Discovering DNA motifs and genomic variants associated with DNA methylation

Haoyang Zeng **David K. Gifford**

Computer Science and Artificial Intelligence Lab
Massachusetts Institute of Technology
Cambridge, MA 02139
{haoyangz, gifford@mit.edu}

Abstract

DNA methylation plays a crucial role in establishing tissue-specific gene expression. However, our incomplete understanding of the *cis* elements that regulate DNA methylation prevents us from interpreting the functional effects of non-coding variants. We present CpGenie (http://cpgenie.csail.mit.edu), a deep convolutional neural network that learns a regulatory sequence code of DNA methylation and enables allele-specific DNA methylation prediction with single-nucleotide sensitivity. Variant annotations from CpGenie accurately identify methylation quantitative trait loci (meQTL) and contribute to the prioritization of functional non-coding variants including expression quantitative trait loci (eQTL) and disease-associated mutations.

1 Introduction

DNA methylation is an important epigenetic state that is involved in the regulation of key biological processes, including the establishment and maintenance of tissue-specific expression profiles, X-chromosome inactivation, genomic imprinting, transposable element silencing, cell differentiation, and inflammatory processes (Bird, 2002; Bock, 2012; Barlow, 2011; Martin and Herceg, 2012; Meissner, 2010; Bestor, 1998). DNA methylation encodes cellular state information not contained in other epigenetic marks, and can be used to improve the prediction of tissue-specific functional elements such as enhancers (Lee et al., 2015; Hwang et al., 2015). Disease specific changes in DNA methylation have been reported for cancer and certain other diseases (Feinberg, 2007; Baylin and Jones, 2011). While the mechanism for the inheritance of DNA methylation down a cellular lineage is known in detail (Bird, 2002), the regulatory mechanism that determines its tissue-specific state remains largely unknown.

Recent advances in high-throughput sequencing technologies have enabled the observation of DNA methylation at single CpG resolution in a wide variety of cell types. Methods such as restricted representation bisulfite sequencing (RRBS) and whole-genome bisulfite sequencing (WGBS) have enabled genome-wide maps of DNA methylation to be routinely used to profile DNA methylation to help reveal the active genome.

Cell-type specific DNA methylation data has been used to train computational models to predict the sequence features that are associated with cell-type specific differences in DNA methylation. Early work focused on analyzing the overall methylation level of CpG islands (Feltus et al., 2003; Das et al., 2006; Fang et al., 2006; Bock et al., 2006, 2007; Fan et al., 2008; Previti et al., 2009; Zheng et al., 2013). With the introduction of whole genome bisulfite sequencing, recent work has focused on predicting the methylation level of single CpGs using a range of features (Bhasin et al., 2005; Kim et al., 2008; Lu et al., 2010; Zhou et al., 2012; Ma et al., 2014; Zhang et al., 2015; Wang et al., 2016;

Fan et al., 2016; Angermueller et al., 2016). Sequence-only approaches predict DNA methylation directly from DNA sequence by using k-mer counts and other sequence features (Bhasin et al., 2005; Kim et al., 2008; Lu et al., 2010; Zhou et al., 2012). Sequence-plus-state approaches have improved predictive accuracy by using additional functional information including the methylation level of neighboring CpGs, histone marks, transcription factor binding sites and genome topology data (Hi-C) (Zhang et al., 2015; Wang et al., 2016; Fan et al., 2016; Angermueller et al., 2016). However, previous approaches that use sequence features have not provided an interpretable set of sequence features that are associated with proximal DNA methylation.

A complementary perspective we examine here is to predict the effect of genetic variants on proximal DNA methylation. Many quantitative trait loci associated with DNA methylation (meQTLs) have been discovered (Kaplow et al., 2015; Banovich et al., 2014), suggesting a direct causal link between genetic variants and DNA methylation. Given the crucial role of DNA methylation in gene regulation, variants associated with complex traits could impact downstream cellular function through DNA methylation modulation. Thus the capacity to predict the impact of genomic variants on DNA methylation *in silico* is desired for accurate prioritization and interpretation of functional non-coding variants.

We introduce CpGenie (Figure 1), a deep-learning model that predicts the DNA methylation level of a single CpG from sequence context with an accuracy that surpasses existing methods. We find that the sequence determinants learned by CpGenie correspond to the binding motifs of proteins known for their involvement in the regulation of DNA methylation state. CpGenie is able to predict the effect of non-coding variants on DNA methylation at single-nucleotide resolution, including predicting meQTLs that result in an allelic imbalance of DNA methylation. In addition, we show that CpGenie improves the prediction of expression quantitative trait loci (eQTLs) and disease-associated variants by providing complementary functional information to other data such as transcription factor binding. We provide CpGenie as open source software available at http://cpgenie.csail.mit.edu .

2 Results

2.1 A convolutional neural network accurately predicts DNA methylation of single CpGs from sequence

Unlike traditional models, convolutional neural networks (CNN) automatically learn features at different levels, from single nucleotide counts to motif co-occurrences, that help predict DNA methylation. Convolutional neural networks can be efficiently trained using graphical processing units (GPUs), and can easily scale to millions of CpG methylation training examples from high-throughput DNA methylation data to learn sophisticated sequence patterns that determine the methylation status of individual CpGs.

We evaluated 10 different convolutional neural network architectures for predicting DNA methylation (Supplementary Table 1). Our evaluation used our platform for efficiently tuning and comparing different deep learning architectures on the Amazon EC2 Cloud (Zeng et al., 2016). For each CpG training example we extracted the 1001bp sequence centered at the CpG site and encoded each base as a one-hot vector of length four denoting the existence of each of the four nucleotides. Each of the 10 network architectures was trained on the same reduced representation bisulfite sequencing (RRBS) data from the GM12878 lymphoblastoid cell line and then used to predict whether the methylation level for a test set of CpGs was greater than 0.5. Previous work has reported that the methylation status of adjacent CpG sites are highly predictive of a target CpG's status (Fan et al., 2016; Wang et al., 2016; Zhang et al., 2015). With the large sequence-context window we consider, adjacent CpG sites will share similar sequence features. To minimize memorization effects we trained on data from chromosome 1-9 and chromosome 14-22, performed hyper-parameter and model selection on data from chromosome 12-13, and used data from chromosome 10-11 for testing. Thus the CpGs used to evaluate model performance were in distinct contexts from the training data.

We compared the performance of three existing methods to CpGenie: (1) the sequence module of DeepCpG, a deep learning method for imputing missing CpG methylation values in single-cell DNA methylation experiments using both sequence and adjacent methylation level; (2) gkm-SVM, a support vector machine (SVM)-based method with great performance in predicting transcription factor binding and DNase-seq peaks from sequences (Ghandi et al., 2014); and (3) a random forest (RF) classifier trained on 4-mer frequencies of the input sequence, considering that random forest

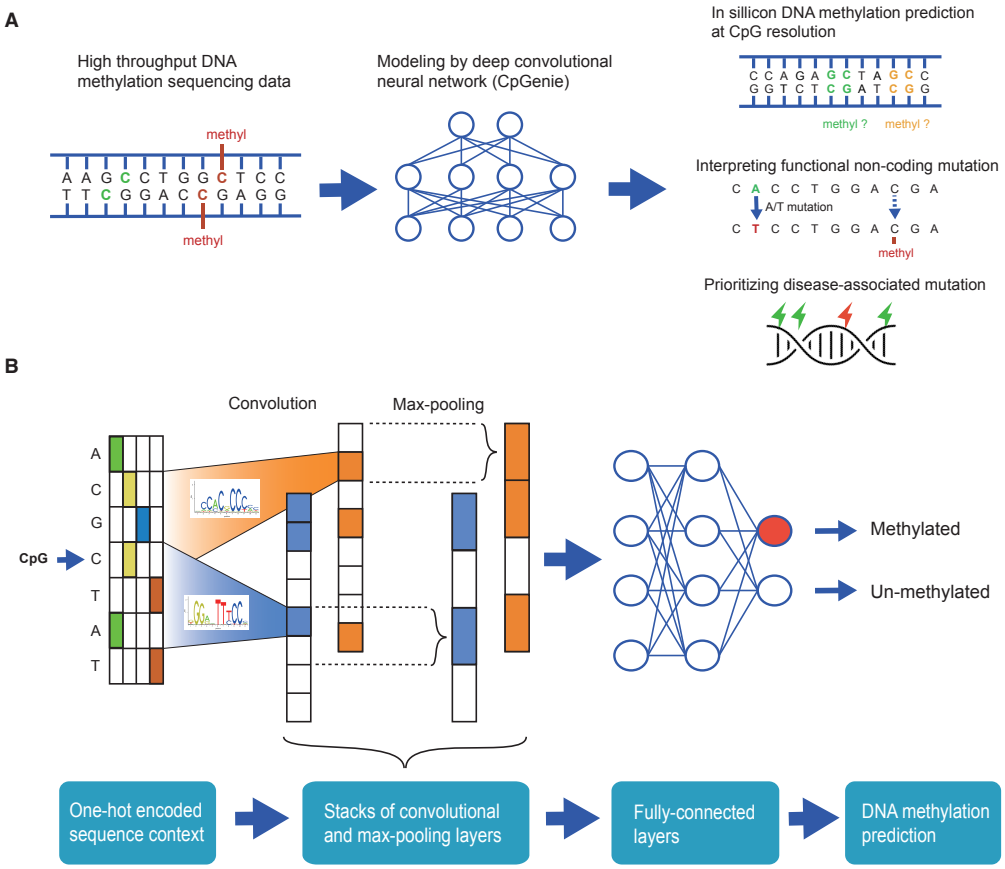

Figure 1: Schematics of CpGenie. (A) CpGenie takes the high-throughput DNA methylation sequencing data, such as restricted representation bisulfite sequencing (RRBS) or whole-genome bisulfite sequencing (WGBS) as input and produces predictions of CpG methylation as output. CpGenie can predict DNA methylation at CpG resolution, interpreting the functional consequence of non-coding sequence variants, and prioritizing causal mutations from GWAS-determined associations. (B) CpGenie converts the sequence context around a CpG into one-hot encoding, and transforms it to higher-level features through three pairs of convolutional and max-pooling layers. Two fully-connected layers follow to make predictions on the methylation status of the queried CpG.

and k-mer frequencies has been reported to accurately predict DNA methylation (Zhang et al., 2015; Fan et al., 2016; Lu et al., 2010) .

All of the convolutional neural network structures we designed achieved good performance, the best of which we called CpGenie and reached an area under receiver operating characteristic (auROC) of 0.854 and an area under precision-recall curve (auPRC) of 0.685 on the held-out test set (Supplementary Table 1). CpGenie surpassed the performance of all of the competing methods (Figure 2A). Having evaluated the performance on a single cell line, we further applied CpGenie and the competing methods to 50 RRBS datasets from ENCODE (Supplementary Table 6) to systematically benchmark their capacity in predicting DNA methylation. CpGenie robustly outperformed all the alternative methods, with a better auROCfor 42 of the 50 experiments (Figure 2B, C) and a better auPRC for 49 of the 50 experiments (Supplementary Figure 1). We also found 1001bp to be the sequence window size that optimized performance (Supplementary Figure 2). This suggests that sequence features over 500bp away may be involved in regulating CpG methylation.

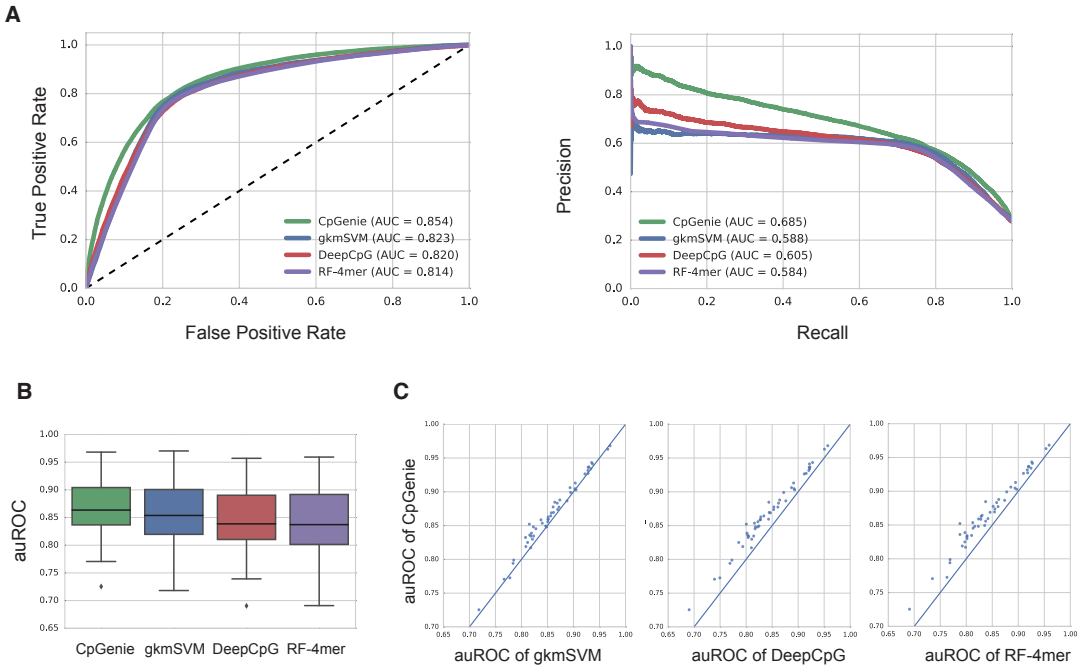

Figure 2: CpGenie outperformed competing methods in predicting DNA methylation at CpG resolution. (A) The receiver operating characteristic (ROC) curve (left) and precision-recall (PRC) curve (right) of CpGenie (green), gkmSVM(blue), DeepCpG (red), and Random Forest (purple) for predicting DNA methylation status of held-out CpGs in GM12878 RRBS data. (B) The box plot of area under the ROC curve (auROC) of CpGenie (green), gkmSVM (blue), DeepCpG (red), and Random Forest (purple) for predicting DNA methylation status of held-out CpGs in 50 RRBS data sets from ENCODE. (C) Pairwise comparison of auROC on the 50 RRBS datasets between CpGenie and the gkmSVM (left), DeepCpG (middle) and Random Forest (right).

2.2 CpGenie learns the binding motifs of proteins known to regulate DNA methylation

We expected that a predictive model of DNA methylation from sequence would learn motifs that correspond to regulators associated with the mechanism of DNA methylation. Previous studies have established that many transcription factors interact with DNA methyltransferases (DNMT) that methylate DNA (Hervouet et al., 2009). As transcription factors are known for binding DNA with strong sequence specificity, we evaluated CpGenie to determine if it learned certain of these sequence motifs.

The basic unit of a convolutional layer is a "kernel" that searches for patterns in the input, analogous to a motif scanner looking for motif matches. Interpreting the convolutional kernels in the first layer of a network is crucial for understanding how the network responds to an input sequence (Kelley et al., 2016; Alipanahi et al., 2015). We collected all of the sequence regions in the training examples that activate the first convolutional layer, and aligned the sequences activating the same convolutional kernel into a positive weight matrix (PWM). All the PWMs were then compared to database of known transcription factor motifs at a false discovery rate of 0.1 as suggested by Kelly et al. (Kelley et al., 2016) (Methods)

We found that 97 out of the 128 PWMs recognized by CpGenie significantly match the motifs of known transcription factors (Figure 3A, Supplementary Table 2). We found the motifs of 21 transcription factors known to strongly interact with a DNA methyltransferase (DMNT) (Hervouet et al., 2009), including EGR-1,GATA1 and SP1. As a point of comparison, Das et al. (Das et al., 2006) found that the motifs of 31 transcription factors, only a small fraction (6/31) of which overlap with the DNMT-interacting factors, help classify hyper- methylated regions from hypo-methylated regions. We found 48% (15/31) of the Das et al. discovered motifs were among the CpGenie-discovered

PWMs including KROX and E2F. Moreover, although without a statistically significant overall match, many CpGenie discovered PWMs capture motif information associated with transcription factors previously reported to be associated with DNA methylation, such as NF-kappaB, MEF-3, and LUN-1 (Supplementary Figure 3).

Interestingly, a large number of the CpGenie discovered PWMs are dedicated to motif variants of PAX4 and SP3 (24 and 23 respectively). Hervouet et al. reported DNMT-interaction with other transcription factors in the same family (PAX6, SP1 and SP4). Certain of the predictive transcription factor motifs discovered with CpGenie are not known for involvement in DNA methylation. Two examples are GFI1 (FDR q-value=0.0025), which is a transcriptional repressor that functions by histone deacetylase (HDAC) recruitment, and THRA (FDR q-value=0.0023), which is a nuclear hormone receptor that can act as a repressor or activator of transcription.

We next scored the importance of CpGenie's 128 first-layer convolution kernels with an optimization-based framework. The framework identifies the first-layer kernel activation pattern that can maximize the network's confidence to classify a sample as one class (methylated / unmethylated) (Methods). To understand the biological relevance of the top-ranking kernels, we chose a more stringent false discovery rate of 0.01 when matching with known motifs. The top 10 convolution kernels for high and low methylation prediction are quite distinct, with the exception that SP3 is important for predicting both high and low methylation (Supplementary Table 3).

Similar to other epigenetic marks, DNA methylation varies across tissues and cell lines. We analyzed the tissue-specific RRBS data from ENCODE, and found that placenta and testis have the most distinct DNA methylation pattern (Methods). To investigate the regulatory mechanism behind cell line-specific methylation, we applied CpGenie to predict whether a CpG is differentially methylated in placenta and testis. We labeled a CpG as tissue-specific if its magnitude of methylation change ranks at top 5% among the differential methylation data combined from all pairs of tissues, and labeled as tissue-invariant if otherwise. With the same model structure and cross-validation scheme as previously described, CpGenie trained on the 1001 bp sequence context can well classify tissue-specific from tissue-invariant CpG sites on held-out chromosomes with an auROC of 0.79.

We interpreted our model of tissue-specific DNA methylation to see if we could find sequence features predictive of tissue-specific DNA methylation. With a stringent FDR of 0.01, 53 out of the 128 PWMs encoded in the model significantly match known transcription factors (Supplementary Table 4). CTCF, which is an important chromatin insulator, and several members of SP family, which contains important transcription regulators, appeared at top of the list as highly predicative for tissue-invariant methylation. Sequence features associated with PAX family and ZFX, a transcription coactivator, were used by the model to predict tissue-specific methylation. Interestingly, slightly different variants of SP1 motifs were involved in the prediction of both tissue-specific and tissue-invariant CpGs. Most of the other top 10 convolution kernels predictive for tissue-specific and tissue-invariant CpG methylation are *de novo* (Figure 3B,C, Supplementary Table 5).

2.3 CpGenie accurately predicts allele-specific DNA methylation and cis-acting meQTL

We next explored CpGenie's capability to identify genetic variants that modulate DNA methylation at single-nucleotide resolution. Kaplow et al. analyzed the DNA methylation level of over 800,000 single nucleotide polymorphism (SNP)-CpG pairs by mapping the bisulfite sequencing reads back to the reference and alternate allele of a variant (Kaplow et al., 2015). They found over 2,000 genetic variants (meQTLs) with statistically significant allelic imbalance of DNA methylation. As only reads that overlap with both the CpG site and the variant locus were counted, the meQTLs discovered from this method tend to act *in cis* (with an average distance of 25.4 bp), making these data a relevant standard to evaluate the ability to predict allelic change of DNA methylation in the presence of a sequence variant.

Data from Kaplow et al. contain more CpG sites with methylation levels distributed around 0.5 compared with RRBS data (Supplementary Figure 4). Thus we re-trained a CpGenie model on this dataset with the same chromosome-based cross-validation scheme as previously described for the RRBS data. The best performing architecture reached a decent auROC of 0.75, surpassing a DeepCpG model trained and tested on the same datasets (auROC = 0.69).

CpGenie accurately predicted the direction of allelic methylation change caused by sequence variants. When applied to the meQTLs on the held-out chr10 and chr11, CpGenie accurately identified the allele

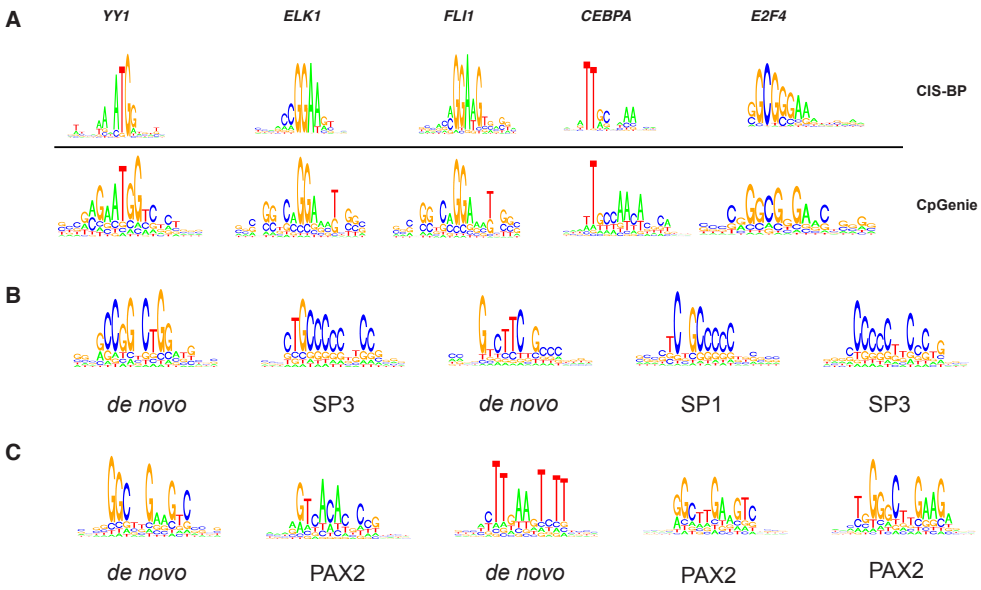

Figure 3: CpGenie learns a regulatory code of DNA methylation. (A) 97 out of 128 of the convolutional filters match motifs of known transcription factors in the human CIS-BP database at an FDR threshold of 0.1. (B, C) The motif logo of the convolutional filters highly predicative (top 5) of tissue-invariant (B) or tissue-specific (C) DNA methylation

with higher DNA methylation level, showing sensitivity and accuracy to single-nucleotide changes (Figure 4A). Moreover, the accuracy quickly and stably increased to 100% when we gradually retained only the high-confidence predictions by increasing the threshold of the absolute allelic difference in the predicted methylation (Figure 4B). For instance, for the variants of which the predicted absolute difference of DNA methylation between the two alleles is greater than 0.03, CpGenie identified the allele with more methylation with an accuracy $> 90\%$.

We applied CpGenie to the 201 meQTLs and 76,532 non-meQTLs on the held-out chromosome 10 and 11, and scored each variant with the absolute allelic difference in the methylation prediction. CpGenie assigned significantly higher scores to meQTLs than those to non-meQTLs (Figure 4C, Mann-Whitney U test p-value $= 3 \times 1e^{-31}$). To simulate different equilibrium linkage structure, we sampled three random subsets of the non-meQTLs that are 10 times, 50 times and 100 times the size of meQTLs. With the CpGenie-assigned scores, we can accurately discriminate meQTLs from non-meQTLs with a precision-recall performance that consistently surpasses that of DeepCpG (Figure 4D). Thus CpGenie can produce high-confidence predictions of genetic variants' impact on DNA methylation.

2.4 CpGenie improves computational interpretation of non-coding variants

We extended CpGenie to evaluate the functional impact of genetic variants on methylation level of nearby CpGs (Methods). We show that these allele-specific features of DNA methylation provide useful information in the downstream interpretation and prioritization of eQTLs and disease-associated variants.

Zhou et al. (Zhou and Troyanskaya, 2015) collected two publicly available datasets of functional variants, one with 78,613 eQTLs from GRASP (Genome-Wide Repository of Associations between SNPs and Phenotypes) database (Leslie et al., 2014) and one with 12,296 disease-associated SNPs from the US National Human Genome Research Institute's GWAS Catalog (Welter et al., 2014). For each dataset, five size-matched negative sets were constructed by sampling from different subsets of 1000 Genome Project SNPs (Consortium et al., 2012). Using predictions from their DeepSEA model, the authors generated 1,838 allele-specific features of DNase sensitivity, histone mark and transcription factor binding for each SNP. A logistic regression model trained on this feature set

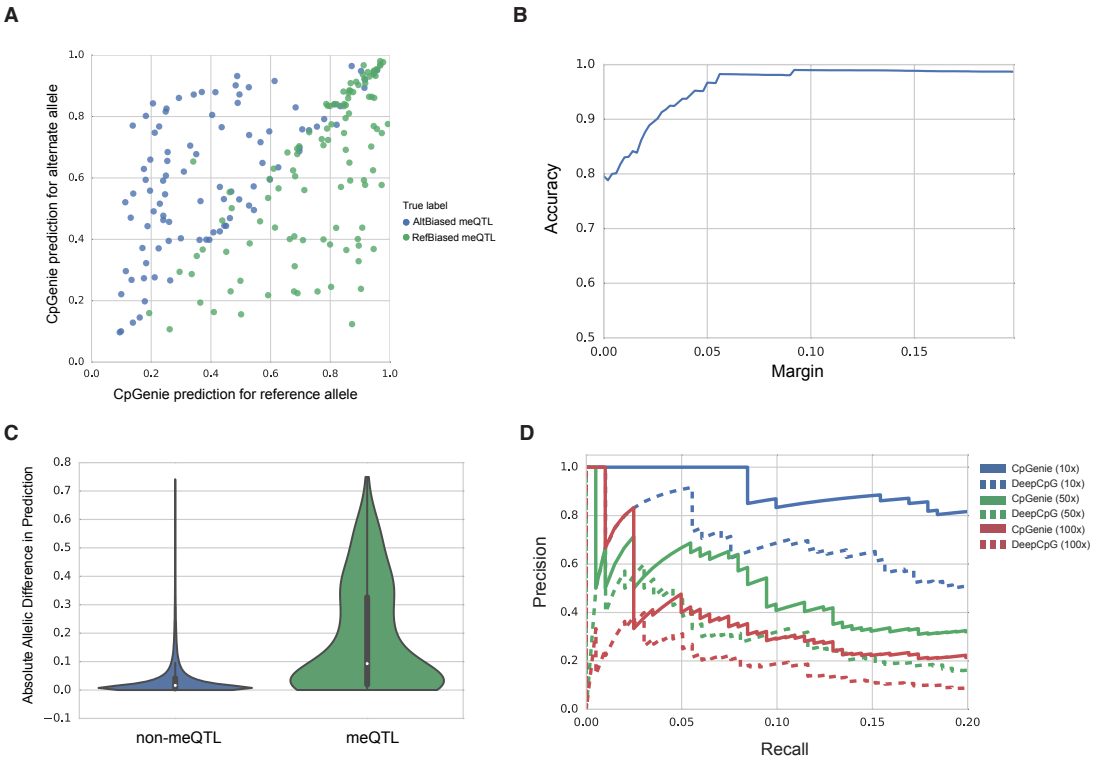

Figure 4: CpGenie accurately predicts the direction of allele-specific (AS) DNA methylation and prioritizes methylation QTLs. (A) CpGenie's DNA methylation prediction for the reference and alternate alleles of 201 meQTLs on held-out chromosome 11 and 12. The x and y axis represents the CpGenie predicted DNA methylation level. The blue and red dots represent reference allele-biased and alternate allele-biased variants respectively as experimentally determined by Kaplow et al. (B) Prediction accuracy quickly and steadily increased to 100% when only the high-confidence predictions were retained. The y-axis denotes accuracy and the x-axis represents margin, or the threshold of predicted absolute allelic difference in methylation to retain high-confidence predictions. (C) The absolute allelic differences of CpGenie-predicted DNA methylation are significantly higher for meQTLs (green) than non-meQTLs (blue). (D) The precision-recall curve (PRC) for classifying the 201 meQTL from three different random subsets of the 76,532 non-meQTL that are 10 times (blue), 50 times (green), and 100 times (red) the size of meQTL. CpGenie (solid line) outperformed DeepCpG (dash line) with better precision at the same recall.

plus conservation scores outperformed previous approaches, including CADD (Kircher et al., 2014), GWAVA (Ritchie et al., 2014) and Funseq2 (Fu et al., 2014), in classifying functional variants (eQTLs or GWAS SNPs) from the negative sets.

We evaluated the contribution of CpGenie's DNA methylation features to eQTL and GWAS SNP prioritization on these two datasets. We applied the CpGenie models trained on the 50 RRBS datasets from ENCODE to generate 250 features for each variant (Methods). A regularized logistic regression model trained on the DNA methylation features alone achieve a performance better than (for eQTL) or comparable to (for GWAS SNPs) the best performer among CADD, GWAVA and Funseq2, all of which were trained on much more annotation information beyond the sequence including histone modification, transcription factor binding, and gene expression (Figure 5A).

To assess the relative importance of DNA methylation features, we combined the prediction from CpGenie and DeepSEA, and trained a Random Forest model in which features importance can be evaluated by mean decrease impurity (Methods). In both eQTL and GWAS SNP prioritization, CpGenie-predicted DNA methylation features are considered significantly more important than the original DeepSEA features as a whole (Mann-Whitney U test, Figure 5B). Thus the allele-

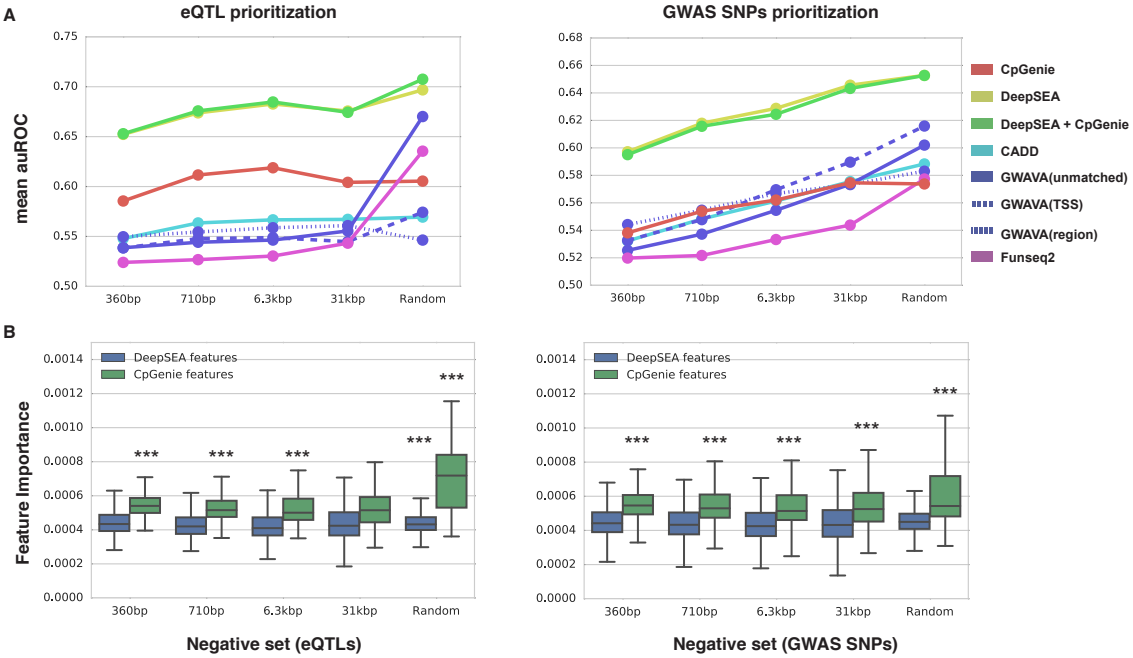

Figure 5: CpGenie's sequence-based DNA methylation predictions prioritize functional non-coding variants. (A) Compared to previous methods that utilize more annotation information, CpGenie achieved better or comparable performance in prioritizing noncoding GRASP eQTLs (left) and noncoding GWAS Catalog SNPs (right) against noncoding 1000 Genome Project SNPs. The x-axis denotes the mean distance of the SNPs in the negative set to the paired positive SNP. The 'Random' group denotes 1,000,000 randomly sampled 1000 Genome Project SNPs. (B) CpGenie's DNA methylation features (green) were considered significantly more important in general than DeepSEA's functional predictions on histone modification, transcription factor binding and DNase hypersensitivity (blue) in eQTL (left) and GWAS SNPs (right) prioritization. The asterisks denote statistical significance calculated from Mann-Whitney U test (p-value < 0.001).

specific DNA methylation prediction contributes as much as other functional interpretations, such as histone modification and transcription factor binding, in functional non-coding variants prioritization. Interestingly, in most of the prioritization tasks, we didn't observe performance improvement after the DNA methylation predictions were combined with the original DeepSEA features, which could suggest great correlation and information redundancy between the features generated from DeepSEA (DNase hypersensitivity, histone mark peak, TF binding peak) and CpGenie (DNA methylation).

3 Methods

3.1 Convolutional neural network

Convolutional neural networks (CNN) are a type of artificial neural network that have been utilized on many tasks ranging from computer vision to DNA-protein binding (Krizhevsky et al., 2012; LeCun et al., 2015; Tompson et al., 2014b; Sainath et al., 2013; Tompson et al., 2014a; Le, 2013; Zhou and Troyanskaya, 2015; Alipanahi et al., 2015; Kelley et al., 2016). As the basic unit of CNN, a convolutional layer has dozens to hundreds of "kernels" that perform convolution followed by a non-linear transformation on a small region of the input to look for a particular pattern. By tiling the same convolutional kernels on the input, the convolutional layer can identify complex patterns wherever they exist in the input. The convolutional layer is usually followed by a max-pooling layer that reduces the number of parameters in the network by only retraining the approximate location of the maximum output in a small window. A "deep" neural network with several pairs of convolutional and max-pooling layer stacked on top of each other is capable to learn higher level features as the combination of lower ones. The last few layers of a convolutional neural network are usually

fully-connected layers, in which each output node connects to every input node, summarizing the features automatically learned from the previous stack of convolutional and max-pooling layers to classify the label of the input. The objective function of a neural network is task-dependent. For classification, it is usually defined as cross-entropy between the predicted probability and the true class label. Several approaches exist to regularize the network to avoid overfitting, where max-norm constraint on the weights in combination with drop-out, which with certain probability randomly sets the output of the each neuron in the applied layer to zero, has been reported to produce the best performance (Srivastava et al., 2014). Back-propagation and stochastic gradient decent (SGD) based optimization algorithms, such as RMSprop (Tieleman and Hinton, 2012), are the standard ways to train neural network.

We transformed each DNA sequence of length L into a 2-D matrix of size $4 \times L$, where each column is a one-hot vector encoding the presence of the four DNA nucleotides A, C, G, and T. We followed the above principles to design and implement our convolutional neural networks with Keras (`https://keras.io/`), a python-based deep learning package. To find the best neural network structure, we adopted the framework introduced in Zeng et al. (2016) to efficiently train 10 different architecture variants on g2.2xlarge instances on Amazon EC2 cloud. The full list of architectures we compared and their descriptions are in the Supplementary Table 1. The best performing structure has three pairs of convolutional layers (128, 256,and 512 convolutional kernels of size 10) and max-pooling layers (with window size 5), followed by two fully-connected layers of 64 neurons and one output layer of two neurons to produce predictions (methylated/unmethylated).

Hyper-parameters, such as learning rate of the optimizer, has a large impact on the performance and the best performing hyper-parameter values are usually architecture-specific (Bergstra and Bengio, 2012; Bergstra et al., 2011). We used Hyperas (`https://github.com/maxpumperla/hyperas`) to perform hyper-parameter tuning on the learning rate and the drop-out ratio of the fully-connected layers. For each architecture, 9 different combinations of hyper-parameters were used to train on the first 10,000 samples in the training set, and tested on the validation set. The best combination was picked to train on the whole dataset for 20 epochs, where an epoch is the time when the model has gone through the whole training set for once. Every 1 epoch, the model was saved and tested on the validation set. The model with the smallest validation loss was picked as final one and evaluated on the test set.

We also extended CpGenie to score functional variants by computing the predicted change in DNA methylation. Given a sequence variant, we score all CpGs within 500 bp for their allele-specific DNA methylation levels. The maximum, mean and sum of the methylation level across all the CpGs within 500 bp of the variant is reported for each allele. In case where no CpG exist in the 500bp vicinity, a pseudo-methylation level of 0.001 is reported. Then we combined the allelic features by calculating the absolute change

$$|ref - alt| \tag{1}$$

in the sum/mean/max, and the absolute change of log odds

$$|log\frac{ref}{1-ref} - log\frac{alt}{1-alt}| \tag{2}$$

in the mean/max of nearby DNA methylation level, resulting five features for each variant. Finally we concatenate the features from 50 CpGenie models trained on RRBS datasets from ENCODE immortal cell lines to form a feature vector of length 250 for each variant.

3.2 High-throughput DNA methylation data

The 50 RRBS datasets for immortal cell lines, including GM12878, were downloaded from ENCODE website (`https://www.encodeproject.org/`). We merged multiple replicates for the same experiments, and where a CpG exists in all replicates we merged the counts of methylated and unmethylated reads and re-calculated the percentage of methylation. We further applied a minimum-read cutoff of 10 to filter out unreliable samples. Samples from chromosome 1-9 and chromosome 14-22 were used for training, samples from chromosome 12-13 were used for hyper-parameter tuning and model selection, and the rest of the data were held-out for testing.

The raw allele-specific DNA methylation data were obtained from Fraser lab (personal communication). They surveyed 823,726 SNP-CpG pairs, among which 2,379 are meQTLs. After filtering out CpGs with read counts less than 10, the whole dataset was split into training, validation and testing

set in the same way as previously described for RRBS data. For simplicity, only data for the reference allele were used in training and evaluating CpGenie. Only data from the held-out chromosome 10 and 11 were used to evaluate allele-specific methylation prediction and mQTL prioritization.

3.3 Comparison with existing methods

gkm-SVM We downloaded the gkm-SVM R package (version 0.65, `https://cran.r-project.org/web/packages/gkmSVM`) and ran it with the default parameters. Consistent with Zhou et al. (Zhou and Troyanskaya, 2015) and Kelly et al. (Kelley et al., 2016) with a sequence length of 1001 bp, the gkm-SVM software does not scale to the around 1 million CpG samples in the RRBS data due to its memory requirement to compute a full kernel matrix. Therefore we adopted the approach of Zhou et al. (Zhou and Troyanskaya, 2015) to randomly sample 5000 positive samples and an equal number of negative samples for training.

DeepCpG We download the DeepCpG package (version 0.0.1, `https://github.com/cangermueller/deepcpg`) and trained all the models with the default parameters. As required by DeepCpG, we extracted the 501bp sequence centered at a CpG as input and embedded sequences into one-hot encoding as input features to the model.

Random Forest We counted the frequency of each possible 4-mer in the 1001bp sequence centered at a CpG with JELLYFISH ((Marçais and Kingsford, 2011), version 2.2.6, `https://github.com/gmarcais/Jellyfish/releases`), generating 256 features for each sample. We used the Random Forest implementation in sciki-learn Python package (`http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`).

3.4 Network interpretation

We used the same visualization method as adopted in previous work (Kelley et al., 2016; Alipanahi et al., 2015) to convert the first layer kernels to PWMs. For each convolutional kernel, we searched through all the samples for all that can activate at least one neuron (output of the neuron > 0.5 of the maximum output among all samples) in the first convolutional layer. Each such activation was mapped back to the input sequence to locate the region that led to the activation. For each convolutional kernel, we aligned all of the activating sequences to generate a PWM. To understand the biological meaning of these PWMs, we used tomtom (Gupta et al. (2007), ver 4.11.1) to match the PWMs to known human motifs in CIS-BP database (Weirauch et al., 2014) with a FDR threshold of 0.1 as suggested in Kelly et al. (Kelley et al., 2016) When combined with importance analysis, we used a more stringent FDR of 0.01.

We interpret the importance of the first layer kernels with an optimization-based framework. We fixed all weights in a trained CpGenie model, and optimized the output of the neuron in the last layer that corresponds to the target label (methylated/unmethylated, or tissue-specific/tissue-invariant) with respect to the input of the second layer (i.e. the output of the first max-pooling layer) under a $L2$ regularization. The resulting optimum input is a $2D$ matrix, representing the spatial activation pattern of each of the first layer convolutional kernels for the network to reach high confidence in the prediction. For each kernel, we assigned the importance as the maximum activation from all locations.

3.5 Analysis of tissue-specific DNA methylation

We obtained 17 RRBS datasets corresponding to 16 different tissues from the ENCODE website (Supplementary Table 6). We first filtered out CpG sites with fewer than 10 reads. For each pair of RRBS datasets, only the CpG sites surveyed in both datasets were retained and the absolute changes in methylation level (between 0 and 1) were recorded. Then we calculated the cumulative probability function (CDF) of absolute methylation change and evaluated the methylation variability between the two tissues by the area under CDF (auCDF), where an auCDF close to one implies that all CpG sites have no change of methylation (identical datasets) and an auCDF close to zero indicates that most CpG sites have the opposite methylation status in the two tissues.

We set the methylation change cutoff for tissue-specific methylation such that the empirical frequency of observing a larger change is less than 5%. We combined and sorted the absolute methylation change

data collected above for each pair of tissues. The magnitude of change ranks at the top 5% was used as the cutoff (0.3337). With this cutoff, 14% of the CpG sites (143330 /1013716) are tissue-specific between placenta and testis. When training CpGenie to identify CpG sites differentially methylated in placenta and testis, we used the samples from chromosome 1-9 and chromosome 14-22 for training after matching the size of positive and negative set by random subsampling the tissue-invariant sites. The data from chromosome 12-13 were used for validation, and the samples from chromosome 10-11 were held-out for evaluation.

3.6 Functional variant prioritization

We obtained the eQTL and GWAS SNPs datasets, as well as their corresponding five negative sets from the supplementary tables in Zhou et al. (Zhou2015) Four of the five negative sets were constructed by finding, for each positive variant, the closest SNP in the full set, 20%, 4% and 0.8% random subset of 1000 Genome variants with minor allele frequency distribution matched to the positive set. The mean distance to the positive set is 360 bp, 1,400 bp, 6,300 bp and 31k bp for these four negative sets respectively. The fifth negative set was constructed by sampling 1,000,000 non-coding 1000 Genome SNPs with minor allele frequency distribution matched to the positive set.

For each variant in the positive and negative set, we applied the CpGenie-extended variant scoring framework to generate 250 DNA methylation features that describe the impact on the proximal DNA methylation levels. As described in Zhou et al., for each negative set we trained $L2$-regularized logistic regression models on CpGenie features, DeepSEA features, and features combined from CpGenie and DeepSEA respectively using the scikit-learn implementation (`http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegressionCV.html`). The performance was evaluated with 10-fold cross-validation. For CADD, GWAVA and Funseq2, the auROC reported in Zhou et al. (Zhou and Troyanskaya, 2015) was directly used as we tested on the same dataset.

To interpret the feature importance, we trained a Random Forest classifier on the same tasks as above with all the features normalized to have mean 0 and variance 1 before training. We used the scikit-learn implementation (`http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`) in which the feature importance is calculated as the "mean decrease impurity" defined as total decrease in node impurity averaged over all trees of the ensemble (Breiman et al., 1984).

4 Discussion

We found that CpGenie learns sequence determinants associated with DNA methylation from large-scale data generated from high-throughput bisulfite sequencing technology. Compared with existing models, CpGenie demonstrated superior performance when systematically evaluated on 50 restricted representation bisulfite sequencing (RRBS) datasets from the ENCODE project. CpGenie's consistently good performance across datasets shows that it can be reliably applied to DNA methylation data from different cell lines, and that the model learns meaningful sequence determinants rather than fitting to batch effects.

With our visualization framework, we are able to interpret the predictive sequence features learned by CpGenie to suggest connections to the biological mechanisms underlying the regulation of DNA methylation. We recovered the motifs of transcription factors previously reported to interact with DNMT or known to be discriminative of hypo- and hyper-methylated regions. CpGenie also learned important sequence features corresponding to transcription regulators not known for their roles in methylation regulation, as well as *de novo* motifs not matched to any known TFs. CpGenie models trained to distinguish tissue-specific and tissue-invariant methylation revealed sequence features that are potentially important for tissue-specific regulation of DNA methylation, and certain of these sequence features do not match the motifs of known TFs. Understanding if these *de novo* sequence codes are involved in the tissue-specific regulation of DNA methylation will require more experimental analysis.

CpGenie achieved high accuracy in predicting allele-specific DNA methylation, a difficult task that demands methods with sensitivity to single-base changes in sequence context. The accurate allele-specific predictions from CpGenie allow us to well identify methylation quantitative trait

loci (meQTLs) from variants that exhibit no impact on DNA methylation. As a sequence-based model, the ability of CpGenie to generate allele-specific predictions enables more precise and *in situ* interpretation of sequence variants than alternative approaches that are solely based on non-allele-specific features, for instance the overall DNase hypersensitivity level in the nearby regions.

We provide a variant-scoring framework that extends on CpGenie's allele-specific DNA methylation prediction. This framework assesses the effect of a variant on the DNA methylation level of the nearby CpG sites, and can play important roles in explorative and interpretive analyses of non-coding variants, such as causal variants fine-mapping and unveiling the pathogenic pathways of a known causal mutation. We demonstrate that a simple logistic regression model trained on the scores from this framework can prioritize eQTLs and disease-associated SNPs with a performance competitive to several more complicated models that utilize additional information. Moreover, the DNA methylation scores were considered more important than many other functional annotations, such as DNase sensitivity, histone marks and transcription factor binding, when jointly included in a model that is trained to predict functional non-coding variants such as eQTLs and GWAS SNPs.

We envision CpGenie to be a resource to help understand the regulatory mechanism encoded in the non-coding region of the genome, and contribute to interpreting the effect of non-coding variants associated with complex traits and diseases.

Competing interests

The authors declare no competing interests.

Acknowledgments

We are grateful for insight suggestions from other members in Gifford Lab. We acknowledge funding from the National Institutes of Health under grants R01HG008363 and U01HG007037 to D.K.G. and an equipment grant from NVIDIA.

References

Alipanahi B, Delong A, Weirauch MT, and Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature biotechnology* **33**: 831–838.

Angermueller C, Lee H, Reik W, and Stegle O. 2016. Accurate prediction of single-cell dna methylation states using deep learning. *bioRxiv* p. 055715.

Banovich NE, Lan X, McVicker G, Van de Geijn B, Degner JF, Blischak JD, Roux J, Pritchard JK, and Gilad Y. 2014. Methylation qtls are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet* **10**: e1004663.

Barlow DP. 2011. Genomic imprinting: a mammalian epigenetic discovery model. *Annual review of genetics* **45**: 379–403.

Baylin SB and Jones PA. 2011. A decade of exploring the cancer epigenome—biological and translational implications. *Nature Reviews Cancer* **11**: 726–734.

Bergstra J and Bengio Y. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* **13**: 281–305.

Bergstra JS, Bardenet R, Bengio Y, and Kégl B. 2011. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, pp. 2546–2554.

Bestor TH. 1998. The host defence function of genomic methylation patterns. In *Novartis Found. Symp*, volume 214, pp. 187–195.

Bhasin M, Zhang H, Reinherz EL, and Reche PA. 2005. Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS letters* **579**: 4302–8.

Bird A. 2002. Dna methylation patterns and epigenetic memory. *Genes & development* **16**: 6–21.

Bock C. 2012. Analysing and interpreting dna methylation data. *Nature Reviews Genetics* **13**: 705–719.

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

Bock C, Walter J, Paulsen M, and Lengauer T. 2007. CpG island mapping by epigenome prediction. *PLoS computational biology* **3**: e110.

Breiman L, Friedman J, Stone CJ, and Olshen RA. 1984. *Classification and regression trees*. CRC press.

Consortium GP et al.. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

Das R, Dimitrova N, Xuan Z, Rollins RA, Haghighi F, Edwards JR, Ju J, Bestor TH, and Zhang MQ. 2006. Computational prediction of methylation status in human genomic sequences. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 10713–6.

Fan S, Huang K, Ai R, Wang M, and Wang W. 2016. Predicting CpG methylation levels by integrating Infinium HumanMethylation450 BeadChip array data. *Genomics* .

Fan S, Zhang MQ, and Zhang X. 2008. Histone methylation marks play important roles in predicting the methylation status of CpG islands. *Biochemical and biophysical research communications* **374**: 559–64.

Fang F, Fan S, Zhang X, and Zhang MQ. 2006. Predicting methylation status of CpG islands in the human brain. *Bioinformatics (Oxford, England)* **22**: 2204–9.

Feinberg AP. 2007. Phenotypic plasticity and the epigenetics of human disease. *Nature* **447**: 433–440.

Feltus FA, Lee EK, Costello JF, Plass C, and Vertino PM. 2003. Predicting aberrant CpG island methylation. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 12253–8.

Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, Khurana E, and Gerstein M. 2014. Funseq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome biology* **15**: 1.

Ghandi M, Lee D, Mohammad-Noori M, and Beer MA. 2014. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS computational biology* **10**: e1003711.

Gupta S, Stamatoyannopoulos JA, Bailey TL, and Noble WS. 2007. Quantifying similarity between motifs. *Genome biology* **8**: 1.

Hervouet E, Vallette FM, and Cartron PF. 2009. Dnmt3/transcription factor interactions as crucial players in targeted dna methylation. *Epigenetics* **4**: 487–499.

Hwang W, Oliver VF, Merbs SL, Zhu H, and Qian J. 2015. Prediction of promoters and enhancers using multiple dna methylation-associated features. *BMC genomics* **16**: 1.

Kaplow IM, MacIsaac JL, Mah SM, McEwen LM, Kobor MS, and Fraser HB. 2015. A pooling-based approach to mapping genetic variants associated with dna methylation. *Genome research* **25**: 907–917.

Kelley DR, Snoek J, and Rinn JL. 2016. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research* .

Kim S, Li M, Paik H, Nephew K, Shi H, Kramer R, Xu D, and Huang TH. 2008. Predicting DNA methylation susceptibility using CpG flanking sequences. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 315–26.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, and Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**: 310.

Krizhevsky A, Sutskever I, and Hinton GE. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25* (eds. F Pereira, CJC Burges, L Bottou, and KQ Weinberger), pp. 1097–1105. Curran Associates, Inc.

Le QV. 2013. Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 8595–8598. IEEE.

LeCun Y, Bengio Y, and Hinton G. 2015. Deep learning. *Nature* **521**: 436–444.

Lee HJ, Lowdon RF, Maricque B, Zhang B, Stevens M, Li D, Johnson SL, and Wang T. 2015. Developmental enhancers revealed by extensive dna methylome maps of zebrafish early embryos. *Nature communications* **6**.

Leslie R, O'Donnell CJ, and Johnson AD. 2014. Grasp: analysis of genotype–phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* **30**: i185–i194.

Lu L, Lin K, Qian Z, Li H, Cai Y, and Li Y. 2010. Predicting DNA methylation status using word composition. *Journal of Biomedical Science and Engineering* **03**: 672–676.

Ma B, Wilker EH, Willis-Owen SAG, Byun HM, Wong KCC, Motta V, Baccarelli AA, Schwartz J, Cookson WOCM, Khabbaz K, et al.. 2014. Predicting DNA methylation level across human tissues. *Nucleic acids research* **42**: 3515–28.

Marçais G and Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**: 764–770.

Martin M and Herceg Z. 2012. From hepatitis to hepatocellular carcinoma: a proposed model for cross-talk between inflammation and epigenetic mechanisms. *Genome medicine* **4**: 1.

Meissner A. 2010. Epigenetic modifications in pluripotent and differentiated cells. *Nature biotechnology* **28**: 1079–1088.

Previti C, Harari O, Zwir I, and del Val C. 2009. Profile analysis and prediction of tissue-specific cpg island methylation classes. *BMC bioinformatics* **10**: 1.

Ritchie GR, Dunham I, Zeggini E, and Flicek P. 2014. Functional annotation of noncoding sequence variants. *Nature methods* **11**: 294–296.

Sainath TN, Mohamed Ar, Kingsbury B, and Ramabhadran B. 2013. Deep convolutional neural networks for lvcsr. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 8614–8618. IEEE.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, and Salakhutdinov R. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **15**: 1929–1958.

Tieleman T and Hinton G. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning* **4**.

Tompson J, Goroshin R, Jain A, LeCun Y, and Bregler C. 2014a. Efficient object localization using convolutional networks. *arXiv preprint arXiv:1411.4280* .

Tompson JJ, Jain A, LeCun Y, and Bregler C. 2014b. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems*, pp. 1799–1807.

Wang Y, Liu T, Xu D, Shi H, Zhang C, Mo YY, and Wang Z. 2016. Predicting DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep Networks. *Scientific reports* **6**: 19598.

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al.. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**: 1431–1443.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al.. 2014. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research* **42**: D1001–D1006.

Zeng H, Edwards MD, Liu G, and Gifford DK. 2016. Convolutional neural network architectures for predicting dna–protein binding. *Bioinformatics* **32**: i121–i127.

Zhang W, Spector TD, Deloukas P, Bell JT, and Engelhardt BE. 2015. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome biology* **16**: 14.

Zheng H, Wu H, Li J, and Jiang SW. 2013. CpGIMethPred: computational model for predicting methylation status of CpG islands in human genome. *BMC medical genomics* **6 Suppl 1**: S13.

Zhou J and Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods* **12**: 931–934.

Zhou X, Li Z, Dai Z, and Zou X. 2012. Prediction of methylation CpGs and their methylation degrees in human DNA sequences. *Computers in biology and medicine* **42**: 408–13.