Zou et al. BMC Genomics (2018) 19:390
https://doi.org/10.1186/s12864-018-4766-y

BMC Genomics

METHODOLOGY ARTICLE Open Access

BoostMe accurately predicts DNA methylation values in whole-genome bisulfite sequencing of multiple human tissues

Luli S. Zou[1], Michael R. Erdos[1], D. Leland Taylor[1,2], Peter S. Chines[1], Arushi Varshney[3], The McDonnell Genome Institute[4], Stephen C. J. Parker[3,5], Francis S. Collins[1*†] and John P. Didion[1†]

Abstract

Background: Bisulfite sequencing is widely employed to study the role of DNA methylation in disease; however, the data suffer from biases due to coverage depth variability. Imputation of methylation values at low-coverage sites may mitigate these biases while also identifying important genomic features associated with predictive power.

Results: Here we describe BoostMe, a method for imputing low-quality DNA methylation estimates within whole-genome bisulfite sequencing (WGBS) data. BoostMe uses a gradient boosting algorithm, XGBoost, and leverages information from multiple samples for prediction. We find that BoostMe outperforms existing algorithms in speed and accuracy when applied to WGBS of human tissues. Furthermore, we show that imputation improves concordance between WGBS and the MethylationEPIC array at low WGBS depth, suggesting improved WGBS accuracy after imputation.

Conclusions: Our findings support the use of BoostMe as a preprocessing step for WGBS analysis.

Keywords: DNA methylation, XGBoost, Whole-genome bisulfite sequencing (WGBS), EPIC, Imputation, Adipose, Skeletal muscle, Pancreatic islets

Background

DNA methylation is an epigenetic mark that is known to play a role in many fundamental biological processes, including differentiation, development, and gene regulation [1, 2]. In mammals, DNA methylation occurs primarily on cytosines of CG dinucleotides (CpGs). CpG methylation marks convey epigenetic information across the lifespan, as they can be stably propagated through mitosis, and in special circumstances even through meiosis [3–7]. DNA methylation is an important mechanism for gene-environment interaction, and can thus influence health of cells, organs, and organisms.

DNA methylation is most commonly measured in cell lines or bulk tissue samples using microarrays or sequencing of bisulfite-converted DNA. These assays provide an estimate of the fraction of chromosomes in the cell population that are methylated at each CpG of interest (“beta” values). Microarrays such as the Illumina Infinium Methylation450k, and more recently the MethylationEPIC [8] (“EPIC”), allow beta value estimation using fluorescent probe technology, and are cost-effective platforms for measuring methylation in genes, promoters, and enhancers. However, the EPIC array measures only ~ 3% of all CpGs in the genome and has relatively little coverage of intergenic regions. In contrast, whole-genome bisulfite sequencing (WGBS) provides coverage of most of the ~ 28 million CpGs in the genome of an average tissue, giving it a clear advantage over the EPIC array. However, due to the high cost and sample input requirements of WGBS, it is often infeasible to generate deep-coverage data for a large number of replicates. Since beta values are estimated from

* Correspondence: collinsf@od.nih.gov
†Francis S. Collins and John P. Didion contributed equally to this work.
[1] National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA
Full list of author information is available at the end of the article

WGBS by calculating the ratio of methylated and unmethylated reads at each CpG, CpGs sequenced at low read depth are subject to error and are typically removed or imputed before performing downstream analyses, such as detecting differentially methylated regions [9, 10]. One potential remedy for inefficiencies with WGBS is the generation of a small number of high-coverage reference samples in relevant tissues and disease states. These reference samples could be used to facilitate lower coverage and/or lower density methylation profiling in a larger number of samples. Similar techniques have already been used to increase the power of GWAS studies by leveraging data from sparse yet cost-effective SNP arrays [11–13].

Machine and deep learning algorithms have shown promise in providing accurate beta value estimates after training on sparse data sets [14, 15]. Prediction accuracy, however, is still far from what is achieved with SNP imputation in GWAS [14], leading to the need for algorithm improvement. The most recent beta value imputation methods were based on either random forests [16] or deep neural networks [17]. A relatively new algorithm called extreme gradient boosting (XGBoost) has been shown to outperform both methods in accuracy and computational efficiency when highly predictive features can be constructed [18]. Previous imputation methods have also only classified beta values as fully unmethylated or methylated. This binarization of the data not only represents a loss of information but also ignores the possible significance of intermediate beta values as a conserved and biologically relevant genomic signature [19]. Specifically, although single chromosome methylation is binary, intermediate methylation in a population of cells has been shown to be a predominantly tissue specific signature that is enriched in genes, enhancers, and evolutionarily conserved regions [19, 20]. Furthermore, although previous algorithms have constructed features that capture the local correlation structure of beta values [14, 21] as well as information from the surrounding DNA sequence context [14, 15, 22–24], no algorithms have created features that incorporate information from multiple samples in the same tissue and/or disease state. This adaptation could improve prediction for CpGs that are not highly correlated to neighboring CpGs or strongly associated with their surrounding DNA context.

Importantly, machine and deep learning algorithms not only can impute missing values in sparse methylation data sets but can also identify genomic features and sequence motifs associated with methylation patterns in different tissues [14, 15, 22, 25–27]. For example, a previous random forest [14] algorithm applied to whole blood identified co-localized active transcription factor binding sites (TFBS), including those for ELF1, MAZ, MXI1, and RUNX3, to be predictive of beta values in whole blood. A recent deep learning algorithm [15] found that transcription factor motifs such as Foxa2 and Srf, which are both implicated in cell differentiation and embryonic development, were important to beta value prediction in mouse embryonic stem cells. These algorithms are therefore useful for characterizing methylation regulatory networks.

Methylation regulatory networks may have particular significance in complex diseases such as type 2 diabetes (T2D). The complexity of T2D is characterized by interactions between genetic and environmental factors acting in multiple tissues over time. Implicated tissues include pancreatic islets, skeletal muscle, adipose, liver, intestine, and brain. Genome-wide association studies (GWAS) have shown that the majority of T2D-associated loci lie in non-coding regions of the genome [28–30]. These loci therefore lack a clear relationship with any potential causal genes, underscoring the importance of identifying the epigenetic mechanisms by which they could affect gene expression.

In this work, we generated EPIC and WGBS data on 58 human samples from adipose, skeletal muscle, and pancreatic islets (Additional file 1: Table S1). Samples from adipose and skeletal muscle included those from patients with normal glucose tolerance (NGT) and T2D; all pancreatic islet samples were NGT. We found 1) a high rate of missingness in the WGBS data and 2) discordance between WGBS and EPIC, biased towards low coverage and intermediate methylation sites. To address these issues, we developed an imputation method based on XGBoost called BoostMe, which is designed to leverage information from multiple independent samples from the same tissue type and disease state to impute low-coverage CpGs in WGBS data. We find that, for all tissues and all genomic contexts, BoostMe outperforms other methods, achieving the lowest error as well as the highest computational efficiency. We also examine the effect of imputation on WGBS accuracy by comparing raw WGBS and imputed beta values to those of the EPIC array. We find that discordance between EPIC and WGBS measurements at low WGBS depth is mitigated after imputation using BoostMe, supporting the use of imputation as an important preprocessing step for WGBS data analyses.

Results

Characterizing beta values in WGBS of adipose, skeletal muscle, and pancreatic islets

We generated WGBS and EPIC data from 58 samples of human adipose, skeletal muscle, and pancreatic islets. We discovered that, despite deep mean sequence coverage across samples (~ 30× genome-wide), there was a relatively high rate of missingness in DNA methylation (beta) values (CpG sequencing depth < 10×) (Fig. 1a). The number of missing beta values across all samples ranged from 2.6 million to 10.5 million, or roughly 10 to 40% of all ~ 25.5 million autosomal CpGs. We next explored the overlap between beta value missingness and the underlying tissue-specific epigenomic architecture. We used previously published chromatin state segmentations for the corresponding tissues [31]. We found that missingness was spread across chromatin states (Fig. 1b) , with the highest raw numbers of missing beta values located in the Quiescent/Low Signal and Weak Transcription states (Additional file 1: Figure S1).

Previous imputation work using array-based technology has shown that the beta value of a CpG is correlated with the beta values of its neighboring CpGs [14]. The average distance between neighboring CpGs in the human genome is 50 bp (Additional file 1: Figure S2); however, there is a high degree of variance in inter-CpG distance among chromatin states. To determine the extent to which neighboring beta values in WGBS are correlated, we quantified neighboring CpG similarity as a function of distance by calculating pairwise differences between beta values within chromatin states for each tissue and disease state combination (Fig. 2, Additional file 1: Figure S3, S4). The majority (~ 70%) of CpG pairs genome-wide were highly similar, with an absolute difference in beta values less than 0.1 (Fig. 2a). As distance between CpGs increased, chromatin states such as active and bivalent/poised transcription start site (TSS), strong and weak transcription, and quiescent/low signal had generally low differences, suggesting that neighboring beta values may be highly informative for prediction in these regions. In contrast, enhancer, flanking TSS, and repressed polycomb states exhibited larger differences as distance increased, suggesting that neighboring information alone may not be enough to make accurate predictions in these states, particularly when the nearest neighboring CpG is located at some distance.

Since ~ 70–80% of CpGs are invariantly methylated across tissues and samples [32], we also calculated pairwise differences within regions of higher across-tissue variance. In contrast to the bimodal distribution of beta values genome-wide, average beta values in these high-variance blocks were more highly enriched for intermediate values (Additional file 1: Figure S5). Pairwise differences within these blocks exhibited less drastic changes over distance compared to the genome-wide analysis and were more similar across chromatin states (Fig. 2b). However, they had slightly larger magnitudes at low distances, where the bulk of differences occurred, indicating that even proximal CpGs may be less informative in these regions.

BoostMe outperforms random forests and DeepCpG for methylation imputation

To address the high rate of missingness in our data, we developed BoostMe, a method for imputing beta values using WGBS data from at least three samples. Previous attempts at beta value imputation based on penalized functional regression [24], random forests [14], and deep neural networks [15] yielded relatively poor predictive accuracy genome-wide (RMSE > 0.23, AUROC < 0.93) (Additional file 1: Table S2). To improve on those methods, we implemented predictive models optimized for WGBS data using both random forest and gradient boosting [18] algorithms.

We constructed a total of 648 features designed to both parallel and improve upon previous work [14]. Prediction features constructed from the WGBS data included the nearest non-missing neighboring CpG beta values upstream and downstream of the CpG of interest,

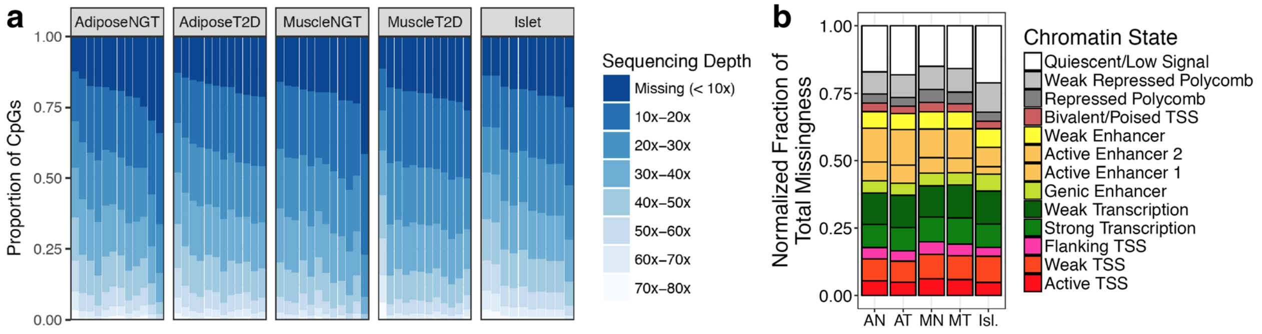

Fig. 1 Characterization of WGBS missingness. (**a**) WGBS coverage (sequencing depth) distributions across tissue types and samples, visualized as the proportion of the ~ 25.5 million autosomal CpGs lying in each coverage interval. Each column is one sample from that tissue type. (**b**) WGBS missingness is distributed across chromatin states in all tissues. The normalized fraction of total missingness (y-axis) was calculated as the number of CpGs in each chromatin state that had missing methylation (beta) values (sequencing depth < 10×) normalized by the total number of CpGs in that chromatin state for each tissue. Abbreviations: AN, adipose NGT; AT, adipose T2D; MN, muscle NGT; MT, muscle T2D; Isl., islets; NGT, normal glucose tolerance; T2D, type 2 diabetes

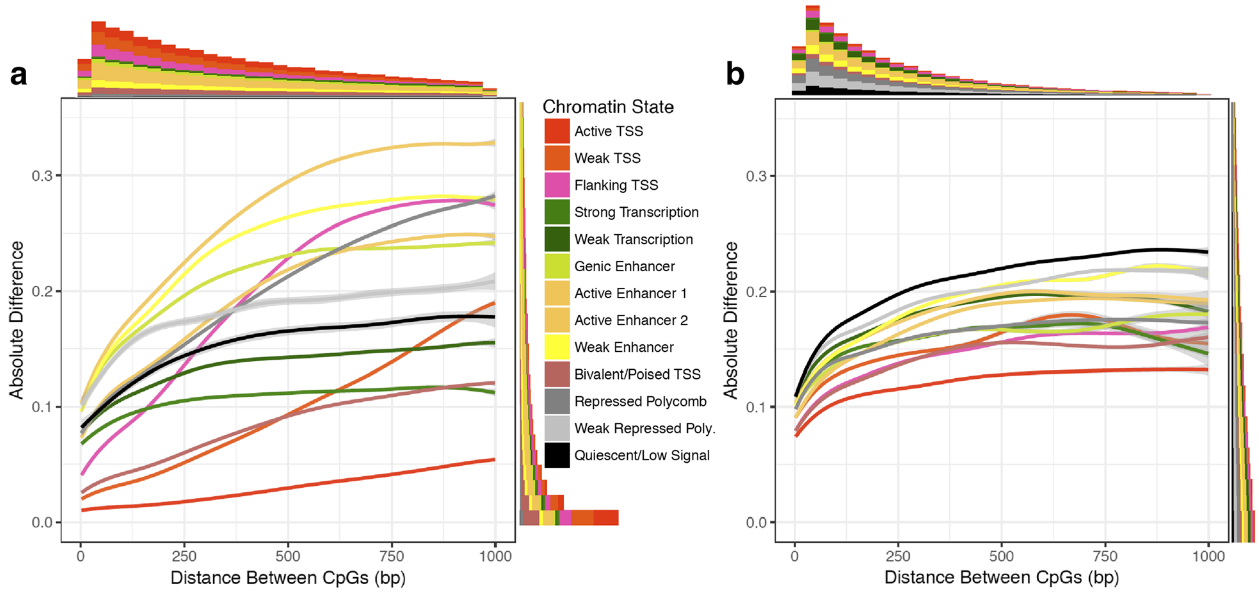

Fig. 2 CpG methylation pairwise differences increase with distance and have different average behaviors across chromatin states. Differences were calculated using the average methylation (beta) value of each CpG across all 12 muscle NGT samples. Smoothed lines were fit using a generalized additive model. Average behaviors across chromatin states were similar across all tissue and disease state combinations examined in this work; see Additional file 1: Figure S3, S4. (**a**) Absolute pairwise differences within chromatin states genome-wide. We randomly sampled blocks of chromatin states genome-wide and used ~ 2 million pairwise differences for each chromatin state. Marginal histograms indicate the number of pairwise differences across chromatin states within the range of the graph. (**b**) Pairwise differences in blocks of consecutive CpGs with higher across-tissue variance. Variances were calculated across all 58 tissue samples; CpGs with beta value variances above the third quartile of variances were considered as having higher variances. Pairs of CpGs lying in the same high-variance region but different chromatin states were excluded from this analysis

base-pair distance to the neighboring CpGs, and the average beta value of the CpG of interest in other samples from the same tissue and disease state (sample average). We also used tissue-specific reference data to create features that describe the genomic context of individual CpGs such as histone marks ($n = 7$), computational predictions of transcription factor binding sites (TFBSs) ($n = 608$), chromatin states ($n = 13$), and ATAC-Seq peaks (as a measure of DNA accessibility; see Methods, Additional file 1: Table S3 for a full list of features).

We tested the inclusion of different features and selected the best combination by assessing performance on a held-out validation set (Table 1). We found that the

Table 1 Performance of BoostMe using different feature combinations

Features	RMSE (all)	RMSE (int.)	AUROC	AUPRC	Accuracy
Nearest non-missing upstream and downstream neighboring beta values and distances (N)	0.15046 ± 0.00940	0.21429 ± 0.01133	0.94983 ± 0.00774	0.98595 ± 0.00337	0.93743 ± 0.01065
Sample average (A)	0.09594 ± 0.00478	0.14304 ± 0.00649	0.98954 ± 0.00177	0.99424 ± 0.00486	0.96237 ± 0.00464
A, N	0.09330 ± 0.00461	0.13768 ± 0.00620	0.99019 ± 0.00160	0.99769 ± 0.00049	0.96389 ± 0.00457
A, N, transcription factor binding sites	0.09333 ± 0.00459	0.13776 ± 0.00617	0.99018 ± 0.00159	0.99769 ± 0.00049	0.96384 ± 0.00457
A, N, recombination rate	0.09330 ± 0.00462	0.13774 ± 0.00621	0.99018 ± 0.00160	0.99769 ± 0.00049	0.96386 ± 0.00459
A, N, ATAC-seq peaks (P)	0.09327 ± 0.00461	0.13768 ± 0.00620	0.99019 ± 0.00160	0.99769 ± 0.00049	0.96389 ± 0.00457
A, N, histone marks (H)	0.09322 ± 0.00461	0.13758 ± 0.00619	0.99020 ± 0.00159	0.99769 ± 0.00049	0.96393 ± 0.00456
A, N, GENCODE annotations (G)	0.09323 ± 0.00461	0.13759 ± 0.00619	0.99019 ± 0.00159	0.99769 ± 0.00049	0.96390 ± 0.00457
A, N, chromatin states (C)	0.09318 ± 0.00461	0.13759 ± 0.00619	0.99019 ± 0.00159	0.99769 ± 0.00049	0.96390 ± 0.00457
A, N, P, H, G, C[a]	0.09311 ± 0.00459	0.13735 ± 0.00616	0.99022 ± 0.00158	0.99770 ± 0.00049	0.96401 ± 0.00454

Feature selection performance was evaluated on holdout validation sets by repeating the training and validation process ten times using ten different random seeds. All metrics were calculated by averaging across all 58 samples and are displayed as mean ± standard deviation. RMSE, root-mean-squared error; int., intermediate beta values, defined as having a sample average between 0.2 and 0.8; AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve. Accuracy was calculated as the number of beta values correctly predicted as methylated or unmethylated divided by the total number of beta values. [a]Final set of features used to benchmark performance

RMSE obtained using sample average alone was within one standard deviation of the RMSE obtained using sample average and neighbor CpG features combined; however, the AUPRC of the latter was higher and had a standard deviation that was approximately one order of magnitude smaller. We also found that some genomic features had negligible effect on accuracy. In particular, TFBS and recombination rate did not improve performance. The lowest RMSE was obtained when using the sample average, neighboring beta values and distances, ATAC-seq peaks, histone marks, GENCODE annotations, and chromatin states. Although using this set of features provided the best RMSE, improvements were within one standard deviation of RMSE using only sample average and neighboring CpG features. This result suggests that comparably low RMSE can be achieved without using additional genomic features, though such features, if available, may provide modest improvements in performance.

Using the full set of features, and after applying additional quality control exclusion criteria (Methods), the average number of CpGs usable for training and testing per sample was 20 million (range: 14.7 million - 21.2 million), and the average number of missing CpGs able to be imputed per sample (sequencing depth < 10×) was 2.6 million (range: 750,000–7.7 million).

We compared the performance of BoostMe and random forests with DeepCpG for predicting continuous beta values in WGBS data from adipose NGT ($n = 12$), adipose T2D (n = 12), muscle NGT (n = 12), muscle T2D (n = 12), and pancreatic islet ($n = 10$) tissue (Fig. 3a). Due to memory limits, we replicated a previously-described form of repeated random subsample validation [14] for both BoostMe and random forests, training on 1,000,000

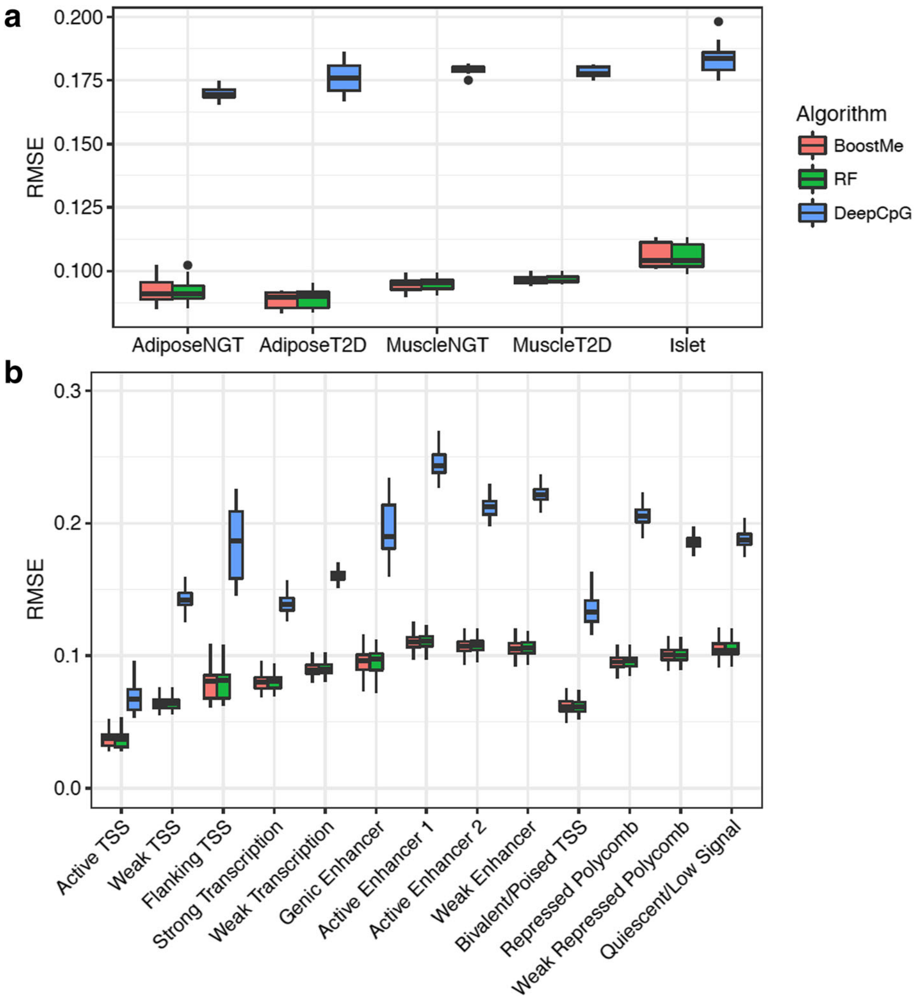

Fig. 3 BoostMe and random forests outperform DeepCpG for predicting methylation values genome-wide. (**a**) Root-mean-squared error (RMSE) of BoostMe, random forests (RF) and DeepCpG for predicting methylation (beta) values in all tissue and disease state combinations examined in this study. Data points represent performance on individual samples. NGT, normal glucose tolerance; T2D, type 2 diabetes. (**b**) RMSE of all algorithms by chromatin state

randomly selected CpGs, validating on a hold-out set of 1,000,000 CpGs, and testing on a hold-out set of 1,000,000 CpGs (Methods). DeepCpG was trained for each tissue and disease state combination as described in Angermueller et al. [15], using a total of ~ 10 million CpGs for training and ~ 5 million for validation. We evaluated DeepCpG models on a held-out random sample of 1,000,000 CpGs that also fit the BoostMe criteria for training and testing. Unlike BoostMe and random forests, DeepCpG learns features from the DNA sequence and neighboring CpGs surrounding the CpG of interest and does not explicitly use manually constructed features. We found that both BoostMe and random forests outperformed DeepCpG, achieving an average root-mean-squared error (RMSE) of 0.09, area under the receiver operating characteristic curve (AUROC) of 0.99, area under the precision-recall curve (AUPRC) of 0.99, and an accuracy of 0.96 (Table 2). This result held true when training all methods on a smaller sample of 500,000 CpGs (Additional file 1: Table S4). Unlike previous methods [14, 15], we trained on continuous beta values rather than binary values because of the available depth in our WGBS data. We found that this change improved overall RMSE by at least 0.06 and performed similarly for AUROC, AUPRC, and accuracy (Additional file 1: Table S5).

To characterize performance patterns in different genomic contexts, we compared the performance of each algorithm within tissue-specific chromatin states (Fig. 3b). Again, BoostMe and random forests outperformed DeepCpG in all chromatin states. In addition, all three algorithms exhibited the same trend across chromatin states, with the best predictive performance in TSS-associated states, which were strongly correlated with low beta values (Additional file 1: Figure S6).

Beta values are bimodally distributed, with the majority of CpGs being either fully methylated or unmethylated; however, there is evidence that intermediate methylated CpGs are a conserved genomic signature that is often tissue-specific [19, 20]. Furthermore, given our finding that regions of higher across tissue-variance tended to have average beta values in the intermediate range (Additional file 1: Figure S5), CpGs with intermediate average beta values may be more biologically significant, and therefore more important to predict accurately. Therefore, we also benchmarked all algorithms on intermediate beta values, defined as having a sample average methylation between 0.20 and 0.80 inclusive. We found similar trends in algorithm performance, with both BoostMe and random forests having an RMSE of 0.13 and DeepCpG an RMSE of 0.23 (Table 2).

Because CpG methylation is an example of extreme class imbalance (only ~ 28% of CpGs have intermediate methylation), we hypothesized that artificially creating a uniform distribution of beta values in our training set would improve prediction at CpGs with intermediate beta values. We tested this by generating a training set using a biased sampling procedure that drew CpGs from each beta value decile with a frequency inversely proportional to its size. Contrary to our expectation, we found that this sampling procedure did not improve significantly the performance of BoostMe (Additional file 1: Figure S7).

We further examined BoostMe error as a function of distance to the nearest non-missing CpG across chromatin states, both genome-wide (Fig. 4a) and within regions of higher across-tissue variance (Fig. 4b). Trends in RMSE across distance strongly paralleled our previous analysis of pairwise differences in CpG methylation within chromatin states (Fig. 2). As expected, the absolute prediction error was lowest for all chromatin states when there was a non-missing, neighboring CpG within 100 bp of the CpG of interest, which was true for the majority (~ 87%) of CpGs. The error increased for the smaller subset of CpGs where the nearest non-missing neighbor was farther away to varying degrees for each chromatin state: error in TSS states increased rapidly; transcribed states (strong transcription, weak transcription) remained relatively stable and low; and enhancer and inactive chromatin states had higher but generally stable error rates. Similar to the pairwise differences within regions of high across-tissue variance (Fig. 2b), all chromatin states in these regions exhibited stably higher error rates and had similar behaviors. Due to the small average block size for regions of high across-tissue variance (~ 533 bp on average), there was a lack of data past 200 bp, which led to larger confidence intervals and less accurate smoothed line estimates.

Finally, we benchmarked the computational performance of all algorithms. BoostMe had a training runtime

Table 2 Genome-wide performance of different algorithms on predicting methylation values, averaged across tissues and samples

Algorithm	RMSE (all)	RMSE (int.)	AUROC	AUPRC	Accuracy	Resources	Time (hrs)
BoostMe	0.09 ± 0.005	0.13 ± 0.006	0.99 ± 0.002	0.99 ± 0.0005	0.96 ± 0.005	16 CPUs	0.50 ± 0.15
Random Forests	0.09 ± 0.005	0.13 ± 0.006	0.99 ± 0.002	0.99 ± 0.0005	0.96 ± 0.005	16 CPUs	14 ± 2
DeepCpG	0.17 ± 0.007	0.27 ± 0.014	0.94 ± 0.003	0.98 ± 0.002	0.91 ± 0.010	1 GPU	140 ± 30

RMSE, root-mean-squared error; int., intermediate beta values, defined as having a sample average methylation between 0.2 and 0.8; AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve. Time is the average number of computational hours it took to train on all samples within a tissue

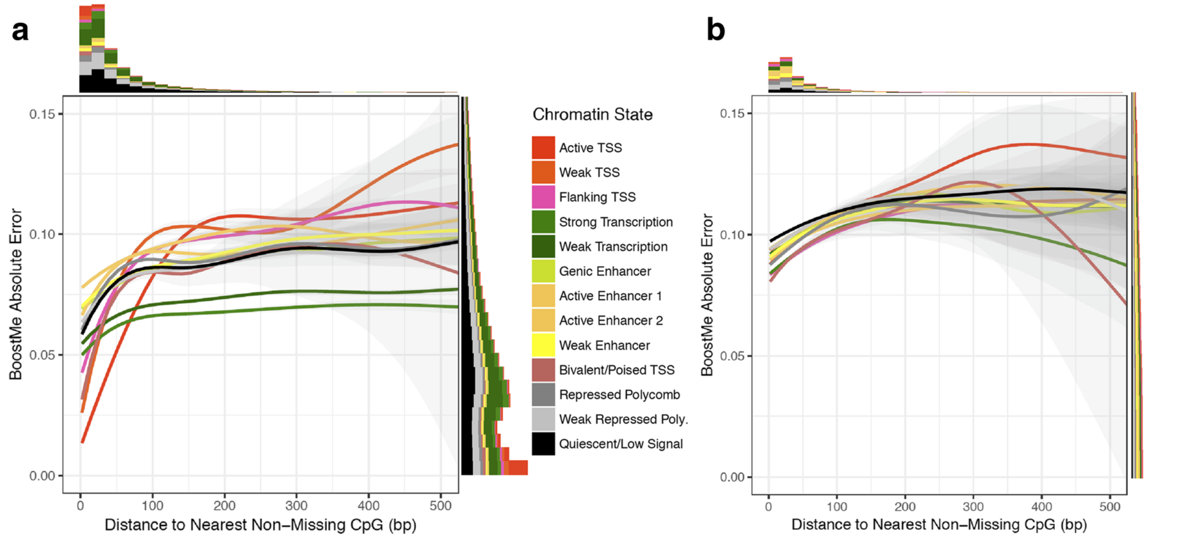

Fig. 4 BoostMe error across chromatin states as a function of distance to the nearest non-missing CpG. Absolute prediction error was measured as the difference between the predicted and actual methylation (beta) values for (**a**) a holdout test set of 5 million CpGs and (**b**) a holdout set of ~ 1.3 million CpGs in high-variance regions that met BoostMe criteria for training and testing. Gray shaded areas indicate confidence intervals for each smoothed line created using a generalized additive model. Marginal histograms display the distribution of predicted CpGs across chromatin states within the range of the graph

that was up to 28× faster than random forests using identical computational resources and up to 280× faster than DeepCpG (Table 2). Both BoostMe and random forests training times outperformed DeepCpG, the latter of which took multiple days due to the need to train CpG and DNA modules separately before training the joint module (Methods).

Imputation reduces WGBS discordance with EPIC at low sequencing depth

To assess the effect of imputation on the quality of WGBS data, we first characterized the concordance of WGBS and EPIC array beta estimates at the same CpGs in the same samples (Fig. 5a). As reported previously [33], WGBS and EPIC beta values were

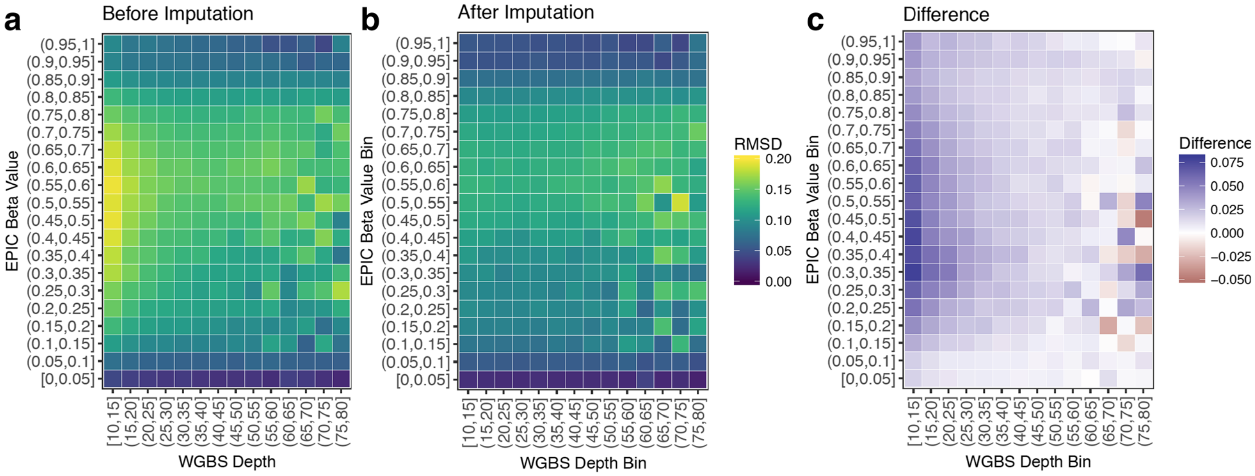

Fig. 5 Imputation reduces discordance between WGBS and EPIC methylation estimates at low sequencing depth. (**a**) Root-mean squared discordance (RMSD) between EPIC and WGBS methylation (beta) estimates at CpGs common between the two platforms. X-axis: depth at which the CpG was sequenced, binned into intervals of 5× coverage. Y-axis: the beta value of the CpG as measured by the EPIC array, binned into intervals of 0.05. Yellow color indicates higher discordance. (**b**) RMSD between EPIC and imputed WGBS values at the same CpGs as in **a**. (**c**) Difference between **a** and **b**

generally well-correlated ($r^2 = 0.92$) (Additional file 1: Figure S8). However, we found that disagreement between the two platforms was concentrated at lower WGBS depth and intermediate beta values, with varying levels of discordance at high sequencing depth. Neither EPIC nor WGBS beta values can be considered the true methylation value of a particular CpG; however, since discordance between the two estimates was a function of sequencing depth, we hypothesized that discordance at low depth is most likely due to WGBS inaccuracy. For example, say a CpG that has a true methylation value of 85% is measured through both WGBS and EPIC, but is only covered by 10 WGBS reads. The WGBS estimate will differ from the true value by at least 5%, since the closest it can get is either 80% (8 out of 10 reads methylated) or 90%. The EPIC estimate, on the other hand, does not directly depend on read depth, but on the intensity of two possible fluorescent signals from probes that hybridize to the methylated or unmethylated alleles at a particular CpG. Therefore, assuming any inaccuracies due to fluorescent measurement are less than 5%, the EPIC estimate will be closer to the true value of 85% and generally more accurate for low-coverage CpGs. For high-coverage CpGs, we hypothesized that WGBS-EPIC discordance is most likely due to EPIC inaccuracy, perhaps due to saturation of the fluorescence signal and the presence of background hybridization to the alternate probe.

We then used BoostMe to impute and replace beta values common between the two platforms. We found that the discordance was mitigated (Fig. 5b), particularly at lower WGBS depth and intermediate beta values (Fig. 5c). Furthermore, we found that discordance mitigation at lower WGBS depth was robust with respect to the EPIC array probe type examined (Additional file 1: Figure S9). Discordance at higher depth was variable and, in some cases, increased after imputation (Fig. 5c).

BoostMe and random forests identify features important to general methylation levels

To interrogate differences in the methylation patterns of the different tissues and disease states, we examined the top variable importance scores output by random forests and BoostMe using all features (Fig. 6, Additional file 1: Figure S10, S11). Both algorithms highly prioritized the sample average and neighboring CpG features, which were well-correlated with the beta value of the CpG of interest.

We found that random forests also ranked highly features that were negatively correlated with beta values, especially those associated with open chromatin and promoter regions such as H3K4me3, ATAC-Seq peaks, CpG islands, and the TSS chromatin states (Fig. 6a). Random forests also identified as important several TFBSs previously shown to be methylation-sensitive such as YY1 [34, 35], REST [36, 37], and EP300 [38]. In concordance with previous results [14], we found that random forests were biased to rank highly features that are positively correlated with each other. This trend was particularly evident in the correlations among the top TFBSs identified, with all of them having some degree of overlap with each other (Fig. 6a).

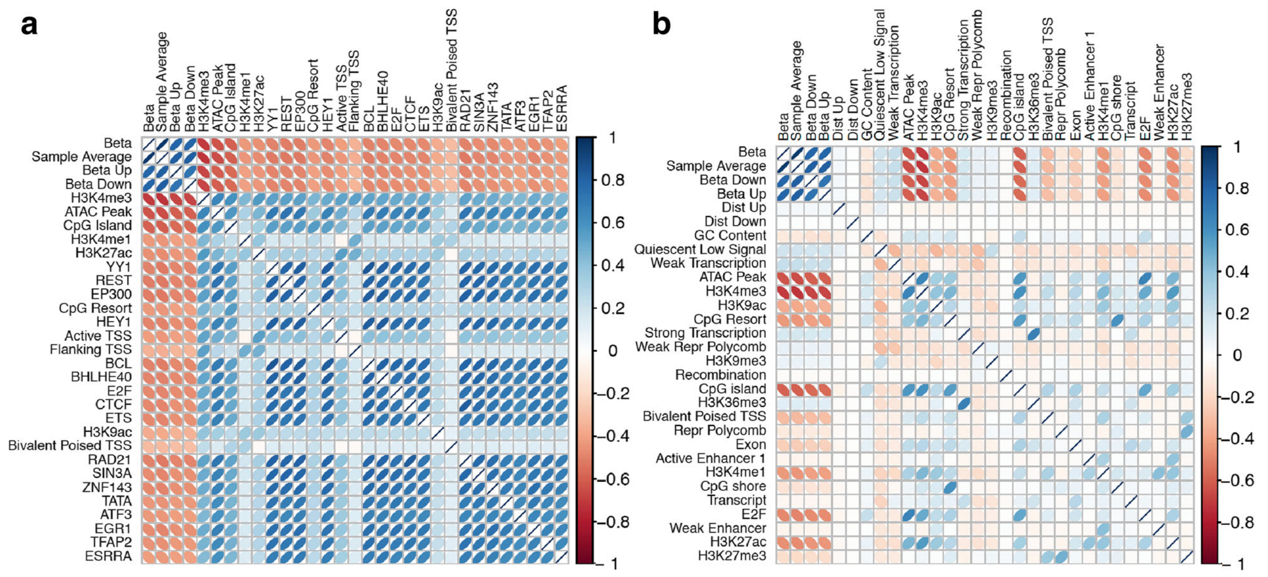

Fig. 6 Random forests exhibit greater bias in favor of positively correlated features compared to BoostMe. Correlation among methylation (beta) value and top 30 features in descending order for adipose T2D as reported by (**a**) random forests and (**b**) BoostMe. Ranking was determined by aggregating the variable importance scores across 10 runs from all adipose T2D samples

In contrast, BoostMe did not exhibit the same bias as random forests in favor of positively correlated features (Fig. 6b). Since gradient boosting trees are trained sequentially, with each subsequent tree designed to reduce error from the previous tree, BoostMe is less likely to rank highly features that exhibit strong positive cross-correlations. Therefore, in addition to highly predictive features that were negatively correlated with methylation identified by random forests, such as ATAC-Seq peaks and H3K4me3, BoostMe also prioritized chromatin states that were positively correlated with methylation such as the quiescent/low signal and weak transcription chromatin states. Compared with random forests, BoostMe did not report as many methylation-associated TFBSs to be highly predictive, likely because of their high positive correlations with each other and with other features indicative of open chromatin, such as ATAC-Seq peaks and H3K4me3.

To further determine whether BoostMe or random forests could identify features that were important to specific tissues, we trained both algorithms using only TFBSs in tissue-specific regions of open chromatin as determined by ATAC-seq peaks. We found that the rankings were generally the same across tissues for both algorithms (Additional file 1: Tables S6, S7).

Discussion

Here we introduce BoostMe, a method for imputing low-quality DNA methylation (beta) values within whole-genome bisulfite sequencing (WGBS) data. BoostMe is based on XGBoost, a computationally efficient gradient boosting algorithm [18]. Importantly, BoostMe leverages information from at least three samples and both trains and predicts on continuous beta values. This framework allows BoostMe to outperform existing imputation methodology, including DeepCpG [14], a deep neural network method, in both speed and accuracy across tissues and genomic contexts. BoostMe also achieves lower RMSE than DeepCpG for intermediately methylated CpGs, which we found to be enriched in regions of high across-tissue methylation variance. Furthermore, using matched EPIC and WGBS data from the same samples, we have shown that BoostMe imputation reduces discordance between the two platforms, particularly at low WGBS depth. Overall, our results support the use of BoostMe as a preprocessing step to improve WGBS quality when multiple samples are available.

A notable limitation of BoostMe is its interpretability. Although it was previously reported that random forests identified TFBSs associated with methylation in whole blood [14], we found that neither BoostMe nor random forests identified noteworthy differences in variable importance scores between different tissues. On the other hand, DeepCpG, a deep neural network method, was able to identify differences in transcription factor motifs associated with prediction among the different tissues. For example, DeepCpG identified motifs of TFs important to a tissue type, such as EBF1 in adipose, ASCL2 in muscle, and FOXA1 in pancreatic islets, which have all been reported to be involved in regulating differentiation and development in their respective cell types [39–43]. Thus, despite its relatively poor performance, DeepCpG may be superior for identifying tissue-specific differences. No algorithm examined in this work readily identified differences between NGT and T2D.

Similar to previous methodology [14, 15], BoostMe relies on the locally correlated structure of neighboring CpGs to identify sample-specific differences. Although using neighboring information leads to an overall more accurate prediction for all algorithms examined in this work, the accuracy of this approach may not be robust for a subset of CpGs. To determine the local similarity of CpG methylation within WGBS and its effect on algorithm performance, we first calculated pairwise differences between beta values across chromatin states, finding that enhancer states generally had the largest differences, while TSS and transcribed states had the lowest. In concordance with this result, we found that all algorithms performed worst in active and weak enhancer chromatin states and best in TSS and transcribed states. Performance was also slightly worse in regions of high across-tissue variance, where pairwise differences were generally larger.

To further characterize the relationship between informative neighboring CpGs and algorithm performance, we examined BoostMe error as a function of distance to the nearest non-missing CpG. Notably, the nearest non-missing beta values upstream and downstream were the second and third-most informative features for BoostMe after the sample average. We found that error was lowest when the CpG of interest had a non-missing neighbor within 100 bp. This result parallels a previous study [44], which reported that methylation haplotype blocks, defined as areas of consecutive CpGs with $r^2 > 0.5$, measure just 95 bp long on average. Although the majority (~ 87%) of CpGs did have a non-missing neighbor within 100 bp, the decreased performance for the remaining subset of CpGs may be a significant shortcoming of BoostMe and all neighbor-dependent imputation methods in general.

An important question of interest is whether BoostMe is able to predict whole-genome methylation values using methylation information from the EPIC array. The EPIC array consists of a sparse subset of ~ 3% of CpGs in the genome (with mean spacing of 1.5 kb; Additional file 1: Figure S2) and is much cheaper to use than WGBS. However, only ~ 12% of WGBS CpGs are within 100 bp of an EPIC CpG. Given the limitations in accuracy described above, additional work must be done to develop improved imputation methodology that can identify sample-specific differences for prediction without depending heavily on informative neighboring CpGs, perhaps through incorporation of long-range interactions and other high-dimensional genomic and epigenomic features not considered in this work. Such methodology could potentially facilitate whole-genome imputation from a sparse subset of CpGs, with accuracy independent of neighboring CpG distance. Furthermore, the experience with BoostMe imputation described here could be applied usefully to the future design of a tissue-specific methylation array, with array features optimally chosen from reference WGBS data from that tissue, allowing genome-wide imputation of unmeasured CpGs.

Conclusions

We characterized WGBS and EPIC data from 58 samples of human adipose, skeletal muscle, and pancreatic islets. We found high rates of missing methylation (beta) values spread across chromatin states in the WGBS data. To address this missingness, we developed BoostMe, a method for imputing beta values using WGBS data from at least three samples based on a gradient boosting algorithm called XGBoost. We found that our method outperforms random forests and DeepCpG, a previous deep neural network approach to imputing missing beta values. To assess the effect of imputation on the quality of WGBS, we compared the concordance of WGBS and EPIC beta values before and after imputation. We found that concordance between the two platforms increased after imputation, particularly at low WGBS depth. To interrogate the limitations of neighboring CpG-dependent methylation imputation, we characterized general patterns of neighboring CpG similarity and measured imputation performance as a function of distance to the nearest CpG. We found that BoostMe performance decreased when distance to the nearest CpG was higher, and that performance was generally worse in regions with more variability in CpG distances, such as enhancer chromatin states. To evaluate the ability of imputation methods to identify biologically-relevant features, we examined the variable importance scores output by BoostMe and random forests. Both algorithms identified features that were highly correlated with general methylation levels, such as indicators of open chromatin, but did not identify notable differences among tissue types. Overall, our findings support the use of BoostMe as a preprocessing step for WGBS analysis and inform the development of future methylation imputation methodology.

Methods

Sample collection

Muscle and adipose NGT and T2D samples were collected as previously described [4]. Briefly, we attempted to contact participants and participants' relatives from previous diabetes-related studies [45–48] and also recruited subjects by newspaper advertisements. We excluded individuals with any diseases or drug treatments that might confound analyses. We defined glucose tolerance categories of NGT and T2D using World Health Organization (WHO) criteria [49]. Biopsies were performed by 9 experienced and well-trained physicians from 2009 to 2013 in 3 different study sites (Helsinki, Kuopio, and Savitaipale). The study was approved by the coordinating ethics committee of the Hospital District of Helsinki and Uusimaa. A written informed consent was obtained from all subjects.

Islet samples were collected as previously described [31]. Briefly, samples were procured from the Integrated Islet Distribution Program, the National Disease Research Interchange (NDRI), or ProdoLabs. Islets were shipped overnight from distribution centers, prewarmed in shipping media for 1–2 h before harvest, and cultured in tissue culture-treated flasks. Genomic DNA was then isolated from islet explant cultures and used for sequencing.

Whole-genome sequencing

Whole genome sequencing libraries were generated from 50 ng genomic DNA fragmented by Covaris sonication. DNA end repair achieved using Lucigen DNA Terminator Repair Enzyme Mix. Sequencing adapters were added according to Illumina PE Sample Prep instruction. Libraries were size-selected on Invitrogen 4–12% polyacrylamide gels excising 200–250 bp fragments. Libraries were amplified with 10 PCR cycles and purified using AMPure beads (Beckman).

Whole-genome bisulfite sequencing

Whole-genome bisulfite sequencing was performed using Epigenome/TruSeq DNA Methylation Kit (Illumina). Libraries were prepared for each sample using 50 ng of input DNA by denaturing the DNA at 98 °C for 10 min. Bisulfite conversion was generated at 64 °C for 2.5 h and DNA purified using EZ DNA Methylation Gold Kit (Zymo Research). Bisulfite converted libraries were generated by random-primed DNA synthesis, 3′ tagging, and purification using AMPure beads (Beckman). Sample-specific index sequences were added with 10 cycles of amplification.

Library quality was assessed using Qubit (Thermo Fisher Scientific) and Agilent Bioanalyzer. Paired-end 125 bp sequencing was performed on Illumina HiSeq 2500 instruments to 30× genome coverage.

EPIC array

Genomic DNA was extracted from each tissue using DNeasy Blood and Tissue Kits (QIAGEN), according to the manufacturer's recommendations. 200 ng of genomic DNA per sample was submitted to the Center for Inherited Disease Research at The Johns Hopkins University, where they were bisulfite-converted using EZ DNA methylation Kits (Zymo research), as part of the TruSeq DNA Methylation protocol (Illumina). DNA methylation was measured using the Illumina Infinium HD Methylation Assay with Infinium MethylationEPIC BeadChips according to manufacturer's instructions.

WGS data processing

Raw FASTQ files were evaluated with FastQC [50]. Adapter sequences were trimmed using Atropos [51], and reads with at least one pair shorter than 25 bp were excluded. Reads were aligned to the reference genome (GRCh37) using BWA MEM [52], followed by Samblaster [53] for marking duplicates.

WGS variant calling

SNPs and indels were called separately for all sample BAM files using GATK HaplotypeCaller [54]. Variants were filtered using GATK Variant Quality Score Recalibration. Quality score cutoffs were chosen by comparing rates of discordance with SNP array genotypes.

WGBS data processing

Raw FASTQ were pre-processed as above and aligned using bwa-meth [55]. Methylation values were extracted using the MethylDackel 'extract' command, including bias correction based on the values recommended by the 'mbias' command, and forward- and reverse-strand CpGs were merged with a minimum coverage cutoff of 10 (https://github.com/dpryan79/methyldackel). Methylation level data from the X and Y chromosomes were excluded.

EPIC Array data processing

The EPIC data are part of a much larger, unpublished study. As such, all samples were processed jointly with other samples from the larger study. We processed raw signal idat files using minfi v1.20.2 [56] with the Illumina normalization method. We analyzed the quality of each sample looking for outliers across a variety of measures including fraction of failed probes (detection p-value > 0.05), median methylated and un-methylated intensity, control probe signal (using the returnControlStat function from shinyMethyl v1.10.0 [57]), distribution of the overall methylation profile, and principal component analysis. None of the samples included in this study were flagged as outliers. In addition, we verified the identity of each sample by comparing genotypes assayed on the EPIC array to imputed genotypes using the HRC reference panel r1.1 [58] and Illumina Omni2.5 array genotypes.

For both the earlier 450k and recent EPIC Illumina methylation array, previous studies [59–62] have identified poor quality probes that either do not uniquely map to the reference genome or contain common genetic variation. These properties make the signal at these probes un-reliable. We removed such probes from the EPIC array. First, we removed cross-reactive probes on the EPIC chip by mapped non-control probes back to the entire bisulfite-converted genome, using Novoalign's -b4 option, with allowance for up to three mismatches in the probe alignment (–R120 option). We kept only unique mapping probes. Second, we removed probes with a SNP within 10 bp of the 3′ end of the probe, within the target CpG itself, and finally, in the case of type I probes, if the variant overlapped the single base extension site. We used 10 bp as this cutoff is consistent with previous studies [60]. For SNPs we used common (MAF ≥ 1%) SNPs, indels, or structural variation in the phase 3 1000 Genomes European dataset, common (MAF ≥ 1%) SNPs in the HRC reference panel r1.1, and SNPs appearing at all our own samples, even at low frequency, after imputation to the Haplotype Reference Consortium (HRC) reference panel. As a final step, we combined our blacklist with a previously published blacklist [62] for a total of 120,627 probes which were removed from analysis. In addition, we removed probes per tissue with a high detection p-value (p-value > 0.05 in ≥ 5% of samples from the larger study). After blacklist filters, we removed 578 adipose probes, 733 muscle probes, and 2206 islet probes based on the per sample filters.

Identification of higher across-tissue variance regions in WGBS

Using all 58 WGBS samples, we calculated the variance in beta values for each CpG. We then searched for blocks of consecutive CpGs that had 1) variance above the third quartile of variance levels and 2) a non-missing methylation value in at least 20 samples, a cutoff which was determined by looking at the distribution of variance values as a function of missing values (Additional file 1: Figure S12). This analysis identified approximately 200,000 blocks of high across-tissue variance CpGs genome-wide. Blocks contained an average of eight CpGs, spanned an average of 533 bp, and had higher relative enrichment in enhancer chromatin states (Additional file 1: Figure S5).

Feature construction for BoostMe and random forests

We used the same 648 features in the BoostMe and random forest algorithms (see Additional file 1: Table S3 for a detailed list). Prior to feature construction, we applied a further set of exclusion criteria to filter the CpGs included in training, validation, and testing. Only autosomal CpGs were used ($n = 25,586,776$). We overlapped WGS data with the WGBS data from all samples and excluded CpGs for which the CG dinucleotide on either strand was disturbed by a SNP or indel that was 2 bp long (indels longer than 2 bp were not considered). We also excluded all CpGs located in ENCODE blacklist regions [63].

CpG features

Features constructed from the WGBS data included the nearest non-missing neighboring CpG beta values, the distances to these CpGs, and the sample average feature. Neighboring CpG features were taken within the sample of interest. The sample average feature was created by taking the average beta value of all samples within each tissue at the CpG of interest, not including the sample being interrogated. Samples in which the CpG was not sequenced above 10× coverage were excluded from the calculation. CpGs without a measurement above 10× coverage from at least two additional samples were also excluded.

Genomic features

We constructed both general and tissue-specific genomic features. General genomic features were the same across all tissues and included GC content, recombination rate, GENCODE annotations, and CpG island (CGI) information. GC content data was downloaded from the raw data used to encode the gc5Base track on hg19 from the UCSC Genome Browser [64, 65]. DNA recombination rate annotations from HapMap were downloaded from the UCSC hg19 annotation database (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/). CGI coordinates were obtained from UCSC browser. CGI shores and shelves were calculated from CGI coordinates by taking 2 kb flanking regions. GENCODE v25 transcript annotations were downloaded from the GENCODE data portal (ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_25).

Tissue-specific genomic features included ATAC-seq, chromatin states, histone marks, and transcription factor binding sites (TFBS). These features were all binary, with 0 indicating that the CpG of interest did not overlap that feature, and 1 indicating overlap. Chromatin state annotations were obtained from a previously published 13 chromatin state model for 31 diverse tissues that included islets, skeletal muscle, and adipose [30]. This model was generated from cell/tissue ChIP-seq data for H3K27ac, H3K27me3, H3K36me3, H3K4me1, and H3K4me3, and input from a diverse set of publicly available data [20, 66–68]. ATAC-seq data was obtained from previously published studies for islets [31], skeletal muscle [69], and adipose [70]. TFBS data was obtained as described in [69], with additional PWMs from [71]. TFBS data was filtered for each tissue by the ATAC-seq feature to only include hits overlapping an ATAC-seq peak. We merged hits from multiple motifs of the same transcription factor to reduce the number of variables included in the algorithm and optimize computational efficiency.

BoostMe and random forests implementation

For BoostMe, we used the xgboost package (version 0.6–4) [18] in R [72] (version 3.3.1). For random forests, we used the ranger package (version 0.6.0) in R, which facilitates random forest training and testing on multiple CPUs [73]. For both algorithms, we used regression trees to predict a continuous methylation value between 0 and 1 for CpGs of interest. Algorithms were trained on individual samples within each tissue and disease state combination. We trained only on CpGs with at least 10× coverage and no more than 80× coverage. Random forest variable importance was calculated using the mean decrease in variance at each split as implemented in the ranger package. BoostMe variable importance was evaluated for each variable as the loss reduction after each split using that variable as implemented in the xgboost package.

Due to the computational constraints of decision tree algorithms, we replicated a previously-described form of repeated random subsampling validation [14] to evaluate both BoostMe and random forests. Briefly, within each sample, algorithms were trained on a random sample of 1,000,000 CpGs and validated on a hold-out set of 1,000,000 CpGs. Final performance benchmarking was done on another held-out test set of 1,000,000 CpGs (Table 2). This process of random sampling, algorithm training, and benchmarking was repeated with ten different random seeds, and prediction performance was calculated by averaging the performance statistics across each of the ten algorithms. Hyperparameters tuned for random forests included the number of trees grown and number of variables to possibly split on at in each node, however, similar to previous work [14], we found that performance was robust to different settings, and thus did not estimate parameters.

For BoostMe, we used the mlr package (version 2.9) [74] in R to perform a hyperparameter grid search to tune the following xgboost package parameters: learning rate (eta), minimum loss reduction required to make a further partition on a leaf node of the tree (gamma), maximum depth of a tree (max_depth), minimum sum of instance weights needed in a child node (min_child_weight), subsample ratio of the training instance (subsample), subsample ratio of columns when constructing each tree (colsample_bytree), L2 regularization term (lambda), and L1 regularization term (alpha). We found that optimal hyperparameters had varying combinations for each sample, however, improvements in RMSE were only marginally better than using all default settings for each sample. Additionally, due to the number of hyperparameters available to tune, an exhaustive search of hyperparameter space was relatively computationally expensive. Therefore, we used all default settings as described in the xgboost package to benchmark BoostMe performance.

DeepCpG implementation

We implemented DeepCpG (version 1.0.4) as described in [15]. Briefly, for each of the five tissue and T2D status combinations (adipose NGT, adipose T2D, muscle NGT, muscle T2D, and islet) the data was first divided by chromosome into training (chr. 1, 3, 5, 7, 9, 11, 13, 15), validation (chr. 16, 17, 18, 19, 20, 21, 22), and test sets, corresponding to a rough 40–20-40 split. The DNA module and CpG module were trained on separate NVIDIA Tesla K80 GPUs and the performance of each module was evaluated individually on the test set. The joint module was trained with the best-performing DNA and CpG modules, and its predictions were used for final benchmarking. In contrast to original single-cell bisulfite implementation of DeepCpG which was trained and tested on binary methylation values, we trained and tested on continuous methylation values to parallel our implementation of BoostMe and random forests. We found that this change made no difference in the accuracy of the model (Additional file 1: Table S4).

We experimented with six different hyperparameter combinations for each DNA model, including three architectures (CnnL2h128, CnnL2h256, CnnL3h256) and two dropout rates (0, 0.2). We then selected the best-performing combination based on AUC and reported the motifs significantly matching the filters from the first convolutional layer of that model [75]. Similarly, we tested both RnnL1 and RnnL2 for the CpG model for each tissue. For the joint module, we tested JointL1h512, JointL2h512, and JointL3h512. The best-performing joint model was selected to evaluate RMSE, AUROC, AUPRC, and accuracy for each tissue. We used a default learning rate of 0.001 for all models. Similar to random forests and BoostMe, performance was generally robust with respect to different architectures. For a detailed explanation of all model architectures, see http://deepcpg.readthedocs.io/.

Additional file

Additional file 1: Figure S1. Distribution of WGBS missingness across chromatin states, not normalized for total number of CpGs in that chromatin state. **Figure S2.** Distribution of distance from a WGBS or EPIC CpG to the nearest WGBS or EPIC CpG. **Figure S3.** CpG methylation pairwise differences as a function of distance genome-wide and in regions of higher across-tissue variance. **Figure S4.** CpG methylation pairwise differences for pancreatic islets as a function of distance genome-wide and in regions of higher across-tissue variance. **Figure S5.** Comparison of all CpGs and CpGs that had higher across-tissue variance. **Figure S6.** Joyplot showing distribution of WGBS methylation (beta) values within each chromatin state for all tissues. **Figure S7.** Performance at intermediate CpGs does not improve when using a balanced training distribution. **Figure S8.** Smooth scatterplot of beta values of CpGs shared between EPIC and WGBS. **Figure S9.** Imputation mitigates discordance between WGBS at EPIC at low WGBS depth regardless of EPIC probe type. **Figure S10.** Correlation among the top 30 features ranked by BoostMe. **Figure S11.** Correlation among the top 30 features ranked by random forests. **Figure S12.** Distribution of across-sample CpG variance values vs. number of missing values for each CpG. **Table S1.** Summary of the data used in this work. **Table S2.** Previously reported imputation metrics and those reported in this work. **Table S3.** All features included in BoostMe and random forests and their source. **Table S4.** Genome-wide performance of algorithms, trained on 500,000 CpGs, for predicting methylation values. **Table S5.** RMSE performance of BoostMe and random forests improves when training on continuous values. **Table S6.** Top 100 transcription factors ranked in descending order as reported by BoostMe, trained only using TFBS features. **Table S7.** Top 100 transcription factors ranked in descending order as reported by random forests trained only using TFBS features. (PDF 5703 kb)

Abbreviations

ATAC-seq: Assay for transposase-accessible chromatin using sequencing; AUPRC: Area under the precision-recall curve; AUROC: Area under the receiver operating characteristic curve; CGI: CpG island; CpG: CG dinucleotide; EPIC: Illumina Infinium MethylationEPIC BeadChip; GWAS: Genome wide association study; NGT: Normal glucose tolerance; RMSE: Root-mean-squared error; SNP: Single nucleotide polymorphism; T2D: Type 2 diabetes; TF: Transcription factor; TFBS: Transcription factor binding site; TSS: Transcription start site; WGBS: Whole genome bisulfite sequencing; WGS: Whole genome sequencing

Acknowledgements

The authors would like to thank members of the Collins laboratory for helpful discussions, Christof Angermueller for his assistance with DeepCpG, and the members of MGI involved in this study: Vincent Magrini, Richard K. Wilson, Ryan Demeter, and Robert S. Fulton. This manuscript is dedicated to Peter Chines.

Funding

LSZ, MRE, DLT, PSC, FSC, JPD: NHGRI. JPD: American Diabetes Association Fellowship. DLT: NIH-Oxford Cambridge Scholars Program. AV: American Association for University Women (AAUW) International Doctoral Fellowship. SCJP: National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Grant R00DK099240 and American Diabetes Association Pathway to Stop Diabetes Grant 1–14-INI-07.

Availability of data and materials

The WGBS and EPIC datasets analyzed from the current study will be deposited in dbGaP under study accession numbers phs001048 (FUSION tissue biopsy study) and phs001188 (pancreatic islets). BoostMe is freely available as an R package on Github at https://github.com/lulizou/boostme.

Authors' contributions

JPD conceived of the study. LSZ developed the software and ran experiments. MRE and MGI generated the data. DLT, PSC, AV, SCJP, and FSC contributed to analysis and interpretation of data. All authors contributed to writing or revising the manuscript, and all authors have approved the final manuscript.

Ethics approval and consent to participate

The study was approved under IRB protocol number OH95-HG-N030 and by the coordinating ethics committee of the Hospital District of Helsinki and Uusimaa. A written informed consent was obtained from all subjects.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA. [2]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK. [3]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA. [4]Washington University School of Medicine, St. Louis, MO 63108, USA. [5]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA.

Received: 6 February 2018 Accepted: 8 May 2018
Published online: 23 May 2018

References

1. Robertson KD. DNA methylation and human disease. Nat Rev Genet. 2005;6:597–610.
2. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet. 2008;9:465–76.
3. Bird AP. DNA methylation patterns and epigenetic memory. Genes Dev. 2002;16:6–21.
4. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat. Rev. Genet. 2013;14:204–20.
5. Trerotola M, Relli V, Simeone P, Alberti S. Epigenetic inheritance and the missing heritability. Hum Genomics. 2015;9:17.
6. Heard E, Martienssen RA. Transgenerational epigenetic inheritance: myths and mechanisms. Cell. 2014;157:95–109.
7. Lim JP, Brunet A. Bridging the transgenerational gap with epigenetic memory. Trends Genet. 2013;29:176–86.
8. Illumina Support. https://support.illumina.com/. Accessed 8 Feb 2018.
9. Ziller MJ, Hansen KD, Meissner A, Aryee MJ. Coverage recommendations for methylation analysis by whole genome bisulfite sequencing. Nat Methods. 2015;12:230–2.
10. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. 2012;13:R83.
11. Das S, Foerer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48:1284–7.
12. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009;10:387–406.
13. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 2010;11:499–511.
14. Zhang W, Spector T, Deloukas P, Bell JT, Engelhardt BE. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. Genome Biol. 2015;16:14.
15. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. 2017;18:67.
16. Breiman L. Random forests. Mach Learn. 2001;45:5.
17. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
18. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM; 2016. p. 785–794.
19. Elliott G, Hong C, Xing X, Zhou X, Li D, Coarfa C, et al. Intermediate DNA methylation is a conserved signature of genome regulation. Nat Commun. 2015;6:6363.
20. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.
21. Lövkvist C, Dodd IB, Sneppen K, Haerter JO. DNA methylation in human epigenomes depends on local topology of CpG sites. Nucleic Acids Res. 2016;44:5123–32.
22. Zeng H, Gifford DK. Predicting the impact of non-coding variants on DNA methylation. Nucleic Acids Res. 2017;45:e99.
23. Ma B, Wilker EH, Willis-Owen SAG, Byun H, Wong KCC, Motta V, et al. Predicting DNA methylation level across human tissues. Nucleic Acids Res. 2014;42:3515–28.
24. Zhang G, Huang K, Xu Z, Tzeng Y, Conneely KN, Guan W, et al. Across-platform imputation of DNA methylation levels incorporating nonlocal information using penalized functional regression. Genet Epidemiol. 2016;40:333–40.
25. Fan S, Huang K, Ai R, Wang M, Wang W. Predicting CpG methylation levels by integrating Infinium HumanMethylation450 BeadChip array data. Genomics. 2016;107:132–7.
26. Wang Y, Liu T, Xu D, Shi H, Zhang C, Mo YY, Wang Z. Predicting DNA methylation state of CpG dinucleotide using genome topological features and deep networks. Sci Rep. 2016;6:19598.
27. Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat Biotech. 2015;33:364–76.
28. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, et al. The genetic architecture of type 2 diabetes. Nature. 2016;536:41–7.
29. McCarthy MI, Zeggini E. Genome-wide association studies in type 2 diabetes. Curr Diab Rep. 2009;9:164–71.
30. Saxena R, Saleheen D, Been LF, Garavito ML, Braun T, Bjonnes A, et al. Genome-wide association study identifies a novel locus contributing to type 2 diabetes susceptibility in Sikhs of Punjabi origin from India. Diabetes. 2013;62:1746–55.
31. Varshney A, Scott LJ, Welch RP, Erdos MR, Chines PS, Narisu N, et al. Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. Proc Natl Acad Sci. 2017;114:2301–6.
32. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500:477–81.
33. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016;17:208.
34. Kim J, Kollhoff A, Bergmann A, Stubbs L. Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, Peg3. Hum Mol Genet. 2003;12:233–45.
35. Sekimata M, Murakami-Sekimata A, Homma Y. CpG methylation prevents YY1-mediated transcriptional activation of the vimentin promoter. Biochem Biophys Res Commun. 2011;414:767–72.
36. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011;480:490–5.
37. Marchal C, Miotto B. Emerging concept in DNA methylation: role of transcription factors in shaping DNA methylation patterns. J Cell Physiol. 2015;230:743–51.
38. Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, et al. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Res. 2013;23:555–67.
39. Gao H, Mejhert N, Fretz JA, Arner E, Lorente-Cebrián S, Ehrlund A, et al. Early B cell factor 1 regulates adipocyte morphology and lipolysis in white adipose tissue. Cell Metab. 2014;19:981–92.
40. Petrus P, Mejhert N, Gao H, Bäckdahl J, Arner E, Arner P, Rydén M. Low early B-cell factor 1 (EBF1) activity in human subcutaneous adipose tissue is linked to a pernicious metabolic profile. Diabetes Metab. 2015;41:509–12.
41. Wang C, Wang M, Arrington J, Shan T, Yue F, Nie Y, et al. Ascl2 inhibits myogenesis by antagonizing the transcriptional activity of myogenic regulatory factors. Development. 2017;144:235–47.
42. Gao N, Le Lay J, Qin W, Doliba N, Schug J, Fox AJ, et al. Foxa1 and Foxa2 maintain the metabolic and secretory features of the mature beta-cell. Mol Endocrinol. 2010;24:1594–604.
43. Vatamaniuk MZ, Gupta RK, Lantz KA, Doliba NM, Matschinsky FM, Kaestner KH. Foxa1-deficient mice exhibit impaired insulin secretion due to uncoupled oxidative phosphorylation. Diabetes. 2006;10:2730–6.
44. Guo S, Diep D, Plongthongkum N, Fung HL, Zhang K, Zhang K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nat Genet. 2017;49:635–42.
45. Valle T, Tuomilehto J, Bergman RN, Ghosh S, Hauser ER, Eriksson J, et al. Mapping genes for NIDDM. Design of the Finland-United States Investigation of NIDDM genetics (FUSION) study. Diabetes Care. 1998;21:949–58.
46. Väätäinen S, Keinänen-Kiukaanniemi S, Saramies J, Uusitalo H, Tuomilehto J, Martikainen J. Quality of life along the diabetes continuum: a cross-sectional view of health-related quality of life and general health status in middle-aged and older Finns. Qual Life Res. 2014;23:1935–44.
47. Kouki R, Schwab U, Lakka TA, Hassinen M, Savonen K, Komulainen P, et al. Diet, fitness and the metabolic syndrome - the DR's EXTRA study. Nutr Metab Cardiovasc Dis. 2012;22:553–60.

48. Stančáková A, Kuulasmaa T, Paananen J, Jackson AU, Bonnycastle LL, Collins FS. Association of 18 confirmed susceptibility loci for type 2 diabetes with indices of insulin release, proinsulin conversion, and insulin sensitivity in 5,327 nondiabetic Finnish men. Diabetes. 2009;58:2129–36.
49. World Health Organization (WHO), International Diabetes Federation (IDF). Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF consultation. Geneva, Switzerland: WHO; 2006.
50. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010; available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
51. Didion JP, Martin M, Collins FS. Atropos: specific, sensitive, and speedy trimming of sequencing reads. PeerJ. 2017;5:e3720.
52. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013. Preprint at arXiv:1303.3997v2 [q-bio.GN].
53. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 2014;30:2503–5.
54. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
55. Pedersen BS, Eyring K, De S, Yang IV, Schwartz DA. Fast and accurate alignment of long bisulfite-seq reads. 2014. Preprint at arXiv:1401.1129 [q.bio.GN].
56. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30:1363–9.
57. Fortin JP, Fertig E, Hansen K. shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R. F1000Res. 2014;3:175.
58. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;48:1279–83.
59. Chen Y, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics. 2013;8:203–9.
60. Price ME, Cotton AM, Lam LL, Farré P, Emberly E, Brown CJ, et al. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. Epigenetics Chromatin. 2013;6:4.
61. Zhang X, Mu W, Zhang W. On the analysis of the Illumina 450k array data: probes ambiguously mapped to the human genome. Front Genet. 2012;3:73.
62. McCartney DL, Walker RM, Morris SW, McIntosh AM, Porteous DJ, Evans KL. Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. Genom Data. 2016;9:22–4.
63. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
64. Golden path track of the University of Santa Cruz Genome Browser. http://hgdownload.cse.ucsc.edu/goldenPath/hg19/gc5Base/.
65. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, et al. The UCSC genome browser database: extensions and updates 2013. Nucleic Acids Res. 2013;41:D64–9.
66. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473:43–9.
67. Mikkelsen TS, Xu Z, Zhang X, Wang L, Gimble JM, Lander ES, Rosen ED. Comparative epigenomic analysis of murine and human adipogenesis. Cell. 2010;143:156–69.
68. Parker SCJ, Stitzel ML, Taylor DL, Orozco JM, Erdos MR, Akiyama JA, et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc Natl Acad Sci. 2013;110:17921–6.
69. Scott LJ, Erdos MR, Huyghe JR, Welch RP, Beck AT, Wolford BN, et al. The genetic regulatory signature of type 2 diabetes in human skeletal muscle. Nat Commun. 2016;7:11764.
70. Allum F, Shao X, Guénard F, Simon MM, Busche S, Caron M, et al. Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. Nat Commun. 2015;6:7211.
71. Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature. 2015;527:384–8.
72. R project. http://www.r-project.org/.
73. Wright M, Ziegler A. Ranger: a fast implementation of random forests for high dimension data in C++ and R. J. Stat Softw. 2017;77:1–17.
74. Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, et al. Mlr: machine learning in R. Journal J Mach Learn Res. 2016;17:1–5.
75. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016;26:990–9.