Supplementary tables

Supplementary Table 1 | Prediction performance of alternative methods and contexts
This table is provided as separate xlsx file (tab1_perf.xlsx.xlsx).

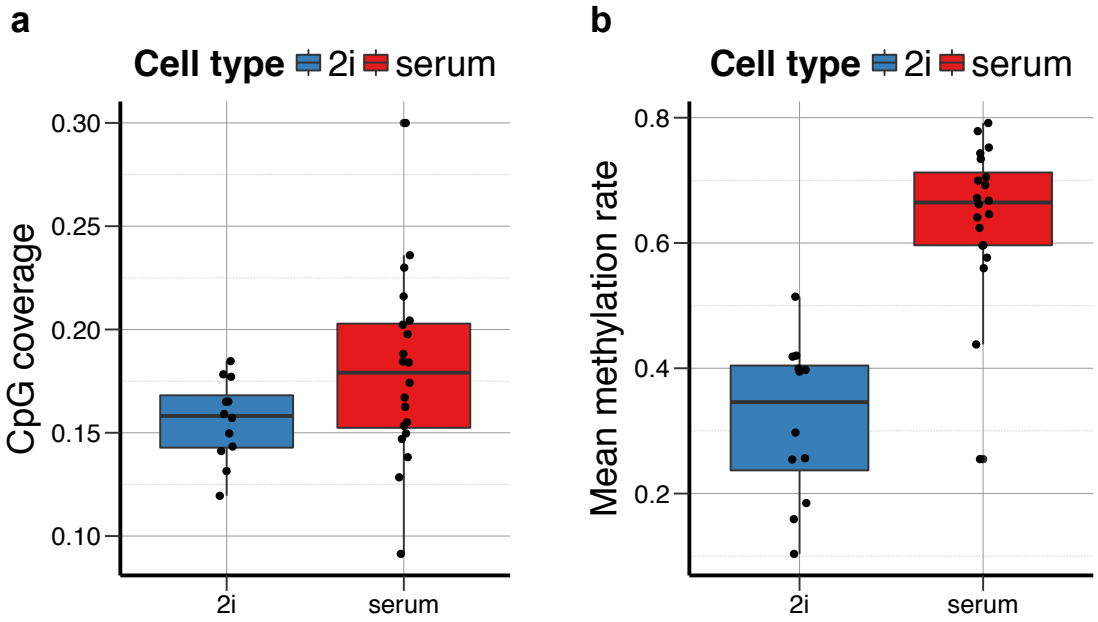
Supplementary Table 2 | Summary table of learnt motifs
This table is provided as separate html file (tab2_motifs.html).

Supplementary Table 3 | DeepCpG hyper-parameters
This table is provided as separate xlsx file (tab3_hyper-params.xlsx).

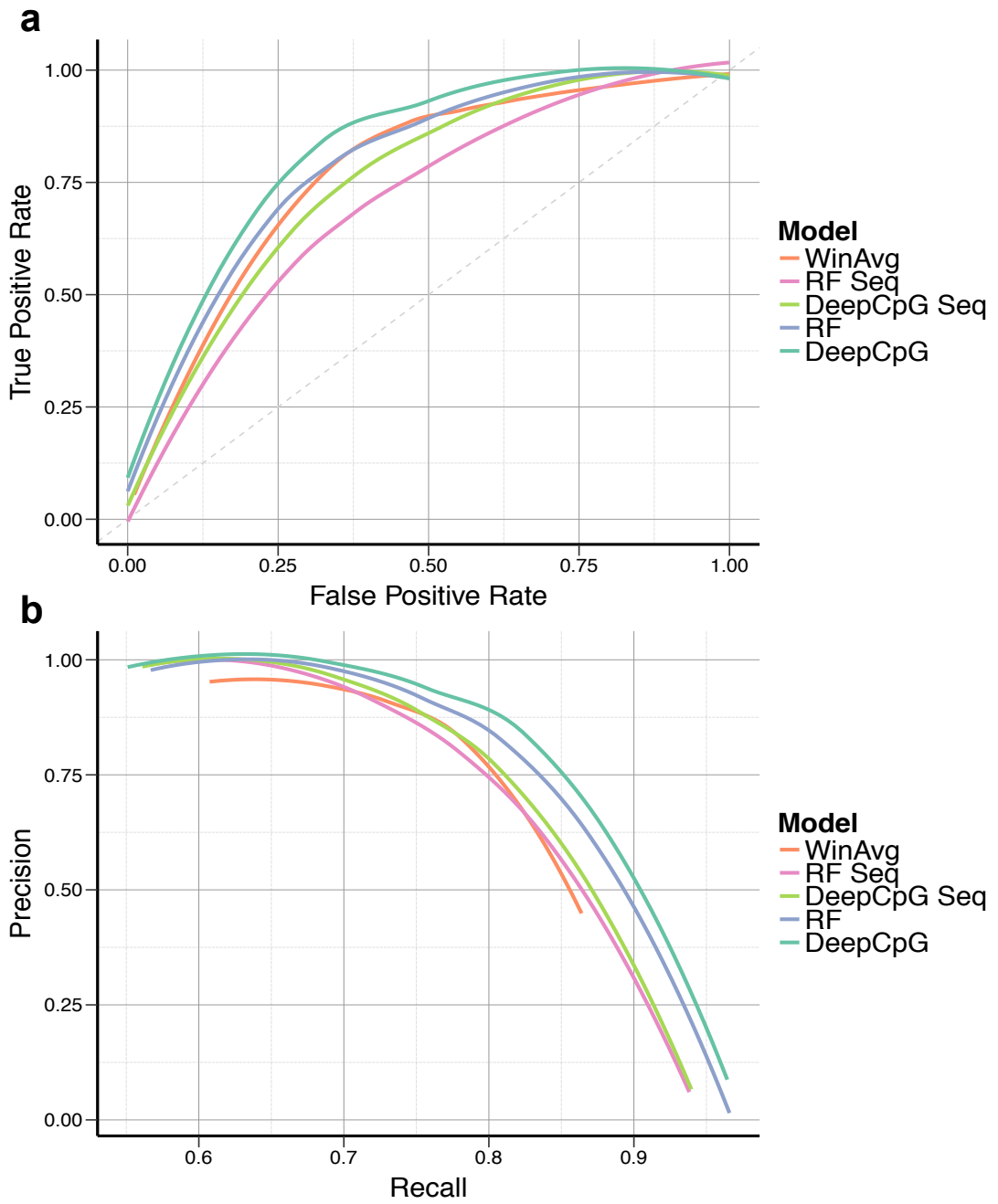
Supplementary Table 4 | Features RF Zhang model
This table is provided as separate xlsx file (tab4_rf-zhang.xlsx).

Supplementary Table 5 | Description of genomic contexts
This table is provided as separate xlsx file (tab5_contexts.xlsx).

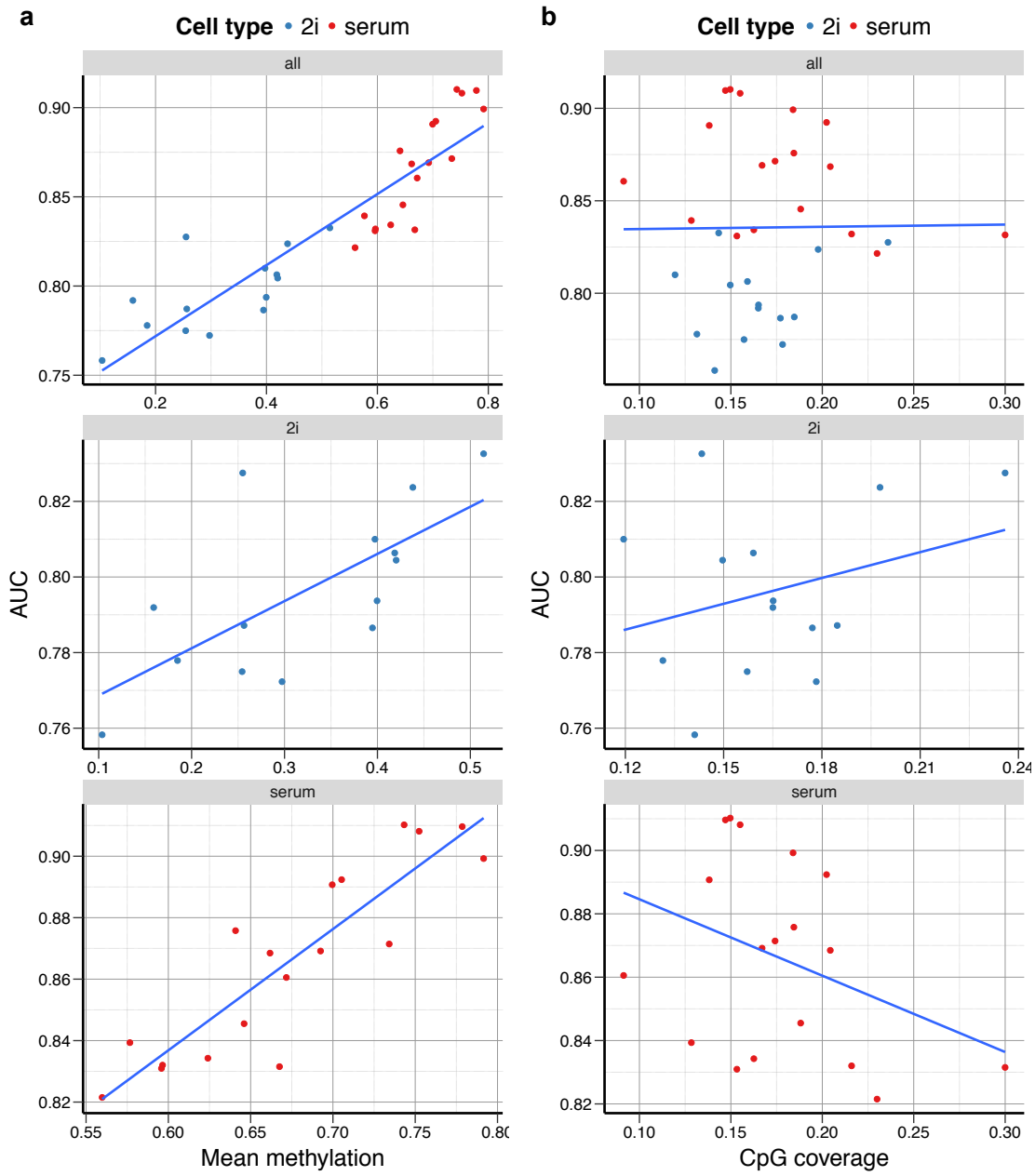
Supplementary figures



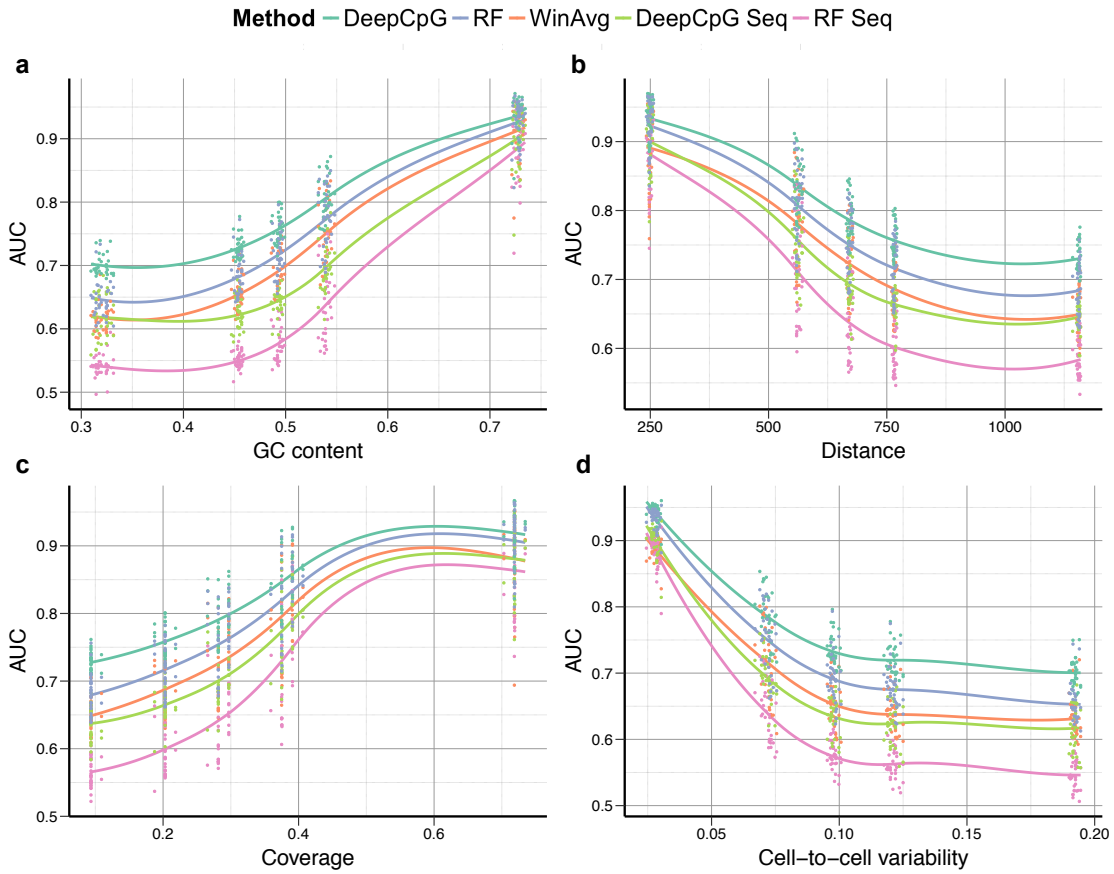
Supplementary Figure 1 | Quality metrics of cells profiled using scBS-Seq. (**a**) Genome wide CpG coverage in 2i cells (blue, 16%) and serum cell (red, 0.18%). (**b**) Genome-wide mean methylation rate in 2i cells (blue, 32%), and serum cell (red, 64%).



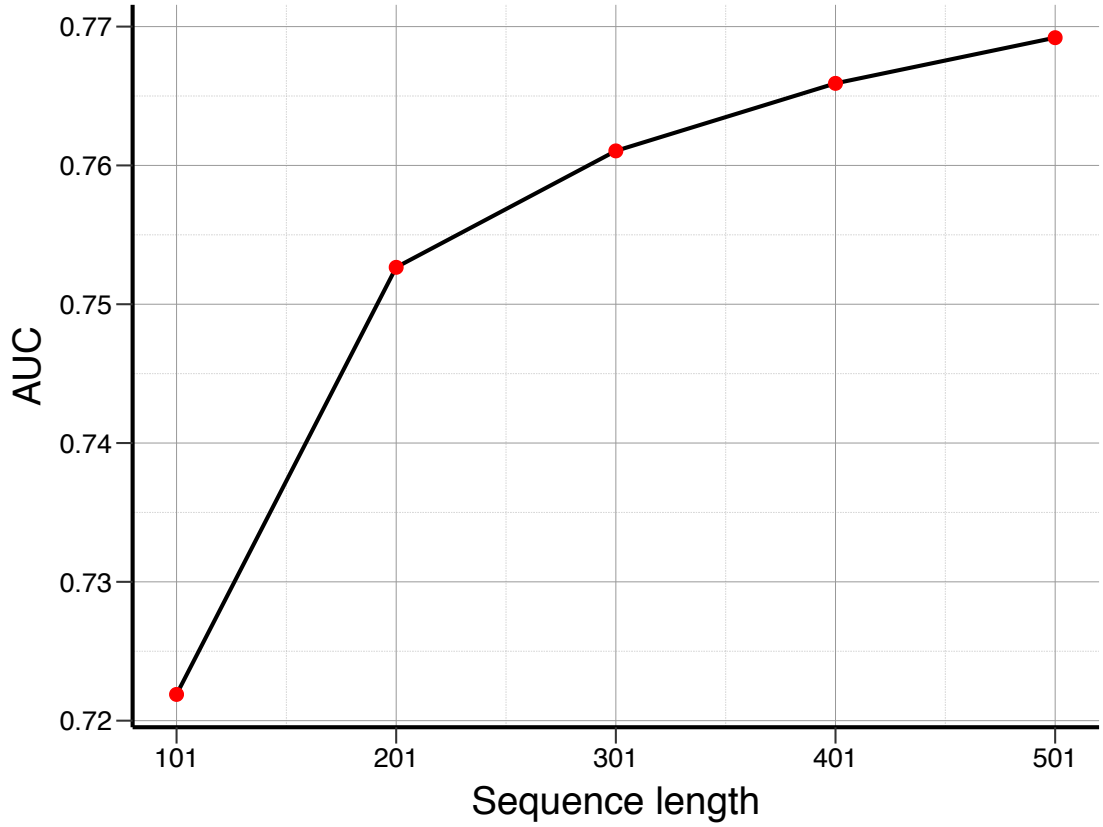
Supplementary Figure 2 | Receiver operating characteristic and precision recall for methylation predictions using alternative methods. Prediction performance comparision by receiver operating characterisic curve (**a**) and precision recall curve (**b**). Shown are results for DeepCpG, window averaging in consecutive (3 kb) regions (WinAvg), a random forest model (RF), and the corresponding models when trained using sequence information only (DeepCpG Seq, RF Seq). Individual curves show aggregate results across all cells.



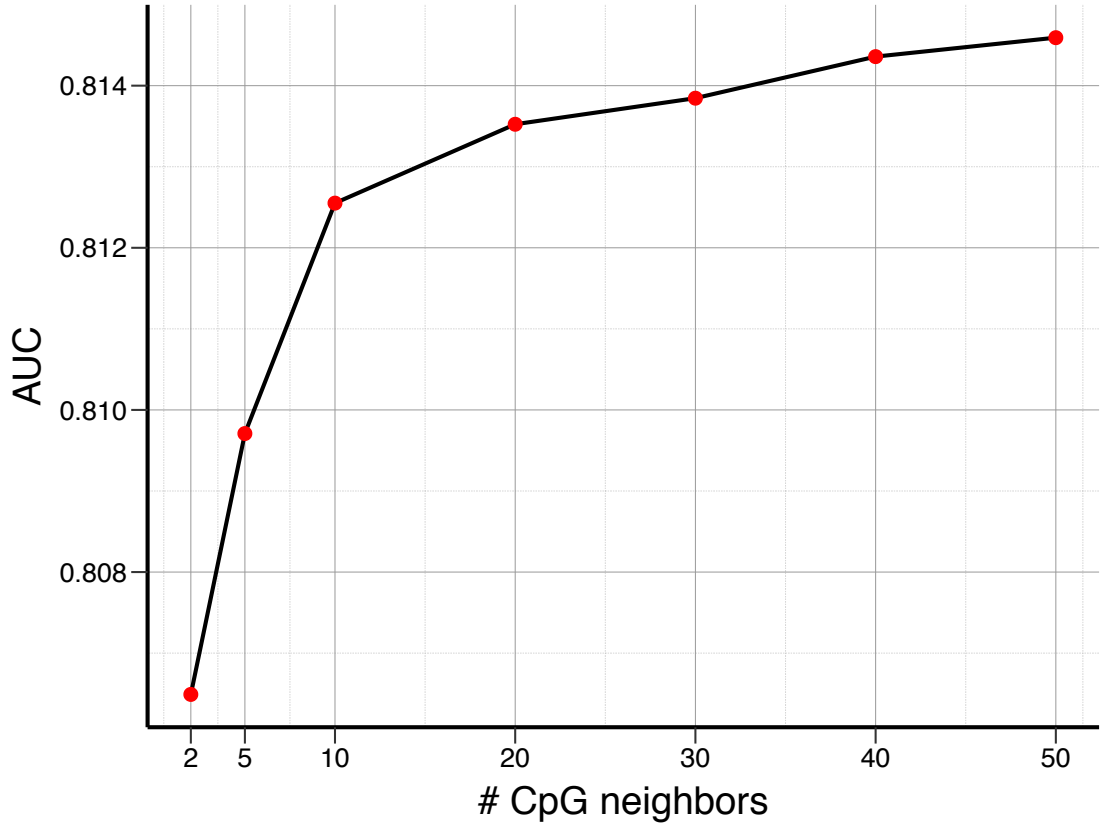
Supplementary Figure 3 | Correlation between prediction accuracy and genome-wide methylation rate in individual cells. AUC prediction performance for individual cells vs. the corresponding genome wide methylation rate (**a**), and CpG coverage (**b**) for all, 2i (blue), and serum (red) cells separately. We find a clear positive correlation (R=0.91, 0.69, 0.89 for all, 2i, and serum cells) between the methylation rate per cell and the corresponding prediction accuracy.



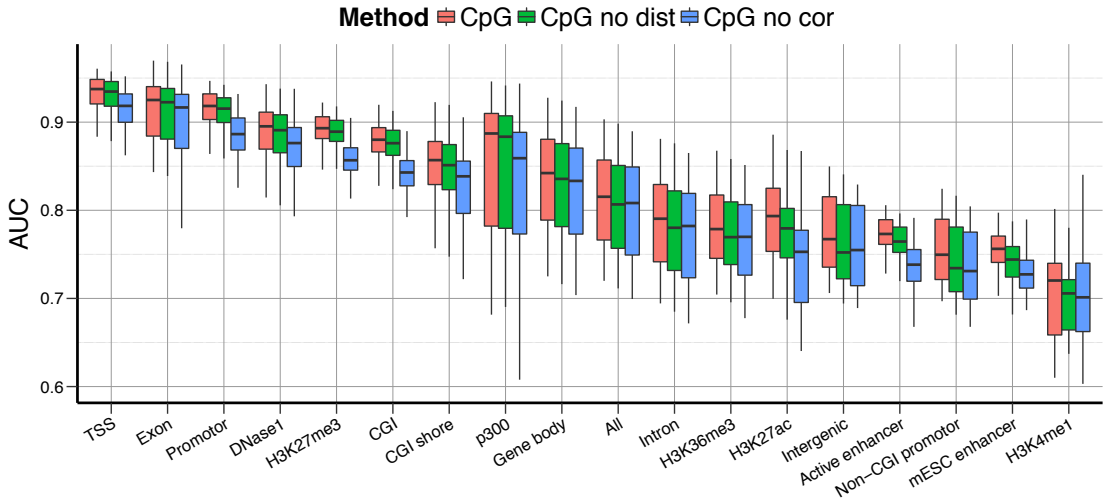
Supplementary Figure 4 | Prediction performance stratified for differnet genomic regions. Prediction performances of alternative methods stratified by (**a**) GC content, (**b**) distance to neighboring CpG sites, (**c**) fraction of cells by which the target CpG site is covered, (**d**) cell-to-cell variability within 3 kb windows centered on the target CpG site.



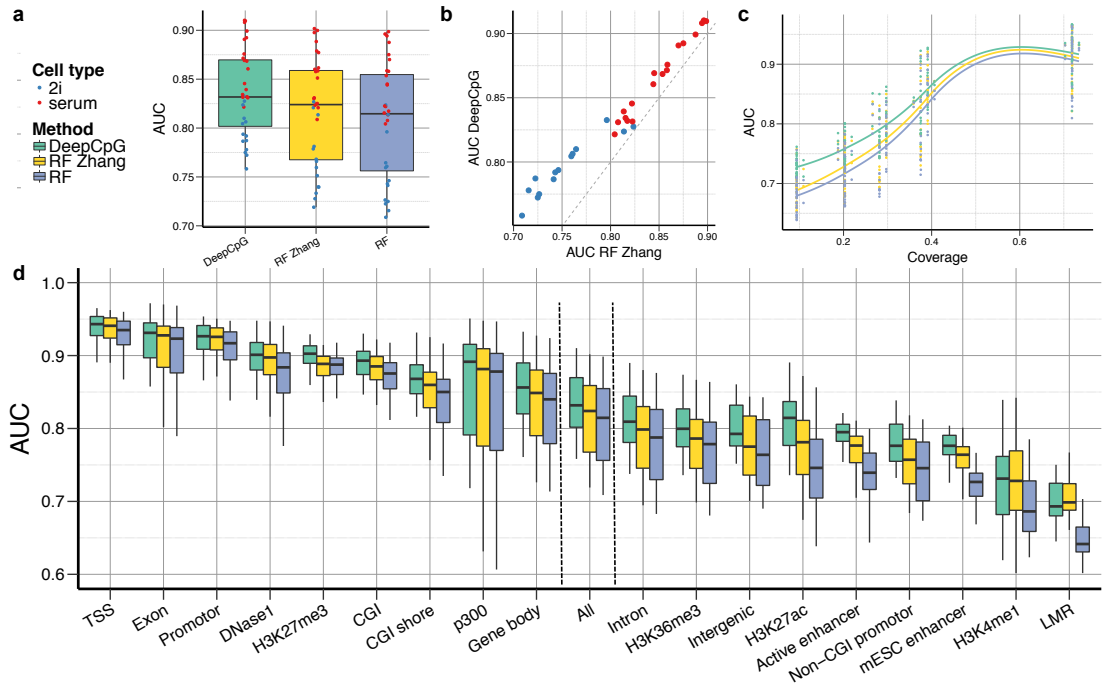
Supplementary Figure 5 | Prediction performance of the DeepCpG DNA module when considering alternative input DNA sequence lengths. Shown is AUC of the DeepCpG DNA module (DeepCpG Seq) when using DNA sequences of increasing lengths (101 bp to 501 bp) centred on the target CpG site.



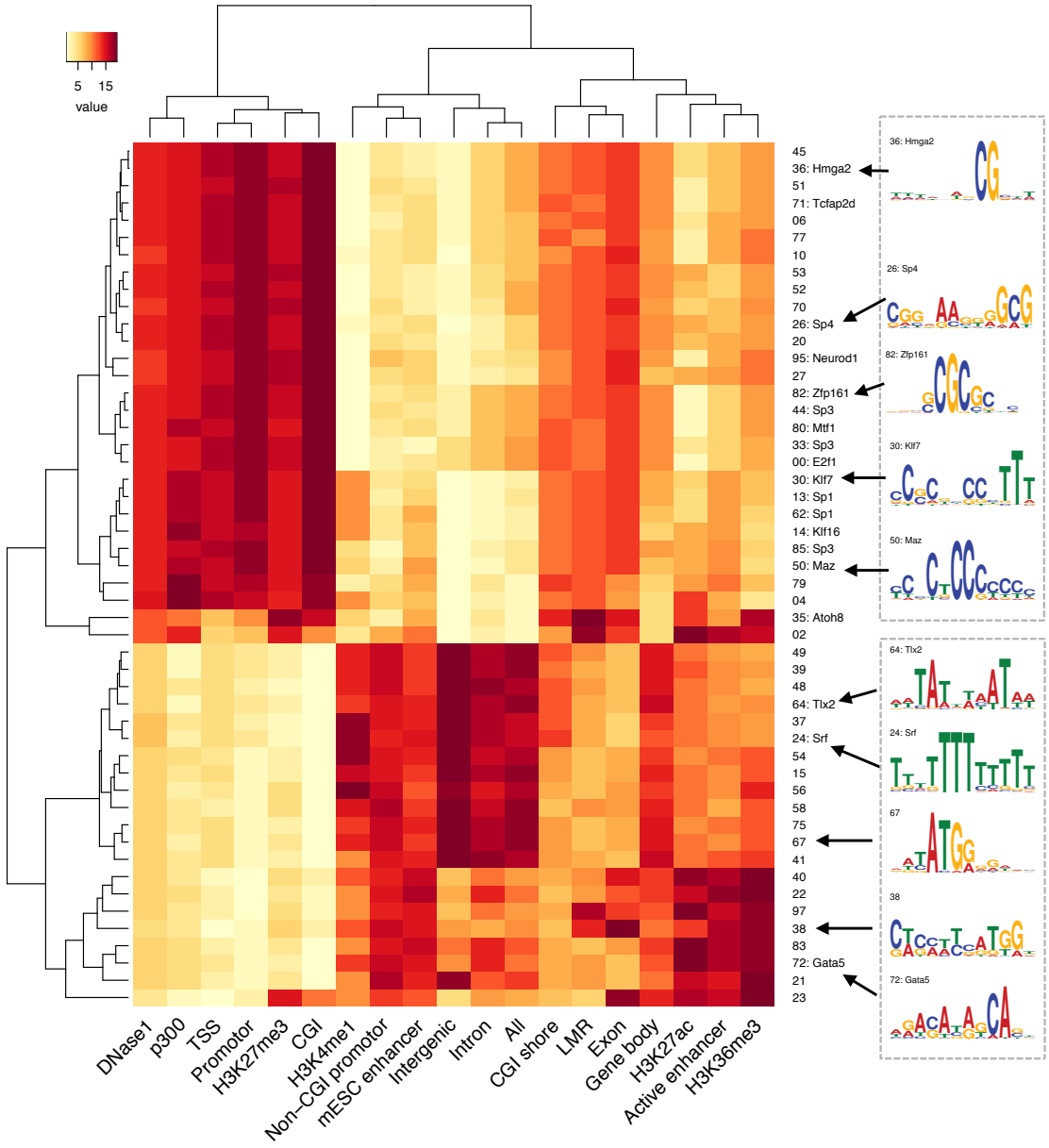
Supplementary Figure 6 | Prediction performance of the DeepCpG CpG module when considering increasing numbers of neighbouring CpG sites for training. Shown is AUC of the DeepCpG CpG module, when using between 2 and 50 binary CpG sites adjacent to the target CpG site from all cells as input (at increasing distances, symmetric around the target site).



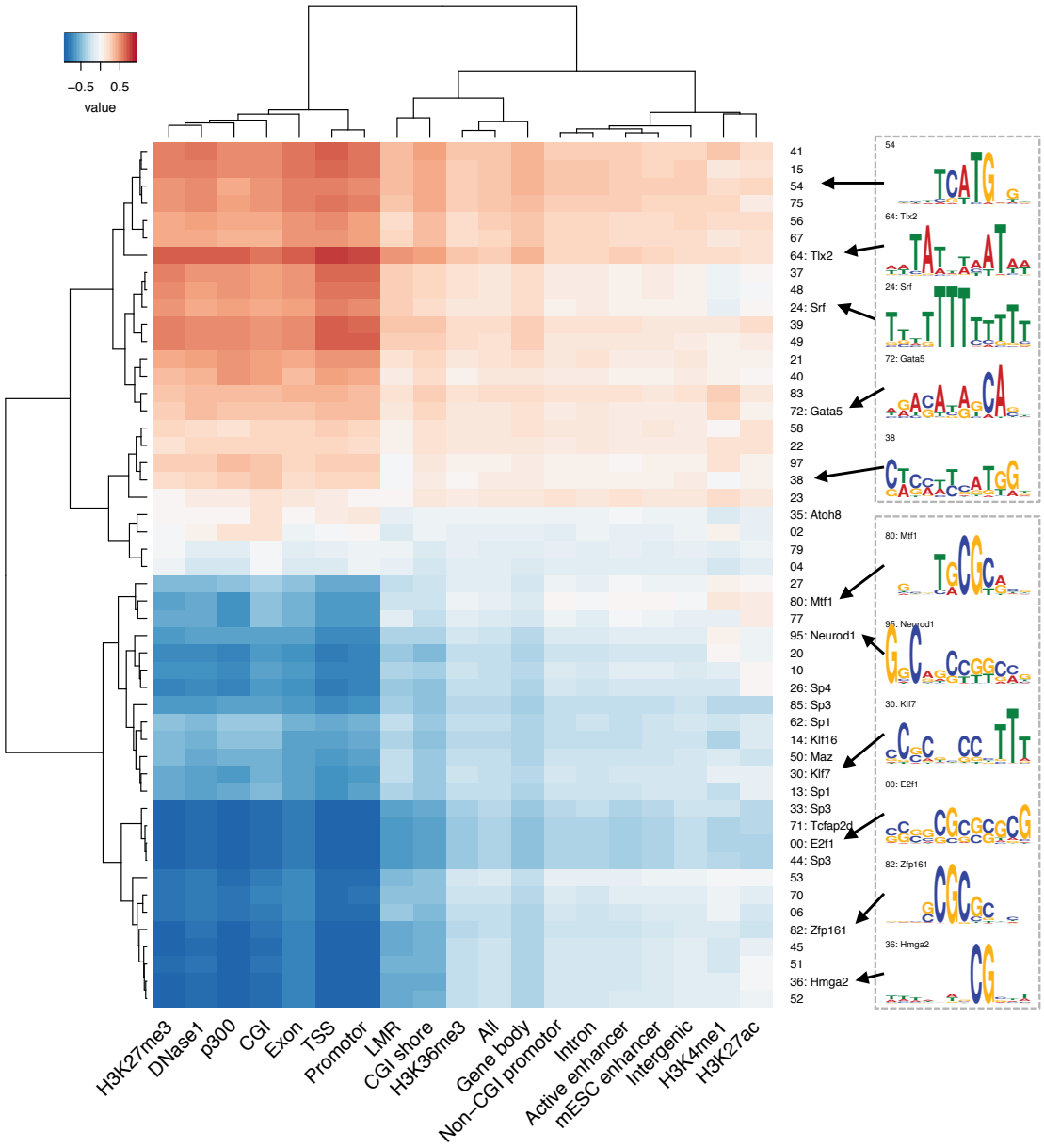
Supplementary Figure 7 | Feature analysis of the DeepCpG CpG module
Shown is AUC of the DeepCpG CpG model when using the binary methylation state and distance of 50 neighbouring CpG sites of all cells (CpG), using binary methylation states of all cells without distance information (CpG no dist), and using binary methylation states and distances of the target cell only (CpG no cor).



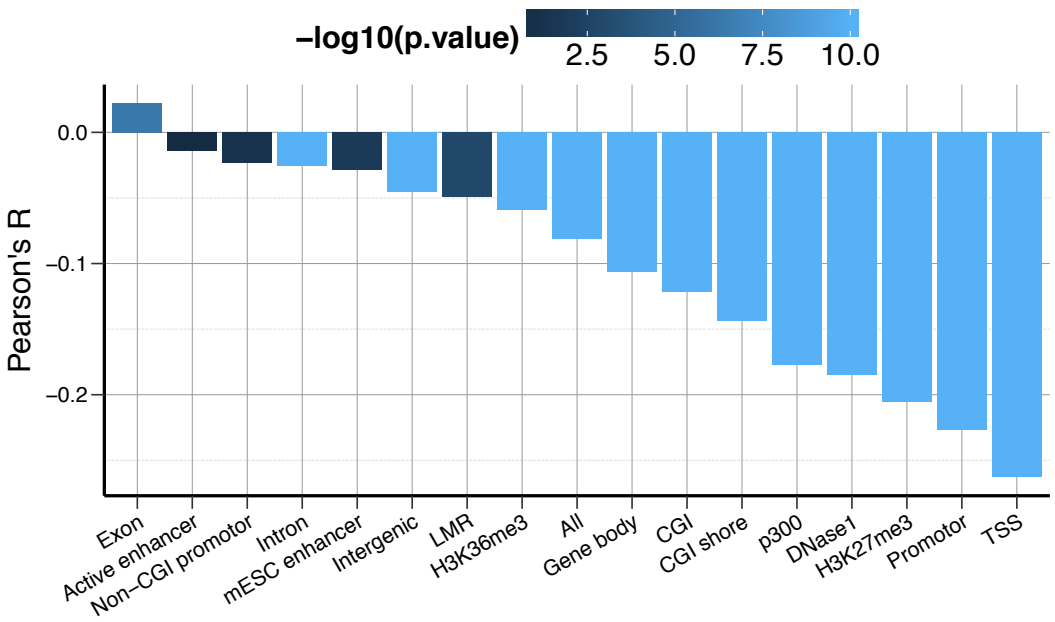
Supplementary Figure 8 | DeepCpG prediction performance compared to a random forest models with additional sequence annotations. Shown is the AUC performance for DeepCpG, the random forest model (RF) described in the main text, as well a random forest with additional sequence annotations as described in Zhang et al. 2015 (RF Zhang), including transcription factor binding sites, DNase1 hypersensitivity sites, and histone modifications marks (Online methods, **Supplementary Table 4**). (**a**) Genome-wide hold-out prediction accuracy quantified using the area under the receiver operating characteristic curve (AUC) in 12 2i cells (blue) and 20 serum cells (red). (**b**) Relative comparison of DeepCpG and RF Zhang. Each dot denotes one single cell. (**c**) AUC stratified by the fraction of cells by which the target CpG site is covered, and (**d**) when considering alternative sequence contexts.



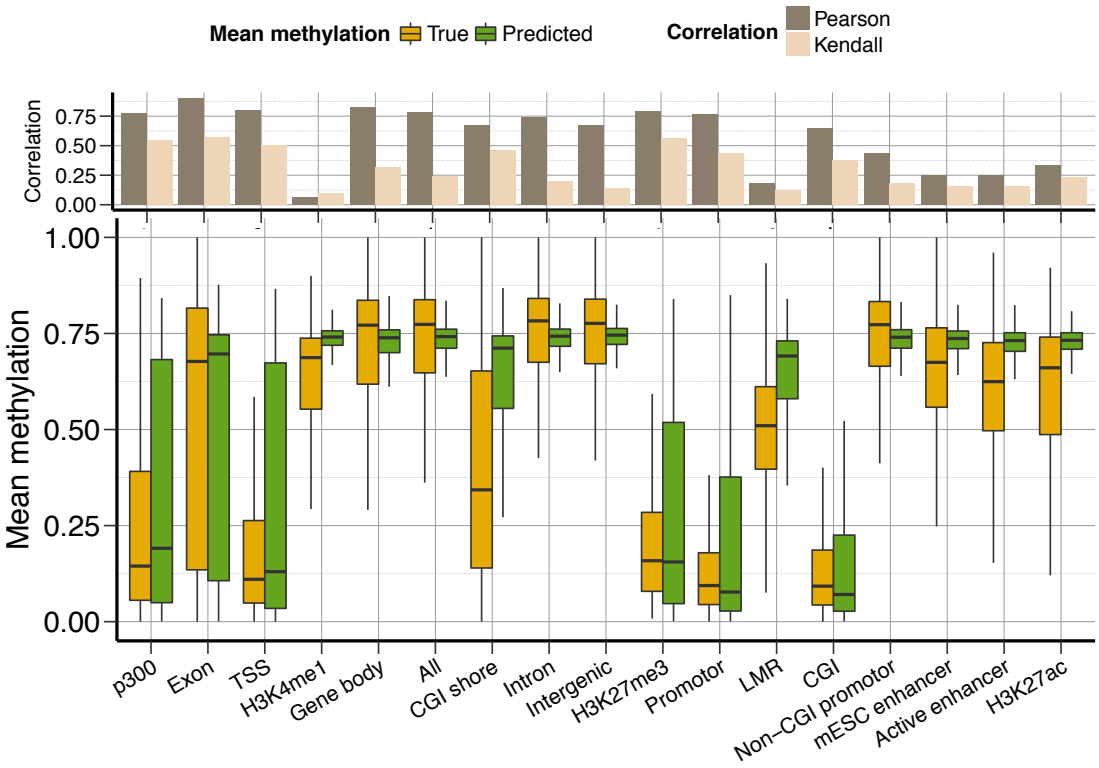
Supplementary Figure 9 | Motif activities in genomic contexts. Rank of motif activies (occurrence frequencies) in different genomic contexts. The highest activity of CG rich motifs is observed in regions with high CG content such as promotors and CpG islands (CGI). Conversely, the highest activity of AT rich motifs is observed in low CG content regions such as intergenic or active enhancer regions.



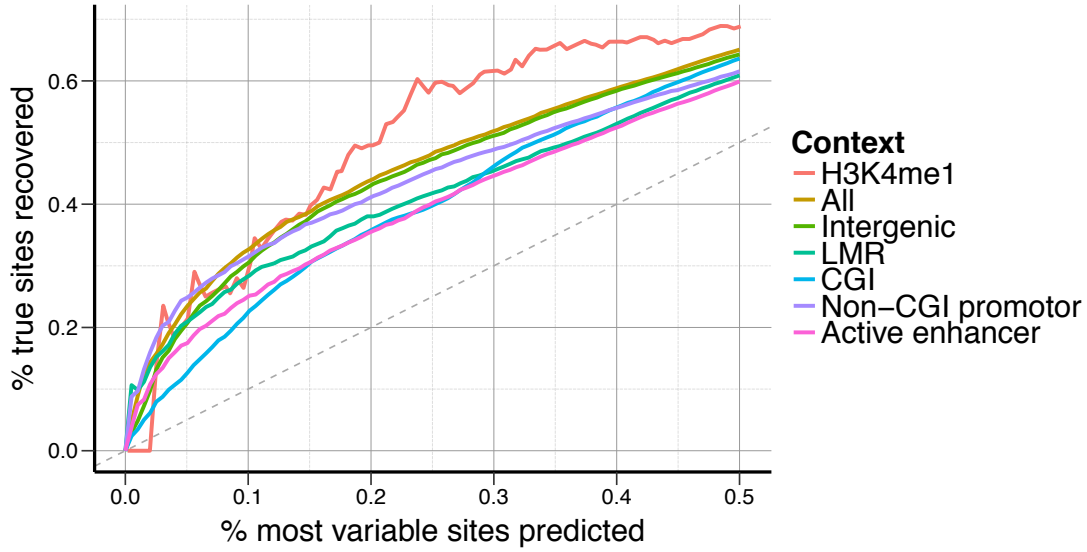
Supplementary Figure 10 | Motif influences in genomic contexts. Effect of discovered motifs on CpG methylation in different genomic contexts, quantified by Spearman's correlation between motif activities and predicted methylation levels (Online methods). Motifs cluster into CG-rich methylation-increasing motifs and AT-rich motifs that decrease methylation levels.



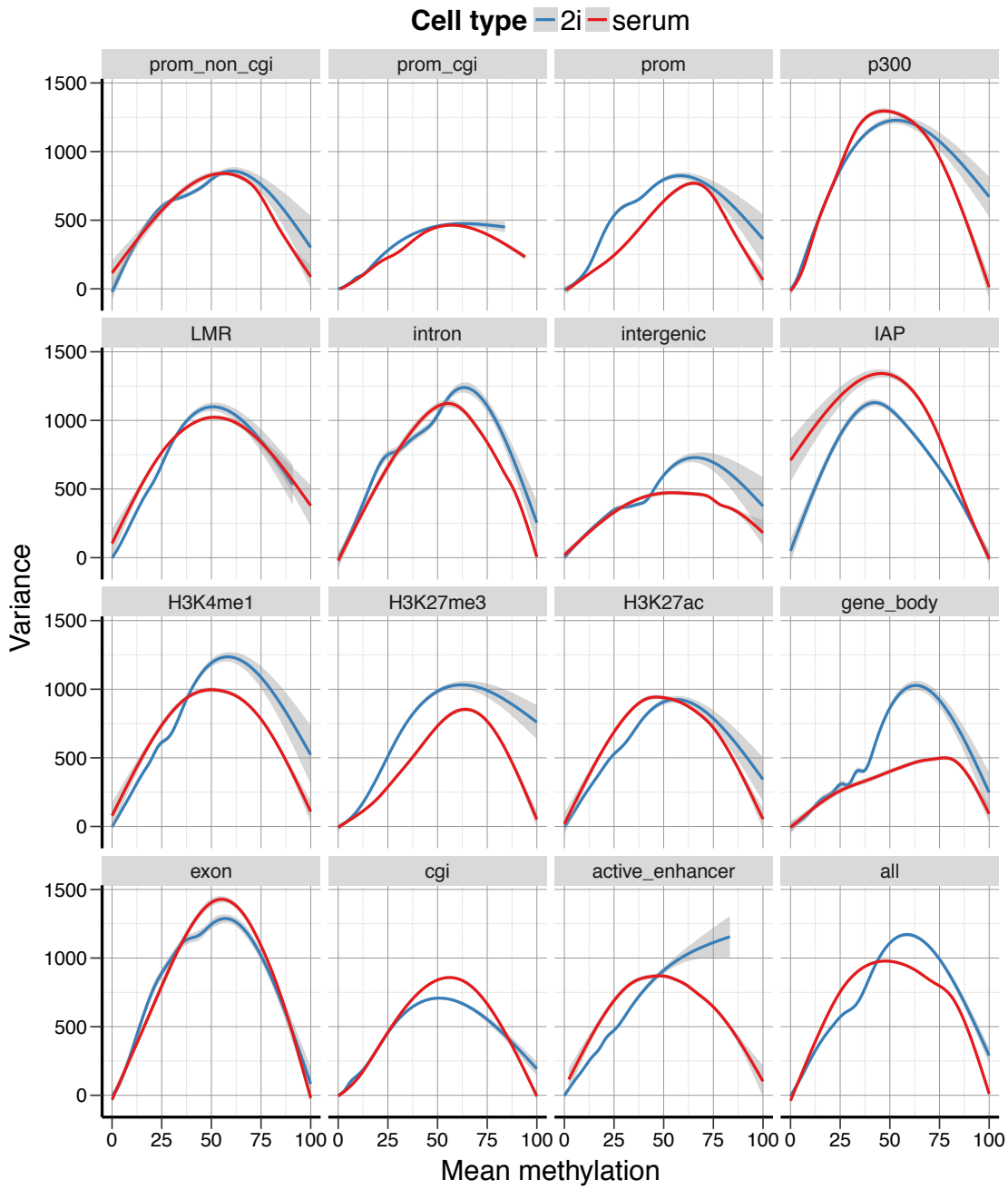
Supplementary Figure 11 | Correlation mutation effects and DNA sequence conservation. Correlation between predicted effect of nucleotide changes and *PhastCons* conservation score for alternative contexts. Nucleotide changes are significantly anti-correlated overall ('All', $P < 1.0\times10^{-15}$) in CG dense regions.



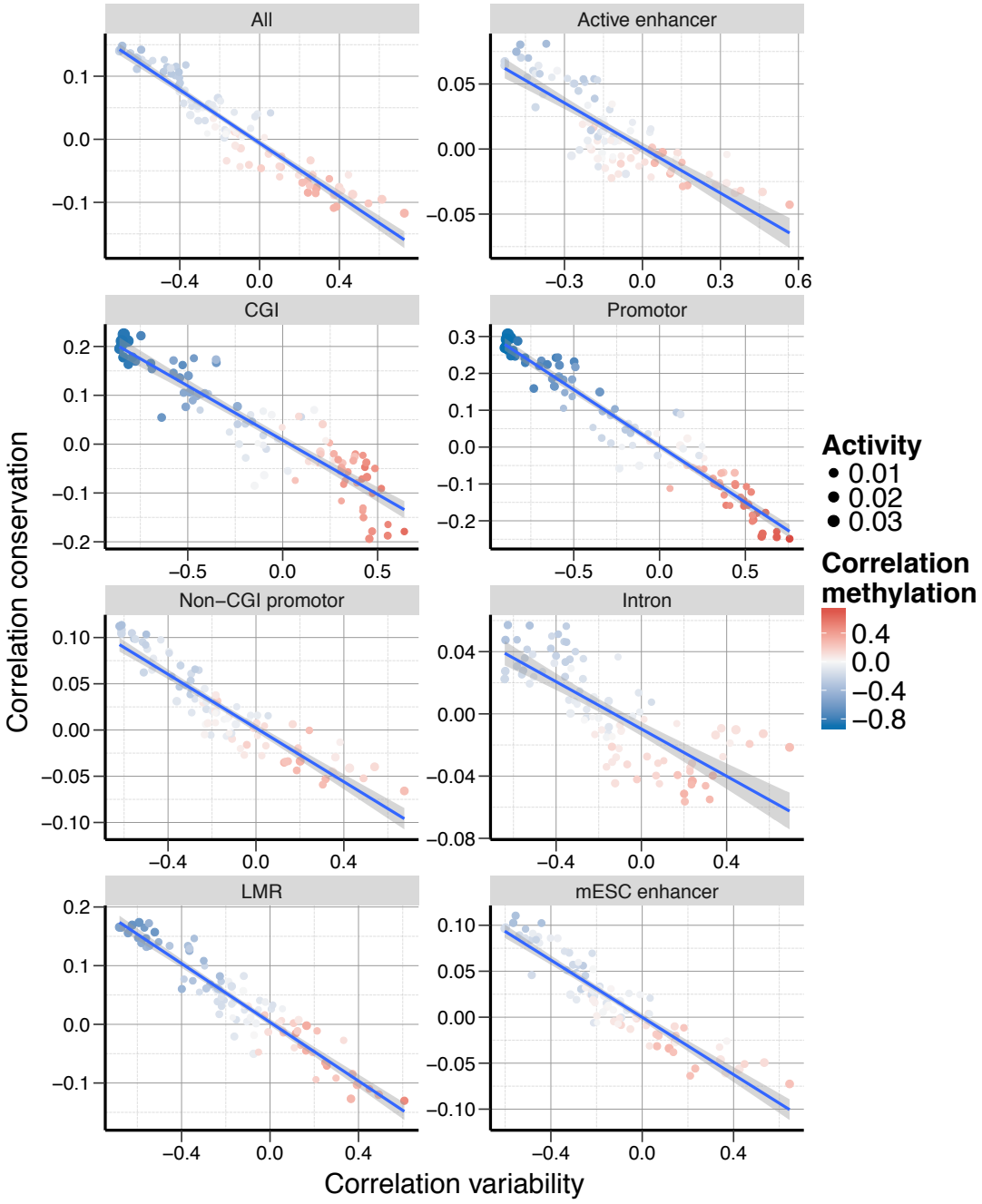
Supplementary Figure 12 | Prediction performance of mean methylation levels. Histogram of predicted (green) and observed mean methylation levels in 3 kb windows centered on indivual CpG sites, along with Pearon's and Kendall's correlation coefficients on top. Shown are regions from test set chromosomes only (Onlinet Methods).



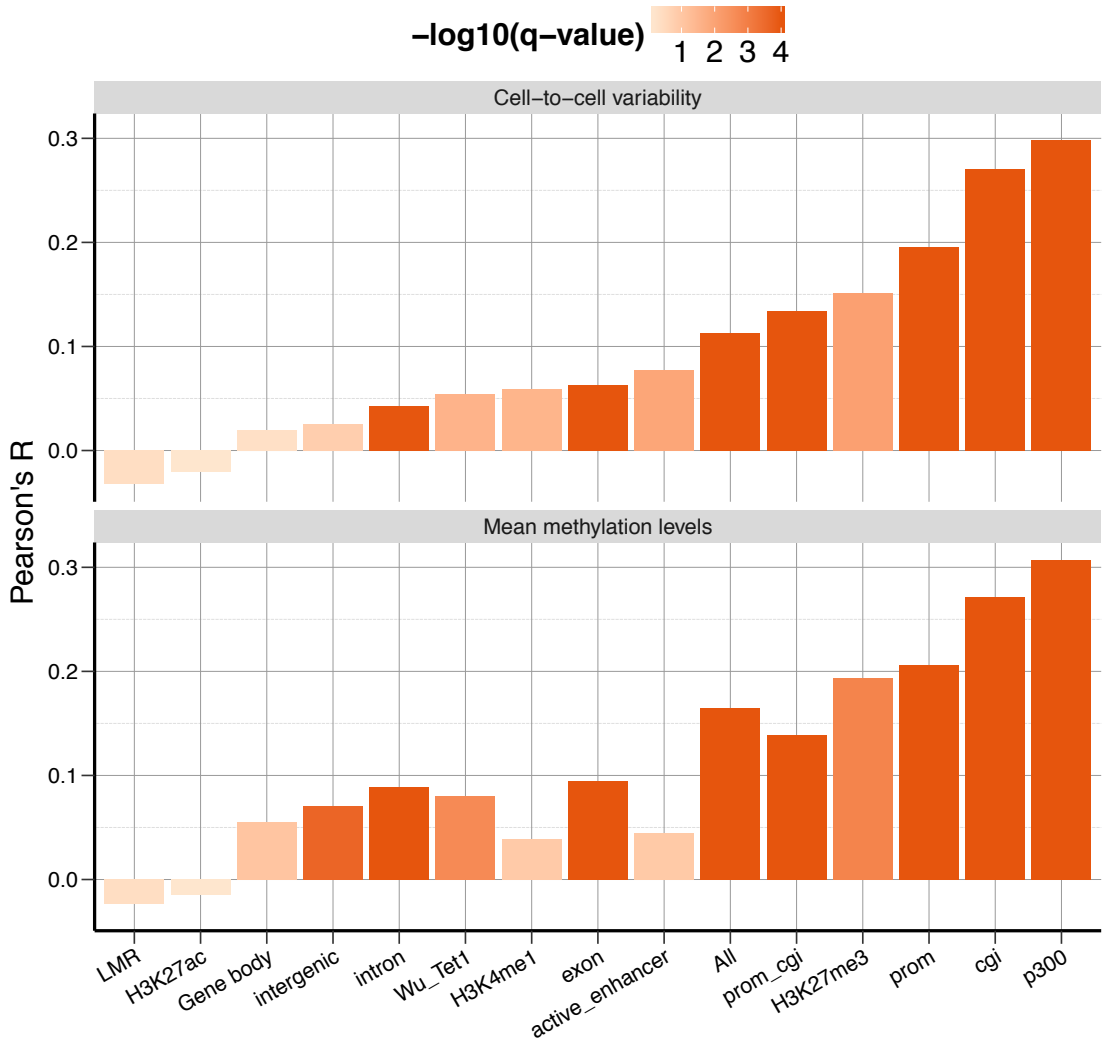
Supplementary Figure 13 | Identification of most variable CpG sites. Overlap between true and predicted most variable sites for different thresholds and genomic contexts. Single CpG sites were ranked by the true variance estimated for 3kb windows, or the predicted variance, and the overlap computed for the fraction of most variables sites shown on the x-axis. Dashed line indicates random ranking performance. Ranking accuracy was assessed on test set chromosomes (Onlinet Methods).



Supplementary Figure 14 | Dependency between mean methylation levels and cell-to-cell variance. Smoothed representation of mean methylation levels on x-axis verus cell-to-cell variance on the y-axis, for 2i and serum cells in different genomic contexts. Cell-to-cell variance is linked to the mean methylation levels and highest for an intermediate methylation level of 50%.



Supplementary Figure 15 | Linkage between motif correlation with sequence conservation and cell-to-cell variability. Correlation of motif activities (occurrence frequencies) with cell-to-cell variability (x-axis), and DNA conservation (y-axis), showing that variability increasing motifs are most active in non-conserved regions. Individual dots correspond to motifs, whose mean activity is represented by size, and influence on CpG methylation by colour.



Supplementary Figure 16 | Functional validation of predicted cell-to-cell variability. Pearson correlation coefficient between methylome-transcriptome linkage as reported in Angermueller *et al* (2016), and predicted cell-to-cell variability (top), as well as predicted mean methylation levels (bottom). Colour denotes significance (q-value, Benjamini Hochberg adjusted). Correlations are significant (FDR < 0.01) overall ('All') and in most genomic contexts.