Spring 5-2016

Predicting DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep Networks

Yiheng Wang
University of Southern Mississippi

PREDICTING DNA METHYLATION STATE OF CPG DINUCLEOTIDE USING GENOME TOPOLOGICAL FEATURES AND DEEP NETWORKS

by

Yiheng Wang

A Thesis
Submitted to the Graduate School
and the School of Computing
at The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Master of Science

Approved:

_____
Dr. Zheng Wang, Committee Chair
Assistant Professor, School of Computing

_____
Dr. Zheng Sun, Committee Member
Associate Professor, School of Computing

_____
Dr. Nan Wang, Committee Member
Associate Professor, School of Computing

_____
Dr. Karen S. Coats
Dean of the Graduate School

May 2016

ABSTRACT

PREDICTING DNA METHYLATION STATE OF CPG DINUCLEOTIDE USING GENOME TOPOLOGICAL FEATURES AND DEEP NETWORKS

by Yiheng Wang

May 2016

The hypo- or hyper-methylation of the human genome is one of the epigenetic features of leukemia. However, experimental approaches have only determined the methylation state of a small portion of the human genome. I developed a deep learning based (stacked denoising autoencoders, or SdA) software named “DeepMethyl” to predict the methylation state of DNA CpG dinucleotides using features inferred from three-dimensional genome topology (based on Hi-C) and DNA sequence patterns. I used the experimental data from immortalized myelogenous leukemia (K562) and healthy lymphoblastoid (GM12878) cell lines to train the learning models and assess prediction performance. I have tested various SdA architectures with different configurations of hidden layer(s) and amount of pre-training data and compared the performance of deep networks relative to support vector machines (SVM). Using the methylation states of sequentially neighboring regions as one of the learning features, SdA achieved a blind test accuracy of 89.7% for GM12878 and 88.6% for K562. When the methylation states of sequentially neighboring regions are unknown, the accuracies are 84.82% for GM12878 and 72.01% for K562. I also analyzed the contribution of genome topological features inferred from Hi-C. DeepMethyl can be accessed at http://dna.cs.usm.edu/deepmethyl/.

ACKNOWLEDGMENTS

The authors thank Mr. Joseph Luttrell, an undergraduate student from the Honors College of The University of Southern Mississippi, for editing the English writing of the manuscript.

This work has been published: Wang, Y., Liu, T., Xu, D., Shi, H., Zhang, C., Mo, Y. Wang, Z. (2016) Predicting DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep Networks. Scientific Reports, 6, 19598; doi: 10.1038/srep19598.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF ILLUSTRATIONS

Figure

CHAPTER I

INTRODUCTION

DNA methylation represents the addition of a methyl group to the fifth carbon of the cytosine or adenine[1]. DNA methylation occurs more frequently at CpG sites, where a guanine nucleotide follows a cytosine nucleotide in the sequence of the genome[2,3]. In some regions, the frequency of CpG sites is up to 10 times greater than the average. These regions are called CpG islands (CGIs)[4]. CpG islands have a GC percentage greater than 50% with at least 200 base pairs long. Generally speaking, CpG sites outside the CGIs are mostly methylated, whereas CpG sites within CGIs are mostly unmethylated[5]. This difference indicates that CGIs usually have distinguished patterns of methylation, which may be important in gene regulation or gene mutation[6,7].

DNA methylation has been found to have influences on the expression of gene and functional regulation of proteins[8,9]. According to recent studies[10-12], DNA methylation can affect the onset and progress of various cancers and complex diseases. There are more methylated promoters and suppressors found in abnormal cell lines[13]. The aberrance of DNA methylation is one of the typical features of cancers such as acute myeloid leukemia[14]. However, the mechanistic link between aberrance of DNA methylation and leukemia is not well understood. Recent studies investigated DNA methylation in various cancers such as breast cancer[15,16]. The results indicate that abnormal DNA methylation usually occurs at some specific genomic locations[17,18].

Recent advances in methylation sequencing technologies allow the identification of genome-wide methylated sites in DNA[19,20]. One way of profiling methylation patterns of DNA is via the use of bisulfite treatment of DNA followed by next-generation sequencing, which is known as bisulfite sequencing[21]. The current bisulfite sequencing methods include whole-genome bisulfite sequencing (WGBS)[22] and reduced representation bisulfite sequencing (RRBS)[23]. Comparing to WGBS, RRBS reduces the amount of sequencing by using representative fractions of the genome. Therefore, RRBS specifically profiles and analyzes the methylation pattern for the regions with a high CpG content[24].

Methods have been developed to predict the methylation states at CpG sites, windows or segments of a genome[25-28]. Most of the current methods for methylation prediction assume that the methylation states are binary classes, that is, a CpG site or a window is either methylated or un-methylated (methylation-resistant)[29]. However, some other methods classified the methylation level to multiple classes[30]. Among these methods, predictions were usually limited to specific regions such as CGIs[28,31]. Predictive features used by these methods included DNA composition[32], GC content[28], sequence patterns[33], and methylation state of neighboring region[30]. Recent methods also used pseudo nucleotide composition to predict the methylations sites of a genome[32,34]. The DNA composition and methylation state of sequential neighbors are the two most common features among these methods[30,33].

One of the features that has not been used in predicting DNA methylation is chromosome interaction. The Hi-C technique enables the investigation of both

intra- and inter-chromosomal contacts in a genome[35]. The analysis of the genome at 1-1000 kilo-base resolution captures the overall genome spatial conformational arrangements. The 1 kilo-base resolution would further capture the contacts between the genes within the genome[36]. The Hi-C experiments cut the crosslink DNA with restriction enzyme and ligate them under extremely dilute conditions that favor intermolecular ligation. The experiments then purify and shear the ligated DNA segments to obtain paired-end reads. The paired-end Hi-C reads are mapped to the reference genome. After mapping, the data are binned and normalized into the Hi-C contacts library, which indicates that certain positions are spatially close in the three-dimensional space.

Although many methods have been developed to predict the methylation state of specific regions, the prediction of the methylation state of CpG sites in the loci of long non-coding RNAs (lncRNA) has received little attention. LncRNA are transcripts of non-coding genes ranging from 200 bases to 100 kilo-bases (kb) [37], yet their potential activities in human diseases have not been significantly unveiled. Recent studies on gene expression indicate that lncRNA may function as the connector between DNA and specific chromatin remodeling activities[38], and the expression level of lncRNA usually is lower than the ones of protein-coding genes[39]. Furthermore, lncRNA expression might be a main factor in carcinogenesis[40]. The exact mechanism of how lncRNAs influence cancer is unknown, but abnormal lncRNA expression may be a factor causing cancer by affecting major genetic processes. I evaluated my methylation predictions on the CpG sites in lncRNA loci.

In this study, I applied a deep learning algorithm, stacked denoising autoencoders (SdA), to predict DNA methylation status of CpG sites. Different from traditional learning algorithms, the training of SdA contains two stages: an unsupervised pre-training stage using unlabeled training data and a supervised fine-tuning stage using labeled data (data with known target values). I used sequential features generated within a window of the genome and features generated from the three-dimensional topology of a genome indicated by the Hi-C experiment[41]. I did extensive tests of my method through several benchmarks. In the first benchmark (Benchmark 1), I included the methylation level of sequential neighboring regions as features, whereas in the second benchmark (Benchmark 2), I excluded this type of features to increase prediction difficulty. I also benchmarked the influences of unlabeled data in deep learning and the influences of the genome topological features on the prediction accuracy.

CHAPTER II

RESULTS

Overview

I built SVM and SdA models to predict binary DNA methylation status of CpG sites (methylated or unmethylated). I applied my predictive models on lymphoblastic cell lines (GM12878) and chronic myelogenous leukemia cell lines (K562) to compare the performance of predictions on the healthy and cancer cell lines. Two types of windows were defined to generate features from sequentially and topologically neighboring regions of the genome (details see Methods). To better understand the factors influencing the performance of SdA, I applied different amounts of unlabeled pre-training samples, hidden layers, numbers of denoising autoencoders in each layer, and pre-training and training epochs. I also tested the performance of predicting methylation states of the CpG sites in lncRNAs loci.

In Benchmark 1, I measured the performance of my predictors using the metrics accuracy (Acc), specificity (Sp), sensitivity (Se), Matthews's correlation coefficient (MCC), and Receiver Operating Characteristic (ROC) curve using leave-one-out cross-validation. For the blind test data, only test accuracy was applied to evaluate the performance of SdA and SVM models. Different window-A sizes of each target CpG site, from 500 to 1000 nt, were tested. The definition of window-A can be found in the Methods section.

Moreover, I conducted Benchmark 2 by eliminating the features containing the methylation state of sequential neighboring regions but only using features of

(1) the methylation level of three-dimensional (3D) neighboring regions and (2) sequential composition patterns of the DNA sequence. In order to find the impact of the Hi-C based (3D genome topology) features, I tested the performance of using features generated from randomly selected windows that do not have any Hi-C contact to the target region. Different “Hi-C ranges” (from 10K to 50K) were used to benchmark the impact of including different amounts of Hi-C window-B (definition see Methods) features. In addition, I performed a blind test by randomly combining the samples from chromosome 1 with the ones from chromosome 21.

Chromosome-wide Analysis of Methylation Patterns and Preparing Training and Testing data sets

The methylated and unmethylated samples were defined based on parameters α and β (details see Methods). In order to balance my training dataset, I examined the distribution of PercentMeth (explained in Methods) values from RRBS experiments on chromosomes 1, 2 and 3 for GM12878 and K562 (Figure 1). I found that the majority of CpG sites were either hyper-methylated or hypo-methylated. Specifically, for GM12878 on chromosomes 1, 2 and 3, 67.73% of CpG sites have methylation level < 0.1 (hypo-methylated), and 14.40% of CpG sites have methylation level > 0.9 (hyper-methylated). Similarly, for K562, 60.42% are hyper-methylated and 14.43% are hypo-methylated. Based on this analysis, in order to balance the number of samples in methylated and un-methylated classes for leave-one-out cross-validation and blind test, I set the threshold β to be 0.01 making about half (46.25%) of the samples labeled as un-

methylated. The threshold α was set accordingly to ensure the number of methylated samples was equal to the one of un-methylated samples. In this way, most of the samples were labeled into one of the binary classes, and no sample was labeled twice.

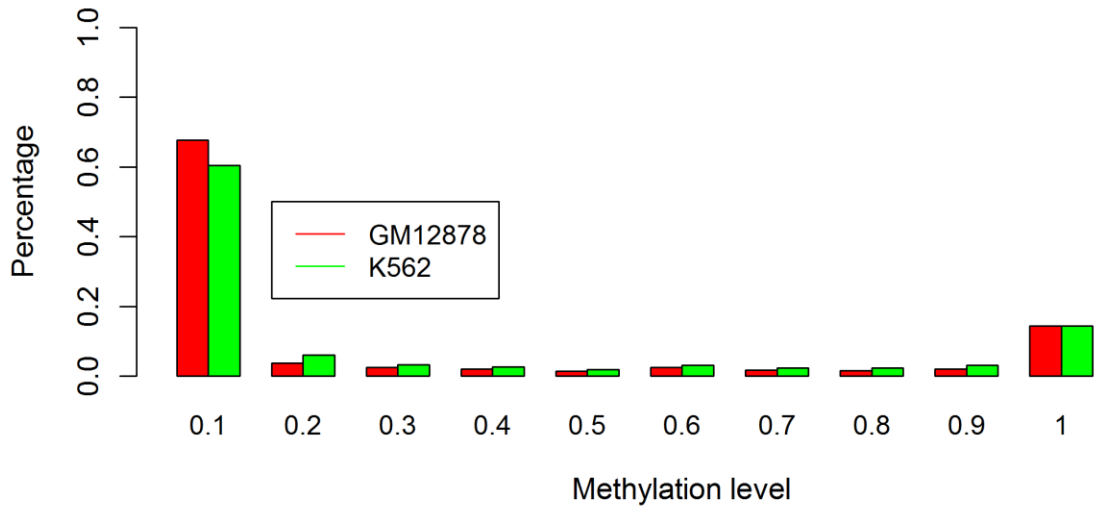

Figure 1. Distribution of DNA methylation levels on CpG sites for chromosomes 1, 2 and 3 for GM12878 and K562

Optimizing Stacked Denoising Autoencoders (Benchmark 1)

The parameters of stacked denoising autoencoders include number of hidden layers, number of hidden units in each layer, pre-training learning rate, number of pre-training epochs, fine-tuning learning rate, and the maximum of training epochs.

I optimized the parameters to obtain the best average performance on the individual test samples in every round of leave-one-out cross-validation (Table 1). In each round, one sample was chosen as the test sample, and the other samples were equally split into one training set and one validation set. The training set was also used for unsupervised pre-training of the SdAs. I found that

after the SdA architecture reached 23-500-500-2 (23 input unites, two hidden layers each with 500 hidden units, and two output nodes), the performance no longer changed when increasing the number of hidden layers and number of nodes in each layer. Therefore, the number of hidden layers was set to two, and the number of hidden units in each layer was set to 500 for the leave-one-out cross-validation and blind test. The setup of other parameters of SdA can be found in the Methods section.

Table 1

Performance of SdA for GM12878 on chromosome 21 under different numbers of hidden layers and different numbers of hidden units using leave-one-out cross-validation.

Number of hidden units and hidden layers	200	200-200	500	500-500	500-500-500
Accuracy	0.889	0.891	0.896	0.935	0.935

Leave-one-out Cross-validations for Support Vector Machine and Stacked Denoising Autoencoders (Benchmark 1)

In order to compare the performance of SVM, leave-one-out cross-validation was conducted on chromosomes 1 and 21 for GM12878 and K562 with different window-A sizes (Figure 2). The output of the SVM classifier is a

continuous number. Thus, I defined a cutoff μ to classify the output to discrete classes (for details see Methods).

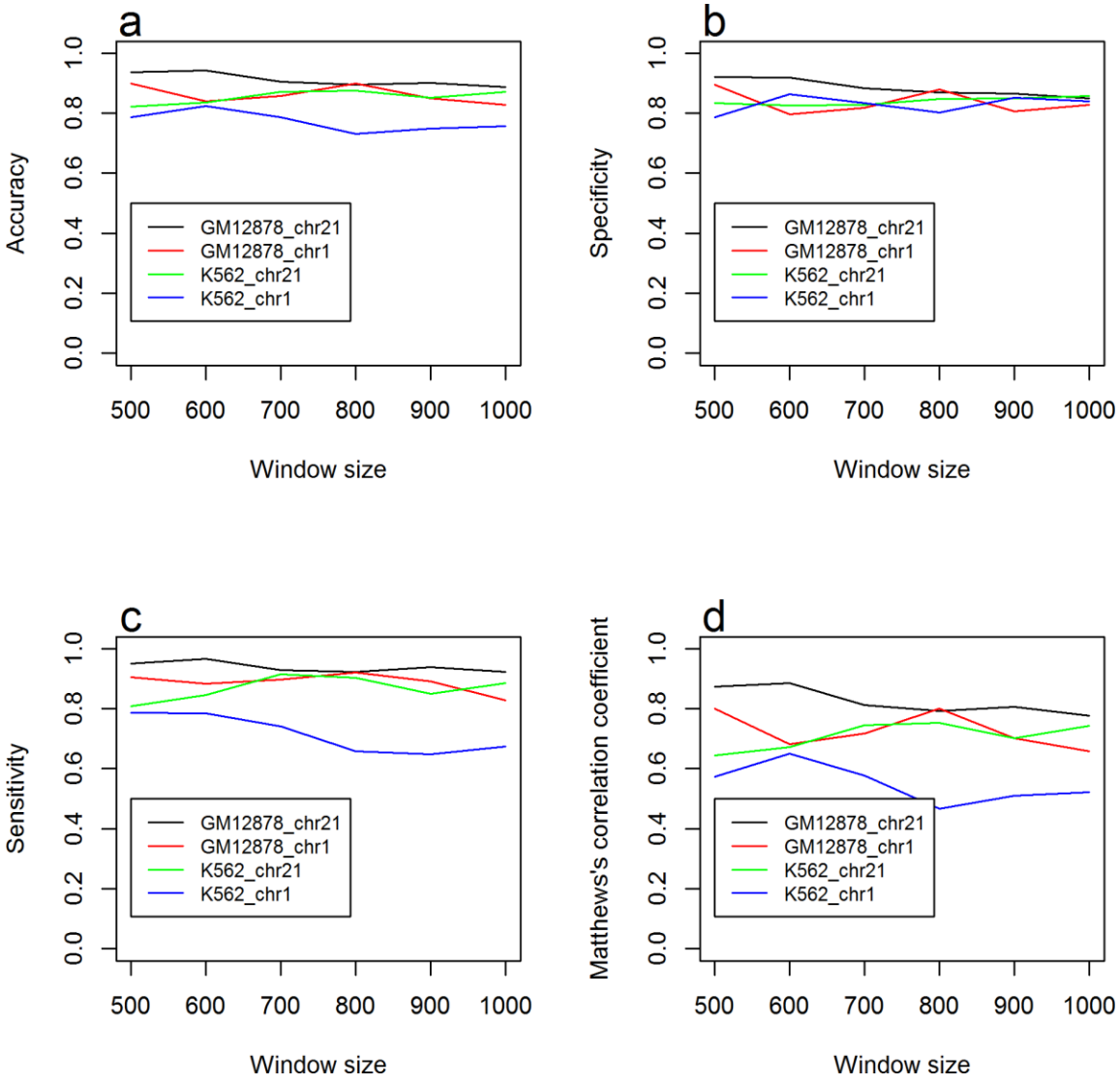

Figure 2. Leave-one-out cross-validation performance of SVM with different window sizes, chromosomes, and cell lines: (A) prediction accuracy, (B) specificity, (C) sensitivity, and (D) Matthews's correlation coefficient.

By the comparison of accuracies (Figure 2A) and Matthew's correlation coefficient (Figure 2D) of the two cell lines on chromosome 21 (black and green line) and chromosome 1 (red and blue line), I found that the performance for GM12878 is better overall than the performance for K562 on both chromosomes

1 and 21 on SVM model. One of the reasons may be that the number of samples for GM12878 is higher than K562 (Table 2), resulting from the different Hi-C coverages in the two cell lines. The average number of Hi-C reads for each nucleotide on GM12878 (chromosome 1 with 1.415 and chromosome 21 with 0.822) is higher than K562 (chromosome 1 with 0.371 and chromosome 21 with 0.206). In Benchmark 1, I included all the window-Bs that have at least one Hi-C contact with window-A. A higher Hi-C coverage results in more samples having at least one Hi-C contact and a higher number of window-Bs on average for each target CpG site.

Table 2

Number of samples used in leave-one-out cross-validation on chromosomes 1 and 21 for cell lines GM12878 and K562 with window size 600nt.

Cell Line	**Chromosome**	**Number of Samples**
GM12878	21	296
K562	21	230
GM12878	1	2616
K562	1	1988

For both K562 and GM12878, SVM achieves better performance on chromosome 21 than on chromosome 1 with most window sizes (Figure 2A, D). However, the average specificity of prediction for K562 chromosome 1 does not have a significant difference with prediction on chromosome 21 (Figure 2B). Together with the lower sensitivity on chromosome 1 (Figure 2C), it indicates that

the worse performance on chromosome 1 may be due to the worse performance on predicting true positive (methylated) samples.

The ROC curves were generated for chromosome 21. I calculated the values in the ROC curves by varying the cutoff μ from -2 to 2 (Figure 3). There is not a common window size to obtain the best performance for all cell lines. Figure 3 suggests that 600 nt is the best window size for GM12878 chromosome 21, which achieves an accuracy of 94.3%, Matthews's correlation coefficient of 0.886, specificity of 0.919, and sensitivity of 0.966 (based on Figure 2). For K562, 800 nt is the best window size, which achieves an accuracy of 87.6%, Matthews's correlation coefficient of 0.753, specificity of 0.848, and sensitivity of 0.904. Table 3 summarizes the best performance that SVM achieves and the corresponding window sizes.

Table 3

The best performance achieved from leave-one-out cross-validation using SVM on chromosomes 1 and 21 for cell lines GM12878 and K562. The threshold α was used to ensure equal number of samples in methylated class and un-methylated class.

Cell Line	Chromosome	α	Window Size	Acc	Sp	Se	MCC
GM12878	CHR1	0.55	500	0.900	0.894	0.905	0.800
GM12878	CHR21	0.99	600	0.942	0.918	0.966	0.886
K562	CHR1	0.07	600	0.823	0.863	0.784	0.649
K562	CHR21	0.43	800	0.876	0.848	0.904	0.753

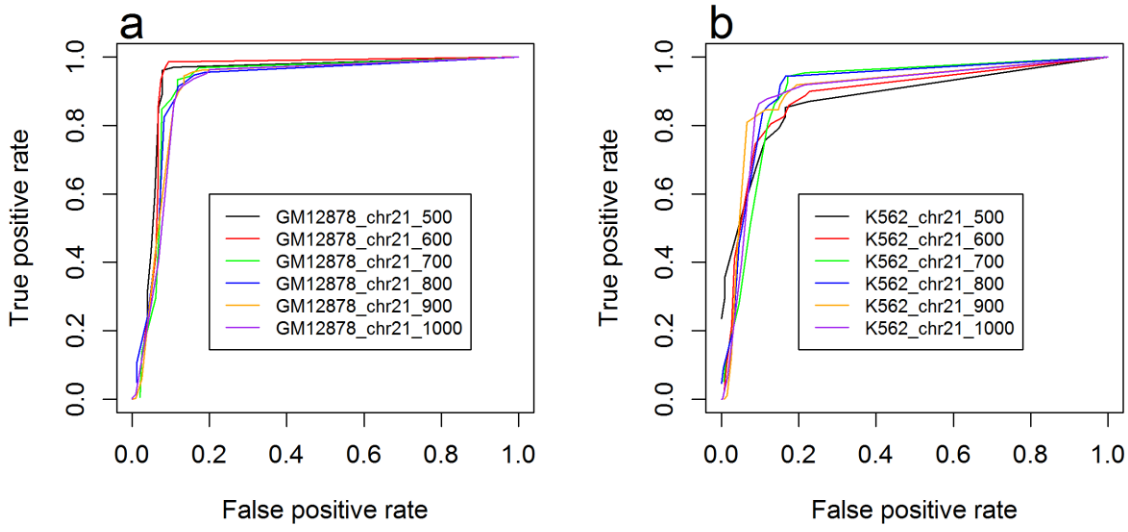

Figure 3. ROC curves of leave-one-out cross-validation using SVM with different window sizes for (A) GM12878 and (B) K562.

To compare the performance between SdA and SVM, I conducted leave-one-out cross-validations for SdA on chromosomes 1 and 21 with window size 600nt (Table 4). On chromosome 21, SdA obtained a worse performance for GM12878 with an accuracy of 0.935 compared to SVM's accuracy 0.943. However, on chromosome 1, the SdA model achieved a better performance with an accuracy of 0.885, which is higher than SVM's 0.839. Table 4 shows that the number of samples in chromosome 1 (2,616) is about six times higher than the ones for chromosome 21 (296), which may be one of the reasons for the performance difference. This indicates that the SdA algorithm may need more training samples to achieve better performance, whereas the SVM algorithm can achieve a decent performance with a much smaller size of training data.

Table 4

Performance of leave-one-out cross-validation using SVM and SdA on chromosomes 1 and 21 for cell line GM12878 on the window size 600nt.

Classifier	Cell Line	Chromosome	Acc	Number of Samples
SdA	GM12878	CHR21	0.934	296
SVM	GM12878	CHR21	0.942	296
SdA	GM12878	CHR1	0.885	2616
SVM	GM12878	CHR1	0.839	2616

Evaluating SVM and Stacked Denoising Autoencoders on Blind Test Data Sets (Benchmark 1)

I further evaluated the performance of SVM and SdA using two blind test data sets. The predictive models were trained using CpG sites on chromosomes 1, 2, and 3 with different window sizes for both healthy (GM12878) and cancer (K562) cell lines. I used CpG sites on chromosomes 1, 2, and 3 as the training set because chromosomes 1, 2, and 3 are the largest three chromosomes in humans. The numbers of training samples associated with various window sizes are shown in Figure 4. All features for the SVM and SdA classifiers were generated from the same dataset. The CpG sites on chromosomes 21 and X were selected as two independent test sets, considering that chromosome 21 is a smaller chromosome and chromosome X can be inactivated by the lncRNAs called Xist for female[42]. It would be interesting to study the methylation pattern in X chromosome and compare it with chromosome 21.

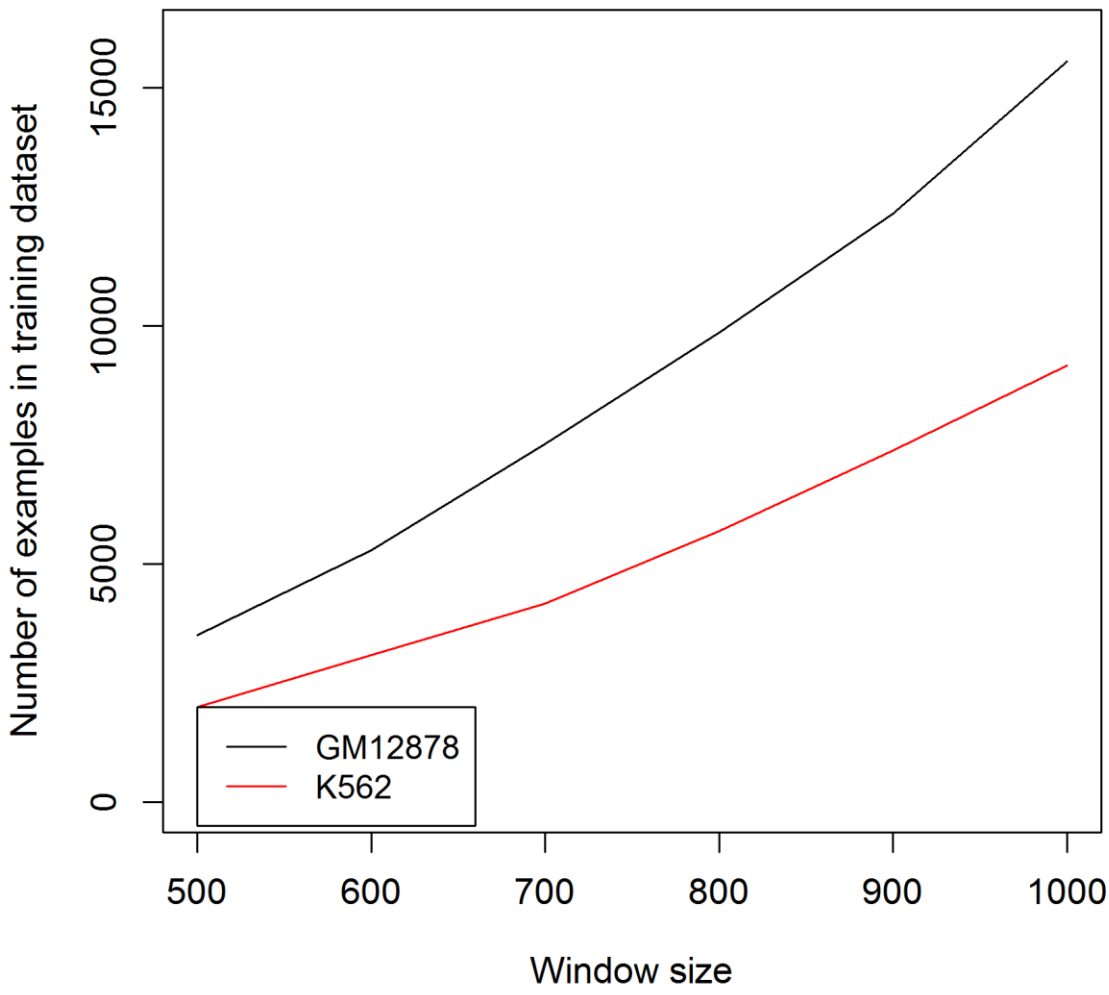

Figure 4. Number of training samples generated from chromosomes 1, 2 and 3.

Comparing to chromosome X, predictions on chromosome 21 for GM12878 achieved a better performance on most window sizes (Figures 5 and 6). The difference may be due to the chromosome-specific methylation patterns. I explored the distribution of the methylation level on chromosomes 21 and X (Figures 7 and 8), which suggests that for both GM12878 and K562, methylation distributions on chromosome 21 share similar patterns with the distribution on chromosomes 1, 2, and 3 (Figure 1), which were used as the training data. For example, on chromosome 21, 51.98% of the CpG sites have a methylation level

< 0.1, and 15.44% of the CpG sites have methylation level >0.9 (Figure 7), which is similar to 67.73% and 14.40% in chromosomes 1, 2, and 3 (Figure 1), respectively. However, for GM12878 on chromosome X, CpG sites with methylation level < 0.1 take a much lower proportion, that is, 33.24% (Figure 8), which indicates that the methylation distribution in chromosome X is significantly different from the distribution in chromosomes 1, 2, 3 (training data set), and 21 (the other test data set).

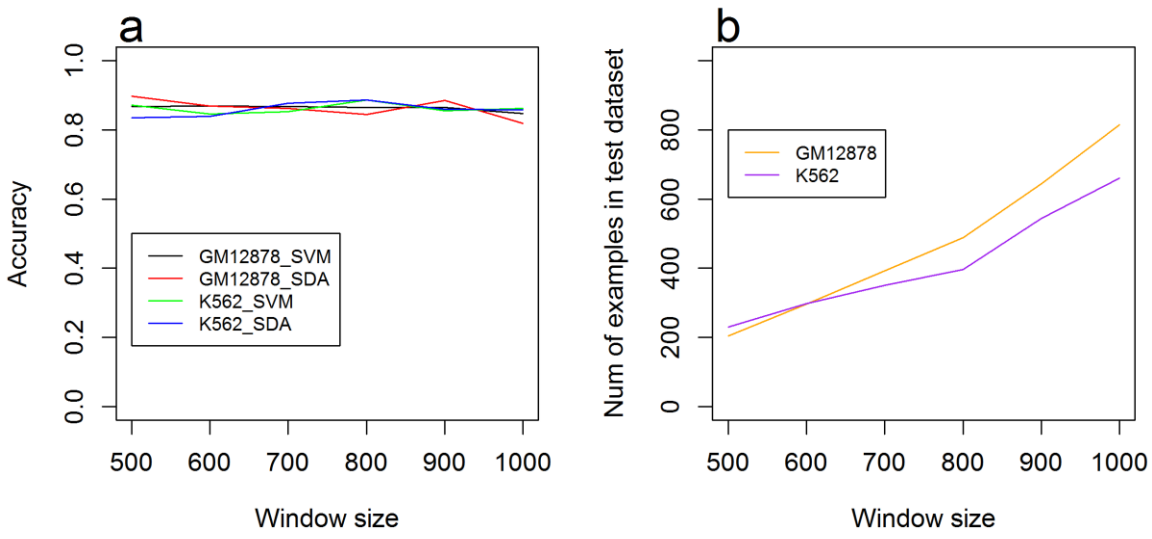

Figure 5. (A) Accuracy of blind test on chromosome 21 using SdA and SVM. (B) Number of samples in the test dataset with different window sizes in chromosome 21.

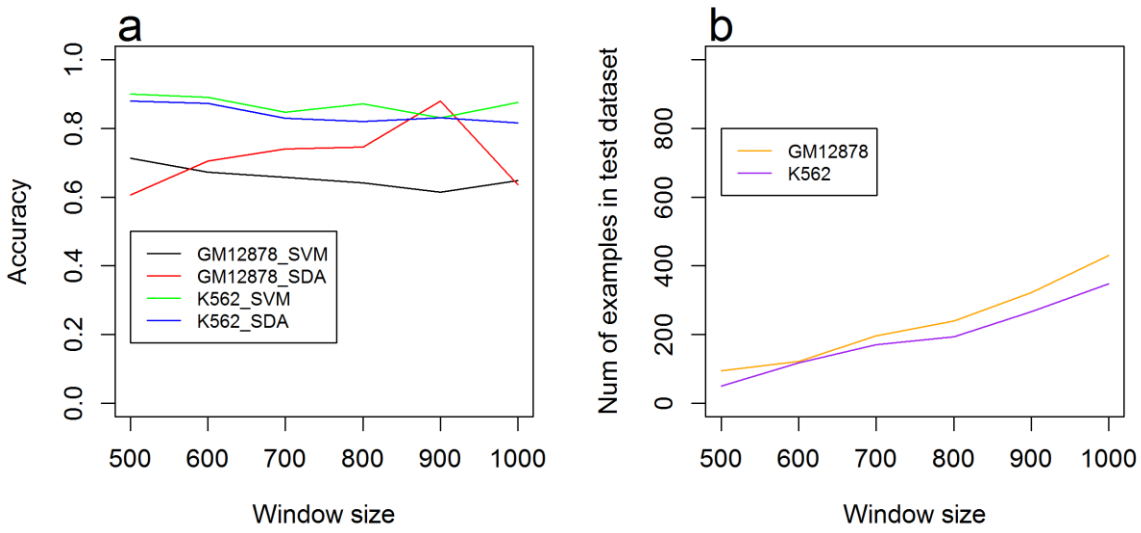

Figure 6. (A) Accuracy of blind test on chromosome X using SdA and SVM. (B) Number of samples in the test dataset with different window sizes in chromosome X.

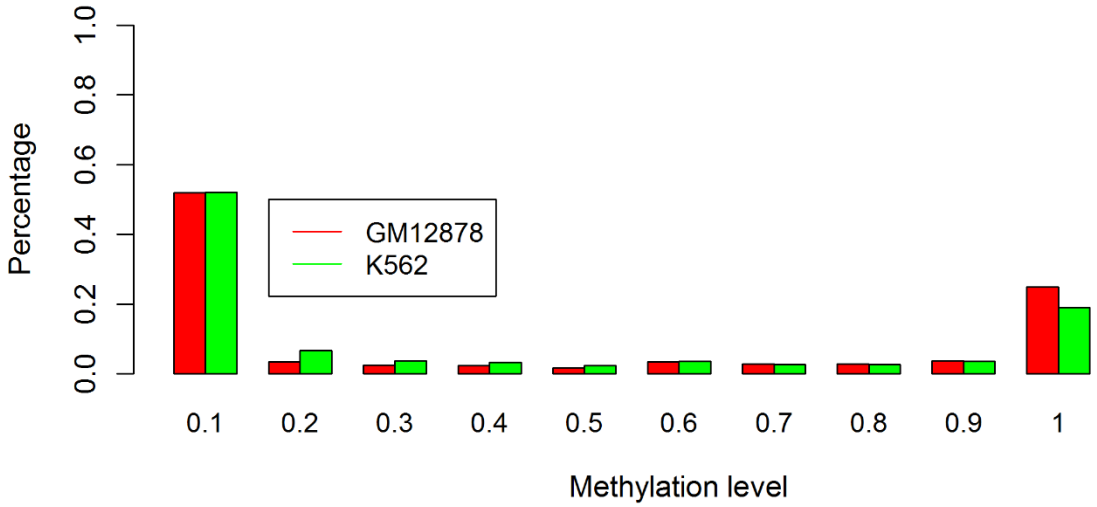

Figure 7. DNA methylation level distribution on chromosome 21 for GM12878 and K562

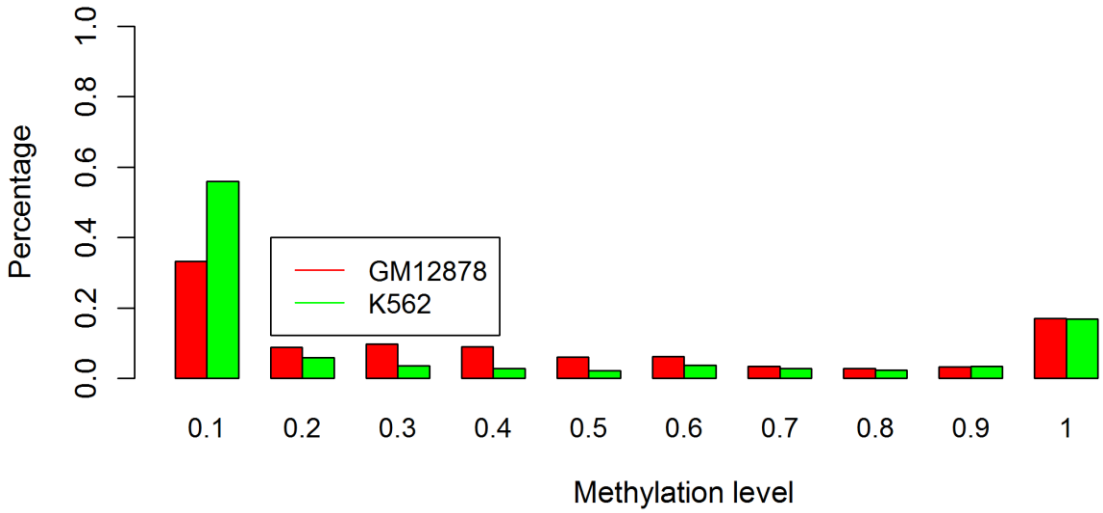

Figure 8. DNA methylation level distribution on chromosome X for GM12878 and K562.

Predicting Methylation State of lncRNA loci (Benchmark 1)

I investigated DNA methylation prediction for CpGs sites located within lncRNAs genes. I used the same training data set, which is the combination of CpG sites on chromosomes 1, 2, 3, and the test data set that contains the CpG sites within lncRNAs genes on chromosomes 21 and X.

Benchmarking on chromosome 21 lncRNA loci: for both GM12878 and K562 on chromosome 21, predictions for CpG sites within lncRNAs (Figure 9) achieved better performance than the ones without region-specific limitation (that is, both CpG sites within lncRNA genes and outsides lncRNA genes) (Figure 5). Specifically, for K562, SdA reached the best accuracy of 0.977, while the best accuracy is 0.886 for predictions on all CpG sites. This improvement in accuracy may be because the methylation distribution patterns of chromosome 21 lncRNAs are more similar to the training dataset (chromosomes 1, 2 and 3) as compared to the ones of all CpG sites. Thus, I explored the methylation patterns

of CpG sites within lncRNA on chromosome 21 (Figure 10). I found that 60.42% of the CpG sites within lncRNAs had a methylation level < 0.1 (Figure 10), which is closer to the training dataset's 67.73% (Figure 1) than 51.98% in CpG sites without region-specific limitation (Figure 7).

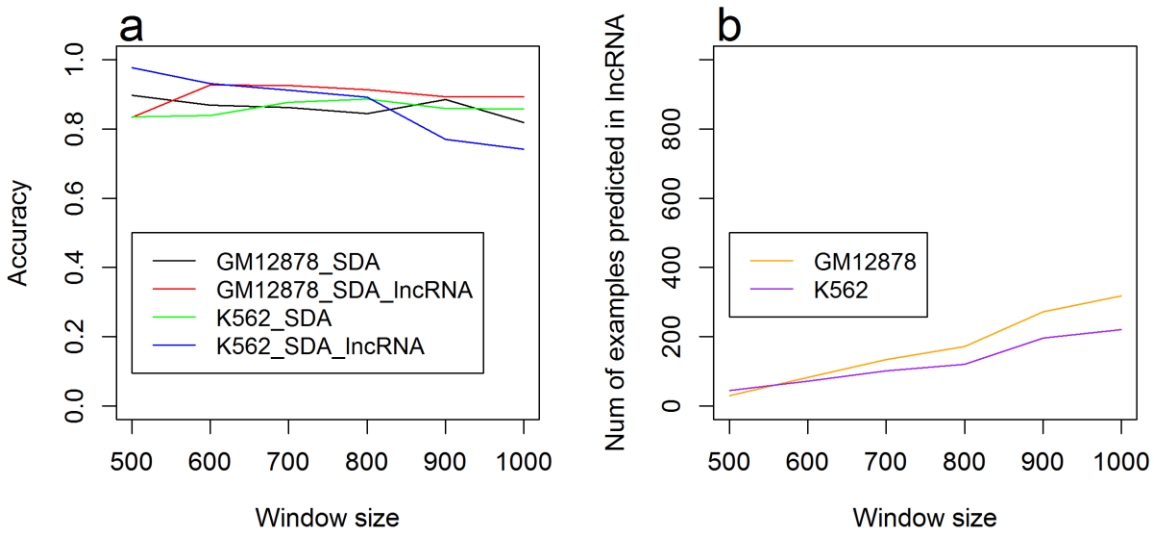

Figure 9. (A) Performance of SdA for the prediction of methylation for lncRNAs and CpG sites without region-specific limitation on chromosome 21. (B) Number of samples in the test dataset on different window sizes in chromosome 21.

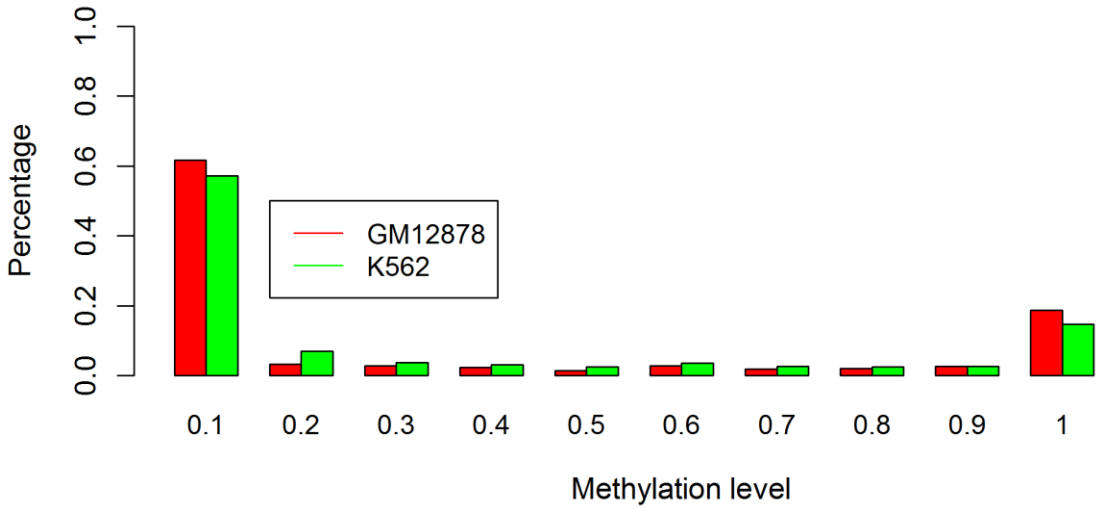

Figure 10. DNA methylation level distribution of CpG sites within lncRNA on chromosome 21 for GM12878 and K562.

Benchmarking on chromosome X lncRNA loci: furthermore, I found that the performance for lncRNAs genes of GM12878 on chromosome X is worse than the one on chromosome 21 (Figures 11 and 9). The difference of performance for GM12878 on chromosome 21 and X may result from the different characteristics of methylation for chromosomes 21 and X. Therefore, I explored the distributions of the methylation levels of lncRNAs for chromosomes 21 and X (Figures 10 and 12). It can be found that for GM12878, the methylation distribution of chromosome 21 lncRNAs shares similar patterns with the methylation levels of all CpG sites (not only lncRNAs) in chromosomes 1, 2, and 3, which were used as the training data. Specifically, on chromosome 21, there are 61.57% of lncRNA CpG sites have a methylation level < 0.1 (Figure 10), which is similar to 67.73% on chromosomes 1, 2 and 3 (Figure 1). In contrast, on chromosome X, only 31.65% of the lncRNA CpG sites have a methylation level < 0.1 (Figure 12), which indicates that the methylation distribution for lncRNA on

chromosome X is quite different from the training data set comprised of chromosomes 1, 2 and 3.

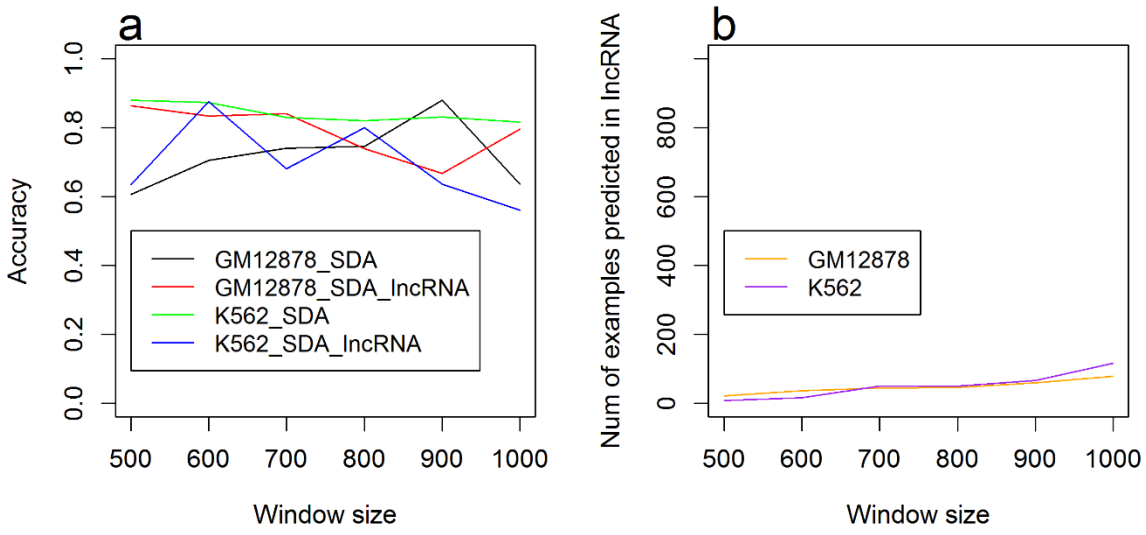

Figure 11. (A) Performance of SdA for the prediction of methylation for lncRNAs and CpG sites without region-specific limitation on chromosome X. (B) Number of samples in the test dataset on different window sizes in chromosome X.

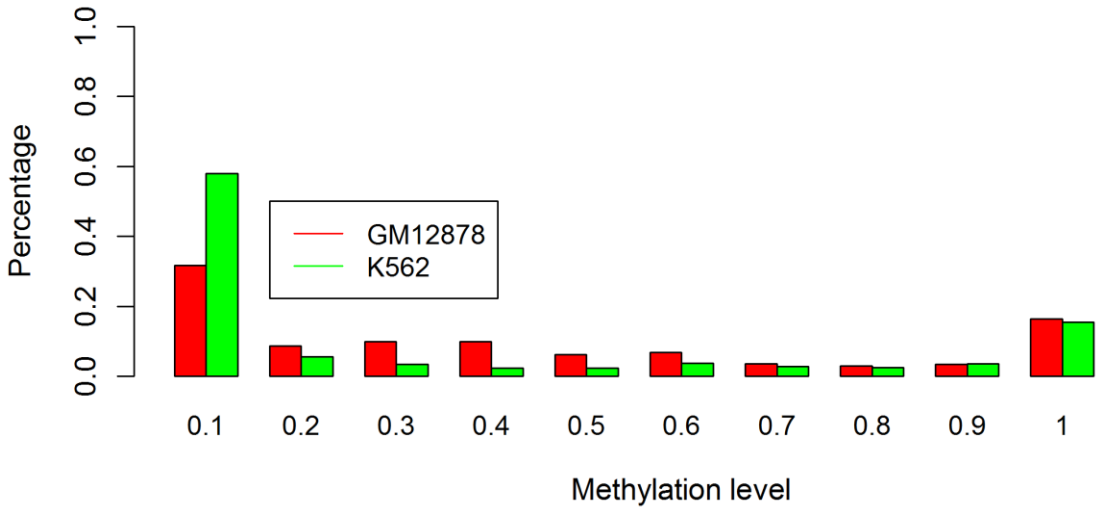

Figure 12. DNA methylation level distribution of CpG sites within lncRNA on chromosome X for GM12878 and K562.

Moreover, since both K562 and GM12878 are samples from female, it is possible that the X chromosome may be inactivated or in the process of inactivation by an lncRNA called Xist[42] that packs the three-dimensional structure of X chromosome to disable the expressions of most X-chromosome genes. The change of three-dimensional genome structure of X chromosome influences the genome structural based features used in my methods and may also alter the DNA methylation patterns in chromosome X. Moreover, another reason may be that the test dataset of X chromosome is relatively small, compared to the one of chromosome 21. Therefore, the influence of error becomes more significant (Figure 11 B).

The Impact of Hi-C Based Genome Topological Features (Benchmark 2)

Benchmark 1 used the methylation level of sequential neighboring region of a target CpG site. In Benchmark 2, I eliminated that feature in order to benchmark the performance only based on the sequence composition of window-A and window-B (three-dimensional topological neighboring regions) in addition to methylation levels in window-B. Compared to Benchmark 1, I added 74 PseTNC features for window-A and eight features for window B. All of these newly added features indicate sequence composition. In Benchmark 2, I used both up-sampling and down-sampling to balance training data. Details can be found in the Methods section.

I also changed Hi-C based window-B to a randomly generated window-B in order to observe the impact of Hi-C inferred topological neighbors. I only used the randomly generated windows that do not have any Hi-C contact with the Hi-C

ranges (the region surrounding a target CpG site that Hi-C neighbors were collected from, see the Methods section for details). In this way, I eliminated topological neighbors from the random windows. Different sizes of Hi-C ranges were tested.

Table 5 shows the 5-fold cross-validation accuracy and MCC scores of SVM on using both Hi-C based window-Bs and randomly selected windows. The performance on Hi-C based and random windows is similar in this case with random windows performing slightly worse. Tables 6-11 show the performance of SdAs. I benchmarked one, two, and three hidden layer(s) and found that more hidden layers result in significantly worse performance for randomly selected windows.

Table 5

The SVM's 5-fold cross-validation accuracy and MCC scores of using Hi-C based topological neighboring window-Bs and random window-Bs on chromosome 1 with different Hi-C ranges.

Hi-C range	Acc (Hi-C based)	Acc (random)	MCC (Hi-C based)	MCC random
10K	0.831	0.828	0.616	0.600
20K	0.833	0.810	0.618	0.584
30K	0.830	0.815	0.614	0.586
40K	0.837	0.832	0.623	0.606
50K	0.838	0.824	0.628	0.601

Table 6

The 5-fold cross-validation accuracies of SdA on chromosome 1 with different Hi-C ranges.

Hi-C range	Hi-C_1L	Random_1L	Hi-C_2L	Random_2L	Hi-C_3L	Random_3L
10K	0.829	0.830	0.837	0.714	0.835	0.406
20K	0.839	0.839	0.828	0.668	0.829	0.376
30K	0.840	0.835	0.832	0.823	0.830	0.376
40K	0.828	0.835	0.831	0.831	0.828	0.565
50K	0.841	0.819	0.826	0.828	0.834	0.326

Note. The SdA model was trained with 10 pre-training epochs (unsupervised learning, learning rate 0.01) and 100 fine-tuning epochs (supervised learning, learning rate 0.01). The 1L, 2L and 3L are the number of hidden layers with corruption levels of all layers set to 0.1. All the layers have 100 hidden nodes. Features based on genome topological neighbors (window-Bs, indicated as “Hi-C” in the table) and features based on randomly selected regions (random windows, indicated as “Random” in the table) were used to benchmark the impact of Hi-C based features.

Table 7

The MCC scores for the same set up as in Table 6

Hi-C range	Hi-C_1L	Random_1L	Hi-C_2L	Random_2L	Hi-C_3L	Random_3L
10K	0.611	0.600	0.628	0.549	0.627	0.040
20K	0.626	0.624	0.613	0.372	0.614	0.032
30K	0.627	0.619	0.620	0.604	0.614	0.018
40K	0.612	0.619	0.623	0.614	0.615	0.265
50K	0.635	0.602	0.617	0.615	0.623	0.050

Table 8

The accuracy of the same SdA architectures as in Table 6 with pre-training epochs set to 10 and training epochs set to 10.

Hi-C range	Hi-C_1L	Random_1L	Hi-C_2L	Random_2L	Hi-C_3L	Random_3L
10K	0.794	0.795	0.777	0.6545	0.765	0.692
20K	0.787	0.776	0.768	0.7117	0.796	0.682
30K	0.793	0.773	0.768	0.766	0.798	0.712
40K	0.795	0.778	0.778	0.7668	0.777	0.699
50K	0.792	0.770	0.778	0.7807	0.789	0.777

Table 9

The MCC of the same configuration as in Table 8.

Hi-C range	Hi-C_1L	Random_1L	Hi-C_2L	Random_2L	Hi-C_3L	Random_3L
10K	0.530	0.439	0.501	0.040	0.474	0.008
20K	0.525	0.495	0.488	0.056	0.324	0.036
30K	0.531	0.505	0.480	0.425	0.357	0.008
40K	0.528	0.501	0.501	0.426	0.000	0.066
50K	0.534	0.499	0.510	0.376	0.196	0.000

Table 10

The accuracy of the same SdA architectures as in Table 6 with pre-training epochs set to 100 and training epochs set to 10.

Hi-C range	Hi-C_1L	Random_1L	Hi-C_2L	Random_2L	Hi-C_3L	Random_3L
10K	0.773	0.690	0.763	0.639	0.765	0.642
20K	0.801	0.699	0.768	0.641	0.776	0.759
30K	0.791	0.67	0.764	0.715	0.771	0.662
40K	0.777	0.699	0.765	0.643	0.759	0.649
50K	0.795	0.739	0.795	0.634	0.763	0.625

Table 11

The MCC scores for the same configurations in Table 10.

Hi-C range	Hi-C_1L	Random_1L	Hi-C_2L	Random_2L	Hi-C_3L	Random_3L
10K	0.484	0.003	0.493	-0.003	0.495	-0.04
20K	0.375	0.044	0.508	-0.001	0.519	0.011
30K	0.392	0.026	0.502	0.017	0.507	0.015
40K	0.103	0.091	0.503	0.009	0.484	0.008
50K	0.276	0.027	0.499	-0.003	0.494	-0.005

In order to benchmark the influence of unsupervised pre-training of SdAs, I conducted three independent 5-fold cross-validations, in which the epochs of unsupervised pre-training and supervised training were set to (10, 100) (Tables 6 and 7), (10, 10) (Tables 8 and 9), and (100, 10) (Tables 10 and 11) while unifying all the other factors, including pre-training, training, validation, and testing data and other SdA parameters. The results show that larger epochs for unsupervised pre-training and smaller epochs of supervised training may decrease performance and make the SdAs perform significantly worse for random windows. The epochs of unsupervised pre-training and supervised training of (10, 100) generated the best performance.

Blind Test on Chromosome 21 and LncRNA Loci (Benchmark 2)

I tested the performance on randomly combined samples from chromosomes 1 and 21 (details of data generation see Methods). For SdAs, the

epochs of unsupervised pre-training and supervised training was set to (10, 100), and the other parameters remained the same as the 5-fold cross-validation that generated the best results. Similar findings as in the 5-fold cross-validation were observed. That is, a higher number of hidden layers makes the SdA perform significantly worse on the random windows in GM12878 (Table 12). For K562, the SdA model achieved an accuracy of 72.01%. I also benchmarked the performance on the CpG sites without genome topological features (no Hi-C signals); and in this case, the accuracies of GM12878 and K562 are 84.25% and 69.95%, respectively.

Table 12

The blind test accuracy and MCC scores for SdA and SVM on randomly combined training and testing samples from chromosomes 1 and 21 with Hi-C range 10K.

Classifier	Features	SdA architecture	Acc	MCC
SdA	Hi-C based window-B	109-100-2	0.871	0.666
SdA	Random window-B	109-100-2	0.810	0.612
SdA	Hi-C based window-B	109-100-100-2	0.867	0.659
SdA	Random window-B	109-100-100-2	0.631	0.058
SVM	Hi-C based window-B	NA	0.860	0.685
SVM	Random window-B	NA	0.858	0.725

Note. The ration of fine-tuning, validation, and testing samples for SdA is 3:1:1. With two hidden layers, the MCC score of SdA is 0.058. I found that the predictions are highly biased to negative samples. This causes the false negative to be a value close to 1. Therefore, it has a very low MCC score.

Using the optimized SdA configuration and SVM model found in the 5-fold cross-validations, I tested their performance on GM12878 chromosome 21 lncRNA loci (Table 13). Two hidden layers of SdA generated the best testing accuracy that is similar to SVM. In terms of MCC score, two-hidden-layer SdA (0.6427) performed slightly better than SVM (0.6385).

Table 13

Performance of SdA and SVM for predicting methylation level of CpG sites within lncRNA regions.

Classifier	SdA architecture	Acc	MCC	Number of test samples
SdA	109-100-2	0.796	0.5678	2138 (551 positive, 1587 negative)
SdA	109-100-100-2	0.784	0.5617	2138
SdA	109-100-100-100-2	0.832	0.6427	2138
SVM	NA	0.837	0.6385	2138

Note. The SdA architecture and SVM model used were the one with the best test accuracy in 5-fold cross-validation on chromosome 1 (Tables 6-11). The number of testing lncRNA samples are 2,138 (551 positive and 1587 negative).

Benchmarking the Parallel Algorithm for Generating Features and Training SVMs

A parallel algorithm was used to reduce the execution time of the entire feature generation process. The parallel algorithm was implemented using C++ and MPICH, and the performance tests were conducted on my own shared memory server equipped with 48 CPUs with speed 1200 MHz and 126 gigabytes

of memory. A test result is given in Table 14, which shows that my parallel method dramatically saves computational time.

Table 14

Execution time (seconds) and corresponding Speedup (time of using one process divided by the time using x processors, x = 2, 4, 6, and 16) on chromosome 21 of K562.

Number of Processes		P=1	P=2	P=4	P=8	P=16
Chr21	Time	703.05	370.69	201.41	99.97	56.58
	Speedup	-	1.89	3.49	7.03	12.43

CHAPTER III

DISCUSSIONS

I developed SVM and SdA models to predict binary methylation state of CpG sites on GM12878 and K562 on different chromosomes with different window sizes. In the leave-one-out cross-validation for SVM classifier, the accuracy reaches 0.943 on chromosome 21 of GM12878, while the accuracy reaches 0.876 on chromosome 21 of K562. The distinction of performance between GM12878 and K562 on the SVM model may result from the different numbers of samples and Hi-C coverage. This indicates that the Hi-C reads coverage plays an important role, as a higher Hi-C coverage can increase the resolution of the three-dimensional genome structure and provide more neighboring CpG sites as features for the machine learning models.

Furthermore, I evaluated the SdA classifier using a leave-one-out cross-validation. For SdA classifier tested on 296 CpG sites of GM12878 chromosome 21, the accuracy reaches 0.935, which is slightly lower than SVM classifier's 0.943. However, on chromosome 1 for GM12878, in which the total number of leave-one-out samples reaches 6,516, the accuracy of SdA classifier reaches 0.885, which is obviously higher than SVM classifier's 0.839. The difference of performance between SVM and SdA may suggest that the SdA algorithm needs more training samples to achieve better performance. Moreover, by comparing the performance with features excluding methylation level of neighbors and GC contents, I found that, especially for SdA, neighboring methylation levels and GC content are influential to the prediction performance.

Moreover, I evaluated the performance of SVM and SdA classifiers using two blind test sets. My experiments used chromosomes 1, 2 and 3 as the training set, and chromosomes 21 and X as two independent test data sets. SdA reaches the best accuracy of 0.897 on chromosome 21 of GM12878 with window size 500nt. On chromosome 21, both SVM and SdA have a stable performance over different window sizes for both K562 and GM12878. For chromosome X, SdA achieved a best accuracy of 0.880 for GM12878 with window size 900 nt. Overall, the accuracies of GM12878 on chromosome X are lower than the ones on chromosome 21 for most window sizes. This may be because the distributions on chromosome X are largely different from the distributions in the training dataset of chromosomes 1, 2, and 3.

I investigated the performance of predicting the DNA methylation state for CpG sites within lncRNA DNA locus. The best accuracy, 0.977, was obtained when using SdA on chromosome 21 of K562 with window size 500nt. I further found that the performance on chromosome X was overall worse than the performance on chromosome 21. By analysis, I found that the methylation distribution of lncRNA genes in chromosome X of GM12878 was largely different from the distributions found in both chromosome 21 and the training chromosomes 1, 2, and 3. This may result from the existence of an lncRNA called Xist that packs and inactivates the chromosome X of female causing the different methylation patterns. My data indicates methylation patterns of lncRNA may be chromosome- and cell line-specific.

In order to benchmark the influence of Hi-C based genome topological features, I replaced Hi-C neighbors with randomly selected windows and then benchmarked the performance. I found that using random windows significantly decreased the performance of SdAs with two or more hidden layers. I also tested it with different numbers of epochs for pre-training and fine-tuning and found that a larger number of fine-tuning increases performance, whereas a larger number of pre-training decreases the performance.

CHAPTER IV

METHODS

Datasets

Human cell lines: the cell lines GM12878 and K562 were selected for my study because of their accessibility and sufficient experimental data associated with them. GM12878 is a B-lymphocyte cell line from a female, while K562 is an immortalised cell line from a female patient with chronic myelogenous leukemia (CML) (for description of these two cell lines see http://www.genome.gov/26524238). Thus, investigating the methylation prediction on these two cell lines may help me characterize the methylation patterns of cancer and healthy cell lines.

DNA methylation data: DNA methylation state at each CpG dinucleotide is measured by Reduced Representation Bisulfite Sequencing (RRBS) data. RRBS methylation data for cell lines GM12878 and K562 were obtained from the ENCODE project (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibMethylRrbs/).

Genome topology: the Hi-C paired reads[36] for GM12878 and K562 cell lines were obtained from the public accessible NCBI GEO database (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1181867) and NCBI SRA database (accessible at http://sra.dnanexus.com/experiments/SRX011614/runs), respectively. The paired-end Hi-C reads were mapped to the human reference genome (UCSC version

hg19) using the read sequence alignment tool Maq[43]. The contact library containing spatial contacts between pairs of genomic positions were generated by parsing the Maq mapping outputs. Each contact between two positions on genome implies that they are spatially proximate in three-dimensional structure.

Support Vector Machines (SVM)

There are four types of kernel functions in SVM-Light[44]: linear, polynomial, radial basis function, and sigmoid. In Benchmark 1, I selected the polynomial kernel function for my SVM classification model because this kernel function achieves the best performance based on the 23 features using leave-one-out cross-validation (data not shown). Based on the optimization of SVM model, the parameter C (trade-off between training error and margin) was set to 5, and the polynomial kernel function parameter d was set to 3. In Benchmark 2, the radical basis function was selected as the kernel function based on the cross-validations on 109 features. I used the default value of parameter C in SVM-light and set the parameter gamma in radical basis function to 9 based on optimization.

Deep Learning - Stacked Denoising Autoencoder

The deep learning architecture applied to this research is Stacked Denoising Autoencoder (SdA)[45] implemented with Theano (http://deeplearning.net/software/theano/). Theano is a Python-based library enabled GPU-based high performance computing for deep networks. The SdA algorithm composed of two phases of learning. The first phase is unsupervised pre-training carried out by layers of denoising autoencoders, which learn a

reconstruction *Z* from corrupted version of data *X* by minimizing the cross-entropy of the reconstruction:

$$L_H(X,Z) = -\sum_{k=1}^{d}[X_k \, log \, Z_k + (1 - X_k) log(1 - Z_k)] \quad (1)$$

for all the training samples in a minibatch. The *Z*, which is the reconstruction of the corrupted version of data *X*, was computed from

$$Z = S(W'y + b') \quad (2)$$

where W' is the reconstruction weighting matrix, b' is the reconstruction bias, and function $S()$ is a sigmoid function:

$$S(t) = \frac{1}{1+ e^{-t}} \quad (3)$$

Also, Z, the reconstruction of the corrupted input *X*, can be considered as the prediction of *X* because it tries to have the same shape of *X* given *y*, where

$$y = s(Wx + b) \quad (4)$$

in which $s()$ is a sigmoid function, *W* is the weighting matrix, and *b* is the bias. Formula 4 maps the corrupted input $X \, \epsilon \, [0,1]^d$ to a hidden representation $y \, \epsilon \, [0,1]^{d'}$, which is reversely mapped by Formula (2) to build a reconstruction of corrupted data X by minimizing Formula (1). The corrupted input *X* is a sparse version of the original input *X-orig*. There are multiple ways to generate *X* from *X-orig*, and I used a parameter called corruption level to set it. The hidden units in a hidden layer were randomly selected to be disabled from an input node based on the probability set by the corruption level parameter. This corrupted version of autoencoders does not only learn the identifiers of the input data but also learns the features that are more useful to the problem. Therefore,

it is also named denoising autoencoders[46]. This corruption process was applied in each layer of hidden units in stacked denoising autoencoders.

The learning process computes the cost based on Formula (1) for each layer of stacked denoising autoencoders and updates the weights and biases by gradient descent. The training process starts from the first layer directly connecting to the input data, and continues layer by layer. The trained *m* layers enable the computation of latent representation in layer *m+1*. In this way, all stacked layers of denoising autoencoders were trained, and the outputs from these layers of denoising autoencoders are the reconstruction of input *X-orig* or the features selected from the original data. With that, the unsupervised pre-tuning part is finished.

A supervised fine-tuning is applied after the unsupervised pre-tuning. A logistic regression model was added on top of the layers of denoising autoencoders that calculates:

$$\mathrm{P(Y = i|x, W, b) = softmax_i(Wx + b) = \frac{e^{W_i x + b_i}}{\sum_j e^{W_j x + b_j}}} \quad . \tag{5}$$

This formula calculates the probability of an input vector *x* having the class *i* of *Y*. W is the weighting matrix; b is the bias; and j can be all the possible classes in *Y*. After calculating the probabilities for all possible classes in *Y*, an input vector x is assigned or predicted to the class that gives highest probability as:

$$y_{pred} = argmax_i P(Y = i \,|x, W, b) \; . \tag{6}$$

A multilayer perceptron is constructed that shares the same number of layers, number of neurons in each layer, weight, and bias as previously trained

stacked denoising autoencoders. Label value *Y* was used to train the multilayer perceptron by a backpropagation algorithm with logistic function as activation function. In this way, the multilayer perceptron was trained, and the entire learning architecture was fine-tuned. Therefore, the weights and bias in each hidden layer of the deep network were updated again based on the class label *y* of each training sample.

Based on my Benchmark 1 (see Results), the best configuration achieving optimal performance contains two hidden layers each with 500 hidden units; the pre-training learning rate and epochs were set as 0.01 and 100; the fine-tuning learning rate was set as 0.1; and the maximum of training epochs was set as 1000. In Benchmark 2, the learning rates of pre-training and fine-tuning were set to 0.01; corruption level was set to 0.1 for all hidden layers. Different epochs for pre-training and fine-tuning were tested (see Results). The SdA algorithm was implemented on a NVIDIA Quadro K5100 GPU with 1,536 CUDA parallel processor cores.

Machine Learning Features

Overview: I defined two types of windows for each CpG site to generate features. The first type of window, window-A, is a DNA sequence window with the target CpG site as the center whose size varies from 500 to 1000nt. Window-A was used to generate features from the sequence that are immediately adjacent to the target CpG sites. The second type of window, window-B, is a sequence window with point X in the center, whereas point X and a point in window-A (for Benchmark 1) or “Hi-C range” (for Benchmark 2, definition see below) must be in

contact indicated by a Hi-C paired ends read. The coordinates of CpG sites and corresponding window sequences were determined based on human reference genome hg19.

Features from window-A: there are four types of DNA nucleotides: adenine (A), thymine (T), guanine (G), and cytosine (C). Both the ratio and order of these four nucleotides indicate important features of the DNA sequence. Studies[27,33] have proved that the occurrence of certain DNA patterns may be related to the methylation level. Hence, for Benchmark 1, the ratios of A, T, G, C and eight specific fragments (sequential signatures, Table 15), which have been proven to be useful features for methylation prediction[25], were used as features for my prediction. In some recent studies[30], the methylation state of neighboring regions was incorporated as one of the features. Hence, the “percentMeth” values from RRBS experiments indicating averaged methylated percentage were gathered and averaged in window-A, and then were included as a type of feature in Benchmark 1.

Table 15

Features used for machine learning algorithms and their descriptions.

Feature name	Feature description	Used in benchmark:
Ra_A	Ratio of adenine in window-A	1, 2
Ra_B	Ratio of thymine in window-A	1, 2
Ra_C	Ratio of guanine in window-A	1, 2

Ra_D	Ratio of cytosine in window-A	1, 2
Pa_AAWGGR	Pattern frequency of AAWGGR in window-A	1, 2
Pa_TGRAAT	Pattern frequency of TGRAAT in window-A	1, 2
Pa_AAT	Pattern frequency of AAT in window-A	1, 2
Pa_ATGVAA	Pattern frequency of ATGVAA in window-A	1, 2
Pa_ACG	Pattern frequency of ACG in window-A	1, 2
Pa_GC	Pattern frequency of GC in window-A	1, 2
Pa_CG	Pattern frequency of CG in window-A	1, 2
Pa_TG	Pattern frequency of TG in window-A	1, 2
Pa_CCGC	Pattern frequency of CCGC in window-A	2
Pa_CCCC	Pattern frequency of CCCC in window-A	2
Pa_CGCC	Pattern frequency of CGCC in window-A	2
Pa_AAAG	Pattern frequency of AAAG in window-A	2
Pa_CTCC	Pattern frequency of CTCC in window-A	2
Ave_ meth	Average methylation level in window-A	1
PseTNC	74 pseudo tri-nucleotide composition features (Detail see Methods)	2
Ave_meth_Hi_C	Average methylation level in window-Bs	1, 2
Ave_Ra_A_Hi_C	Average Ra_A in window-Bs	1, 2

Ave_Ra_B_Hi_C	Average Ra_B in window-Bs	1, 2
Ave_Ra_C_Hi_C	Average Ra_C in window-Bs	1, 2
Ave_Ra_D_Hi_C	Average Ra_D in window-Bs	1, 2
Ave_Pa_AAWGGR_Hi_C	Average Pa_ AAWGGR in window-Bs	1, 2
Ave_Pa_TGRAAT_Hi_C	Average Pa_ TGRAAT in window-Bs	1, 2
Ave_Pa_AAT_Hi_C	Average Pa_ AAT in window-Bs	1, 2
Ave_Pa_ATGVAA_Hi_C	Average Pa_ ATGVAA in window-Bs	1, 2
Ave_Pa_ACG_Hi_C	Average Pa_ ACG in in window-Bs	1, 2
Ave_Pa_CCGC _Hi_C	Average Pa_CCGC in window-Bs	2
Ave_Pa_CCCC _Hi_C	Average Pa_CCCC in window-Bs	2
Ave_Pa_CGCC _Hi_C	Average Pa_CGCC in window-Bs	2
Ave_Pa_AAAG _Hi_C	Average Pa_AAAG in window-Bs	2
Ave_Pa_CTCC _Hi_C	Average Pa_CTCC in window-Bs	2
Ave_Pa_GC _Hi_C	Average Pa_GC in window-Bs	2
Ave_Pa_CG _Hi_C	Average Pa_CG in window-Bs	2
Ave_Pa_TG _Hi_C	Average Pa_TG in window-Bs	2

Note. The feature names containing “Hi_C” were generated in window-B, that is, the topological neighbors indicated by Hi-C experiments.

For Benchmark 2, that is, the 5-fold cross-validation on chromosome 1 and blind test on random combination of chromosomes 1 and 21, I incorporated more sequential features and eliminated the features indicating methylation level in neighboring region. As introduced by some recent publications[47-62], some useful statistical features for biological systems have been developed and presented. These features include pseudo amino acid composition (PseAAC)[63], pseudo k-tuple nucleotide composition (PseKNC) and pseudo trinucleotide composition (PseTNC)[64]. I implemented 74 PseTNC features as DNA sequence property features. The pseTNC (pseudo trinucleotide composition) is a statistical feature, which incorporates the occurrence frequencies of all the pseudo trinucleotide compositions. The features are defined as

$$\mathrm{D} = [d_1\ d_2 \ldots d_{64} \ldots\ d_{64+\lambda}\ \ ] \quad (7)$$

in which the first 64 features measure the local or short-range sequence pattern and the next λ = 10 components measure the global effect. The 74 features were generated by incorporating the frequency and multiple physical properties of each pseudo trinucleotide composition. The detail of calculating these features can be found in the reference[64].

Features based on three-dimensional genome topology - Benchmark 1: for each target CpG site, I gathered all the Hi-C contact pairs with one end falling into the window-A region. Using the other Hi-C end as the center, a window-B was defined with the same size of window-A. I only included the window-Bs that are > 1000nt away from the target CpG sites ensuring they are sequentially a long-distance away but proximate in three-dimensional space. In this way, I

eliminated the methylation level of the sequentially neighboring region for a target CpG site. Because multiple Hi-C pairs may have one end falling into the window-A region of each target CpG site, I usually gathered multiple window-Bs. The number of available window-Bs is influenced by the size of window-A and the Hi-C reads coverage, which was calculated by: multiplying the length of Hi-C read by the number of Hi-C reads and then dividing by the total length of the reference genome. I benchmarked my performance with different sizes of window-A. For each window-B, I generated the DNA sequence properties (Table 15) and averaged methylation PercentMeth, and then averaged these values for multiple window-Bs.

Benchmark 2: I eliminated the methylation level in window-A, but only kept the methylation levels in window-B for every target CpG sites. In this way, the prediction models no longer know the methylation level in the sequential neighboring region of a target CpG site, increasing the prediction difficulty. In order to observe how the number of Hi-C neighboring regions impact prediction performance, a “Hi-C range” was defined with the target CpG site as the center of it. The Hi-C pairs with one end fell into this “Hi-C range” and were collected; and the other end was used as the center of window-B. Only the Hi-C contacts whose two ends have a sequential distance longer than the “Hi-C range” were included so that only long-range spatial neighbors were kept.

Evaluation Methods

Evaluation criteria: the specificity (Sp), sensitivity (Se), accuracy (Acc), and Matthews's correlation coefficient (MCC) were used to evaluate prediction performance. These parameters were calculated using the following equations[65]:

$$\mathrm{Sn} = 1 - \frac{N_-^+}{N^+} \tag{8}$$

$$\mathrm{Sp} = 1 - \frac{N_+^-}{N^-} \tag{9}$$

$$\mathrm{Acc} = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \tag{10}$$

$$\mathrm{MCC} = \frac{1 - (\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-})}{\sqrt{(1 + \frac{N_+^- - N_-^+}{N^+})(1 + \frac{N_-^+ - N_+^-}{N^-})}} \tag{11}$$

where $N^+$is the total number of the positive samples (methylated samples), and N_-^+ is the number of the positive samples incorrectly predicted as negative samples (un-methylated samples), N^- is the total number of the negative samples, and N_+^- is the number of negative samples incorrectly predicted as the positive samples.

Receiver Operating Characteristic (ROC) curves for leave-one-out cross-validation were plotted with different values of threshold μ, which was used as the cutoff for methylated and unmethylated classes based on the SVM output real-number value.

Leave-one-out Cross-validation (Benchmark 1)

The performance of SVM model and SdAs were evaluated by the leave-one-out cross-validation. For the prediction of methylation state of each CpG site, the rest of the CpG sites were used as training samples. For SdA, the rest of the samples were split so that 50% of the samples were used as fine-tuning set and 50% as validation set. The same fine-tuning samples were also used in the pre-training stage (unsupervised learning), in which the target values Y were not used. The final evaluation of the prediction performance was obtained by averaging the results from all round of cross-validation.

The methylation value for each CpG site was indicated by the value of percentMeth from RRBS experiments. Methylation level of a CpG site is a continuous value ranging from 0 (un-methylated) to 1 (methylated). Because I tried to classify the methylation status of a CpG site into binary classes (methylation state), that is, either methylated or un-methylated, I incorporated two thresholds α and β to convert the continuous value of PercentMeth into binary classes. Specifically, if the PercentMeth value of a CpG site is larger than α, the CpG site is classified as methylated, and if the methylation level of a CpG site is less than β, the CpG site is classified as un-methylated (methylation-resistant). The threshold β was set first to 0.01, and then the threshold α was calculated based on β to ensure these two binary classes would have equal numbers of samples. Balancing the number of samples in each class avoids bias in training.

Blind Test on Chromosomes 21 and X (Benchmark 1)

The chromosomes 1, 2, and 3 were used as training data sets because of their relatively larger size, and chromosomes 21 and X were selected as two independent blind testing data sets because of their smaller size and the possible inactivation of female X chromosome.

Five-fold Cross-validation (Benchmark 2)

In Benchmark 2, I eliminated the feature “Ave_meth” (methylation level in the neighboring region of target CpG sites) and added 74 PseTNC features (Table 15). I collected all the CpG sites with “PercentMeth” value in the RRBS experiment equal to 0 and assigned them as un-methylated samples; the CpG sites with >= 0.9 were used as positive samples. In this way, I collected in total 559 positive samples and 1,959 negative samples. These samples were evenly split into five folds. For the training of SVM, down-sampling (cut samples from the majority class) was performed on the four training folds. The up-sampling technique (randomly picking up the same number of samples for the minority class) was performed for SdA in order to balance the positive and negative samples in the training folds. The data in the testing fold was not balanced. For SdAs, three folds were used as fine-tuning data (up-sampling balanced), one fold as validation (up-sampling balanced), and one fold as test (not balanced). Benchmark 2 was performed on chromosome 1 of the GM12878 cell line.

In order to benchmark the contribution of unsupervised pre-training of SdAs, I randomly collected 2,330 samples in chromosome 1 with unknown target value. For every round in the 5-fold cross-validation and blind test with

chromosome 21, this data set was used as the pre-training sample for training SdAs. This 5-fold cross-validation was performed with multiple “Hi-C range” (definition see Machine Learning Features section).

Blind test with Chromosome 21 (Benchmark 2)

I collected 1,039 positive and 1,746 negative samples from chromosome 21 in the same way as from chromosome 1, and randomly combined them with all the samples from chromosome 1 used in the 5-fold cross-validation. All of the randomly combined data set was split into five folds. For SVM, four folds were used to train the model and one for test. For SdAs, three folds were used as fine-tuning, one fold for validation, and one for testing. The same un-labeled data set was used for unsupervised pre-training. No up-sampling or down-sampling was performed on any of the folds. Only 10K “Hi-C range” was used in this blind test stage

Test with randomly selected windows (Benchmark 2)

To benchmark the contributions of Hi-C related features, I replaced Hi-C based window-Bs with same-size random windows, which do not have any Hi-C contacts with the “Hi-C range” of a target CpG site. All the same features were generated on the random window as for Hi-C window-B.

Parallelization of Feature Generation and SVM Classification

A parallel algorithm was designed to reduce the execution time of feature generating and SVM-light classification. First, multiple processors simultaneously read the Hi-C contact files using MPI (Message Passing Interface), and then a parallel version of SVM-light was developed to make each processor perform

learning and classification simultaneously. This parallel algorithm was designed and tested on an early version of my methods that targeted on predicting average methylation level of a segment of the genome instead of each CpG site. However, the feature types and SVM classification are the same. Execution time decreased with the increase of number of processors (see the Results section).

Statement for Experiments Involving Vertebrates and Human Subjects

This research was conducted with purely computational methods and did not use any animals, human subjects, or tissue samples. This work did not conduct any wet lab biological experiments that used vertebrates, human subjects, or tissue samples. The data of all cell-lines were downloaded from the public database ENCODE that has already been previously published.

REFERENCES

1 Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261-282 (1987).

2 Cedar, H. DNA methylation and gene activity. *Cell* **53**, 3-4 (1988).

3 Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* **33**, 245-254 (2003).

4 Bird, A. P. CpG-rich islands and the function of DNA methylation. *Nature* **321**, 209-213 (1985).

5 Takai, D. & Jones, P. A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci.* **99**, 3740-3745 (2002).

6 Bird, A. The essentials of DNA methylation. *Cell* **70**, 5-8 (1992).

7 Bird, A., Taggart, M., Frommer, M., Miller, O. J. & Macleod, D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40**, 91-99 (1985).

8 Das, P. M. & Singal, R. DNA methylation and cancer. *J. Clin. Oncol.* **22**, 4632-4642 (2004).

9 Rivenbark, A. G. *et al.* Epigenetic reprogramming of cancer cells via targeted DNA methylation. *Epigenetics* **7**, 350-360 (2012).

10 Iguchi-Ariga, S. & Schaffner, W. CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor

binding as well as transcriptional activation. *Genes Dev.* **3**, 612-619 (1989).

11 Iannello, R. C. *et al.* Methylation-dependent silencing of the testis-specific Pdha-2 basal promoter occurs through selective targeting of an activating transcription factor/cAMP-responsive element-binding site. *J. Biol. Chem.* **275**, 19603-19608 (2000).

12 Inamdar, N. M., Ehrlich, K. C. & Ehrlich, M. CpG methylation inhibits binding of several sequence-specific DNA-binding proteins from pea, wheat, soybean and cauliflower. *Plant Mol. Biol.* **17**, 111-123 (1991).

13 Kalantari, M. *et al.* Methylation of human papillomavirus 16, 18, 31, and 45 L2 and L1 genes and the cellular DAPK gene: considerations for use as biomarkers of the progression of cervical neoplasia. *Virology* **448**, 314-321 (2014).

14 Schoofs, T., Berdel, W. & Müller-Tidow, C. Origins of aberrant DNA methylation in acute myeloid leukemia. *Leukemia* **28**, 1-14 (2014).

15 Figueroa, M. E. *et al.* DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell* **17**, 13-27 (2010).

16 Akalin, A. *et al.* Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. *PLoS Genet.* **8**, DOI:10.1371/journal.pgen.1002781 (2012).

17 Timothy J, L., Christopher, M., Li, D., Baty, J. & Lucinda, F. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059-2074 (2013).

18 Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.* **43**, 768-775 (2011).

19 Grunau, C., Clark, S. & Rosenthal, A. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucl. Acids Res.* **29**, DOI:10.1093/nar/29.13.e65 (2001).

20 Smith, Z. D., Gu, H., Bock, C., Gnirke, A. & Meissner, A. High-throughput bisulfite sequencing in mammalian genomes. *Methods* **48**, 226-232 (2009).

21 Levin, J. Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7**, 709-715 (2010).

22 Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinform.* **10**, DOI:10.1186/1471-2105-10-232 (2009).

23 Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucl. Acids Res.* **33**, 5868-5877 (2005).

24 Chatterjee, A., Rodger, E., Stockwell, P., Weeks, R. & Morison, I. Technical considerations for reduced representation bisulfite sequencing with multiplexed libraries. *J. Biomed. Biotechnol.* **2012**, DOI:10.1186/s13059-015-0581-9. (2011).

25 Das, R. *et al.* Computational prediction of methylation status in human genomic sequences. *Proc. Natl. Acad. Sci.* **103**, 10713-10716 (2006).

26 Feltus, F., Lee, E., Costello, J., Plass, C. & Vertino, P. Predicting aberrant CpG island methylation. *Proc. Natl. Acad. Sci.* **100**, 12253-12258 (2003).

27 Bhasin, M., Zhang, H., Reinherz, E. L. & Reche, P. A. Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Lett.* **579**, 4302-4308 (2005).

28 Fang, F., Fan, S., Zhang, X. & Zhang, M. Q. Predicting methylation status of CpG islands in the human brain. *Bioinformatics* **22**, 2204-2209 (2006).

29 S.KiM *et al.* Predicting DNA methylation susceptibility using CpG flanking sequences. *Pac. Symp. Biocomput.* **13**, 315-326 (2008).

30 Zhang, W., Spector, T., Deloukas, P., Bell, J. & Engelhardt, B. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biol.* **16**, DOI:10.1186/s13059-015-0581-9 (2015).

31 Yamada, Y. & Satou, K. Prediction of genomic methylation status on CpG islands using DNA sequence features. *BAB* **5**, 153-162 (2008).

32 Liu, Z., Xiao, X., Qiu, W.-R. & Chou, K.-C. iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. *Anal. Biochem.* **474**, 69-77 (2015).

33 Bock, C. *et al.* CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet.* **2**, 243-252 (2006).

34 Chen, W., Feng, P., Ding, H., Lin, H. & Chou, K.-C. iRNA-Methyl: Identifying N 6-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.* **490**, 26-33 (2015).

35 Wang, Z. *et al.* The properties of genome conformation and spatial gene interaction and regulation networks of normal and malignant human cell types. *PLoS One* **8**, DOI:10.1371/journal.pone.0058793 (2013).

36 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).

37 Pan, Y. *et al.* Role of long non-coding RNAs in gene regulation and oncogenesis. *Chin. Med. J.* **124**, 2378-2383 (2011).

38 Gibb, E. A., Brown, C. J. & Lam, W. L. The functional role of long non-coding RNA in human carcinomas. *Mol. Cancer* **10**, 38-55 (2011).

39 Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* **5**, DOI:10.1371/journal.pcbi.1000598 (2009).

40 Cheetham, S., Gruhl, F., Mattick, J. & Dinger, M. Long noncoding RNAs and the genetics of cancer. *Br. J. Cancer* **108**, 2419-2425 (2013).

41 Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.* **43**, 1059-1065 (2011).

42 Engreitz, J. M. *et al.* The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341**, DOI:10.1126/science.1237973 (2013).

43 Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851-1858 (2008).

44 Joachims, T. *Making large scale SVM learning practical. Advances in Kernel Methods - Support Vector Learning*. (MIT Press, 1999).

45 Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504-507 (2006).

46 Vincent, P., Larochelle, H., Bengio, Y. & Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. *Proc. Int. conf. Mach. Learn.* **2008**, 1096-1103 (2008).

47 Qiu, W.-R., Xiao, X., Lin, W.-Z. & Chou, K.-C. iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. *J. Biomol. Struct. Dyn.* **2014**, 1731-1742 (2014).

48 Xu, Y., Shao, X.-J., Wu, L.-Y., Deng, N.-Y. & Chou, K.-C. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *PeerJ* **1**, DOI:10.7717/peerj.171 (2013).

49 Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **273**, 236-247 (2011).

50 Jia, J., Liu, Z., Xiao, X., Liu, B. & Chou, K.-C. iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their

physicochemical properties and wavelet transforms into PseAAC. *J. Theor. Biol.* **377**, 47-56 (2015).

51 Qiu, W.-R., Xiao, X. & Chou, K.-C. iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. *Int. J. Mol. Sci.* **15**, 1746-1766 (2014).

52 Qiu, W., Xiao, X., Lin, W. & Chou, K. iMethyl-PseAAC: Identification of Protein Methylation Sites via a Pseudo Amino Acid Composition Approach. *BioMed Res. Int.* **2014**, DOI:10.1155/2014/947416 (2013).

53 Xu, R. *et al.* Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach. *J. Biomol. Struct. Dyn.* **33**, 1720-1730 (2014).

54 Lin, H., Deng, E.-Z., Ding, H., Chen, W. & Chou, K.-C. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucl. Acids Res.* **42**, 12961-12972 (2014).

55 Chou, K.-C. Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst.* **9**, 1092-1100 (2013).

56 Chou, K.-C., Wu, Z.-C. & Xiao, X. iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.* **8**, 629-641 (2012).

57 Xiao, X., Wang, P., Lin, W.-Z., Jia, J.-H. & Chou, K.-C. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem.* **436**, 168-177 (2013).

58 Fan, Y.-N., Xiao, X., Min, J.-L. & Chou, K.-C. iNR-Drug: Predicting the interaction of drugs with nuclear receptors in cellular networking. *Int. J. Mol. Sci.* **15**, 4915-4937 (2014).

59 Liu, B. *et al.* Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucl. Acids Res.* **43**, W65-W71 (2015).

60 Chou, K.-C. Impacts of bioinformatics to medicinal chemistry. *Med. Chem.* **11**, 218-234 (2015).

61 Chen, W., Lin, H. & Chou, K.-C. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. *Mol. Biosyst.* **11**, 2620-2634 (2015).

62 Guo, S.-H. *et al.* iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics* **30**, 1522-1529 (2014).

63 Chou, K. C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Struct. Funct. Bioinform.* **43**, 246-255 (2001).

64 Chen, W., Feng, P.-M., Deng, E.-Z., Lin, H. & Chou, K.-C. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Anal. Biochem.* **462**, 76-83 (2014).

65 Chen, W., Feng, P., Lin, H. & Chou, K. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucl. Acids Res.* **41**, DOI:10.1093/nar/gks1450 (2013).