# Accurate prediction of single-cell DNA methylation states using deep learning

Christof Angermueller[1,*], Heather J. Lee[2,3], Wolf Reik[2,3,] Oliver Stegle[1,#]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK
[2] Epigenetics Programme, Babraham Institute, Cambridge, UK
[3] Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

* | These authors contributed equally
# | Corresponding author (oliver.stegle@ebi.ac.uk)

**Recent technological advances have enabled assaying DNA methylation in single cells. Current protocols are limited by incomplete CpG coverage and hence methods to predict missing methylation states are critical to enable genome-wide analyses. We here report DeepCpG, a computational approach based on deep neural networks to predict DNA methylation states from DNA sequence and incomplete methylation profiles in single cells. We validate DeepCpG on mouse embryonic stem cells, where we report substantially more accurate predictions than previous methods. Additionally, we show that DeepCpG provides new insights for interpreting the sources of epigenetic diversity. Our model can be used to estimate the effect of single nucleotide changes and we uncover sequence motifs that are associated with DNA methylation level and epigenetic heterogeneity.**

## Introduction

DNA methylation is one of the most extensively studied epigenetic marks, and is known to be implicated in a wide range of biological processes, including chromosome instability, X-chromosome inactivation, cell differentiation, cancer progression and gene regulation[1–4].

Well-established protocols exist for quantifying average DNA methylation levels in populations of cells. Recent technological advances have enabled profiling DNA methylation at single-cell resolution, either using genome-wide bisulfite sequencing (scBS-seq[5]) or reduced representation protocols (scRRBS[6,7]). These protocols have already provided unprecedented insights into the regulation and the dynamics of DNA methylation in single cells[6,8], and have uncovered new linkages between epigenetic and transcriptional heterogeneity [9–11].

Because of the small amounts of genomic DNA starting material per cell, single-cell methylation analyses are intrinsically limited by moderate CpG coverage (**Fig. 1a,** 20-40% for current protocols[5]). Consequently, a first critical step is to predict missing methylation states to enable genome-wide analyses. While methods exist for predicting average DNA methylation profiles in cell populations[12–16], these approaches do not account for cell-to-cell variability, which is critical for studying epigenetic diversity. Additionally, existing approaches require *a priori* defined features and genome annotations, which are typically limited to a narrow set of cell types and conditions.

Here we report DeepCpG, a computational method based on deep neural networks[17–19] for predicting single-cell methylation states and for modelling the sources of DNA methylation heterogeneity. DeepCpG leverages associations between DNA sequence patterns and methylation states as well as between neighbouring CpG sites, both within individual cells and across cell populations. Unlike previous methods[12,13,15,20–23], our approach does not separate the extraction of DNA sequence features and model training. Instead, DeepCpG uses a modular architecture to learn predictive sequence patterns in a data-driven manner. We evaluate DeepCpG on mouse embryonic stem cells profiled using scBS-Seq, finding that our model yields substantially more accurate predictions of methylation states than previous approaches. Additionally, we show that by interpreting trained model parameters, DeepCpG uncovers both previously known and novel sequence motifs that are associated with methylation changes and epigenetic heterogeneity between cells.

# Results

DeepCpG is trained to predict binary CpG methylation states from local DNA sequence windows and observed neighbouring methylation states (**Fig. 1a**). A major feature of the model is its modular architecture, comprising of a CpG module to account for correlations between CpG sites within and across cells, a DNA module to detect informative sequence patterns, and a fusion module that integrates the evidence from the CpG and DNA module to predict methylation states at target CpG sites (**Fig. 1b**).

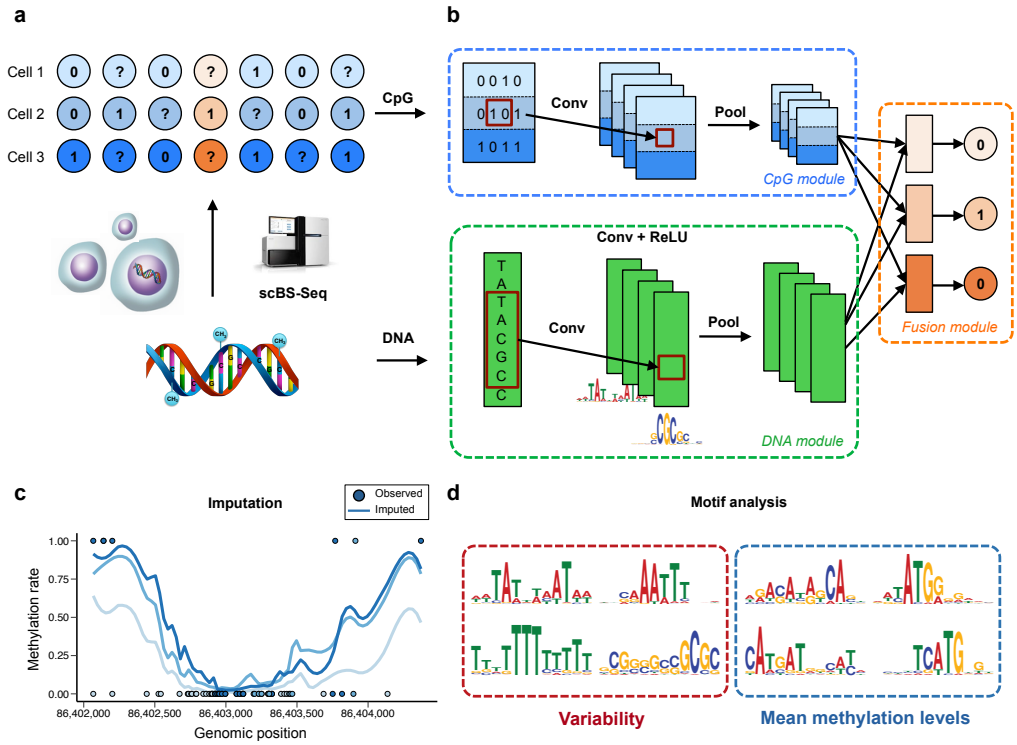



**Figure 1 | DeepCpG model training and applications. (a)** Sparse single-cell CpG profiles, for example as obtained by scBS-Seq[5]. Methylated CpG sites are denoted by ones, unmethylated CpG sites by zeros, and CpG sites with missing methylation state by question marks. **(b)** Modular architecture of DeepCpG. The DNA module uses convolutional filters and pooling operations to identify predictive sequence motifs. The CpG module identifies patters in the CpG neighbourhood across multiple cells, using cell-specific convolution and pooling operations (rows in **b**). The fusion module models interactions between higher-level features derived form the DNA- and CpG module to predict methylation states in all cells. **(c,d)** The trained DeepCpG model is used for downstream analyses, including the genome-wide imputation of missing CpG sites **(c)** and the discovery of DNA sequence motifs that are associated with DNA methylation level and cell-to-cell heterogeneity **(d)**.

Briefly, the DNA and CpG module are built using convolutional neural networks, which have been successfully applied in different domains[24–27], including bioinformatics[28–32]. DNA sequences in windows centred on target CpG sites form the input to the DNA module, which uses convolutional filters to scan for sequence motifs, analogous to conventional position weight matrices[33,34] (**Online methods**). Similarly, the CpG module has a convolutional architecture to detect predictive patterns in neighbouring CpG sites based on methylation states and their distances to the target site. Finally, the fusion module learns cell-specific interactions between output features of the DNA- and CpG module using a multi-task architecture, to predict the methylation state of target sites in all cells. The trained DeepCpG model is then used for different downstream analyses, including i) to impute low-coverage methylation profiles for sets of cells and ii) to discover DNA sequence motifs that are associated with DNA methylation states and cell-to-cell heterogeneity (**Fig. 1c**).

## DeepCpG accurately predicts single-cell methylation states

We applied DeepCpG to 32 mouse embryonic stem cells profiled using scBS-seq[5] (12 2i-cultured cells, 20 serum-cultured cells; average CpG coverage 17.7%; **Supplementary Fig. 1**). We compared the prediction accuracy of DeepCpG to a baseline model that estimates average methylation levels in consecutive 3 kb regions (WinAvg)[5], as well as a prediction model based on random forests (RF)[35] . Methods were trained, selected, and tested on distinct chromosomes via holdout validation (**Online methods**), and we quantified prediction accuracies using the area under the receiver operating characteristics curve (AUC). DeepCpG yielded more accurate predictions than alternative methods (**Fig. 1a**), which was consistent for all individual cells (**Fig. 1a,b**) and when considering alternative metrics such as precision-recall (**Supplementary Fig. 2 and Supplementary Table 1**). The average methylation rate was highly correlated with cell-specific prediction accuracies (R=0.91, $P = 1.03 \times 10^{-12}$ ), consistently within 2i and serum cells (**Supplementary Fig. 3).** This relationship explains the different prediction

accuracies between 2i-cultured cells (AUC 0.80; 31.1% average methylation) and serum-grown cells (AUC 0.87; 63.9% average methylation; **Fig. 2a,b)**.

To explore to which extent DeepCpG can impute methylation in domains without observed methylation states, we evaluated the DNA module separately (DeepCpG Seq). Notably, although this reduced model does not account for local CpG correlations, its performance was similar to conventional window averaging (AUC 0.78 WinAvg, AUC 0.77 DeeCpG Seq), and it was considerably more accurate than a random forest model trained on k-mer frequencies derived from the same sequence windows (RF Seq, AUC 0.72). Consistent with these results, we also found that the relative gains of the full DeepCpG model compared to other methods were largest in regions with low CpG coverage (**Fig. 1c**). This suggests that DeepCpG can be used to extrapolate to larger uncovered genomic regions, for example as obtained when using reduced representation sequencing data[7].

In line with previous findings[12,13], we also observed a relationship between GC content and prediction accuracy, where GC-rich regions tended to be associated with increased accuracy (**Supplementary Fig. 4**). Notably, DeepCpG performed markedly better than alternative methods in GC-poor genomic contexts, including non-CGI promoters, enhancer regions, and histone modification marks (H3K4me1, H3K27ac) —contexts that are known to be associated with higher epigenetic heterogeneity between cells[5].

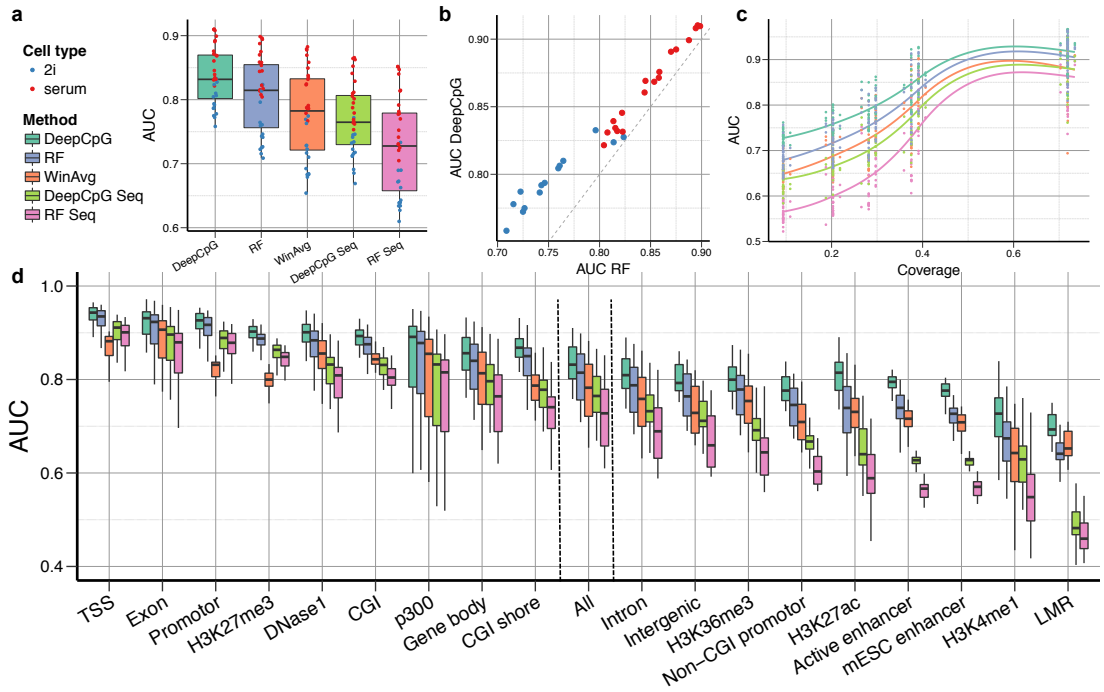



**Figure 2 | DeepCpG accurately predicts single-cell CpG methylation states. (a)** Genome-wide prediction accuracy for imputing CpG sites in mouse embryonic stem cells (20 serum-cultured cells, red; 12 2i-cultured cells, blue). Shown is the prediction accuracy for individual cells and for alternative methods, quantified using the area under the receiver-operating characteristic curve (AUC) based on holdout validation. Considered were DeepCpG, window averaging in consecutive 3 kb regions (WinAvg), a random forest model (RF), and the corresponding models when trained with DNA sequence windows only (DeepCpG Seq, RF Seq). **(b)** Prediction accuracy for individual cells, comparing DeepCpG and RF. **(c,d)** AUC in genomic regions stratified by the fraction of cells with sequence coverage **(c)** and when considering alternative sequence contexts **(d).**

A major advantage of DeepCpG compared to previous methods is its convolutional architecture, which allows for discovering predictive motifs in larger DNA sequence contexts, as well as for capturing complex methylation patterns in neighbouring CpG sites. Indeed, the accuracy of DeepCpG was markedly reduced when considering smaller sequence windows (AUC 0.77 vs. 0.72 for 501 bp and 101 bp contexts; **Supplementary Fig. 5),** or when limiting the number of neighbouring CpG sites in the model (**Supplementary Fig. 6**). We also confirmed that models that include methylation states from

other cells were more accurate than equivalent models applied to individual cells (**Supplementary Fig. 7)**.

Finally, we compared DeepCpG to a random forest model based on rich DNA annotations, including genomic contexts, and tissue-specific elements such as DNase1 hypersensitivity sites, histone modification marks, and transcription factor binding sites[12] (**Online methods**). Even though DeepCpG does not use external annotations, the model clearly outperformed such an approach (**Supplementary Fig. 8**). This strongly suggests that DeepCpG learns higher-level annotations from the DNA sequence. This ability is particularly important for analysing single-cell datasets, where individual cells may be from different cell types and cell states, making it difficult to choose appropriate annotations.

## Analysis the effect of DNA sequence features on DNA methylation

In addition to imputing missing methylation states, DeepCpG can be used to discover methylation-associated motifs, and investigate effects of DNA mutations and neighbouring CpG sites on CpG methylation.

To explore this, we analysed the filters of the first convolutional layer of the DNA module, which recognize DNA sequence motifs similarly to conventional position weight matrices (**Fig. 3a**). 18 out of 98 discovered motifs could be matched to known motifs in CIS-BP[36] and UniPROPE[37] (FDR<0.01). We considered two complementary metrics to score the importance of individual motifs: i) their occurrence frequency in DNA sequence windows (activity), and ii) the estimated association with single-cell methylation states (**Online Methods**; see **Supplementary Table 2, Supplementary Fig. 9, 10**).

A principal component analysis on the motif activity revealed that motifs with similar nucleotide composition tend to co-occur in the same sequence windows, where two major motif clusters were associated with increased or decreased methylation levels (**Fig. 3a, Supplementary Fig. 10**). Consistent with previous findings[16,38,39], we observed that motifs associated with

decreased methylation were CG rich and most active in CG rich promoter regions, transcription start sites, as well as in contexts with active promoter marks such as H3K4me3 and p300 sites (**Supplementary Fig. 9**). These include known transcription factors and regulators of cell differentiation such as Max[40], E2f[41], and members of the Sp/KLF family[42]. Conversely, motifs associated with increased methylation levels tended to be AT rich and most active in CG poor genomic contexts (**Supplementary Figure 9**). Examples include the serum response factor (Srf)[43,44], Tlx2[45], and Gata5[46], with known implications for cell differentiation.

DeepCpG further allows estimating the effect of single nucleotide mutations on CpG methylation. For this purpose, we developed an approach that exploits gradient information[47], a strategy that is computationally more efficient than previous approaches[29,30,32] (**Online methods**). Sequence changes in the direct vicinity of the target site had the largest predicted effects (**Fig. 3b**). Mutations in CG dense regions such as CpG islands or promoters tended to have smaller effects, suggesting that DNA methylation in these regions is more robust to single base-pair mutations. Globally, we observed that the predicted effect of single-nucleotide changes was significantly anti-correlated with DNA sequence conservation ($P < 1.0\times10^{-15}$, **Supplementary Fig. 11**), providing evidence that the model-based effect size estimates of DeepCpG capture genuine effects.

We used the analogous approach to quantify the effect of epimutations on neighbouring CpG sites, again finding a clear relationship between distance and the mutational effect (**Fig. 3c**). Similar to DNA mutations, we found that alterations in CG dense regions had smaller effects, whereas CpG island shores and CG poor regions had larger mutational effects close to the target site, which is consistent with previous findings[12,48].

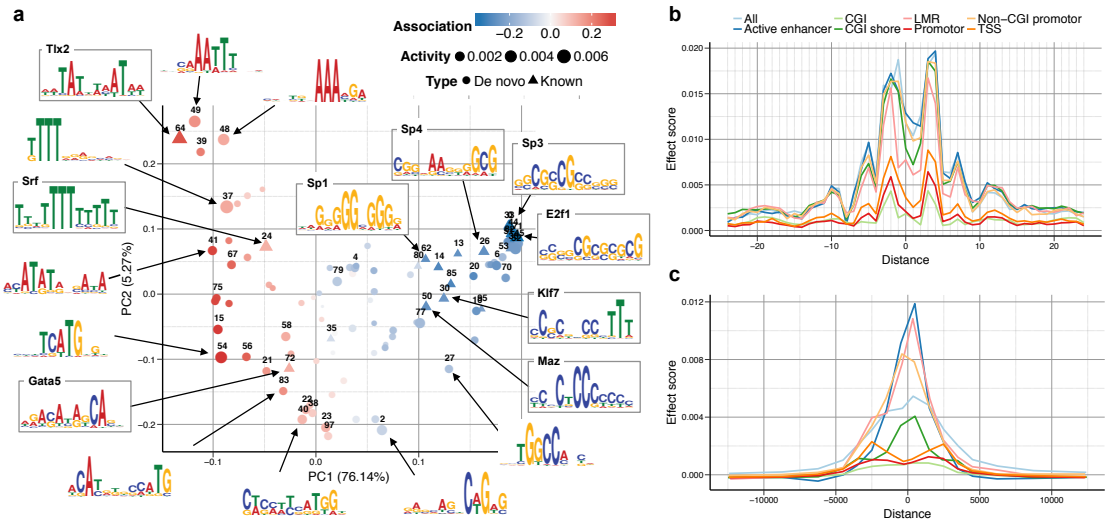



**Figure 3: Analysis of DNA sequence features and impact of genetic and epigenetic mutations.** (**a**) Clustering of 98 discovered motifs. Shown are the two first principal components of the motif occurrence frequencies in genome-wide sequence windows. Triangles denote annotated motifs (matched to CIS-BP[36] and UniPROPE[37]; FDR<0.01); circles denote de novo motifs. The symbol size indicates the average occurrence frequency (activity); the estimated motif effect on methylation level is shown in colour. Sequence logos are shown for representative motifs with larger effects, including 9 annotated motifs. PC1 separates GC rich motifs (PC 1 high, decreased methylation) and AT rich motifs (PC 1 low, increased methylation). (**b,c**) Average effect of DNA sequence mutations (**b**) and epimutations (**c**) on methylation levels, as a function of distances to the CpG site and for different sequence contexts.

## Discovery of DNA sequence motifs that are associated with epigenetic variability

Uniquely, single-cell methylation studies allow quantifying the epigenetic diversity to study the sources of this variation.

To discern motifs that affect variability between cells from those that affect overall methylation level, we trained a second neural network, reusing the learnt motifs from the DNA module of DeepCpG, however using a multi-task objective to jointly predict the variability across cells and the average methylation level for each CpG site (**Online methods**).

Notably, this model could predict both global changes in DNA methylation level (Pearson's R=0.77; MAD=0.11; **Supplementary Fig. 12**; hold-out validation), and cell-to-cell variability (Pearson's R=0.48; MAD=0.03; **Fig. 4c,** or Kendall's R=0.3, **Supplementary Fig. 13**).

To identify motifs that drive epigenetic variability between cells, we estimated the effect of individual motifs on both cell-to-cell variability and methylation levels. Although there is an intrinsic mean-variance relationship of methylation levels (**Supplementary Fig. 14**), we identified 12 motifs that were primarily associated with differences in cell-to-cell variability (**Fig. 4a**). These motifs were most active in CG-poor and active enhancer regions — sequence contexts with increased epigenetic variability between cells[5]. Six AT-rich motifs were associated with increased variability, including Tlx2 and the serum response factor Srf, both known transcription factors that play a role in cell-differentiation and gene expression regulation[44,45]. Notably, variance-increasing motifs were more frequent in non-conserved regions such as active enhancers, in contrast to variance- decreasing motifs, which were enriched in evolutionary conserved regions such as gene promoters (**Fig. 4b**, **Supplementary Fig. 15**).

As an indirect validation of model predictions for heterogeneous sites, we overlaid the predicted cell-to-cell variability with methylome-transcriptome linkages obtained using parallel single-cell methylation and transcriptome sequencing in the same cell type[9]. The rationale behind this approach is that regions with increased epigenetic heterogeneity are more likely to harbour associations between transcriptional and epigenetic diversity. Indeed, we observed that the predicted epigenetic heterogeneity was correlated to the level of epigenome-transcriptome coupling (**Supplementary Fig. 16**). This suggests that DNA sequence motifs discovered using DeepCpG explain a genuine component of the epigenetic heterogeneity between cells.

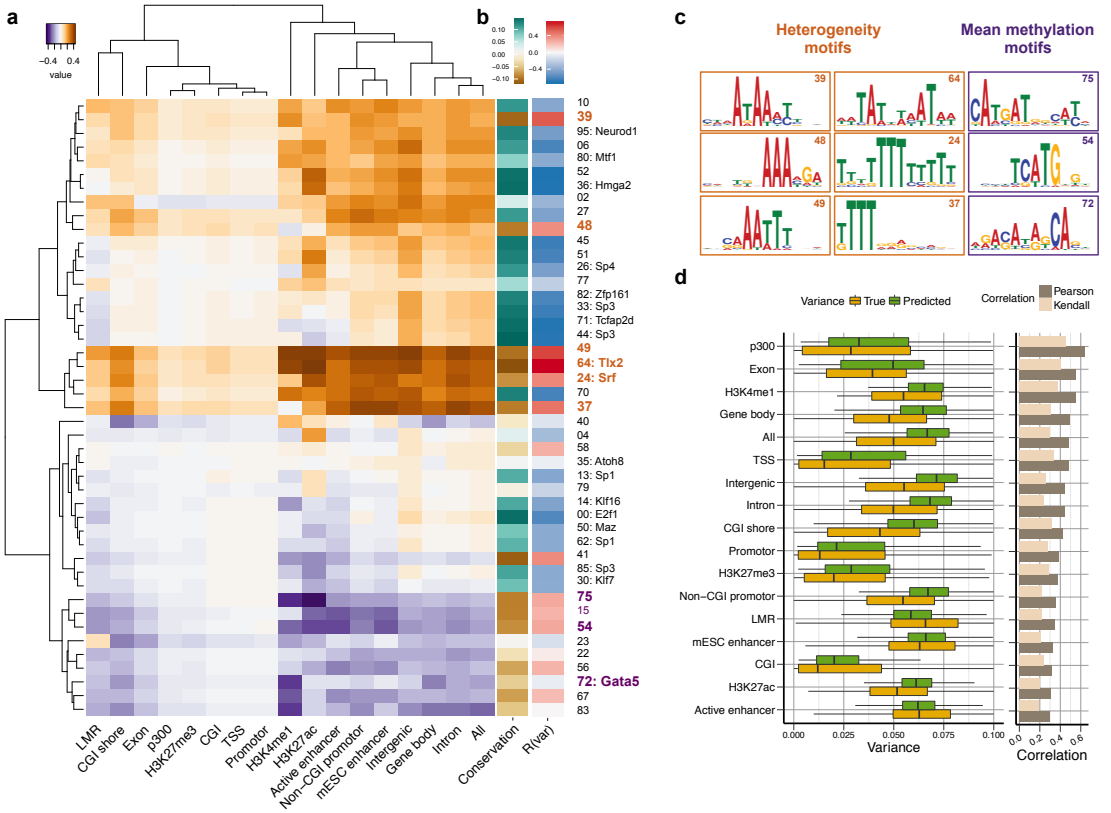



**Figure 4: Prediction of epigenetic heterogeneity from local DNA sequence.** (**a**) Difference of motif effect on cell-to-cell variability and methylation levels for different genomic contexts. Motifs associated with cell-to-cell variability are highlighted in brown; motifs that were primarily associated with changes in methylation level are shown in purple. **(b)** Genome-wide correlation coefficients between motif activity and DNA sequence conservation (left), as well as cell-to-cell variability (right). (**c**) Sequence logos for selected motifs identified in **a**. (**d**) Boxplots of the predicted and the observed cell-to-cell variability in different genomic contexts (left) alongside Pearson's and Kendall's correlation coefficients between predicted and observed heterogeneity within contexts (right).

## Discussion

Here we reported DeepCpG, a computational approach based on convolutional neural networks for modelling low-coverage single-cell methylation data. In applications to mouse embryonic stem cells, we have shown that DeepCpG accurately predicts missing methylation states, and detects sequence motifs that are associated with changes in methylation level and epigenetic heterogeneity.

We have demonstrated that our model enables accurate imputation of missing methylation states, thereby facilitating genome-wide downstream analyses. DeepCpG offers most important advantages in shallow sequenced cells as well as in sparsely covered sequence contexts with increased epigenetic heterogeneity between cells. More accurate imputation methods may also help to reduce the required sequencing depth in single-cell bisulfite sequencing studies, thereby enabling assaying larger numbers of cells at lower cost.

We have further shown that DeepCpG enables the interpretation of DNA sequence and CpG methylation features that are associated with changes in DNA methylation states. We have identified known and de novo sequence motifs that are predictive for DNA methylation level or methylation heterogeneity. Models such as DeepCpG allow discerning pure epigenetic effects from variation that reflect DNA sequence changes. Although we have not considered this in our work, it would also be possible to consider the model residuals to study epigenetic variation that is unlinked to DNA sequence effects.

Finally, we have used additional data obtained from parallel methylation-transcriptome sequencing protocols[9] to annotate regions with increased epigenetic diversity. An important area of future work will be to integrate multiple datasets from parallel-profiling methods[9,10], which are now becoming increasingly available for different molecular layers.

## Methods

Methods and any associated references are available in the online version of the paper.

**Accession codes** The scBS-Seq data from 20 serum and 12 2i ES-cells have previously been described in Smallwood *et al.*[9] and are available under the Gene Expression Omnibus (GEO) accession number GSE56879.

**Availability of code** An implementation of DeepCpG is available at https://github.com/cangermueller/deepcpg.

**Acknowledgements**

We are grateful to Yarin Gal for valuable discussions about quantifying prediction uncertainty using dropout. We are grateful to Leopold Parts for commenting on the manuscript, and Felix Krueger for pre-processing the data. O.S. is supported by the European Molecular Biology Laboratory (EMBL), the Wellcome Trust and the European Union.

**Author contributions**

C.A. and O.S. devised and developed the model. C.A. implemented and evaluated the model. C.A., H.J.L., W.R. and O.S. interpreted the results. C.A. and O.S. wrote the paper.

**Author information** The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to oliver.stegle@ebi.ac.uk

# Online Methods

## DeepCpG model

DeepCpG consists of a DNA module to extract features from the DNA sequence, a CpG module to extract features from the CpG neighbourhood of all cells, and a multi-task fusion model that integrates the evidence from both modules to predict the methylation state of target CpG sites for multiple cells.

### DNA module

The DNA module is built using a convolutional neural network (CNN) with one convolutional, pooling, and fully connected hidden layer. CNNs are designed to extract features from high-dimensional inputs while keeping the number of model parameters tractable by applying a series of convolutional and pooling operations. Unless stated otherwise, the DNA module takes as input a *501 bp* long DNA sequence cantered on a target CpG site $n$, which was represented as a binary matrix $s_n$ by one-hot encoding the *D = 4* nucleotides as binary vectors A=[0, 0, 0, 1], G=[0, 0, 1, 0], T=[0, 1, 0, 0], C=[1, 0, 0, 0]. $s_n$ is first transformed by a 1d-convolutional layer, which computes the *activations* $a_{nfi}$ of multiple convolutional filters $f$ at every position $i$ as follows:

$$a_{nfi} = \mathrm{ReLU}\left(\sum_{l=1}^{L}\sum_{d=1}^{D} w_{fld} s_{n,i+l,d}\right) \qquad (1)$$

Here, $w_f$ are the parameters or *weights* of convolutional filter $f$ of length $L$. This weight vector can be interpreted similarly to a position weight matrix (PWM), which is matched against the input sequence $s_n$ at each position $i$ to recognize a certain motif. The $\mathrm{ReLU}(x) = \max(0, x)$ activation function sets

negative values to zero, such that $a_{nfi}$ indicates to which extend a motif represented by $w_f$ occurs as position $i$.

A pooling layer is used to summarize the activations of *P* adjacent neurons by their maximum value

$$p_{nfi} = \max_{|k|<P/2}(a_{nf,i+k}).$$

Non-overlapping pooling is applied with step size $P$ to decrease the dimension of the input sequence and hence the number of model parameters. The pooling layer is followed by one fully connected hidden layer with ReLU activation function.

### CpG module

Akin to the DNA module, the CpG module has of one convolutional, pooling, and fully connected hidden layer. The methylation state and distance of observed neighbouring CpG sites are inputs to a 2d-convolutional layer. Importantly, this layer convolves each cell separately with the same convolutional filters to unlink the number of model parameters from the number of cells, which can be large. Specifically, for each target site $n$, the binary methylation state and distance of *K = 25* CpG neighbours to the left and right of *T* cells were represented as a *T x 2K x D* tensor $c_n$, by storing methylation states at dimension *d = 1*, and distances at dimension *d = 2*. Distances were transformed to relative ranges $[0; 1]$ by dividing by the maximum distance. A 2d-convolution layer convolves the CpG neighbourhood of cells *t* independently at every position $i$ by using filters $w_f$ of dimension *1 x L x D* and length $L$:

$$a_{nfti} = \mathrm{ReLU}\left(\sum_{l=1}^{L}\sum_{d=1}^{D} w_{fld}c_{t,i+l,d}\right).$$

Non-overlapping pooling of $P$ neurons is performed independently on the activations $a_{nfti}$ of each cell by using a max-pooling with step size *1 x P*:

$$\mathrm{p_{nfti}} = \max_{|k|<P/2}\left(a_{n,f,t,i+k}\right)$$

The pooling layer is followed by one fully connected hidden layer with ReLU activation function.

### Fusion module

To model cell-specific interactions between extracted DNA sequence and CpG neighbourhood features, the fusion module has one hidden layer with ReLU activation function, which is connected to all neurons of the last layer of the DNA and CpG module. Each hidden layer is connect to one output neuron with sigmoid activation function, which predicts the methylation rate $\hat{y}_{nt} \in [0; 1]$ of CpG site $n$ and cell $t$:

$$\hat{y}_{nt}(\mathrm{x})=\mathrm{sigmoid}(\mathrm{x})=\left(\frac{1}{1+\mathrm{e}^{-\mathrm{x}}}\right)$$

### Model training

Model parameters were learnt on the training set by minimizing the following loss function:

$$L(w) = \mathrm{NLL}_w(\hat{y}, y) + \lambda_1\|w\|_1 + \lambda_2\|w\|_2$$

Here, the regularization hyper-parameters $\lambda_1$ and $\lambda_2$ penalize large model weights quantified by the L1 and L2 norm, respectively, and *NLL* is the negative log-likelihood, which measures the fit between the predicted methylation rates $\hat{y}_{nt}$ and true methylation states $y_{nt} \in \{0,1\}$:

$$NLL_w(\hat{y}, y) = -\sum_{n=1}^{N}\sum_{t=1}^{T} o_{nt}\left[y_{nt} log(\hat{y}_{nt}) + (1-\hat{y}_{nt})\log\left(1-\hat{y}_{nt}\right)\right]$$

$o_{nt} = 1$ if the true methylation state $y_{nt}$ is observed and zero otherwise. Dropout[49] with different dropout rates for the sequence, CpG, and fusion module was used for additional regularization. Model parameters were initialized randomly following the approach in Golorot *et al*[50]. The loss function was optimized by mini-batch stochastic gradient descent with Adam[51] learning rate adaptation. The global learning rate was multiplied by 0.5 if the validation loss did not improve over four epochs. Learning was terminated if the validation loss did not improve over five epochs (early stopping). The DNA and CpG module was pre-trained independently to predict methylation from the DNA sequence (DeepCpG Seq) or the CpG neighbourhood (DeepCpG CpG) alone, followed by further joint training with the fusion module afterwards. Training time on a single NVIDIA Tesla K20 GPU was approximately 15 hours for the DNA module, 12 hours for the CpG modules, and 4 hours for the fusion module. Hyper-parameters were optimized on the validation set by random sampling[52], and are summarized in **Supplementary Table 6**. DeepCpG was implemented in Python using Theano[53] 0.7.0 and Keras[54] 0.2.0.

# Prediction performance evaluation

### Data pre-processing

We evaluated DeepCpG using a dataset of 20 serum and 12 2i mouse embryonic stem cells that have been profiled using scBS-Seq[5]. Data were pre-processed as described in Smallwood *et al*[9], and mapped to the GRCm38 mouse genome. Binary CpG methylation states were obtained for CpG sites with mapped reads, by defining sites with more methylated than un-methylated read counts as methylated, and un-methylated otherwise.

The prediction performance of DeepCpG was compared with window averaging (WinAvg), a random forest model with comparable features than DeepCpG (RF), and a random forest classifier with high-level sequence features as described in Zhang *et al*[20] (RF Zhang).

For all prediction experiments and evaluations, we used chromosomes 1, 3, 5 and 10 (4,509,888 CpG sites) as training set, chromosomes 7, 9 and 13 (2,875,909 CpG sites) as validation set, and chromosomes 2, 4, 6 and 12 (4,359,411 CpG sites) as test set.

### Window averaging (WinAvg)

For window averaging, the methylation rate $\hat{y}_{nt}$ of CpG site *n* and cell *t* was estimated as the mean of all observed CpG neighbours $y_{n+k,t}$ in a window of length *W = 3001 bp* cantered on site *n*:

$$\hat{y}_{nt} = \text{mean}_{|k|<\frac{W}{2}, k\neq 0}(y_{n+k,t})$$

$\hat{y}_{nt}$ was set to the mean genome wide methylation rate of cell *t* if no CpG neighbours were present in the window.

### Random forest models (RF, RF Zhang)

Features of the *RF* model are i) the methylation state and distance of 25 CpG neighbours to the left and right of the target site (100 features), and ii) 4-mer frequencies from the *501 bp genomic* sequence cantered on the target site (256 features).

The set of features for the *RF Zhang* model are summarized in **Supplementary Table 5** and include a) the methylation state and distance of 2 CpG neighbours to the left and right of the target site (8 features), b) annotated genomic contexts (23 features), c) transcription factor binding sites (24 features), d) histone modification marks (28 features), and e) DNaseI hypersensitivity sites (1 feature). Features were downloaded from the ChipBase database and UCSC Genome Browser for the GRCm37 mouse genome, and mapped to the GRCm38 mouse genome using the liftOver tool from the UCSC Genome Browser.

Random forest classifiers were trained independently for individual cells, and the number of trees and tree depth optimized on the validation set by random sampling. The implementation is based on the RandomForestClassifer class of the scikit-learn v0.17 Python package.

# Motif analysis

In the following, the filters of the first convolutional layer of the DNA module will be denoted by the motif that they recognize in the input sequence.

### Visualization, motif comparison, GO analysis

Filters of the convolutional layer of the DNA module were visualized by aligning sequence fragments that maximally activated them. Specifically, the activations of all filters were computed for a set of sequences. For each sequence $s_n$ and filter $f$ of length $L$, sequence window $s_{n,i-L}, \ldots, s_{n,i+L}$ were selected, if the activation $a_{nfi}$ of filter $f$ at position $i$ (**Equ. 1**), was greater than 0.5 of the maximum activation of $f$ over all sequences, i.e.

$a_{nfi} > 0.5 \max_{n,i}(a_{nfi})$ . Selected sequence windows were aligned and visualized as sequence motifs using WebLogo[55] version 3.4.

For motif comparison, filter alignments were matched against the Mus Musculus CIS-BP[36] and UniPROBE[37] database (version 12.12, updated 14 Mar 2016) using Tomtom 4.11.1 from the MEME-Suite[56] with a false discovery rate threshold of 0.1.

For Genome Ontology (GO) enrichment analysis, the web interface of the GOMo tool of MEME-Suite was used.

### Quantification motif importance

The importance of each filter was quantified by its activity (occurrence frequency) and influence on model predictions.

Specifically, the activity of filter $f$ for a set of sequences, e.g. within a certain genomic context, was computed by averaging mean sequence activations $\bar{a}_{nf}$, where $\bar{a}_{nf}$ is the weighted mean over all activations $a_{nfi}$ (**Eq. 1**) using a linear weighting function that gives highest weight to the centre position.

The influence of filter $f$ on the predicted methylation states $\hat{y}_{nt}$ of cell $t$ was computed as the Spearman correlation $r_{ft} = \mathrm{cor}_n(\bar{a}_{nf}, \hat{y}_{nt})$ over CpG sites $i$, and the mean influence $r_f$ over all cells by averaging $r_{ft}$.

### Motif co-occurrence

The co-occurrence of filters (**Fig. 3a**) was quantified by principle component analysis on their mean sequence activations $\bar{a}_{nf}$.

### Conservation analysis

For computing the correlation between filter activities $\bar{a}_{nf}$ and sequence conservation, the Spearman correlation was computed as described before. *PhastCons* [57] conservation scores for the Glire subset (phastCons60wayGlire) downloaded from the UCSC Web Browser were used to quantify sequence conservation.

## Effect of sequence and methylation state changes

We used gradient-based optimization as described in Simonyan *et al.*[47] to quantify the effect of changes in the input sequence $s_n$ on predicted methylation rates $\hat{y}_{nt}(s_n)$. Specifically, let $\hat{y}_n(s_n) = \mathrm{mean}_t(\hat{y}_{nt}(s_n))$ be the mean predicted methylation rate across cells $t$. Then the effect $e^s_{n,i,d}$ of changing nucleotide $d$ at position $i$ was quantified as:

$$e^s_{nid} = \frac{d\,\hat{y}_{nt}(s_n)}{ds_{nid}} * (1 - s_{nid})$$

Here, the first term is the first-order gradient of $\hat{y}_n$ with respect to $s_{nid}$, and the second term sets the effect of wild-type nucleotides ($s_{nid} = 1$) to zero. The overall effect score $e^s_{ni}$ at position $i$ was computed as the maximum absolute effect over all nucleotide changes, i.e. $e^s_{ni} = \max_d |e^s_{nid}|$. For mutation analysis shown in **Supplementary Fig. 11**, $e^s_{nii}$ was correlated with *PhastCons* (phastCons60wayGlire) conservation scores. The overall effect of changes at position $i$ as shown in **Fig. 3b** was computed by the mean effect $e^s_i = \mathrm{mean}_n(e^s_{ni})$ over all sequences.

Analogously, the effect $e^c_{nti}$ of changing the methylation state $c_{nti}$ of cell $t$ at position $i$ was quantified by the first-order gradient:

$$e_{nti}^{c} = \frac{d\,\hat{y}_{nt}(c_n)}{dc_{nti}}$$

## Predicting cell-to-cell variability

For predicting cell-to-cell variability (variance) and mean methylation levels, we trained a second neural network with the same architecture as the DNA module, except for the output layer. Specifically, output neurons were replaced by neurons with sigmoid activation function to predict for single CpG sites both the mean methylation rate $\hat{m}_{ns}$ and cell-to-cell variance $\hat{v}_{ns}$ within a window of size $s \in \{1000, 2000, 3000, 4000, 5000\}$ bp. Multiple window sizes were used to make predictions at different scales and to account for noisy estimates in low-coverage regions. For training the resulting model, parameters were initialized with the corresponding parameters of the DNA module and fine-tuned, except for motif parameters of the convolutional layer. Training objective was

$$L(w) = \mathrm{MSE}_w(\hat{m}, m, \hat{v}, v) + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2,$$

where MSE the is mean squared error between model predictions and training labels:

$$MSE_w(\hat{m}, m, \hat{v}, v) = \sum_{n=1}^{N} \sum_{s=1}^{S} (m_{ns} - \hat{m}_{ns})^2 + (v_{ns} - \hat{v}_{ns})^2$$

$m_{ns}$ is the estimated mean methylation level for a window cantered on target site $n$ of a certain size indexed by $s$:

$$m_{ns} = \frac{1}{T} \sum_{t=1}^{T} m_{nst}$$

Here, $m_{nst}$ denotes the estimated mean methylation rate of cell $t$ computed by averaging the binary methylation state $y_{it}$ of all observed CpG sites $Y_{nst}$ in window $s$:

$$m_{nst} = \frac{1}{|Y_{nst}|} \sum_{i \in Y_{nst}} y_{it}$$

$v_{ns}$ is the estimated cell-to-cell variance

$$v_{ns} = \frac{1}{T} \sum_{t=1}^{T} (m_{nst} - m_{ns})^2$$

2i cells were excluded for estimating $m_{ns}$ and $v_{ns}$ due to their low cell-to-cell variance.

### Identifying motifs associated with cell-to-cell variability

The influence $r_{fs}^{v}$ of filter $f$ on cell-to-cell variability was computed as the Spearman correlation between mean sequence filter activities $\bar{a}_{nf}$ and predicted variances $\hat{v}_{ns}$ over sites $n$:

$$r_{fs}^{v} = \mathrm{cor}_n(\bar{a}_{nf}, \hat{v}_{ns})$$

The influence $r_{fs}^{m}$ on predicted methylation levels $\hat{m}_{ns}$ was computed analogously. The difference $r_{fs}^{d} = r_{fs}^{v} - r_{fs}^{m}$ in influences was used to differentiate between motifs that were associated with either high cell-to-cell variance ($r_{fs}^{d} > 0.2$), or changes in mean methylation levels ($r_{fs}^{d} < -0.2$).

### Function validation of predicted variability

For functional validation, methylation-transcriptome linkages as reported in Angermueller *et al.*[9] were correlated with the predicted cell-to-cell variability. Specifically, let $r_{ij}^{e}$ be the linkage between expression levels of gene $i$ and the mean methylation levels of an adjacent region $j$ (see Angermueller *et al.*[9]).

Then we correlated $r_{ij}^{e}$ with $v_j$, which is the average predicted variability over all CpG sites within context $j$, and FDR adjusted p-values over genes $i$ and contexts $j$.

## Bibliography


1. Robertson, K. D. DNA methylation and human disease. *Nat. Rev. Genet.* **6,** 597–610 (2005).
2. Suzuki, M. M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* **9,** 465–476 (2008).
3. Laird, P. W. Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.* **11,** 191–203 (2010).
4. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13,** 484–492 (2012).
5. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11,** 817–820 (2014).
6. Farlik, M. *et al.* Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *Cell Rep.* **10,** 1386–1397 (2015).
7. Guo, H. *et al.* Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.* **23,** 2126–2135 (2013).
8. Peat, J. R. *et al.* Genome-wide Bisulfite Sequencing in Zygotes Identifies Demethylation Targets and Maps the Contribution of TET3 Oxidation. *Cell Rep.* **9,** 1990–2000 (2014).

9. Angermueller, C. *et al.* Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13,** 229–232 (2016).

10. Hou, Y. *et al.* Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* **26,** 304–319 (2016).

11. Hu, Y. *et al.* Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol.* **17,** (2016).

12. Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T. & Engelhardt, B. E. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biol.* **16,** 14 (2015).

13. Stevens, M. *et al.* Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome Res.* **23,** 1541–1553 (2013).

14. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.* **33,** 364–376 (2015).

15. Liu, Z., Xiao, X., Qiu, W.-R. & Chou, K.-C. iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. *Anal. Biochem.* **474,** 69–77 (2015).

16. Whitaker, J. W., Chen, Z. & Wang, W. Predicting the human epigenome from DNA motifs. *Nat. Methods* **12,** 265–272 (2015).

17. LeCun, Y. *et al.* Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* **1,** 541–551 (1989).

18. Bengio, Y. Learning Deep Architectures for AI. (2008).

19. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015).

20. Bhasin, M., Zhang, H., Reinherz, E. L. & Reche, P. A. Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Lett.* **579,** 4302–4308 (2005).

21. Lu, L. Predicting DNA methylation status using word composition. *J. Biomed. Sci. Eng.* **03,** 672–676 (2010).

22. Zhou, X., Li, Z., Dai, Z. & Zou, X. Prediction of methylation CpGs and their methylation degrees in human DNA sequences. *Comput. Biol. Med.* **42,** 408–413 (2012).

23. Li, Z., Chen, L., Lai, Y., Dai, Z. & Zou, X. The prediction of methylation states in human DNA sequences based on hexanucleotide composition and feature selection. *Anal. Methods* **6,** 1897 (2014).

24. Jarrett, K., Kavukcuoglu, K., Ranzato, M. & LeCun, Y. What is the best multi-stage architecture for object recognition? in *2009 IEEE 12th International Conference on Computer Vision* 2146–2153 (2009).

25. Zhang, X., Zhao, J. & LeCun, Y. Character-level Convolutional Networks for Text Classification. *arXiv* (2015).

26. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *arXiv* (2015).

27. Szegedy, C., Ioffe, S. & Vanhoucke, V. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv* (2016).

28. Denas, O. & Taylor, J. Deep modeling of gene expression regulation in an erythropoiesis model. in *Representation Learning, ICML Workshop* (Citeseer, 2013).

29. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33,** 831–838 (2015).

30. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12,** 931–934 (2015).

31. Xiong, H. Y. *et al.* The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347,** 1254806 (2015).

32. Kelley, D. R., Snoek, J. & Rinn, J. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *bioRxiv* (2015).

33. Stormo, G. D., Schneider, T. D., Gold, L. & Ehrenfeucht, A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res.* **10,** 2997–3011 (1982).

34. Sinha, S. On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics* **22,** e454–e463 (2006).

35. Breiman, L. Random forests. *Mach. Learn.* **45,** 5–32 (2001).

36. Weirauch, M. T. *et al.* Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158,** 1431–1443 (2014).

37. Newburger, D. E. & Bulyk, M. L. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* **37,** D77–D82 (2009).

38. Thomson, J. P. *et al.* CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464,** 1082–1086 (2010).

39. Mendenhall, E. M. *et al.* GC-Rich Sequence Elements Recruit PRC2 in Mammalian ES Cells. *PLoS Genet.* **6,** e1001244 (2010).

40. Grandori, C., Cowley, S. M., James, L. P. & Eisenman, R. N. The Myc/Max/Mad network and the transcriptional control of cell behavior. *Annu. Rev. Cell Dev. Biol.* **16,** 653–699 (2000).

41. Tsai, S.-Y. *et al.* Mouse development with a single E2F activator. *Nature* **454,** 1137–1141 (2008).

42. Fernandez-Zapico, M. E. *et al.* A functional family-wide screening of SP/KLF proteins identifies a subset of suppressors of *KRAS* -mediated cell growth. *Biochem. J.* **435,** 529–537 (2011).

43. Marais, R., Wynne, J. & Treisman, R. The SRF accessory protein Elk-1 contains a growth factor-regulated transcriptional activation domain. *Cell* **73,** 381–393 (1993).

44. Arsenian, S., Weinhold, B., Oelgeschläger, M., Rüther, U. & Nordheim, A. Serum response factor is essential for mesoderm formation during mouse embryogenesis. *EMBO J.* **17,** 6289–6299 (1998).

45. Tang, S. J. *et al.* The Tlx-2 homeobox gene is a downstream target of BMP signalling and is required for mouse mesoderm development. *Development* **125,** 1877–1887 (1998).

46. Morrisey, E. E., Ip, H. S., Tang, Z., Lu, M. M. & Parmacek, M. S. GATA-5: a transcriptional activator expressed in a novel temporally and spatially-restricted pattern during embryonic development. *Dev. Biol.* **183,** 21–36 (1997).

47. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv* (2013).

48. Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500,** 477–481 (2013).

49. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15,** 1929–1958 (2014).

50. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. in *International conference on artificial intelligence and statistics* 249–256 (2010).

51. Kingma, D. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* (2014).

52. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13,** 281–305 (2012).

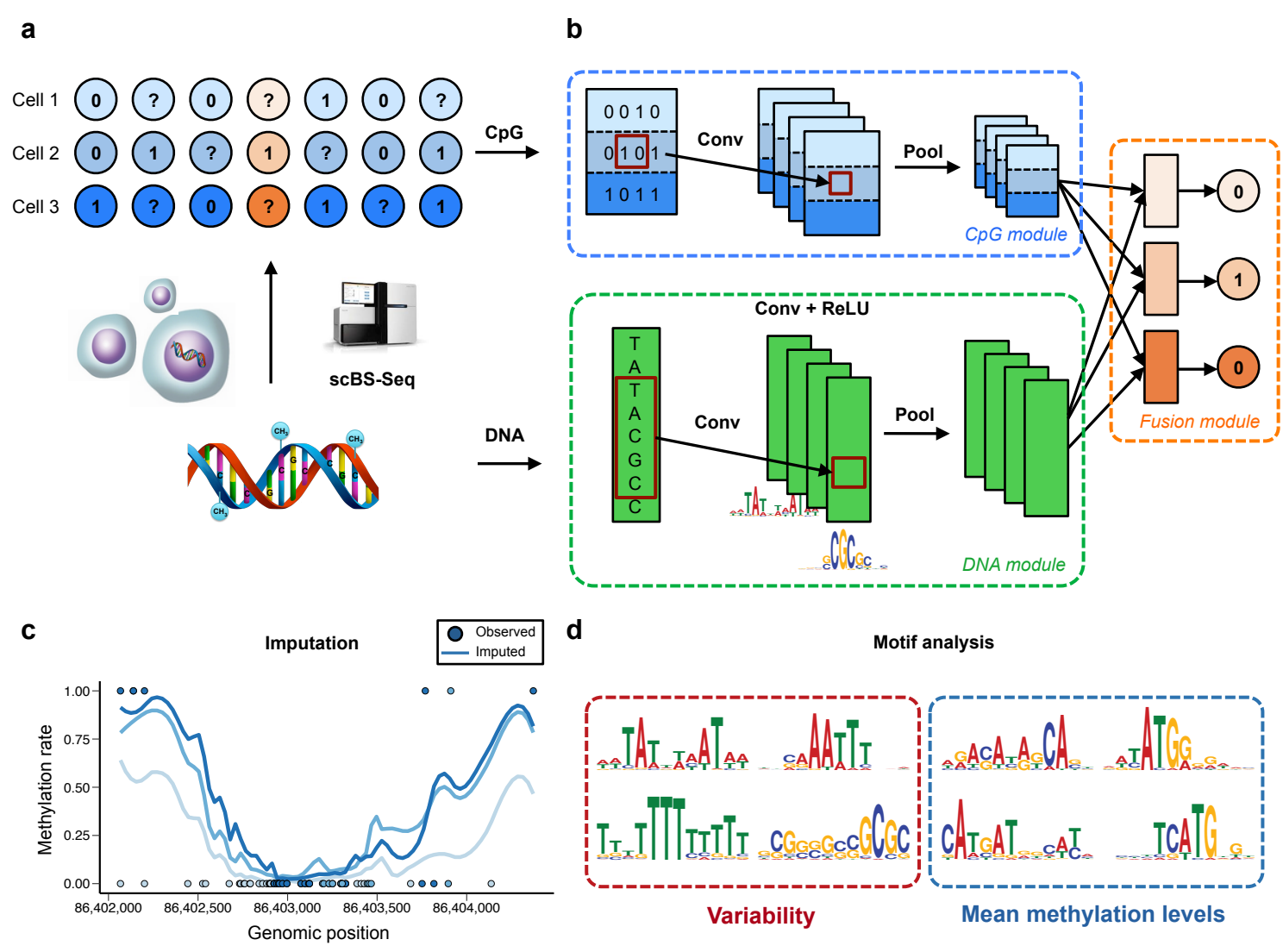
53. Bastien, F. *et al.* Theano: new features and speed improvements. *arXiv* (2012).

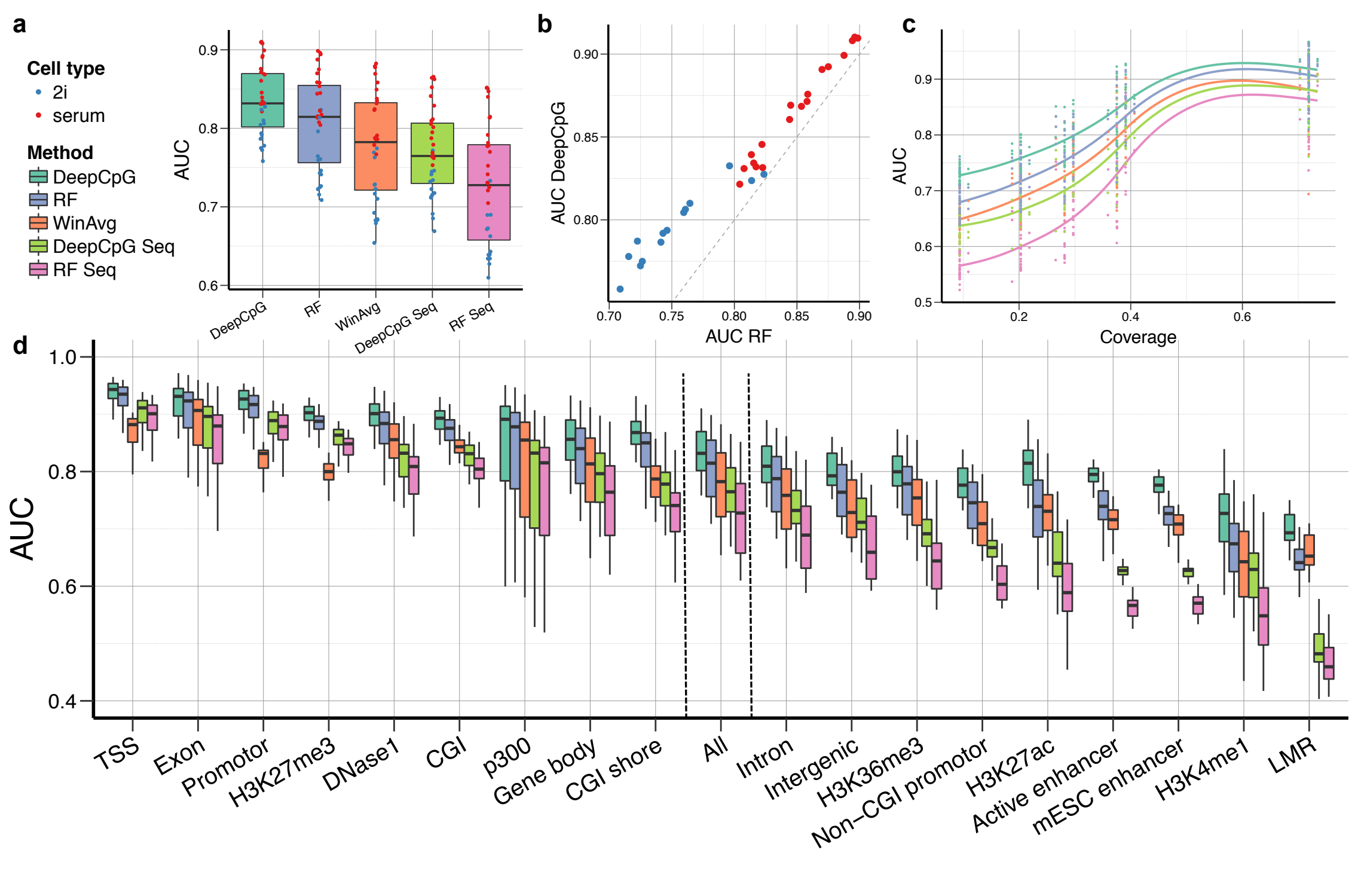
54. Chollet, F. *Keras: Theano-based deep learning library*.

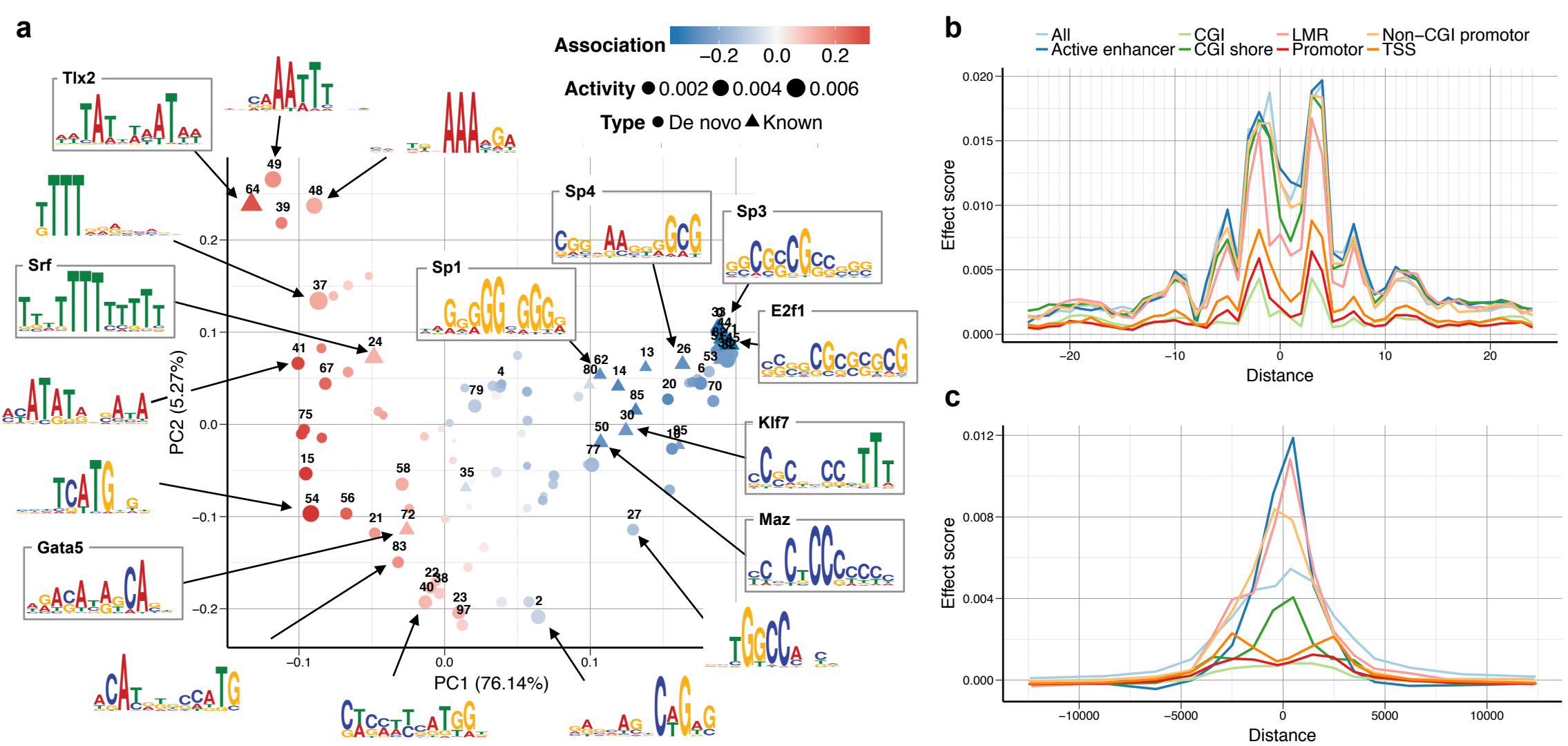
55. Crooks, G. E. WebLogo: A Sequence Logo Generator. *Genome Res.* **14,** 1188–1190 (2004).

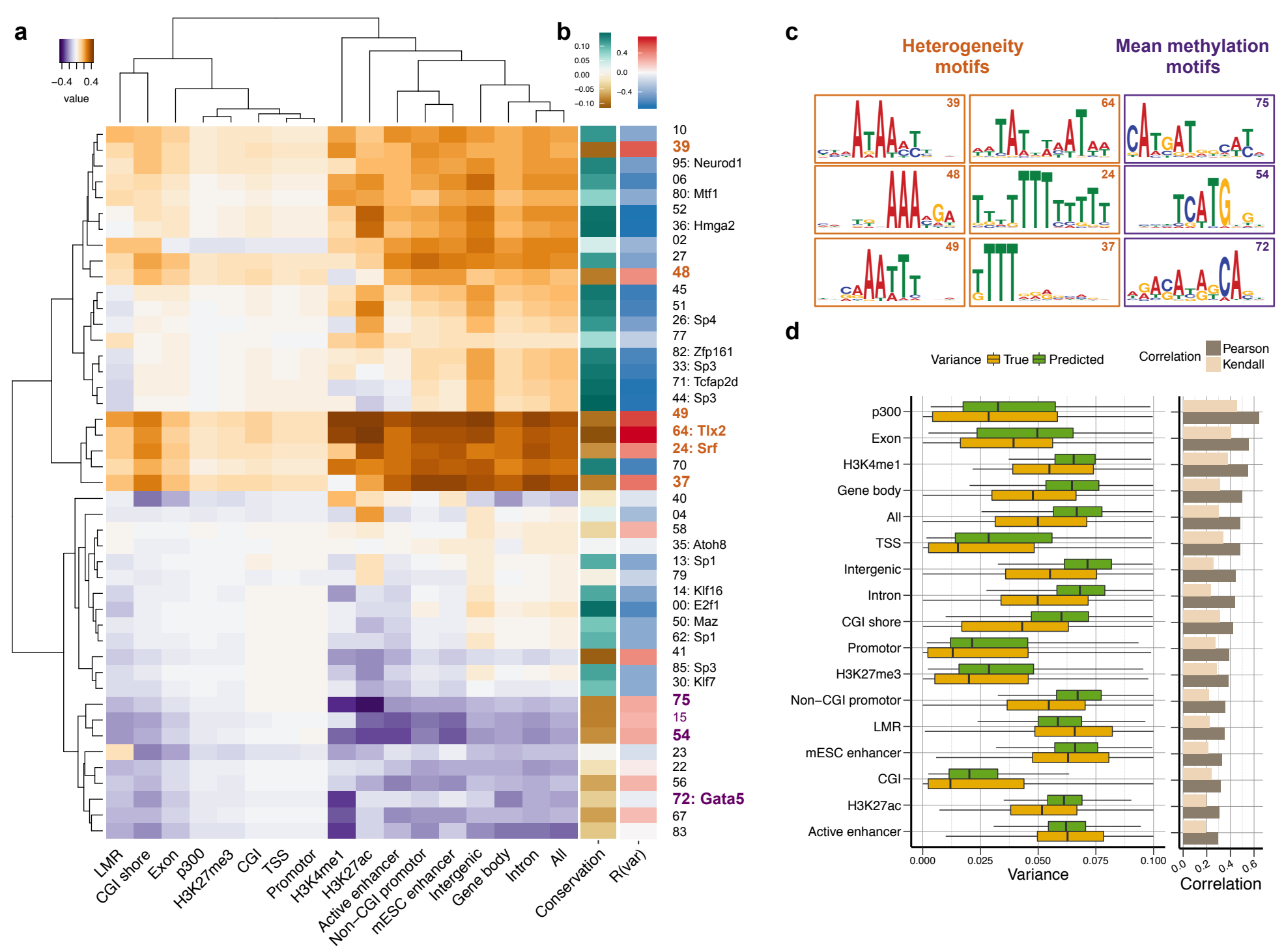
56. Bailey, T. L. *et al.* MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.* **37,** W202–W208 (2009).

57. Siepel, A. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).


a
Cell 1
Cell 2
Cell 3
CpG
DNA
scBS-Seq
b
Conv
Pool
CpG module
Conv + ReLU
Conv
Pool
DNA module
Fusion module
c
Imputation
Observed
Imputed
Methylation rate
Genomic position
d
Motif analysis
Variability
Mean methylation levels


a
Cell type
2i
serum
Method
DeepCpG
RF
WinAvg
DeepCpG Seq
RF Seq
AUC
0.9
0.8
0.7
0.6
DeepCpG
RF
WinAvg
DeepCpG Seq
RF Seq
b
AUC DeepCpG
AUC RF
0.90
0.85
0.80
0.70
0.75
0.80
0.85
0.90
c
AUC
Coverage
0.9
0.8
0.7
0.6
0.5
0.2
0.4
0.6
d
AUC
1.0
0.8
0.6
0.4
TSS
Exon
Promotor
H3K27me3
DNase1
CGI
p300
Gene body
CGI shore
All
Intron
Intergenic
H3K36me3
Non−CGI promotor
H3K27ac
Active enhancer
mESC enhancer
H3K4me1
LMR


a
Association
−0.2
0.0
0.2
Activity 0.002 0.004 0.006
Type De novo Known
Tlx2
Srf
Gata5
Sp4
Sp1
Sp3
E2f1
Klf7
Maz
PC1 (76.14%)
PC2 (5.27%)
b
All
Active enhancer
CGI
CGI shore
LMR
Promotor
Non−CGI promotor
TSS
Effect score
Distance
c
Effect score
Distance


a
−0.4 0.4
value
b
0.10
0.05
0.00
−0.05
−0.10
0.4
0.0
−0.4
10
39
95: Neurod1
06
80: Mtf1
52
36: Hmga2
02
27
48
45
51
26: Sp4
77
82: Zfp161
33: Sp3
71: Tcfap2d
44: Sp3
49
64: Tlx2
24: Srf
70
37
40
04
58
35: Atoh8
13: Sp1
79
14: Klf16
00: E2f1
50: Maz
62: Sp1
41
85: Sp3
30: Klf7
75
15
54
23
22
56
72: Gata5
67
83
LMR
CGI shore
Exon
p300
H3K27me3
CGI
TSS
Promotor
H3K4me1
H3K27ac
Active enhancer
Non−CGI promotor
mESC enhancer
Intergenic
Gene body
Intron
All
Conservation
R(var)
c
Heterogeneity motifs
Mean methylation motifs
39
64
75
48
24
54
49
37
72
d
Variance
True
Predicted
Correlation
Pearson
Kendall
p300
Exon
H3K4me1
Gene body
All
TSS
Intergenic
Intron
CGI shore
Promotor
H3K27me3
Non−CGI promotor
LMR
mESC enhancer
CGI
H3K27ac
Active enhancer
0.000
0.025
0.050
0.075
0.100
Variance
0.0
0.2
0.4
0.6
Correlation