Supplementary Text, Figures and Tables for DIRECTION: A machine learning framework for predicting and characterizing DNA methylation and hydroxymethylation in mammalian genomes

S1. Binary labeling of cytosines in BS-seq and TAB-seq data

For methylation prediction, we chose a CCR threshold of 0.5 (equidistant from the extremal CCRs of 0 and 1) for labeling methylation status for individual cytosines as high or low (≥0.5 or <0.5 respectively). For analyzing 5-hmC levels, a naive analysis yields a distribution with maximal frequency at the CCR of 0, and no other observable secondary modes in the distribution (Supp Fig 1B). However, it is well characterized that while most CpG sites have a CCR of 0, statistically significant hydroxymethylated CpG sites have a CCR frequency distribution with a mode of 0.18 (Supp Fig 1C) (Yu,M., et al. 2012). Thus, in order to label individual cytosines into significantly hydroxymethylated versus non-hydroxymethylated/weakly hydroxymethylated classes, we choose a threshold of 0.09, equidistant from 0 and 0.18.

S2. Feasibility of 5-hmC status prediction

Strong preference of 5-hmC for open chromatin regions, as well as its positive correlation with gene expression and bias towards exon inclusion were previously documented in literature (Marina,R.J., et al. 2015, Mellen,M., et al. 2012), suggesting a functional role and consistency of 5-hmC modifications across biological replicates. 5-hmC modifications have also been shown to be temporally stable (Bachman,M., et al. 2014) further suggesting a strong signal to noise ratio in hydroxymethylation assays. At the outset, we performed pairwise comparison of hydroxymethylation levels across biological replicates in NPC, using binary discretization of hydroxymethylation levels. TAB-seq CCR correlation across biological replicates is less faithful than BS-seq by exhibiting some stochasticity in the signal. However, we obtain a concordance rate (fraction of cytosines where 5-hmC status between replicates agree) of 82% (in CpG sites with coverage >60) for 5-hmC status between biological replicates in NPC (Supp Fig 1D). For practical purposes, this may be considered as an approximate upper bound of possible predictive accuracy when evaluating 5-hmC status predictions.

Further we looked at consistency of our BS-seq and TAB-seq datasets in NPC. MLML (Qu,J., et al. 2013) is a method that uses read counts from data obtained by TAB-seq (or oxBS-seq), and BS-seq; to estimate CCRs for the 5-mC and 5-hmC modifications jointly. It identifies indices exhibiting “overshoot” where the sum of estimated CCRs for 5-mC and 5-hmC sum to greater than 1. Upon running MLML on our BS-seq and TAB-seq datasets in NPC we obtained the maximum likelihood distribution of 5-mC levels (Supp Fig 1E) that strongly resembled the one of BS-seq levels (Supp Fig 1A). Additionally, out of the 52,531,101 CpG sites being analyzed (sites without coverage in either of the experiments are discarded) the number of overshoot indices was only 3,186 or 0.006% in NPC. Most of the overshoot indices contained very low coverage (2,654 CpG sites) in both BS-seq and TAB-seq experiments and were systematically discarded prior to training our model.

S3. Design decisions for training and testing the SVM and Random Forest

SVM model decisions: The parameters used to train the SVM are as follows: kkt violation fraction =0.05, maximum number of sampled training sets used for training in order to achieve SVM convergence =3, maximum number of iterations in each training for SVM convergence =10^7. The average number of support vectors per 8000 training examples within different BS-seq optimal feature sets varied between 1200-1300, suggesting an upper bound of the experimental error rate range of 0.15-0.1625.

RF model decisions: When training the RF, we randomly sample one third of all available features in the training set, and perform sampling of training data-points with replacement. Splitting on input features is performed in a way that minimizes Gini Impurity score. Depending on the prediction paradigm we grow between 50 and 150 decision trees in the forest (for example CGI methylation status predictions can be successfully performed using 50 decision trees: when classification error reaches its minimum). Additional information about different modes implemented in our toolkit can be found in Supp Table T2.

Training and testing set sizes: In order to evaluate the performance of our SVM and RF based predictive model we perform 5 fold cross-validation using balanced sets having 10,000 data points. The balanced sets are comprised of 5,000 positive and 5,000 negative examples, where 4,000 of each class are used for training and the remaining 1,000 of each class for testing. We discovered that the aforementioned design decisions govern the best trade-off between stably and accurately estimating prediction metrics versus computational time (Supp Figs 2A, 2B, Supp Table T12). We thus chose k=5 for k-fold cross-validation on 10,000 sampled training examples (5,000 of each class) to balance out the trade-off between the testing set size. Namely, if k is too high the testing set size will be too small, and conversely if k is too small the training set size is too small and the number of experiments may not be enough to estimate the prediction performance. If k is too large, it is worth noting that using more than 20,000 data points to train the SVM may cause the MATLAB built-in function svmtrain to be very slow, which may effectively result in non-convergence from a practical point of view.

Increasing label fidelity for training and testing samples: We identified the sequencing depth required for cytosines used for training (inclusion in the training set) and evaluating (inclusion in the testing set) our models based on the minimum sequencing depth that would always distinguish unmethylated (or non-hydroxymethylated) cytosine from marginally methylated or hydroxymethylated (CCR of 0.5 or 0.09 for BS-seq and TAB-seq respectively) given representative sampling. Due to sampling variance at low sample sizes causing small sample sizes to often not be representative, we performed a non-parametric categorical test (Fisher’s Exact Test) between categorical distributions where for one sample the CCR is zero, versus another sample where CCR of the marginally methylated or hydroxymethylated sample is faithfully represented in the sample. We perform this over a range of sequencing depths fixed for both samples to identify when Fisher’s Exact Test is able to identity a statistically significant difference between the two samples. This was performed to ensure label fidelity of training and testing samples. For BS-seq datasets, we need to minimally differentiate between completely unmethylated cytosines with a CCR of 0 with respect to marginally methylated cytosines with a CCR of 0.5. Given representative sampling, the minimum sequencing depth at a cytosine required to differentiate between the cases is two. However, we find that for the Fisher’s Exact Test, we get a statistically significant p-value ($p \leq 0.05$) when sequencing depth for both samples is 10. In practice, for the SVM and RF models, both balanced set predictions and whole genome predictions were performed with cytosines where coverage ≥ 20. We find that out of 56,434,896 annotated CpG cytosines, 50,379,832 have coverage ≥ 20 in H1, and 49,134,499 have coverage ≥ 20 in NPC, suggesting that even in datasets with high sequencing depth, between 11% and 13% of cytosines do not have satisfactory coverage depth and can be imputed using DIRECTION. For the Reference Methylome predictor variable based predictions, and SVM model is compared with the Reference Methylome predictor, since sequencing depth ≥ 20 across all reference methylomes causes a large drop in the number of cytosines eligible for training and testing, a more modest sequencing depth constraint of ≥ 5 was used. Similarly, when Nearest Neighbor evaluations were performed, the more modest sequencing depth constraint of ≥ 5 was used in order to capture more cytosines in the evaluation process.

Additionally, Consensus Reference Methylome and Nearest Neighbor were introduced as input predictor variables into our toolkit, and cytosines with sequencing depth ≥ 5 were chosen for this purpose.

For TAB-seq datasets, we need to minimally differentiate between completely non-hydroxymethylated cytosines with a CCR of 0 with respect to marginally hydroxymethylated cytosines with a CCR of 0.09. Given representative sampling, the minimum sequencing depth at a cytosine required to differentiate between these cases is 20. We find that for the Fisher's Exact Test, we get a statistically significant p-value (p< or ~0.05) when sequencing depth for both samples is 60. In practice, for the SVM model, both balanced set predictions and whole genome predictions were performed with cytosines where coverage ≥ 60. See Supp Table T13A for p-values.

S4. Relevant metrics for evaluating classification of cytosine modifications

The metrics commonly used to assess the performance of a supervised learning algorithm belong to one of the following three categories: threshold metrics, rank metrics, or probability metrics (Caruana,R. and Niculescu-Mizil,A. 2006). Since we perform classification using non-likelihood based approaches (SVM and RF), we use appropriate metrics in the "threshold-based" metrics category. The decision of which one to chose mostly depends on the nature of the problem that needs to be addressed. For prediction of skewed classes, special care needs to be taken such that the metric does not get inflated by simply predicting one class more often than the other. Concretely, we perform both methylation and 5-hmC predictions using balanced sets (avoiding skewed classes) and report the performance using Precision, Recall, F-Score (harmonic mean of Precision and Recall), and AUC while whole-genome prediction performance (where the frequency of the two classes are skewed for both methylation and 5-hmC status prediction) is evaluated using True Positive Rate (Sensitivity or Recall), True Negative Rate (Specificity) and Accuracy. Evaluation metrics used in our analyses are formulated as follows:

1. $Precision = \frac{TP}{TP+FP}$
2. $Recall\ or\ Sensitivity\ or\ True\ Positive\ Rate = \frac{TP}{TP+FN}$
3. $Specificity\ or\ True\ Negative\ Rate = \frac{TN}{TN+FP}$
4. $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
5. $F - score\ or\ F1 - score = \frac{2*Sensitivity*Specificity}{Sensitivity+Specificty}$
6. AUC: Which is created by plotting the False Positive Rate (1-True Negative Rate) on the x-axis and the True Positive Rate on the y-axis, and the resulting area under the curve is calculated.

where TP=number of true positives, TN=number of true negatives, FP=number of false positives, and FN=number of false negatives.

AUC calculations were only performed for RF models due to the ease of ranking predictions using the MATLAB native random forest function.

S5. Feature preprocessing

We use a variety of genomic and epigenomic traits as input to train our classifier (Supp Table T3). Features we do not model include gene annotation because histone modification data implicitly contain this information and enable us to discern between active, poised, and repressed cis-regulatory (Ernst,J. and Kellis,M. 2012) and transcribed regions. Such annotation-based features may be incorporated when histone modification datasets are not available. Additionally, we do not model spatial contiguity explicitly into our predictive model. Since DNA methylation response variable (thresholded BS-seq CCRs) and various input features (e.g. histone modifications) are very well correlated spatially, our predictions are able to identify stretches of similar methylation without a need for explicit spatial auto-correlative models like Hidden Markov Model (HMM) or explicit spatial input features. TAB-seq CCRs are not spatially auto-correlated as well as BS-seq CCRs, but 5-hmC enriched regions and large stretches of 5-hmC depletion can be identified. Finally, features such as discriminative k-mers and motifs or ChIP-seq datasets of TF binding that can predict the methylation status were not used since the expression of such TFs are likely to be cell-type specific and accordingly not suitable for transfer learning purposes in the context of whole methylome reconstruction. Only the near ubiquitously expressed CTCF and p300 TF ChIP-seq data were used in the Initial Feature Set for predicting H1 methylation status, and these features were not used for NPC methylation status prediction, transfer learning for methylation status prediction, or 5-hmC status prediction.

All genomic features (tracks) such as Alu repeats, CGI as well as the genomic positions of CpG sites in the human genome (hg19 assembly) were obtained from the UCSC genome browser (Speir,M.L., et al. 2016), or calculated based on the downloaded sequence and annotation. Histone mark ChIP-seq, DNase-seq and Transcription Factor binding ChIP-seq data (CTCF, p300) were obtained from the Roadmap Epigenome consortium (http://egg2.wustl.edu/roadmap/web_portal/processed_data.html) (Kundaje,A., et al. 2015) under the NCBI GEO GSE16256 accession. Genome-wide signal coverage tracks (negative log10 transform of the p-value) based on the uniformly processed Roadmap Epigenome Consortium datasets were used for ChIP-seq and DNase-seq features (Kundaje,A., et al. 2015). BS-seq and TAB-seq preprocessing is detailed in Supp Text S11. For 5-hmC status prediction, BS-seq CCR (and not the predicted methylated status) was used as an input feature. All the raw features were matched against the list of available CpG sites using the IntersectBed tool from the Bedtools toolkit (Quinlan,A.R. and Hall,I.M. 2010). After initial processing all the features were stored into a single matrix. The features were normalized to zero mean and variance=1 (also called standardization), before training the model (Bishop,C. 2007).

An additional feature was created for methylation status imputation based on the methylation status of the CpG cytosine nearest to the cytosine in question (nearest neighbor feature). However, a similar feature was not used for 5-hmC status imputation since 5-hmC modifications do not occur in long stretches even though they can be somewhat locally enriched. We find that the 5-hmC status of the nearest CpG cytosine has poor predictive ability when performing imputation (Supp Table T14).

S6. Recursive feature elimination and beam search pseudocode

Typically, machine learning models with more parameters tend to fit the response variable better, occasionally resulting in overfitting (Bishop,C. 2007). This leads to a trade-off between predictive power and feature sparsity. Some previous approaches to perform optimal feature selection include dimensionality reduction (Zheng,H., et al. 2013), and removal of individual features from the full feature set to create the Gini index (Yan,H., et al. 2015, Zhang,W., et al. 2015) which will rank the features according to their contributions to the prediction metric. In order to gain additional insight about features and their additive effects we implemented a modified version of the recursive feature elimination algorithm that provides information about the discriminative nature of individual features and features subsets. Recursive feature elimination is a well established strategy that was successfully used to determine the most predictive features and feature sets for methylation prediction (Das,R., et al. 2006). However, performing a top-down exhaustive search given a high number of input features (N) can be extremely time consuming and computationally demanding since the number of explored feature sets may reach 2^N-1, leading us to consider heuristic approaches in determining the OFS .

Our recursive feature elimination algorithm is implemented using beam search :

```
BeamSearch(beam width b, Initial Feature Set I, cross-validation fold k, maximum number of iterations m, number of top feature sets returned t )

fit and test models on Initial Feature Set I on training data using k-fold cross-validation ;
calculate evaluation metric score $e_I$ by comparing predictions and known labels ;
update list of top optimal feature sets L based on evaluation ;
initialize priority queue Q with feature set I using priority $e_I$ ;
initialize number of evaluations n ;
while ( Q is not empty AND n < maximum number of iterations m )
{
    S = feature set dequeued from head of Q having highest priority ;
    for each feature f in S
    {
        initialize candidate feature set list C = { } ;
        S' = S – { f } ;
        if  (S' has not been evaluated previously)
        {
            fit and test models on feature set S' on training data using k-fold cross-validation;
            if (model converges on training)
            {
                calculate evaluation metric score $e_{S'}$ by comparing predictions to known
                labels ;
                update list of evaluated feature sets L with (S', $e_{S'}$) ;
                add (S', $e_{S'}$) to candidate feature set list C ;
                increment n ;
            }
        }
        sort C based on evaluated metric scores ;
        choose the top b (beam width) feature sets with highest evaluation metric scores from C
        and enqueue into priority queue Q using evaluated metric score as priority ;
    }
}
sort L based on evaluation metric score and return top t feature sets
```

An integrated picture of Initial Feature Set selection and the identification of the OFS using beam search is shown in Supp Fig 3D.

S7. Initial elimination of redundant features before beam search

We identified and eliminated redundant features based on feature clustering and reduced the size of the full feature set, and ultimately created the "Initial Feature Sets" (IFS) (Supp Table T15). We identify clusters of highly correlated features and keep only one representative feature for each cluster and eliminate the others. The total set of predictor variables include several features that were engineered at multiple genomic resolutions (in bins of 50bp, 100bp, 200bp, 400bp, and 800bp) to predict DNA methylation and hydroxymethylation in genomic regions of corresponding size, and these naturally cluster in redundant groups. Since DIRECTION is trained to classify methylation and 5-hmC status at single nucleotide resolution, engineered features at the smallest resolution (50bp) were kept for the IFS, and the lower resolution features were discarded. These decisions resulted in saving a reasonable amount of computational time, and significantly reduced the possibility of overfitting our model (Supp Fig 3D).

S8. Characterization of feature subset contributions to predictive ability of the OFS

While creating the IFSs eliminated highly correlated features, OFSs identified by the beam search algorithm can still contain somewhat correlated, partially redundant features. For performance issues, we want to have some degree of redundancy in the OFS to make the prediction robust, but on the other hand we want to also assess the contribution to the predictive ability by subsets of features in the OFS. We thus performed the following assessment. We performed a standardization (Z-transformation (Bishop,C. 2007)) across all features and hierarchically clustered them to identify similarity across features. Based on the feature clustering in the OFS, we left out individual features and feature subsets according to the nodes of the dendrogram, and retrained our classifier. The difference in performance metrics with respect to the OFS provides a clear indication of both feature redundancy and contributions of subsets of features to the OFS prediction metric.

While max-margin models do not explicitly posses a likelihood-based inferential framework to directly apply information theoretic approaches to sparse model selection like the Aikake Information Criterion (Bishop,C. 2007), our approach provides an intuitive platform to identify smaller subsets of the OFS having comparable predictive power, and also identifies subsets of features that have major contributions to the precision and recall (Fig 4C and Supp Figs 2C, 2D).

The notion behind identifying a "minimal" feature set was based on the notion of several correlated input features potentially being part of the OFS, each only contributing a limited amount of predictive power to the overall OFS. By clustering the individual features in the OFS and eliminating them one at a time, we identified the effect each (or a subset) possesses on the predictive power, in a manner agnostic to the classification algorithm. The tradeoff between obtaining a smaller feature set versus improving classification performance metrics can thus be clearly identified, allowing the user to decide on a choice of the input feature set for related experiments.

S9. Predictive power of neighboring CpG sites

Since the predictive power of neighboring CpG sites drops with distance, we wanted to determine what fraction of CpG sites with low coverage (< 5) has high coverage (≥ 5) neighboring CpG sites within 500bp, making them a good candidate for imputation. To answer this question, we computed the Cumulative Distributive Function (CDF) of the fraction of low coverage sites with respect to distance to the nearest high coverage neighboring site (Fig 3D), in high coverage (NPC) and low coverage (Fetal Small Intestine) Roadmap Epigenome consortium datasets. Even in a low coverage methylome such as Fetal Small Intestine (Supp Fig 3B), more than 60% of low coverage CpG sites had a corresponding high coverage neighbor within 500bp, suggesting high probability of them being correctly imputed (Fig 3C).

S10. Defining a consensus reference methylome

The high predictive ability of DNA methylation predictors by using only sequence derived features (in multiple datasets) suggests that a portion of DNA methylation status in CpG sites is governed by the underlying sequence, and should be unchanged across tissue and cell types and across conditions. By utilizing the concept of similar methylation status across different tissues, we identified the regions of invariant methylation and implemented it into our prediction framework. We obtained 25 publicly available cell line and tissue WGBS datasets (Supp Table T9B) from the Roadmap Epigenome consortium excluding H1 and H1-derived cells, estimated its methylation status by thresholding the CCR at

0.5 for each cytosine, and compared the respective binary methylation statuses (high and low methylation) across all the CpG sites (with coverage ≥ 5 for all 25 reference methylomes not based on the H1 lineage).

S11. DNA methylation and hydroxymethylation data sourcing and performance of DIRECTION for different BS-seq CCR values

DNA methylation and hydroxymethylation data sourcing:

Traditionally, Next Generation Sequencing (NGS) techniques require either a *de novo* assembly of the sequenced reads or mapping the reads to a known reference genome. Various tools like BSMAP, RMAP, BS-Seeker, and BISMARK (Guo,W., et al. 2013, Krueger,F. and Andrews,S.R. 2011, Smith,A.D., et al. 2009, Xi,Y. and Li,W. 2009) perform end to end mapping analysis or build "wrappers" around state-of-the-art generic NGS read mapping tools like Bowtie (Langmead,B. and Salzberg,S.L. 2012) for this purpose. Typically, most such methylation calling strategies use a filtering scheme to count bases with high quality sequencing and alignment scores, followed by a simple binomial probability test (Yu,M., et al. 2012). We have devised the pipeline for end to end mapping and variant calling of raw BS-seq and TAB-seq reads using the BISMARK BS-seq read mapper (Krueger,F. and Andrews,S.R. 2011). Scripts that were used to calculate the reads sequencing depth and hydroxymethylation levels were coverage2cytosine and bismark methylation extractor. The final output to the .bed format was performed by the bismark2bedGraph. This was performed to generate H1 and NPC TAB-seq CCRs. H1, NPC, MSC, and IMR90 BS-seq CCRs were obtained from the uniformly processed datasets of the NIH Roadmap Epigenome Consortium (http://egg2.wustl.edu/roadmap/web_portal/processed_data.html) processed from GEO series GSE16256 datasets by the Consortium as fractional methylation value and read coverage for each CpG cytosine.

Performance of DIRECTION for different BS-seq CCR values: We analyzed the results for whole methylome predictions in H1 and NPC by binning all high-coverage cytosines (sequencing depth ≥ 20) in the BS-seq datasets based on their CCRs. We created 5 bins, based on intervals of 0.2 from 0 (completely unmethylated) to 1 (completely methylated) based on the CCR. We find that for the SVM model (tested on the H1 dataset) and the RF model (tested on the NPC dataset), our accuracy for the extremal values of BS-seq CCRs are accurate (Supp Figs 2I, 2J, respectively, Supp Table T6), while the performance is limited in the interval [0.4, 0.6) corresponding to intermediate methylation. This suggests that cytosines in these regions correspond to data points near the classification boundary, and are prone to be misclassified due to their proximity to the boundary. However, intermediate methylation is relatively uncommon in *in vitro* cell lines due to their homogeneity, and in mammalian systems (H1 and NPC: <3%, (Zeng,J., et al. 2012a)). Thus, we find that the lower predictive ability of DIRECTION in cytosines with intermediate methylation only has a modest effect on the overall prediction metric by contrasting precision and recall in balanced sets sampled from the methylome by including or withholding cytosines with intermediate methylation (Supp Figs 2K, 2L). For datasets with significantly higher amounts of intermediate methylation, we recommend using regression based approaches (Zhang,W., et al. 2015).

S12. Additional discussion on OFS feature contributions to DNA methylation status prediction and transfer learning

OFS feature contributions to DNA methylation status prediction: We discovered that DNase and histone state features impacted the recall in CGI regions significantly, whereas high precision values were predominantly governed by H2AK5ac histone modification, known to be associated with regions of active chromatin and insulator region shores (Wang,T., et al. 2012). Similarly, if any of the aforementioned features or the clusters they belong to are removed, the DNA methylation prediction in non-CGI regions drops (Supp Fig 2D), suggesting similar informational content of predictors in CGI and non-CGI OFSs. In summary, a small set of features (H3K4me3; either DNase or Histone states; and H2AK5ac, along with Repeats for Non-CGI regions) can near optimally predict methylation status at single nucleotide resolution.

Many aspects of our learned models are consistent with previous findings: a significant gain in prediction accuracy when highly discriminative epigenomic features are included (Figs 2B and 2C) (Yan,H., et al. 2015), and significantly improved prediction performance in CGI regions with respect to non-CGI regions.

Transfer learning: Successful transfer learning between two cell types requires that a set of discriminative features and its associated model decision boundary in one cell type, also have comparable predictive power in the other. In order to identify methylomes that are significantly dissimilar with respect to H1 and NPC, we performed clustering for all reference methylomes in the Roadmap Epigenome Consortium datasets (top eight principal components accounting for 81% of variation in the data, Euclidean distance measure and average linkage were used, Supp Fig 3C). We chose to use the methylomes for Mesenchymal Stem Cells (MSC), and fetal fibroblast cell line IMR90, which show distinct divergence from the H1 and NPC methylomes. We analyzed the predictive performance for the NPC-trained predictive model on H1, MSC, and IMR90 methylomes.

We find that H1 predictions using the NPC-trained model are comparable to the NPC whole methylome predictions. The metrics for the MSC cell line (totipotent, but nearly terminally differentiated) are still fairly accurate (TPR: 0.87, TNR: 0.71, Accuracy: 0.85) (Supp Table T6). However, we find that for the terminally differentiated IMR90, the metrics for the predictions are very modest (TPR: 0.86, TNR: 0.23, Accuracy: 0.69) (Supp Table T6). This suggests that transfer learning only works within similar methylation paradigms, where the relationship between methylation and discriminative input features are similar. Given that the methylation profile and prevalence in stem cells and terminally differentiated cells are very distinct, we find that such transfer learning is not feasible.

NB: For evaluating IMR90, the sex chromosomes were left out during evaluation, as IMR90 is a female cell line, as opposed to H1, NPC, and MSC.

S13. BS-seq driven 5-hmC status identification

There is a vast number of BS-seq datasets which are publicly accessible, and only a handful of these have an accompanying TAB-seq counterpart. We used the NPC BS-seq and TAB-seq datasets to train and test a 5-hmC status prediction classifier using only CpG sites where BS-seq CCR could be reliably estimated (coverage ≥ 20). We trained our model using the 5-hmC OFS, where the BS-seq level feature was excluded. Such a classifier performs comparably to our previously reported classifiers, achieving a precision of 0.74, recall of 0.8 and an F-score of 0.77 (Supp Table T10). Hence, we show that our method has the capability of performing *de novo* 5-hmC modification map reconstruction based on the BS-seq dataset and a handful of other features. Such an approach trades off the size and diversity of the training data for a smaller, higher quality training set, and can likely be useful in reconstructing 5-hmC maps of experimental conditions with published BS-seq data.

S14. Enhancer identification

The ChIP-seq H3K27ac raw SRA file for calling enhancer regions in NPC was obtained under the accession

(GSM818031). Raw SRA files were mapped to the reference human hg19 genome using Bowtie2 to create the bam file. The obtained bam file was used as an input to the enhancer calling tool ROSE (Whyte,W.A., et al. 2013).

S15. 5-hmC prediction feasibility in enhancers for reduced representation datasets

Simulations were performed in enhancer regions to create downsampled TAB-seq datasets. The overall number of TAB-seq reads (approx. 84,000,000) were downsampled to different downsampling levels (75, 50, 25, 12, 5, 1 percentage of original-not all shown) (Fig 5A). A linear regression was used to fit the number of reads mapping to the enhancer to the sum of sequencing depth (Fig 5B) across all cytosines in it. 25 downsampling operations for each downsampling level were performed, and the obtained variance was low as shown in the box plot (Fig 5A). The histogram was divided into 3 categories: low, medium and high sequencing depth. Enhancers with the sum of cytosine sequencing depths > 60 were regarded as high, based on the Fisher test obtained p-value < 0.05 for discriminating against high and low hydroxymethylation coverage (Supp Table T13A).

As shown in Fig 5A, downsampling to 5% of the total number of reads still leaves more than 10,000 enhancer regions with the sum of cytosine sequencing depths (≥ 60) and over 2,000 cytosines with individual sequencing depths ≥ 20 (corresponding to our sequencing depth levels required for assigning class labels, (details on filtering training data based on sequencing depth in Supp Text S1) which suffices for the purpose of training our classifier at the resolution of individual enhancers. The downsampling size of 12% contains ~10 million reads, which corresponds to the amount of RRBS-seq reads obtained in previous studies (Bock,C., et al. 2010, Chatterjee,A., et al. 2012). This suggests that even for RRBS-seq datasets, it is possible to train a model to successfully reconstruct the hydroxymethylome in enhancer regions.

S16. 5-hmC enrichment ratio calculation

Unlike DNA methylation status, hydroxymethylation status cannot be successfully imputed using neighboring CpG site information (Supp Table T14), suggesting that CpG sites of similar hydroxymethylation status do not occur in as frequent and long stretches as similarly DNA methylated CpG sites. Hence, we devised a metric for identifying 5-hmC enrichment in a given genomic region. We used this to identify 5-hmC enrichment in enhancers.

GTF hg19 files were obtained from UCSC Genome Browser (Speir,M.L., et al. 2016), and further intersected with an available list of annotated enhancers (Supp Data 1 on the tool website). The regions that contained less than 10 CpG sites upon intersecting with enhancer and gene annotations were discarded from analyses.

We define the ratio of the number of 5-hmC modified cytosines to the sum of 5-hmC and 5-mC modified cytosines in an enhancer as the 5-hmC enrichment ratio. We performed calculation of 5-hmC enrichment ratio in the intragenic enhancer regions, using 5,000bp sliding windows spanning intragenic enhancers. 5-hmC enrichment ratio in a given region is defined as the ratio of the number of cytosines with 5-hmC modification to the number of cytosines with 5-hmC or 5-mC modification. This may be estimated using BS-seq data, or based on SVM predictions. Genes depicted in Fig 5C were sorted based on the gain of 5-hmC enrichment ratio in intragenic enhancers in NPC vs H1 (Supp Data 2, Supp Data 3 on the tool website).

S17. Gene ontology analysis

Gene sets that were used in gene ontology analysis were previously sorted based on the criteria described in the results section. We performed our analysis using the DAVID gene ontology toolkit (Huang,D.W., et al. 2009) using Gene Ontology database (The Gene Ontology Consortium. 2017) biological processes. The first 500 gene or pseudogene entries were used as input to DAVID to calculate the GO terms related to 5-hmC enrichment ratio in intragenic enhancer regions. Pseudogene entries were discarded by DAVID. All of the included GO terms had Benjamini corrected p-value < 0.05 Supp Data 2, Supp Data 3 on the tool website).

S18. Strengths, limits and relevance of DIRECTION

Use of DIRECTION in reduced representation datasets: Both BS-seq (Meissner,A., et al. 2005) and TAB-seq (Plongthongkum,N., et al. 2014) protocols have reduced representation versions where assays query a limited set of CpGs. DIRECTION is ideally suited to impute methylation or hydroxymethylation status in such reduced representation datasets (as well as existing low coverage whole genome datasets), being able to make use of relevant genome-wide traits (based on genomic annotation, DNA sequence and relevant publicly available genome-wide assays) to create whole-genome scale datasets.

Starting point for 5-hmC functional studies: One of the key differences that sets DIRECTION apart from other predictors is the ability to predict 5-hmC modifications. 5-hmC modifications are known to be cell-type or developmental stage specific (Wang,T., et al. 2012), and hence *in silico* detection of differentially hydroxymethylated regions can be performed by integrating reduced representation datasets and available genomic and epigenomic traits using DIRECTION. *In silico* detection of such differentially hydroxymethylated regions (as we show in H1 and NPC) can be the starting point of deeper functional studies in such regions.

5-hmC status prediction and correlative studies: The molecular mechanisms underlying 5-hmC creation and potential maintenance in the genome, its stability and regulatory potential, are presently all subject to a lot of scientific debate (Hahn,M.A., et al. 2014, Shen,L. and Zhang,Y. 2013). As we have shown, DIRECTION is capable of testing predictive powers of different sets of genomic and epigenomic features with respect to 5-hmC status prediction. Such correlative studies, in conjunction with perturbation models, can lead to a better understanding of 5-hmC.

Potential for use in oxBS-seq datasets: The oxBS-seq protocol (Booth,M.J., et al. 2012) allows for positive readouts of 5-mC modifications (as opposed to 5-hmC modifications in TAB-seq experiments). As future work, we will consider additional experiments to train a model for directly predicting 5-mC modifications. However, likelihood based models like MLML (Qu,J., et al. 2013) can integrate datsets from any two of BS-seq, oxBS-seq and TAB-seq datasets, to estimate CCRs for the third. Estimated CCRs for TAB-seq or BS-seq datasets generated in this fashion can then be used for analysis in DIRECTION.

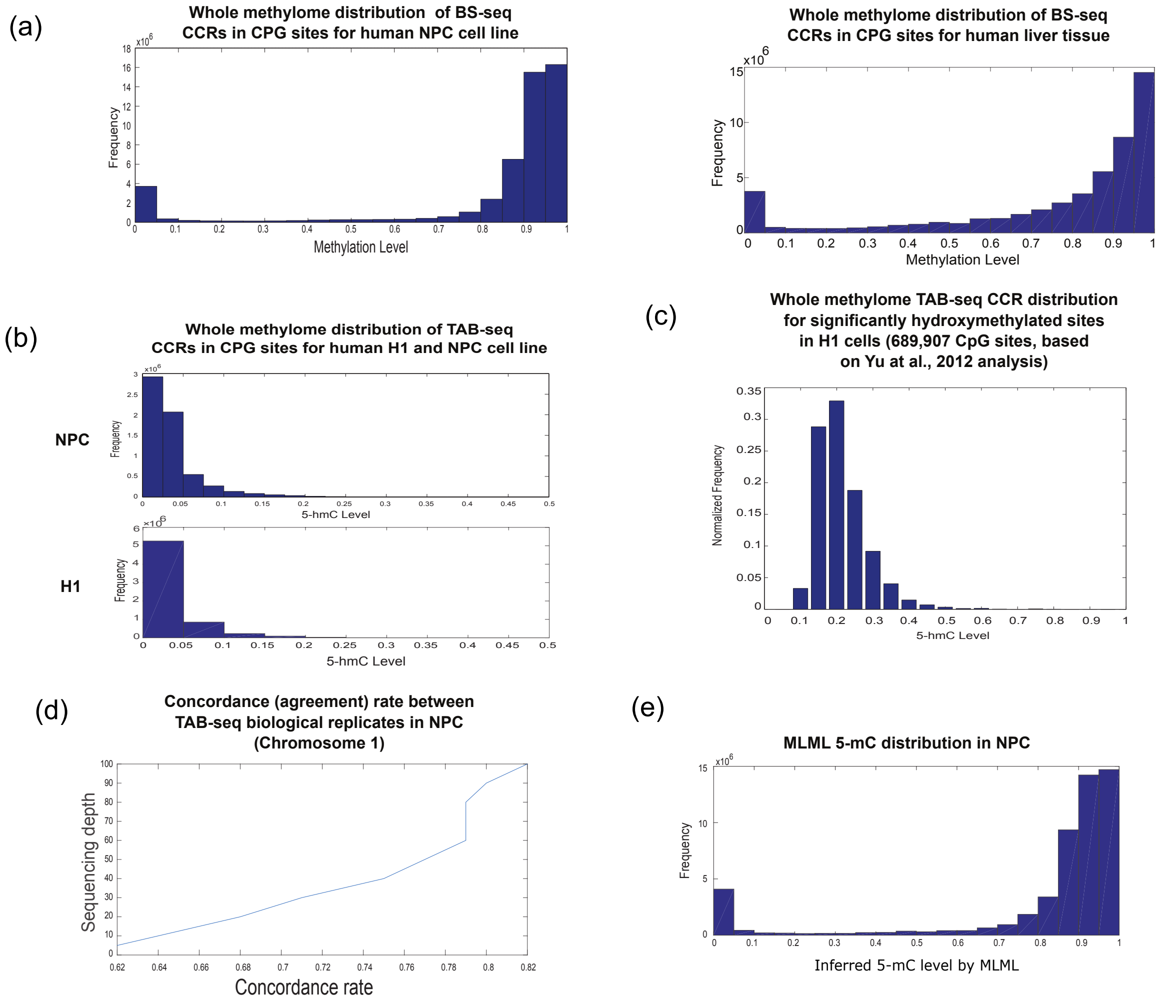
Limits of transfer learning using DIRECTION: Transfer learning for the purposes of prediction requires that the set of input features used for prediction in the source dataset, are discriminative in the target dataset and have similar correlational structure (Pan,S.J. and Yang,Q. 2010). While an in-depth analysis of transfer learning for methylation prediction is beyond the scope of this paper, we used the NPC-trained methylation prediction SVM to predict the methylome in H1, MSC and IMR90. Based on our NPC-trained SVM's performance in the whole genome NPC dataset, we find drops in accuracy in the pluripotent H1 (7% decrease) and near-differentiated, totipotent MSC cell lines (11% decrease). However, the accuracy for the NPC-trained SVM in the terminally differentiated IMR90 cell line drops by over 25%, suggesting that the OFS and SVM decision boundary for NPC is not suited for predicting the IMR90 methylome. Such results are in agreement with studies showing large-scale epigenetic reprogramming during differentiation (Teif,V.B., et al. 2014) that likely cause a change in the correlational structure between the input features and the response variable

(DNA methylation status). The limited number of input features in the OFS used by DIRECTION, while practical, does not lend itself to transfer learning in such scenarios. However, transfer learning is potentially feasible in closely related cell types or conditions where methylation paradigms remain unchanged.

Tradeoffs underlying classification frameworks for methylation prediction: DNA methylation is essentially a discrete phenomenon at the level of individual alleles. Hence, methylation prediction lends itself naturally to a classification framework. However, bisulfite-based sequencing assays typically agglomerate signal from both alleles across millions of cells, giving rise to CCRs that may be closer to 0.5 (intermediate methylation) than 0 or 1 (completely unmethylated or methylated). Classification algorithms excel at predicting methylation status in cytosines with CCRs that have extremal values (near 0 or 1). However, their performance degrades for predicting methylation status in cytosines with CCRs commensurate with intermediate methylated levels due to the near-arbitrariness of class label assignments for such intermediate CCRs. DIRECTION is designed for predicting methylation and hydroxymethylation at CpG cytosines in mammalian genomes, which are well known to show a bimodal distribution of CCRs even in highly heterogeneous mammalian tissues like muscle (Couldrey,C., et al. 2014) and brain (Zeng,J., et al. 2012b). However, some mammalian datasets can possess some intermediate methylation (Lister,R., et al. 2009). Cancer datasets, with their underlying mixture of cell-types and genome heterogeneity, can be a source of such intermediate methylation in mammals (Ahn,J.B., et al. 2011). Additionally, for invertebrates, the degree of intermediate methylation is known to be higher (Elango,N. and Yi,S.V. 2008). For such situations, a regression based approach is possibly more suitable (Zhang,W., et al. 2015). Given the flexibility of our learning framework, such regression-based predictors can be incorporated into our learning framework if needed.

Relevance of in silico prediction: Widespread use of epigenome-querying assays like BS-seq naturally leads to a discussion of relevance of *in silico* epigenome prediction. However, for an *in vivo* sourced sample with a limited DNA yield (like clinical samples), only a few assays can be performed, necessitating the *in silico* prediction of some assays based on the outcome of others. Secondly, paralleling the rise of whole genome assays, are reduced representation BS-seq (Meissner,A., et al. 2005) and TAB-seq (Plongthongkum,N., et al. 2014) assays, for which in silico prediction is especially relevant. Recent developments in single cell technologies allow BS-seq assays to be performed on individual cells (Kantlehner,M., et al. 2011), with some studies contemplating single-cell TAB-seq as future work (Guo,H., et al. 2013). Given the destructive nature of next generation sequencing, *in silico* prediction tools can be potentially useful for using single-cell methylation data and underlying genomic sequence to make a model-based prediction for 5-hmC status, or for imputing methylation status.

Supplementary Figures and Captions



Whole methylome distribution of BS-seq CCRs in CPG sites for human NPC cell line

Whole methylome distribution of BS-seq CCRs in CPG sites for human liver tissue

Whole methylome distribution of TAB-seq CCRs in CPG sites for human H1 and NPC cell line

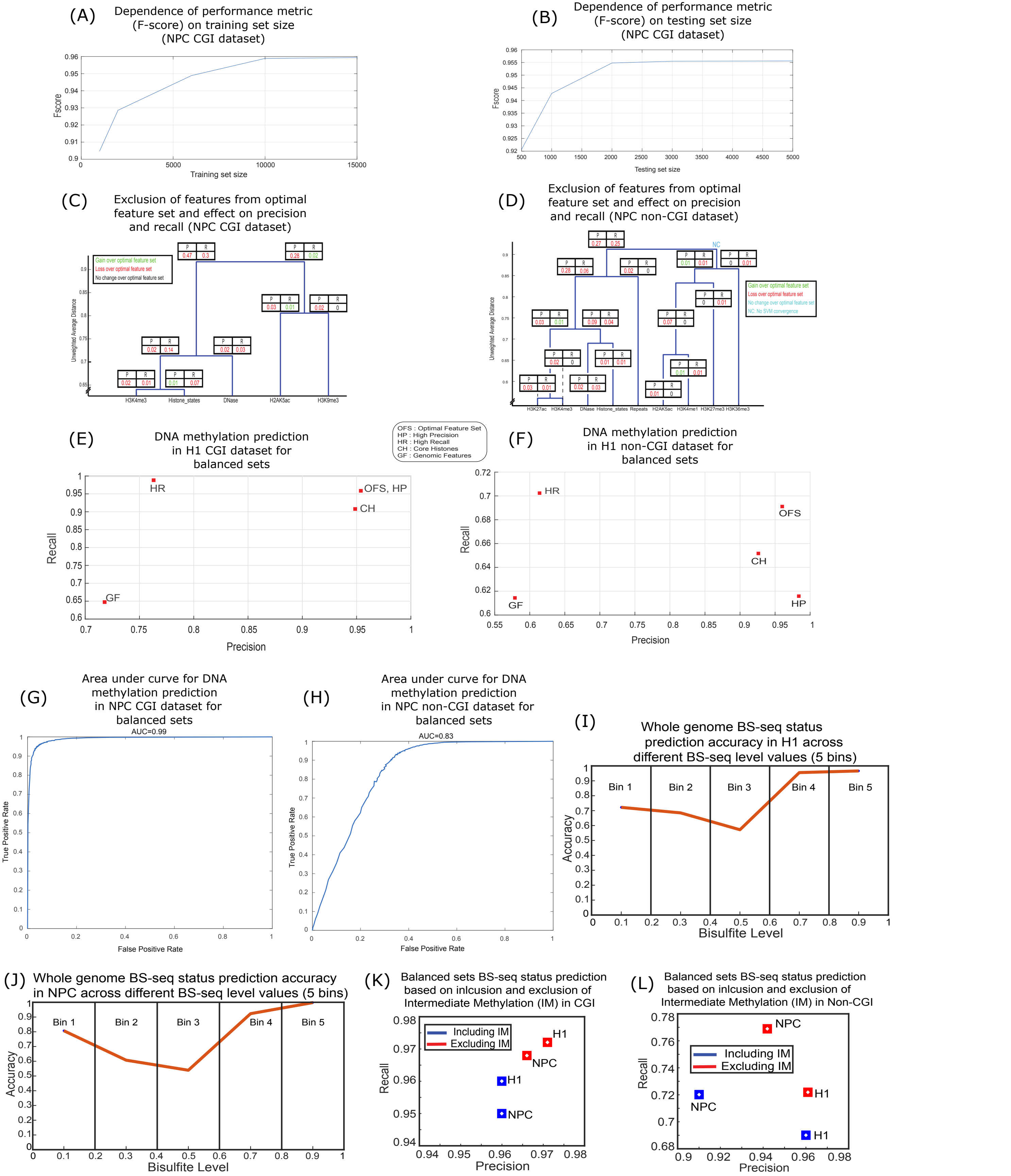
Whole methylome TAB-seq CCR distribution for significantly hydroxymethylated sites in H1 cells (689,907 CpG sites, based on Yu at al., 2012 analysis)

Concordance (agreement) rate between TAB-seq biological replicates in NPC (Chromosome 1)

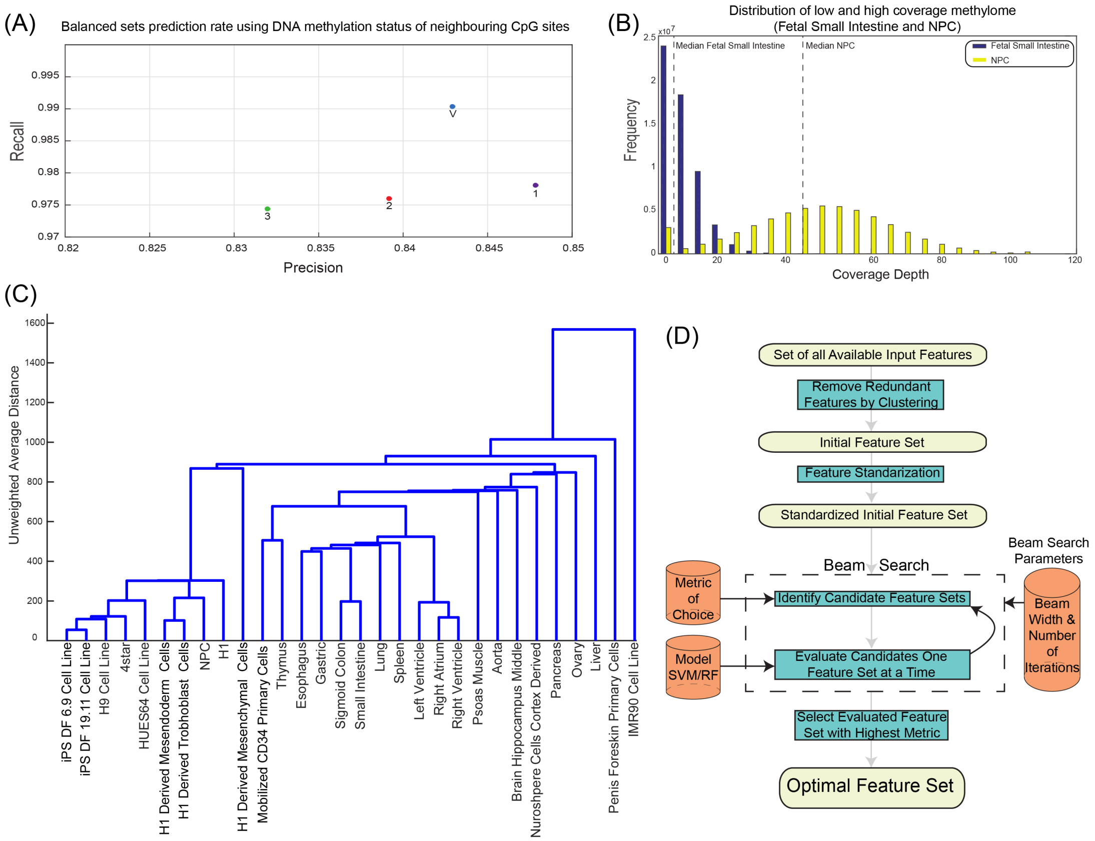
MLML 5-mC distribution in NPC

Supplementary Figure 1: (a) Empirical distributions of methylation levels in NPC and liver tissue. (b) Distribution of hydroxymethylation levels in H1 and NPC. (c) Distribution of 5-hmC levels in the set of CpG sites identified as significantly hydroxymethylated in (Yu,M., et al. 2012). (d) Concordance rate between NPC replicates as a function of minimum depth of mapping at either replicate. (e) Inferred 5-mC level distribution in NPC by the tool MLML (Qu,J., et al. 2013) by jointly analyzing BS-seq and TAB-seq CCRs.

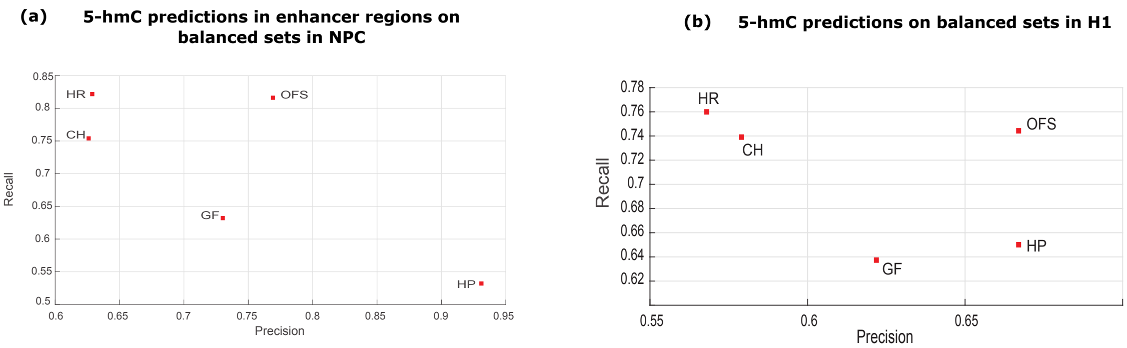
Machine learning framework for DNA hydroxymethylation prediction



Supplementary Figure 2: (A, B) F-score trajectories with increase of training (A) and testing (B) set size for SVM classification on balanced sets in NPC for DNA methylation status. (c, d) Hierarchical clustering of features in OFS for predicting methylation status in NPC CGI (C) and non-CGI (D) regions, and corresponding changes in precision and recall with respect to OFS (E, F) Prediction metrics for DNA methylation status prediction in H1 cells CGI (E) and non-CGI (F) regions. (G, H) AUC curves for DNA methylation status predictions in NPC CGI (G) and NPC non-CGI (H) regions. (I, J) Methylation status prediction accuracy obtained by binning the whole genome based on the BS-seq level in H1 (I), and NPC (J). (K, L) Balanced sets predictions in H1 and NPC based on exclusion and inclusion of intermediate methylation [0.4 0.6) in CGI (K), and Non-CGI (L).



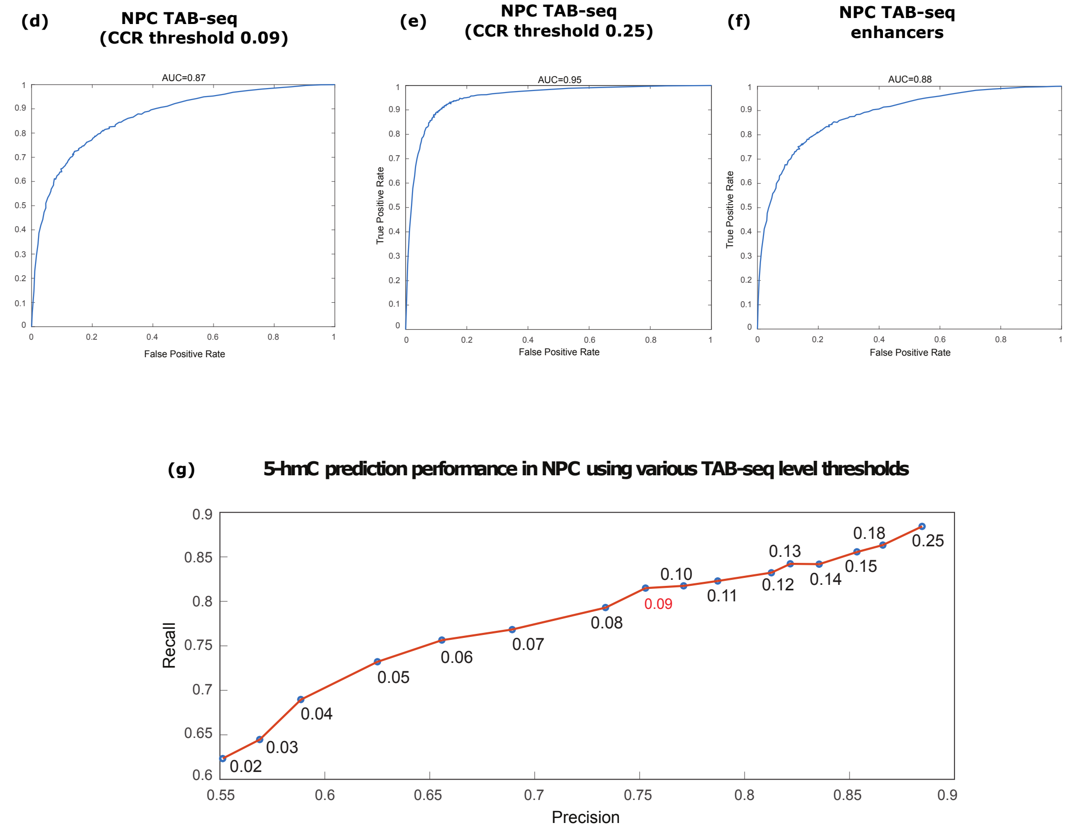
Supplementary Figure 3: (a) Precision/Recall plot for prediction of DNA methylation status using the methylation status of the 1st, 2nd, and 3rd nearest neighbor, and a vote amongst all three. (b) Empirical distributions of high coverage methylome such as NPC (yellow) and low coverage methylome such as Fetal Small Intestine (blue) c) Hierarchical clustering of reference methylomes from the Roadmap Epigenome Consortium depicting distinctive methylation profile of IMR90 with respect to H1 and NPC cell lined d) Schematic representation of optimal feature set finding algorithm embedded within DIRECTION.


(a) **5-hmC predictions in enhancer regions on balanced sets in NPC**

(b) **5-hmC predictions on balanced sets in H1**

(c) **Prediction metrics for H1 5-hmC whole methylome analysis**

Prediction Metric	H1 5-hmC Whole Genome
True Positive Rate	0.67
True Negative Rate	0.75
Accuracy	0.75


(d) **NPC TAB-seq (CCR threshold 0.09)**

(e) **NPC TAB-seq (CCR threshold 0.25)**

(f) **NPC TAB-seq enhancers**

(g) **5-hmC prediction performance in NPC using various TAB-seq level thresholds**

Supplementary Figure 4: (a) Precision/Recall plot for 5-hmC status predictions in NPC in enhancer regions using balanced sets for SVM. (b, c) Precision/Recall plot for 5-hmC status predictions in H1 in balanced sets (b) and whole genome (c) for SVM. (d) ROC curve for 5-hmC status predictions in NPC RF model. (e) ROC curve for 5-hmC status predictions using threshold of 0.25 in NPC RF model. (f) ROC curve for 5-hmC predictions in enhancer regions in NPC RF model. (g) Precision/Recall plot for 5-hmC status predictions in NPC using various 5-hmC level thresholds for SVMs. Threshold of 0.09 is marked red to symbolize the default value that was used in this paper (Underlying data: Supp Table T13B).

Supplementary Tables

Citations	Samples	Model	Features	Response variable	Performance metric
(Feltus,F.A., et al. 2003)	Restriction Landmark Genome Scanning for control fibroblast and DNMT1-overexpressed fibroblast cell lines	LDA	k-mer and consensus motifs in CGI	Methylation prone CGIs vs Methylation resistant CGIs for DNMT1 overexpression among unmethylated CGIs in controls	ACC: 0.82
(Bhasin,M., et al. 2005)	MethDB (curated database of ~5,000 experimentally determined methylation of DNA fragments in species from plants to humans)[1]	SVM (best), ANN, NB, LR, k-NN, decision tree	Genomic features (binary sparse encoding of sequence)	Methylation status of DNA fragments of 39bp	SVM (polynomial kernel degree 6) metrics: ACC: 0.7506, MCC: 0.504, AUC: 0.82
(Feltus,F.A., et al. 2006)	Restriction Landmark Genome Scanning for control fibroblast and DNMT1-overexpressed fibroblast cell lines	LDA	Discriminative motifs in CGI obtained using MAST	Methylation prone CGIs vs Methylation resistant CGIs for DNMT1 overexpression among unmethylated CGIs in controls	ACC: 0.84
(Bock,C., et al. 2006)	Methylation status of CGI in the non-repetitive parts of human Chromosome 21 (HpaII-McrBC PCR method)- 149 CGIs[2]	SVM linear kernel (best), RBF SVM, Decision tree, AdaBoost	k-mer and nucleotide content, predicted DNA structure, repeat regions, TFBS, evolutionary conservation, SNP frequency	CGI methylation status for whole CGI	Linear SVM metrics: CC:0.74, ACC:0.915
(Das,R., et al. 2006)	Human brain data[3] with methylation status of ~5,500 genomic domains	SVM RBF kernel (best), K-means, LDA, LR	k-mer content and repeat regions	Methylation status of 800bp regions	RBF SVM metrics: ACC: Overall: 0.86, CGIs: 0.965, non-CGIs: 0.84
(Fang,F., et al. 2006)	Human brain data[3] with methylation status of ~5,500 genomic domains	SVM (linear kernel)	Nucleotide and dinucleotide content, Alu element, TFBSs	Methylation status of CpG-rich 200-500bp regions (CGI fragments)	ACC: 0.8303-0.8499, CC: 0.567-0.686
(Kim,S., et al. 2008)	Bisulfite treated tumor and normal human samples followed by targeted 454 sequencing of 25 gene-related CGIs	NB (best), SVM (SMO), ANN, kNN (k=3)	30bp flanking sequence of each CpG site	Methylation status of randomly selected 41 CpG sites from sequenced dataset (methylation level ≥0.5 or ≤ 0.01)	NB metrics: ACC:>0.75
(Bock,C., et al. 2007)	Methylation status of CGI in the non-repetitive parts of human Chromosome 21 (HpaII-McrBC PCR method)[2]	SVM (linear kernel)	DNA sequence patterns, repeat distribution, predicted DNA helix structure, predicted TFBS, genetic variation, and CGI attributes	Methylation status of CGI	CC: 0.698, ACC: 0.868
(Fan,S., et al. 2008)	Human Epigenome Project[4] data for chromosomes 6, 20, and 22, using methylation status in human CD4+ T lymphocytes	SVM (linear kernel)	Nucleotide content, Alu annotation, TFBS, and histone methylation (H3K4me1, H3K4me2, H3K4me3, and H3K9me1)	CGI methylation status	ACC: 0.8994
(Carson,M.B., et al. 2008)	Methylation status of CGI in the non-repetitive parts of human Chromosome 21 (HpaII-McrBC PCR method)[2]	Alternative decision tree (best), decision tree, AdaBoost, SVM	4-mer frequencies in CGI	Methylation status of CGIs on chromosome 21	Alternating decision tree metrics: ACC: 0.9063, AUC: 0.8906, MCC: 0.742
(Bock,C., et al. 2009)	Various vertebrate epigenomic datasets[5]	AdaStump, Decision Tree, RF, NB, LR, SVM (linear, RBF kernels)	DNA sequence content, predicted DNA structure, evolutionary history and population variation, annotation of repeats, genes, regulatory regions, chromosomal bands and isochores, histone modification	Prediction of various epigenetic features (including DNA methylation)	AdaStump metrics: for all epigenome predictions: CC: 0.498, ACC: 0.749
(Previti,C., et al. 2009)	Human Epigenome Project[4] data for chromosomes 6, 20, and 22, using methylation status for all samples, and Epigraph datasets[5]	Decision tree (best), SVM	Nucleotide content, evolutionary conservation, DNA structure prediction	CGI methylation status (2-way: methylated/unmethylated, or 4-way: methylation patterns across tissues)	Decision tree metrics: 2-way: CC:0.775, ACC: 0.9167; 4-way: CC: 0.707, ACC: 0.8939
(Lu,L., et al. 2010)	Human Epigenome Project[4] data for chromosomes 6, 20, and 22, using methylation status in human CD4+ T lymphocytes	k-NN	5-mer frequency in 499bp upstream and downstream of CpG site	Methylation status of CpG sites	ACC: 0.7745
(Fan,S., et al. 2010)	Human Epigenome Project[4] data for chromosomes 6, 20, and 22 across 1.9 million CpG sites, using methylation status in human CD4+ T lymphocytes	SVM (linear kernel)	DNA sequence derived features: GC content, GC observed/expected ratio, Alu repeats, and repeat masker. 214 TFBS and 38 histone marks.	CGI methylation status in chromosomes 6, 20, and 22	ACC: 0.94, CC: 0.81
(Zhang,W., et al. 2011)	Human Epigenome Project[4] data for chromosomes 6, 20, and 22, using methylation status in human CD4+ T lymphocytes	SVM	Sequence length, nucleotide and dinucleotide content, promoter and TFBS annotation, nucleosome positioning	Methylation status of CGI in chromosome 22	ACC: 0.9059, CC: 0.65
(Zhou,X., et al. 2012)	MethDB (curated database of ~5,000 experimentally determined methylation of DNA fragments in species from plants to humans)[1]	SVM (RBF kernel)	3-mer composition of DNA fragments	Methylation status and level for 400 human DNA fragments in MethDB	Methylation status prediction: ACC: 0.8207, MCC: 0.6411 Methylation level prediction: R: 0.8223, RMSE: 0.2042
(Zheng,H., et al. 2013)	Human Epigenome Project[4] data for chromosomes 6, 20, and 22, using methylation status in several human tissue or cell types	SVM	Gardiner-Garden criteria, 4-mer composition, conserved TFBSs and conserved elements, predicted DNA structure, functional annotation of proximal genes, nucleosome positioning, histone methylation and acetylation	Methylation status of CGI	Metric in human CD4+ lymphocyte: ACC: 0.9313, CC: 0.8302
(Gaidatzis,D., et al. 2014)	BS-seq for H1 and IMR90 cell lines	Linear regression	Dinucleotide sequence derived features created using the sequence environment of 78bp. Each nucleotide interpreted as a categorical variable with 16 states.	DNA methylation levels at CpG nucleotides within partially methylated domains	R=0.86 (for the sequence context of 140bp)
(Ma,B., et al. 2014)	Methylation array data of multiple human tissues	Support vector regression (RBF kernel) (best), linear regression	Methylation beta values in surrogate tissue	Methylation beta values for different tissues	Methylation level prediction: For probes in beta-value range 0.2 to 0.8: R^2: 0.89-0.98
(Yan,H., et al. 2015)	BS-seq for H1, NPC, IMR90 cell lines	RF (best), SVM (RBF kernel), LR, Decision Tree, NB	Nucleotide composition, 16 histone marks, RNA-seq	Methylation status of genomic segments (based on CpG_MPs tool)	RF metrics: H1: AUC: 0.99, NPC: AUC: 0.99, IMR90: AUC: 0.92
(Zhang,W., et al. 2015)	100 blood samples for 450K arrays	RF	Sequence composition, evolutionary rate, copy number variation, haplotype score, recombination rate, SNP presence, annotation of gene body, promoters, CGIs, repeats, DNase, Pol2 and TF ChIP-seq, histone marks, neighboring CpG site methylation level and distance, chromatin states	Methylation status and levels at single CpG sites	Classification: CGI: ACC: 0.98, Whole genome: ACC: 0.92, Regression: R=0.9, RMSE=0.19
(Wang,Y., et al. 2016)	GM12878 and K562 cell lines (RRBS-seq)	Deep Nets (ANN) and SVM	Genomic features, neighboring CpG sites, and Hi-C	Methylation status at CpG dinucleotides across 1kb windows	ACC: 0.721-0.897
(Fan,S., et al. 2016)	BS-seq and methylation arrays for H1 and H9 cell lines	RF (best), LR, SVM	Nucleotide, dinucleotide frequencies and NpN ratios for 500bp flanks, methylation data for 1000bp flanks, histone marks, chromosome organization, chromatin structure, evolutionary features, repeats, TFBS	Methylation status and levels at CpG sites	Metrics for RF: Classification: ACC: 0.93, MCC: 0.86, Regression: Spearman correlation coefficient: 0.7602

Supplementary Table T1: Literature survey of methylation prediction (Methods: NB: Naive Bayes, LR: Logistic Regression, k-NN: k Nearest Neighbor, RF: Random Forest, SVM Support Vector Machine, LDA: Linear Discriminant Analysis, ANN: Artificial Neural Network) (Metrics: ACC: Accuracy, MCC: Matthews Correlation Coefficient, CC: Correlation Coefficient, R: Regression Coefficient, RMSE: Root Mean Square Error) [1](Amoreira,C., et al. 2003) [2](Yamada,Y., et al. 2004) [3](Rollins,R.A., et al. 2006) [4](Eckhardt,F., et al. 2006) [5](Bock,C., et al. 2009)

Mode	**Description**
training	This mode outputs a trained model (SVM or RF) for a given feature set, by sampling balanced sets from given input data
testing	This mode performs testing to provide prediction metrics. It outputs the precision/recall metric on a testing data using user-specified previously trained model with feature set.
whole-genome	This mode performs testing to provide predictions. It outputs the actual predictions of the model for a given input dataset, using user-specified previously trained model with feature set. It can be used to perform imputation, prediction limited to a certain set of genomic regions, or whole genome prediction.
cross-validation	This mode samples balanced sets from the training data, partitions them into k sets for cross validation purposes, and performs testing for each partition, by training on the remaining partition, for a given feature set. It outputs precision/recall metrics across different test partitions, along with the associated model and feature set.
beam_search	This mode takes as input an initial feature set, and training data. It performs beam search to identify the best feature set (a subset of the initial feature set) according to user specified prediction metric, and outputs the best feature set, trained model along with the metric. It also outputs all evaluated feature sets with respective metrics. It has the cross-validation mode embedded inside it.
bed_to_binary	This mode is for database management. It takes as input multiple feature sets in bed format, and merges them together into 1 matrix and saves it in binary file (.mat binary MATLAB format) format. This ensures a consistent, compact input matrix which enhances overall computational performance.
append_bed_to_binary	This mode is also for database management. It takes a single feature in .bed format as an input, appends it to a larger matrix in a user-specified .mat file and eventually updates the .mat binary file.

Supplementary Table T2: Summary of different modes in our epigenome prediction toolkit

Feature	Type	Description	Motivation
Genome-derived features (processed from UCSC Genome Browser datasets (Speir,M.L., et al. 2016)**)**			
CpG island (CGI)	Binary	Presence/absence of CGI annotation at CpG site	CGIs tend to be significantly unmethylated in comparison to non-CGI regions of the genome
Distance to nearest CGI (in bps)	Non-negative integer	Helps distinguish CpGs in CGI, CGI "shores" and non-CGI	Cytosines on the CGI shores (near CGIs) tend to be highly methylated and govern most of methylation within non-CGI regions)
Distance to nearest CGI (in CpGs)	Non-negative integer	Alternative feature for distance to CGI, measured in number of intervening CpGs, rather than genomic coordinates	As above
GC content	Continuous ∈ [0,1]	Percentage of nucleotides which are G/Cs in centered window around CpG site (window sizes: 50, 100, 200, 400, 800bp used)	Higher GC content empirically shows lower methylation levels: fact corroborated in CGIs
CpG density	Continuous ∈ [0,1]	Percentage of dinucleotides which are CpGs in centered window around CpG site (window sizes: 50, 100, 200, 400, 800bp used)	As above
Strand-specific guanine density	Continuous ∈ [0,1]	Percentage of guanines in centered window around CpG site (window sizes: 50, 100, 200, 400, 800bp used)	5-hmC levels can be asymmetrically distributed in a CpG site between strands (Yu,M., et al. 2012)
Repeats (SINEs, LTRs)	Binary	Presence/absence of SINE or LTR annotation at the CpG site	Higher methylation suppresses transcription in repeat regions (Hackett,J.A., et al. 2013)
Alu	Binary	Presence/absence of Alu annotation at the CpG site	As above
Epigenome-derived features			
Enhancers	Binary	Created using a cutoff value of the ChIP-seq H3K27ac and H3K4me3 signal generated using MACS tool[1]	5-hmC is known to be overrepresented in enhancers (Stadler,M.B., et al. 2011)
Core histone modification ChIP-seq signal	Continuous	-log$_{10}$ transformed ChIP-seq p-values based on ChIP binding and input control, as calculated by the MACS tool[1]. (H3K9me3, H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K27ac: available for 109 epigenomes)[2]	Repressive marks like H3K9me3 and H3K27me3 are often mutually exclusive with DNA methylation
Auxiliary histone modification ChIP-seq signal	Continuous	Similarly processed data for additional histone modifications available for a limited number of epigenomes (H2AK5ac, H2AZ, H2BK120ac, H2BK12ac, H2BK15ac, H2BK20ac, H2BK5ac, H3K14ac, H3K18ac, H3K23ac, H3K23me2, H3K4ac, H3K4me1, H3K4me2, H3K56ac, H3K79me1, H3K79me2, H4K20me1, H4K5ac, H4K8ac, H4K91ac)[2]	As above
Histone states	Discrete: 1-15	Using core histone modification signal for core marks to segment data into posterior decoded 15-state HMM annotation tool ChromHMM[3], based on (Chadwick, 2012)	Histone states have been shown to be well correlated with DNA methylation (Kundaje,A., et al. 2015)
BS-seq CCR	Continuous ∈ [0,1]	Percentage of cytosines remaining unchanged based on the Roadmap Epigenome consortium datasets[2]	Used only for predicting 5-hmC status, since 5-hmC modifications show up as part of the BS-seq CCRs
ChIP-seq TF binding-derived features			
DNase-seq signal	Continuous	Regions of open chromatin characterized by DNase digestion and sequencing: coverage signal contrasted with uniformly distributed read set simulation, and -log$_{10}$ transform of p-value used[2]	DNase hypersensitive regions positively correlated to active regulatory regions, negatively correlated to 5-mC
CTCF ChIP-seq signal	Continuous	-log$_{10}$ transformed ChIP-seq p-values based on ChIP binding and input control for CTCF binding[2]	Well-known insulator. Used only for H1 methylation and 5-hmC status prediction.
p300 ChIP-seq signal	Continuous	-log$_{10}$ transformed ChIP-seq p-values based on ChIP binding and input control for p300 binding[2]	p300 marks active transcription sites. Used only for H1 methylation and 5-hmC status prediction

Supplementary Table T3: List of features used for predicting DNA methylation and hydroxymethylation. All features for methylation prediction were used for 5-hmC prediction as well, since 5-hmC is on the demethylation pathway. [1](Zhang,Y., et al. 2008), [2](Chadwick,L.H. 2012), [3](Ernst,J. and Kellis,M. 2012)

Optimal Feature Selection in NPC CGI using SVM for different values of beam width parameter			
Beam Width			
2	3	4	5
DNase	DNase	DNase	DNase
H2AK5ac	H2AK5ac	H2AK5ac	H2AK5ac
H3K4me3	H3K4me3	H3K4me3	H3K4me3
H3K9me3	H3K9me3	H3K9me3	H3K9me3
Histone_states	Histone_states	Histone_states	Histone_states
Bp_to_CGI		Bp_to_CGI	

Supplementary Table T4: Similarity of OFS across different beam widths for beam search using SVM model for methylation status prediction in NPC CGI dataset.

Evaluation on genomic loci subsets by sampling balanced sets

Comparison of different predictive models in NPC dataset

				Datasets			Evaluation metrics	
Data type	**Sampling loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**Precision**	**Recall**	**Fscore**
BS-seq	CGI cytosines	SVM	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 20)	0.96	0.95	0.95
BS-seq	CGI cytosines	RF	(RF OFS)	NPC (depth >= 20)	NPC (depth >= 20)	0.95	0.96	0.95
BS-seq	CGI cytosines	Classification Tree	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 20)	0.94	0.95	0.94
BS-seq	CGI cytosines	Ensemble model with SVM and	SVM + N + C = (SVM features: SVM OFS + nearest neighbor	NPC (depth >= 20)	NPC (depth >= 20)	0.97	0.96	0.96
BS-seq	non-CGI cytosines	SVM	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 20)	0.91	0.72	0.80
BS-seq	non-CGI cytosines	RF	(RF OFS)	NPC (depth >= 20)	NPC (depth >= 20)	0.89	0.74	0.81
BS-seq	non-CGI cytosines	Classification Tree	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 20)	0.71	0.71	0.71
BS-seq	non-CGI cytosines	with SVM and consensus	SVM OFS + nearest neighbor feature) + Consensus	NPC (depth >= 20)	NPC (depth >= 20)	0.93	0.78	0.85

Comparison of predictive abilities for different feature sets in NPC dataset

				Datasets			Evaluation metrics	
Data type	**Sampling loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**Precision**	**Recall**	**Fscore**
BS-seq	CGI cytosines	SVM	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 20)	0.96	0.95	0.95
BS-seq	CGI cytosines	SVM	GF	NPC (depth >= 20)	NPC (depth >= 20)	0.75	0.61	0.67
BS-seq	CGI cytosines	SVM	CH	NPC (depth >= 20)	NPC (depth >= 20)	0.95	0.85	0.90
BS-seq	CGI cytosines	SVM	HP	NPC (depth >= 20)	NPC (depth >= 20)	0.97	0.88	0.92
BS-seq	CGI cytosines	SVM	HR	NPC (depth >= 20)	NPC (depth >= 20)	0.78	0.97	0.86
BS-seq	CGI cytosines	SVM	SVM + N = (SVM features: SVM OFS + nearest neighbor feature)	NPC (depth >= 20)	NPC (depth >= 20)	0.97	0.95	0.96
BS-seq	CGI cytosines	with SVM and consensus reference methylome based	SVM + N + C = (SVM features: SVM OFS + nearest neighbor feature) + Consensus Reference Methylome	NPC (depth >= 20)	NPC (depth >= 20)	0.97	0.96	0.96
BS-seq	non-CGI cytosines	SVM	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 20)	0.91	0.72	0.80
BS-seq	non-CGI cytosines	SVM	GF	NPC (depth >= 20)	NPC (depth >= 20)	0.70	0.67	0.68
BS-seq	non-CGI cytosines	SVM	CH	NPC (depth >= 20)	NPC (depth >= 20)	0.88	0.65	0.75
BS-seq	non-CGI cytosines	SVM	HP	NPC (depth >= 20)	NPC (depth >= 20)	0.94	0.60	0.73
BS-seq	non-CGI cytosines	SVM	HR	NPC (depth >= 20)	NPC (depth >= 20)	0.87	0.72	0.79
BS-seq	non-CGI cytosines	SVM	SVM + N = (SVM features: SVM OFS + nearest neighbor feature)	NPC (depth >= 20)	NPC (depth >= 20)	0.93	0.77	0.84
BS-seq	non-CGI cytosines	with SVM and consensus reference methylome based predictor	SVM + N + C = (SVM features: SVM OFS + nearest neighbor feature) + Consensus Reference Methylome	NPC (depth >= 20)	NPC (depth >= 20)	0.93	0.78	0.85

Comparison of predictive abilities for different feature sets in H1 dataset

				Datasets			Evaluation metrics	
Data type	**Sampling loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**Precision**	**Recall**	**Fscore**
BS-seq	CGI cytosines	SVM	OFS	H1 (depth >=20)	H1 (depth >=20)	0.96	0.96	0.96
BS-seq	CGI cytosines	SVM	GF	H1 (depth >=20)	H1 (depth >=20)	0.72	0.65	0.68
BS-seq	CGI cytosines	SVM	CH	H1 (depth >=20)	H1 (depth >=20)	0.95	0.91	0.93
BS-seq	CGI cytosines	SVM	HP	H1 (depth >=20)	H1 (depth >=20)	0.96	0.96	0.96
BS-seq	CGI cytosines	SVM	HR	H1 (depth >=20)	H1 (depth >=20)	0.76	0.99	0.86
BS-seq	non-CGI cytosines	SVM	OFS	H1 (depth >=20)	H1 (depth >=20)	0.96	0.69	0.80
BS-seq	non-CGI cytosines	SVM	GF	H1 (depth >=20)	H1 (depth >=20)	0.56	0.62	0.59
BS-seq	non-CGI cytosines	SVM	CH	H1 (depth >=20)	H1 (depth >=20)	0.93	0.65	0.77
BS-seq	non-CGI cytosines	SVM	HP	H1 (depth >=20)	H1 (depth >=20)	0.98	0.62	0.76
BS-seq	non-CGI cytosines	SVM	HR	H1 (depth >=20)	H1 (depth >=20)	0.61	0.70	0.65

Comparisons of predictions involving the Consensus Reference Methylome

				Datasets			Evaluation metrics	
Data type	**Sampling loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**Precision**	**Recall**	**Fscore**
BS-seq	Cytosines with disagreement threshold = 0	SVM	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 5)	0.87	0.99	0.93
BS-seq	Cytosines with disagreement threshold = 0	Single predictor variable	Consensus Reference Methylome	NPC (depth >= 5)	NPC (depth >= 5)	0.98	0.99	0.98
BS-seq	Cytosines with disagreement threshold <= 4	Single predictor variable	Consensus Reference Methylome	NPC (depth >= 5)	NPC (depth >= 5)	0.93	0.99	0.96
BS-seq	Cytosines with disagreement threshold <= 8	Single predictor variable	Consensus Reference Methylome	NPC (depth >= 5)	NPC (depth >= 5)	0.88	0.98	0.93
BS-seq	Cytosines with disagreement threshold <= 12	Single predictor variable	Consensus Reference Methylome	NPC (depth >= 5)	NPC (depth >= 5)	0.85	0.97	0.91
BS-seq	CGI cytosines	with SVM and consensus reference methylome based predictor	SVM + N + C = (SVM features: SVM OFS + nearest neighbor feature) + Consensus Reference Methylome	NPC (depth >= 20)	NPC (depth >= 20)	0.97	0.96	0.96
BS-seq	non-CGI cytosines	with SVM and consensus reference methylome based predictor	SVM + N + C = (SVM features: SVM OFS + nearest neighbor feature) + Consensus Reference Methylome	NPC (depth >= 20)	NPC (depth >= 20)	0.93	0.78	0.85

Comparisons of predictions involving the Nearest Neighbor Methylation Status predictor

				Datasets			Evaluation metrics	
Data type	**Sampling loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**Precision**	**Recall**	**Fscore**
BS-seq	All cytosines with nearest neighbor distance within 2 - 20 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	0.96	0.98	0.97
BS-seq	All cytosines with nearest neighbor within distance 20 - 50 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	0.96	0.98	0.97
BS-seq	All cytosines with nearest neighbor within distance 50 - 100 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	0.95	0.97	0.96
BS-seq	All cytosines with nearest neighbor within distance 100 - 200 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	0.95	0.97	0.96
BS-seq	All cytosines with nearest neighbor within distance 200 - 500 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	0.91	0.95	0.93
BS-seq	All cytosines with nearest neighbor within distance 500 - 1000 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	0.81	0.92	0.86
BS-seq	All cytosines with nearest neighbor within distance 1000 - 1500 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	0.70	0.89	0.78
BS-seq	All cytosines with nearest neighbor within distance 1500 - 2000 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	0.62	0.89	0.73
BS-seq	All cytosines with nearest neighbor within distance 2000 - 2500 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	0.59	0.88	0.71

Prediction metrics with intermediate methylation removed

				Datasets			Evaluation metrics	
Data type	**Sampling loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**Precision**	**Recall**	**Fscore**
BS-seq	CGI cytosines	SVM	(SVM OFS)	H1 (depth >= 20, no intermediate methylation sites)	H1 (depth >= 20, no intermediate methylation sites)	0.97	0.97	0.97
BS-seq	non-CGI cytosines	SVM	(SVM OFS)	H1 (depth >= 20, no intermediate methylation sites)	H1 (depth >= 20, no intermediate methylation sites)	0.96	0.72	0.82
BS-seq	CGI cytosines	SVM	(SVM OFS)	NPC (depth >= 20, no intermediate methylation sites)	NPC (depth >= 20, no intermediate methylation sites)	0.97	0.97	0.97
BS-seq	non-CGI cytosines	SVM	(SVM OFS)	NPC (depth >= 20, no intermediate methylation sites)	NPC (depth >= 20, no intermediate methylation sites)	0.94	0.77	0.85

Supplementary Table T5: Balanced set evaluations for DNA methylation status predictions

Evaluation on genomic loci subsets

Comparison of SVM predictive model in NPC and H1 datasets

				Datasets		Evaluation metrics						
Data type	**Evaluation loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**TP**	**TN**	**FP**	**FN**	**TPR**	**TNR**	**Accuracy**
BS-seq	All cytosines	SVM	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 20)	43477143	3808315	1628764	220277	0.99	0.70	0.96
BS-seq	All cytosines	SVM	(SVM OFS)	H1 (depth >=20)	H1 (depth >=20)	43625145	3253354	1440859	2060474	0.95	0.69	0.93

Transfer learning between datasets using SVM predictive model

				Datasets		Evaluation metrics						
Data type	**Evaluation loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**TP**	**TN**	**FP**	**FN**	**TPR**	**TNR**	**Accuracy**
BS-seq	All cytosines	SVM	(SVM NPC OFS)	NPC (depth >= 20)	H1 (depth >=20)	41317737	3727930	966283	4367882	0.90	0.79	0.89
BS-seq	All cytosines	SVM	(SVM H1 OFS)	H1 (depth >=20)	NPC (depth >= 20)	42286042	2999939	2437140	1411378	0.97	0.55	0.92
BS-seq	All cytosines	SVM	(SVM NPC OFS)	NPC (depth >= 20)	MSC (depth >=20)	23703248	2779019	1132119	3528772	0.87	0.71	0.85
BS-seq	All cytosines	SVM	(SVM NPC OFS)	NPC (depth >= 20)	IMR90 (depth >=20, no sex chromosomes)	24457244	2292688	7853852	3990412	0.86	0.23	0.69

Comparison of different predictive models in NPC dataset

				Datasets		Evaluation metrics						
Data type	**Evaluation loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**TP**	**TN**	**FP**	**FN**	**TPR**	**TNR**	**Accuracy**
BS-seq	All cytosines	SVM	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 20)	43477143	3808315	1628764	220277	0.99	0.70	0.96
BS-seq	All cytosines	with SVM and	SVM OFS) + Consensus	NPC (depth >= 20)	NPC (depth >= 20)	43516018	4072279	1364800	181402	0.99	0.75	0.97
BS-seq	All cytosines	RF	(RF OFS)	NPC (depth >= 20)	NPC (depth >= 20)	43409038	3928194	1508885	288382	0.99	0.72	0.96

Comparisons of predictions involving the Consensus Reference Methylome

				Datasets		Evaluation metrics						
Data type	**Evaluation loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**TP**	**TN**	**FP**	**FN**	**TPR**	**TNR**	**Accuracy**
BS-seq	Cytosines with disagreement threshold = 0	SVM	(SVM OFS)	NPC (depth >= 20)	NPC (depth >= 5)	22892862	1724692	298403	48224	0.99	0.85	0.99
BS-seq	Cytosines with disagreement threshold = 0	Single predictor variable	Consensus Reference Methylome	NPC (depth >= 5)	NPC (depth >= 5)	22931737	1988656	34439	9349	0.99	0.98	0.99

Comparisons of predictions involving the Nearest Neighbor Methylation Status predictor

				Datasets		Evaluation metrics						
Data type	**Evaluation loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**TP**	**TN**	**FP**	**FN**	**TPR**	**TNR**	**Accuracy**
BS-seq	All cytosines with nearest neighbor distance within 2 - 20 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	29550709	14763379	564837	564881	0.98	0.96	0.98
BS-seq	All cytosines with nearest neighbor within distance 20 - 50 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	49875774	24255756	1059991	1063485	0.98	0.96	0.97
BS-seq	All cytosines with nearest neighbor within distance 50 - 100 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	74673162	38853553	1929159	1939160	0.97	0.95	0.97
BS-seq	All cytosines with nearest neighbor within distance 100 - 200 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	134860662	72343025	4301602	4328057	0.97	0.94	0.96
BS-seq	All cytosines with nearest neighbor within distance 200 - 500 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	343604957	178749445	18992938	19013419	0.95	0.90	0.93
BS-seq	All cytosines with nearest neighbor within distance 500 - 1000 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	535404825	184117716	49462992	49353293	0.92	0.79	0.88
BS-seq	All cytosines with nearest neighbor within distance 1000 - 1500 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	516462743	97126961	61567274	61614285	0.89	0.61	0.83
BS-seq	All cytosines with nearest neighbor within distance 1500 - 2000 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	506944740	57313501	65247592	64967145	0.89	0.47	0.81
BS-seq	All cytosines with nearest neighbor within distance 2000 - 2500 bp	Single predictor variable	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	502844979	40255457	65773735	65603667	0.88	0.38	0.81

Comparisons of predictions involving the Nearest Neighbor Methylation Status predictor

				Datasets		Evaluation metrics						
Data type	**Evaluation loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**TP**	**TN**	**FP**	**FN**	**TPR**	**TNR**	**Accuracy**
BS-seq	All cytosines	Nearest neighbor status (N1)	Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	92474170	9903311	2108509	2075216	0.98	0.82	0.96
BS-seq	All cytosines	2nd Nearest neighbor status (N2)	2nd Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	92279738	9765000	2246820	2269648	0.98	0.81	0.96
BS-seq	All cytosines	3rd Nearest neighbor status (N3)	3rd Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	92127642	9647770	2364050	2421744	0.97	0.80	0.96
BS-seq	All cytosines	Vote among 3 Nearest neighbor status (V)	Vote among 3 Nearest neighbor status	NPC (depth >= 5)	NPC (depth >= 5)	46819091	4897408	1108502	455602	0.99	0.82	0.97

Comparisons of predictions in H1 and NPC for different ranges of BS-seq CCRs

				Datasets		Evaluation metrics
Data type	**Evaluation loci constraints**	**Predictive Model**	**Input Features Used**	**Trained on**	**Tested on**	**Accuracy**
BS-seq	All cytosines	RF	(RF OFS)	NPC (depth >= 20, methylation CCR range [0,0.2)	NPC (depth >= 20, methylation CCR range [0,0.2)	0.81
BS-seq	All cytosines	RF	(RF OFS)	NPC (depth >= 20, methylation CCR range [0.2,0.4)	NPC (depth >= 20, methylation CCR range [0.2,0.4)	0.61
BS-seq	All cytosines	RF	(RF OFS)	NPC (depth >= 20, methylation CCR range [0.4,0.6)	NPC (depth >= 20, methylation CCR range [0.4,0.6)	0.54
BS-seq	All cytosines	RF	(RF OFS)	NPC (depth >= 20, methylation CCR range [0.6,0.8)	NPC (depth >= 20, methylation CCR range [0.6,0.8)	0.92
BS-seq	All cytosines	RF	(RF OFS)	NPC (depth >= 20, methylation CCR range [0.8,1.0]	NPC (depth >= 20, methylation CCR range [0.8,1.0]	0.99
BS-seq	All cytosines	SVM	(SVM OFS)	H1 (depth >= 20, methylation CCR range [0,0.2)	H1 (depth >= 20, methylation CCR range [0,0.2)	0.72
BS-seq	All cytosines	SVM	(SVM OFS)	H1 (depth >= 20, methylation CCR range [0.2,0.4)	H1 (depth >= 20, methylation CCR range [0.2,0.4)	0.69
BS-seq	All cytosines	SVM	(SVM OFS)	H1 (depth >= 20, methylation CCR range [0.4,0.6)	H1 (depth >= 20, methylation CCR range [0.4,0.6)	0.57
BS-seq	All cytosines	SVM	(SVM OFS)	H1 (depth >= 20, methylation CCR range [0.6,0.8)	H1 (depth >= 20, methylation CCR range [0.6,0.8)	0.96
BS-seq	All cytosines	SVM	(SVM OFS)	H1 (depth >= 20, methylation CCR range [0.8,1.0]	H1 (depth >= 20, methylation CCR range [0.8,1.0]	0.97

Supplementary Table T6: Whole genome evaluations for DNA methylation status predictions

(A): Biologically meaningful feature sets H1 BS-seq using SVM												
	CGI						Non-CGI					
	OFS	HR	HP	CH	GF		OFS	HR	HP	CH	GF	
Alu_repeat					✓						✓	Alu_repeat
Bp_to_CGI		✓			✓				✓		✓	Bp_to_CGI
CG_sat_50bp	✓		✓		✓		✓	✓			✓	CG_sat_50bp
CpG_sat_50bp					✓						✓	CpG_sat_50bp
CpG_to_CGI		✓			✓		✓	✓			✓	CpG_to_CGI
DNase	✓	✓	✓				✓	✓				DNase
G_sat_50bp					✓						✓	G_sat_50bp
H2AK5ac												H2AK5ac
H3K27ac				✓			✓	✓	✓	✓		H3K27ac
H3K27me3	✓	✓	✓	✓			✓	✓		✓		H3K27me3
H3K36me3	✓		✓	✓						✓		H3K36me3
H3K4me1				✓			✓			✓		H3K4me1
H3K4me3	✓	✓	✓	✓			✓	✓	✓	✓		H3K4me3
H3K79me1												H3K79me1
H3K9ac							✓		✓			H3K9ac
H3K9me3	✓		✓					✓				H3K9me3
Histone_states	✓	✓	✓					✓				Histone_states
Repeats					✓						✓	Repeats
CTCF							✓	✓				CTCF

(B): Random Forest Optimal Feature Sets in NPC		
Feature List	CGI	Non-CGI
CG_sat_50bp		
CpG_sat_50bp	✓	
CpG_to_CGI		✓
G_sat_50bp		✓
Alu_repeat		
Bp_to_CGI	✓	✓
DNase	✓	✓
Repeats		✓
H2AK5ac	✓	✓
H3K27ac		
H3K27me3	✓	✓
H3K36me3	✓	✓
H3K4me1		✓
H3K4me3	✓	✓
H3K79me1	✓	
H3K9ac		
H3K9me3	✓	✓
Histone_states	✓	✓

Supplementary Table T7: (A) Feature sets for methylation status prediction using SVM in H1 CGI and non-CGI datasets. (B) OFS for methylation status prediction using RF in NPC CGI and non-CGI datasets

(A): Various beam search feature sets used to show metric improvement in H1 BS-seq non-CGI using SVM							
	Feature Sets						
	A	B	C	D	E	F	G
Alu_repeat	✓						
Bp_to_CGI	✓					✓	
CG_sat_50bp	✓			✓		✓	
CpG_sat_50bp	✓						
CpG_to_CGI	✓		✓		✓		✓
DNase	✓	✓	✓		✓	✓	✓
G_sat_50bp	✓						
H2AK5ac	✓	✓	✓	✓		✓	✓
H3K27ac	✓	✓	✓	✓	✓	✓	✓
H3K27me3	✓	✓		✓			✓
H3K36me3	✓				✓	✓	
H3K4me1	✓			✓	✓	✓	✓
H3K4me3	✓				✓	✓	✓
H3K79me1	✓						
H3K9ac	✓						
H3K9me3	✓						
Histone_states	✓			✓		✓	✓
Repeats	✓	✓	✓	✓	✓	✓	✓

(B): Prediction metric for H1 Beam Search BS-seq in non-CGI using SVM		
	Precision	Recall
A	0.7970	0.7292
B	0.8170	0.7262
C	0.8180	0.7363
D	0.8470	0.7340
E	0.8570	0.7439
F	0.8970	0.7334
G	0.9200	0.7342

Supplementary Table T8: Underlying data for Figure 2E, showing F-score (for H1 non-CGI methylation status prediction) trajectory as beam search algorithm searches through feature space. (A) For incrementally improved F-scores, the feature sets are shown. (B) For each improved F-score, the corresponding precision and recall values are shown.

(A): Reference methylome sizes	Disagreement threshold			
Size of the reference methylome relative to the whole methylome	0	4	8	12
	0.44236051	0.664482316	0.716133503	0.751079765

(B): Cell lines & tissues used to create reference methylome	
H9 Cell Line	Gastric
HUES64 Cell Line	Left Ventricle
iPS DF 6.9 Cell Line	Lung
iPS DF 19.11 Cell Line	Ovary
4star	Pancreas
IMR90 Cell Line	Psoas Muscle
Mobilzied CD34 Primary Cells Female	Right Atrium
Neurosphere Cultured Cells Cortex Derived	Right Ventricle
Penis Foreskin Keratinocycte Primary Cells skin03	Sigmoid Colon
Aorta	Small Intestine
Adult Liver	Thymus
Brain Hippocampus Middle	Spleen
Esophagus	

Supplementary Table T9: (A) Underlying data for Figure 3B, showing the size of the consensus reference methylome with disagreement thresholds 0, 4, 8, and 12 as a fraction of the entire methylome (in terms of CpG cytosines). (B) List of reference methylomes used to create the consensus reference methylome.

Evaluation on genomic loci subsets by sampling balanced sets

Comparison of predictive abilities for different feature sets in NPC dataset

				Datasets		Evaluation metrics		
Data type	Sampling loci constraints	Predictive Model	Input Features Used	Trained on	Tested on	Precision	Recall	Fscore
TAB-seq	All cytosines	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	0.75	0.80	0.77
TAB-seq	All cytosines	SVM	GF	NPC (depth >= 60)	NPC (depth >= 60)	0.73	0.63	0.68
TAB-seq	All cytosines	SVM	CH	NPC (depth >= 60)	NPC (depth >= 60)	0.62	0.76	0.68
TAB-seq	All cytosines	SVM	HP	NPC (depth >= 60)	NPC (depth >= 60)	0.80	0.73	0.76
TAB-seq	All cytosines	SVM	HR	NPC (depth >= 60)	NPC (depth >= 60)	0.65	0.84	0.73

Comparison of predictive abilities for different feature sets in H1 dataset

				Datasets		Evaluation metrics		
Data type	Sampling loci constraints	Predictive Model	Input Features Used	Trained on	Tested on	Precision	Recall	Fscore
TAB-seq	All cytosines	SVM	(SVM OFS)	H1 (depth >= 60)	H1 (depth >= 60)	0.67	0.74	0.70
TAB-seq	All cytosines	SVM	GF	H1 (depth >= 60)	H1 (depth >= 60)	0.62	0.64	0.63
TAB-seq	All cytosines	SVM	CH	H1 (depth >= 60)	H1 (depth >= 60)	0.58	0.74	0.65
TAB-seq	All cytosines	SVM	HP	H1 (depth >= 60)	H1 (depth >= 60)	0.67	0.65	0.66
TAB-seq	All cytosines	SVM	HR	H1 (depth >= 60)	H1 (depth >= 60)	0.57	0.76	0.65

Comparison of predictive abilities for different feature sets in NPC enhancer dataset

				Datasets		Evaluation metrics		
Data type	Sampling loci constraints	Predictive Model	Input Features Used	Trained on	Tested on	Precision	Recall	Fscore
TAB-seq	Enhancer cytosines	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	0.77	0.82	0.79
TAB-seq	Enhancer cytosines	SVM	GF	NPC (depth >= 60)	NPC (depth >= 60)	0.73	0.63	0.68
TAB-seq	Enhancer cytosines	SVM	CH	NPC (depth >= 60)	NPC (depth >= 60)	0.63	0.75	0.68
TAB-seq	Enhancer cytosines	SVM	HP	NPC (depth >= 60)	NPC (depth >= 60)	0.93	0.53	0.68
TAB-seq	Enhancer cytosines	SVM	HR	NPC (depth >= 60)	NPC (depth >= 60)	0.63	0.82	0.71

Comparison of different predictive models in NPC dataset

				Datasets		Evaluation metrics		
Data type	Sampling loci constraints	Predictive Model	Input Features Used	Trained on	Tested on	Precision	Recall	Fscore
TAB-seq	All cytosines	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	0.75	0.80	0.77
TAB-seq	All cytosines	RF	(RF OFS)	NPC (depth >= 60)	NPC (depth >= 60)	0.78	0.82	0.80

Comparison of different predictive models in NPC dataset

				Datasets		Evaluation metrics		
Data type	Sampling loci constraints	Predictive Model	Input Features Used	Trained on	Tested on	Precision	Recall	Fscore
TAB-seq	Cytosines with high BS-seq levels	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	0.74	0.80	0.77
TAB-seq	All cytosines	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	0.75	0.80	0.77

Evaluation on genomic loci subsets

Comparison of SVM models in NPC and H1 datasets

				Datasets		Evaluation metrics						
Data type	Evaluation loci constraints	Predictive Model	Input Features Used	Trained on	Tested on	TP	TN	FP	FN	TPR	TNR	Accuracy
TAB-seq	All cytosines	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	339376	7269832	1578026	114373	0.75	0.82	0.82
TAB-seq	All cytosines	SVM	(SVM OFS)	H1 (depth >= 60)	H1 (depth >= 60)	84682	1664105	552163	40920	0.67	0.75	0.75

Comparison of SVM models in NPC enhancer regions, non-enhancer regions, & whole genome

				Datasets		Evaluation metrics						
Data type	Evaluation loci constraints	Predictive Model	Input Features Used	Trained on	Tested on	TP	TN	FP	FN	TPR	TNR	Accuracy
TAB-seq	Enhancer cytosines	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	67008	311652	123148	12030	0.85	0.72	0.74
TAB-seq	Non-enhancer cytosines	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	272368	6958180	1454878	102343	0.73	0.83	0.82
TAB-seq	All cytosines	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	339376	7269832	1578026	114373	0.75	0.82	0.82

Comparison of different predictive models in NPC dataset

				Datasets		Evaluation metrics						
Data type	Evaluation loci constraints	Predictive Model	Input Features Used	Trained on	Tested on	TP	TN	FP	FN	TPR	TNR	Accuracy
TAB-seq	All cytosines	SVM	(SVM OFS)	NPC (depth >= 60)	NPC (depth >= 60)	339376	7269832	1578026	114373	0.75	0.82	0.82
TAB-seq	All cytosines	RF	(RF OFS)	NPC (depth >= 60)	NPC (depth >= 60)	348195	7260210	1587648	105554	0.77	0.82	0.82

Transfer learning between H1 and NPC datasets using SVM predictive model

Data type	Evaluation loci constraints	Predictive Model	Input Features Used	Trained on	Tested on	TP	TN	FP	FN	TPR	TNR	Accuracy
TAB-seq	All cytosines	SVM	(SVM NPC OFS)	NPC (depth >= 60)	H1 (depth >=60)	80328	1510583	705685	45364	0.64	0.68	0.68
TAB-seq	All cytosines	SVM	(SVM H1 OFS)	H1 (depth >=60)	NPC (depth >= 60)	295697	6559183	2288675	158052	0.65	0.74	0.74

Supplementary Table T10: Balanced set and whole genome evaluations for 5-hmC status predictions

(A) Biologically meaningful feature sets for 5-hmC status prediction on balanced sets using SVM

	NPC enhancers					NPC					H1				
	OFS	HR	HP	CH	GF	OFS	HR	HP	CH	GF	OFS	HR	HP	CH	GF
Alu_repeat	✓	✓	✓		✓	✓				✓	✓				✓
BS-seq_CCR	✓	✓				✓	✓				✓	✓	✓		
Bp_to_CGI					✓					✓					✓
CG_sat_50bp	✓				✓	✓	✓	✓		✓	✓		✓		✓
CpG_sat_50bp					✓			✓		✓					✓
CpG_to_CGI					✓			✓		✓					✓
DNase	✓	✓				✓					✓				
G_sat_50bp					✓			✓		✓	✓				✓
H2AK5ac															
H3K27ac	✓								✓		✓			✓	
H3K27me3				✓					✓					✓	
H3K36me3				✓					✓			✓	✓	✓	
H3K4me1	✓			✓		✓	✓		✓		✓	✓		✓	
H3K4me3				✓					✓					✓	
H3K79me1															
H3K9ac															
H3K9me3	✓			✓		✓					✓				
Histone_states		✓						✓			✓				
Repeats		✓			✓		✓	✓		✓	✓				✓
CpG_Island								✓							
CTCF											✓	✓	✓		

(B) Random Forest OFS for NPC 5-hmC status prediction

	Features
Alu_repeat	
Bp_to_CGI	
CG_sat_50bp	✓
CpG_sat_50bp	✓
CpG_to_CGI	✓
DNase	✓
G_sat_50bp	
H2AK5ac	
H3K27ac	✓
H3K27me3	✓
H3K36me3	
H3K4me1	✓
H3K4me3	✓
H3K79me1	✓
H3K9ac	
H3K9me3	✓
Histone_states	✓
Repeats	
BS-seq_CCR	✓
CpG_Island	

Supplementary Table T11: Feature sets for 5-hmC status prediction in NPC, NPC enhancers, and H1 using SVM model (A) and in NPC dataset using RF model (B)

Dependence of BS-seq prediction metric in NPC CGI on training and testing set size using SVM

	Training				Testing	
Size	Precision	Recall		Size	Precision	Recall
1000	0.92	0.89		500	0.956	0.888
2000	0.948	0.91		1000	0.956	0.9428
6000	0.958	0.94		2000	0.967	0.9548
10000	0.968	0.95		3000	0.9677	0.9555
15000	0.68	0.951		5000	0.97	0.9555

Supplementary Table T12: Underlying data for Supplementary Figures 2A and 2B showing how precision and recall are affected by training and testing set size.

(A): Distinguishing between unmethylated (or non-hydroxymethylated) and marginally methylated (or hydroxymethylated) CpG sites in BS-seq and TAB-seq experiments based on sequencing depth using Fisher Exact Test

BS-seq			TAB-seq		
Sequencing Depth			*Sequencing Depth*		
Sample1: #C	*Sample1: #T*		*Sample1: #C*	*Sample1: #T*	
Sample2: #C	*Sample2: #T*	*p-value*	*Sample2: #C*	*Sample2: #T*	*p-value*
Sequencing Depth 8			Sequencing Depth 48		
4	4		4	44	
0	8	0.077	0	48	0.117
Sequencing Depth 10			Sequencing Depth 60		
5	5		5	55	
0	10	0.0325	0	60	0.057
Sequencing Depth 12			Sequencing Depth 72		
6	6		6	66	
0	12	0.014	0	72	0.028

(B): Precision/Recall values across different 5-hmC thresholds in NPC

Threshold	Precision	Recall
0.01	0.5838	0.588152327
0.02	0.5514	0.623473541
0.03	0.569	0.644686155
0.04	0.5886	0.689550141
0.05	0.625	0.732021551
0.06	0.6558	0.756401384
0.07	0.6892	0.768338907
0.08	0.7338	0.792954398
0.09	0.7528	0.814894999
0.10	0.771	0.817430025
0.11	0.7872	0.822914489
0.12	0.8128	0.832275241
0.13	0.8218	0.842180775
0.14	0.8356	0.841829539
0.15	0.8536	0.855482061
0.16	0.8564	0.845910707
0.17	0.8602	0.85371179
0.18	0.8658	0.863210369
0.19	0.8652	0.857312723
0.20	0.882	0.870337478
0.25	0.8846	0.884246301

Supplementary Table T13: (A) Fisher's Exact Test shows statistical significance (p< or ~0.05) for distinguishing between a sample that is unmethylated (or non-hydroxymethylated) versus a sample that is marginally methylated (or hydroxymethylated) at sequencing depths of 10 for BS-seq data and 60 for TAB-seq data. (B) Underlying data for Supp Figure 4G, showing smooth change in prediction metric with change in CCR threshold for identifying 5-hmC status in training and testing sets for NPC datasets.

(A) Counts for genome wide imputation of 5-hmC using neighbouring sites across different window sizes (A-I evaluation sets: same as table below)

Ids	TP	TN	FP	FN
A	418464	41328087	1241926	1241202
B	579873	69248852	2140558	2139104
C	738763	105991323	3422785	3424418
D	1082363	193305489	6536477	6529338
E	2296306	497242071	18672602	18674921
F	2737617	718888372	30939446	31002174
G	2278368	642428371	30193001	30290416
H	2090107	603801513	29347926	29401784
I	1978227	585763859	28813642	28723890

(B) Predicting Tab-seq level status using neigbouring CpG sites with respect to the distance to the predicted site: results on balanced sets

	Window_size	Precision	Recall
A	2-20bp	0.8963	0.2521
B	20-50bp	0.8767	0.2133
C	50-100bp	0.8501	0.1775
D	100-200bp	0.813	0.1422
E	200-500bp	0.7516	0.1095
F	500-1000bp	0.6629	0.0811
G	1000-1500bp	0.6091	0.07
H	1500-2000bp	0.5888	0.0664
I	2000-2500bp	0.5788	0.0644

Supplementary Table T14: Prediction metric: TP, TN, FP, FN (A) and Precision and Recall (B) for 5-hmC status prediction based on nearest neighbor's 5-hmC status, showing that such an approach is not feasible for 5-hmC status imputation.

Initial feature sets for BS-seq predictions in NPC and H1	
Initial Feature Set NPC BS-seq	**Initial Feature Set H1 BS-seq**
Alu_repeat	Alu_repeat
Bp_to_CGI	Bp_to_CGI
CG_sat_50bp	CG_sat_50bp
CpG_sat_50bp	CpG_sat_50bp
CpG_to_CGI	CpG_to_CGI
DNase	DNase
G_sat_50bp	G_sat_50bp
H2AK5ac	H2AK5ac
H3K27ac	H3K27ac
H3K27me3	H3K27me3
H3K36me3	H3K36me3
H3K4me1	H3K4me1
H3K4me3	H3K4me3
H3K79me1	H3K79me1
H3K9ac	H3K9ac
H3K9me3	H3K9me3
Histone_states	Histone_states
Repeats	Repeats
	CTCF
	p300

Supplementary Table T15: Initial Feature Sets for NPC and H1 methylation status prediction

References

Ahn,J.B., Chung,W.B., et al. (2011) DNA methylation predicts recurrence from resected stage III proximal colon cancer. *Cancer*, 117, 1847-1854.

Amoreira,C., Hindermann,W., et al. (2003) An improved version of the DNA Methylation database (MethDB). *Nucleic Acids Res.*, 31, 75-77.

Bachman,M., Uribe-Lewis,S., et al. (2014) 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nature chemistry*.

Bhasin,M., Zhang,H., et al. (2005) Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Lett.*, 579, 4302-4308.

Bishop,C. (2007) Pattern Recognition and Machine Learning (Information Science and Statistics), 1st edn. 2006. corr. 2nd printing edn. *Springer, New York*.

Bock,C., Paulsen,M., et al. (2006) CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet*, 2, e26.

Bock,C., Halachev,K., et al. (2009) EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi) genomic data. *Genome Biol.*, 10, 1.

Bock,C., Tomazou,E.M., et al. (2010) Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat.Biotechnol.*, 28, 1106-1114.

Bock,C., Walter,J., et al. (2007) CpG island mapping by epigenome prediction. *PLoS Comput Biol*, 3, e110.

Booth,M.J., Branco,M.R., et al. (2012) Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science*, 336, 934-937.

Carson,M.B., Langlois,R., et al. (2008) Mining knowledge for the methylation status of CpG islands using alternating decision trees. , 3787-3790.

Caruana,R. and Niculescu-Mizil,A. (2006) An empirical comparison of supervised learning algorithms. , 161-168.

Chadwick,L.H. (2012) The NIH roadmap epigenomics program data resource.

Chatterjee,A., Rodger,E.J., et al. (2012) Technical considerations for reduced representation bisulfite sequencing with multiplexed libraries. *BioMed Research International*, 2012.

Couldrey,C., Brauning,R., et al. (2014) Genome-wide DNA methylation patterns and transcription analysis in sheep muscle. *PloS one*, 9, e101853.

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

Eckhardt,F., Lewin,J., et al. (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat.Genet.*, 38, 1378-1385.

Elango,N. and Yi,S.V. (2008) DNA methylation and structural and functional bimodality of vertebrate promoters. *Mol.Biol.Evol.*, 25, 1602-1608.

Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*, 9, 215-216.

Fan,S., Huang,K., et al. (2016) Predicting CpG methylation levels by integrating Infinium HumanMethylation450 BeadChip array data. *Genomics*, 107, 132-137.

Fan,S., Zhang,M.Q., et al. (2008) Histone methylation marks play important roles in predicting the methylation status of CpG islands. *Biochem.Biophys.Res.Commun.*, 374, 559-564.

Fan,S., Zou,J., et al. (2010) Predicted methylation landscape of all CpG islands on the human genome. *Chinese Science Bulletin*, 55, 2353-2358.

Fang,F., Fan,S., et al. (2006) Predicting methylation status of CpG islands in the human brain. *Bioinformatics*, 22, 2204-2209.

Feltus,F.A., Lee,E.K., et al. (2006) DNA motifs associated with aberrant CpG island methylation. *Genomics*, 87, 572-579.

Feltus,F.A., Lee,E.K., et al. (2003) Predicting aberrant CpG island methylation. *Proc.Natl.Acad.Sci.U.S.A.*, 100, 12253-12258.

Gaidatzis,D., Burger,L., et al. (2014) DNA sequence explains seemingly disordered methylation levels in partially methylated domains of Mammalian genomes. *PLoS Genet*, 10, e1004143.

Guo,W., Fiziev,P., et al. (2013) BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC Genomics*, 14, 774.

Guo,H., Zhu,P., et al. (2013) Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.*, 23, 2126-2135.

Hackett,J.A., Sengupta,R., et al. (2013) Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. *Science*, 339, 448-452.

Hahn,M.A., Szabó,P.E., et al. (2014) 5-Hydroxymethylcytosine: a stable or transient DNA modification? *Genomics*, 104, 314-323.

Huang,D.W., Sherman,B.T., et al. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4, 44-57.

Kantlehner,M., Kirchner,R., et al. (2011) A high-throughput DNA methylation analysis of a single cell. *Nucleic Acids Res.*, 39, e44.

Kim,S., Li,M., et al. (2008) Predicting DNA methylation susceptibility using CpG flanking sequences. , 13, 315-326.

Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27, 1571-1572.

Kundaje,A., Meuleman,W., et al. (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, 518, 317-330.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9, 357-359.

Lister,R., Pelizzola,M., et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462, 315-322.

Lu,L., Lin,K., et al. (2010) Predicting DNA methylation status using word composition. *Journal of Biomedical Science and Engineering*, 3, 672.

Ma,B., Wilker,E.H., et al. (2014) Predicting DNA methylation level across human tissues. *Nucleic Acids Res.*, 42, 3515-3528.

Marina,R.J., Sturgill,D., et al. (2015) TET-catalyzed oxidation of intragenic 5-methylcytosine regulates CTCF-dependent alternative splicing. *EMBO J.*, e201593235.

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

Mellen,M., Ayata,P., et al. (2012) MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell*, 151, 1417-1430.

Pan,S.J. and Yang,Q. (2010) A survey on transfer learning. *IEEE Trans.Knowled.Data Eng.*, 22, 1345-1359.

Plongthongkum,N., Diep,D.H., et al. (2014) Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nature Reviews Genetics*, 15, 647-661.

Previti,C., Harari,O., et al. (2009) Profile analysis and prediction of tissue-specific CpG island methylation classes. *BMC Bioinformatics*, 10, 1.

Qu,J., Zhou,M., et al. (2013) MLML: consistent simultaneous estimates of DNA methylation and hydroxymethylation. *Bioinformatics*, 29, 2645-2646.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 841-842.

Rollins,R.A., Haghighi,F., et al. (2006) Large-scale structure of genomic methylation patterns. *Genome Res.*, 16, 157-163.

Shen,L. and Zhang,Y. (2013) 5-Hydroxymethylcytosine: generation, fate, and genomic distribution. *Curr.Opin.Cell Biol.*, 25, 289-296.

Smith,A.D., Chung,W.Y., et al. (2009) Updates to the RMAP short-read mapping software. *Bioinformatics*, 25, 2841-2842.

Speir,M.L., Zweig,A.S., et al. (2016) The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.*, 44, D717-25.

Stadler,M.B., Murr,R., et al. (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*.

Teif,V.B., Beshnova,D.A., et al. (2014) Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development. *Genome Res.*, 24, 1285-1295.

The Gene Ontology Consortium. (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.*, 45, D331-D338.

Wang,T., Pan,Q., et al. (2012) Genome-wide DNA hydroxymethylation changes are associated with neurodevelopmental genes in the developing human cerebellum. *Hum.Mol.Genet.*, 21, 5500-5510.

Wang,Y., Liu,T., et al. (2016) Predicting DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep Networks. *Scientific Reports*, 6, 19598.

Whyte,W.A., Orlando,D.A., et al. (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153, 307-319.

Xi,Y. and Li,W. (2009) BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, 10, 232.

Yamada,Y., Watanabe,H., et al. (2004) A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. *Genome Res.*, 14, 247-266.

Yan,H., Zhang,D., et al. (2015) Chromatin modifications and genomic contexts linked to dynamic DNA methylation patterns across human cell types. *Scientific reports*, 5.

Yu,M., Hon,G.C., et al. (2012) Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell*, 149, 1368-1380.

Zeng,J., Konopka,G., et al. (2012a) Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. *The American Journal of Human Genetics*, 91, 455-465.

Zeng,J., Konopka,G., et al. (2012b) Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. *The American Journal of Human Genetics*, 91, 455-465.

Zhang,W., Zheng,H., et al. (2011) Nucleosome positioning plays an important role in predicting the methylation status of CpG islands. , 3, 1580-1583.

Zhang,W., Spector,T.D., et al. (2015) Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biol.*, 16, 14.

Zhang,Y., Liu,T., et al. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9, R137.

Zheng,H., Wu,H., et al. (2013) CpGIMethPred: computational model for predicting methylation status of CpG islands in human genome. *BMC medical genomics*, 6, S13.

Zhou,X., Li,Z., et al. (2012) Prediction of methylation CpGs and their methylation degrees in human DNA sequences. *Comput.Biol.Med.*, 42, 408-413.