# 4 RNA and chromatin modification patterns around promoters

## Patterns of gene expression can be modelled using histone modifications and transcription factor binding at promoters

We developed predictive models to explore the interaction between histone modifications and measures of transcription at promoters, distinguishing between modifications known to be added as a consequence of transcription (such as H3K36me3 and H3K79me2) and other categories of histone marks[59]. In our analyses, the best models had two components: an initial classification component (on/off) and a second quantitative model component. Our models showed activating acetylation marks (H3K27ac and H3K9ac) are roughly as informative as activating methylation marks (H3K4me3 and H3K4me2) (Figure 2A). Although repressive marks, such as H3K27me3 or H3K9me3, show negative correlation both individually and in the model, removing these marks produces only a small reduction in model performance. However, for a subset of promoters in each cell line repressive histone marks (H3K27me3 or H3K9me3) must be used to accurately predict their expression. We also examined the interplay between the H3K79me2 and H3K36me3 marks, both of which mark gene bodies, likely reflecting recruitment of modification enzymes by polymerase isoforms. As described previously, H3K79me2 occurs preferentially at the 5' ends of gene bodies and H3K36me3 occurs more 3', and our analyses support the previous model in which the H3K79me2 to H3K36me3 transition occurs at the first 3' splice site[60].

Few previous studies have attempted to build qualitative or quantitative models of transcription genome-wide from TF levels because of the paucity of documented TF-binding regions and the lack of coordination around a single cell line. We thus examined the predictive capacity of TF-binding signals for the expression levels of promoters (Figure 2B). In contrast to the profiles of histone modifications, most TFs show enriched binding signals in a narrow DNA region near the TSS, with relatively higher binding signals in promoters with higher CpG content. Most of this correlation could be recapitulated by looking at the aggregate binding of TFs without specific TF terms. Together, these correlation models suggest both that a limited set of chromatin marks are sufficient to "explain" transcription and that a variety of TFs might have broad roles in general transcription levels across many genes. It is important to note that this is an inherently observational study of correlation patterns, and is consistent with a variety of mechanistic models with different causal links between the chromatin, TF and RNA assays. However it does indicate that there is enough information present at the promoter regions of genes to explain the majority of variation in RNA expression.

Figure 4 shows some of the characteristic chromatin features of the different regions. For each type of data, we have picked a particular dataset from the K562 cell line for illustration, but the general trends are also observed in other datasets in K562 and in other cell lines. BARs, PRMs and DRMs have strong open chromatin signals (Figures 4A and 4B), consistent with their expected roles as active gene regulatory elements[21,23,42]. PRMs have stronger H3K4me3 signals and DRMs have stronger H3K4me1 (Figure 4C and 4E), which are expected since H3K4me3 is a signature of active promoters while H3K4me1 is an indicator of enhancers[43]. Both PRMs and DRMs have enriched H3K4me2 signals over the whole genome, which is also consistent with previous observations [40]. PRMs have stronger H3K36me3 and H3K79me2 (Additional file 2, Figure S8) signals than DRMs. These histone marks are found in transcribed regions[44-46], and are thus good features for distinguishing between regulatory elements that are close to and those that are far away from transcribed genes.

Correlation with Gene Expression. We calculated the average expression levels of TFs across 34 tissues[26]; highly connected TFs tend to be highly expressed. We further examined the relationship between connectivity and expression by calculating, for each TF, the correlation between its binding signal around its targets and

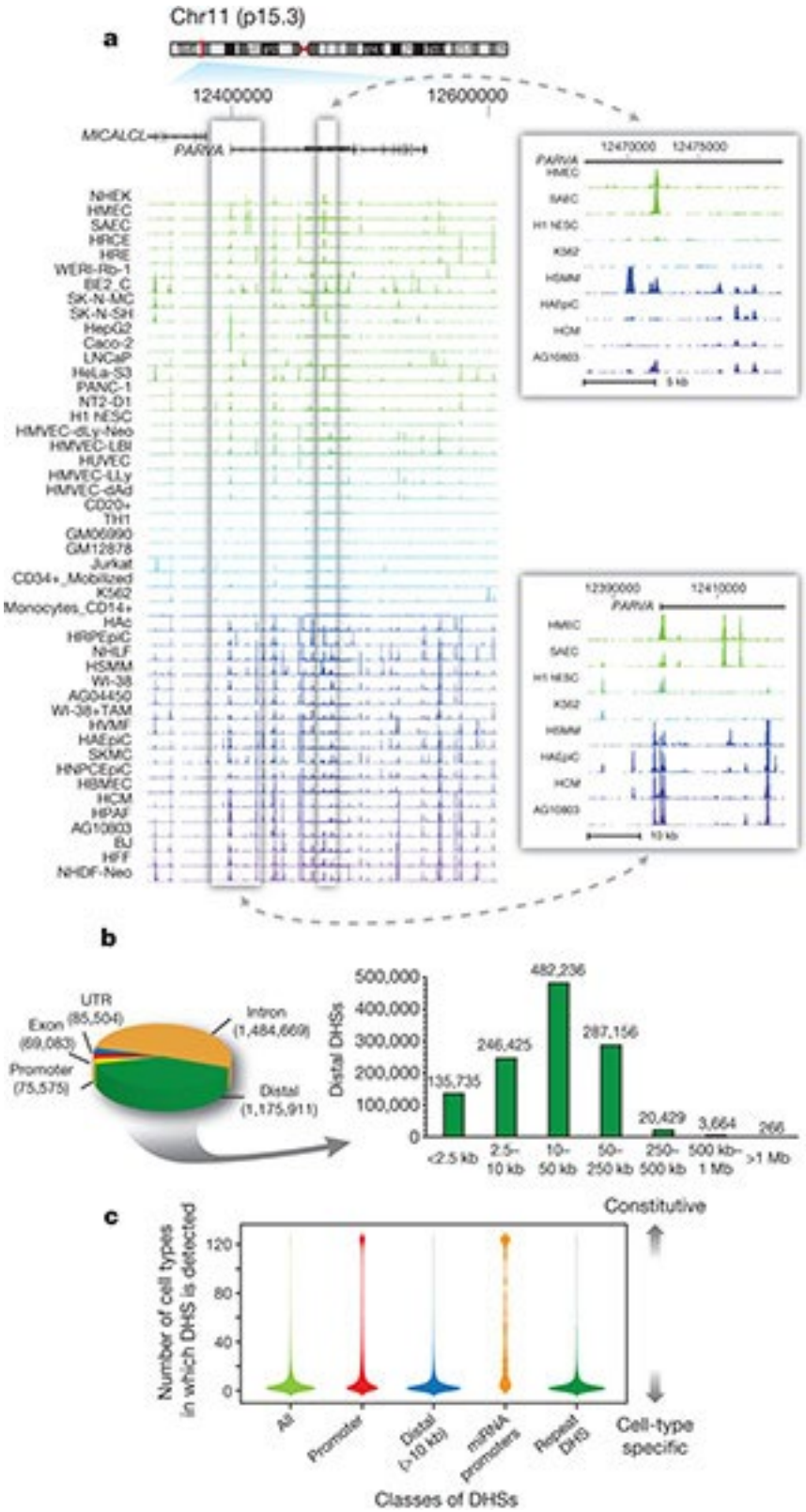



**Figure 1** | **General features of the DHS landscape.** (**a**) Density of DNase I cleavage sites for selected cell types, shown for an example ≥350-kb region. Two regions are shown to the right in greater detail. **(b)** Left: distribution of 2,890,742 DHSs with respect to GENCODE gene annotations. Promoter DHSs are defined as the first DHS localizing within 1 kb upstream of a GENCODE TSS. Right: distribution of intergenic DHSs relative to Gencode TSSs. (**c**) Distributions of the number of cell types, from 1 to 125 (*y* axis), in which DHSs in each of four classes (*x* axis) are observed. Width of each shape at a given *y* value shows the relative frequency of DHSs present in that number of cell types.

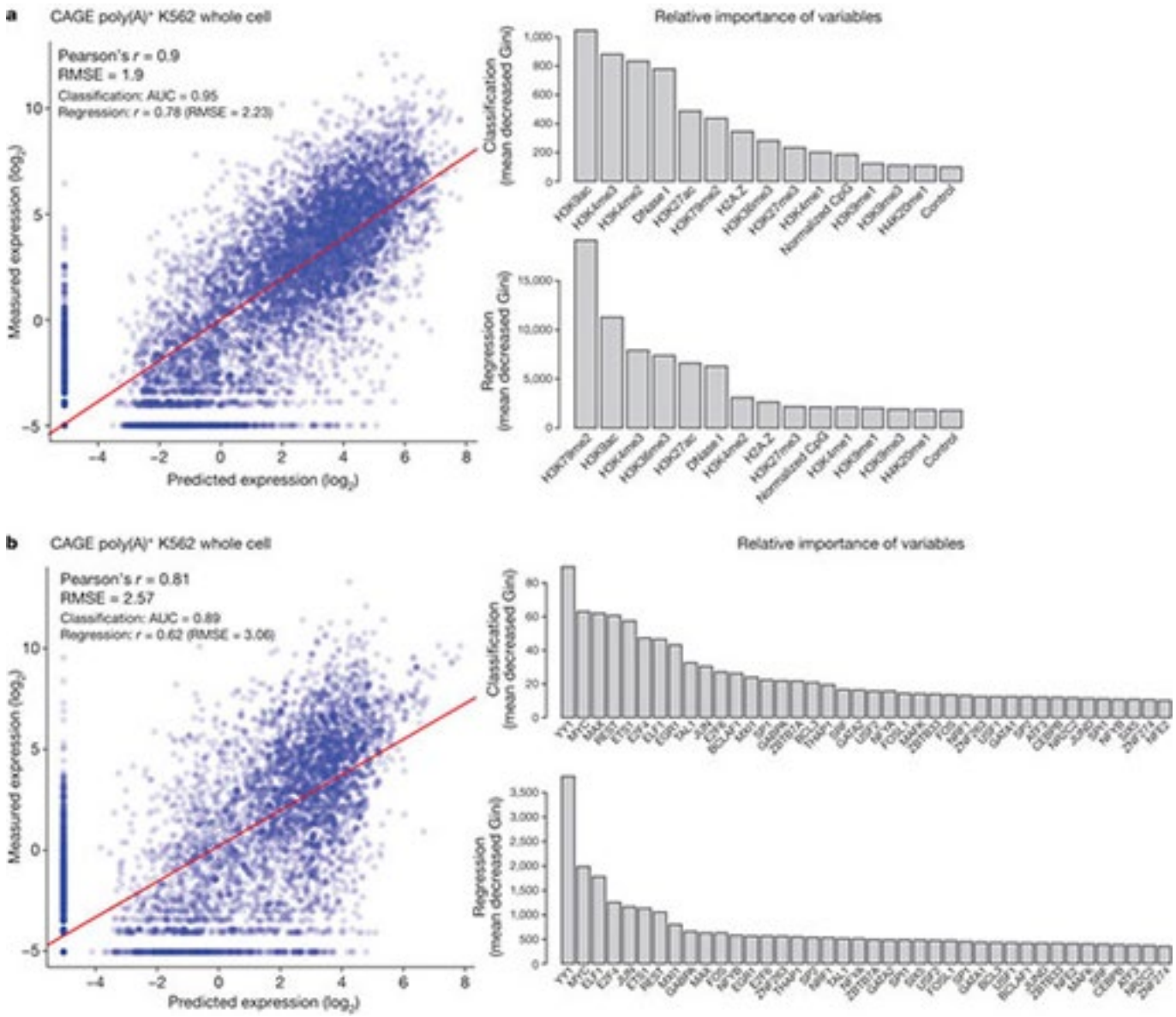



**Figure 2** | **Modelling transcription levels from histone modification and transcription-factor-binding patterns.** (**a**,**b**) Correlative models between either histone modifications or transcription factors, respectively, and RNA production as measured by CAGE tag density at TSSs in K562 cells. In each case the scatter plot shows the output of the correlation models (*x* axis) compared to observed values (*y* axis). The bar graphs show the most important histone modifications (**a**) or transcription factors (**b**) in both the initial classification phase (top bar graph) or the quantitative regression phase (bottom bar graph), with larger values indicating increasing importance of the variable in the model. Further analysis of other cell lines and RNA measurement types is reported elsewhere[59,79]. AUC, area under curve; Gini, Gini coefficient; RMSE, root mean square error.

the level of target expression (SOM/F.3.4). This binding-expression correlation is positively correlated with TF connectivity. Moreover, TFs at the top and middle levels exhibit a greater correlation. Thus, more "influential" TFs tend to be better connected and higher in the hierarchy. (This degree of "influence" becomes even clearer when one considers weighting the correlation by the number of TF targets, given that higher-level TFs tend to have more targets.) However, somewhat surprisingly, a model integrating the binding-expression relationships of all the highly connected TFs has about the same predictive power for expression as a model integrating all the less connected ones, indicating that the weak binding-expression relationships of the less influential TFs are collectively quite influential (SOM/F.3.4)[38].

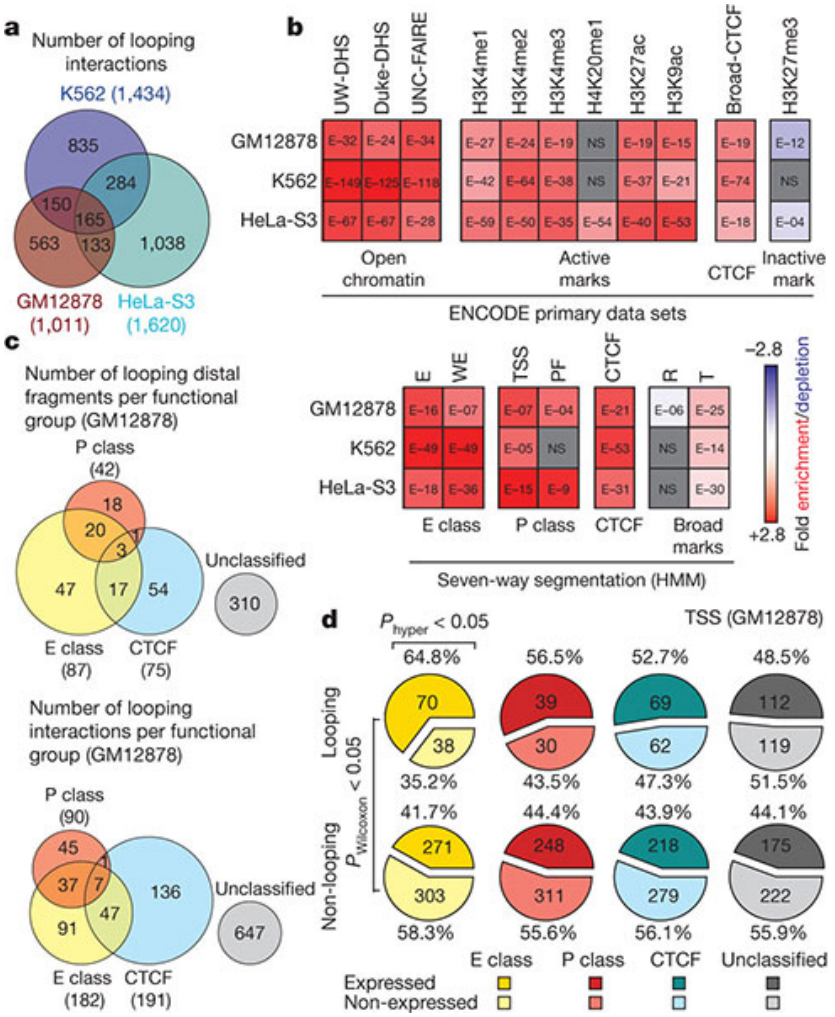



**Figure 2 | Distribution of looping interactions across cell types and their relationship with chromatin features and gene expression.** (**a**) Venn diagram showing the number of unique and overlapping looping interactions across three cell types. (**b**) Heat map showing the enrichment/depletion of chromatin features in looping fragments compared to all interrogated fragments based on genome-wide data sets from the ENCODE consortium (Supplementary Table 7). Features include open chromatin (UW-DHS (UW, University of Washington), Duke-DHS and UNC-FAIRE (UNC, University of North Carolina; FAIRE, formaldehyde-assisted isolation of regulatory elements)); active marks (Broad Institute histone H3K4me1/2/3, H4K20me1, H3K27ac, H3K9ac); CTCF (Broad Institute CTCF ChIP peaks); inactive marks (Broad Institute histone H3K27me3); and seven-way segmentation[4] (based on HMM prediction for indicated cells). We further grouped segmentation categories E and WE into 'E class', TSS and PF into 'P class', and R and T into 'broad marks'. The colour scale represents the fold enrichment (red) or depletion (blue). The numbers listed inside each box represent *P* values of the significant ($P < 0.05$) enrichment/depletion for that mark, where (for example) E−32 indicates $\times 10^{-32}$ (NS, not significant, grey; two-tailed hypergeometric test and corrected for multiple testing using Bonferroni). (**c**) Venn diagram showing the number of unique and overlapping looping distal fragments (top) and looping interactions (bottom) among four functional groups in GM12878 cells. Distal fragments are classified into four non-exclusive groups based on the seven-way segmentation. Similarly, TSS-distal fragment interactions are classified based on the functional grouping of the distal fragments. The four functional groups are E class (yellow), P class (magenta), CTCF (cyan) and unclassified (grey). **d**, Pie charts showing percentages and numbers of expressed/non-expressed TSSs looping or not looping to a particular group (E, P, CTCF or unclassified; coloured as in **c**) of distal fragments in GM12878 cells. TSSs with a CAGE value >0 are deemed expressed. Significant enrichment for expressed TSSs in the looping or non-looping categories is indicated on top (hypergeometric test; $P_{hyper} < 0.05$). Significant differences in expression levels between TSS in the looping versus the non-looping category is indicated on the left (Wilcoxon signed-rank test; $P_{Wilcoxon} < 0.05$).

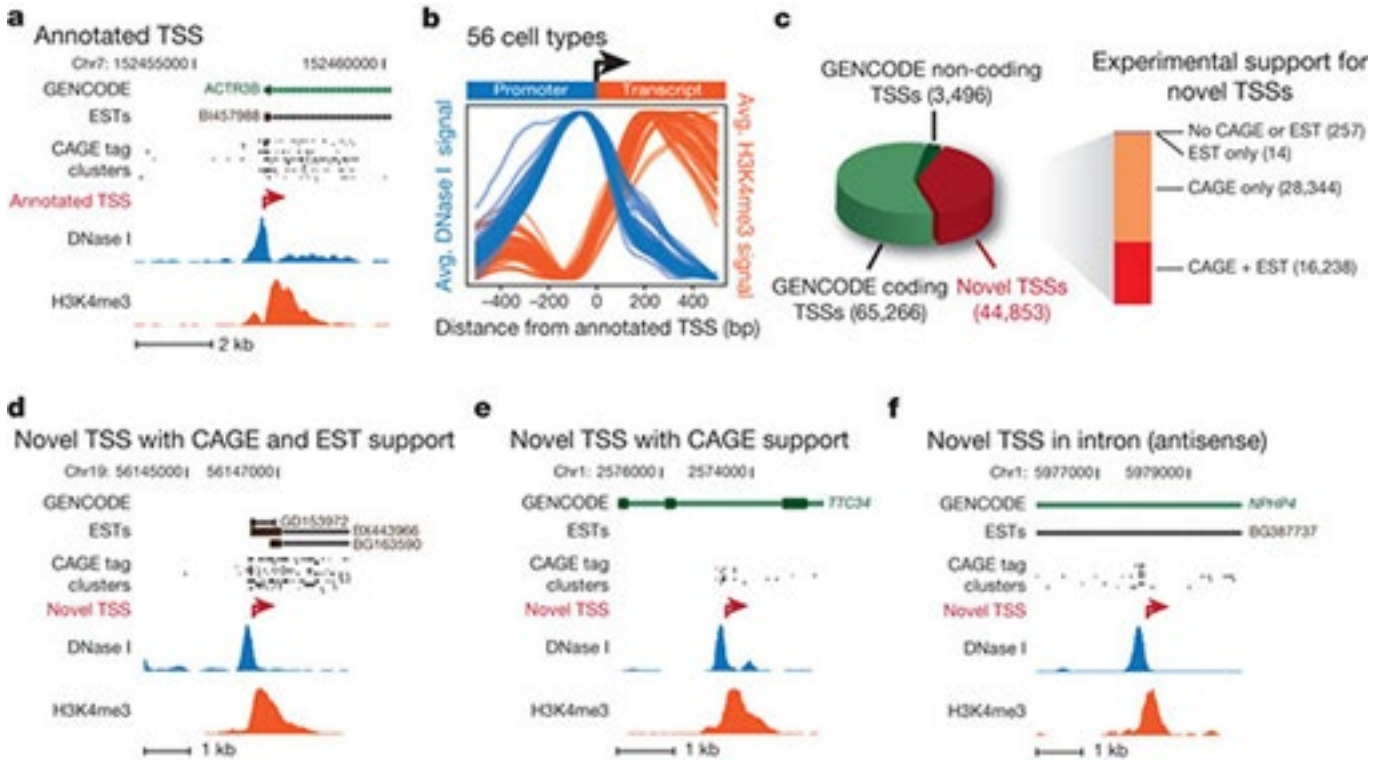



**Figure 3 | Identification and directional classification of novel promoters. (a) DNase I (blue) and H3K4me3 (red) tag densities for K562 cells around annotated TSS of *ACTR3B*. (b) Averaged H3K4me3 tag density (red, right *y* axis) and log DNase I tag density (blue, left *y* axis) across 10,000 randomly selected GENCODE TSSs, oriented 5′→3′. Each blue and red curve is for a different cell type, showing invariance of the pattern. (c) Relation of 113,615 promoter predictions to GENCODE annotations, with supporting EST and CAGE evidence (bar at right). (d-f) Examples of novel promoters identified in K562; red arrow marks predicted TSS and direction of transcription, with CAGE tag clusters, spliced ESTs and GENCODE annotations above. (d) Novel TSS confirmed by CAGE and ESTs. e, Novel TSS confirmed by CAGE, no ESTs. Note intronic location. (f) Antisense prediction within annotated gene.**

In this study, we aim at validating this result using data from CAGE that directly measures the expression levels of TSSs, and to investigate the influences of different technologies and RNA extraction methods on TSS expression quantification.We constructed models to quantify the ability of TF-binding signals to statistically predict the expression levels of promoters. Unless stated otherwise, we represent the binding strength of a TF in a promoter by its average ChIP-seq signal in a 100-bp region centered on the TSS. We combined the TSS expression data with TF-binding data and then divided them into a training data set and a test data set. A model was trained on the training data set and then applied to the test data to predict the expression levels of TSSs (see Methods for details). The relationship between expression and TF binding was quantified by the correlation between predicted and actual expression levels (R), or by the coefficient of determination ($R^2$), the percentage of variance of gene expression explained by the model. In order to evaluate the stability of our results, we built models using four different machine-learning methods: random forest (RF), support vector regression (SVR), multivariate adaptive regression splines (MARS), and multiple linear regression (MLR).

Previous studies used a mean signal of the TSS-flanking region ([-2k, +2k] around the TSS[20,10]) to estimate the level of histone modifications for a gene. However, this strategy could result in bias since modification marks have different density distributions along the gene[11]. For instance, H3K4me3 and H3K36me3 peak at 5' and 3' ends respectively[21]. To better estimate the representative signal for each chromatin feature, we divided specific genetic regions into bins following the approach by Cheng *et al.*[11] and searched for the bin(s) showing the best correlation between the chromatin feature signal and the expression level, namely 'bestbin'. The bestbin was determined using one-third of all genes (D1) and applied to the remaining two-third of genes (D2) for further analysis.

Additionally, instead of using a fixed bin for different chromatin features, we used the 'bestbin' strategy to capture maximal effects from different chromatin features. We have compared the performance of the

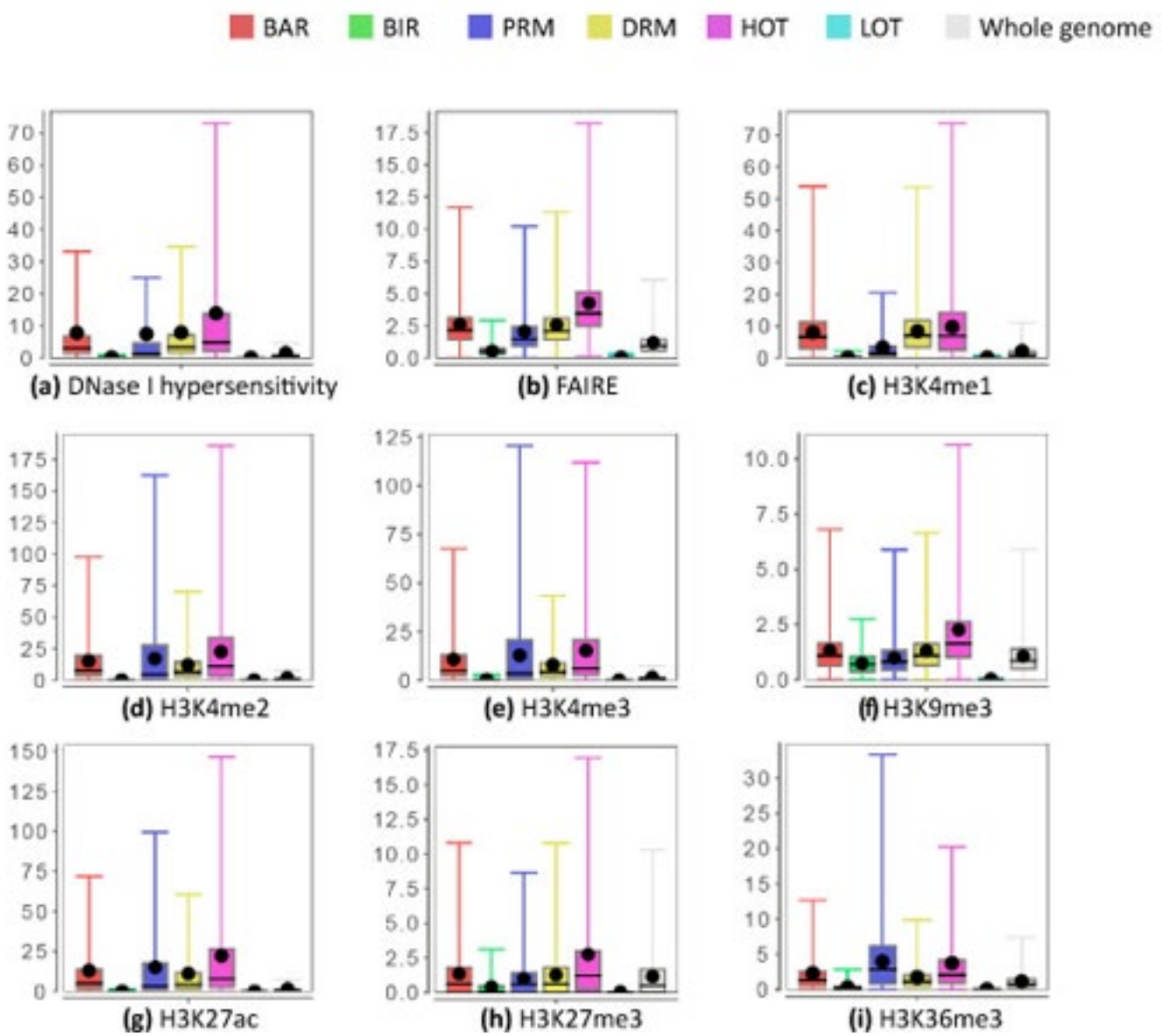
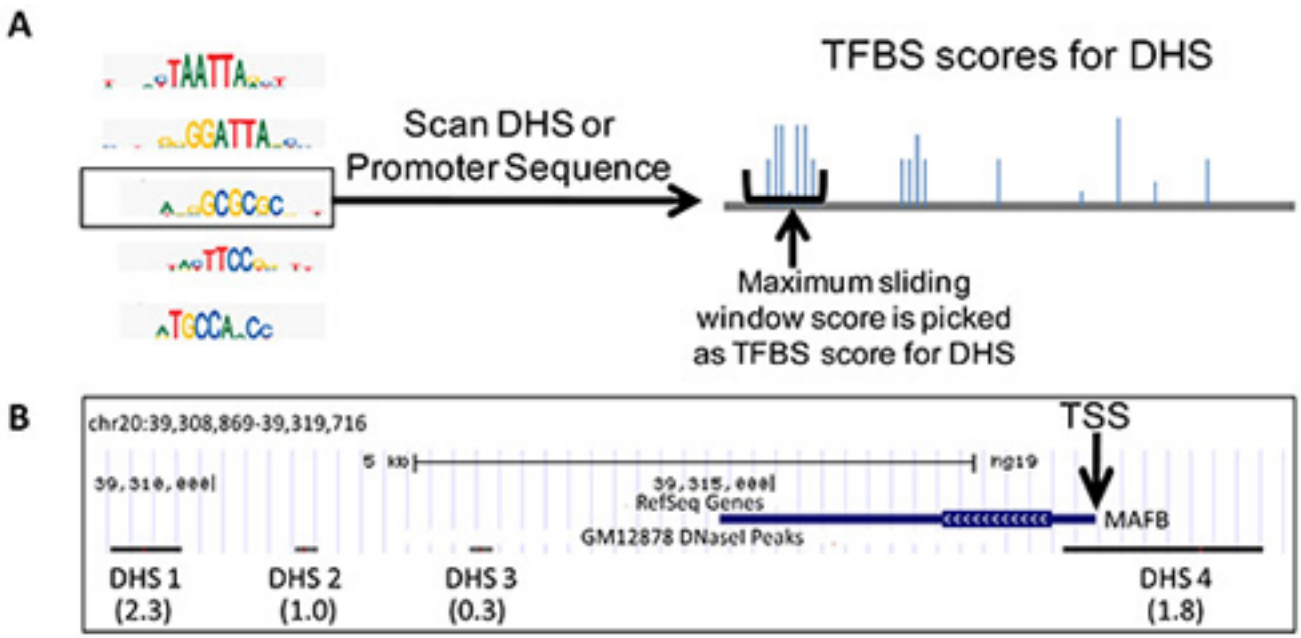



**Figure 4** | **Chromatin features of the six types of regions in K562.** (**a**) DNase I hypersensitivity from the dataset Uw.OpenChrom.K562.Dnase.Na (cf. Additional file 2, Figure S8E). (**b**) FAIRE signals from the dataset Unc.OpenChrom.K562.Faire.Na. (**c**) H3K4me1 signals from the dataset Broad.Histone.K562.H3K4me1.Std. (**d**) H3K4me2 signals from the dataset Broad.Histone.K562.H3K4me2.Std. (**e**) H3K4me3 signals from the dataset Broad.Histone.K562.H3K4me3.Std. (**f**) H3K9me3 signals from the dataset Broad.Histone.K562.H3k9me3. Std. (**g**) H3K27ac signals from the dataset Broad.Histone.K562.H3k27ac.Std. (**h**) H3K27me3 signals from the dataset Uw.Histone.K562.H3k27me3.Std. (**i**) H3K36me3 signals from the dataset Uw.Histone.K562.H3k36me3. Std. Each dataset ID has the format <Data source>.<Experiment type>.<Cell line>.<Open chromatin method/ histone modification/ TF>.<Experiment details>. The dot in each box-and-whisker plot is the average value. Some outlier values are not shown. See Materials and methods for details.

'bestbin' strategy with that of several other bin-selection methods. Table 1 shows that the 'bestbin' approach improves the performance by 2-13% compared to fixed-bin or no binning, and that overall, 'bestbin' has the best performance. Moreover, most chromatin marks show very stable 'bestbin', such as H3K36me3, DNase, H3K27me3, H4K20me1, and H3K9me1 (see Figure S9 in Additional file 2).

We used CAGE expression data[25] to assign an expression level to each TSS. We find that looping interactions with elements containing enhancer-like E elements are significantly enriched for those that involve expressed




| Closest Gene DHS | | Split DHS | |
|---|---|---|---|
| DHS associated with MAFB | TFBS score | DHS associated with MAFB | TFBS score |
| DHS 1, 2, 3, 4 | 2.3 | DHS 1, 2, 3 | 2.3 |
| | | DHS 4 | 1.8 |

**Figure 4 | Transcription factor binding site features. (a) DHS and promoter sequences are scanned with PWMs. TFBS scores are log-likelihood ratios of PWM over the background model. A sliding window is used to identify the score for each DHS or promoter. (b) Example to show association of DHSs with genes. Numbers in the brackets are example TFBS scores for the DHS for a specific DHS. Two methods of association were used. In closest gene DHS, DHSs 1-4 from the GM12878 cell line are associated with the gene *MAFB*. For the TF in consideration, the maximum of all TFBS scores is 2.3. In Split DHS, we separated DHSs overlapping the TSS and other DHSs. This resulted in two features for each gene for each TF.**

TSSs. (Figure 2d, Supplementary Figure 6). In addition, the subset of TSSs that interact with fragments containing E-elements were significantly more highly expressed compared to TSSs that do not interact with E-elements. Interactions with other classes of elements (CTCF, P, and Unclassified) are in some cell lines, but not all, significantly enriched for actively expressed genes (Supplementary Figure 6).

The nucleosome occupancy profile dips at the peak summits of most TFs (Fig. 5ab and Fig. S10), indicating that TFs prefer to bind nucleosome-depleted regions or that the binding of a TF excludes nucleosomes. In the vicinity of TSS-proximal summits, lower nucleosome occupancy is seen in the direction of transcription than upstream of transcription. We define nucleosome depletion as the amount that nucleosome occupancy dips at the peak summit, as compared with the nucleosome occupancy at 2 kb from the summit (considered as background). TSS-proximal summits show significantly greater nucleosome depletion than TSS-distal summits (Fig. 5c). It is well known that the binding of the transcriptional machinery to the TSS excludes nucleosomes to

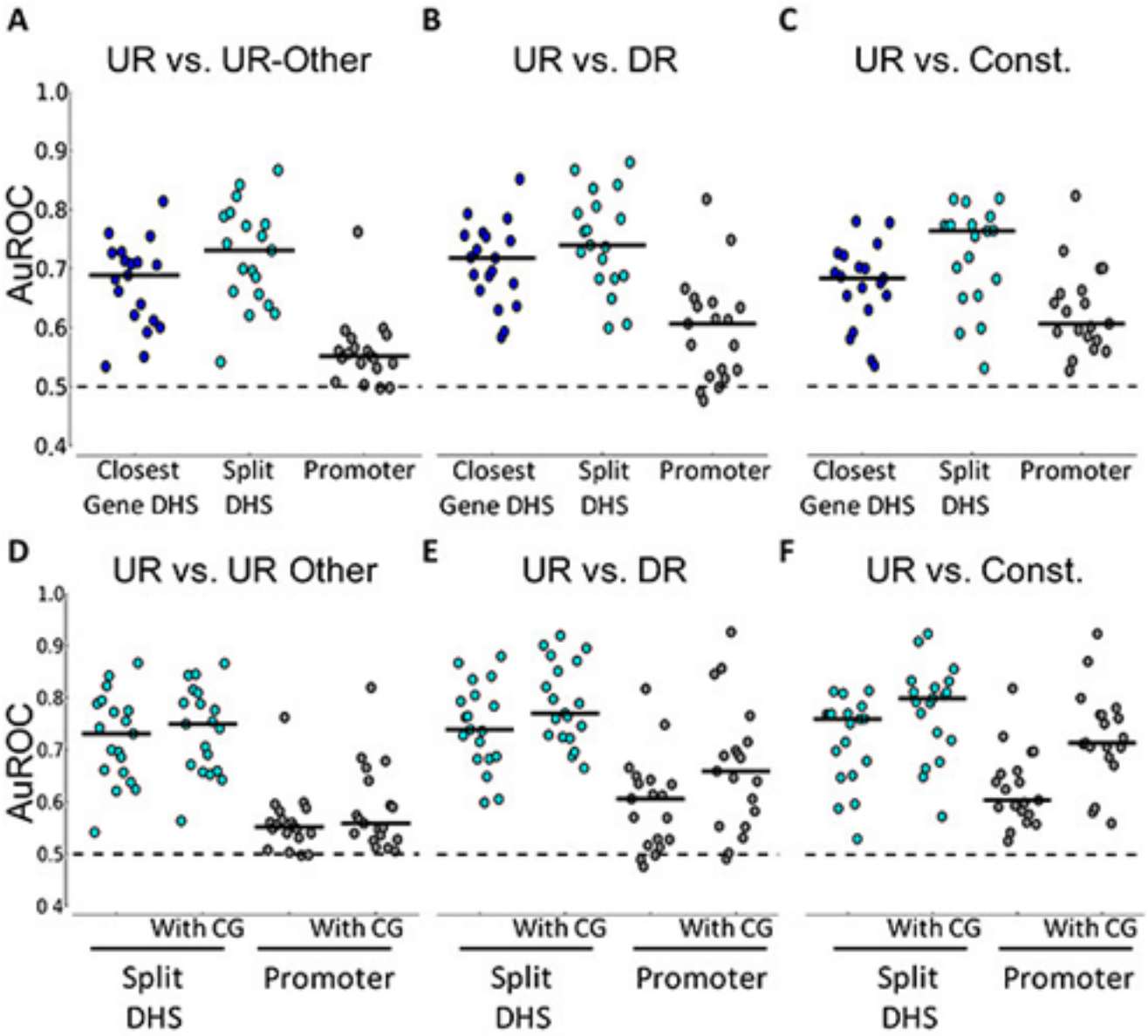
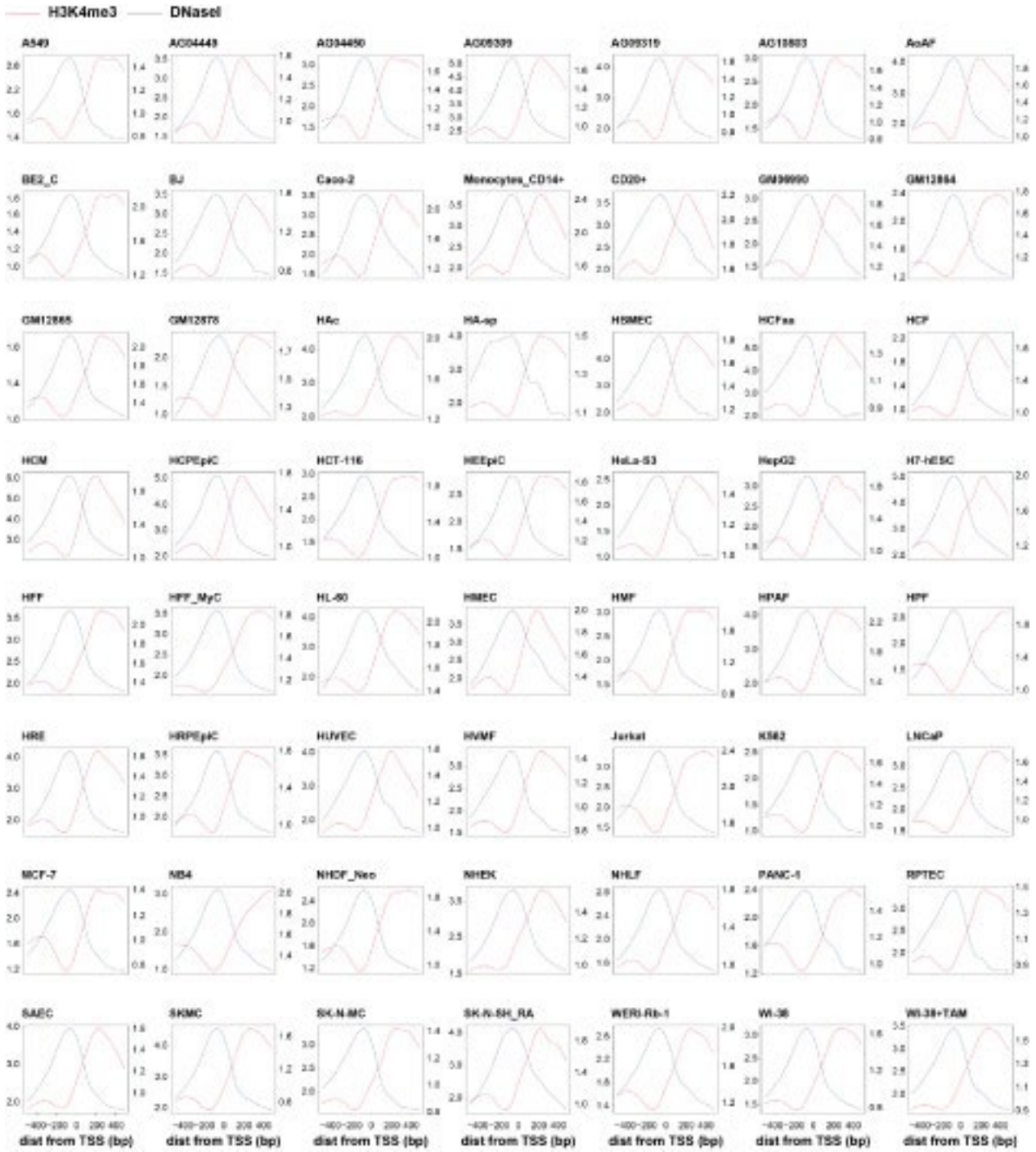


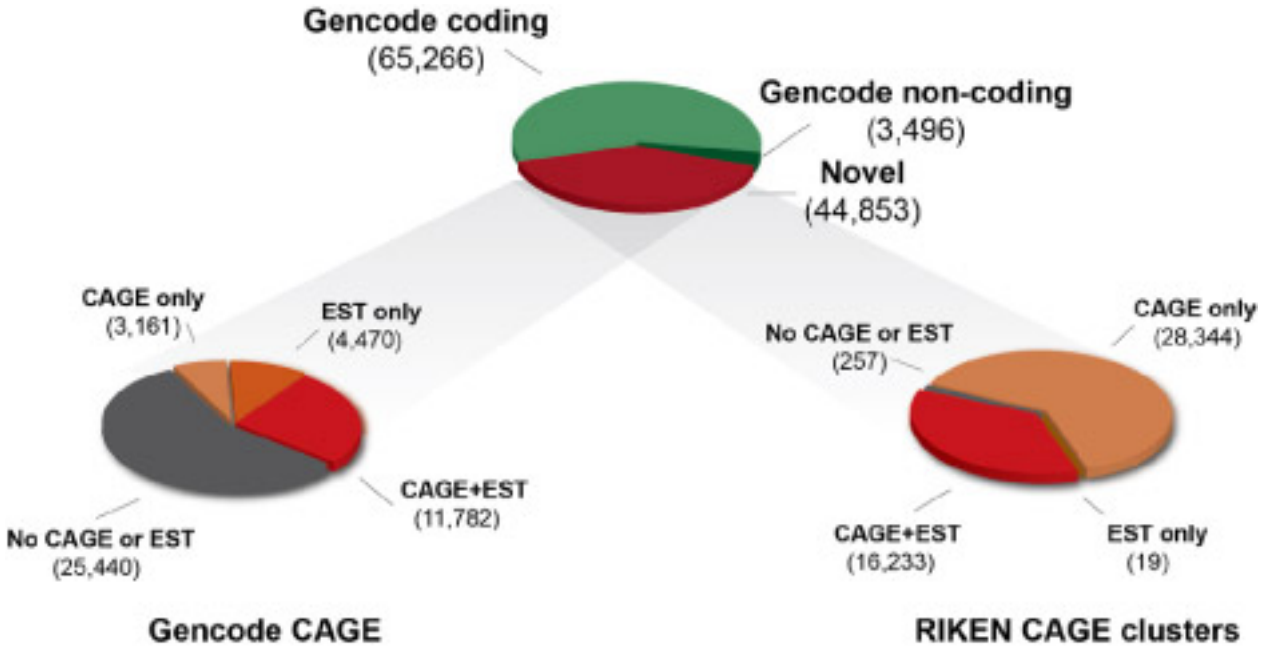

**Figure 5 | Classifier performance for various classification tasks. (a-c) Performance of the classifier using all PWMs. Each figure compares the performance of two methods of associating DHSs to genes (Closest Gene DHS and Split DHS) with the proximal promoter. The solid black lines across the dots indicate the median. Across all figures, the promoter sequence classifier does not perform as well as the performance achieved by using Closest Gene DHS and Split DHS and is significant at the 0.05 level (paired $t$-test). (d-f) Impact of normalized CG dinucleotide content on classifier performance. Results using the Split DHS and promoter sequence are shown. Without CG, columns are the same as in a-c. All figures show average results from five iterations of fourfold cross-validation. The dotted line indicates an AuROC of 0.5, which is the performance of a random classifier.**

a considerable extent (Radman-Livaja and Rando 2010). Indeed, average nucleosome occupancy anchored on the TSS shows an overall loss of nucleosomes (Fig. S12). Interestingly, we observed that TSS-proximal TF peak summits show a significantly greater depletion in nucleosome occupancy than do TSSs (Fig. 5d). The median nucleosome depletion at the summits of TSS-proximal peaks is 0.56 for GM12878 cells and 0.59 for K562 cells, significantly greater than the maximal nucleosome depletion around TSS (0.42 for GM12878 cells and 0.48 for K562 cells; Wilcoxon rank-sum test p-value=7.1e-28 and 1.1e-22 respectively). Within the proximal and distal categories, the top, middle, and bottom third peaks showed greatest, medium, and weakest nucleosome depletion respectively (Fig. 5c). This result indicates that TFs and nucleosomes compete for the genomic DNA and that stronger TF binding is correlated with greater nucleosome depletion, above and beyond the effect of transcription.




**Supplementary Figure 8 | DNaseI and H3K4me3 patterns around promoters in 56 cell types. This is the same as Fig. 3c, broken out for each of the 56 cell-types for which we have both DNaseI and H3K4me3 data, showing the stereotypical pattern of DNaseI and H3K4me3 around annotated promoters. Tag density for H3K4me3 (red) and log tag density for DNaseI (blue), averaged and centered across 10,000 randomly-selected Gencode v7 TSSs, oriented with respect to the transcription direction (gene body to the right). The *x*-axis is the distance in bp from the TSS. Left *y*-axis scale is for DNaseI; right *y*-axis scale is for H3K4me3.**

Approximately 3% ($n = 75{,}575$) of DHSs localize to transcriptional start sites (TSSs) defined by Gencode10 and 5% ($n = 135{,}735$, including the aforementioned) lie within 2.5 kilobases (kb) of a TSS. The remaining 95% of DHSs are positioned more distally, and are roughly evenly divided between intronic and intergenic regions (Fig. 1b). Promoters typically exhibit high accessibility across cell types, with the average promoter DHS detected in 29 cell types (Fig. 1c, second column). By contrast, distal DHSs are largely cell selective (Fig. 1c, third column).



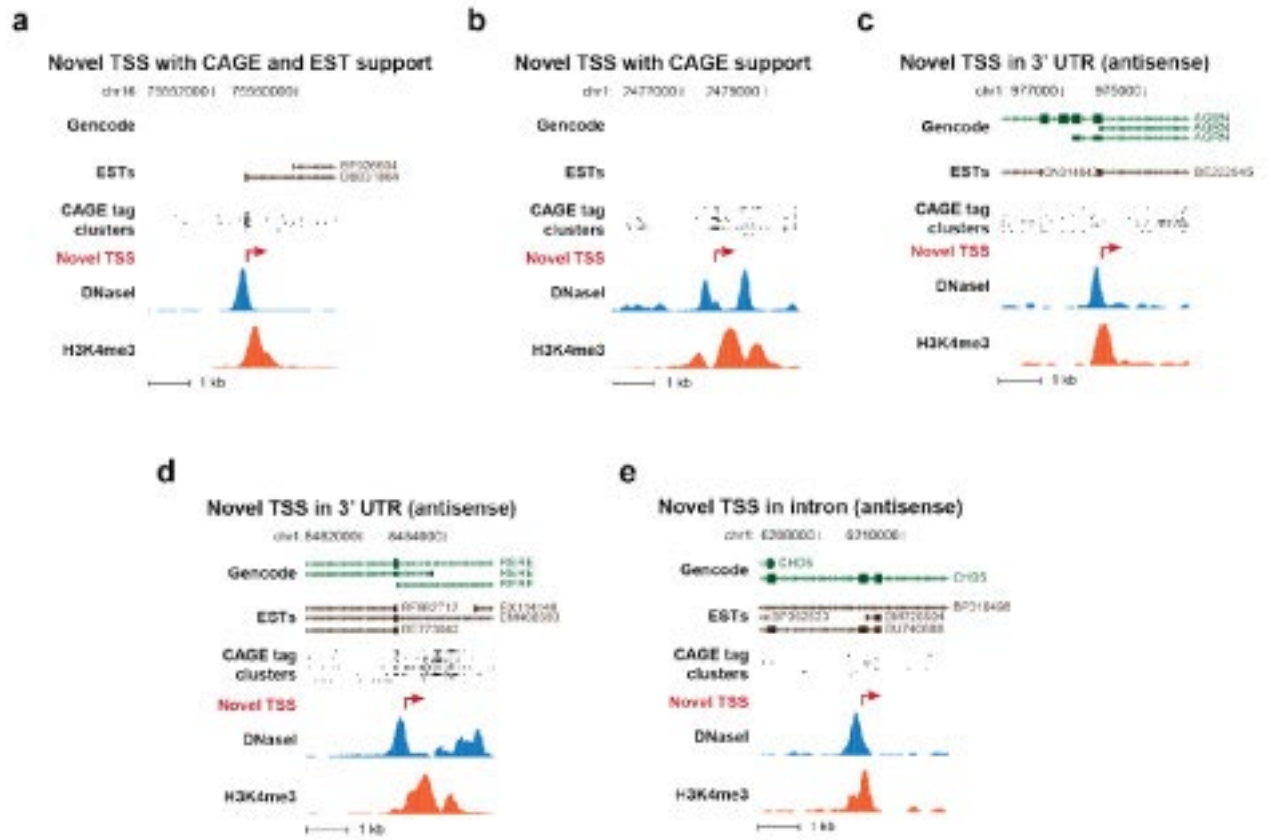

**Supplementary Figure 9 | Overlaps between novel promoters, CAGE clusters, and ESTs. This is a refinement of Fig. 3d. The top pie charts are identical in both figures. The bottom two pie charts here show the breakdown of novel promoter predictions with regard to their overlap separately with Gencode CAGE cluster TSS (left), and RIKEN CAGE cluster TSS (right), both of which datasets are described in the Supplementary Methods.**

The annotation of sites of transcription origination continues to be an active and fundamental endeavour[13]. In addition to direct evidence of TSSs provided by RNA transcripts, H3K4me3 modifications are closely linked with TSSs24. We therefore explored systematically the relationship between chromatin accessibility and H3K4me3 patterns at well-annotated promoters, its relationship to transcription origination, and its variability across ENCODE cell types.

We performed ChIP-seq for H3K4me3 in 56 cell types using the same biological samples used for DNaseI data (Supplementary Table 1, column D). Plotting DNaseI cleavage density against ChIP-seq tag density around TSSs reveals highly stereotyped, asymmetrical patterning of these chromatin features with a precise relationship to the TSS (Fig. 3a, b). This directional pattern is consistent with a rigidly positioned nucleosome immediately downstream from the promoter DHS, and is largely invariant across cell types (Fig. 3b and Supplementary Fig. 8).

To map novel promoters (and their directionality) not encompassed by the Gencode consensus annotations, we applied a pattern-matching approach to scan the genome across all 56 cell types (Supplementary Methods). Using this approach we identified a total of 113,622 distinct putative promoters. Of these, 68,769 correspond to previously annotated TSSs, and 44,853 represent novel predictions (versus Gencode v7). Of the novel sites, 99.5% are supported by evidence from spliced expressed sequence tags (ESTs) and/or cap analysis of gene expression (CAGE) tag clusters (Fig. 3c and Supplementary Fig. 9, $P > 0.0001$; see Supplementary Methods). We found novel sites in every configuration relative to existing annotations (Fig. 3d-f and Supplementary Fig. 10). For example, 29,203 putative promoters are contained in the bodies of annotated genes, of which 17,214 are oriented antisense to the annotated direction of transcription, and 2,794 lie immediately downstream of an annotated gene's 3' end, with 1,638 in antisense orientation. The results indicate that chromatin data can systematically inform RNA transcription analyses, and suggest the existence of a large pool of cell-selective transcriptional promoters, many of which lie in antisense orientations.

To predict gene expression patterns from sequence, approaches have frequently used features contained within fixed-size proximal promoter sequences. We used DHS data from a large number of cell types to determine whether using both proximal and distal regulatory regions with open chromatin would improve predictive




**Supplementary Figure 10 | Additional examples of novel promoters identified in K562 cells. Additional examples of novel promoters identified in K562 cells. (a) Novel prediction confirmed by CAGE and ESTs. (b) Novel prediction confirmed by CAGE annotation, no ESTs. (c), (d) Antisense promoter predictions at 3' end of annotated genes. (e) Antisense promoter prediction within Gencode-annotated genes.**

models for cell-type specific expression patterns. Position Weight Matrices (PWMs) for TFs in vertebrates were compiled from Transfac, JASPAR and UniProbe databases (Matys *et al.* 2006; Bryne *et al.* 2008; Newburger and Bulyk 2009). For each DHS, 789 PWMs were used to calculate TFBS scores that accounted for local dinucleotide composition. The maximum sliding window score for each PWM was used as the TFBS score for that DHS (Figure 4A). To associate DHS with specific genes that they are likely to regulate, we applied a simple approach of associating each DHS with the closest TSS (closest gene DHS). For each TF, we then chose the maximum TFBS score across all DHS associated with a gene (Figure 4B). As an alternative approach, we split DHS into distal sites (a set including both Gene-Body and Intergenic DHS) and TSS DHS sites and used the maximum TFBS in each set as individual features (split DHS). This doubled the number of features and allowed us to identify different characteristics of TSS-overlapping vs. distal DHS. To compare our models to previous approaches, we also used TFBS features calculated in proximal promoters, defined here as -900 to +100 nucleotides surrounding the TSS (Landolin *et al.* 2010).

We used the TFBS scores as features for sparse logistic regression classifiers to discriminate between different gene classes. These classifiers balance the use of many available features against model complexity, effectively selecting a small subset of informative features which are used in the classification. We trained cell-type specific classifiers on the task to discern whether a gene belonged to a specific expression pattern (e.g., UR vs. UR-Other, UR vs. DR, UR vs. constitutive, etc.). The area under the Receiver Operating Characteristic curve (AuROC) metric was used to evaluate the performance of a model, where a value of 0.5 indicates random assignments and 1.0 indicates perfect classification (see Methods). To not bias results due to different amounts of training data, the positive sets of up- and down-regulated genes were all of the same size.

The performance of the classifier using only proximal promoter information is close to that of a random classifier, across all tasks. All the classifiers using DHS sequences display strong improvements in performance over this baseline in discriminating genes that are up-regulated in different cell types (UR vs. UR-Other, Figure 5A), with a greater improvement in performance coming from the Split DHS approach with separate features for the TSS and Distal DHSs (median AuROC ~0.73). Similar results were obtained when training classifiers to distinguish between specifically up- and down-regulated genes from the same cell types (UR vs. DR, Figure 5B), and to distinguish up-regulated from constitutively expressed genes (UR vs. Const., Figure 5C). Discriminating down-regulated genes from different cell types (DR vs. DROther), and down-regulated from constitutively expressed genes (DR vs. Const.), resulted in lower accuracies but still showed the trend of better performance with DHS compared to proximal promoter sequence (Supplemental Figure 2A-B). All results clearly indicate that strong performance improvement is achieved by scanning for TFBS matches in open chromatin regions.

**Identifying candidate regulators**

In addition to classifying genes belonging to different groups, we inspected the classifiers to identify motifs that were most informative in the classification task, i.e., those PWMs that had large regression coefficients (Supplemental Table 4). This identified several TFs with known impact on transcriptional output in the cell line of interest. For example, YY1, SPI1 and IRF8 are crucial in the specification of B-cells (GM12878 cell line) (Lu *et al.* 2003; Liu *et al.* 2007; Sokalski *et al.* 2011). We also identified the REST motif as a positive regulator of UR genes in medulloblastoma cell line that is of neural origin (Supplemental Table 6). REST specifically down-regulates neuron-specific genes in many non-neuronal cell lines, and its expression is suppressed in neurons (Schoenherr and Anderson 1995). As a result, the model identified the ciselements that are present in the DHS associated with neuron specific genes as the factor that separates these genes from the genes up-regulated elsewhere. This example illustrates that the inactivation of a repressor can also explain up-regulation of genes. Other well characterized factors included ETS1 in HUVEC cells and HNF4A for HepG2 cells (Cereghini 1996; Oda *et al.* 1999; Yordy *et al.* 2005).

The feature set described thus far was comprehensive in that it used available PWM information from multiple sources, independent of the expression levels of transcription factors or the potential redundancy of features. To assess how much celltype specific regulation can be explained by the cell-type specific expression of transcription factors themselves, we selected the top 10 TFs with highest absolute zscores from each cell line and had PWMs that were not similar to each other (Supplementary Table 7).

Our results indicate that TF binding signals around the TSS are informative for "predicting" their expression levels. For example, Figure 1A shows the consistency between predicted and actual expression levels of TSSs measured by CAGE of whole cell Poly A+ RNA in K562 cells. TF binding accounts for at least 67% of the variance of expression levels ($R^2$=0.67). In total, there are 267 promoter expression profiles representing 12 different human cell lines in our dataset. The performance of the model is not directly comparable between cell lines, because different numbers of TF binding datasets are available for different cell lines. Since the most complete data were from K562, we chose this cell line for further analysis. The expression levels of a large fraction of TSSs (~50% on average) are not detected (RPKM=0) in any of these K562 datasets. Thus, we developed a more complicated model that first classifies TSSs into expressed and non-expressed categories and then adopts a regression model to predict the expression levels for the expressed TSSs only (The-ENCODE-Consortium 2012). When applied to the TF data, this model achieves results very consistent with the methods without a classification step in terms of the R2 value and the relative importance of different TFs. We therefore focus on the classification-free models in the rest of this analysis.

For each of the 40 TFSSs assayed in K562, we investigated its individual predictive power in a degenerate model that uses this TF as a single predictor (Figure 2B). Strikingly, each TF alone can predict TSS expression levels of all genes with fairly high accuracy. As shown, the binding signal of MAX alone can explain 55% of the variance

in expression of all TSS, which is only ~12% lower than the variance explained by the full model (67%). The R2 in a degenerate model indicates the power of a TF for predicting expression individually. In the full model, the relative importance of TFs for predicting the expression levels of promoters is roughly reflected by their Relative Importance score (RI score, see "Methods") (Figure 2C). We use the standard RI metrics of different machine learning methods, which indicate the contribution of TFs after considering their inter-correlations in a model, and thus provide complementary information to the individual predictive power. Specifically, in a random forest model the RI of a TF is calculated as the increase of prediction error (%IncMSE) when binding data for this TF is permuted. In general, highly predictive TFs have more binding peaks, particularly in the TSS proximal regions. We found in the full model that the top five most important TFs in K562 are YY1, E2F4, MYC, MAX and ELF1. We also examined the effect of TF-TF interaction on the predictive accuracy. Our results indicated that including interaction terms in the model did not lead to further improvement.

We next examined the effectiveness of predicting differential expression based on differential binding of TFs in promoter regions. The binding differences (log2) in K562 versus GM12878 were calculated for 22 TFs for which the ChIP-seq data were available in both cell lines. A model using those differences as predictors explains 53% of the variance in expression differences (log2 ratios) of TSSs between K562 and GM12878 (whole cell Poly A+ RNA extraction) (Figure 5B). We also explored the relative importance of TFs in the differential expression model. Interestingly, we find that the TFs important for differential expression (e.g YY1) are in general those that are important in both the K-model and the G-model. TFs with higher RI scores in only one cell line (e.g. SP1, MAX and ETS1) show quite limited contributions to predicting differential expression of promoters (Figure 5C).

We find that histone modification can be predicted accurately by the binding signals of TFs at the TSS regions. As shown in Figure 6, the TF binding signal at the TSS of genes can predict H3K4me3 signals around the TSS with very high accuracy ($R^2$= 0.85). It is also highly predictive of the signals of other histone marks, such as H3K9ac and H3K79me3 (see Supplementary Figure S3). More interestingly, the TF binding signals can predict the patterns of histone marks, i.e. the positions where they are located. For example, the best prediction accuracy was achieved right at the TSS for H3K4me3, which is known to be a mark for active promoters (Koch *et al.* 2007). In contrast, high predictive accuracy was obtained at the TSS and in the transcribed region of genes for H3K36me3, which is a histone mark for the gene body (Kolasinska-Zwierz *et al.* 2009). The relative importance of TFs is different for predicting different histone modification types, but MAX, YY1, ETS1 and E2F6 are generally the most informative ones (see Supplementary Figure S4 and Supplementary Table S3).

After considering the histone modification data, binding of TFSS accounts for a further 13% of additional variance in gene expression levels (the TFSS|HM model), and 8% vice versa (the HM|TFSS model). This suggests that the contributions of TFSS binding and histone modification to aggregate expression of TSS are highly but not completely redundant. Each provides extra information that is not accounted for by the other. We note that here we only use histone modification signals at the TSS regions (100bp). Since histone modifications affect a broad region around genes, the actual variance that can be explained by the HM model should be even larger (Cheng *et al.* 2011b; Dong *et al.* 2012).

**Promoter-proximal regulatory modules (PRMs) and gene-distal regulatory modules (DRMs)**

Among the TRF binding sites, one subset of particular interest is the ones close to the TSSs of active genes, as they are likely actively involved in the regulation of these genes in the corresponding cell lines. Depending on the distance from a transcription start site, these regions may contain core promoters and proximal promoter elements[2]. We call these regions promoter-proximal regulatory modules (PRMs) in general. To define PRMs, instead of using an arbitrary distance threshold from TSSs, we determined distance cutoffs according to chromatin feature patterns using a machine learning framework. Specifically, for each cell line, we took TSSs of genes expressed in the cell line as positive examples, and random non-TRF binding sites and distal TRF binding sites as negative examples (Materials and methods). Expression of TSSs was determined by ENCODE data from

Cap-Analysis of Gene Expression (CAGE)[27], Paired-End diTag (PET)[28], and RNA sequencing (RNA-seq)[29-30]. Based on the examples, a discriminative model was learned using chromatin features and TRF binding data of the cell line as explanatory variables. The resulting models separated positive and negative examples well in all cell lines (Additional file 2, Figure S3 and Additional file 2, Figure S4). Finally we used the learned models to give PRM scores to all regions in the whole genome. Since in this case we have a relatively complete set of positive examples from annotated genes, we used a more stringent threshold to call PRMs (Materials and methods).

EP300, which is found at some enhancers[58], has a slight enrichment at DRMs. The same trend is also observed for GATA1 and GATA2 (Figure 5D and Additional file 2, Figure S8), which were reported to enhance the expression of some genes[59-60]. In comparison, some TRFs (such as E2F4) are strongly enriched at PRMS as compared to DRMs, and some (such as USF2) have almost the same enrichment at PRMs and DRMs.