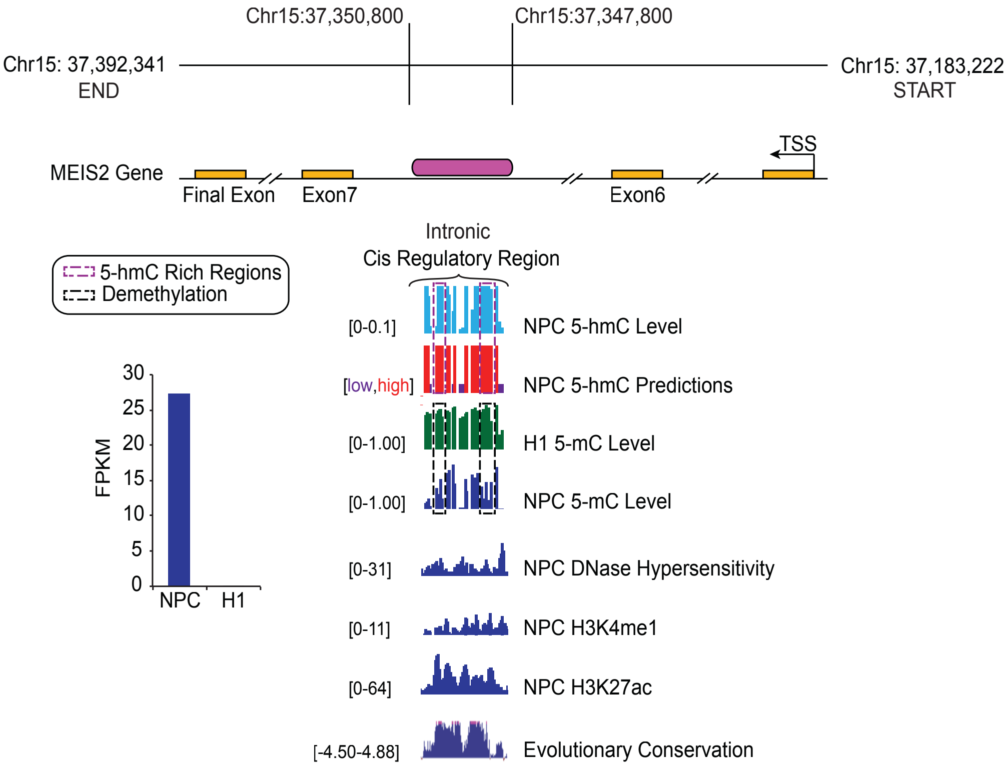
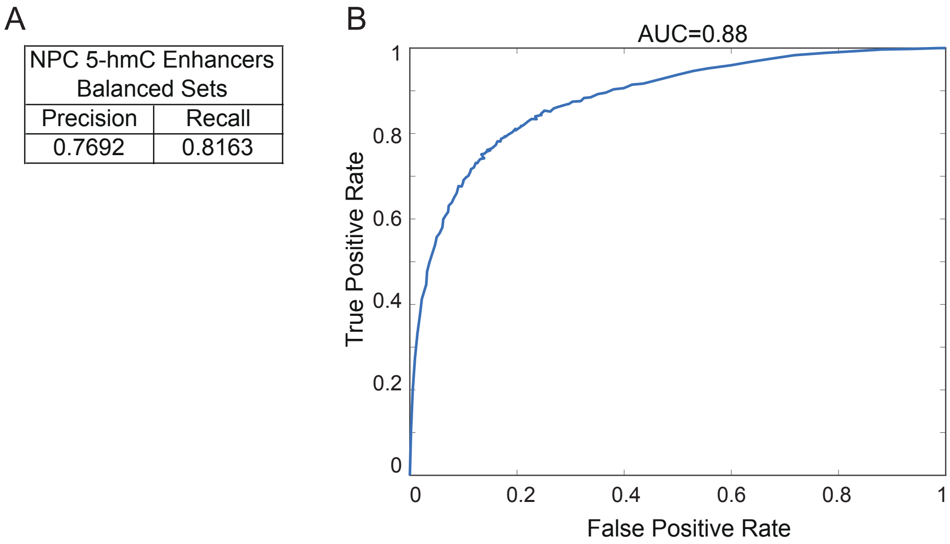


Fig. 4.10: Visualization of 5-hmC status prediction and discriminative input features in MEIS2

4.5.3 Overall 5-hmC prediction in enhancer regions

H3K27ac raw SRA file (accession GSM818031) was used for the purpose of identifying enhancers in NPC. Raw SRA files were mapped to the reference human hg19 genome using Bowtie2 to create the bam file. The obtained bam file was used as an input to the enhancer-calling tool ROSE (139). Thus, since 5-hmC is differentially enriched in functionally important enhancers (116), we trained and tested our model by restricting it only to NPC enhancers, obtaining 0.77 precision, 0.82 recall (Fig. 4.11 (A)) and a high AUC (Fig. 4.11 (B)).



NPC 5-hmC Enhancers Balanced Sets	
Precision	Recall
0.7692	0.8163

Fig. 4.11: 5-hmC status prediction in enhancer regions using SVM and RF models. (A) 5-hmC status prediction in NPC enhancers using SVM model. (B) ROC curve for 5-hmC predictions in enhancer regions in NPC RF model.

The active enhancer mark H3K27ac was present in the OFS (Table 4.1) suggesting a correlation of 5-hmC with enhancer activation. A significant improvement in the maximum precision feature set (HP) was found in models constrained to enhancers (Fig. 4.12), due to 5-hmC overabundance in enhancers.

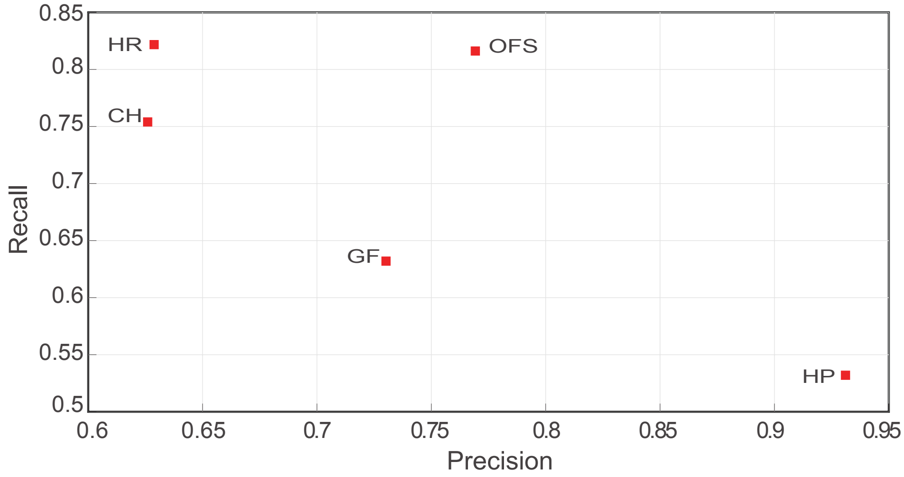


Fig. 4.12: Precision/Recall plot for 5-hmC status predictions in NPC in enhancer regions using balanced sets for SVM.

4.5.4 5-hmC prediction in small TAB-seq datasets

BS-seq and TAB-seq datasets require high sequencing depth to reliably determine CpG methylation and 5-hmC status across the genome, but as coverage decreases in smaller datasets, the ability to do so is diminished. The feasibility of training a model (like SVM) does not decrease proportionally to dataset size, as we can train SVMs with as few as 2000-2500 training

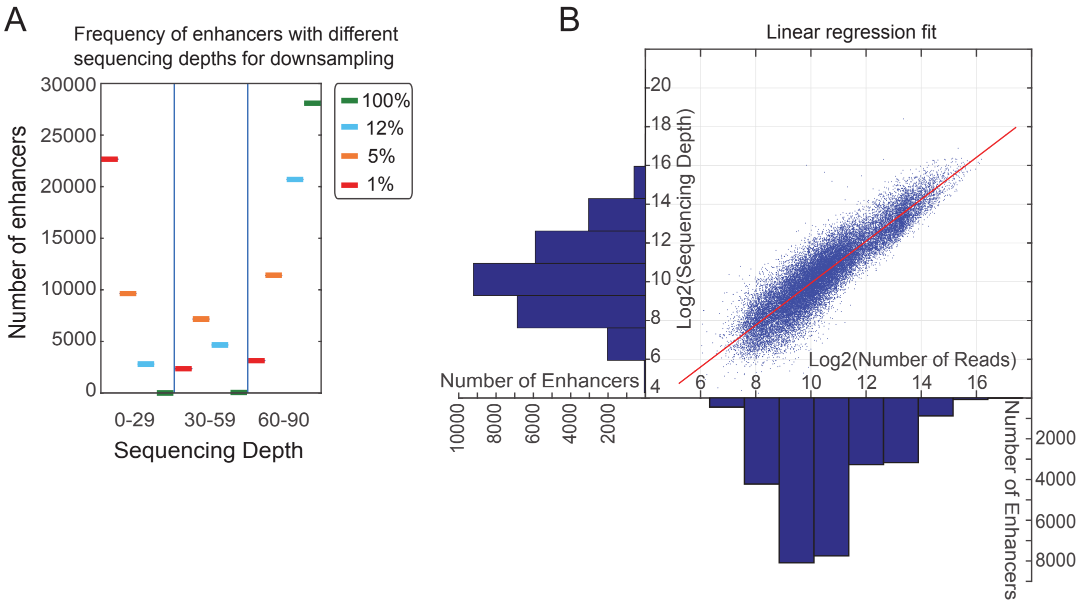


Fig. 4.13: Sequencing depth in NPC enhancers. (A) Sequencing depth across cytosines in enhancers after downsampling. B) Log-log linear regression fit (mapped read count vs. sequencing depth) in NPC enhancers

examples (Fig. 3.2). We downsampled one of our NPC datasets to 12% (commensurate with RRBS-seq dataset sizes (35)) of the original number of reads, and predicted the corresponding sequencing depth in enhancer CpG cytosines (Fig. 4.13 (A), 4.13 (B)). Finally, we find sufficient training examples (>2000) at resolutions of both whole enhancers and individual cytosines with sequencing depths suited for reliable CCR estimation in training SVMs, suggesting feasibility of robustly training 5-hmC status prediction models in enhancers for reduced representation TAB-seq data (see *5-hmC prediction feasibility in enhancers for reduced representation datasets*).

5-hmC prediction feasibility in enhancers for reduced representation datasets: Simulations were performed in enhancer regions to create downsampled TAB-seq datasets. The overall number of TAB-seq reads (approx. 84,000,000) were downsampled to different downsampling levels (75, 50, 25, 12, 5, 1 percentage of original-not all shown) (Fig. 4.13 (A)). A linear regression was used to fit the number of reads mapping to the enhancer to the sum of sequencing depth (Fig. 4.13 (B)) across all cytosines in it. 25 downsampling operations for each downsampling level were performed, and the obtained variance was low as shown in the box plot (Fig. 4.13 (A)). The histogram was divided into 3 categories: low, medium and high sequencing depth. Enhancers with the sum of cytosine sequencing depths > 60 were regarded as high, based on the Fisher test obtained p-value < 0.05 for discriminating against high and low hydroxymethylation coverage (Table 3.2).

As shown in Fig. 4.13 (A), downsampling to 5% of the total number of reads still leaves more than 10,000 enhancer regions with the sum of cytosine sequencing depths (≥ 60) and over 2,000 cytosines with individual sequencing depths ≥ 20 (corresponding to our sequencing depth levels required for assigning class labels: details on filtering training data based on sequencing depth in Methods section Chapter 3) which suffices for the purpose of training our classifier at the resolution of individual enhancers. The downsampling size of 12% contains ~10 million reads, which corresponds to the amount of RRBS-seq reads obtained in previous studies (140, 141). This suggests that even for RRBS-seq datasets, it is possible to train a model to successfully reconstruct the hydroxymethylome in enhancer regions, especially if the underlying implementation of a predictive model is predicated upon SVMs, which can be adequately trained and tested with as low as 1000-2000 training examples (Fig. 3.2).

4.5.5 *In silico* framework for high-throughput hypothesis testing

Hypothesis testing using TAB-seq data to identify 5-hmC rich regions or differential 5-hmC enrichment across conditions, naturally leads to a feasibility study of performing such tests on *in silico* predictions. 5-hmC is an intermediate in the demethylation pathway and low DNA methylation levels are the hallmark of active enhancers (114). Thus, we hypothesized that increase in an enhancer's 5-hmC enrichment (quantified as 5-hmC enrichment ratio, see *5-hmC ratio calculation*) from H1 to NPC differentiation corresponded to changes in proximal gene expression, putatively indicative of functional differences between H1 and NPC. We identified enhancers with the largest changes in 5-hmC enrichment ratio using both experimental TAB-seq data and our 5-hmC predictions. Gene set enrichment analysis (see *Gene ontology analysis)* on proximal genes to the identified enhancers reveal similar results for the two gene sets, enriched in neurodevelopmental processes. We find differential expression between H1 and NPC in the prediction-based gene set, suggesting our prediction-based functional study yields biologically relevant findings (Fig. 4.14, and Supp Data 1, Supp Data 2, Supp Data 3 which can be downloaded under the following URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btx316#supplementary-data).

5-hmC ratio calculation: Unlike DNA methylation status, hydroxymethylation status cannot be successfully imputed using neighboring CpG site information (Table 4.4), suggesting that CpG sites of similar hydroxymethylation status do not occur in as frequent and long stretches as similarly DNA methylated CpG sites. Hence, we devised a metric for identifying 5-hmC enrichment in a given genomic region. We used this to identify 5-hmC enrichment in enhancers.

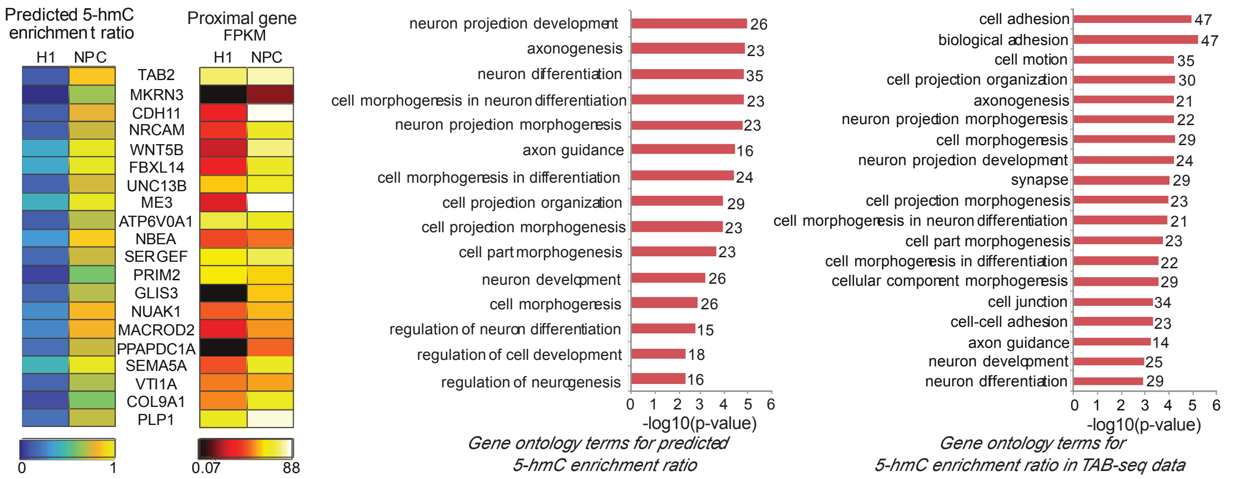
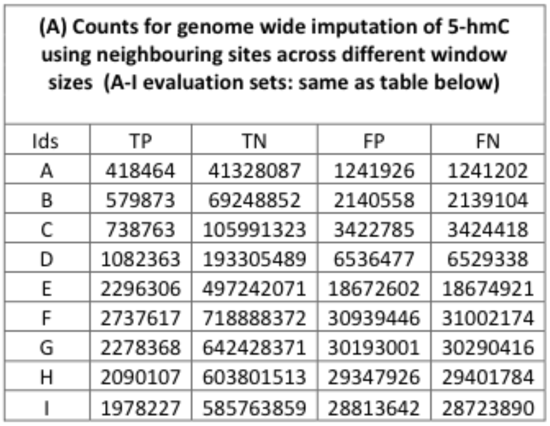


Fig. 4.14: Enhancers with high 5-hmC enrichment and their proximal genes. Heatmap of predicted 5-hmC enrichment ratio and proximal gene expression for enhancers with highest predicted gain in 5-hmC enrichment ratio (NPC vs. H1). GO term enrichment for genes with highest 5-hmC enrichment ratio (NPC vs. H1) using predictions and TAB-seq data.

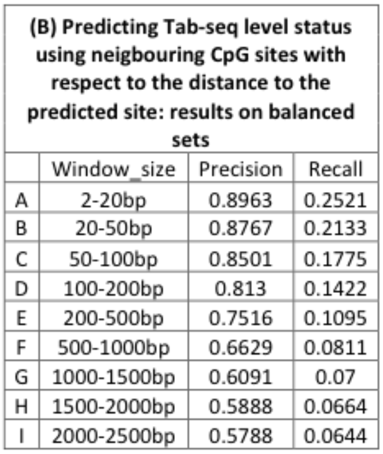
GTF hg19 files were obtained from UCSC Genome Browser (109), and further intersected with an available list of annotated enhancers (Supp Data 1). The regions that contained less than 10 CpG sites upon intersecting with enhancer and gene annotations were discarded from analyses. We define the ratio of the number of 5-hmC modified cytosines to the sum of 5-hmC and 5-mC modified cytosines in an enhancer as the 5-hmC enrichment ratio. We performed calculation of 5-hmC enrichment ratio in the intragenic enhancer regions, using 5,000 bp sliding windows spanning intragenic enhancers. 5-hmC enrichment ratio in a given region is defined as the ratio of the number of cytosines with 5-hmC modification to the number of cytosines with 5-hmC or 5-mC modification. This may be estimated using BS-seq data, or based on SVM predictions. Genes depicted in Fig. 4.14 were sorted based on the gain of 5-hmC enrichment ratio in intragenic enhancers in NPC versus H1 (Supp Data 2, Supp Data 3).

Table 4.4: Prediction metric: TP, TN, FP, FN (A) and Precision and Recall (B) for 5-hmC status prediction based on nearest neighbor's 5-hmC status, showing that such an approach is not feasible for 5-hmC status imputation.



(A) Counts for genome wide imputation of 5-hmC using neighbouring sites across different window sizes (A-I evaluation sets: same as table below)

Ids	TP	TN	FP	FN
A	418464	41328087	1241926	1241202
B	579873	69248852	2140558	2139104
C	738763	105991323	3422785	3424418
D	1082363	193305489	6536477	6529338
E	2296306	497242071	18672602	18674921
F	2737617	718888372	30939446	31002174
G	2278368	642428371	30193001	30290416
H	2090107	603801513	29347926	29401784
I	1978227	585763859	28813642	28723890



(B) Predicting Tab-seq level status using neigbouring CpG sites with respect to the distance to the predicted site: results on balanced sets

	Window_size	Precision	Recall
A	2-20bp	0.8963	0.2521
B	20-50bp	0.8767	0.2133
C	50-100bp	0.8501	0.1775
D	100-200bp	0.813	0.1422
E	200-500bp	0.7516	0.1095
F	500-1000bp	0.6629	0.0811
G	1000-1500bp	0.6091	0.07
H	1500-2000bp	0.5888	0.0664
I	2000-2500bp	0.5788	0.0644

4.5.6 Summary and Discussion

Here we performed a pioneering work by using DIRECTION to predict 5-hmC status in NPC and H1 cell lines to obtain high whole-genome accuracy (0.82 in NPC, Table 4.3). Upon performing optimal feature selection we discovered that most predictive features for 5-hmC status prediction involve enhancer-like histone modifications such as H3K27ac and H3K4me1. These findings naturally lead to predicting 5-hmC status in enhancer regions and collating our predictions in analyzed cell types for the purpose of identifying and characterizing differentially hydroxymethylated regions in analyzed cell types. We noticed that regions of differential 5-hmC (calculated using 5-hmC enrichment, see *5-hmC ratio calculation*) exhibit high correlation with expression profiles of proximal genes, and that our predictions in these regions are highly congruent with the ground truth obtained by TAB-seq experiments (Fig. 4.14).

Our work opens up new directions in DNA methylation and hydroxymethylation studies. Discriminative feature sets for predicting 5-hmC status include features engineered to leverage idiosyncrasies of hydroxymethylation, like strand asymmetry, G-rich sequence bias, and enrichment in open chromatin and gene bodies. Such correlative descriptions of 5-hmC modification with respect to genomic and epigenomic features can help create fine-grained "epigenome states" by integrating 5-mC and 5-hmC modifications with histone mark based chromatin states (57) in the future.

DIRECTION is the first *in-silico,* whole-epigenome predictor of DNA methylation and 5-hmC status at single nucleotide resolution, with results comparable to state-of-the-art DNA methylation prediction tools. One of the key differences that sets DIRECTION apart from other predictors is the ability to predict 5-hmC modifications. Our tool allows us to identify candidate genomic regions for differential hydroxymethylation as a first step in functional studies. 5-hmC modifications are known to be cell-type or developmental stage specific (120), and hence *in silico* detection of differentially hydroxymethylated regions can be performed by integrating reduced representation datasets and available genomic and epigenomic traits using DIRECTION. *In silico* prediction of epigenetic biomarkers is currently not performed in the field. However, cell-type specific methylation or 5-hmC regions can be predicted by collating predictions in contiguous genomic regions. The key to such predictions is that one or more of the input features need to contain signal for enabling such predictions. Here, similar analysis was performed based on the predicted 5-hmC ratio changes between H1 and NPC. Consequently, *in silico* prediction of epigenetic marks (i.e. 5-mC and 5-hmC) can be potentially considered as the first high-throughput step in discovering molecular mechanisms that are specific to a certain cell or tissue type, and such predictive analyses can be key to identifying starting "molecular targets" for in depth functional analyses.

5-hmC status prediction and correlative studies: The molecular mechanisms underlying 5-hmC creation and potential maintenance in the genome, its stability and regulatory potential, are presently all subject to a lot of scientific debate (142, 143). As we have shown, DIRECTION is capable of testing predictive powers of different sets of genomic and epigenomic features with respect to 5-hmC status prediction. Such correlative studies, in conjunction with perturbation models, can lead to a better understanding of 5-hmC.

Potential for use in oxBS-seq datasets: The oxBS-seq protocol (40) allows for positive readouts of 5-mC modifications (as opposed to 5-hmC modifications in TAB-seq experiments). As future work, we will consider additional experiments to train a model for directly predicting 5-mC modifications. However, likelihood based models like MLML (31) can integrate datasets from any two of BS-seq, oxBS-seq and TAB-seq datasets, to estimate CCRs for the third. Estimated CCRs for TAB-seq or BS-seq datasets generated in this fashion can then be used for analysis in DIRECTION.

CHAPTER 5

DNA METHYLATION AND HYDROXYMETHYLATION TRANSFER LEARNING

Authors: Milos Pavlovic, Pradipta Ray, Kristina Pavlovic, Aaron Kotamarti,
Min Chen and Michael Q. Zhang
Department of Biological Sciences
The University of Texas at Dallas
800 W. Campbell Road
Richardson, TX 75080-3021

5.1 Prior Publication

Milos Pavlovic (M.P.) performed the majority of experiments, and Pradipta Ray (P.R.) designed the majority of experiments. M.P. and P.R. wrote the manuscript. Kristina Pavlovic (K.P.) performed methylome clustering. Min Chen (M.C.) advised M.P. and Michael Zhang (M.Q.Z.) supervised the project. This chapter covers various aspects of DNA methylation and hydroxymethylation transfer learning predictions across various cell types, such as *de novo* methylome and hydroxymethylome reconstruction, feasibility of performing such predictions and the limit of such approach. Per the policy of OUP Bioinformatics, the publication of material in a PhD thesis is permitted with the publication of a peer-reviewed manuscript in their journal. The original manuscript (1) "DIRECTION: A machine learning framework for predicting and charactering DNA methylation and hydroxymethylation in mammalian genomes" by Milos Pavlovic, Pradipta Ray, Kristina Pavlovic, Aaron Kotamarti, Min Chen and Michael Q. Zhang, published in 2017, is reproduced by permission of Oxford University Press and appears online at the following web address: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btx316. Supplementary information is available online at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btx316. The main text is partially altered compared to the online version of the manuscript, and figures and tables do not chronologically correspond to the online manuscript numbering.

5.2 Abstract

Successful transfer learning between two cell or tissue types requires that a set of discriminative input features and its associated model decision boundary in one cell type, also have comparable

predictive power in the other. In order to demonstrate feasibility and applicability of such approach we performed a supervised version of transfer learning in H1 and NPC cell lines using DNA methylation and hydroxymethylation as response variables, and beam search identified OFSs as predictor variables. For DNA methylation prediction we successfully transferred learning between NPC and H1, and from NPC to MSC, while NPC to IMR90 transfer learning accuracy was modest. For 5-hmC status prediction, we successfully performed transfer learning between H1 and NPC since the whole-genome accuracy dropped by only a few percentage points. Our results show that transfer learning of epigenomic characteristics in developmentally non-divergent cell types is accurate and feasible, allowing *de novo* methylome and hydroxymethylome reconstruction, suggesting a significant degree of conservation among the epigenomes in question.

5.3 Introduction

Transfer learning is a machine learning technique in which a set of input features which was used to predict a certain response variable is reused to predict another response variable (106). Such learning can be performed in unsupervised and supervised fashion using both classification and regression based predictive models. Unsupervised learning classification is a procedure in which a set of input features that was used to predict status of one response variable is used to predict the missing class labels of another response variable. Unlike unsupervised transfer learning classification algorithms which major goal is an assignment of the missing class labels, supervised transfer learning algorithms aim to discover if such learning is feasible, assign class labels for the missing data, and determine what predictor variables are contributing the most to the predictive power. We performed a supervised version of transfer learning in cell lines (which genomic sequence is the same unless otherwise stated) using DNA methylation and hydroxymethylation status as response variables and beam search identified OFSs (see Methods Chapter 3) as predictor variables. The aforementioned experimental setup was primarily used as a feasibility study of performing transfer learning among assorted cell lines for the purpose of *de novo* methylome and hydroxymethylome reconstruction, and to discover what epigenetic traits are most predictive and thereby “likely conserved ” among the analyzed cell types.

5.4 DNA methylation transfer learning

Successful transfer learning between two cell types requires that a set of discriminative features and its associated model decision boundary in one cell type, also have comparable predictive power in the other. Given that one of our goals is to perform whole-methylome reconstructions, we trained our classifier on H1 cells and tested its performance on NPC and vice versa. The results of the testing are only a few percentage points worse than the corresponding results in the same cell type (Table 3.9), due to the fact that our approach relies on a minimal set of discriminative features (OFS) which are similar in H1 and NPC, and therefore has great promise for “transfer learning” scenarios like *de novo* reconstruction of the methylome.

Since H1 and NPC cell lines are not significantly divergent in their developmental stages, their respective epigenomes do not differ much. In order to identify methylomes that are significantly dissimilar with respect to H1 and NPC, we performed clustering for all reference methylomes in the Roadmap Epigenome Consortium datasets (top eight principal components accounting for 81% of variation in the data, Euclidean distance measure and average linkage were used, Fig. 5.1).

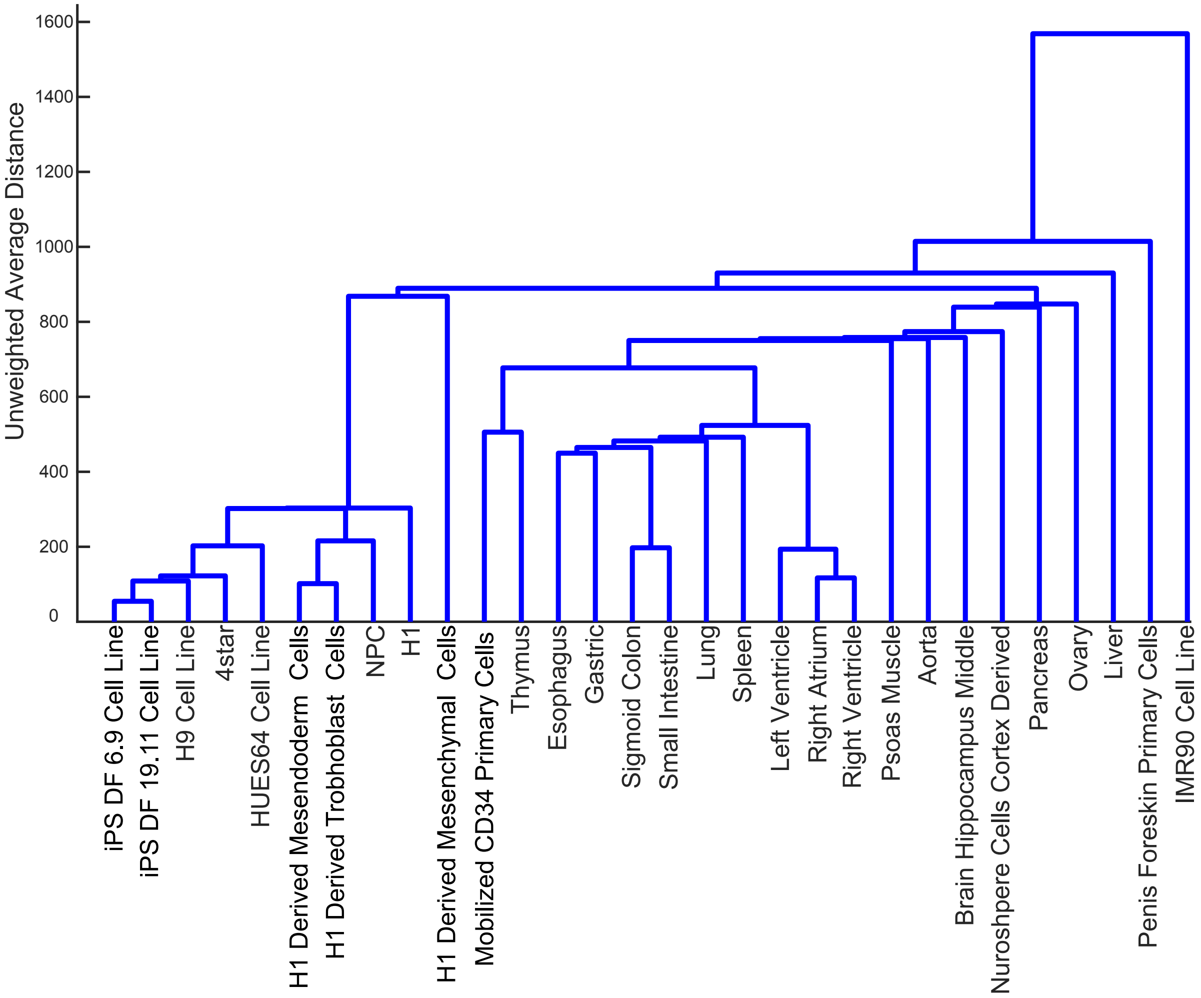


Fig. 5.1: Hierarchical clustering of reference methylomes from the Roadmap Epigenome Consortium depicting distinctive methylation profile of IMR90 with respect to H1 and NPC cell lines.

In order to test the limits of such transfer learning, we used the NPC-trained SVM to perform whole methylome predictions in the totipotent MSC cell line and in the terminally differentiated fetal fibroblast cell line IMR90. We chose to use the methylomes for Mesenchymal Stem Cells (MSC), and fetal fibroblast cell line IMR90, which show distinct divergence from the H1 and NPC methylomes. We analyzed the predictive performance for the NPC-trained predictive model on H1, MSC, and IMR90 methylomes. Since loss of pluripotency is associated with epigenome reprogramming involving DNA methylation, we find that the NPC-trained SVM performs well on the MSC dataset, but performs only modestly in IMR90 (Table 3.9).

We find that H1 predictions using the NPC-trained model are comparable to the NPC whole methylome predictions. The metrics for the MSC cell line (totipotent, but nearly terminally differentiated) are still fairly accurate (TPR: 0.87, TNR: 0.71, Accuracy: 0.85) (Table 3.9). However, we find that for the terminally differentiated IMR90, the metrics for the predictions are very modest (TPR: 0.86, TNR: 0.23, Accuracy: 0.69) (Table 3.9). This suggests that transfer learning only works within similar methylation paradigms, where the relationship between methylation and discriminative input features are similar. Given that the methylation profile and prevalence in stem cells and terminally differentiated cells are very distinct, we find that such transfer learning is not feasible. It is noteworthy that for evaluating IMR90, the sex chromosomes were left out during evaluation, as IMR90 is a female cell line, as opposed to H1, NPC, and MSC.

5.5 5-hmC transfer learning

In analogous fashion to our methylation data, we trained our classifier on H1 cells and tested its performance on NPC and vice versa. The results of the testing suggest that transfer learning across H1 and NPC is feasible (Table 4.3). We find that accuracies of 0.74 (H1 to NPC) and 0.68 (NPC to H1) as opposed to 0.82 and 0.75 by training and testing on the same cell type in NPC and H1 respectively. These results firmly suggest that 5-hmC transfer learning between developmentally non-divergent cell types such as H1 and NPC is feasible.

5.6 Limits of transfer learning using DIRECTION

Transfer learning for the purposes of prediction requires that the set of input features used for prediction in the source dataset, are discriminative in the target dataset and have similar correlational structure (144). While an in-depth analysis of transfer learning for methylation prediction is beyond the scope of this dissertation, we used the NPC-trained methylation prediction SVM to predict the methylome in H1, MSC and IMR90. Based on our NPC-trained SVM's performance in the whole genome NPC dataset, we find drops in accuracy in the pluripotent H1 (7% decrease) and near-differentiated, totipotent MSC cell lines (11% decrease). However, the accuracy for the NPC-trained SVM in the terminally differentiated IMR90 cell line drops by over 25%, suggesting that the OFS and SVM decision boundary for NPC is not suited for predicting the IMR90 methylome. Such results are in agreement with studies showing large-scale epigenetic reprogramming during differentiation (145) that likely causes a change in the correlational structure between the input features and the response variable (DNA methylation status). The limited number of input features in the OFS used by DIRECTION, while practical, does not lend itself to transfer learning in such scenarios. However, transfer learning is potentially feasible in closely related cell types or conditions where methylation paradigms remain unchanged.

CHAPTER 6

SUMMARY

Authors: Milos Pavlovic, Pradipta Ray, Kristina Pavlovic, Aaron Kotamarti,
Min Chen and Michael Q. Zhang
Department of Biological Sciences
The University of Texas at Dallas
800 W. Campbell Road
Richardson, TX 75080-3021

6.1 Prior Publication

Milos Pavlovic (M.P) wrote this chapter. This chapter represents a comprehensive summary of the previous chapters, and relates this work to the previous work in the field of DNA methylation prediction. Future studies, as well as the relevance of this work in the future are discussed as well. Per the policy of OUP Bioinformatics, the publication of material in a PhD thesis is permitted with the publication of a peer-reviewed manuscript in their journal. The original manuscript (1) "DIRECTION: A machine learning framework for predicting and charactering DNA methylation and hydroxymethylation in mammalian genomes" by Milos Pavlovic, Pradipta Ray, Kristina Pavlovic, Aaron Kotamarti, Min Chen and Michael Q. Zhang, published in 2017, is reproduced by permission of Oxford University Press and appears online at the following web address: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btx316. Supplementary information is available online at: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btx316. The main text is partially altered compared to the online version of the manuscript.

Summary: This work opens several new directions in mammalian DNA methylation studies, offering novel and incremental improvements. Here we introduced DIRECTION, the first *in-silico* whole-epigenome predictor of DNA methylation and 5-hmC status at single nucleotide resolution, with results comparable to state-of-the-art DNA methylation prediction tools.

For the purpose of performing whole methylome reconstructions we identified a portion of CpG sites which respective methylation statuses were invariant across a set of reference methylomes (Chapter 2) to create the *consensus reference methylome,* which we optionally use as a dictionary-based lookup to make methylation prediction (Fig. 2.1). Therefore, in order to achieve highly accurate *de novo* whole-methylome reconstruction we propose a novel framework, which combines the *consensus reference methylome* predictions and model-based predictions on the invariant and variant portions of the methylome respectively (achieving incremental improvement: 0.97 accuracy in NPC), and splits methylation prediction into separate paradigms (CGI vs. non-CGI) based on the underlying biology of DNA methylation. However, the use of the *consensus reference methylome* is likely most useful in resource-scarce scenarios (lack of available input features), and in the studies that aim to identify differentially methylated regions in a reconstructed methylome. It is noteworthy that the consensus reference methylome was not envisioned to be a fixed entity but rather a flexible one, which can be built using a different set of reference methylomes depending on the reconstructed methylome in question. Such approach can potentially provide insights into aberrant CpG methylation patterns in disease and perturbation studies which are known to affect DNA methylation.

The advent and subsequent improvement of NGS has lead to a development of experimental techniques such as sc-BS-seq, which has allowed us to quantify DNA methylation in single cells, where DNA methylation is essentially a discrete phenomenon, thereby lending itself naturally into a classification framework. The most recent work (80) successfully predicted DNA methylation status in single cells. Given the flexibility of DIRECTION, single cell derived features can be readily incorporated, and such predictions can be performed, making DIRECTION agnostic to the underlying input data (single cell vs. bulk of cells).

Discriminative feature sets for predicting 5-hmC status include features engineered to leverage idiosyncrasies of hydroxymethylation, like strand asymmetry, G-rich sequence bias, and enrichment in open chromatin and gene bodies. Such correlative descriptions of 5-hmC modification with respect to genomic and epigenomic features can help create fine-grained “epigenome states” by integrating 5-mC and 5-hmC modifications with histone mark based chromatin states (57) in the future. Often, the major goal of assorted whole-genome methylation studies is to detect differentially methylated regions either as putative biomarkers or for functional downstream studies. We demonstrated that DIRECTION is capable of identifying candidate genomic regions for differential hydroxymethylation, which serve as putative cis-regulatory elements of proximal protein coding genes (Fig. 4.14). In addition to this, the *in silico* detection of these regions can potentially lead to a development of less labor intensive protocols for the quantification of 5-hmC modification (reduced representation sequencing or potentially arrays), analogous to RRBS-seq which is routinely used for the quantification of DNA methylation.

Unlike previous feature-intensive approaches for predicting DNA methylation, DIRECTION uses a sophisticated feature selection technique adopted from artificial intelligence and identifies a small subset of non-redundant, discriminative, predictive features. This allows for greater biological interpretability of generated results, superior performance in resource-scarce scenarios, making the model sparse without explicit regularization. DIRECTION is an open-source, agile, scalable ensemble predictor using biologically and practically motivated genome partitioning and training a predictive model per partition, allowing us to deconvolute inevitably mixed biological signals in whole-epigenome studies (Fig. 2.1).

In the future, we aim to extend DIRECTION by predicting DNA methylation and 5-hmC status in additional genomic contexts (like non-CpG cytosines), other methylation paradigms (like epigenetic reprogramming in gametes), and in non-mammalian species where methylation plays distinct functional roles. Finally, we aim to expand a set of *consensus reference methylomes*, in which each individual *consensus reference methylome* would correspond to a particular phenotype of interest, such as disease (cancer, non-cancer) or a developmental stage (differentiated, non-differentiated). Such biologically motivated database of *consensus reference methylomes* would provide an instant insight into potential aberrant patterns involving DNA methylation, and would significantly facilitate a *de novo* methylome reconstruction process of a cell or tissue type in question.

REFERENCES

1. Pavlovic M, Ray P, Pavlovic K, Kotamarti A, Chen M, Zhang MQ. DIRECTION: A machine learning framework for predicting and characterizing DNA methylation and hydroxymethylation in mammalian genomes. Bioinformatics. 2017.

2. Weinhold B. Epigenetics: the science of change. Environ Health Perspect. 2006 Mar;114(3):A160-7.

3. Portela A, Esteller M. Epigenetic modifications and human disease. Nat Biotechnol. 2010;28(10):1057-68.

4. Zhang G, Pradhan S. Mammalian epigenetic mechanisms. IUBMB Life. 2014;66(4):240-56.

5. Esteller M. Epigenetics in cancer. N Engl J Med. 2008;358(11):1148-59.

6. Narayan P, Dragunow M. Pharmacology of epigenetics in brain disorders. Br J Pharmacol. 2010;159(2):285-303.

7. Rasool M, Malik A, Naseer MI, Manan A, Ansari SA, Begum I, et al. The role of epigenetics in personalized medicine: challenges and opportunities. BMC medical genomics. 2015;8(1):S5.

8. Rideout WM,3rd, Eggan K, Jaenisch R. Nuclear cloning and epigenetic reprogramming of the genome. Science. 2001 Aug 10;293(5532):1093-8.

9. Crider KS, Yang TP, Berry RJ, Bailey LB. Folate and DNA methylation: a review of molecular mechanisms and the evidence for folate's role. Adv Nutr. 2012 Jan;3(1):21-38.

10. Rajakumara E, Nakarakanti NK, Nivya MA, Satish M. Mechanistic insights into the recognition of 5-methylcytosine oxidation derivatives by the SUVH5 SRA domain. Sci Rep. 2016 Feb 4;6:20161.

11. Vasu K, Nagaraja V. Diverse functions of restriction-modification systems in addition to cellular defense. Microbiol Mol Biol Rev. 2013 Mar;77(1):53-72.

12. Jin B, Li Y, Robertson KD. DNA methylation: superior or subordinate in the epigenetic hierarchy? Genes & cancer. 2011;2(6):607-17.

13. Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell. 1999;99(3):247-57.

14. Hermann A, Goyal R, Jeltsch A. The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites. J Biol Chem. 2004 Nov 12;279(46):48350-9.

15. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Reviews Genetics. 2012;13(7):484-92.

16. Wyatt G, Cohen S. A new pyrimidine base from bacteriophage nucleic acids. Nature. 1952;170(4338):1072-3.

17. Wu H, D'Alessio AC, Ito S, Wang Z, Cui K, Zhao K, et al. Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. Genes Dev. 2011 Apr 1;25(7):679-84.

18. Ficz G, Branco MR, Seisenberger S, Santos F, Krueger F, Hore TA, et al. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature. 2011;473(7347):398-402.

19. Branco MR, Ficz G, Reik W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. Nature reviews Genetics. 2012;13(1):7-13.

20. Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, et al. Global epigenomic reconfiguration during mammalian brain development. Science. 2013 Aug 9;341(6146):1237905.

21. Bachman M, Uribe-Lewis S, Yang X, Williams M, Murrell A, Balasubramanian S. 5-Hydroxymethylcytosine is a predominantly stable DNA modification. Nature chemistry. 2014;6(12):1049-55.

22. Yu M, Hon GC, Szulwach KE, Song C, Jin P, Ren B, et al. Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. Nature protocols. 2012;7(12):2159-70.

23. Khare T, Pai S, Koncevicius K, Pal M, Kriukiene E, Liutkeviciute Z, et al. 5-hmC in the brain is abundant in synaptic genes and shows differences at the exon-intron boundary. Nature structural & molecular biology. 2012;19(10):1037-43.

24. Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. Cell. 2014;156(6):1324-35.

25. Kim S, Li M, Paik HH, Nephew KP, Shi H, Kramer R, et al. Predicting DNA methylation susceptibility using CpG flanking sequences. Pacific Symposium on Biocomputing; Citeseer; 2008.

26. Yang H, Liu Y, Bai F, Zhang J, Ma S, Liu J, et al. Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. Oncogene. 2013;32(5):663-9.

27. Singer-Sam J, Robinson MO, Bellvé AR, Simon Ml, Riggs AD. Measurement by quantitative PCR of changes in HPRT, PGK-1, PGK-2, APRT, MTase, and Zfy gene transcripts during mouse spermatogenesis. Nucleic Acids Res. 1990;18(5):1255-9.

28. Suzuki M, Greally JM. DNA methylation profiling using HpaII tiny fragment enrichment by ligation-mediated PCR (HELP). Methods. 2010;52(3):218-22.

29. Mohn F, Weber M, Schübeler D, Roloff T. Methylated DNA immunoprecipitation (medip). DNA methylation: methods and protocols. 2009:55-64.

30. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, et al. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol. 2008;26(7):779-85.

31. Qu J, Zhou M, Song Q, Hong EE, Smith AD. MLML: consistent simultaneous estimates of DNA methylation and hydroxymethylation. Bioinformatics. 2013;29(20):2645-6.

32. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A. 1992 Mar 1;89(5):1827-31.

33. Li Y, Tollefsbol TO. DNA methylation detection: bisulfite genomic sequencing analysis. Epigenetics Protocols. 2011:11-21.

34. Clark SJ, Statham A, Stirzaker C, Molloy PL, Frommer M. DNA methylation: bisulphite modification and analysis. Nature protocols. 2006;1(5):2353-64.

35. Gu, Tian-Peng and Guo, Fan and Yang, Hui and Wu, Hai-Ping and Xu, Gui-Fang and Liu, Wei and Xie, Zhi-Guo and Shi, Linyu and He, Xinyi and Jin, Seung-gi and others. The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. Nature. 2011;477(7366):606-10.

36. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. 2005;33(18):5868-77.

37. Fraga MF, Esteller M. DNA methylation: a profile of methods and applications. BioTechniques. 2002 Sep;33(3):632, 634, 636-49.

38. Ehrich M, Zoll S, Sur S, Van Den Boom D. A new method for accurate assessment of DNA quality after bisulfite treatment. Nucleic Acids Res. 2007;35(5):e29.

39. Capra JA, Kostka D. Modeling DNA methylation dynamics with approaches from phylogenetics. Bioinformatics. 2014;30(17):i408-14.

40. Booth MJ, Branco MR, Ficz G, Oxley D, Krueger F, Reik W, et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. Science. 2012 May 18;336(6083):934-7.

41. Cohen-Karni D, Xu D, Apone L, Fomenkov A, Sun Z, Davis PJ, et al. The MspJI family of modification-dependent restriction endonucleases for epigenetic studies. Proc Natl Acad Sci U S A. 2011 Jul 5;108(27):11040-5.

42. Nestor CE, Meehan RR. Hydroxymethylated DNA immunoprecipitation (hmeDIP). Functional Analysis of DNA and Chromatin. 2014:259-67.

43. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27(11):1571-2.

44. Smith AD, Chung W, Hodges E, Kendall J, Hannon G, Hicks J, et al. Updates to the RMAP short-read mapping software. Bioinformatics. 2009;25(21):2841-2.

45. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009;10(1):232.

46. Guo W, Fiziev P, Yan W, Cokus S, Sun X, Zhang MQ, et al. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics. 2013;14(1):774.

47. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012;9(4):357-9.

48. Stevens M, Cheng JB, Xie M, Costello JF, Wang T. MethylCRF, an algorithm for estimating absolute methylation levels at single CpG resolution from methylation enrichment and restriction enzyme sequencing methods. Annual International Conference on Research in Computational Molecular Biology; Springer; 2013.

49. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011.

50. Burger L, Gaidatzis D, Schübeler D, Stadler MB. Identification of active regulatory regions from DNA methylation data. Nucleic Acids Res. 2013;41(16):e155-.

51. Shen L, Kondo Y, Ahmed S, Boumber Y, Konishi K, Guo Y, et al. Drug sensitivity prediction by CpG island methylation profile in the NCI-60 cancer cell line panel. Cancer Res. 2007 Dec 1;67(23):11335-43.

52. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. The American Journal of Human Genetics. 2012;90(1):7-24.

53. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends in genetics. 2008;24(8):408-15.

54. Banovich NE, Lan X, McVicker G, Van de Geijn B, Degner JF, Blischak JD, et al. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. PLoS genetics. 2014;10(9):e1004663.

55. Gibbs JR, van der Brug, Marcel P, Hernandez DG, Traynor BJ, Nalls MA, Lai S, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS genetics. 2010;6(5):e1000952.

56. Karmaus W, Ziyab AH, Everson T, Holloway JW. Epigenetic mechanisms and models in the origins of asthma. Curr Opin Allergy Clin Immunol. 2013 Feb;13(1):63-9.

57. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nature methods. 2012;9(3):215-6.

58. Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat Biotechnol. 2015;33(4):364-76.

59. Bhasin M, Zhang H, Reinherz EL, Reche PA. Prediction of methylated CpGs in DNA sequences using a support vector machine. FEBS Lett. 2005;579(20):4302-8.

60. Bock C, Paulsen M, Tierling S, Mikeska T, Lengauer T, Walter J. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. PLoS Genet. 2006;2(3):e26.

61. Das R, Dimitrova N, Xuan Z, Rollins RA, Haghighi F, Edwards JR, et al. Computational prediction of methylation status in human genomic sequences. Proc Natl Acad Sci U S A. 2006 Jul 11;103(28):10713-6.

62. Fang F, Fan S, Zhang X, Zhang MQ. Predicting methylation status of CpG islands in the human brain. Bioinformatics. 2006 Sep 15;22(18):2204-9.

63. Carson MB, Langlois R, Lu H. Mining knowledge for the methylation status of CpG islands using alternating decision trees. Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE; IEEE; 2008.

64. Feltus FA, Lee EK, Costello JF, Plass C, Vertino PM. DNA motifs associated with aberrant CpG island methylation. Genomics. 2006;87(5):572-9.

65. Goh L, Murphy SK, Muhkerjee S, Furey TS. Genomic sweeping for hypermethylated genes. Bioinformatics. 2006;23(3):281-8.

66. Ali I, Seker H. Detailed methylation prediction of CpG islands on human chromosome 21. 10th WSEAS International Conference on Mathematics and Computers In Biology and Chemistry; ; 2009.

67. Whitaker JW, Chen Z, Wang W. Predicting the human epigenome from DNA motifs. Nature methods. 2015;12(3):265-72.

68. Flores KB, Amdam GV. Deciphering a methylome: what can we read into patterns of DNA methylation? J Exp Biol. 2011 Oct 1;214(Pt 19):3155-63.

69. Zheng H, Wu H, Li J, Jiang S. CpGIMethPred: computational model for predicting methylation status of CpG islands in human genome. BMC medical genomics. 2013;6(1):S13.

70. McCabe MT, Brandes JC, Vertino PM. Cancer DNA methylation: molecular mechanisms and clinical implications. Clin Cancer Res. 2009 Jun 15;15(12):3927-37.

71. Wrzodek C, Büchel F, Hinselmann G, Eichner J, Mittag F, Zell A. Linking the epigenome to the genome: correlation of different features to DNA methylation of CpG islands. PloS one. 2012;7(4):e35327.

72. Shen L, Kantarjian H, Guo Y, Lin E, Shan J, Huang X, et al. DNA methylation predicts survival and response to therapy in patients with myelodysplastic syndromes. Journal of Clinical Oncology. 2009;28(4):605-13.

73. Luu PL, Scholer HR, Arauzo-Bravo MJ. Disclosing the crosstalk among DNA methylation, transcription factors, and histone marks in human pluripotent cells through discovery of DNA methylation motifs. Genome Res. 2013 Dec;23(12):2013-29.

74. Craig JM, Bickmore WA. The distribution of CpG islands in mammalian chromosomes. Nat Genet. 1994;7(3):376-82.

75. Ma B, Wilker EH, Willis-Owen SA, Byun H, Wong KC, Motta V, et al. Predicting DNA methylation level across human tissues. Nucleic Acids Res. 2014;42(6):3515-28.

76. Zhang W, Spector TD, Deloukas P, Bell JT, Engelhardt BE. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. Genome Biol. 2015;16(1):14.

77. Yan H, Zhang D, Liu H, Wei Y, Lv J, Wang F, et al. Chromatin modifications and genomic contexts linked to dynamic DNA methylation patterns across human cell types. Sci Rep. 2015 Feb 12;5:8410.

78. Wang Y, Liu T, Xu D, Shi H, Zhang C, Mo YY, et al. Predicting DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep Networks. Sci Rep. 2016 Jan 22;6:19598.

79. Fan S, Huang K, Ai R, Wang M, Wang W. Predicting CpG methylation levels by integrating Infinium HumanMethylation450 BeadChip array data. Genomics. 2016;107(4):132-7.

80. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. 2017;18(1):67.

81. Amoreira C, Hindermann W, Grunau C. An improved version of the DNA Methylation database (MethDB). Nucleic Acids Res. 2003;31(1):75-7.

82. Yamada Y, Watanabe H, Miura F, Soejima H, Uchiyama M, Iwasaka T, et al. A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. Genome Res. 2004;14(2):247-66.

83. Rollins RA, Haghighi F, Edwards JR, Das R, Zhang MQ, Ju J, et al. Large-scale structure of genomic methylation patterns. Genome Res. 2006;16(2):157-63.

84. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet. 2006;38(12):1378-85.

85. Bock C, Halachev K, B\uch J, Lengauer T. EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi) genomic data. Genome Biol. 2009;10(2):1.

86. Feltus FA, Lee EK, Costello JF, Plass C, Vertino PM. Predicting aberrant CpG island methylation. Proc Natl Acad Sci U S A. 2003 Oct 14;100(21):12253-8.

87. Bock C, Walter J, Paulsen M, Lengauer T. CpG island mapping by epigenome prediction. PLoS Comput Biol. 2007;3(6):e110.

88. Fan S, Zhang MQ, Zhang X. Histone methylation marks play important roles in predicting the methylation status of CpG islands. Biochem Biophys Res Commun. 2008;374(3):559-64.

89. Previti C, Harari O, Zwir I, del Val C. Profile analysis and prediction of tissue-specific CpG island methylation classes. BMC Bioinformatics. 2009;10(1):1.

90. Lu L, Lin K, Qian Z, Li H, Cai Y, Li Y. Predicting DNA methylation status using word composition. Journal of Biomedical Science and Engineering. 2010;3(7):672.

91. Fan S, Zou J, Xu H, Zhang X. Predicted methylation landscape of all CpG islands on the human genome. Chinese Science Bulletin. 2010;55(22):2353-8.

92. Zhang W, Zheng H, Zhang J. Nucleosome positioning plays an important role in predicting the methylation status of CpG islands. Biomedical Engineering and Informatics (BMEI), 2011 4th International Conference on; IEEE; 2011.

93. Zhou X, Li Z, Dai Z, Zou X. Prediction of methylation CpGs and their methylation degrees in human DNA sequences. Comput Biol Med. 2012;42(4):408-13.

94. Gaidatzis D, Burger L, Murr R, Lerch A, Dessus-Babus S, Sch\ubeler D, et al. DNA sequence explains seemingly disordered methylation levels in partially methylated domains of Mammalian genomes. PLoS Genet. 2014;10(2):e1004143.

95. Sulewska A, Niklinska W, Kozlowski M, Minarowski L, Naumnik W, Niklinski J, et al. DNA methylation in states of cell physiology and pathology. Folia histochemica et cytobiologica. 2007;45(3):149-58.

96. Clark SJ, Smallwood SA, Lee HJ, Krueger F, Reik W, Kelsey G. Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). Nature protocols. 2017;12(3):534-47.

97. Schrader A, Gross T, Thalhammer V, Längst G. Characterization of Dnmt1 binding and DNA methylation on nucleosomes and nucleosomal arrays. PloS one. 2015;10(10):e0140076.

98. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317-30.

99. Kulis M, Esteller M. DNA methylation and cancer. Adv Genet. 2010;70(10):27-56.

100. Bartolomei MS, Ferguson-Smith AC. Mammalian genomic imprinting. Cold Spring Harb Perspect Biol. 2011 Jul 1;3(7):10.1101/cshperspect.a002592.

101. Couldrey C, Brauning R, Bracegirdle J, Maclean P, Henderson HV, McEwan JC. Genome-wide DNA methylation patterns and transcription analysis in sheep muscle. PloS one. 2014;9(7):e101853.

102. Zeng J, Konopka G, Hunt BG, Preuss TM, Geschwind D, Soojin VY. Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. The American Journal of Human Genetics. 2012;91(3):455-65.

103. Elango N, Yi SV. DNA methylation and structural and functional bimodality of vertebrate promoters. Mol Biol Evol. 2008;25(8):1602-8.

104. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009 Nov 19;462(7271):315-22.

105. Ahn JB, Chung WB, Maeda O, Shin SJ, Kim HS, Chung HC, et al. DNA methylation predicts recurrence from resected stage III proximal colon cancer. Cancer. 2011;117(9):1847-54.

106. Bishop CM. Pattern recognition and machine learning. springer; 2006.

107. Breiman L. Random forests. Mach Learning. 2001;45(1):5-32.

108. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. Proceedings of the 23rd international conference on Machine learning; ACM; 2006.

109. Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, et al. The UCSC genome browser database: 2016 update. Nucleic Acids Res. 2015;44(D1):D717-25.

110. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2.

111. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.

112. Chadwick LH. The NIH roadmap epigenomics program data resource. . 2012.

113. Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, et al. The UCSC Genome Browser database: 2016 update. Nucleic Acids Res. 2016 Jan 4;44(D1):D717-25.

114. Yu M, Hon GC, Szulwach KE, Song C, Zhang L, Kim A, et al. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell. 2012;149(6):1368-80.

115. Hackett JA, Sengupta R, Zylicz JJ, Murakami K, Lee C, Down TA, et al. Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. Science. 2013 Jan 25;339(6118):448-52.

116. Stroud H, Feng S, Kinney SM, Pradhan S, Jacobsen SE. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. Genome Biol. 2011;12(6):R54.

117. Zhang W, Xia W, Wang Q, Towers AJ, Chen J, Gao R, et al. Isoform Switch of TET1 Regulates DNA Demethylation and Mouse Development. Mol Cell. 2016;64(6):1062-73.

118. Song C, Szulwach KE, Fu Y, Dai Q, Yi C, Li X, et al. Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. Nat Biotechnol. 2011;29(1):68-72.

119. Tan L, Xiong L, Xu W, Wu F, Huang N, Xu Y, et al. Genome-wide comparison of DNA hydroxymethylation in mouse embryonic stem cells and neural progenitor cells by a new comparative hMeDIP-seq method. Nucleic Acids Res. 2013 Apr;41(7):e84.

120. Wang T, Pan Q, Lin L, Szulwach KE, Song CX, He C, et al. Genome-wide DNA hydroxymethylation changes are associated with neurodevelopmental genes in the developing human cerebellum. Hum Mol Genet. 2012 Dec 15;21(26):5500-10.

121. Bishop C. Pattern Recognition and Machine Learning (Information Science and Statistics), 1st edn. 2006. corr. 2nd printing edn. Springer, New York. 2007.

122. Koller D, Sahami M. Toward optimal feature selection. Proceedings of 13th International Conference on Machine Learning. 1996:284-92.

123. Nguyen MH, De la Torre F. Optimal feature selection for support vector machines. Pattern Recognit. 2010;43(3):584-91.

124. Zhang W. Complete anytime beam search. AAAI/IAAI; ; 1998.

125. Zheng F, Zhou X, Moon C, Wang H. Regulation of brain-derived neurotrophic factor expression in neurons. Int J Physiol Pathophysiol Pharmacol. 2012;4(4):188-200.

126. Cuddapah S, Jothi R, Schones DE, Roh T, Cui K, Zhao K. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res. 2009;19(1):24-32.

127. Lazarovici A, Zhou T, Shafer A, Dantas Machado AC, Riley TR, Sandstrom R, et al. Probing DNA shape and methylation state on a genomic scale with DNase I. Proc Natl Acad Sci U S A. 2013 Apr 16;110(16):6376-81.

128. Ooi SK, O'Donnell AH, Bestor TH. Mammalian cytosine methylation at a glance. J Cell Sci. 2009;122(16):2787-91.

129. Kang M, Rhyu M, Kim Y, Jung Y, Hong S, Cho C, et al. The length of CpG islands is associated with the distribution of Alu and L1 retroelements. Genomics. 2006;87(5):580-90.

130. Plongthongkum N, Diep DH, Zhang K. Advances in the profiling of DNA modifications: cytosine methylation and beyond. Nature Reviews Genetics. 2014;15(10):647-61.

131. Kantlehner M, Kirchner R, Hartmann P, Ellwart JW, Alunni-Fabbroni M, Schumacher A. A high-throughput DNA methylation analysis of a single cell. Nucleic Acids Res. 2011 Apr;39(7):e44.

132. Guo H, Zhu P, Wu X, Li X, Wen L, Tang F. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Res. 2013 Dec;23(12):2126-35.

133. Marina RJ, Sturgill D, Bailly MA, Thenoz M, Varma G, Prigge MF, et al. TET-catalyzed oxidation of intragenic 5-methylcytosine regulates CTCF-dependent alternative splicing. EMBO J. 2015:e201593235.

134. Mellen M, Ayata P, Dewell S, Kriaucionis S, Heintz N. MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. Cell. 2012;151(7):1417-30.

135. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nature Reviews Genetics. 2014;15(4):272-86.

136. Yamaguchi S, Hong K, Liu R, Inoue A, Shen L, Zhang K, et al. Dynamics of 5-methylcytosine and 5-hydroxymethylcytosine during germ cell reprogramming. Cell Res. 2013;23(3):329-39.

137. Batista-Brito R, Rossignol E, Hjerling-Leffler J, Denaxa M, Wegner M, Lefebvre V, et al. The cell-intrinsic requirement of Sox6 for cortical interneuron development. Neuron. 2009;63(4):466-81.

138. Hoshina N, Tanimura A, Yamasaki M, Inoue T, Fukabori R, Kuroda T, et al. Protocadherin 17 regulates presynaptic assembly in topographic corticobasal Ganglia circuits. Neuron. 2013;78(5):839-54.

139. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013;153(2):307-19.

140. Bock C, Tomazou EM, Brinkman AB, M\uller F, Simmer F, Gu H, et al. Quantitative comparison of genome-wide DNA methylation mapping technologies. Nat Biotechnol. 2010;28(10):1106-14.

141. Chatterjee A, Rodger EJ, Stockwell PA, Weeks RJ, Morison IM. Technical considerations for reduced representation bisulfite sequencing with multiplexed libraries. BioMed Research International. 2012;2012.

142. Hahn MA, Szabó PE, Pfeifer GP. 5-Hydroxymethylcytosine: a stable or transient DNA modification? Genomics. 2014;104(5):314-23.

143. Shen L, Zhang Y. 5-Hydroxymethylcytosine: generation, fate, and genomic distribution. Curr Opin Cell Biol. 2013;25(3):289-96.

144. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowled Data Eng. 2010;22(10):1345-59.

145. Teif VB, Beshnova DA, Vainshtein Y, Marth C, Mallm JP, Hofer T, et al. Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development. Genome Res. 2014 Aug;24(8):1285-95.

BIOGRAPHICAL SKETCH

Milos Pavlovic received his BS degree from the University of Belgrade, Serbia majoring in Biology. He holds an MS degree in molecular and cell biology from The University of Texas at Dallas. His doctoral research explored the field of epigenetics mainly focusing on predicting mammalian DNA methylation using assorted computational methods. He contributed to the field by developing a software package that predicts both DNA methylation and hydroxymethylation in mammalian genomes. Additionally, he identified and accurately predicted the regions of differential cell type specific hydroxymethylation, which serve as putative cis-regulatory regions. Finally, he transferred learning of epigenomic characteristics across multiple cell types and identified a portion of mammalian DNA methylation that is invariant across mammalian methylomes for the purpose of genome-wide DNA methylation reconstruction.

CURRICULUM VITAE

Milos Pavlovic

Education

2013-present	University of Texas at Dallas	Degree: Doctor of Philosophy
	Major: Computational Biology	GPA: 3.559
	Dissertation Defense: October 2017	
2011-2013	University of Texas at Dallas	Degree: Master of Science
	Major: Molecular and Cell Biology	GPA: 3.613
2003-2009	University of Belgrade, Serbia	Degree: Bachelor of Science
	Major: Biology	GPA: 3.72

Research Experience

08/13/2013-present Worked in the computational biology lab under Dr. Michael Zhang

- Developed a machine learning algorithm for prediction of DNA methylation and hydroxymethylation status using Support Vector Machines and Random Forest
    - Paper published in OUP Bioinformatics
- Designed a Convolutional Neural Network and Random Forest predictive models for classification of eRNA producing and non-producing enhancers
- Designed a statistical framework using Linear Models for Genome Wide Association Studies and Expression Quantitative Trait Loci analysis of prostate cancer data

Publication

Direction: A machine learning framework for predicting and characterizing DNA methylation and hydroxymethylation in mammalian genomes.
Milos Pavlovic, Pradipta Ray, Kristina Pavlovic, Aaron Kotamarti, Min Chen, Michael Q. Zhang
Bioinformatics, doi: 10.1093/bioinformatics/btx316, 2017
Paper and tool can be found at the following URL: http://utdallas.edu/~mxp114330/direction/

Presentations

Regions of invariant DNA methylation in human epigenomes in the context of methylation prediction, and functional significance
Milos Pavlovic, Pradipta Ray, Michael Q. Zhang
Epigenomics 2016 Conference, San Juan, Puerto Rico, 2016

The immutable methylome: characterizing regions of invariant methylation in the mammalian genome as a counterfoil to discovering tissue specific methylation patterns
Pradipta_Ray , Milos Pavlovic , Michael Q. Zhang
Probabilistic Modeling in Genomics 2015, New York City, New York, 2015

Teaching Experience

Spring 2017	Teaching Assistant-Computational Biology

Honors & Scholarships

2011-2013	UT Dallas Chess Graduate Scholarship
2012	Pan American Chess Team Champion with UT Dallas
2007	International Master of Chess and a holder of 2 Grand Master norms
2002	Represented Serbia in European and World Junior Chess Championship

Technical Skills

Programming

- Python
- Matlab
- Unix shell
- SQL

Machine learning

- Support Vector Machines (RBF, Linear kernels)
- Random Forest
- Deep Learning (CNN, Forward Nets, Autoencoders)
- Logistic regression
- Linear regression
- Decision Trees
- K-means
- Collaborative filtering and recommendation systems

Big Data Managing

- Data dimensionality reduction
    - Linear techniques (PCA, LDA)
    - Non-Linear techniques (MDS, Autoencoders)
- Big data processing
    - HDFS, Hadoop (MAP/REDUCE)
    - Apache Hive
    - Apache Pig

Statistical modeling

- Bayesian Nets
- Markov chains
- Hidden Markov Models (Viterbi, Baum-Welch)
- EM algorithm
- Gibbs sampling