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], Peter S. Chines[1], Arushi Varshney[2], Stephen C. J. Parker[2,3], Francis S. Collins[1]* and John P. Didion[1]*

[1] National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892
[2] Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109
[3] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109
*Equal contributor

Correspondence to Francis Collins (collinsf@od.nih.gov).

Abstract

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

Results: Here we describe BoostMe, a novel method for imputation of DNA methylation within whole-genome bisulfite sequencing (WGBS) data based on a gradient boosting algorithm. Importantly, we designed a new feature that leverages information from multiple samples in the same tissue and disease state, enabling BoostMe to outperform existing imputation methods in speed and accuracy. We show that imputation improves WGBS concordance with the Infinium MethylationEPIC array at low WGBS sequencing depth, suggesting improvement in WGBS accuracy after imputation. Furthermore, we compare the ability of BoostMe and DeepCpG - a deep neural network method - to identify interesting features and motifs associated with methylation in three human tissues implicated in type 2 diabetes (T2D) etiology. We find that while BoostMe only identifies features important to general methylation levels across tissues, DeepCpG is able to learn differences in methylation-associated sequence motifs among different tissues and identify tissue-specific regulators of differentiation such as EBF1 in adipose, ASCL2 in muscle, and FOXA1, TCF12, and NRF1 in islets. Neither algorithm readily identified T2D-associated features.

Conclusions: Our findings demonstrate the current power and limitations of machine and deep learning algorithms to both improve the quality of, and infer biological meaning from, genome-wide DNA methylation data.

Keywords: DNA methylation, XGBoost, whole-genome bisulfite sequencing (WGBS), EPIC, DeepCpG, imputation, adipose, muscle, pancreatic islets, type 2 diabetes (T2D)

Background

Type 2 diabetes (T2D) is a complex disorder involving interactions between genetic and environmental factors acting in multiple tissues over time. Signatures of T2D are well-characterized and include impaired insulin production in pancreatic islets, insulin resistance in skeletal muscle, inflammation and dysregulation of lipid metabolism and endocrine signaling in adipose tissue, and significant contributions from liver, intestine, and brain. Although T2D risk factors such as lifestyle and genetic variants have been identified, the mechanisms by which these factors interact to cause T2D remain poorly understood.

Genome-wide association studies (GWAS) show that the majority of T2D-associated variants lie in non-coding regions of the genome [1-3]. These variants 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. Preliminary cis-eQTL studies have shown strong enrichment for regulatory variants among T2D GWAS loci in skeletal muscle [4] and islets [5,6]; however, the specific epigenetic mechanisms that mediate cis-genetic variant regulation of T2D-associated genes remain unclear.

DNA methylation is an epigenetic mark that is thought to integrate environmental signals with gene expression [7-9] and may contribute to regulation of T2D-associated genes [10-13]. In mammals, DNA methylation occurs primarily on cytosines of CG dinucleotides (CpGs). The extent to which methylation plays a causal role in altering gene expression, however, remains debated [14-18]. There is experimental evidence that DNA methylation can prevent [19-21] or promote [22,23] binding of some transcription factors (TFs) to particular binding sites (TFBS), supporting the hypothesis that methylation may directly affect TF affinity and therefore gene expression. However, there is also evidence that methylation of TFBS may occur after TF release rather than cause it [24-31]. In this model, methylation is merely a byproduct of other epigenetic processes, serving as a means of crosstalk among other epigenetic factors such as nucleosomes and TFs that may act non-concurrently [21, 32-34].

Nevertheless, methylation remains a crucial epigenetic mark to consider in the context of T2D. There is strong evidence that methylation is a mode for heritable transmission of epigenetic information, as it can be stably propagated through both mitosis and meiosis [35-39]. This transmission across cells could contribute to the maintenance of the T2D phenotype, particularly in T2D-associated tissues such as adipose, skeletal muscle, and pancreatic islets.

Most previous work to characterize the role of methylation in T2D has important design limitations, including suboptimal choice of tissue (blood) [40,41] and reliance on array-based assays [42-44]. Sparse genome-wide measurements of methylation such as the Illumina-450k array, and to a lesser extent the new Illumina Infinium MethylationEPIC (EPIC) array [45], are biased towards gene and promoter regions, with relatively little coverage of intergenic regions where a large fraction of T2D GWAS loci are located. Despite this, headway has been made in identifying interesting differentially methylated

genes in the T2D state. For example, T2D GWAS variants such as *TCF7L2* and *KCNQ1* have been shown to be differentially methylated in both adipose tissue [43,46] and pancreatic islets [44]. A recent whole-genome bisulfite sequencing (WGBS) analysis of methylation in pancreatic islets in healthy and T2D individuals found that *CTCF*, *FOXA2*, and *NEUROD1* transcription factor motif sites were enriched in differentially methylated regions in pancreatic islets [47]. Further whole-genome work is needed to characterize methylation patterns in adipose, skeletal muscle, pancreatic islets, and other T2D-relevant tissues.

WGBS is advantageous in that it provides unbiased coverage of most of the ~28 million CpGs in the genome. However, quantitative methylation estimates from CpGs sequenced at low depth are subject to error and are typically removed before performing downstream analysis [48], which can lead to the exclusion of millions of CpGs in any given data set. Furthermore, 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.

One potential remedy for these inefficiencies 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. Such techniques have already been used to increase the power of GWAS studies by leveraging data from sparse yet cost-effective SNP arrays [49-51].

Machine and deep learning algorithms have shown promise in providing accurate methylation estimates after training on sparse data sets [52,53]. Prediction accuracy, however, is still far from the currently expected SNP imputation accuracy in GWAS [52], leading to the need for algorithm improvement. The most recent methylation imputation methods used either random forest [54] or deep neural network [55] algorithms. A relatively new algorithm called extreme gradient boosting (XGBoost) has been shown to outperform both methods in accuracy and computational efficiency in data science competitions when highly predictive features can be constructed [56]. Previous imputation methods have also only classified CpG sites 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 methylation levels as a conserved genomic signature [57]. Furthermore, although previous algorithms have constructed features that capture the local correlation structure of CpGs [52,58] as well as information from the surrounding DNA sequence context [52,53,59-61], 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 [52,53,59,62-64]. For

example, a previous random forest [52] algorithm identified TFBS including *ELF1*, *MAZ*, *MXI1*, and *RUNX3* to be predictive of methylation levels in whole blood. A recent deep learning algorithm [53] found that transcription factor motifs such as Foxa2 and Srf, which are both implicated in cell differentiation and embryonic development, were important to methylation prediction in mouse embryonic stem cells. These algorithms are therefore useful for characterizing methylation regulatory networks, but they have not yet been applied to study methylation in T2D-implicated tissues such as adipose, skeletal muscle, and pancreatic islets.

In this work, we generated EPIC and WGBS data on 58 human samples from adipose, skeletal muscle, and pancreatic islets. Samples from adipose and skeletal muscle included those from patients with normal glucose tolerance (NGT) and T2D. 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. Therefore, we developed an imputation method based on XGBoost called BoostMe, which is designed to leverage information from multiple samples in the same tissue 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 methylation 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 a preprocessing step for WGBS data pipelines. Finally, we compare the ability of BoostMe and DeepCpG, a deep learning method, to identify biologically interesting features and sequence motifs for each tissue. We find that, while BoostMe identifies features important to general methylation levels across tissues, DeepCpG also discovers differences in sequence motifs associated with methylation between tissues. However, neither algorithm readily identifies differences between NGT and T2D tissues. Our findings demonstrate the current power and limitations of machine and deep learning algorithms to both improve the quality of, and infer biological meaning from, genome-wide DNA methylation data.

Results

BoostMe outperforms random forests and DeepCpG for methylation imputation

To address the high rate of missingness spread across the genome in our WGBS data (**Figure 1, Additional file 1: Figure S1, S2**), we developed BoostMe, a method for imputing methylation values using WGBS data from multiple samples. Previous attempts at DNA methylation imputation based on penalized functional regression [61], random forests [52], and deep neural networks [53] yielded relatively poor predictive accuracy genome-wide (RMSE > 0.23, AUROC < 0.93) (**Additional file 1: Table S1**). To improve on those methods, we implemented predictive models optimized for WGBS data using both random forest and gradient boosting [56] algorithms.

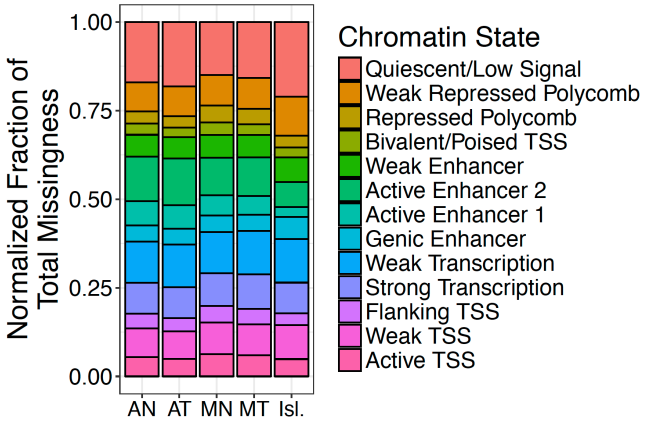

Figure 1 | WGBS missingness is distributed across chromatin states in all tissues. The normalized fraction of total missingness (y-axis) was calculated as the number of CpGs in each chromatin state that were missing (sequencing depth < 10x) normalized by the total number of CpGs in that chromatin state 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.

We constructed a total of 648 features designed to both parallel and improve upon previous work [52]. Prediction features constructed from the WGBS data included the nearest non-missing neighboring CpG methylation (beta) values upstream and downstream of the CpG of interest, base-pair distance to the neighboring CpGs, and the average methylation value of the CpG of interest in other samples from the same tissue and disease state. We also included features that describe the genomic context of individual CpGs such as histone marks (n=7), computational predictions of TFBS (n=608), chromatin states (n=13), and ATAC-Seq peaks (as a measure of DNA accessibility; see **Methods**, **Additional file 1: Table S2** for a full list of features). Using these 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 < 10x) was 2.6 million (range: 750,000 - 7.7 million) (**Additional file 2**).

We compared the performance of BoostMe with random forests and DeepCpG for predicting continuous methylation 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 (**Figure 2A**). We found that both BoostMe and random forests outperformed DeepCpG, achieving an average root-mean-squared error (RMSE) of 0.10, 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 1**). Unlike previous methods [52,53], we trained on continuous methylation values rather than binary values because of the available depth in our WGBS data, and found that this change improved overall RMSE by at least 0.06 and performed similar for AUROC, AUPRC, and accuracy (**Additional file 1: Table S3**).

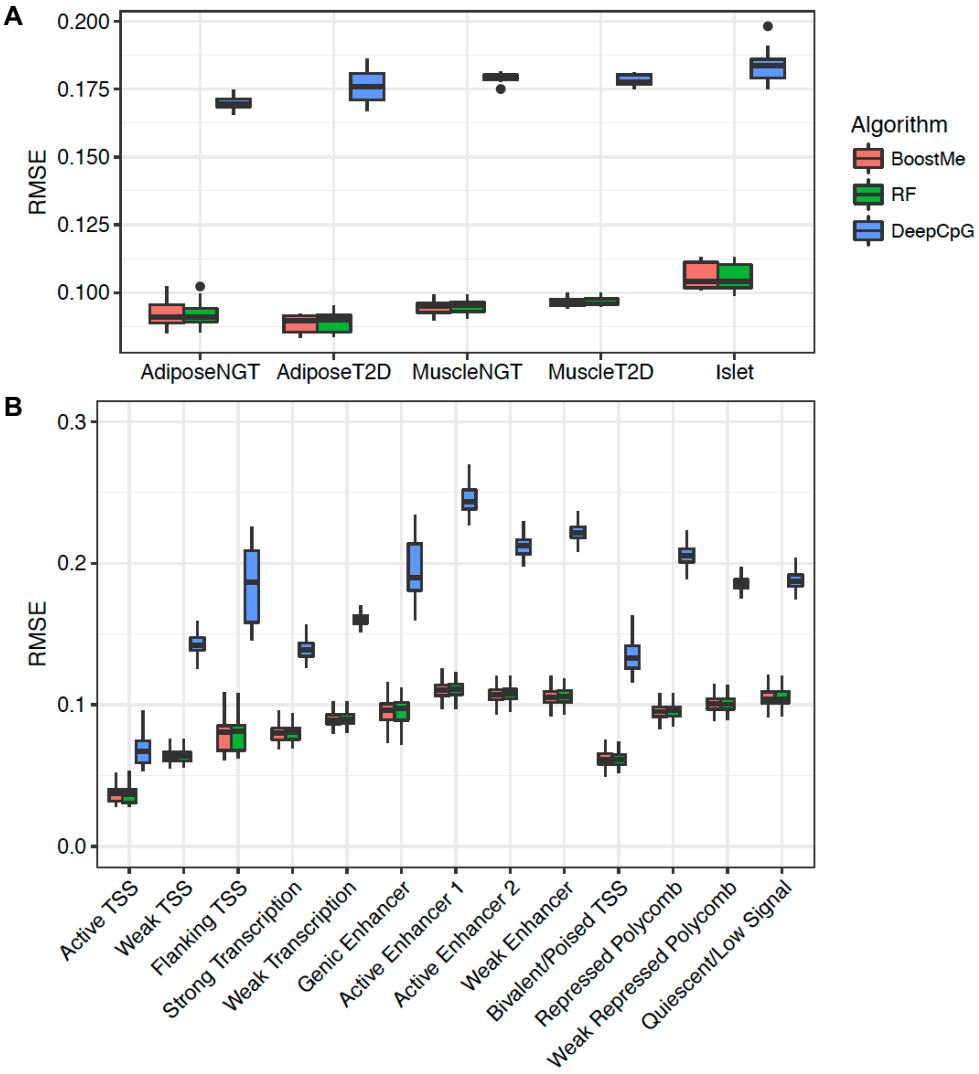

Figure 2 | BoostMe and random forests outperform DeepCpG for predicting methylation values genome-wide. (A) Root-mean-squared error (RMSE) of BoostMe, random forests (RF) and DeepCpG for predicting methylation in all tissue and disease state combinations examined in this study. Data points represent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

Table 1 | 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	0.10	0.13	0.99	0.99	0.96	16 CPUs	0.33
Random Forests	0.10	0.13	0.99	0.99	0.96	16 CPUs	14
DeepcpG	0.18	0.23	0.94	0.98	0.91	1 GPU	144

RMSE, root-mean-squared error; int., intermediate CpGs, 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.

To investigate whether BoostMe could produce accurate predictions for intermediate CpGs as well, we measured performance specifically at CpGs that had an average methylation value across samples 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 1**). We hypothesized that altering the training set distribution to include more CpGs with average methylation values in the intermediate range would improve prediction at these CpGs; however, performance did not improve significantly (**Additional file 1: Figure S3**).

To characterize performance patterns in different genomic contexts, we compared the performance of each algorithm within tissue-specific chromatin states (**Figure 2B**). 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 associated with low methylation levels (**Additional file 1: Figure S4**).

Finally, we benchmarked the computational performance of all algorithms. BoostMe had a training runtime that was up to 40x faster than random forests using identical computational resources and up to 400x faster than DeepCpG (**Table 1**). Both BoostMe and random forests training times outperformed DeepCpG, which took multiple days to train 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 determine the effect of imputation on WGBS accuracy, we first characterized the concordance of WGBS and EPIC array methylation estimates at the same CpGs in the same samples (**Figure 3A**). We found that disagreement between the two platforms was concentrated at lower sequencing depth, with varying levels of discordance at high sequencing depth. Neither EPIC nor WGBS estimates can be considered the true methylation level of a particular CpG; however, since discordance between the two estimates was a function of sequencing depth, we hypothesized that discordance at low depth could be mostly attributed to WGBS inaccuracy, and discordance at high depth to EPIC inaccuracy.

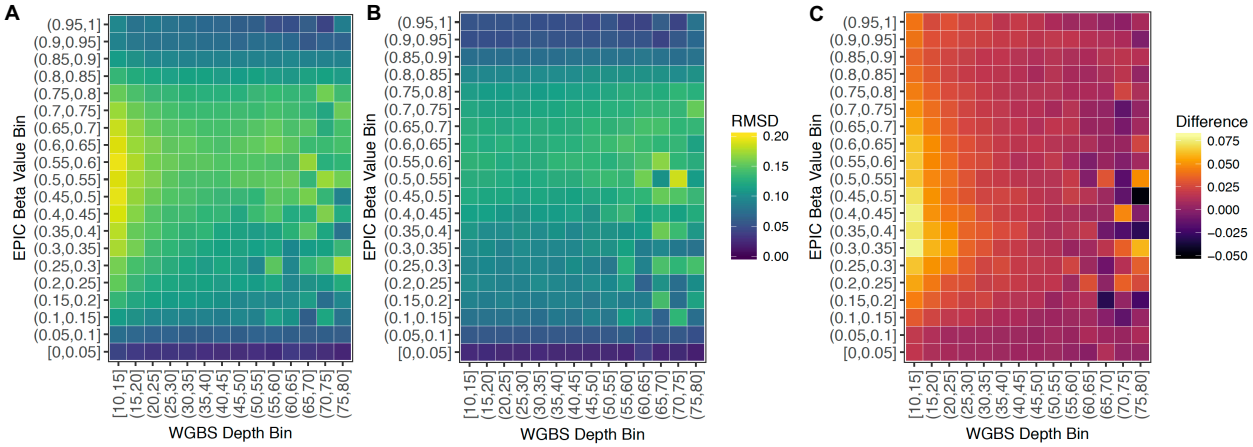

Figure 3 | Imputation reduces discordance between WGBS and EPIC methylation estimates at low sequencing depth. (A) Root-mean squared discordance (RMSD) between EPIC and WGBS methylation estimates at CpGs common between the two platforms. X-axis: depth at which the CpG was sequenced, binned into intervals of 5. Y-axis: the methylation (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.

We then used BoostMe to impute and replace CpGs common between the two platforms. We found that the discordance was mitigated (**Figure 3B**), particularly at lower WGBS depth (**Figure 3C**). 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 S5**). Discordance at higher depth was variable and in some cases increased after imputation (**Figure 3C**).

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 (**Figure 4, Additional file 1: Figures S6, S7**). Both algorithms highly prioritized the sample average and neighboring CpG features, which were well-correlated with the methylation value of the CpG of interest. To further characterize CpG-neighbor correlation, we measured the extent to which neighboring CpGs differed in methylation profile from the CpG of interest. We found that the difference was generally small (< 0.2) when the neighbors were within 50bp, but that a small fraction of CpG-neighbor pairs within this distance differed drastically (> 0.9) (**Additional file 1: Figure S8**).

We found that random forests ranked highly features that were negatively correlated with methylation values, especially those associated with open chromatin and promoter regions such as H3K4me3, ATAC-Seq peaks, CpG islands, and the TSS-related chromatin states (**Figure 4A**). Random forests also identified several methylation-associated transcription factors such as YY1, REST, and EP300. In concordance with previous results [52], 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 transcription factors identified, with all of them having some degree of overlap with each other (**Figure 4A**).

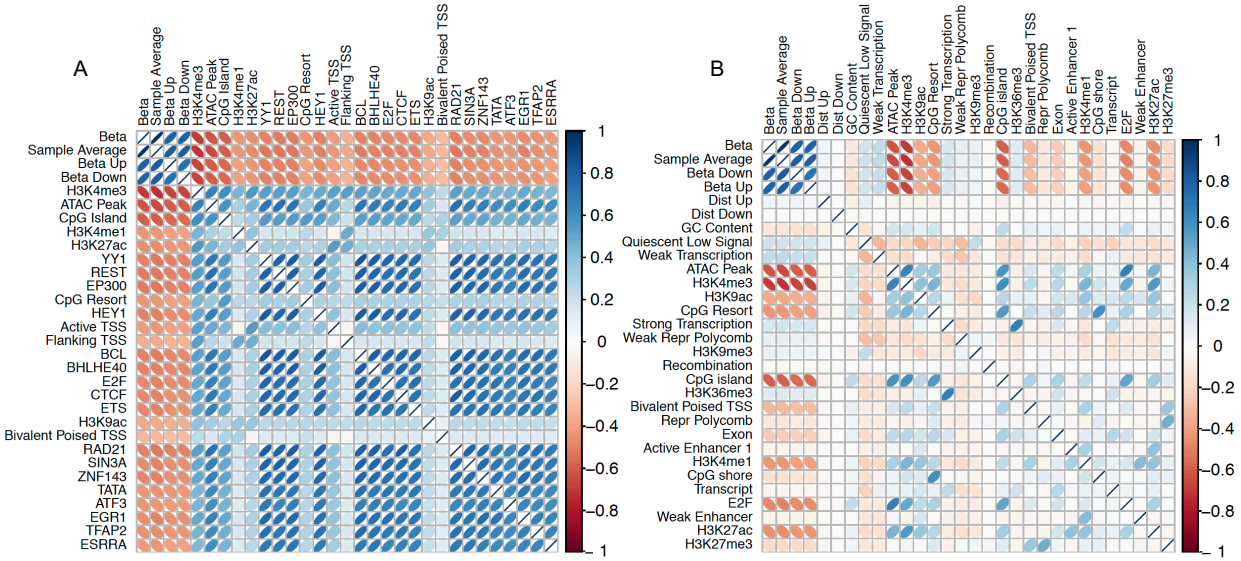

Figure 4 | Random forests exhibit greater bias in favor of positively correlated features compared to BoostMe. Correlation among methylation beta value and top 30 features in descending order for adipose T2D as reported by **(A)** random forests and **(B)** BoostMe. Ranking was determined by aggregating the variable importance scores across 10 runs from all adipose T2D samples. Beta - methylation value.

In contrast, BoostMe did not exhibit the same bias in favor of positively correlated features (**Figure 4B**). 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, BoostMe more highly prioritized chromatin states that were positively correlated with methylation such as "Quiescent/low signal" and "Weak transcription" in addition to highly predictive features that were negatively correlated with methylation identified by random forests, such as ATAC-Seq peaks and H3K4me3. BoostMe did not report as many methylation-associated transcription factors as 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 models using only transcription factor binding sites in tissue-specific regions of open chromatin. We found that the rankings were generally the same across tissues for both algorithms (**Additional file 1: Tables S3, S4**).

DeepCpG identifies sequence motifs important to general and tissue-specific methylation values

We examined the sequences that maximally activated filters from the first convolutional layer of the top-performing DNA modules from DeepCpG as described in Angermueller

et al. (**Methods**) [53]. We found that motif discovery using DeepCpG was inherently noisy, such that two networks trained on the same data could report slightly different motifs, or the most significantly matching TF to a particular motif could vary. To account for this noise, we measured the reproducibility of each TF discovery by training ten DNA models with identical architectures for each tissue and disease state combination (**Figure 5**). In our analysis, we consider only known TF motifs that were found in at least half of trained DNA modules. For a full, unfiltered list of TFs identified and corresponding motif logos, see **Additional file 3**.

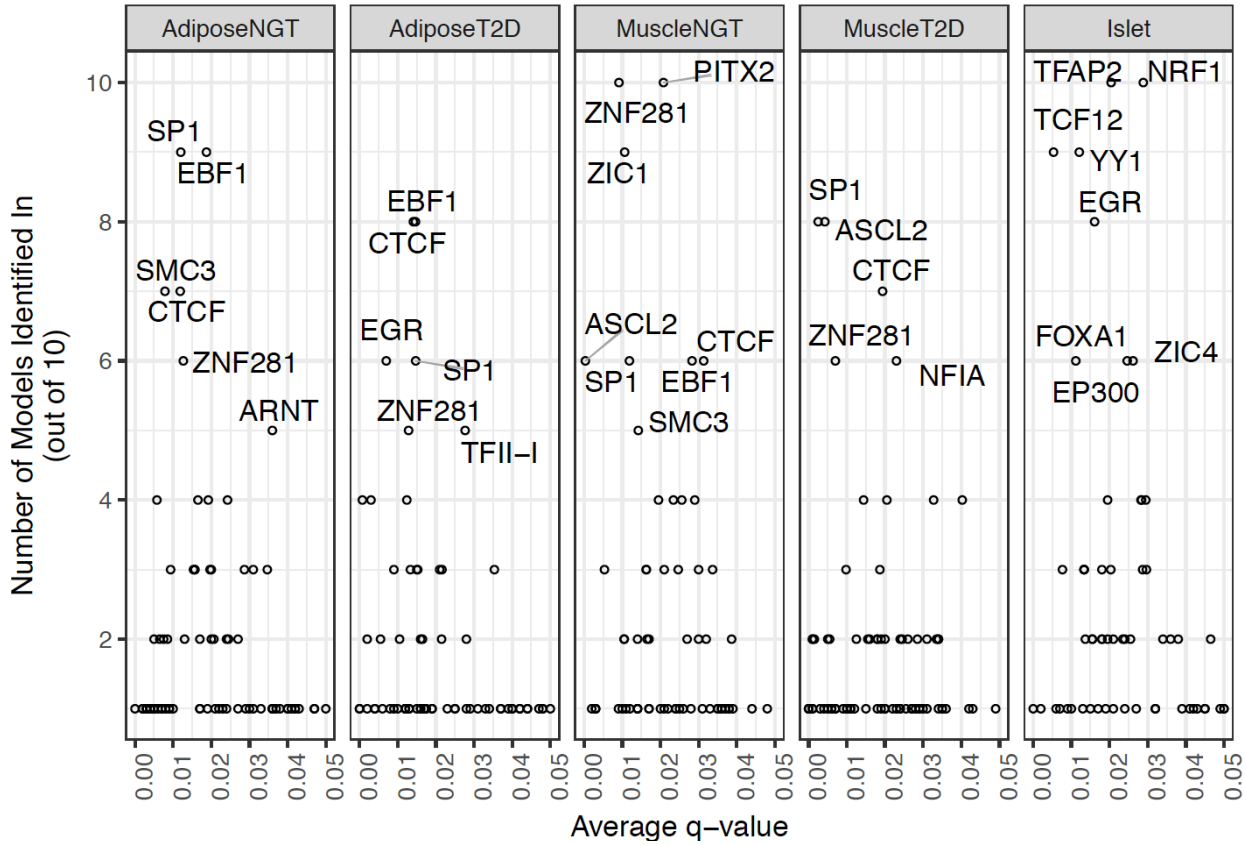

Figure 5 | Reproducible motifs identified by DeepCpG DNA models in different tissues and disease states. Known sequence motifs discovered for each tissue and disease state are plotted by (x-axis) the average q-value* of the motif across all DNA models it was discovered in and (y-axis) the total number of models the motif was identified in. *q-value is the p-value indicating the significance of similarity between the learned motif and the ENCODE [71] database match, corrected for false discovery rate.

Similar to BoostMe and random forests, DeepCpG learned sequence motifs important to methylation prediction across multiple tissues and disease states. For example, SP1, which was reproducible in adipose and muscle, has a GC-rich motif, is often associated with promoters of housekeeping genes, and is known to protect CpG islands from de novo methylation [24,65]. ZNF281 (ZBP-99), also discovered in adipose and muscle, binds GC-rich promoters, and may play important roles in regulating cell proliferation, differentiation, and oncogenesis [66,67]. CTCF, identified in four of the five tissues, is a multifunctional regulator of gene expression [68,69], and its binding may be influenced

by differential methylation [70]. The majority of all reproducible known motifs had a negative correlation with methylation values (**Figure 6**).

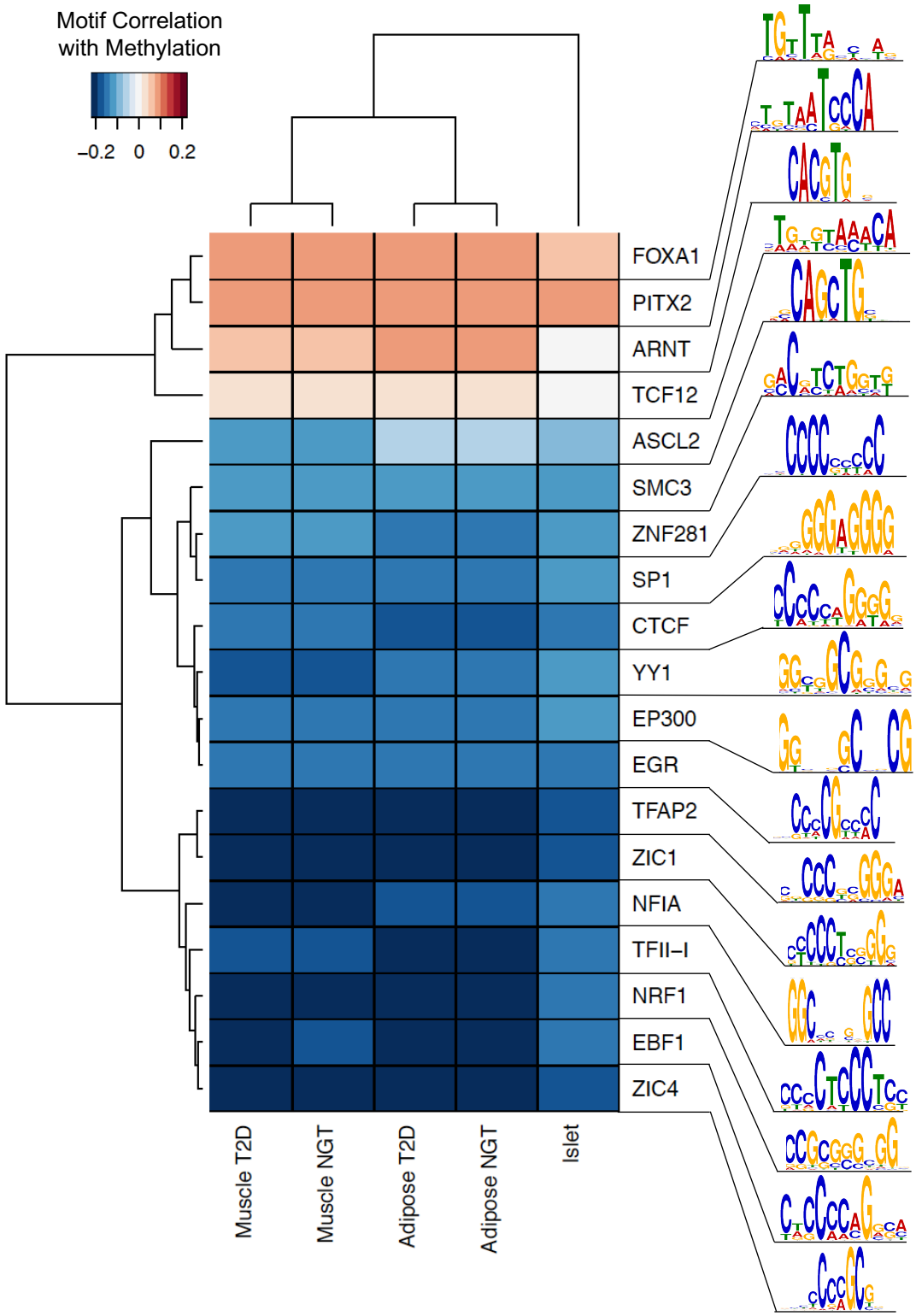

Figure 6 | Heat map of correlations between reproducible motifs and methylation values. Correlations were calculated as the Pearson's correlation coefficient between the weighted activations of

each motif from the first convolutional layer and the WGBS methylation value of CpGs (**Methods**). Motifs are labeled with their corresponding sequence logo representation of the learned DeepCpG DNA model filter. Hierarchical clustering shows that motifs fall into two major categories: GC-rich and negatively correlated with methylation, and less GC-rich and positively correlated with methylation.

Unlike BoostMe and random forests, DeepCpG was also able to learn motifs that are known to play key regulatory roles in their respective tissues based on previous biological findings. For example, EBF1 was highly reproducible in both adipose tissues, and is known to be an adipocyte-expressed transcription factor that regulates adipocyte morphology and lipolysis [72,73]. ASCL2, which may inhibit myogenesis by antagonizing myogenic regulatory factors [74], was identified in both muscle tissues. The same filter also matched with MYOD, which is considered the master regulator of myogenesis [75,76]. Finally, in pancreatic islets, DeepCpG identified FOXA1, which plays a role in maintaining the metabolic and secretory features of beta cells [77, 78]. The same filter also matched significantly to FOXA2, which is required for the differentiation of pancreatic alpha cells and controls PDX1 gene expression in pancreatic beta cells [78,79]. DeepCpG also identified TCF12, which is involved in pancreatic development [81, 82], and NRF1, which plays a role in beta cell insulin secretion and glucose metabolism [83, 84].

We also detected reproducible differences in motifs between NGT and T2D. For example, PITX2, a key regulator of skeletal myogenesis [85, 86], was highly reproducible in Muscle NGT and absent in Muscle T2D (**Figure 5**). However, neither PITX2 nor any other motifs showed significant differences in their correlation with methylation between the NGT and T2D states (**Figure 6**).

Discussion

To better understand the extent to which DNA methylation regulates tissue-specific gene expression, and the specific contribution of methylation dysregulation to T2D etiology, we generated WGBS and EPIC data from 58 samples of human adipose, skeletal muscle, and pancreatic islets. We discovered that, despite the relatively deep mean sequence coverage in these samples (~40x), our ability to interpret these data was limited by a relatively high rate of missingness (**Figure 1, Additional file 1: Figure S1, S2**) and by significant bias in the level of discordance between methylation estimates from WGBS versus the EPIC array (**Figure 3**). To address these limitations, we developed BoostMe, an imputation method for WGBS that improves upon previous algorithms by (1) using a gradient boosting algorithm based on XGBoost [56], (2) leveraging a new sample average feature, and (3) training on and predicting continuous methylation values. BoostMe outperformed both random forests and DeepCpG in computational efficiency while achieving similar accuracy to random forests across tissues and chromatin states (**Figure 2, Table 1**).

Since 70-80% of CpGs in the genome are invariantly methylated across tissues and samples [48,87], it is no surprise that most of the gains in accuracy came from using the sample average feature. For this reason, we also benchmarked all algorithms on intermediate CpG methylation values, which are more likely to be biologically significant

and vary across tissues and samples. Both tree-based algorithms still had lower RMSE than DeepCpG (**Additional file 1: Table S3**), despite DeepCpG being designed to detect sample-specific differences. However, DeepCpG was also originally optimized for single-cell bisulfite sequencing, not WGBS. Thus, we expect that a deep neural architecture designed specifically for WGBS data would perform better and will explore such developments in a future work.

To assess the effect of imputation on the quality of WGBS data, we compared WGBS and imputed methylation estimates from BoostMe to EPIC data from the same samples. Interestingly, we find that imputation mitigates discordance between WGBS and EPIC estimates of the same CpGs, particularly at low coverage sites (**Figure 3**). This result suggests an improvement in the quality of WGBS estimates, making BoostMe a potentially beneficial preprocessing step for WGBS data pipelines. However, we did not attempt to validate that this result holds for non-EPIC sites.

One limitation of BoostMe is that, like previous methodology [52,53], it relies heavily on the locally correlated structure of neighboring CpGs to identify sample-specific differences. Average methylation haplotype blocks, defined as consecutive CpGs with methylation correlation of $r^2 > 0.5$, measure just 95 bp long on average [88]. To achieve the goal of highly accurate methylation imputation from sparse data, further work must be done to develop methodology that can detect sample-specific differences without depending as heavily on neighboring CpGs. For example, the DNA module of DeepCpG, which uses only DNA sequence to predict methylation levels, could be improved to achieve higher accuracy in CpG-sparse regions, eliminating the need for dense CpG input.

We also investigated the ability of all algorithms to identify differences in features important to methylation across tissues and disease states. We found that the variable importance score rankings output by BoostMe and random forests do not readily identify tissue- or disease-specific features. Instead, both algorithms rank highly features important to general methylation levels such as ATAC-Seq peaks and other measures of open chromatin, which are negatively correlated with methylation values, as well as features associated with inaccessible chromatin, which are positively correlated with methylation values (**Figure 4**).

Consistent with previous experimental findings, random forests identified methylation-sensitive TFs such as YY1, REST, and EP300 [30, 89-92]. However, many of the features identified by random forests were also positively correlated with each other (**Figure 4**), which is a known bias of random forests [52,93]. Although there are existing methods to correct for this bias such as conditional random forests [93], they scale poorly to large data sets with hundreds of variables. BoostMe does not have the same bias to highly rank positively correlated features, but it also does not specifically correct for the correlation when it calculates variable importance [56]. Instead, BoostMe highly ranks just one feature from a group of positively correlated features, making it unclear which feature or features are actually important.

DeepCpG, on the other hand, learned important sequence motifs in different tissues related to methylation prediction without any feature construction (**Figure 5**), thus representing an unbiased approach to characterizing methylation regulatory networks. DeepCpG learned common motifs such as SP1, ZNF281, and CTCF, which have roles in cell differentiation and development [24, 65-70]. This result reinforces the previously known importance of methylation in these biological processes and validates the ability of DeepCpG to identify motifs that are specifically associated with methylation. DeepCpG also identified important tissue-associated motifs such as EBF1 in adipose, ASCL2 in muscle, and FOXA1, TCF12, and NRF1 in pancreatic islets, which have also been shown to be involved in regulating differentiation and development in their respective cell types [72-86].

To see whether BoostMe or random forests could reproduce the same findings as DeepCpG, we trained both algorithms using only TFBS. Neither algorithm identified notable differences among the tissues and disease states (**Additional file 1: Tables S4, S5**). Therefore, despite outperforming DeepCpG in methylation prediction RMSE and speed, neither BoostMe nor random forests are ideal algorithms for learning differences in methylation regulation among tissues.

With DeepCpG, we were able to identify interesting candidates for methylation regulatory networks in different tissues supported by previous experimental results. This result highlights the promise of supervised deep learning methods to identify biological signatures in genome-wide data. However, these results are limited by the noise in DeepCpG outputs due to motif similarity and stochasticity in the model training process. We attempted to address this limitation by training multiple DNA models with identical architectures on each tissue and disease state combination (**Figure 5**), however, it is possible that all motifs identified by all networks have the potential to be significant, not just those that are reproducible.

We attempted to identify significant differences between the NGT and T2D states using motif correlation with methylation (**Figure 6**). Using this measure, we found no differences between the NGT and T2D adipose and muscle. However, the correlation does not take into account the complex, non-linear relationships learned by the deep neural network during training. Understanding these relationships is an active area of research. It is possible that there are reasons why DeepCpG identified motifs reproducible in NGT but not T2D, but that these reasons cannot be interrogated with correlation alone. Due to these limitations, important questions such as the extent to which methylation plays a causal role in altering gene expression cannot be answered from our results. Additional experimental and computational work must be done to understand the mechanisms by which these TFs influence methylation and the T2D state.

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 [94-97] 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 [98]. Biopsies were performed by 9 experienced and well-trained physicians from 2009-2013 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 [5]. 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 minutes. 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 125bp sequencing was performed on Illumina HiSeq 2500 instruments to 30X genome coverage.

EPIC array

Genomic DNA was extracted from each tissue using DNeasy Blood and Tissue Kits (QIAGEN), according to the manufacturer's recommendations. 200ng 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 [99]. Adapter sequences were trimmed using Atropos [100], and reads with at least one pair shorter than 25 bp were excluded. Reads were aligned to the reference genome (GRCh37) using BWA MEM [101], followed by Samblaster [102] for marking duplicates.

WGS variant calling
SNPs and indels were called separately for all sample BAM files using GATK HaplotypeCaller [103]. 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 [104]. Methylation values were extracted using the MethylDackel 'extract' command, including bias correct 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 [105] 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 [106]), distribution of the overall methylation profile, and principle component analysis. None of the WGBS 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 [107] and Illumina Omni2.5 array genotypes.

For both the earlier 450k and recent EPIC Illumina methylation array, previous studies [108-111] 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 overlaps the single base extension site. We used 10 bp as this cutoff is consistent with previous studies [109]. 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 HRC reference panel. As a final step, we combined our blacklist with a previously published blacklist [111] 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 2,206 islet probes based on the per sample filters.

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 S2** 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. We also excluded all CpGs located in ENCODE blacklist regions [112].

CpG features

Features constructed from the WGBS data included neighboring CpG methylation values and the sample average feature. Neighboring CpG methylation values were taken within the sample of interest. For each neighbor, the methylation value as well as the base-pair distance from the neighbor to the CpG of interest were included as features. The sample average feature was created by taking the average 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 10x coverage were excluded from the calculation. CpGs without a measurement above 10x 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 [113,114]. DNA recombination rate annotations from HapMap were downloaded from the UCSC hg19 annotation database (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/). CpG island coordinates were obtained from UCSC browser. CpG island shores and shelves were calculated from CpG island 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 models for 31 diverse tissues that included islets, skeletal muscle, and adipose [5]. 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 [115-118]. ATAC-seq data was obtained from previously published studies for islets [5], skeletal muscle [4], and adipose [87]. TFBS data was obtained as described in [4], with additional PWMs from [119]. 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) [56] in R [120] (version 3.3.1). Hyperparameters were tuned using a random search with the tuneParams function in the mlr package [121], however we found that performance did not improve significantly over the default settings of the xgb.train function in the xgboost package. For random forests we used the ranger package (version 0.6.0) in R, which facilitates random forest training and testing on multiple CPUs [122]. In the final algorithm, we set num_trees to 500 to balance computational time and accuracy, and used default values for other parameters after finding that performance was robust to different settings.

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 10x coverage and no more than 80x 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 memory limits, algorithms were trained on a random sample of 1,000,000 CpGs from all available CpGs within a sample, validated on a hold-out set of 500,000 CpGs, and benchmarked on another hold-out set of 500,000 CpGs. To account for possible biases that would arise from only training on a small subset of the over 20 million CpGs available for training from each sample, we repeated the process of randomly sampling CpGs for training, validation, and testing ten times for each sample using ten different random seeds and averaged the results.

Importantly, including all features was detrimental to the performance of the random forest algorithm. This effect was observed to a much lesser extent in BoostMe because trees are grown sequentially to fit the error of the previous tree, allowing BoostMe to ignore less-predictive variables more easily than the bootstrap aggregating algorithm used by random forest [56]. Therefore, to determine the combination of features for both algorithms, we added features in a stepwise fashion and compared the RMSE of each combination (**Additional file 2**). The best performance was achieved when we included

the sample average feature, neighboring CpG values and distances, histone marks, and recombination rate; we used this abbreviated algorithm to benchmark performance. For the purpose of identifying interesting features (**Figure 4**), all features were included.

WGBS methylation values follow a bimodal distribution, with the vast majority of CpGs being fully methylated or unmethylated. To see if we could improve prediction estimates at intermediate CpGs, we altered the sampling distribution to be more uniform instead of bimodal, randomly drawing an equal number of CpGs from each methylation bin for training (**Additional file 1: Figure S3**). We found no significant improvements in prediction.

DeepCpG implementation

Model training and testing

We implemented DeepCpG (version 1.0.4) as described in Angermueller *et al.* (2017) [53]. 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 find that this change made no difference in the accuracy of the model (**Additional file 1: Table S3**).

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 (**Additional file 2**) and reported the motifs significantly matching the filters from the first convolutional layer of that model [123]. 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, AUC, AUPRC, and accuracy for each tissue. We used a default learning rate of 0.001 for all models. For a detailed explanation of all model architectures, see http://deepcpg.readthedocs.io/.

Motif reproducibility and importance quantification

Filters of the first convolutional layer from the DNA module were visualized as described previously [53]. Briefly, selected sequence windows were aligned and visualized as sequence motifs using WebLogo [124] version 3.4. Discovered motifs were matched to annotated motifs in the *Homo sapiens* CIS-BP database using Tomtom 4.11.1 from both the MEME Suite [125], which is the default for DeepCpG, and the ENCODE [71] database. In the final results, we used only the ENCODE matches to parallel our use of

ENCODE TFBS for the random forest and BoostMe algorithms. Matches at FDR < 0.05 were considered significant. For consistency, we reported only the most significant motif match for each filter, however in some cases multiple motifs matched significantly to a single filter due to sequence similarity. For a complete list of significantly-matching motifs and sequence logos see **Additional file 3.**

Motif importance was quantified using the Pearson correlation coefficient with methylation values as previously described [53]. Briefly, we used DeepCpG to extract the activations of each filter from the first convolutional layer. We then averaged these activities across the 1,001bp window centered at the CpG of interest for a sample of 50,000 CpGs. The mean activations were then weighted by a linear weighting function that assigns the highest relative weight to the center position. We then calculated the Pearson correlation between the mean activations and the actual methylation value as measured by WGBS. CpGs lying in CpG islands were excluded from this analysis.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
JPD conceived of the study. LSZ developed the software and ran experiments. LSZ, JPD, and FSC wrote the manuscript. MRE and MGC generated the data. DLT, PSC, AV, and SCJP contributed to data analysis. All authors read and approved the final manuscript.

Code Availability
We will release BoostMe as an R package through Github and Bioconductor upon publication.

Acknowledgements
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.
The authors would like to thank Christof Angermueller for his assistance with DeepCpG and members of the Collins laboratory for helpful discussions.
This manuscript is dedicated to Peter Chines.

References
1. 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.
2. McCarthy MI, Zeggini E. Genome-wide association studies in type 2 diabetes. Curr. Diab. Rep. 2009;9:164-71.
3. 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.
4. 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.
5. 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.
6. Pasquali L, Gaulton KJ, Rodríguez-Seguí SA, Mularoni L, Miguel-Escalada I, et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nat. Genet. 2014;46:136-43.
7. Robertson KD. DNA methylation and human disease. Nat. Rev. Genet. 2005;6:597-610.
8. Lim U, Song MA. Dietary and lifestyle factors of DNA methylation. Methods Mol. Biol. 2012;863:359-76.
9. Aguilera O, Fernández AF, Muñoz A, Fraga MF. Epigenetics and environment: a complex relationship. J. Appl. Physiol. 2010;109:243-51.
10. Das S. Integrating transcriptome and epigenome: putting together the pieces of the type 2 diabetes pathogenesis puzzle. Diabetes. 2014;63:2901-3.
11. Scully T. Diabetes in numbers. Nature. 2012;485:S2-3.
12. Ling C, Groop L. Epigenetics: a molecular link between environmental factors and type 2 diabetes. Diabetes. 2009;58:2718-25.
13. Waki H, Yamauchi T, Kadowaki T. The epigenome and its role in diabetes. Curr. Diab. Rep. 2012;12:673-85.
14. Suzuki MM, Bird AP. DNA methylation landscapes: provocative insights from epigenomics. Nat. Rev. Genet. 2008;9:465-76.
15. Relton CL, Davey Smith G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. Int. J. Epidemiol. 2012;41:161-76.
16. Michels KB, Binder AM, Dedeurwaerder S, Epstein CB, Greally JM, Gut I, et al. Recommendations for the design and analysis of epigenome-wide association studies. Nat. Methods. 2013; 10:949-55.
17. van Dijk S, Tellam RL, Morrison JL, Muhlhausler BS, Molloy PL. Recent developments on the role of epigenetics in obesity and metabolic disease. Clin. Epigenetics. 2015;7:66.
18. Maurano MT, Wang H, John S, Shafer A, Canfield T, Lee K, Stamatoyannopoulos JA. Role of DNA methylation in modulating transcription factor occupancy. Cell. 2015;12:1184-95.
19. Rishi V, Bhattacharya P, Chatterjee R, Rozenberg J, Zhao J, Glass K, et al. CpG methylation of half-CRE sequences creates C/EBPα binding sites that activate some tissue-specific genes. Proc. Natl. Acad. Sci. 2010;107:20311-6.
20. Tate PH, Bird AP. Effects of DNA methylation on DNA-binding proteins and gene expression. Curr. Opin. Genet. Dev. 1993;3:226-31.
21. Medvedeva YA, Khamis AM, Kulakovskiy IV, Ba-Alawi W, Bhuyan MSI, Kawaji H, et al. Effects of cytosine methylation on transcription factor binding sites. BMC Genomics. 2014;15:119.

22. Hu S, Wan J, Su Y, Song Q, Zeng Y, Nguyen HN, et al. DNA methylation presents distinct binding sites for human transcription factors. Elife. 2013;2:e00726.
23. Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science. 2017;356:eaaj2239.
24. Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Nemes A, et al. Sp1 elements protect a CpG island from de novo methylation. Nature. 1994;371:435-8.
25. Feldmann A, Ivanek R, Murr R, Gaidatzis D, Burger L, Schübeler D. Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. PLoS Genet. 2013;9:e1003994.
26. Liener F, Wirbelauer C, Som I, Dean A, Mohn F, Schübeler D. Identification of genetic elements that autonomously determine DNA methylation states. Nat. Genet. 2011;43:1091-7.
27. Lin IG, Tomzynski TJ, Ou Q, Hsieh CL. Modulation of DNA binding protein affinity directly affects target site demethylation. Mol. Cell Biol. 2000;20:2343-9.
28. Macleod D, Charlton J, Mullins J, Bird AP. Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes Dev. 1994;8:2282-92.
29. Matsuo K, Silke J, Georgiev O, Marti P, Giovannini N, Rungger D. An embryonic demethylation mechanism involving binding of transcription factors to replicating DNA. EMBO J. 1998;17:1446-53.
30. 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.
31. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489:75-82.
32. Blattler A, Farnham PJ. Crosstalk between site-specific transcription factors and DNA methylation states. J. Biol. Chem. 2013;288:34287-94.
33. Jimenez-Useche I, Ke J, Tian Y, Shim D, Howell SC, Qiu X, Yuan C. DNA methylation regulated nucleosome dynamics. Sci. Rep. 2013;3:2121.
34. Du J, Johnson LM, Jacobsen SE, Patel DJ. DNA methylation pathways and their crosstalk with histone methylation. Nat. Rev. Mol. Cell Biol. 2015;16:519-32.
35. Bird AP. DNA methylation patterns and epigenetic memory. Genes Dev. 2002;16:6-21.
36. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat. Rev. Genet. 2013;14:204-20.
37. Trerotola M, Relli V, Simeone P. Alberti S. Epigenetic inheritance and the missing heritability. Hum. Genomics. 2015;9:17.
38. Heard E, Martienssen RA. Transgenerational epigenetic inheritance: myths and mechanisms. Cell. 2014;157:95-109.
39. Lim JP, Brunet A. Bridging the transgenerational gap with epigenetic memory. Trends Genet. 2013;29:176-86.
40. Dayeh T, Tuomi T, Almgren P, Perfilyev A, Jansson PA, de Mello VD, et al. DNA methylation of loci within ABCG1 and PHOSPHO1 in blood DNA is associated with future type 2 diabetes risk. Epigenetics. 2016;11:482-8.

41. Kriebel J, Herder C, Rathmann W, Wahl S, Kunze S, Molnos S, et al. Association between DNA methylation in whole blood and measures of glucose metabolism: KORA F4 study. PLoS One. 2016; doi:10.1371/journal.pone.0152314.
42. Bacos K, Gillberg L, Volkov P, Olsson AH, Hansen T, Pedersen O, et al. Blood-based biomarkers of age-associated epigenetic changes in human islets associate with insulin secretion and diabetes. Nat. Commun. 2016;7:11089.
43. Nilsson E, Jansson PA, Perfilyev A, Volkov P, Pedersen M, Svensson MK. Altered DNA methylation and differential expression of genes influencing metabolism and inflammation in adipose tissue from subjects with type 2 diabetes. Diabetes. 2014;63:2962-76.
44. Dayeh T, Volkov P, Salö S, Hall E, Nilsson E, Olsson AH, et al. Genome-wide DNA methylation analysis of human pancreatic islets from type 2 diabetic and non-diabetic donors identifies candidate genes that influence insulin secretion. PLoS Genet. 2014;10:e1004160.
45. Illumina Support. http://support.illumina.com.
46. Benton MC, Johnstone A, Eccles D, Harmon B, Hayes MT, Lea RA, et al. An analysis of DNA methylation in human adipose tissue reveals differential modification of obesity genes before and after gastric bypass and weight loss. Genome Biol. 2015;16:8.
47. Volkov P, Bacos K, Ofori JK, Esguerra JL, Eliasson L, Rönn T, Ling C. Whole-genome bisulfite sequencing of human pancreatic islets reveals novel differentially methylation regions in type 2 diabetes pathogenesis. Diabetes. 2017;66:1074-85.
48. Ziller MJ, Hansen KD, Meissner A, Aryee MJ. Coverage recommendations for methylation analysis by whole genome bisulfite sequencing. Nat. Methods. 2015;12:230-2.
49. 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.
50. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu. Rev. Genomics Hum. Genet. 2009;10:387-406.
51. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 2010;11:499-511.
52. 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.
53. 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.
54. Breiman L. Random forests. Machine Learning. 2001;45:5.
55. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436-44.
56. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. 2016; arXiv:1603.02754v3 [cs.LG]
57. 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.
58. 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.

59. Zeng H, Gifford DK. Predicting the impact of non-coding variants on DNA methylation. Nucleic Acids Res. 2017;45:e99.
60. 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.
61. 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.
62. 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.
63. 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.
64. Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat. Biotech. 2015;33:364-76.
65. Höller M, Westin G, Jiricny J, Schaffner W. Sp1 transcription factor binds DNA and activates transcription even when the binding site is CpG methylated. Genes Dev. 1988;2:1127-35.
66. Law DJ, Du M, Law GL, Merchant JL. ZBP-99 defines a conserved family of transcription factors and regulates ornithine decarboxylase gene expression. Biochem. Biophys. Res. Commun. 1999;262:113-20.
67. Hahn S, Hermeking H. ZNF281/ZBP-99: a new player in epithelial-mesenchymal transition, stemness, and cancer. J. Mol. Med. (Berl). 2014;92:571-81.
68. Herold M, Bartkuhn M, Renkawitz R. CTCF: insights into insulator function during development. Development. 2012;6:1045-57.
69. Kim S, Yu N, Kaang B. CTCF as a multifunctional protein in genome regulation and gene expression. Exp. Mol. Med. 2015;6:e166.
70. Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res. 2012;9:1680-8.
71. Kheradpour P, Kellis M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. 2014;42:2976-87.
72. 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.
73. 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.
74. 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.
75. Tapscott SJ. The circuitry of a master switch: Myod and the regulation of skeletal muscle gene transcription. Development. 2005;132:2685-95.
76. Bentzinger CF, Wang YX, Rudnicki MA. Building muscle: molecular regulation of myogenesis. Cold Spring Harb. Perspect. Biol. 2012;4:a008342.

77. 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.
78. 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.
79. Lee CS, Sund NJ, Behr R, Herrera PL, Kaestner KH. Foxa2 is required for the differentiation of pancreatic alpha-cells. Dev. Biol. 2005;278:484-95.
80. Lee CS, Sund NJ, Vatamaniuk MZ, Matschinsky FM, Stoffers DA, Kaestner KH. Foxa2 controls Pdx1 gene expression in pancreatic beta-cells in vivo. Diabetes. 2002;51:2546-51.
81. White P, May CL, Lamounier RN, Brestelli JE, Kaestner KH. Defining pancreatic endocrine precursors and their descendants. Diabetes. 2008;57:654-68.
82. Benitez CM, Goodyer WR, Kim SK. Deconstructing pancreas developmental biology. Cold Spring Harb. Perspect. Biol. 2012;6:a012401.
83. Zheng H, Fu J, Xue P, Zhao R, Dong J, Liu D, et al. CNC-bZIP protein Nrf1-dependent regulation of glucose-stimulated insulin secretion. Antioxid. Redox. Signal. 2015;10:819-31.
84. Kaufman BA, Li C, Soleimanpour SA. Mitochondrial regulation of beta-cell function: maintaining the momentum for insulin release. Mol. Aspects Med. 2015;42:91-104.
85. L'Honoré A, Coulon V, Marcil A, Lebel M, Lafrance-Vanasse J, et al. Sequential expression and redundancy of Pitx2 and Pitx3 genes during muscle development. Dev. Biol. 2007;2:421-33.
86. Hernandez-Torres F, Rodríguez-Outeriño L, Franco D, Aranega AE. Pitx2 in embryonic and adult myogenesis. Front. Cell Dev. Biol. 2017;5:46.
87. 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.
88. 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.
89. 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.
90. 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.
91. 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.
92. 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.
93. Strobl C, Boulesteix A, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. 2008;9:307.

94. 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.
95. 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.
96. 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.
97. 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.
98. World Health Organization (WHO), International Diabetes Federation (IDF). Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF consultation. 2006; WHO, Geneva, Switzerland.
99. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010; available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
100. Didion JP, Martin M, Collins FS. Atropos: specific, sensitive, and speedy trimming of sequencing reads. PeerJ. 2017;5:e3720.
101. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013; arXiv:1303.3997 [q-bio.GN]
102. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 2014;30:2503-5.
103. 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.
104. Pedersen BS, Eyring K, De S, Yang IV, Schwartz DA. Fast and accurate alignment of long bisulfite-seq reads. 2014; arXiv:1401.1129 [q-bio.GN]
105. 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.
106. Fortin JP, Fertig E, Hansen K. shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R. F1000Res. 2014;3:175.
107. 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.
108. 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.
109. 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.
110. 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.

111. 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.
112. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57-74.
113. Golden path track of the University of Santa Cruz Genome Browser. http://hgdownload.cse.ucsc.edu/goldenPath/hg19/gc5Base/
114. 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.
115. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317-30.
116. 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.
117. 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.
118. 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.
119. 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.
120. R project. http://www.r-project.org/
121. Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, et al. mlr: Machine Learning in R. J. Machine Learning Research. 2016;17:1-5.
122. Wright M, Ziegler A. ranger: A fast implementation of random forests for high dimension data in C++ and R. J. Statistical Software. 2017;77:1-17.
123. 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.
124. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14:1188-90.
125. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37:W202-8.

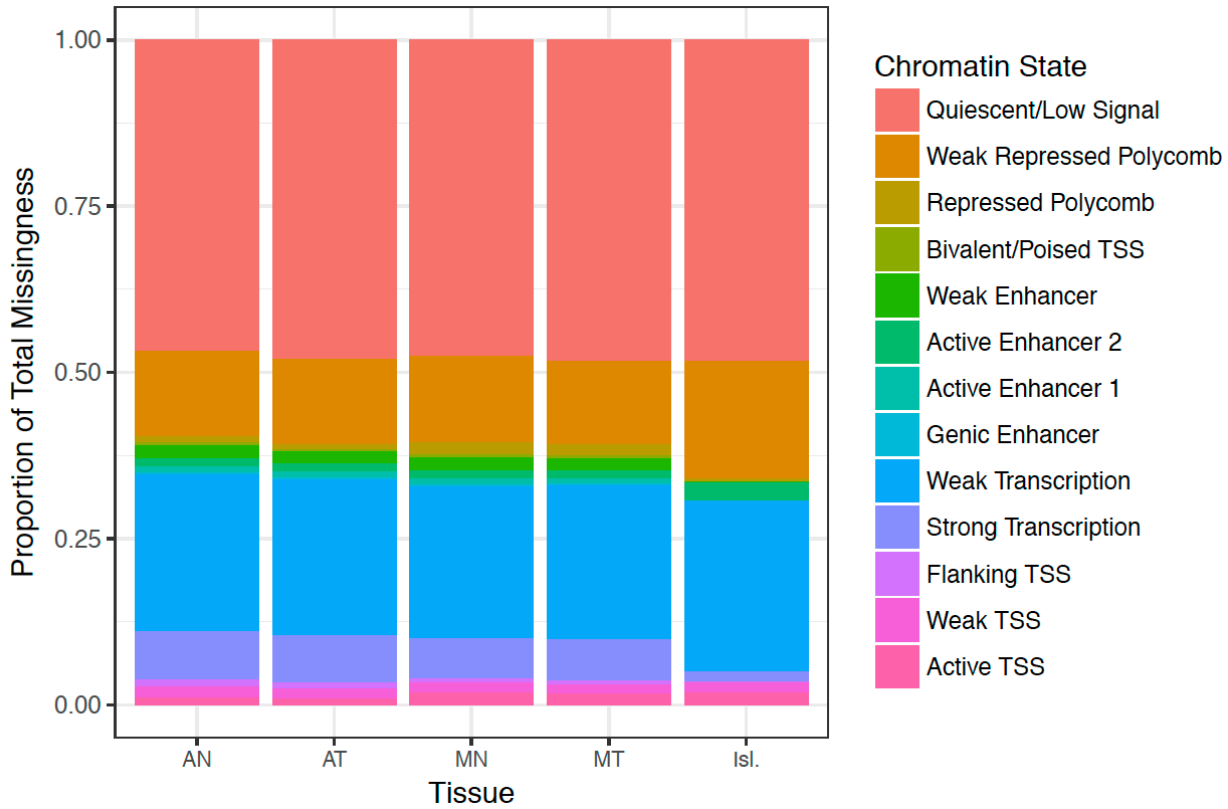

Figure S1: Distribution of WGBS missingness across chromatin states, not normalized for total number of CpGs in that chromatin state. The proportion of total missingness was calculated for each tissue as the number of missing CpGs (sequencing depth < 10x) in that chromatin state divided by the total number of missing CpGs in that tissue.

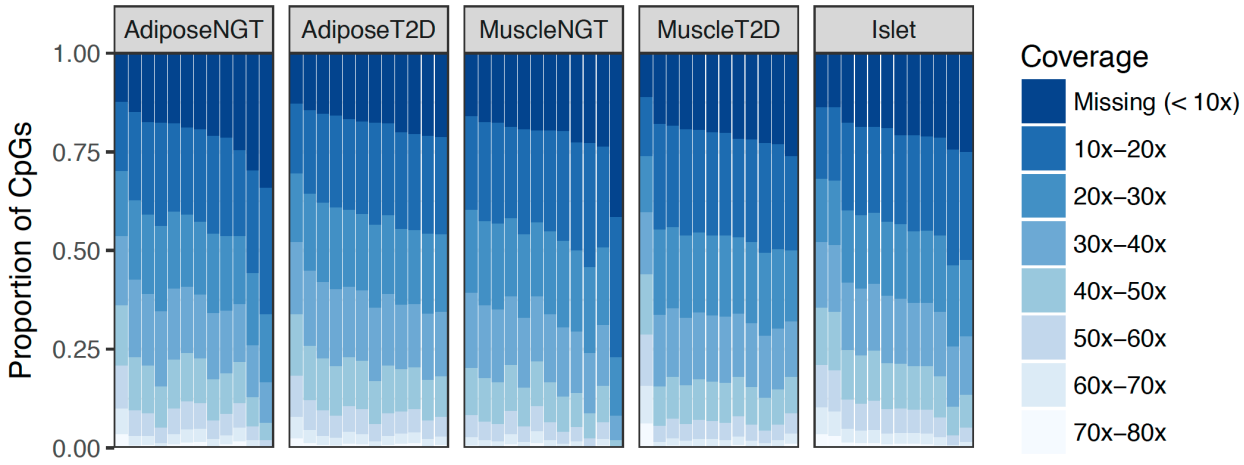

Figure S2: WGBS depth distributions across tissue types and samples. Each column represents one sample from that tissue.

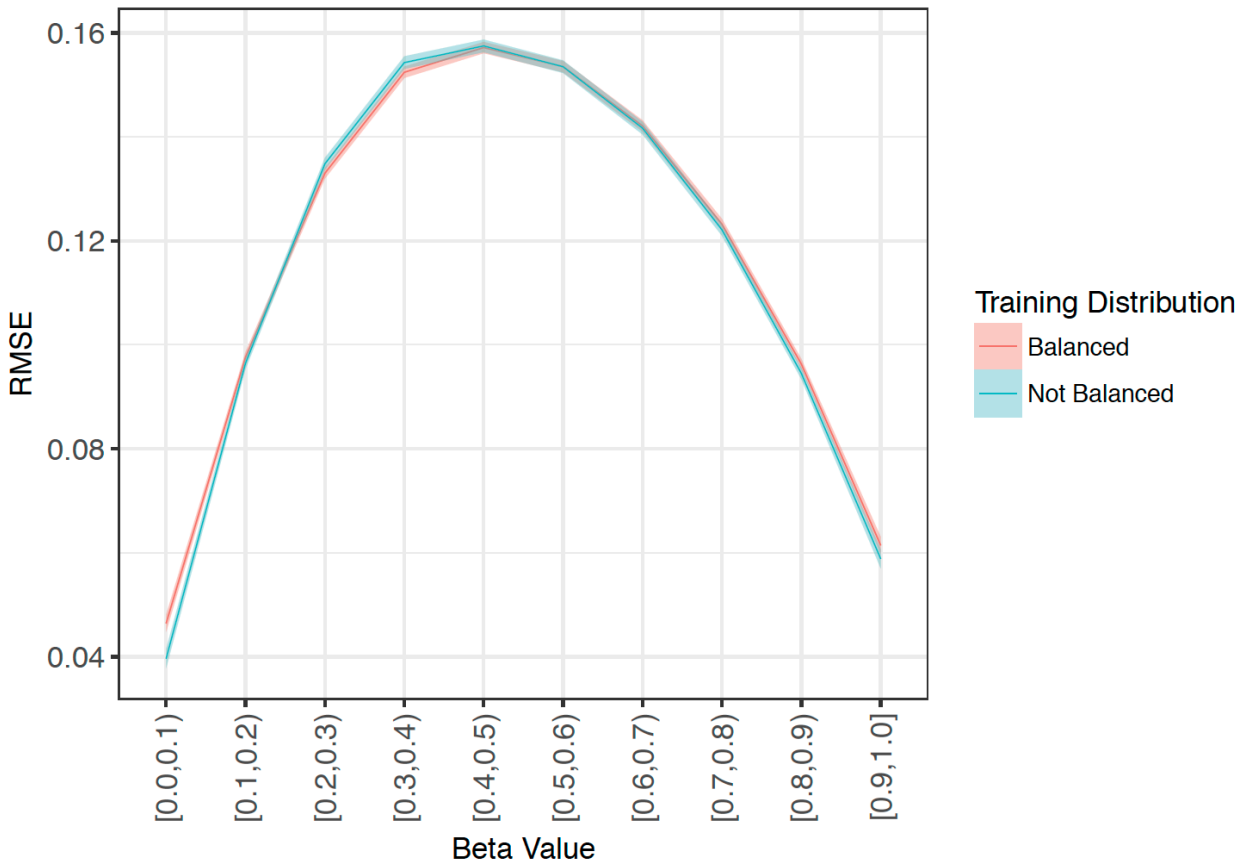

Figure S3: Performance at some intermediate CpGs improves slightly when using a balanced training distribution. Root-mean squared error (RMSE) binned by the average WGBS methylation value of a CpG (sample average) when training on a balanced (red) vs. not balanced (blue) training distribution. Shaded areas indicate 95% confidence interval (n=58).

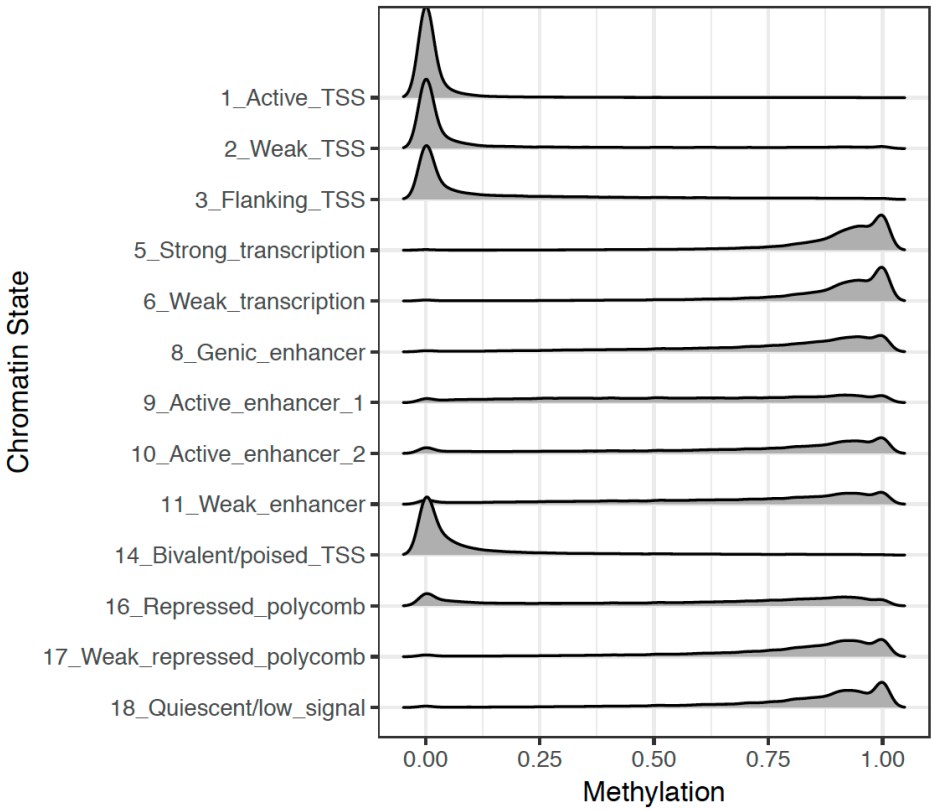

Figure S4: Distribution of WGBS methylation levels across chromatin states in all tissues. TSS-associated states are strongly associated with low methylation levels.

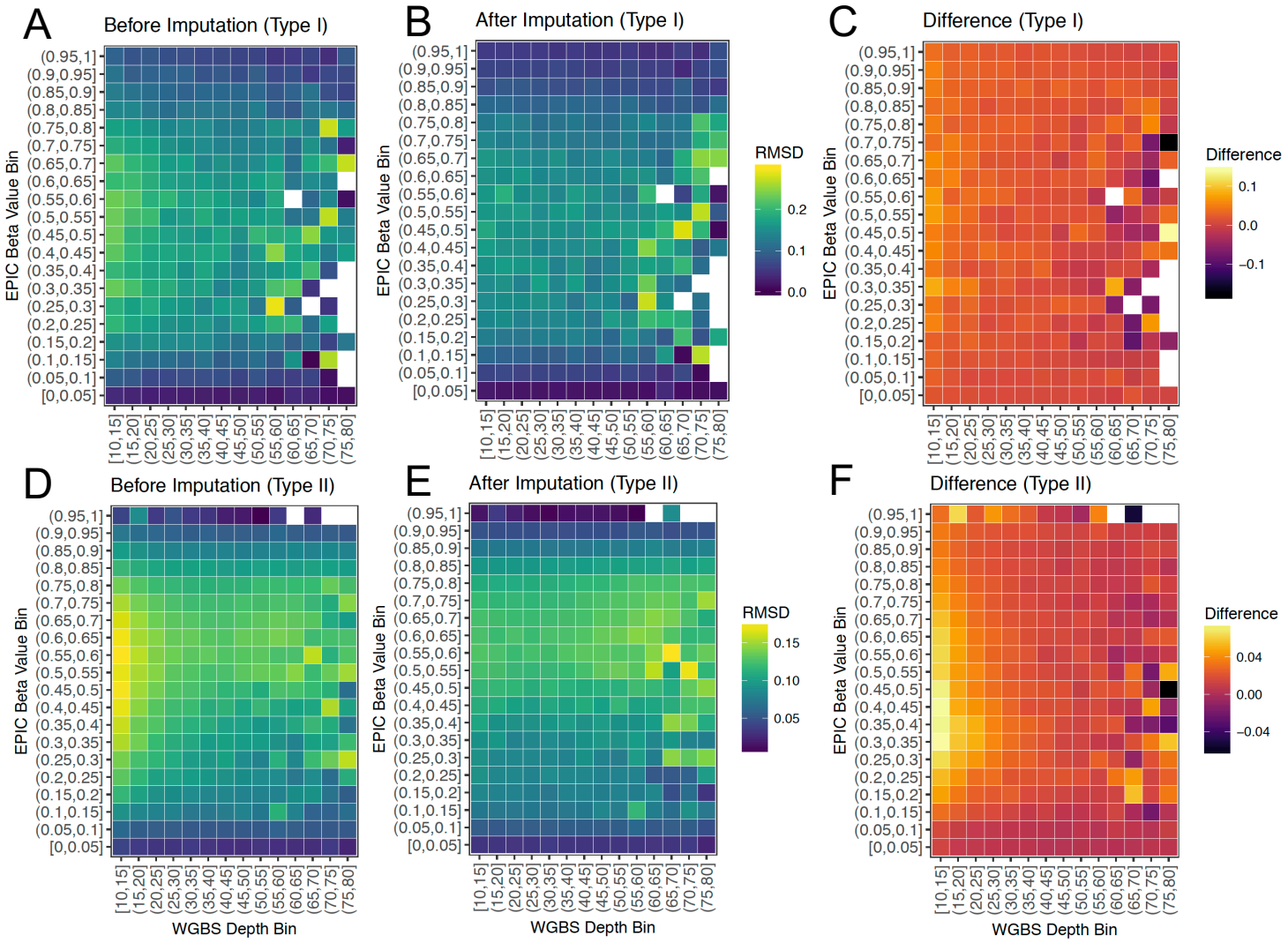

Figure S5: Imputation mitigates discordance between WGBS at EPIC at low WGBS depth regardless of EPIC probe type. (A) Root-mean squared discordance (RMSD) between EPIC and WGBS estimates at CpGs common between the two platforms, binned by WGBS depth of the CpG and EPIC beta value of the CpG for Type I EPIC probes only. **(B)** RMSD between EPIC and imputed WGBS values at the same CpGs as in A. **(C)** Difference between A and B. **(D-F)** Same as A-C for Type II EPIC probes only. Empty bins (white) indicate that no CpGs were present in that bin.

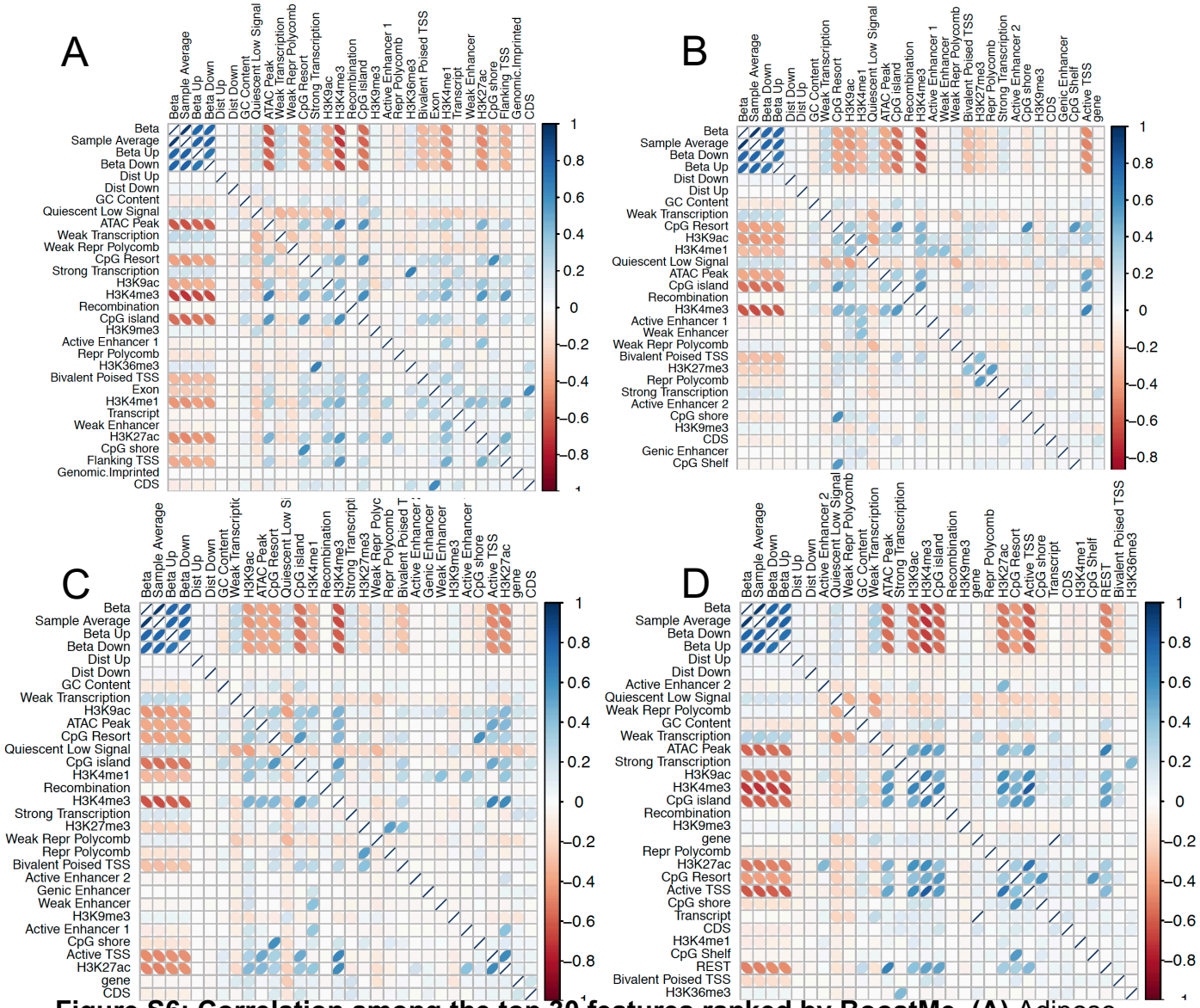

Figure S6: Correlation among the top 30 features ranked by BoostMe. **(A)** Adipose NGT, **(B)** Muscle NGT, **(C)** Muscle T2D, **(D)** Islet.

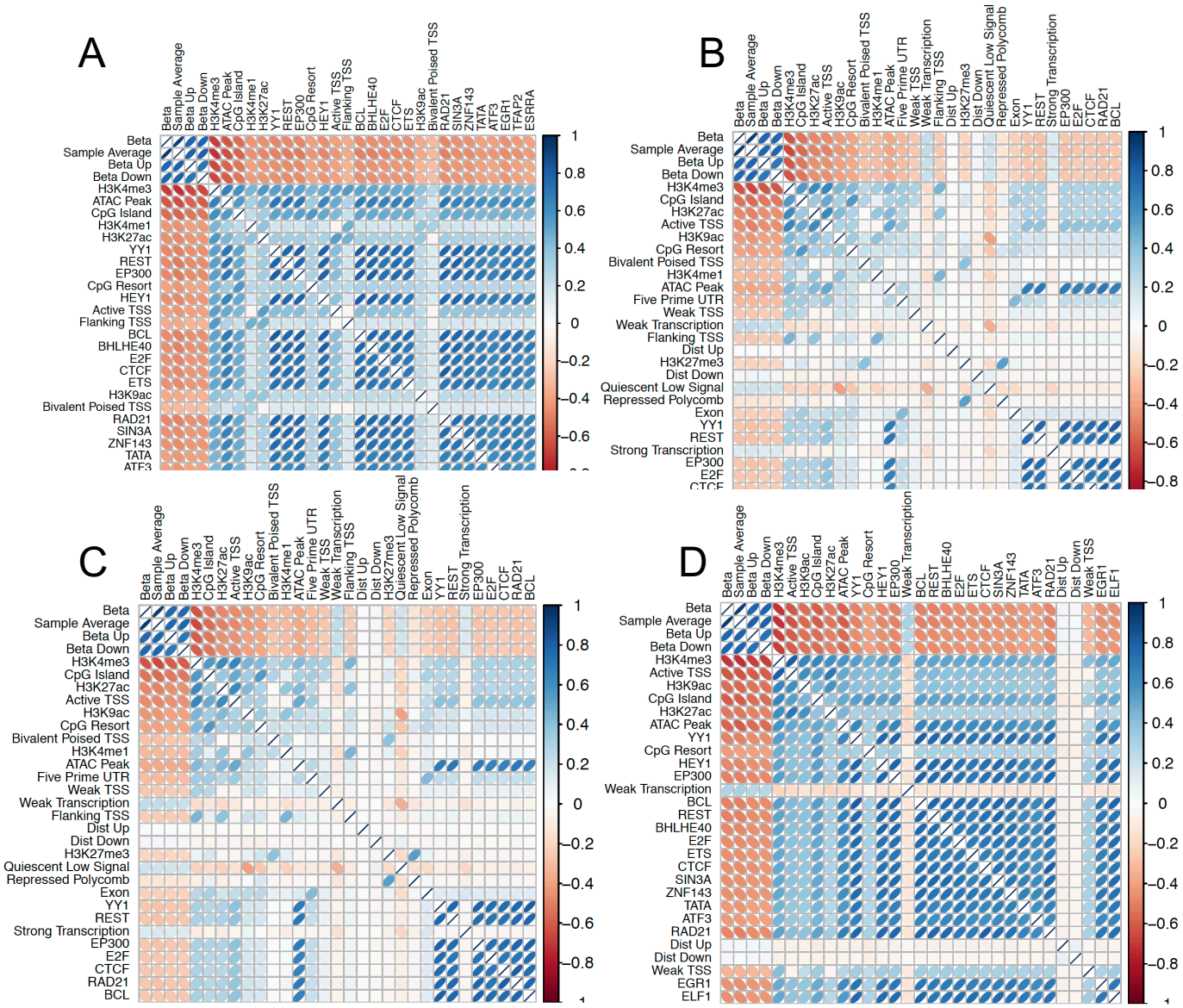

Figure S7: Correlation among the top 30 features ranked by random forests. **(A)** Adipose NGT, **(B)** Muscle NGT, **(C)** Muscle T2D, **(D)** Islet.

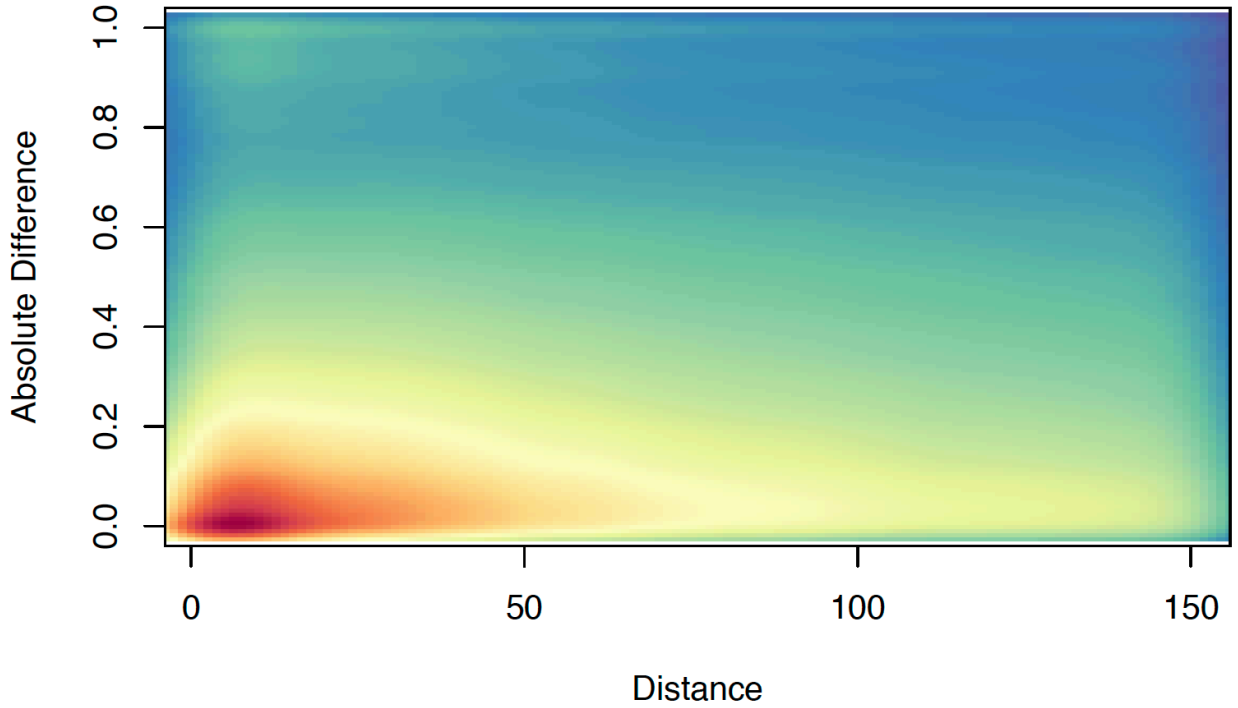

Figure S8: Smooth scatterplot of the absolute difference between CpGs and their neighbors vs. distance to the neighbor. Neighbors for each CpG were taken as the nearest non-missing (sequencing depth ≥ 10x) CpG upstream. Scatterplot was essentially identical across all tissues.

Table S1: Previously reported imputation metrics and those reported in this work.

Method	Train Data/Test Data (if different)	RMSE	AUROC	AUPRC	Accuracy
RF [54]	Illumina-450k sites; blood/WGBS blood	0.23	-	-	0.92
Penalized Functional Regression [61]	Illumina-450k; blood	0.25	-	-	-
DNN (CpGenie) [59]	RRBS; GM12878 cell line	-	0.85	0.69	-
DNN (DeepCpG) [53]	scBS-seq; mouse ES cells	-	0.93	-	-
This work, DeepCpG	WGBS; adipose, muscle, islets	0.18	0.94	0.98	0.91
This work, RF	WGBS; adipose, muscle, islets	0.10	0.99	0.99	0.96

- = not reported. RF - random forests, DNN - deep neural network.

Table S2: All features included in BoostMe and random forests and their source.

Features	Source
Sample average Upstream CpG beta Downstream CpG beta	WGBS
GC content	UCSC Genome Browser
Recombination rate	HapMap (accessed from UCSC)
CpG island (CGI) CpG shore CpG shelf CpG resort	GENCODE
ATAC peaks	ATAC-seq, macs2
H3K27ac H3K27me3 H3K36me3 H3K4me1 H3K4me3 H3K9ac H3K9me3	ChIP-seq, ChromHMM
AFP, AHR, AIRE, ALX1, ALX3, ALX4, AP1, AP3, ARHGEF12, ARID3A, ARID5A, ARID5B, ARNT, ARNTL, ARX, ASCL2, ATF2, ATF3, ATF4, ATF6, ATF7, ATOH1, BACH1, BACH2, BARHL1, BARHL2, BARX1, BARX2, BATF, BBX, BCL, BCL6B, BDP1, BHLHA15, BHLHE22, BHLHE23, BHLHE40, BHLHE41, BPTF, BRCA1, BSX, CACBP, CACD, CBX5, CCDC6, CCNT2, CDC5L, CDX, CDX1, CDX2, CEBPA, CEBPB, CEBPD, CEBPE, CEBPG, CENPB, CHD2, CLOCK, COMP1, CPEB1, CPHX, CREB3, CREB3L1, CREB3L2, CREB5, CRX, CTCF, CTCFL, CUX1, CUX2, DBP, DBX1, DBX2, DLX1, DLX2, DLX3, DLX4, DLX5, DLX6, DMBX1, DMRT1, DMRT2, DMRT3, DMRTA1, DMRTA2, DMRTC2, DOBOX4, DOBOX5, DPRX, DRGX, DUX4, DUXA, E2F, E2F2, E2F7, E2F8, E4F1, EBF1, EGR1, EGR3, EGR4, EHF, ELF1, ELF2, ELF3, ELF4, ELF5, ELK3, ELK4, EMX1, EMX2, EN1, EN2, EOMES, EP300, ERF, ERG, ESR2, ESRRA, ESRRB, ESRRG, ESX1, ETS, ETV1, ETV2, ETV3, ETV4, ETV5, ETV6, ETV7,	ENCODE, FIMO PWM scan

EVX1, EVX2, FEV, FIGLA, FLI1, FOX,
FOXA, FOXB1, FOXC1, FOXC2, FOXD1,
FOXD2, FOXD3, FOXF1, FOXF2,
FOXG1, FOXI1, FOXJ1, FOXJ2, FOXJ3,
FOXK1, FOXL1, FOXM1, FOXN1,
FOXO1, FOXO3, FOXO4, FOXO6,
FOXP1, FOXP3, FOXQ1, GATA, GBX1,
GBX2, GCM, GCM1, GCM2, GFI1,
GFI1B, GLI, GLI2, GLIS1, GLIS2, GLIS3,
GMEB1, GMEB2, GRHL1, GSC, GSC2,
GSX1, GSX2, GTF2A, GTF2I, GZF1,
HAND1, HBP1, HDAC2, HDX,
HERPUD1, HES1, HES5, HES7, HESX1,
HEY1, HEY2, HF1H3B, HIC1, HIC2,
HIF1A, HINFP, HLF, HLTF, HLX,
HMBOX1, HMGA1, HMGN3, HMX1,
HMX2, HMX3, HNF1, HNF1A, HNF1B,
HNF4, HOMEZ, HOXA1, HOXA2,
HOXA3, HOXA4, HOXA5, HOXA6,
HOXA7, HOXA9, HOXA10, HOXA11,
HOXA13, HOXB2, HOXB3, HOXB4,
HOXB5, HOXB6, HOXB7, HOXB8,
HOXB9, HOXB13"
"HOXC4, HOXC5, HOXC6, HOXC8,
HOXC9, HOXC10, HOXC11, HOXC12,
HOXC13, HOXD1, HOXD3, HOXD8,
HOXD9, HOXD10, HOXD11, HOXD12,
HOXD13, HSF, HSF2, HSF4, HSFY2,
ID4, IKZF1, IKZF2, IKZF3, INSM1, IRF,
IRF4, IRF6, IRX2, IRX3, IRX4, IRX5,
IRX6, ISL2, ISX, ITGB2, JDP2, KLF4,
KLF7, KLF12, KLF13, KLF14, KLF16,
LBX2, LHX1, LHX2, LHX3, LHX4, LHX5,
LHX6, LHX8, LHX9, LMO2, LMX1A,
LMX1B, MAF, MAFF, MAFG, MAZ,
MEF2, MEF2B, MEF2D, MEIS1, MEIS2,
MEIS3, MEOX1, MEOX2, MESP1, MGA,
MIXL1, MLX, MLXIPL, MNT, MNX1,
MSC, MSX1, MSX2, MTF1, MXI1, MYB,
MYBL1, MYBL2, MYC, MYCN, MYF,
MYF6, MYOD1, MYOG, MZF1, NANOG,
NEUROD2, NEUROG2, NFAT, NFAT5,
NFATC1, NFATC2, NFE2, NFE2L1,
NFE2L2, NFIA, NFIB, NFIC, NFIL3,
NFIX, NFKB, NFY, NHLH1, NKX1-1,
NKX1-2, NKX2-1, NKX2-2, NKX2-3,

NKX2-4, NKX2-5, NKX2-6, NKX2-8,
NKX3-1, NKX3-2, NKX6-1, NKX6-2,
NKX6-3, NOBOX, NOTO, NR1H, NR1H4,
NR2C2, NR2E1, NR2E3, NR2F2, NR2F6,
NR3C1, NR3C2, NR4A, NR5A1, NR5A2,
NR6A1, NRF1, NRL, OBOX1, OBOX2,
OBOX3, OBOX5, OBOX6, OLIG1,
OLIG2, OLIG3, ONECUT1, ONECUT2,
ONECUT3, OSR1, OSR2, OTP, OTX,
OTX1, OTX2, OVOL2, PATZ1, PAX1,
PAX2, PAX3, PAX4, PAX5, PAX6, PAX7,
PAX8, PAX9, PBX1, PBX3, PDX1,
PHOX2A, PHOX2B, PITX1, PITX2,
PITX3, PKNOX1, PKNOX2, PLAG1,
PLAGL1, POU1F1, POU2F2, POU2F3,
POU3F1, POU3F2, POU3F3, POU3F4,
POU4F1, POU4F2, POU4F3, POU5F1,
POU6F1, POU6F2, PPARA, PRDM1,
PRDM4, PROP1, PROX1, PRRX1,
PRRX2, PTEN, PTF1A, RAD21, RAR,
RARA, RARB, RARG, RAX, RAX2,
RBPJ, REL, REST, RFX2, RFX3, RFX5,
RFX7, RHOXF1, RHOXF2, RORA,
RREB1, RUNX, RUNX1, RUNX2,
RUNX3, RXRA, RXRB, RXRG, SCRT1,
SCRT2, SEF1, SETDB1, SHOX, SHOX2,
SIN3A, SIRT6, SIX5, SMAD, SMAD3,
SMAD4, SMARC, SMC3, SNAI2, SOX,
SOX1, SOX2, SOX3, SOX4, SOX5,
SOX7, SOX8, SOX9, SOX10, SOX11,
SOX12, SOX13, SOX14, SOX15,
SOX17, SOX18, SOX21, SOX30, SP1,
SP2, SP4, SP8, SP100, SPDEF, SPI1,
SPIB, SPIC, SPZ1, SREBP, SRF, SRY,
STAT, T, TAL1, TATA, TBR1, TBX1,
TBX2, TBX4, TBX5, TBX15, TBX19,
TBX20, TBX21, TCF3, TCF4, TCF7,
TCF7L1, TCF7L2, TCF12, TCF21,
TEAD1, TEAD2, TEAD3, TEAD4, TEF,
TFAP2, TFAP2B, TFAP2E, TFAP4,
TFCP2, TFCP2L1, TFE, TFE3, TFEB,
TFEC, TGIF1, TGIF2, TGIF2LX, THAP1,
THRA, THRB, TLX2, TOPORS, TP53,
TP63, TP73, TRIM28, UNCX, VAX1,
VAX2, VDR, VENTX, VSX1, VSX2, WT1,
XBP1, YY1, YY2, ZBED1, ZBTB3,

ZBTB6, ZBTB7A, ZBTB7C, ZBTB12, ZBTB14, ZBTB16, ZBTB18, ZBTB33, ZBTB49, ZEB1, ZFX, ZIC1, ZIC2, ZIC3, ZIC4, ZKSCAN3, ZNF8, ZNF35, ZNF75A, ZNF143, ZNF219, ZNF232, ZNF263, ZNF281, ZNF282, ZNF350, ZNF354C, ZNF384, ZNF410, ZNF423, ZNF524, ZNF589, ZNF628, ZNF652, ZNF691, ZNF713, ZNF740, ZNF784, ZSCAN4, ZSCAN16, ZSCAN26

Table S3: RMSE performance of BoostMe and random forests improves when training on continuous values.

Algorithm	RMSE (bin.)	RMSE (cont.)	AUROC (bin.)	AUROC (cont.)	AUPRC (bin.)	AUPRC (cont.)	Accuracy (bin.)	Accuracy (cont.)
BoostMe	0.16	0.10	0.99	0.99	0.99	0.99	0.96	0.96
Random Forests	0.16	0.10	0.99	0.99	0.99	0.99	0.96	0.96
DeepCpG	0.18	0.18	0.94	0.94	0.98	0.98	0.91	0.91

RMSE - root-mean-squared error, AUROC - area under the receiver operating characteristic curve, AUPRC - area under the precision-recall curve, bin. - performance when training on binary methylation values, cont. - performance when training on continuous methylation values.

Table S4: Top 100 transcription factors ranked in descending order as reported by BoostMe, trained only using TFBS features.

Adipose NGT	Adipose T2D	Muscle NGT	Muscle T2D	Islet
EP300	EP300	REST	REST	EP300
REST	REST	ETS	ETS	REST
ETS	ETS	RAD21	RAD21	ETS
PAX5	PAX5	PAX5	PAX5	HEY1
ATF3	ATF3	E2F	ATF3	E2F
E2F	E2F	ATF3	E2F	PAX5
RXRA	RXRA	RXRA	EGR1	HINFP
RAD21	RAD21	RFX5	RFX5	ATF3
HINFP	RFX5	STAT	STAT	RFX5
RFX5	HINFP	EGR1	HNF4	RAD21
HEY1	HEY1	IRF	NR3C1	RXRA
EGR1	EGR1	GATA	RXRA	STAT
STAT	STAT	EP300	GATA	EGR1
NR3C1	NR3C1	HNF4	IRF	BDP1
HSF	IRF	NR3C1	HSF	HSF
IRF	HSF	HSF	EP300	ZBTB33
BDP1	BDP1	BDP1	BDP1	MYBL2
MYBL1	TFAP2	HINFP	MEF2	BHLHE40
GATA	GATA	MEF2	MYBL1	NR3C1
TFAP2	HNF4	MAF	YY1	GATA
HNF4	IRF4	YY1	HINFP	PAX4
IRF4	MYB	MYBL1	MAF	IRF
PAX4	MYBL1	SMAD3	SMAD3	YY1
YY1	PAX4	MYC	CUX1	IRF4
MYB	MEF2	TFAP2	MYC	MYBL1
GFI1	HSFY2	CUX1	CEBPB	MYB
HSFY2	YY1	NFIC	PAX4	GFI1
SPDEF	GFI1	PAX4	GFI1	CUX1
MEF2	BHLHE40	ESRRA	ZSCAN4	ZNF143
CUX1	SPDEF	ZSCAN4	BCL	CACBP
SMAD3	SMAD3	GFI1	FOXJ3	TFAP2
BHLHE40	KLF4	CTCF	HSFY2	CTCF
CTCF	NFATC1	HSFY2	TFAP2	HOMEZ
NFIC	ZBTB33	CEBPB	NFIC	ZBTB14
RUNX3	MYBL2	BCL	XBP1	MEF2
ZBTB33	XBP1	GMEB2	CTCF	KLF4
NFATC1	CUX1	MYBL2	ESRRA	SMAD3
MAF	MAF	FOXJ3	IRF4	NFATC1
GMEB2	GMEB2	MYB	SPDEF	HSFY2
KLF4	CTCF	SPDEF	MYB	FOXJ3
BARHL1	NFY	TBX1	HEY1	HIC2

FOXN1	FOXN1	XBP1	SP1	TBX5
NFY	ZNF524	IRF4	GMEB2	NFIC
HOMEZ	RUNX3	TFCP2	TBX1	GMEB2
XBP1	TATA	RUNX2	TFCP2	SMC3
VDR	NFIC	NFY	HOMEZ	RUNX2
PAX3	PAX3	NFATC1	FOXN1	SPDEF
PTF1A	BCL6B	HOMEZ	HMX3	HNF4
ZNF524	AHR	ZNF713	BCL6B	BARHL2
GRHL1	VDR	PAX3	RUNX2	TCF12
ZNF713	HOMEZ	FOXN1	NFATC1	GLI
MYBL2	CEBPB	FOXK1	NHLH1	FOXN1
FOXC1	PTF1A	TBX20	TATA	NFY
IKZF2	TBX5	SOX9	NFY	TBX4
TATA	GRHL1	TATA	SOX10	FOXC1
NFKB	RUNX2	RHOXF1	TEAD1	TATA
HMGN3	ZNF713	THAP1	MYBL2	BARHL1
ESRRA	FOXJ3	AHR	VDR	PTF1A
CEBPB	ZNF143	ZBTB33	RHOXF1	SP1
HOXA13	IKZF2	SP1	TP53	HDAC2
TBX5	ZBTB14	FOXC1	ZNF282	IKZF2
TFCP2	HMGN3	BARHL1	NFKB	ZNF713
ZBTB14	HOXA13	MNT	ZNF713	CENPB
BCL6B	MYC	OTX1	ZBTB33	HMGN3
AHR	BCL	HEY1	BARHL2	GRHL1
MYC	TFCP2	NFKB	AHR	NFKB
ZNF143	CACBP	BCL6B	TBX20	PAX3
CACBP	BARHL2	TEAD1	ARHGEF12	HIC1
BCL	ESRRA	HOXA5	SIX5	ZSCAN4
HDAC2	FOXC1	ZNF282	BPTF	CHD2
TBX1	NFKB	BARHL2	IKZF2	XBP1
ZNF282	HDAC2	AIRE	PAX3	RUNX3
CENPB	ZNF282	TP53	THAP1	NANOG
ARHGEF12	NFAT	PTF1A	MEIS1	POU2F2
SP1	SOX4	NHLH1	ZBTB6	INSM1
RUNX2	MNT	BPTF	PRDM1	SP8
TBX4	HIC2	VDR	OTX1	NFAT
NFAT	ARHGEF12	TFAP4	GRHL1	MYC
HIC2	TP53	SIX5	PTF1A	ZNF524
ZSCAN4	PRDM1	RARA	FOXA	AHR
FOXK1	NHLH1	RBPJ	RORA	BCL6B
BARHL2	RHOXF1	ZNF524	CHD2	SREBP
HOXC12	ZSCAN4	ARHGEF12	FOXC1	ZNF384
ZBTB7A	CENPB	IKZF2	AIRE	VDR
ZNF384	ZNF384	HDAC2	HOXA5	TCF3
PRDM1	ZBTB6	NFAT	NFAT	TCF4

PROX1	BARHL1	ZNF384	RARG	ARHGEF12
ZBTB6	HIF1A	FOXG1	HIC1	SPI1
SP2	ESRRG	SOX8	ZNF524	PROX1
MNT	GLI	GRHL1	HOXA13	MAF
SP8	HOXC12	HIC1	ZBTB7A	ZNF282
NHLH1	BPTF	EBF1	TFAP4	NRF1
THAP1	ZKSCAN3	HOXA13	FOXG1	ZBTB7A
BPTF	SP2	CCDC6	CRX	CCDC6
ELF1	TBX4	NKX2-8	RUNX3	TFCP2
SIX5	HSF2	ESRRG	PROX1	BCL
HIF1A	PROX1	HMGN3	SOX9	RFX7
FOXJ3	TBX1	SOX10	HOXA11	TP53
SOX4	ELF1	ZBTB12	PAX6	ZBTB3
TP53	ZBTB7A	GZF1	BHLHE40	HNF1

Table S5: Top 100 transcription factors ranked in descending order as reported by random forests trained only using TFBS features.

Adipose NGT	Adipose T2D	Muscle NGT	Muscle T2D	Islet
REST	REST	REST	REST	EP300
EP300	EP300	EP300	EP300	REST
RAD21	RAD21	RAD21	RAD21	BHLHE40
BHLHE40	BHLHE40	E2F	E2F	E2F
E2F	ETS	ETS	ETS	YY1
ETS	E2F	YY1	YY1	RAD21
YY1	YY1	CTCF	CTCF	ETS
CTCF	CTCF	BHLHE40	EGR1	BCL
ZNF143	ZNF143	EGR1	BHLHE40	HEY1
HEY1	HEY1	ZNF143	ZNF143	ZNF143
TFAP2	TFAP2	TFAP2	TFAP2	CTCF
BCL	BCL	BCL	BCL	TFAP2
EGR1	EGR1	ATF3	ATF3	EGR1
ATF3	ATF3	PAX5	PAX5	ATF3
SIN3A	SIN3A	HEY1	HEY1	SIN3A
PAX5	PAX5	SIN3A	SIN3A	ELF1
ELF1	ELF1	MYC	MYC	PAX5
MYC	BDP1	TATA	TATA	MYC
BDP1	MYC	ELF1	NR3C1	HINFP
TATA	TATA	NR3C1	ELF1	BDP1
NR3C1	NR3C1	BDP1	BDP1	TATA
HINFP	HINFP	ESRRA	ESRRA	NR3C1
ESRRA	ESRRA	SP1	SP1	ESRRA
SP1	SP1	RXRA	RXRA	NRF1
RXRA	RXRA	HINFP	HINFP	POU2F2
POU2F2	POU2F2	IRF	IRF	RXRA
NRF1	NRF1	PAX4	AP1	SP1
AP1	AP1	STAT	PAX4	PAX4
PAX4	PAX4	AP1	STAT	CHD2
HIC1	HIC1	NRF1	HNF4	SRF
IRF	IRF	HIC1	NRF1	AP1
NFE2	STAT	HNF4	HIC1	HIC1
STAT	NFE2	POU2F2	POU2F2	NFE2
SETDB1	CHD2	SETDB1	SETDB1	SETDB1
CHD2	SETDB1	NFE2	NFE2	IRF
HNF4	HNF4	CHD2	CHD2	STAT
SPI1	SPI1	SPI1	SPI1	HNF4
SRF	SRF	SRF	SRF	SPI1
EBF1	EBF1	EBF1	EBF1	ZBTB33
ZBTB33	ZBTB33	ZBTB33	ZBTB33	EBF1
ZBTB14	ZBTB14	TCF12	ZBTB14	ZBTB14

TCF12	TCF12	ZBTB14	TCF12	SMC3
SMC3	SMC3	SMC3	SMC3	TCF12
ZIC3	ZIC3	RFX5	RFX5	ZIC3
HDAC2	HDAC2	NFIC	HDAC2	HDAC2
NFKB	NFKB	HDAC2	NFIC	NFKB
NHLH1	NHLH1	GATA	GATA	NHLH1
ZBTB7A	ZBTB7A	NFKB	NHLH1	ZBTB7A
VDR	VDR	NHLH1	NFKB	VDR
NFIC	NFIC	MAF	MAF	NFIC
RFX5	RFX5	ZBTB7A	ZBTB7A	RFX5
SP4	GATA	ZIC3	VDR	GATA
GATA	SP4	VDR	ZIC3	SP4
ZIC1	ZIC1	PBX3	PBX3	MAF
PBX3	PBX3	SP4	SP4	ZIC1
RUNX3	RUNX3	MYB	MYB	PBX3
MAF	MAF	RUNX3	RUNX3	MTF1
EGR3	EGR3	TFCP2	TFCP2	RUNX3
TFCP2	TFCP2	ZIC1	HSF	EGR3
WT1	WT1	AHR	EGR3	TFCP2
MYB	MYB	HSF	ZIC1	MYB
MTF1	MTF1	EGR3	RUNX2	WT1
RUNX2	RUNX2	RUNX2	AHR	TCF3
AHR	AHR	SP2	TCF3	ZFX
HF1H3B	TCF3	TCF3	MTF1	HF1H3B
TCF3	HSF	MTF1	SPDEF	TP53
HSF	HF1H3B	SPDEF	TP53	TFAP4
EGR4	EGR4	TFAP4	SP2	AHR
SP2	SP2	WT1	WT1	RUNX2
ZFX	TP53	TP53	TFAP4	SP2
SPDEF	ZFX	PAX2	PAX2	TFAP2B
TFAP2B	TFAP2B	EGR4	EGR4	SPDEF
TP53	SPDEF	MEF2	MEF2	SPZ1
TFAP4	TFAP4	CUX1	HF1H3B	EGR4
ZIC4	SPZ1	SMAD3	SMAD3	NR1H
SPZ1	ZIC4	HF1H3B	CEBPB	PLAG1
NR1H	ZNF219	SREBP	CUX1	ZIC4
ZNF219	NR1H	TFAP2B	NR1H	ZNF219
PAX2	PAX2	NR1H	MYBL1	HSF
RREB1	RREB1	ZFX	RARA	RREB1
ZNF281	ZNF281	RARG	ZFX	PAX2
GLI2	PLAG1	RARA	TFAP2B	PPARA
PLAG1	GLI2	RREB1	SREBP	ZBTB3
SREBP	PPARA	MYBL1	RARG	GLI2
PPARA	SREBP	SPZ1	THAP1	NANOG
THAP1	NANOG	CEBPB	ZNF281	SREBP

PLAGL1	TEAD2	THAP1	SPZ1	PLAGL1
TEAD2	THAP1	ZIC4	ZIC4	CACD
KLF4	PLAGL1	PAX6	RREB1	ZNF281
SMAD3	RARA	XBP1	XBP1	MYOD1
ZBTB3	KLF4	GLI2	PPARA	KLF12
CEBPB	ZNF524	ZNF281	PAX6	THAP1
NANOG	ZBTB3	PAX3	GMEB2	TEAD2
RARA	CEBPB	MYBL2	GLI2	RARA
ZNF524	CACD	PPARA	ZNF219	NR2C2
CACD	PATZ1	GFI1	RFX3	SMAD3
PATZ1	NR2C2	TBX20	TBX20	GCM1
RARG	SMAD3	MYOD1	MYOD1	PATZ1
KLF12	KLF12	KLF4	NANOG	PAX6
KLF7	KLF7	GMEB2	GFI1	ELF2