Published online 16 March 2017

Nucleic Acids Research, 2017, Vol. 45, No. 11 ***e99***
doi: 10.1093/nar/gkx177

Predicting the impact of non-coding variants on DNA methylation

Haoyang Zeng and David K. Gifford*

Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology Cambridge, MA 02142, USA

Received December 15, 2016; Revised February 14, 2017; Editorial Decision March 05, 2017; Accepted March 13, 2017

ABSTRACT

DNA methylation plays a crucial role in the establishment of tissue-specific gene expression and the regulation of key biological processes. However, our present inability to predict the effect of genome sequence variation on DNA methylation precludes a comprehensive assessment of the consequences of non-coding variation. We introduce CpGenie, a sequence-based framework that learns a regulatory code of DNA methylation using a deep convolutional neural network and uses this network to predict the impact of sequence variation on proximal CpG site DNA methylation. CpGenie produces allele-specific DNA methylation prediction with single-nucleotide sensitivity that enables accurate prediction of methylation quantitative trait loci (meQTL). We demonstrate that CpGenie prioritizes validated GWAS SNPs, and contributes to the prediction of functional non-coding variants, including expression quantitative trait loci (eQTL) and disease-associated mutations. CpGenie is publicly available to assist in identifying and interpreting regulatory non-coding variants.

INTRODUCTION

A significant portion of the disease and trait-associated variants revealed by genome-wide association studies (GWAS) reside in the non-coding genome where they alter cellular activities and organism phenotype by changing gene regulation (1–3). While GWAS studies can identify thousands of loci that are associated with traits, they are typically underpowered to identify the exact causal variants for a trait of interest, and further analysis of the potential functional consequence of each variant must be performed. Computational methods that analyze candidate variants for their potential contribution to a phenotype of interest are known as variant prioritization methods. Variant prioritization methods that accurately predict which variants influence proximal regulatory elements and thus gene regulation are valuable tools.

Previous variant prioritization methods have considered a diverse set of functional signals (4–8), including DNase hyper-sensitivity sites (DHS), histone modifications, and transcription factor binding. However, DNA methylation, an important epigenetic state that is involved in the regulation of key biological processes (9–14) and encodes cellular state information not contained in other epigenetic marks (15,16) has been largely overlooked. In the few methods where it is considered (5), DNA methylation is used as a low-resolution regional feature that is not allele-specific. While sequence-based methods for DNA methylation exist (17–20), there is no published method to predict the impact of sequence variants on methylation, which makes it difficult to incorporate DNA methylation in functional variant prioritization.

We introduce CpGenie (Figure 1), a deep-learning model that (i) learns a regulatory code of DNA methylation, (ii) predicts the methylation status of a CpG site from the flanking sequence at a single-nucleotide sensitivity and (iii) produces high-confidence predictions of non-coding variants that modulate DNA methylation. We find that CpGenie predicts the impact of sequence variants on DNA methylation with an accuracy that surpasses existing methods for functional variants prioritization. CpGenie also identifies the direction of impact of meQTLs that result in an allelic imbalance of DNA methylation, and prioritizes meQTLs over variants that exhibit no effect on DNA methylation with accuracy higher than alternative methods. We show that predictions from CpGenie improves the prediction of expression quantitative trait loci (eQTLs) and disease-associated variants by providing functional information complementary to other data type. In addition, 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. We provide CpGenie as open source software available at http://cpgenie.csail.mit.edu.

MATERIALS AND METHODS

CpGenie implementation

Me-CpG-prediction. We implemented a three-layer convolutional neural network with Dropout and max-norm

*To whom correspondence should be addressed. Tel: +1 617 253 6039; Email: gifford@mit.edu

© The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nar/article-abstract/45/11/e99/3072752/Predicting-the-impact-of-non-coding-variants-on
by guest
on 30 September 2017

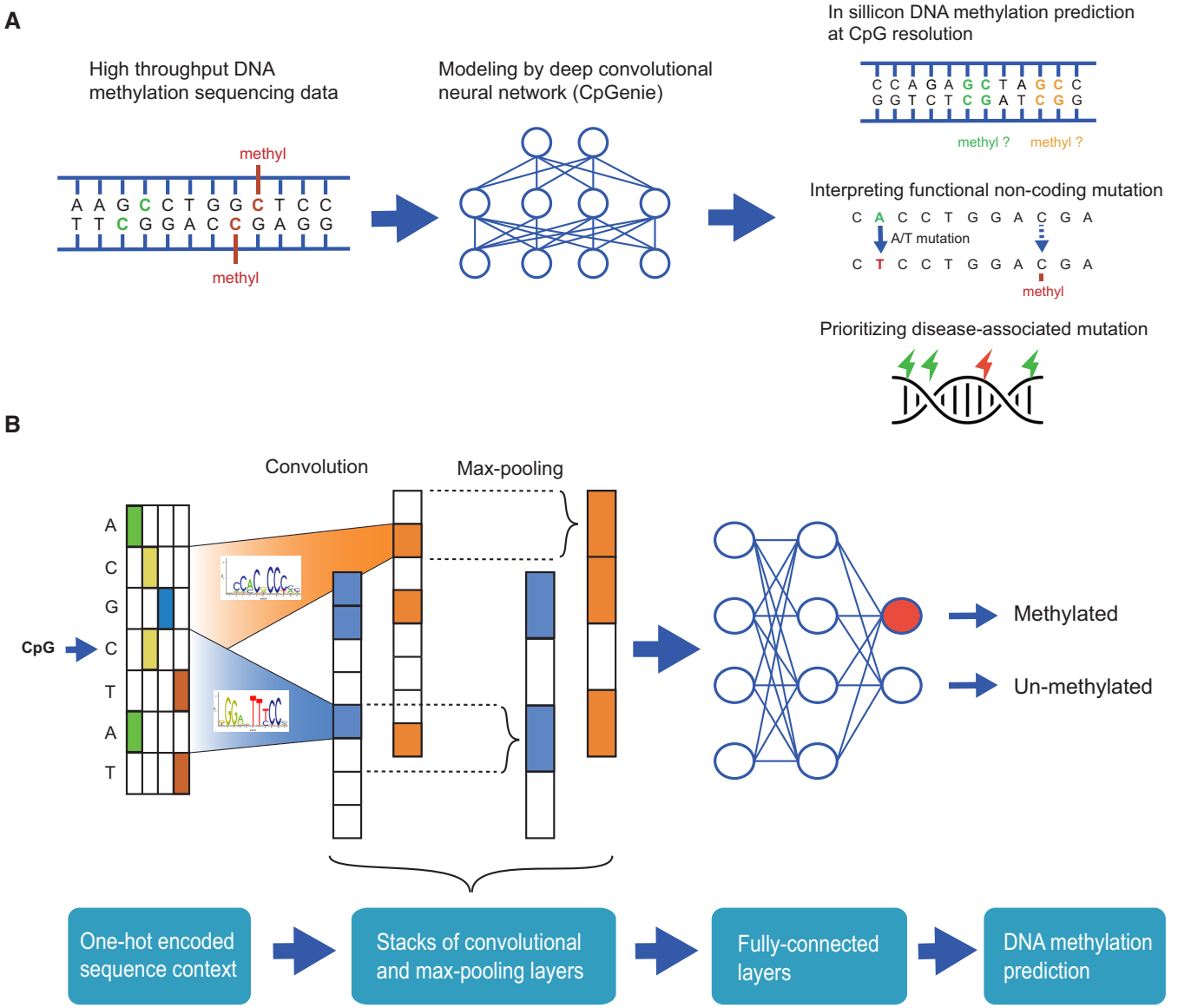

Figure 1. Schematics of CpGenie. (**A**) CpGenie takes the high-throughput DNA methylation sequencing data, such as restricted representation bisulfite sequencing (RRBS) or whole-genome bisulfite sequencing (WGBS) as input and produces predictions of CpG methylation as output. CpGenie can predict DNA 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.

regularization. Our implementation utilizes the Keras library (https://keras.io). To cope with the differences in sample sizes and protocol, we used slightly different network parameters for models trained on RRBS data from ENCODE and the pool-based bisulfite sequencing data from (21). Detailed network structure can be found in Supplementary Table S5. Hyper-parameters, such as learning rate and Dropout ratio, are tuned in a standard cross-validation fashion with test set completely held out. As input, each DNA sequence of length L is converted 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.

Variant-prediction. Given a sequence variant, we predict the methylation status of all CpGs within 500 bp with either of the variant allele. The maximum, mean and sum of the methylation level of adjacent CpGs are reported for each allele. In case where no CpG resides in the 500 bp vicinity of the given variant, a pseudo-methylation level of 0.001 is reported. We evaluate the impact of the variant by calculating the change

$$ref - alt \quad (1)$$

Downloaded from https://academic.oup.com/nar/article-abstract/45/11/e99/3072752/Predicting-the-impact-of-non-coding-variants-on
by guest
on 30 September 2017

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

$$\log \frac{ref}{1-ref} - \log \frac{alt}{1-alt} \quad (2)$$

in the mean/max methylation level of nearby CpG sites, resulting in five features for each variant.

High-throughput DNA methylation data

The 50 RRBS datasets of immortal cell lines, including GM12878, and the WGBS dataset of 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. For WGBS, the original dataset was randomly downsampled to 1 million sites due to the limited scalability of the kmer counter used in the baseline.

The raw allele-specific DNA methylation data were obtained from the authors of Kaplow *et al.* (21)(personal communication). They surveyed 823 726 SNP-CpG pairs, among which 2379 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. Test set was completely held out from training. For simplicity, only methylation levels corresponding to the reference allele were used in the training and evaluation of Me-CpG-prediction. In the analysis of variant-prediction, only meQTL and allele-specific methylation data from the held-out chromosome 10 and 11 were used.

Methylation prediction comparison with random forest

We counted the frequency of each possible 4-mer in the 1001 bp sequence centered at a CpG with JELLYFISH (22) (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).

meQTL prioritization comparison with existing methods

We downloaded DeepSEA (ver. 0.93) from http://deepsea.princeton.edu/, GWAVA (version 1.0) from http://www.sanger.ac.uk/resources/software/gwava/, deltaSVM from http://www.beerlab.org/deltasvm/, and Basset from https://github.com/davek44/Basset. For CADD (version 1.3), we used the online webserver (http://cadd.gs.washington.edu/).

For deltaSVM and Basset, the predictions of which are cell-line specific and direction-included, we used the absolute value of the predictions for the same type of cell line (LCL, lymphoblastoid cell line) from which the meQTL were discovered from. For deltaSVM, we used the gkmSVM weights trained on GM12878 DNase Hyper-sensitive Sites (DHS). For Basset, we used the absolute SAD (SNP Accessibility Difference) scores predicted for GM12878.

Functional variant prioritization

The variants in strong linkage disequilibrium with rs1427407, rs12740374, rs10737680, rs7705033 are determined by finding all variants with $r^2 = 1$ in HaploReg (http://archive.broadinstitute.org/mammals/haploreg/haploreg.php, v4.1). In the case of rs1427407 where no variants match the criteria, we find all variants with $r^2 \geq 0.8$ with it.

We obtained the eQTL and GWAS SNPs datasets, as well as their corresponding five negative sets from the supplementary tables in Zhou *et al.* (6). 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, 1400 bp, 6300 bp and 31 000 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 variant-prediction to generate DNA methylation features that describe the impact on the proximal DNA methylation levels in the 50 ENCODE RRBS dataset, resulting in 250 features for each variant. We further kept only the absolute value of each feature. As described in Zhou *et al.*, for each negative set we trained L_2-regularized logistic regression models on CpGenie and DeepSEA features respectively using scikit-learn library (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.* (6) 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 random forest implementation in scikit-learn library (http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html) in which the feature importance is calculated as 'mean decrease impurity' defined as the total decrease in node impurity averaged over all trees of the ensemble (23). To interpret the features with markedly higher importance, for each model we plotted the sorted feature importance and identified an importance-cutoff corresponding to the 'elbow-point' in the importance distribution. The top features in each model are defined as the ones with importance higher than the corresponding cutoff.

Network interpretation

We adopted a widely-used visualization method (7,24) to convert the first layer kernels to position weight matrices (PWMs). For each convolutional kernel, we searched

Downloaded from https://academic.oup.com/nar/article-abstract/45/11/e99/3072752/Predicting-the-impact-of-non-coding-variants-on
by guest
on 30 September 2017

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 ((25), version 4.11.1) to match the PWMs to known human motifs in CIS-BP database (26) with a FDR threshold of 0.1 as suggested in Kelly *et al.* (7) When combined with importance analysis, we used a more stringent FDR of 0.01. For the analysis of PWMs partially matched with known motifs, we also compared the PWMs against TransFac (27) database.

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) with respect to the input of the second layer (i.e. the output of the first max-pooling layer) under a L_2 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.

RESULTS

A two-step variant-evaluation framework for DNA methylation

CpGenie employs a two-step framework to evaluate the impact of genetic variants on DNA methylation. The first module (Me-CpG-prediction) predicts the DNA methylation status of a CpG from its flanking 1001 bp sequence context. As a fully sequence-based model, Me-CpG-prediction learns a regulatory code of DNA methylation from genomic sequence, which is essential for accurate allele-specific predictions and non-coding variant evaluation. The second module (variant-prediction) uses the regulatory code learned in Me-CpG-prediction to score the impact of a given genetic variant on proximal region. Variant-prediction predicts the methylation modulation caused by a variant by summarizing diverse statistics of the predicted methylation changes in adjacent CpG sites.

Me-CpG-prediction employs a convolutional neural network (CNN) to learn the sequence determinants for DNA methylation. Compared to random forest or support-vector machines (SVM) methods which are often used in existing frameworks (17,19,20,28,29), a CNN is able to learn more effectively from large-scale DNA methylation datasets, such as WGBS and RRBS, and is capable to learn features of different spatial and complexity scales using the hierarchical architecture.

Variant-prediction applies Me-CpG-prediction to characterize the impact of a genetic variant across a region. Variant-prediction first uses Me-CpG-prediction to score the impact of a variant compared with the corresponding wild-type allele on all CpGs within 500 bp of the variant. Variant-prediction then scores a variant's impact by the change in the sum, max and mean of methylation level in a 1001 bp genomic neighborhood around the variant when compared with the wild-type allele. These different statistics, which are often not correlated, describe different aspects of the impact and in sum produce a succinct yet informative picture of how a sequence variant alters the local methylation landscape.

CpGenie predicts DNA methylation from sequence context

We first assessed the ability of CpGenie's Me-CpG-prediction module to predict the methylation status of a CpG site from its flanking sequence. As none of the published models for fully sequence-based DNA methylation prediction (17–20) provide an standalone software for retraining and predicting on a large number of CpG sites , we compared CpGenie with a random forest (RF) classifier trained on 4-mer frequencies of the input sequence, considering that random forest and k-mer frequencies have been used in the literature of sequence-based DNA methylation models (19,20,28,29).

We evaluated Me-CpG-prediction and the random forest method on restricted representation bisulfite sequencing (RRBS) datasets from ENCODE. We trained Me-CpG-prediction and the random forest method on RRBS data from GM12878, a lymphoblastoid cell line (LCL) extensively studied in ENCODE. Me-CpG-prediction resulted in 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. Both of these metrics surpassed the performance of the random forest baseline (auROC of 0.814 and auPRC of 0.584, Figure 2A). The performance of CpGenie and random forest varies across different chromatin states in a similar way (Spearman correlation = 0.47, *P*-value = 0.079, Supplementary Figure S2A and B), with best accuracy for CpG sites in weak promoter, weak enhancer and active promoter state. Yet, we note that the performance correlates with the number of training sample in the corresponding state (Spearman correlation = 0.61, *P*-value = 0.016), with the exception of weak promoter, weak enhancer and heterochromatin state (Supplementary Figure S2C).

We then evaluated Me-CpG-prediction on 50 RRBS datasets from ENCODE (Supplementary Table S1) with our random forest baseline to systematically benchmark their capacity in predicting DNA methylation. Me-CpG-prediction robustly outperformed the alternative methods, with better auROC and auPRC for all 50 experiments (Figure 2C). We also found that a sequence window of 1001 bp optimized performance (Supplementary Figure S1), which could suggest stronger involvement of sequence features within 500 bp away in DNA methylation regulation.

We further characterized Me-CpG-prediction and the competing method using two alternative datasets: a bisulfite sequencing dataset from LCLs derived from 60 Yoruban (YRI) HapMap individuals (21), and a whole-genome bisulfite sequencing (WGBS) dataset from GM12878 cell line. On both datasets, Me-CpG-prediction achieved better performance (auROC = 0.75 and auPRC = 0.79 (Figure 2B) for the first dataset, auROC = 0.786 and auPRC = 0.784 for the second dataset (Supplementary Figure S3)) than the competing model trained and tested on the same

Downloaded from https://academic.oup.com/nar/article-abstract/45/11/e99/3072752/Predicting-the-impact-of-non-coding-variants-on
by guest
on 30 September 2017

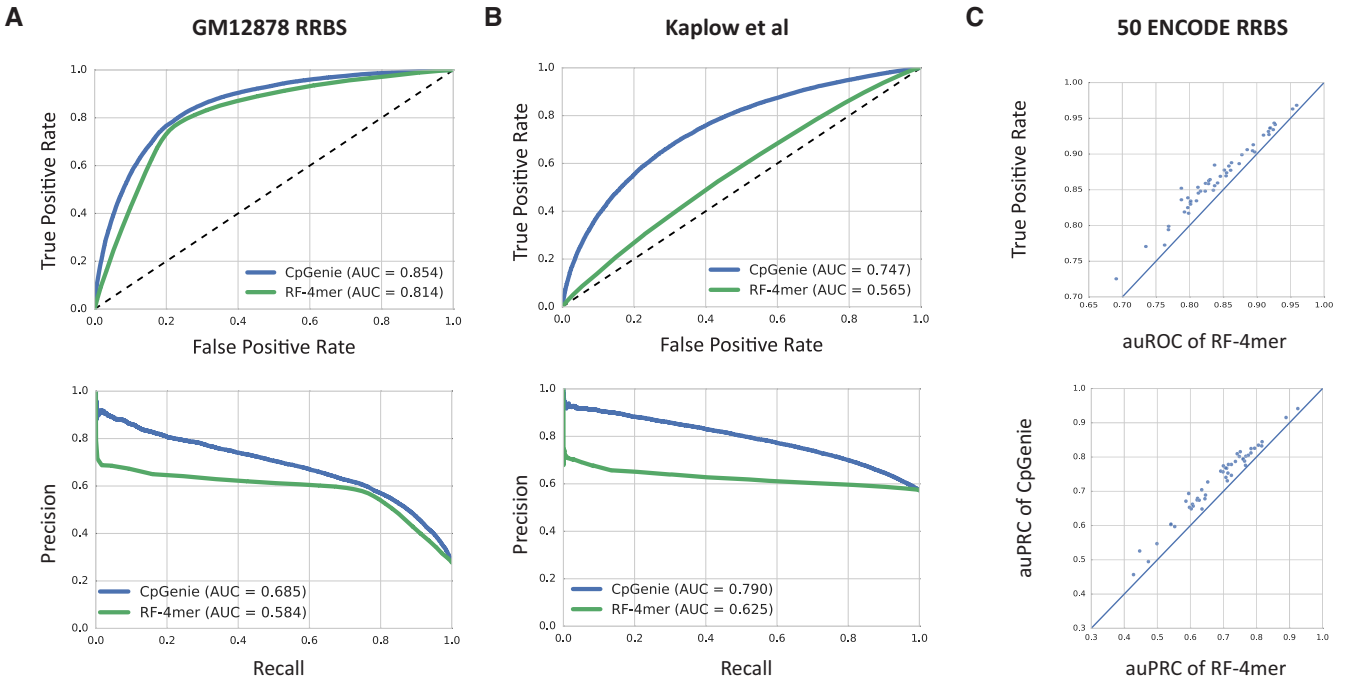

Figure 2. CpGenie predicts DNA methylation at CpG resolution. (**A**, **B**) The receiver operating characteristic (ROC) curve (top) and precision-recall (PRC) curve (bottom) of CpGenie (blue) and random forest using 4-mer counts (green) for predicting DNA methylation status of held-out CpGs in GM12878 RRBS data (A) and bisulfite sequencing data from LCLs derived from 60 Yoruban HapMap individuals (B). (**C**) Pairwise auROC (top) and auPRC (bottom) comparison of CpGenie (y-axis) and random forest using 4-mer counts (x-axis) on 50 RRBS datasets from ENCODE.

datasets (auROC = 0.57 and auPRC = 0.63 for the first dataset, auROC = 0.67 and auPRC = 0.661 for the second dataset).

CpGenie predicts the impact of functional variants on DNA methylation

We next assessed the ability of CpGenie's variant-prediction module to identify genetic variants that modulate DNA methylation. 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 (21). They found over 2000 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 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.

Variant-prediction accurately predicted the direction of allelic methylation change caused by sequence variants. When applied to the meQTLs on chromosomes held out from training, variant-prediction accurately identified the allele with higher DNA methylation level, showing sensitivity and accuracy to single-nucleotide changes (Figure 3A). 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 3B). 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 find that CpGenie variant-prediction accurately classifies variants that are meQTLs from variants that exhibit no impact on DNA methylation. Since CpGenie is the first computational method to predict meQTLs we compare it with several state-of-the-art methods for functional variant prioritization, including DeepSEA (6), Basset (7), deltaSVM (8), GWAVA (5) and CADD (4). We used 201 meQTLs on the chromosomes held out in the training of CpGenie models as positive samples. To simulate different equilibrium linkage structures we constructed three negative sample sets that are 10 times, 50 times and 100 times the size of the positive set by randomly sampling from the 76 532 non-meQTLs on the held-out chromosomes. CpGenie models trained on datasets from Kaplow *et al.* and ENCODE GM12878 RRBS datasets both surpassed the competing methods, with higher accuracy at 10% recall and larger area under precision recall curve (auPRC) when evaluated on all three negative sets (Figure 3C). Thus, CpGenie excels in predicting genetic variants that modulate DNA methylation, an important task that the state-of-the-art frameworks for functional variant prioritization fail in.

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. The basic unit of a convolutional layer is a 'kernel'

Downloaded from https://academic.oup.com/nar/article-abstract/45/11/e99/3072752/Predicting-the-impact-of-non-coding-variants-on
by guest
on 30 September 2017

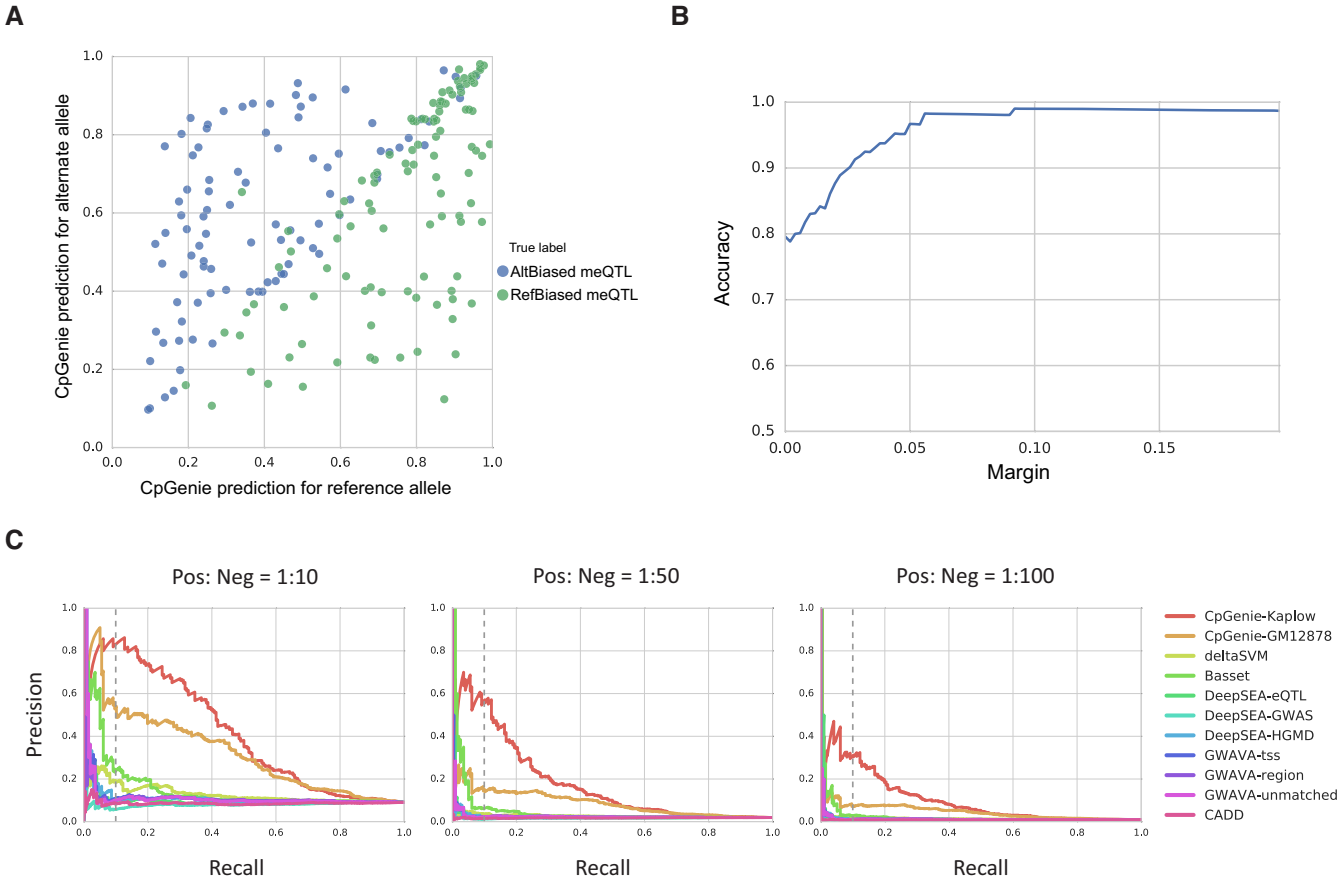

Figure 3. CpGenie accurately predicts the direction of allele-specific (AS) DNA methylation and prioritizes variants that modulate DNA methylation (meQTLs). (**A**) CpGenie's DNA methylation prediction for the reference and alternate alleles of 201 meQTLs on held-out chromosome 11 and 12. The x and y axes represent the CpGenie predicted DNA methylation level. The green and blue 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 precision-recall curve (PRC) for classifying the 201 meQTL from three different random subsets of the 76 532 non-meQTL that are 10 times (left), 50 times (middle), and 100 times (right) the size of meQTL. CpGenie outperformed all the state-of-the-art methods in functional variant prioritization with better precision at the 10% recall and higher area under precision–recall curve.

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 (7,24) . Previous studies have established that many transcription factors interact with DNA methyltransferases (DNMT) that methylate DNA (30). As transcription factors are known for binding DNA with strong sequence specificity, we transformed the first-layer convolutional kernels in CpGenie to position weight matrices (PWMs) (Methods) to determine if it learned certain of these sequence motifs.

We found that 97 out of the 128 PWMs recognized by CpGenie's Me-CpG-prediction significantly match the motifs of known transcription factors (Figure 4A, Supplementary Table S2). We found the motifs of 21 transcription factors known to strongly interact with DNA methyltransferase (DMNT) (30), including ELK1,FLI1 and E2F4. As a point of comparison, Das et al. (31) 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 YY1 and CEBPA. 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 NFKB1, MEF3 and LUN1 (Figure 4B).

Interestingly, a large number of CpGenie-discovered PWMs are variants of PAX4 and SP3 motifs (24 and 23 respectively). Hervouet *et al.* reported DNMT-interaction with other transcription factors in the same family (PAX6, SP1 and SP4). Certain 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.

Downloaded from https://academic.oup.com/nar/article-abstract/45/11/e99/3072752/Predicting-the-impact-of-non-coding-variants-on
by guest
on 30 September 2017

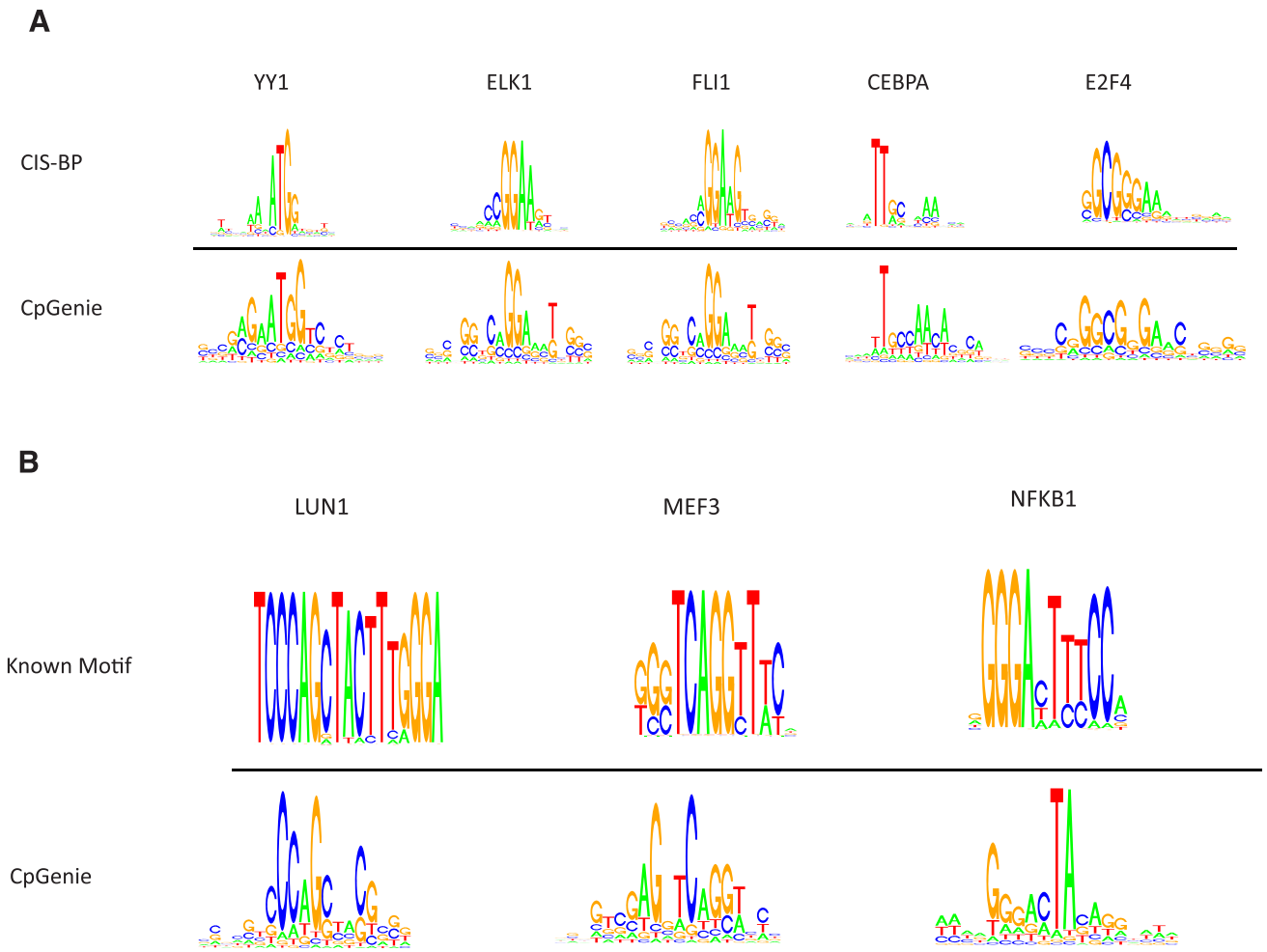

Figure 4. CpGenie learns motifs of regulatory elements involved in DNA methylation. (**A**) 97 out of 128 of the convolutional filters match motifs of known transcription factors in the human CIS-BP database at an FDR threshold of 0.1. (**B**) Examples of convolutional kernels characterizing partial information of transcription factors known for involvement in or predictive for DNA methylation. The logos for LUN1 and MEF3 were generated from motif information in TransFac databse (January 2013) and the logo for NFKB1 was generated from motif information in CIS-BP database.

We next scored the importance of CpGenie's Me-CpG-prediction 128 first-layer convolution kernels with an optimization-based framework (Materials and Methods). 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). 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 S3).

CpGenie assists in downstream analysis of functional variants

We next asked whether CpGenie, which is optimized for meQTL prediction, could shed light on the functional consequence of genetic variants associated with downstream phenotypes. We applied CpGenie on four experimentally validated GWAS SNPs, rs1427407 (fetal hemoglobin levels, (32)), rs12740374 (low-density lipoprotein (LDL) cholesterol levels,(33)), rs10737680 (age-related macular degeneration, (34)), rs7705033 (visceral adipose tissue, (35)), of which the first two have been reported to alter gene expression (BCL11A (36) and SORT1 (37) respectively) and the last two have been reported to alter DNA methylation (21). Compared with linked SNPs in strong linkage disequilibrium (Materials and Methods), all the validated SNPs were scored higher by CpGenie (Figure 5A).

We further applied CpGenie on two much larger GWAS SNPs and eQTL datasets (6), one with 78 613 eQTLs from GRASP (Genome-Wide Repository of Associations between SNPs and Phenotypes) (38) and one with 12 296 disease-associated SNPs from the US National Human Genome Research Institute's GWAS Catalog (39). For each dataset, five size-matched negative sets were constructed by sampling from different subsets of 1000 Genome Project SNPs (40). We found a simple L2-regularized logistic regression model trained on CpGenie's predictions for the 50 ENCODE RRBS datasets performed competitively in

Downloaded from https://academic.oup.com/nar/article-abstract/45/11/e99/3072752/Predicting-the-impact-of-non-coding-variants-on
by guest
on 30 September 2017

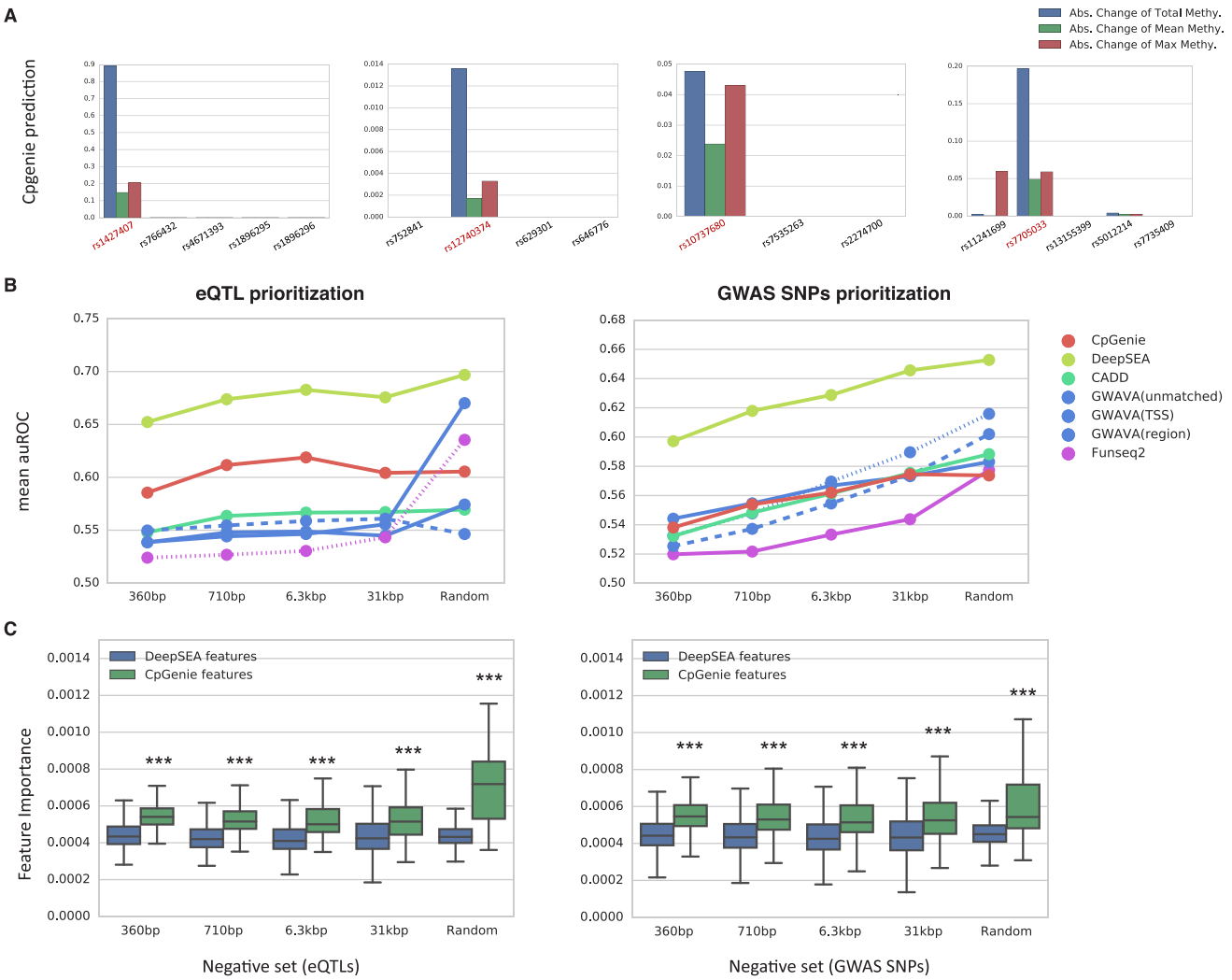

Figure 5. CpGenie's sequence-based DNA methylation predictions assist in downstream analysis of functional variants. (**A**) CpGenie scored the validated GWAS SNPs (red) higher than the SNPs in strong linkage disequilibrium. The three statistics generated from CpGenie are colored in blue (the absolute change of total methylation of proximal CpG sites), green (the absolute change of mean methylation of proximal CpG sites) and red (the absolute change of maximum methylation of proximal CpG sites). (**B**) 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. (**C**) 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).

both eQTL and GWAS SNP prioritization (Figure 5B), compared to several state-of-the-art methods including DeepSEA (6), CADD (4), GWAVA (5) and Funseq2 (41) which were all trained on more diverse sets of functional data such as histone modification, transcription factor binding, and gene expression.

To assess the relative importance of DNA methylation features in eQTL and GWAS SNPs prediction, we combined the functional features predicted from CpGenie and DeepSEA, and trained a random forest model in which feature importance can be evaluated by mean decrease impurity (Materials and Methods). DeepSEA predicts a variant's effect by producing 919 features derived from experiments of DNase hyper-sensitivity (DNase-seq), transcription factor binding (ChIP-seq), and histone modification (ChIP-seq), which is much larger and comprehensive than CpGenie's prediction of methylation alone. However, 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 5C, Supplementary Table S4). To account for the potential inflation of significance from the high correlation in the features, we further only looked at the top features with significantly higher importance (Methods). Consistently across different tasks, CpGenie-predicted methylation-based features account for a significant portion in the top features

Downloaded from https://academic.oup.com/nar/article-abstract/45/11/e99/3072752/Predicting-the-impact-of-non-coding-variants-on
by guest
on 30 September 2017

which the eQTL/GWAS predictor might actually rely on (Supplementary Table S4).

DISCUSSION

Despite the growing number of genetic variants associated with disease and complex traits by genome-wide association studies (GWAS), the identification of causal variants and their pathogenic mechanisms remains a challenge that requires predictive models for accurate interpretation of non-coding variants. CpGenie is a computational framework that is able to assess non-coding variant's effect on DNA methylation, a functional signal largely overlooked by existing models for functional variant prioritization.

With its convolutional neural network-based Me-CpG-prediction, CpGenie is able to learn sophisticated sequence determinants associated with DNA methylation efficiently from large-scale DNA methylation data generated from high-throughput bisulfite sequencing technology. CpGenie predicts the DNA methylation status of a CpG solely from the sequence context with consistently high accuracy on datasets across different cell lines, tissues, and experiment protocols (Figure 2), demonstrating crucial robustness and generalizability of the methodology. Built on convolutional neural network which is a universal computation framework given sufficient training data (42), CpGenie might also be applied to methylation data from other species to learn the species-specific regulatory code.

CpGenie demonstrated high sensitivity to single-base changes in input sequences (Figure 3A and B), a capability that enables the incorporation of allele-specific, rather than regional, DNA methylation information in the interpretation of non-coding variants. CpGenie identifies methylation quantitative trait loci (meQTLs) with a precision that surpasses the state-of-the-art frameworks for functional variants prioritization (Figure 3C), which emphasizes the unique role of CpGenie in providing a more comprehensive and accurate interpretation of the functional consequence of non-coding variants.

We have shown that CpGenie's methylation-based predictions can assist in the downstream analysis of risk-associated variants. By taking into account the influence on all the CpG sites residing in the neighborhood of the target variant, CpGenie summarizes a diverse set of statistics that can both directly instruct the identification of causal variant from candidates in strong linkage disequilibrium (Figure 5A) and be incorporated in secondary models designed to classify sequence variants associated with gene expression or complex traits (Figure 5B). Moreover, CpGenie's methylation-based features are highly favored (Figure 5C), when jointly considered with other functional annotations, such as DNase hyper-sensitivity, histone marks and transcription factor binding, by a predictive model for functional variant classification. This demonstrates the wealth of information captured in the change of DNA methylation and highlights the necessity to include allele-specific DNA methylation predictions in a comprehensive assessment of non-coding variants.

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 the functional interpretation of non-coding variants associated with complex traits and diseases. The website associated with this paper includes the complete source code for CpGenie and detailed instructions on how to use it to predict meQTLs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We acknowledge helpful input from Professor Brendan J. Frey and his lab. We are grateful for insights and suggestions from other members in Gifford Lab.

FUNDING

National Institutes of Health [R01HG008363 and U01HG007037 to D.K.G.]; NVIDIA. Funding for open access charge: National Institutes of Health [R01HG008363 and U01HG007037 to D.K.G.].
Conflict of interest statement. None declared.

REFERENCES

1. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 9362–9367.
2. Maurano,M.T., Humbert,R., Rynes,E., Thurman,R.E., Haugen,E., Wang,H., Reynolds,A.P., Sandstrom,R., Qu,H., Brody,J. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190–1195.
3. Gusev,A., Lee,S.H., Trynka,G., Finucane,H., Vilhjálmsson,B.J., Xu,H., Zang,C., Ripke,S., Bulik-Sullivan,B., Stahl,E. *et al.* (2014) Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.*, **95**, 535–552.
4. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310.
5. Ritchie,G.R., Dunham,I., Zeggini,E. and Flicek,P. (2014) Functional annotation of noncoding sequence variants. *Nat. Methods*, **11**, 294–296.
6. Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, **12**, 931–934.
7. Kelley,D.R., Snoek,J. and Rinn,J.L. (2016) Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, **26**, 990–999.
8. Lee,D., Gorkin,D.U., Baker,M., Strober,B.J., Asoni,A.L., McCallion,A.S. and Beer,M.A. (2015) A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.*, **47**, 955–961.
9. Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.
10. Bock,C. (2012) Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.*, **13**, 705–719.
11. Barlow,D.P. (2011) Genomic imprinting: a mammalian epigenetic discovery model. *Annu. Rev. Genet.*, **45**, 379–403.
12. Martin,M. and Herceg,Z. (2012) From hepatitis to hepatocellular carcinoma: a proposed model for cross-talk between inflammation and epigenetic mechanisms. *Genome Med.*, **4**, 1.
13. Meissner,A. (2010) Epigenetic modifications in pluripotent and differentiated cells. *Nat. Biotechnol.*, **28**, 1079–1088.
14. Bestor,T.H. (1998) The host defence function of genomic methylation patterns. *Novartis Found. Symp.*, **214**, 187–195.
15. Lee,H.J., Lowdon,R.F., Maricque,B., Zhang,B., Stevens,M., Li,D., Johnson,S.L. and Wang,T. (2015) Developmental enhancers revealed

Downloaded from https://academic.oup.com/nar/article-abstract/45/11/e99/3072752/Predicting-the-impact-of-non-coding-variants-on
by guest
on 30 September 2017

by extensive DNA methylome maps of zebrafish early embryos. *Nat. Commun.*, **6**, 6315.
16. Hwang,W., Oliver,V.F., Merbs,S.L., Zhu,H. and Qian,J. (2015) Prediction of promoters and enhancers using multiple DNA methylation-associated features. *BMC Genomics*, **16**, 1.
17. Bhasin,M., Zhang,H., Reinherz,E.L. and Reche,P.A. (2005) Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Lett.*, **579**, 4302–4308.
18. Kim,S., Li,M., Paik,H.-i.H., Nephew,K.P., Shi,H., Kramer,R., Xu,D. and Huang,T.H.-M. (2008) Predicting DNA methylation susceptibility using CpG flanking sequences. *Pacific Symp. Biocomput. Citeseer*, **13**, 315–326.
19. Lu,L., Lin,K., Qian,Z., Li,H., Cai,Y. and Li,Y. (2010) Predicting DNA methylation status using word composition. *J. Biomed. Sci. Eng.*, **3**, 672.
20. Zhou,X., Li,Z., Dai,Z. and Zou,X. (2012) Prediction of methylation CpGs and their methylation degrees in human DNA sequences. *Comput. Biol. Med.*, **42**, 408–413.
21. Kaplow,I.M., MacIsaac,J.L., Mah,S.M., McEwen,L.M., Kobor,M.S. and Fraser,H.B. (2015) A pooling-based approach to mapping genetic variants associated with DNA methylation. *Genome Res.*, **25**, 907–917.
22. Marçais,G. and Kingsford,C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.
23. Breiman,L., Friedman,J., Stone,C.J. and Olshen,R.A. (1984) *Classification and Regression Trees*, CRC Press.
24. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
25. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, 1.
26. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.
27. Matys,V., Fricke,E., Geffers,R., Gößling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
28. Zhang,W., Spector,T.D., Deloukas,P., Bell,J.T. and Engelhardt,B.E. (2015) Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biol.*, **16**, 14.
29. Fan,S., Huang,K., Ai,R., Wang,M. and Wang,W. (2016) Predicting CpG methylation levels by integrating Infinium HumanMethylation450 BeadChip array data. *Genomics*, **107**, 132–137.
30. Hervouet,E., Vallette,F.M. and Cartron,P.-F. (2009) Dnmt3/transcription factor interactions as crucial players in targeted DNA methylation. *Epigenetics*, **4**, 487–499.
31. Das,R., Dimitrova,N., Xuan,Z., Rollins,R.A., Haghighi,F., Edwards,J.R., Ju,J., Bestor,T.H. and Zhang,M.Q. (2006) Computational prediction of methylation status in human genomic sequences. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 10713–10716.
32. Mtatiro,S.N., Singh,T., Rooks,H., Mgaya,J., Mariki,H., Soka,D., Mmbando,B., Msaki,E., Kolder,I., Thein,S.L. *et al.* (2014) Genome wide association study of fetal hemoglobin in sickle cell anemia in Tanzania. *PLoS One*, **9**, e111464.
33. Kathiresan,S., Willer,C.J., Peloso,G.M., Demissie,S., Musunuru,K., Schadt,E.E., Kaplan,L., Bennett,D., Li,Y., Tanaka,T. *et al.* (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.*, **41**, 56–65.
34. AMD Gene Consortium (2013) Seven new loci associated with age-related macular degeneration. *Nat. Genet.*, **45**, 433–439.
35. Fox,C.S., Liu,Y., White,C.C., Feitosa,M., Smith,A.V., Heard-Costa,N., Lohman,K., Johnson,A.D., Foster,M.C., Greenawalt,D.M. *et al.* (2012) Genome-wide association for abdominal subcutaneous and visceral adipose reveals a novel locus for visceral fat in women. *PLoS Genet.*, **8**, e1002695.
36. Bauer,D.E., Kamran,S.C., Lessard,S., Xu,J., Fujiwara,Y., Lin,C., Shao,Z., Canver,M.C., Smith,E.C., Pinello,L. *et al.* (2013) An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science*, **342**, 253–257.
37. Musunuru,K., Strong,A., Frank-Kamenetsky,M., Lee,N.E., Ahfeldt,T., Sachs,K.V., Li,X., Li,H., Kuperwasser,N., Ruda,V.M. *et al.* (2010) From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, **466**, 714–719.
38. Leslie,R., O'Donnell,C.J. and Johnson,A.D. (2014) GRASP: analysis of genotype–phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics*, **30**, i185–i194.
39. 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 Res.*, **42**, D1001–D1006.
40. 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
41. Fu,Y., Liu,Z., Lou,S., Bedford,J., Mu,X.J., Yip,K.Y., Khurana,E. and Gerstein,M. (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.*, **15**, 1.
42. Zeng,H., Edwards,M.D., Liu,G. and Gifford,D.K. (2016) Convolutional neural network architectures for predicting DNA–protein binding. *Bioinformatics*, **32**, i121–i127.

Downloaded from https://academic.oup.com/nar/article-abstract/45/11/e99/3072752/Predicting-the-impact-of-non-coding-variants-on
by guest
on 30 September 2017