11 Machine learning approaches to genomics

ENCODE has applied machine learning approaches to enable integration and exploration of large and diverse data

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.

For discriminative training, we used a three-step process to predict potential enhancers, described in Supplementary Info and ref[68]. Two alternative discriminative models converged on a set of ~13,000 putative enhancers in K562 cells[68]. In the second approach, two methodologically distinct unbiased approaches (see ref [40,69] and Hoffmann *et al.,* manuscript in preparation) converged on a concordant set of histone modification and chromatin-accessibility patterns that can be used to segment the genome in each of the Tier 1 and Tier 2 cell lines, although the individual loci in each state in each cell line are different. With the exception of RNA polymerase II and CTCF, the addition of TF data did not substantially alter these patterns. At this stage, we deliberately excluded RNA and methylation assays, reserving these data as a means to validate the segmentations.

Our integration of the two segmentation methods (Hoffmann *et al.,* manuscript in preparation) established a consensus set of seven major classes of genome states, described in Table 3. The standard view of active promoters, with a distinct core promoter region (TSS and PF states), leading to active gene bodies (T, transcribed state) is rediscovered in this model (Figure 5A and B). There are three "active" distal states. We tentatively labelled two as enhancers (predicted enhancers, E, and predicted weak enhancers, WE) due to their occurrence

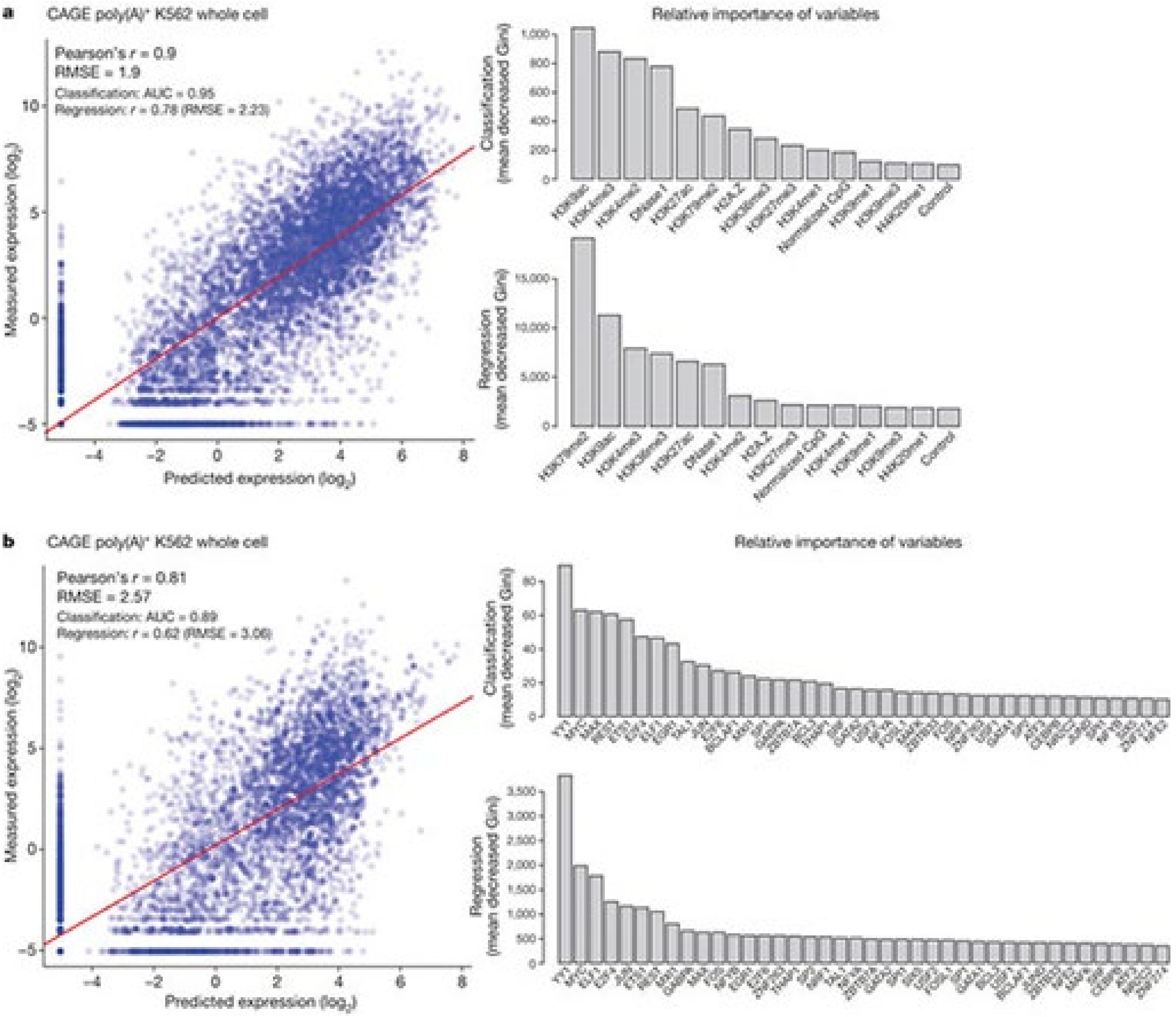


Figure 2 | Modelling transcription levels from histone modification and transcription-factor-binding patterns. (**a**,**b**) Correlative models between either histone modifications or transcription factors, respectively, and RNA production as measured by CAGE tag density at TSSs in K562 cells. In each 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.

in regions of open chromatin with high H3K4me1, although they differ in the levels of marks such as H3K27ac, currently thought to distinguish active from inactive enhancers. The other active state (CTCF) has high CTCF binding and includes sequences that function as insulators in a transfection assay. The remaining repressed state (R) summarises sequences split between different classes of actively repressed or inactive, quiescent chromatin. We found that the CTCF-binding associated state is relatively invariant across cell types, with individual regions frequently occupying the CTCF state across all six cell types (Figure 5C). Conversely, the E and T states have substantial cell-specific behaviour, whereas the TSS state has a bimodal behaviour with similar

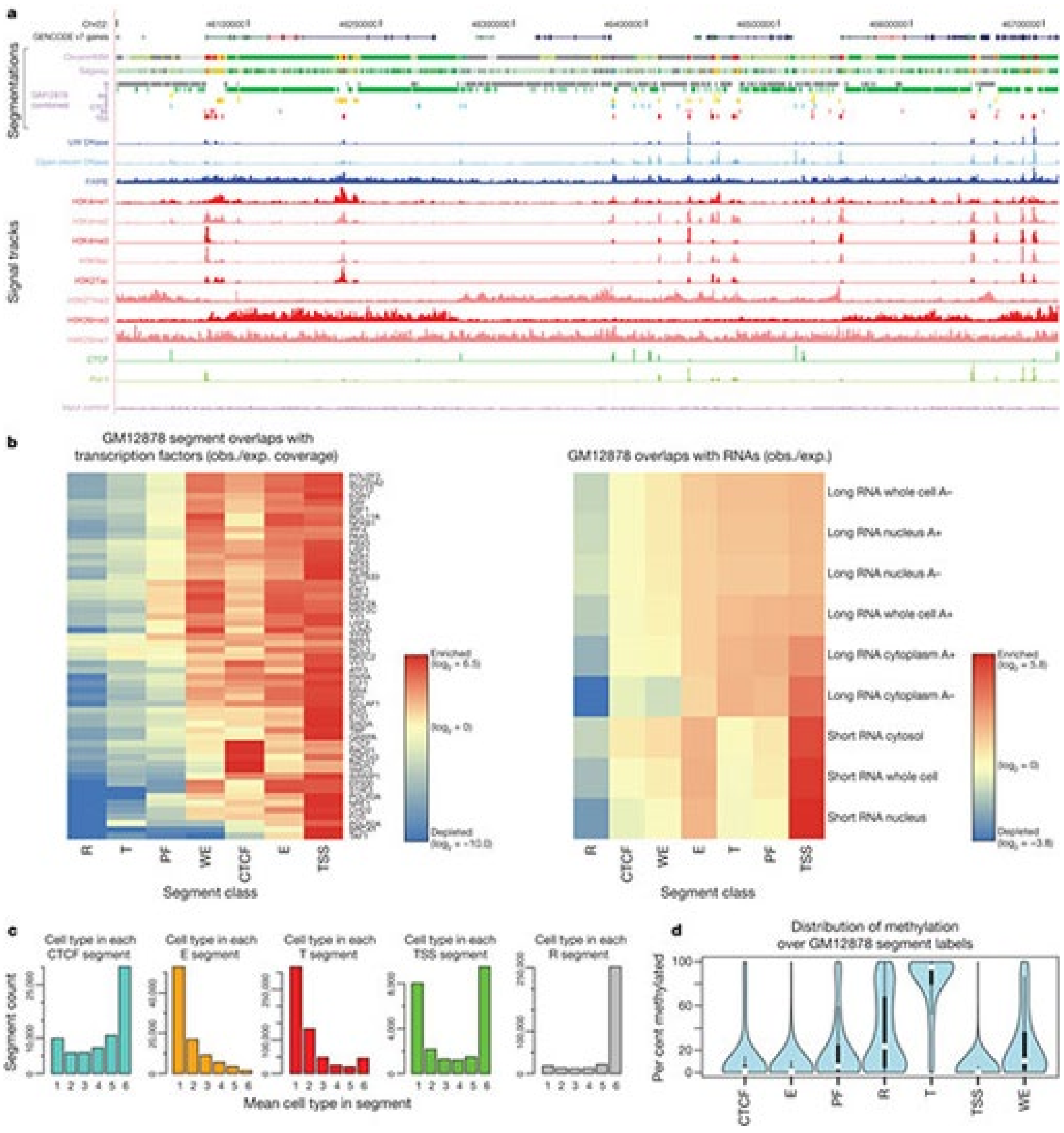


Figure 5 | Integration of ENCODE data by genome-wide segmentation. (**a**) Illustrative region with the two segmentation methods (ChromHMM and Segway) in a dense view and the combined segmentation expanded to show each state in GM12878 cells, beneath a compressed view of the GENCODE gene annotations. Note that at this level of zoom and genome browser resolution, some segments appear to overlap although they do not. Segmentation classes are named and coloured according to the scheme in Table 3. Beneath the segmentations are shown each of the normalized signals that were used as the input data for the segmentations. Open chromatin signals from DNase-seq from the University of Washington group (UW DNase) or the ENCODE open chromatin group (Openchrom DNase) and FAIRE assays are shown in blue; signal from histone modification ChIP-seq in red; and transcription factor ChIP-seq signal for Pol II and CTCF in green. The mauve ChIP-seq control signal (input control) at the bottom was also included as an input to the segmentation. (**b**) Association of selected transcription factor (left) and RNA (right) elements in the combined segmentation states (*x* axis) expressed as an observed/expected ratio (obs./exp.) for each combination of transcription factor or RNA element and segmentation class using the heat-map scale shown in the key besides each heat map. (**c**) Variability of states between cell lines, showing the distribution of occurrences of the state in the six cell lines at specific genome locations: from unique to one cell line to ubiquitous in all six cell lines for five states (CTCF, E, T, TSS and R). (**d**) Distribution of methylation level at individual sites from RRBS analysis in GM12878 cells across the different states, showing the expected hypomethylation at TSSs and hypermethylation of genes bodies (T state) and repressed (R) regions.

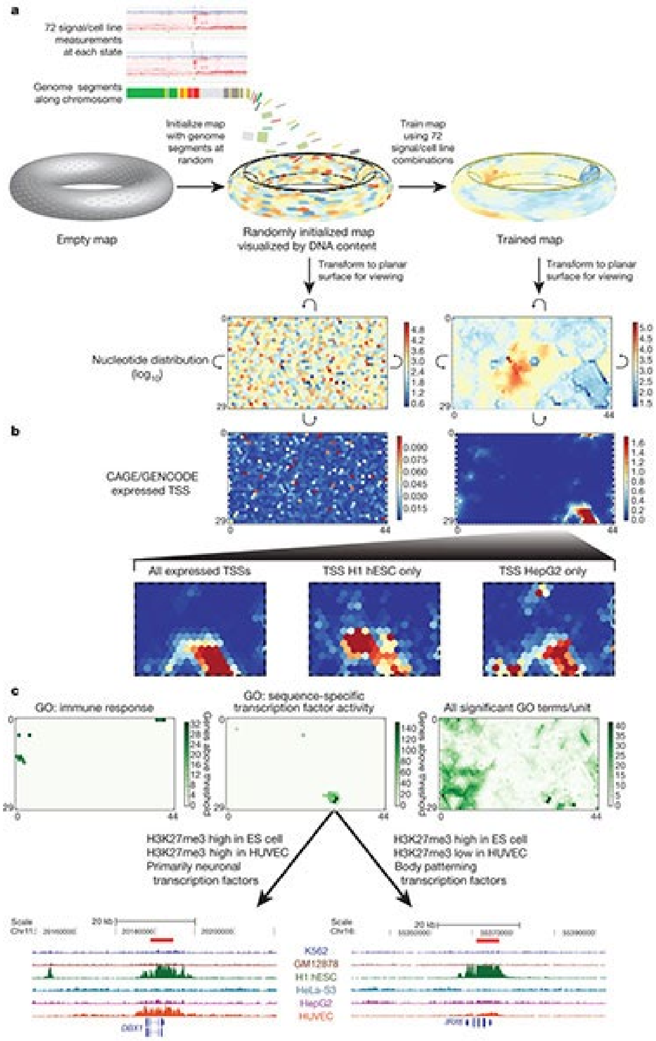


Figure 7 | High-resolution segmentation of ENCODE data by self-organizing maps (SOM). (a-c) The training of the SOM (a) and analysis of the results (b, c) are shown. Initially we arbitrarily placed genomic segments from the ChromHMM segmentation on to the toroidal map surface, although the SOM does not use the ChromHMM state assignments (a). We then trained the map using the signal of the 12 different ChIP-seq and DNase-seq assays in the six cell types analysed. Each unit of the SOM is represented here by a hexagonal cell in a planar two-dimensional view of the toroidal map. Curved arrows indicate that traversing the edges of two dimensional view leads back to the opposite edge. The resulting map can be overlaid with any class of ENCODE or other data to view the distribution of that data within this high-resolution segmentation. In panel a the distributions of genome bases across the untrained and trained map (left and right, respectively) are shown using heat-map colours for log$_{10}$ values. (b) The distribution of TSSs from CAGE experiments of GENCODE annotation on the planar representations of either the initial random organization (left) or the final trained SOM (right) using heat maps coloured according to the accompanying scales. The bottom half of b expands the different distributions in the SOM for all expressed TSSs (left) or TSSs specifically expressed in two example cell lines, H1 hESC (centre) and HepG2 (right). (c) The association of Gene Ontology (GO) terms on the same representation of the same trained SOM. We assigned genes that are within 20 kb of a genomic segment in a SOM unit to that unit, and then associated this set of genes with GO terms using a hypergeometric distribution after correcting for multiple testing. Map units that are significantly associated to GO terms are coloured green, with increasing strength of colour reflecting increasing numbers of genes significantly associated with the GO terms for either immune response (left) or sequence-specific transcription factor activity (centre). In each case, specific SOM units show association with these terms. The right-hand panel shows the distribution on the same SOM of all significantly associated GO terms, now colouring by GO term count per SOM unit. For sequence-specific transcription factor activity, two example genomic regions are extracted at the bottom of panel (c) from neighbouring SOM units. These are regions around the *DBX1* (from SOM unit 26,31, left panel) and *IRX6* (SOM unit 27,30, right panel) genes, respectively, along with their H3K27me3 ChIP-seq signal for each of the tier 1 and 2 cell types. For *DBX1*, representative of a set of primarily neuronal transcription factors associated with unit 26,31, there is a repressive H3K27me3 signal in both H1 hESCs and HUVECs; for *IRX6*, representative of a set of body patterning transcription factors associated with SOM unit 27,30, the repressive mark is restricted largely to the embryonic stem (ES) cell. An interactive version of this figure is available in the online version of the paper.

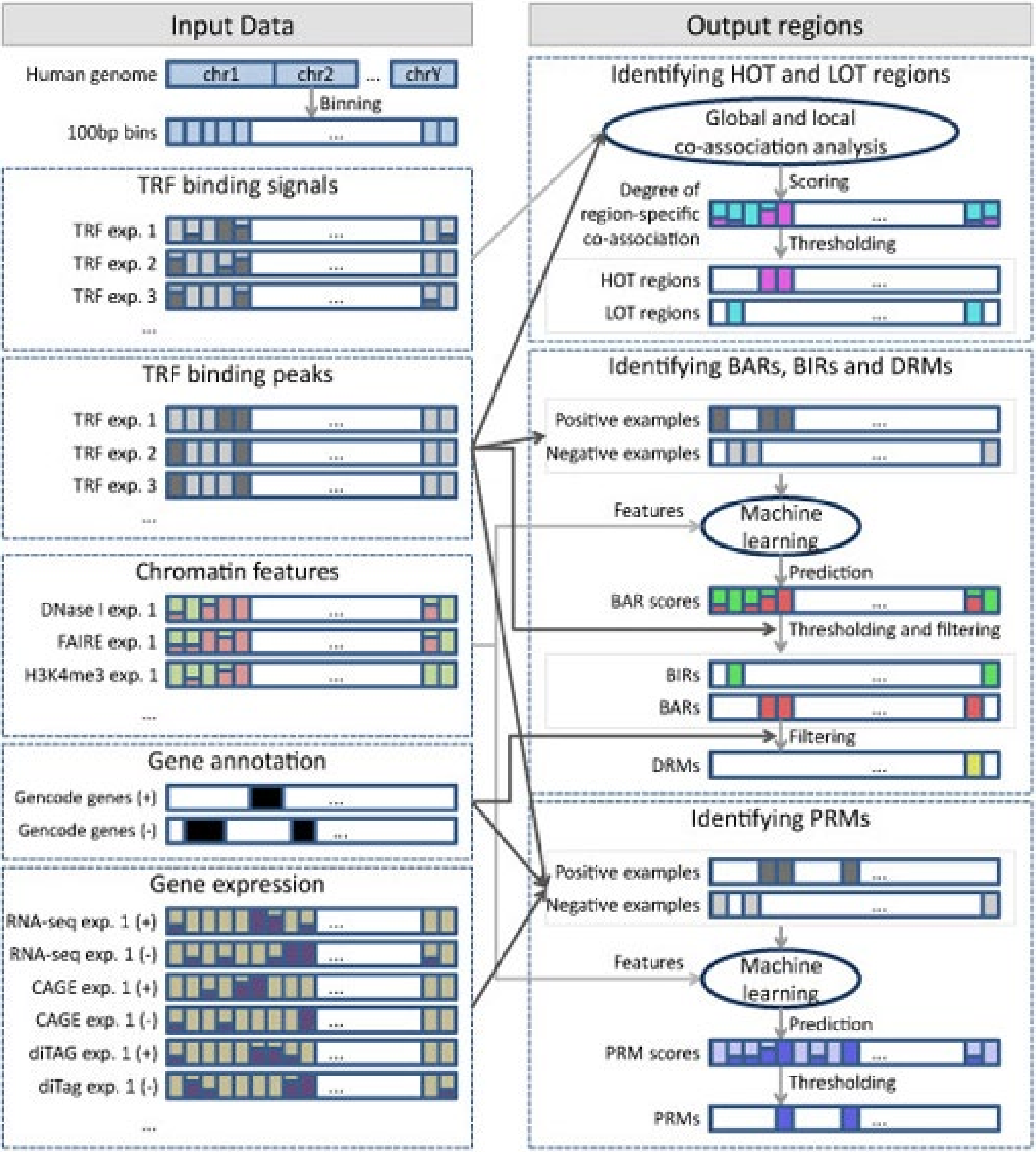


Figure 1 | Overview of the pipeline for identifying the six types of regions for one cell line. The left side shows the input data involved. The right side shows how these datasets were used to identify the regions. The same pipeline was applied to five different cell lines. See Materials and methods for details. The color scheme for the six regions is used in all figures and supplementary figures of the paper.Abbreviation: exp. - experiment.

numbers of cell-invariant and cell-specific occurrences. It is important to note that the consensus summary classes do not capture all the detail discovered in the individual segmentations containing more states.

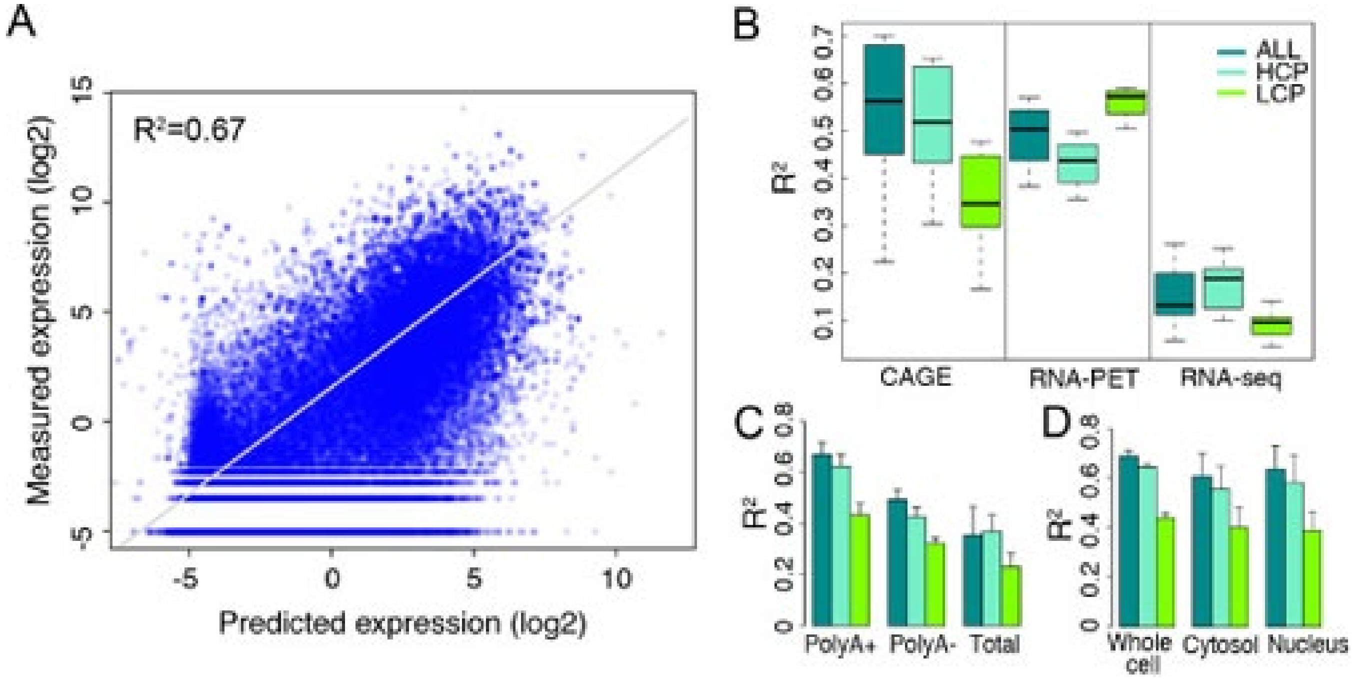


Figure 1 | Accuracy of the TF model for predicting TSS expression levels. (a) Consistency of predicted values with expression levels measured by CAGE in Poly A+ RNA samples extracted from whole cells. (b) Comparison of predictive accuracies of the TF model for expression data generated by three different technologies: CAGE, RNA-PET and RNASeq. (c) Comparison of predictive accuracies of the TF model for expression data from three different RNA extraction protocols: Poly A+, Poly A- and total RNA. (d) Comparison of predictive accuracies of the TF model for expression data in different cellular components. In (b-d.), only data sets from K562 are used. The binding signals of 40 TFSSs are used as predictors. HCP and LCP are high and low CpG content promoters, respectively. Separate models are constructed for ALL, HCP and LCP categories.

The distribution of RNA species across segments is quite distinct, indicating that underlying biological activities are captured in the segmentation. Polyadenylated RNA is heavily enriched in gene bodies. Around promoters, there are short RNA species previously identified as promoter-associated short RNAs (PASRs) (Figure 5B)[16,70]. Similarly, DNA methylation shows marked distinctions between segments, recapitulating the known biology of predominantly unmethylated active promoters (TSS states) followed by methylated gene bodies[42] (T state, Figure 5D). The two enhancer-enriched states show distinct patterns of DNA methylation, with the less active enhancer state (by H3K27ac/H3K4me1 levels) showing higher methylation. These states also have an excess of RNA elements without poly-A tails and methyl-cap RNA as assayed by CAGE sequences compared to matched intergenic controls, suggesting a specific transcriptional mode associated with active enhancers[71]. TFs also showed distinct distributions across the segments (Figure 5B). A striking pattern is the concentration of TFs in the TSS-associated state. The enhancers contain a different set of TFs. For example, in K562, the E state is enriched for binding by the proteins encoded by the *EP300*, *FOS*, *FOSL1*, *GATA2*, *HDAC8*, *JUNB*, *JUND*, *NFE2*, *SMARCA4*, *SMARCB1*, *SIRT6*, and *TAL1* genes. We tested a subset of these predicted enhancers in both Mouse and Fish transgenic models (examples in Figure 6), with over half of the elements showing activity, often in the corresponding tissue type.

To provide a fine-grained regional classification, we turned to a Self Organizing Map (SOM) to cluster genome segmentation regions based on their assay signal characteristics (Figure 7). The segmentation regions were initially randomly assigned to a 1,350-state map in a two-dimensional toroidal space (Figure 7A). This map can be visualised as a two dimensional rectangular plane onto which the various signal distributions can be plotted. For instance, the rectangle at the bottom left of Figure 7A shows the distribution of the genome in the initial randomised map. The SOM was then trained using the 12 different ChIP-seq and DNase-seq assays in the six cell types previously analyzed in the large-scale segmentations (i.e. over 72-dimensional space). After training, the SOM clustering was again visualised in two dimensions, now showing the organized distribution of genome segments (lower right hand, Figure 7A). Individual data sets associated with the genome segments in each SOM map unit (hexagonal cells) can then be visualised in the same framework to learn how each additional

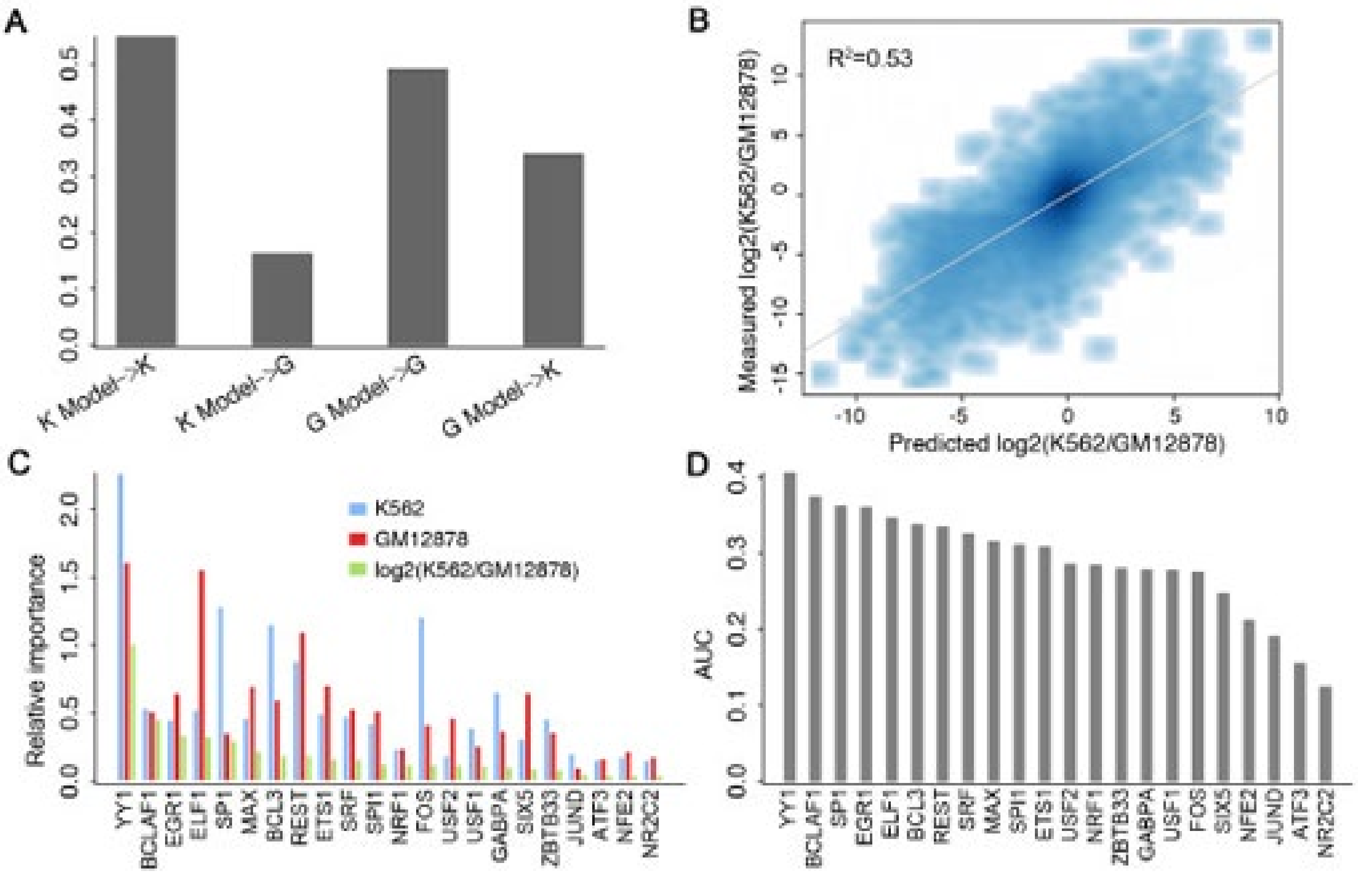


Figure 5 | Cell line specificity of the TF model. (a) Models trained and tested on data from the same cell line result in higher predictive accuracies. K Model and G Model represent models trained with data from K562 and GM12878, respectively. (b) Consistency of predicted log2 fold changes with the experimentally measured differences between K562 and GM12878. Differential binding of 22 TFs are used as the predictors in a predictive model of differential expression. (c) The relative importance of TFs in K562- and GM12878-specific models as well as the predictive model for differential expression. (d) The power of each individual TF for classifying K562- and GM12878-specific promoters (log2 fold change $>$2). CAGE expression data in Poly A$+$ RNA extracted from K562 and GM12878 whole cells were used in the calculation.

kind of data is distributed on the chromatin state map. Figure 7B shows CAGE/TSS expression data overlaid on the randomly initialised (left) and trained map (right) panels. In this way the trained SOM highlighted cell type-specific TSS clusters (bottom panels of Figure 7B), indicating that there are sets of tissue specific TSSs that are distinguished from each other by subtle combinations of ENCODE chromatin data. Many of the ultra-fine-grained state classifications revealed in the SOM are associated with specific gene ontology (GO) terms (right panel of Figure 7C). For instance, the left panel of Figure 7C, identifies 10 SOM map units enriched with genomic regions associated with genes associated with the GO term 'immune response'. The central panel identifies a different set of map units enriched for the GO term "sequence-specific TF activity". The two map units most enriched for this GO term, indicated by the darkest green colouring, contain genes with segments that are high in H3K27me3 in H1 hESC cells, but that differ in H3K27me3 levels in HUVEC cells. Gene function analysis with the GO ontology tool (GREAT[72]) reveals that the map unit with high H3K27me3 in both cell types is enriched in TF genes with known neuronal functions, whereas the neighbouring map unit is enriched in genes involved in body patterning. The genome browser shots at the bottom of Figure 7C pick out an example region for each of the two SOM map units illustrating the difference in H3K27me3 signal. Overall, we have 228 distinct GO terms associated with specific segments across one or more states (Ali Mortazavi, personal communication), and can assign over one third of genes to a GO annotation solely on the basis of its multi-cellular histone patterns. Thus the SOM analysis provides a fine-grained map of chromatin data across multiple cell types, which can then be used to relate chromatin structure to other data-types at differing levels of resolution (for instance, the large

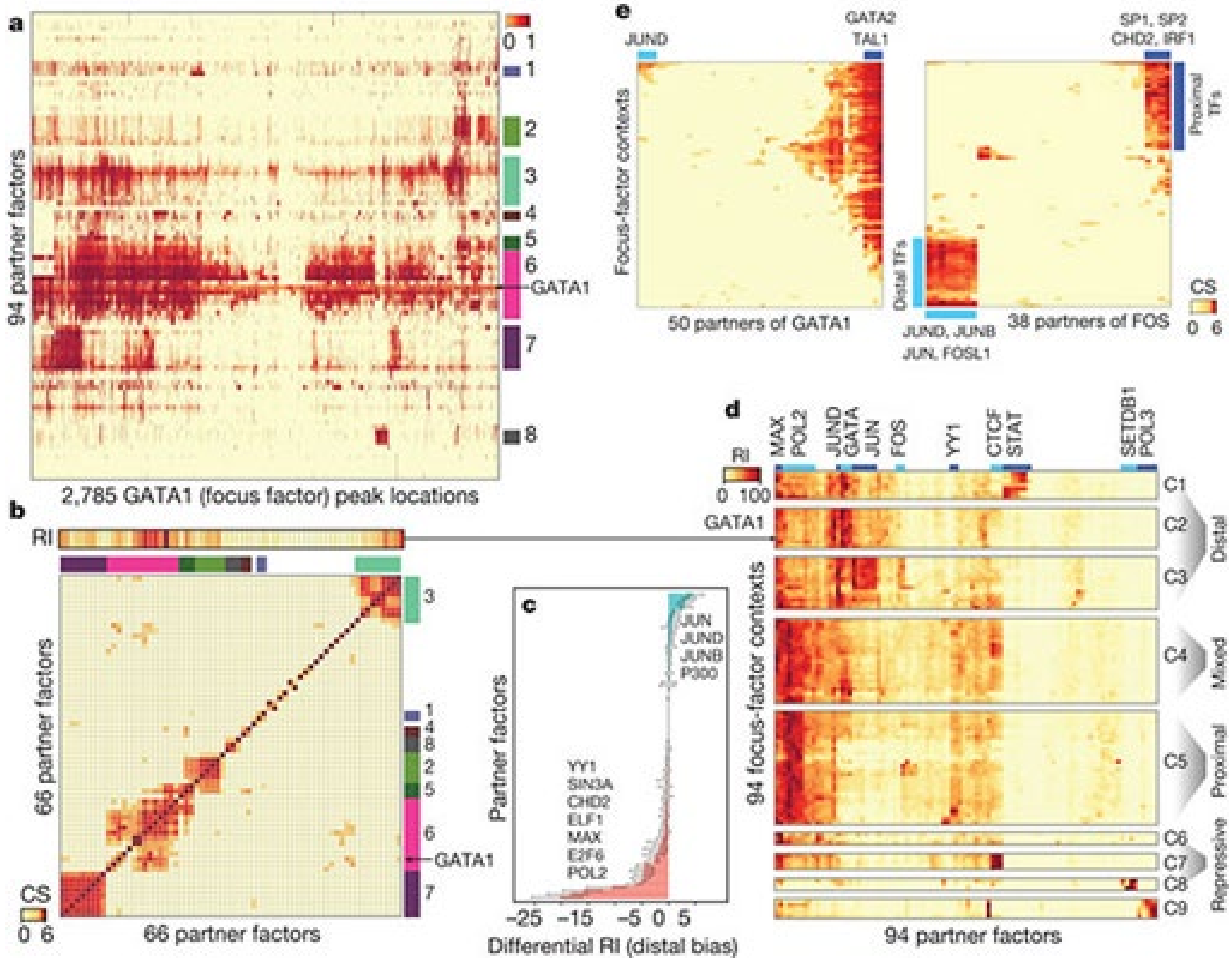
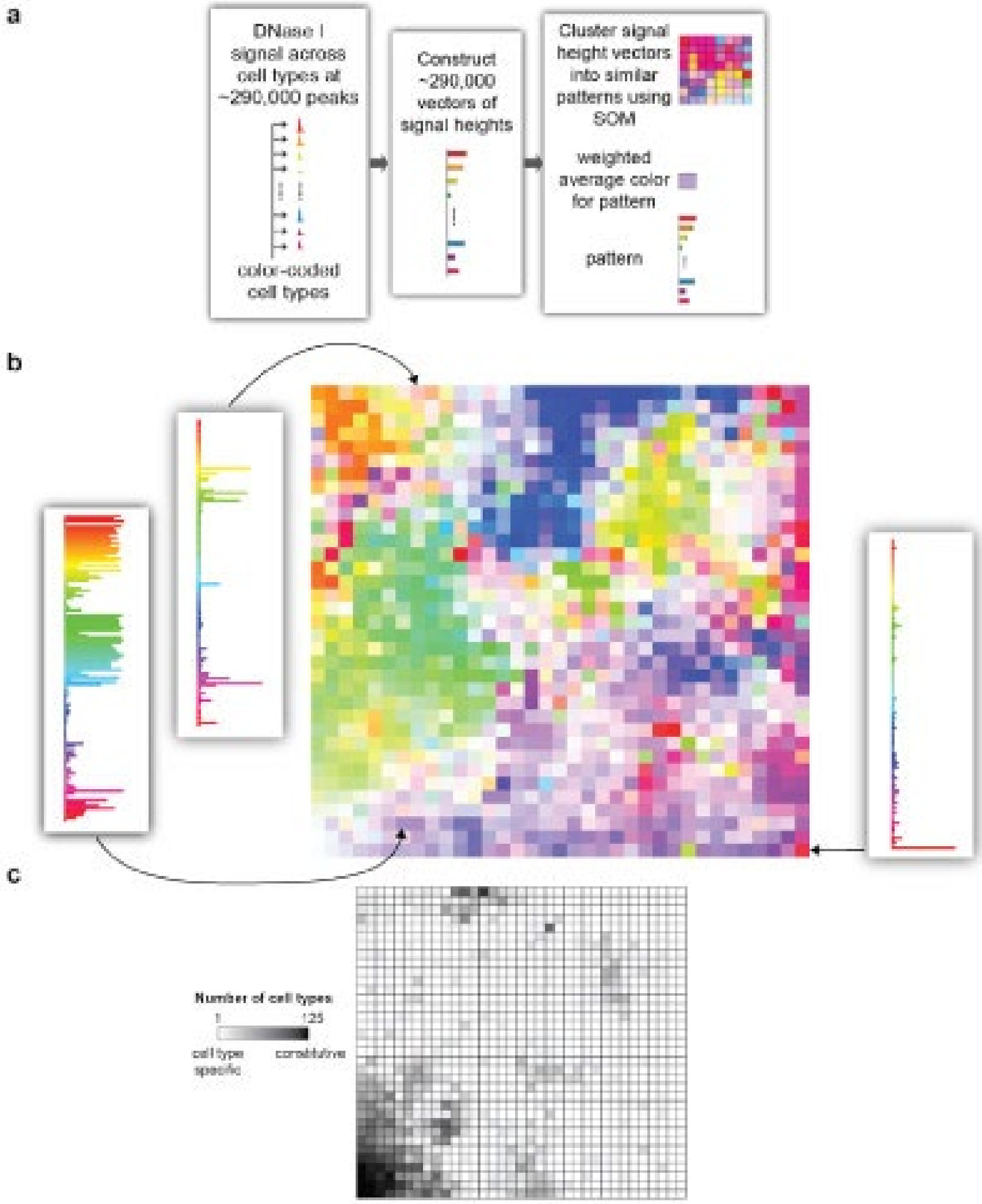


Figure 1 | **TF Co-association** (**a**) The co-binding map for the GATA1 focus-factor context in K562 shows the binding intensity of peaks of all TFs in K562 (rows) that overlap each GATA1 peak (columns). The colored rectangles represent 8 key clusters consisting of different combinations of co-associating partner-factors. (**b**) The GATA1 context-specific relative importance scores (RI) of all partner-factors (top) and the matrix of co-association scores (CS) between all pairs of TFs (bottom). Primary and local partners of GATA have high RI scores. The co-association score matrix captures the 8 clusters observed in (**a**). (**c**) Different partner-factors are preferentially enriched at gene-distal (positive differential RI) and proximal (negative differential RI) GATA1 peaks. (**d**) The aggregate factor importance matrix, obtained by stacking the RI of all partner-factors (columns) from all focus-factor contexts (rows) in K562, shows 9 functionally distinct clusters (C1 to C9) of contexts that can be broadly grouped as distal, proximal, mixed, and repressive. The blue rectangles highlight representative partner-factors with high RI in the clusters. The arrow from (**b**) to (**d**) indicates that the GATA1 context-specific RI scores form one row in this matrix. (**e**) Co-association variability map of partners (columns) of GATA1 (left panel) and FOS (right panel) over all K562 focus-factor contexts (rows). TAL1 and GATA2 show consistently high CS with GATA1 over most focus-factor contexts, but JUND shows context-specific co-association. FOS shows dramatic changes in CS of partner-factors over different contexts (e.g. FOS-JUND in distal contexts and FOS-SP2 in proximal ones). (More details in Fig. S2c, S2f-1, S2d, S2l-2.)

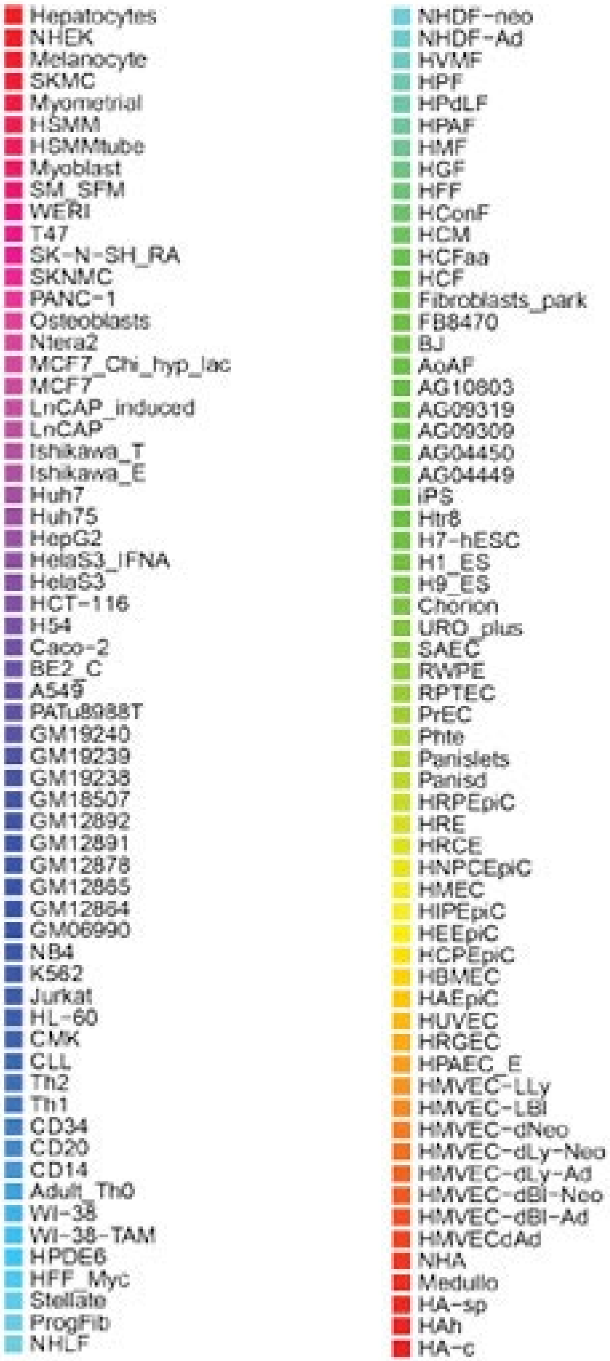
cluster of units containing any active TSS, its sub-clusters composed of units enriched in TSSs active in only one cell type, or individual map units significantly enriched for specific GO terms).



Supplementary Figure 18 | Using a self-organizing map to cluster DHSs by cross-cell-type pattern. Clustering of ~290,000 DHSs by cross-cell-type patterns using a self-organizing map (SOM), which learns patterns in the data and organizes DHSs into stereotyped groups analogous to those shown in Fig. 6a-e. (a) Schematic for SOM clustering and colour coding of patterns; index of cell types with their colours is given in Supplementary Fig. 19. (b) SOM of 1,225 DHS patterns. Each cell in the 35×35 grid represents one stereotyped pattern, with colour coding determined according to the weighted "average" cell type for that pattern. Three example pattern profiles are shown, corresponding to the indicated nodes in the grid. (c) Greyscale heatmap corresponding to that in (b) showing, for each colour-coded pattern, the cell-specificity of that pattern. Shading indicates cell-selectivity; black = DHS is constitutive (i.e. present in all cell types); white = DHS is cell type-specific; greyscale = gradations threreof. Note the concentration of patterns with promiscuous DHSs in the lower right; however, most stereotyped DHS patterns are highly cell-selective.

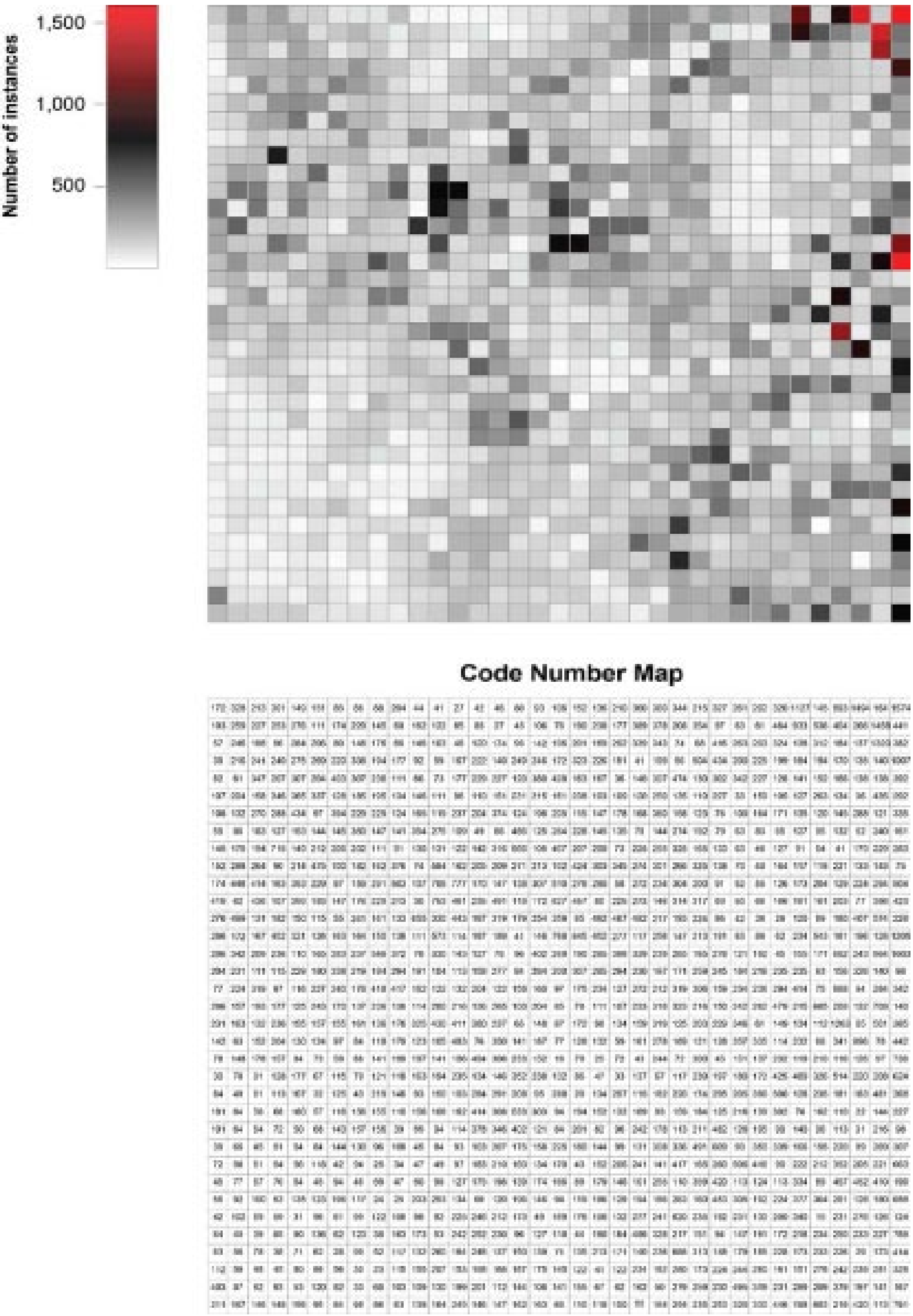
Supervised and semi-supervised prediction of enhancers

Our procedure for identifying enhancers involved the use of "supervised" machine learning methods, i.e., methods that learn model parameters from known examples. However, our overall pipeline is not truly supervised in that we only used supervised models to learn regions needed by the procedure to identify enhancers, such as BARs and PRMs. These regions were then used in an unsupervised manner in the final prediction of enhancers. This design was driven by an insufficient number of cell-type-specific positive and



Supplementary Figure 19 | Colour-coded key to the cell types in Supplementary Fig. 18.

negative examples of enhancers. While there are large enhancer catalogs, such as the VISTA database database[6], most of the validation experiments were done in specific assays (such as embryos of transgenic mouse) that may not be appropriate as examples for other cell types due to the dynamic nature of protein binding and gene regulation. In fact, when we tried to use data from VISTA to learn direct supervised models for enhancers using chromatin data from our cell lines as features, the prediction accuracy was low according to some left-out data not used in model training. We hope that with the larger-scale validation efforts of ENCODE[20] and other groups, more cell-type-specific data will become available and the construction of highly reliable, supervised predictive models of enhancers will become possible.



Supplementary Figure 20 | Instance counts of patterns discovered by the SOM (Supp. Fig. 18) The number of instances of each pattern discovered by the SOM illustrated in Supplementary Fig. 18; the top matrix is simply a heatmap version of the numeric matrix underneath.

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

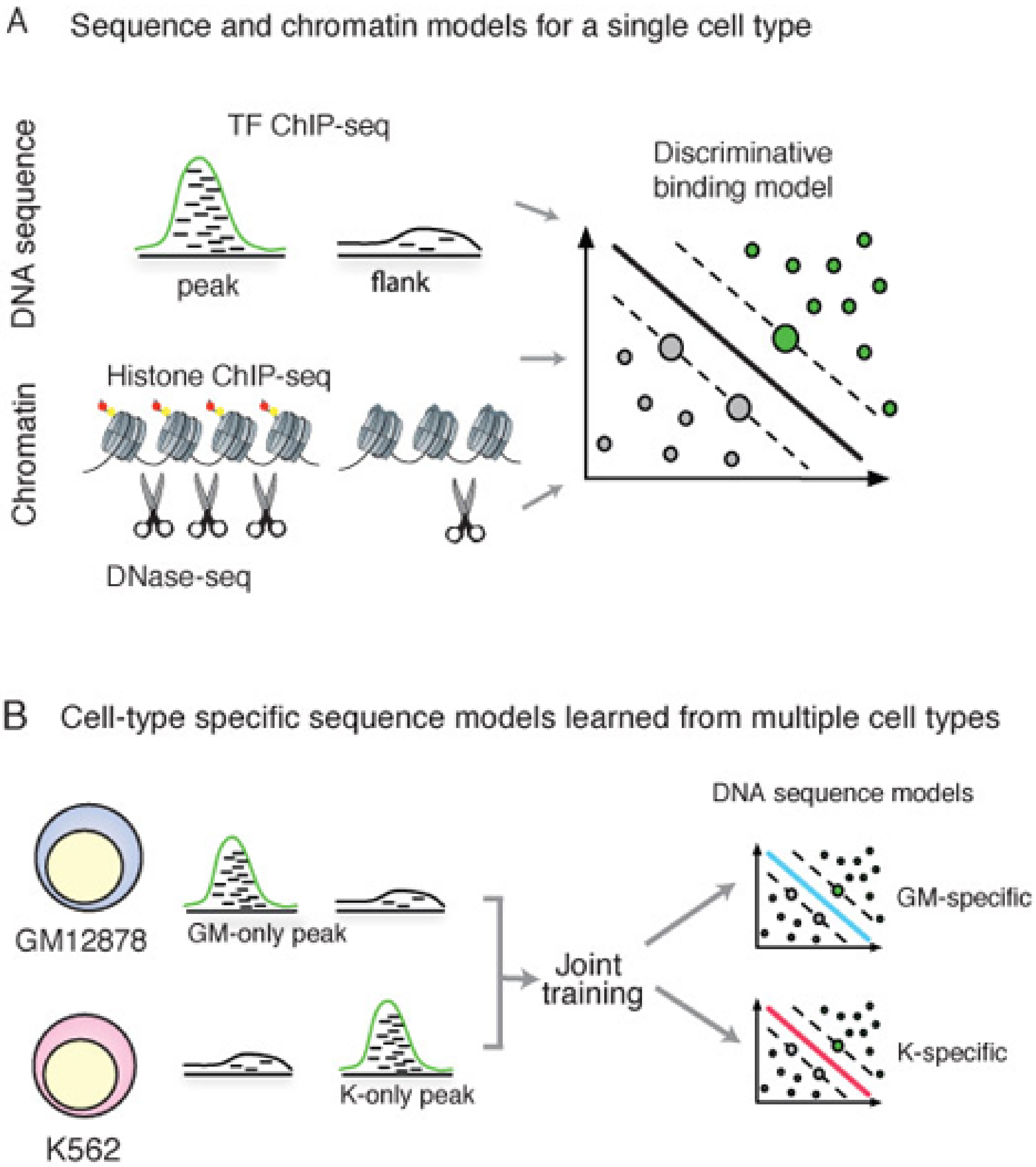


Figure 1 | Schematic of models to predict transcription factor occupancy from sequence and chromatin. (a) We developed DNA sequence and chromatin models based on flexible *k*-mer patterns and spatial organization of histone modifications and DNase accessibility. The models were trained to discriminate between regulatory ChIP-seq peaks and flanking regions within a single cell type using a support vector machine. (b) To study cell-type-specific DNA sequence preferences, we simultaneously train on binding site data from two cell types. This allowed us to jointly learn the cell-type-specific preferences (top and bottom).

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 (R2), the percentage

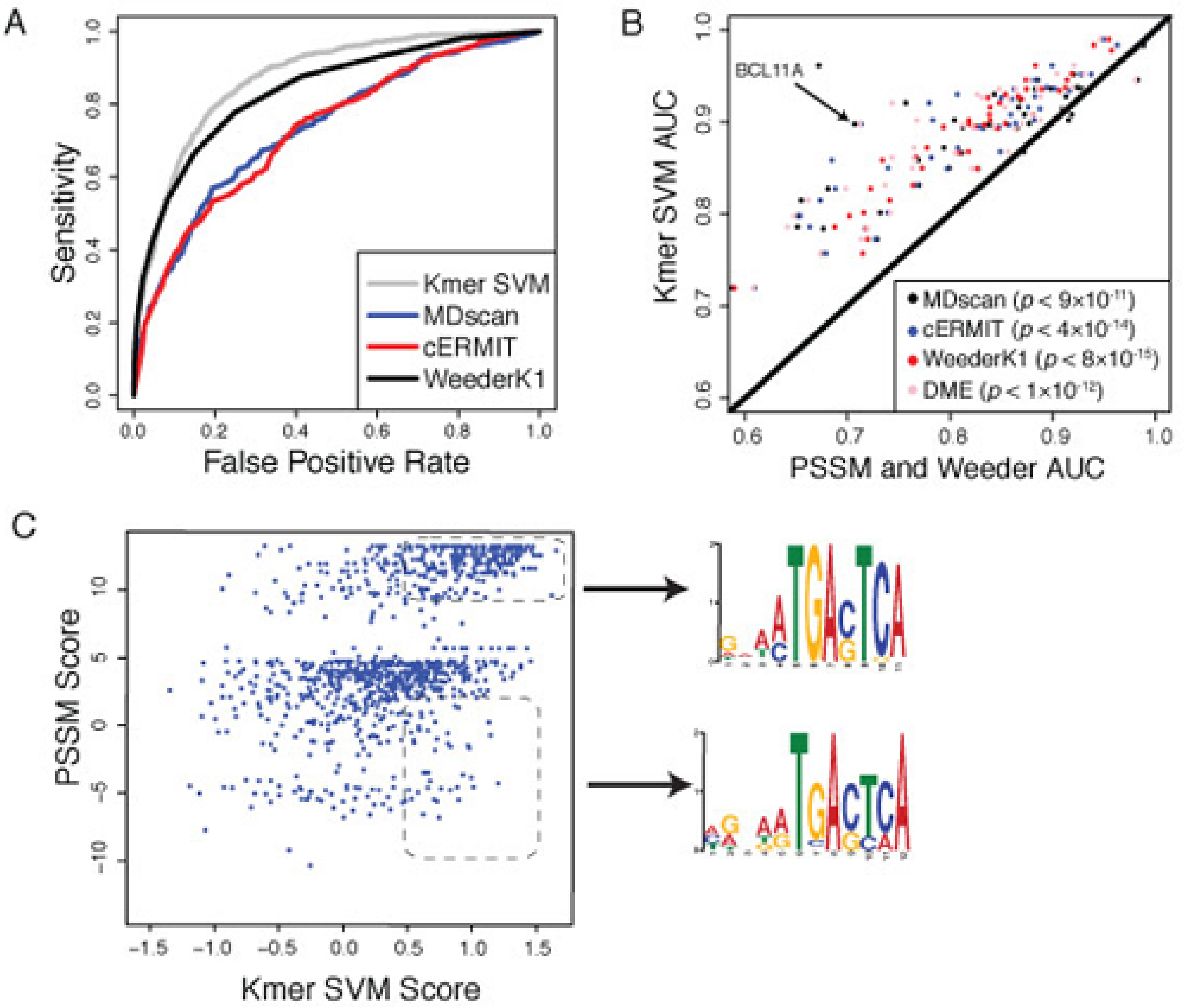


Figure 2 | SVM sequence models better predict binding sites than traditional motif approaches. (a) The accuracy of our method is assessed by the area under the ROC curve, which provides a natural tradeoff between false positives (x-axis) and sensitivity (y-axis). The ROC curve is shown for discriminating BCL11A ChIP-seq peaks from nonpeaks using four approaches: *k*-mer SVM, MDscan, cERMIT, and Weeder. (b) The accuracy (AUC) of *k*-mer SVM models (y-axis) is compared against motif-based algorithms (MDscan, cERMIT, DME, and Weeder; x-axis) for discriminating ChIP-seq peaks from flanking regions. We used training and test sets taken from the same experiment; only accuracy on the test set is shown. Results for transcription factors with multiple ChIP-seq experiments for replicates and cell types were averaged. The SVM models are significantly more accurate than each of the alternative methods (P-values inset and color-coded for each method). (c) The *k*-mer SVM model is able to learn degenerate motifs. We show the *k*-mer SVM scores (y-axis) versus the cERMIT motif score (x-axis) for binding sites of BCL11A in GM12878. Example binding sites that are detected by the SVM but receive low scores by the motif are enriched for a more degenerate motif instance, as found by MEME.

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). Performance of the first three methods was roughly comparable, and was better than MLR, implying a non-linear relationship between TF binding and TSS expression (Supplementary Figure S1). In this article, to simplify presentation we focus on results from the RF method for models with multiple predictors and the SVR method for models with a single predictor (see "Methods" for details). Results from different methods are highly consistent and lead to the same conclusions, e.g. the relative importance of different TFs for predicting gene expression.

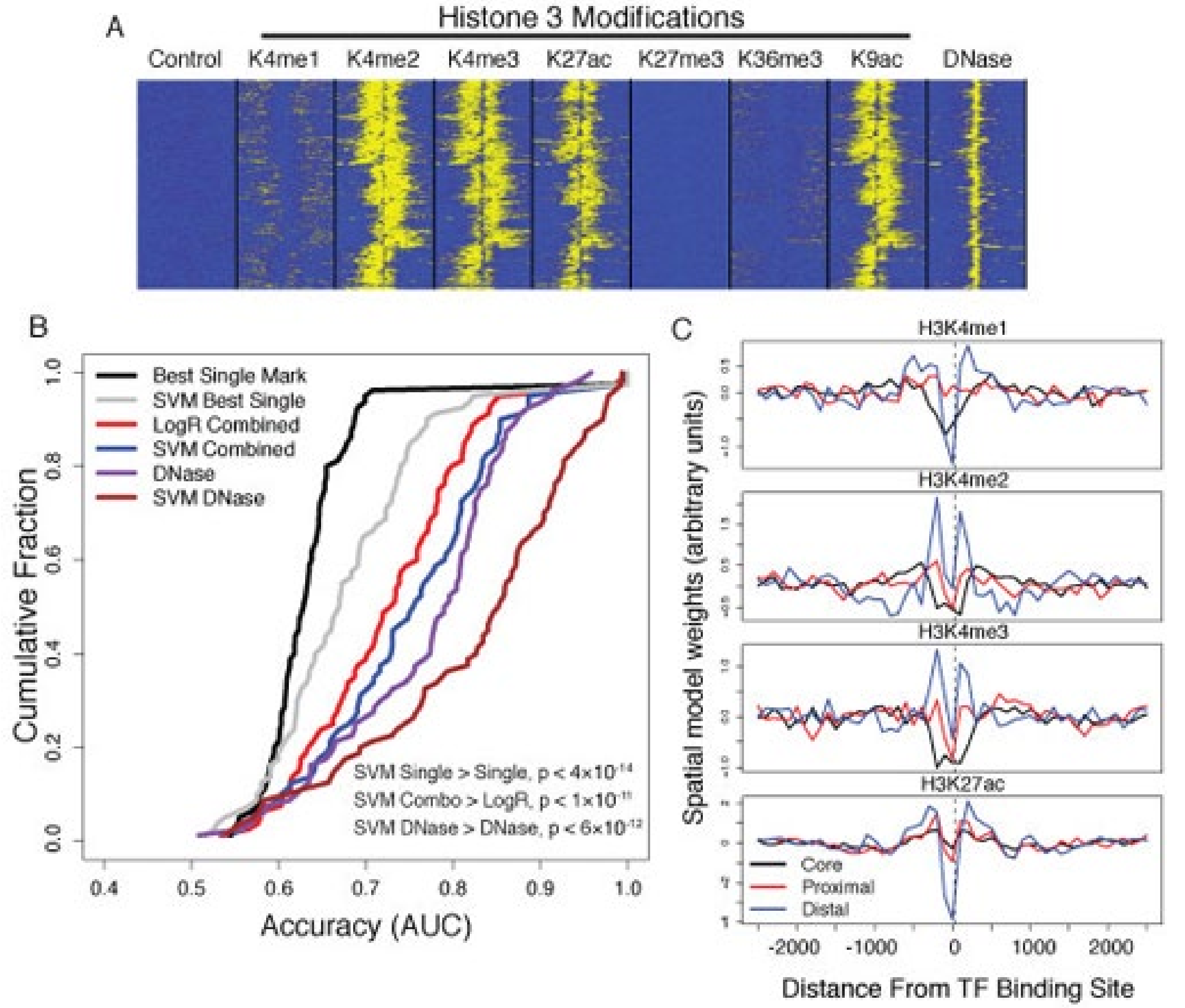


Figure 3 | SVM spatial chromatin models better predict binding sites than simpler models. (**a**) The distribution of histone marks over 5000-bp windows centered at GABPA ChIP-seq peaks in K562 shows spatial organization of multiple correlated signals. (**b**) The accuracy of multiple chromatin models suggests that spatial signatures of DNase accessibility better predict binding sites than other methods. The cumulative distributions of prediction accuracy (AUC; x-axis) across a subset of ChIP-seq experiments are shown for multiple chromatin representations. Shown are an SVM model trained on all spatially binned histone marks (blue), which is more accurate than standard ranking based on best single mark read counts (black) or a logistic regression combination of read counts (red); similarly, an SVM model trained on spatially binned DNase-seq reads (brown) better describes binding sites than use of DNase bin counts (purple). Paired signed rank test P-values are shown. (**c**) Transcription factors that bind the core promoter, proximal to transcript start site, or distal to start site have distinctive spatial patterns of histone modifications. The four plots show spatial coordinates of the learned bin weights arranged along the x-axis, with the values of the weights shown on the y-axis. The bin weightings are averaged across subsets of core, proximal, and distal binding transcription factors. The valleys at the binding site suggests that spatial models are capturing predictive information regarding the differential spacing of nucleosome-depleted regions at core, proximal, and distal binding sites.

Determination of distal regulatory relationship

Interactions involving distal regulatory elements (e.g., enhancers) are more difficult to identify than those

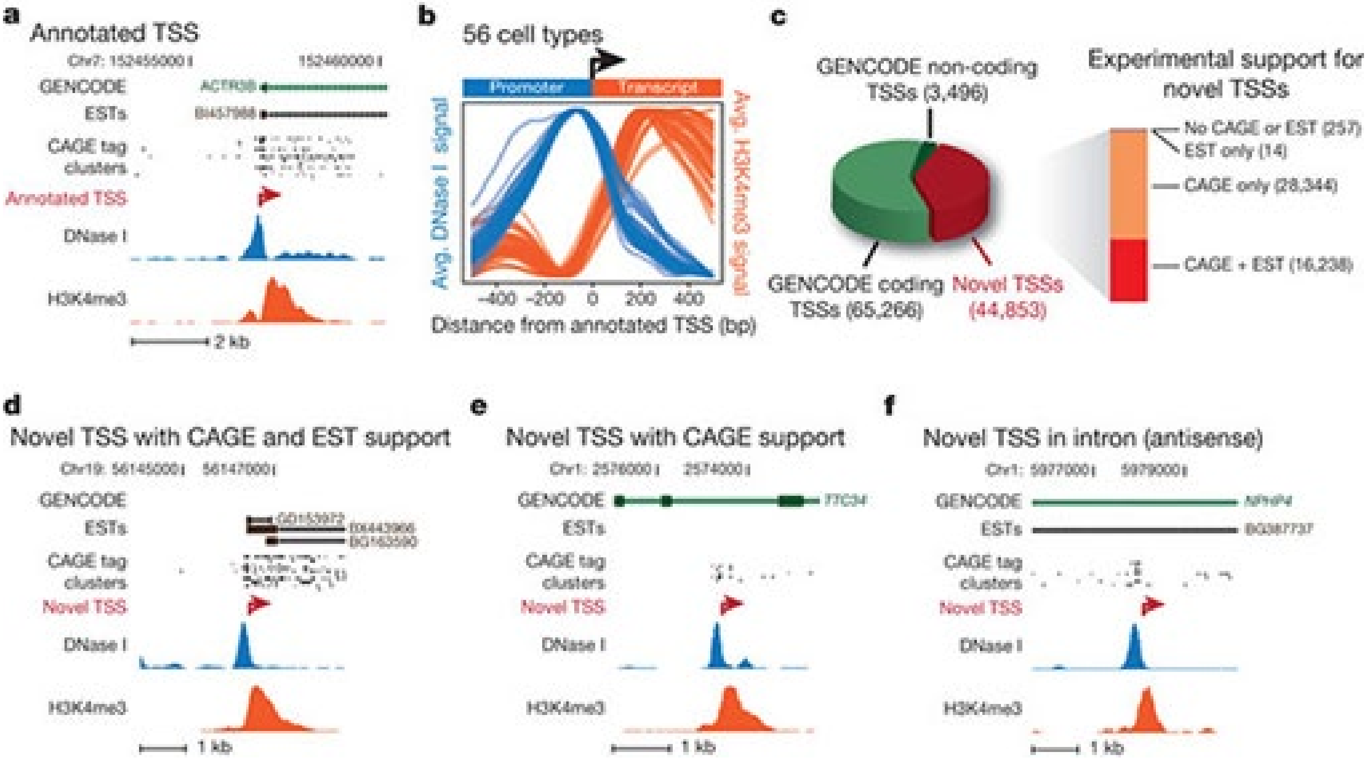


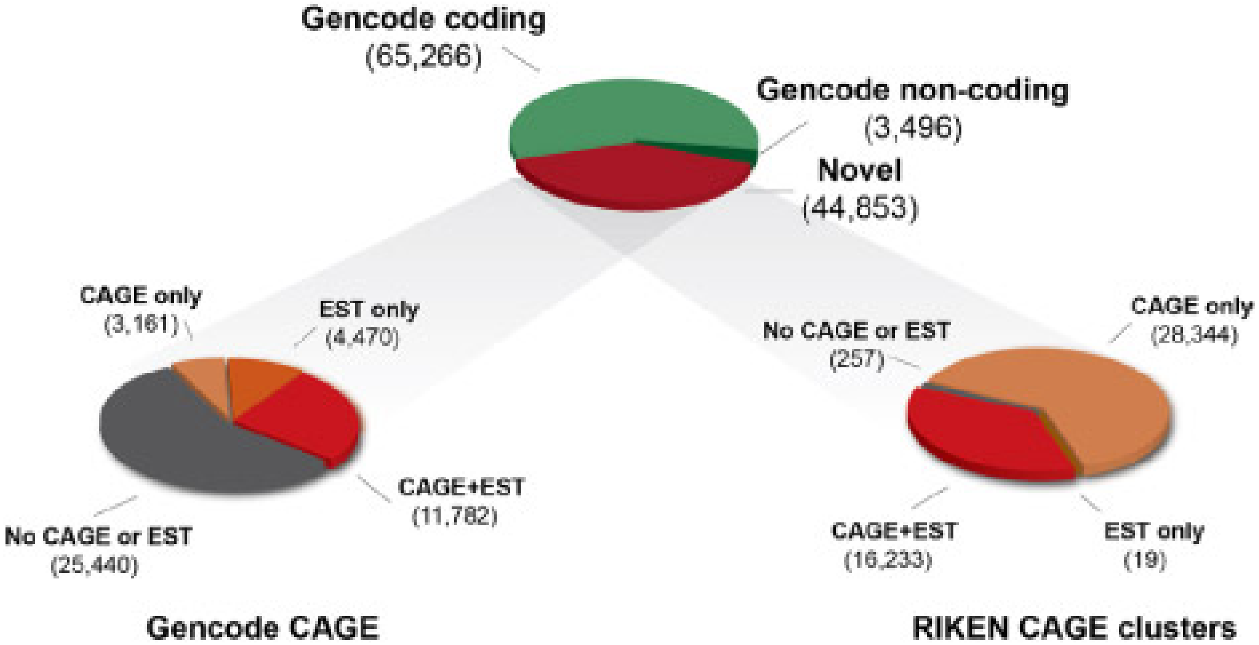
Figure 3 | Identification and directional classification of novel promoters. (a) DNase I (blue) and H3K4me3 (red) tag densities for K562 cells around annotated TSS of *ACTR3B*. (b) Averaged H3K4me3 tag density (red, right *y* axis) and log DNase I tag density (blue, left *y* axis) across 10,000 randomly 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.

involving proximal elements. Here, we employed a statistical model[35]. This identifies distal sites with potentially many binding TFs using chromatin features. These regions were associated with a gene if their changing pattern of chromatin marks across cell lines correlates with the expression of that gene (SOM/E.1). Overall, the model identified 19258 distal edges (Fig. 2a).

Context-specific TF Co-association

We first examined the genome-wide co-association of all pairs of TFs by analyzing the overlap between peaks of all pairs of factors[20]. Although many general trends can be identified, this approach does not take into account the context-specificity of TF binding (i.e., the fact that TFs bind together in distinct combinations at different genomic locations, and that the co-binding of one pair of TFs is often affected by the binding of another TF; SOM/C.1). Therefore, we developed a framework focusing on the specific genomic regions bound by a particular TF (the focus-factor) and examined the co-association of all other TFs (partner-factors) within this context (Fig. S2a). For each ∼350 bp region in the focus-factor context, we extracted normalized binding signals of overlapping peaks of all TFs, generating a co-binding map. Fig. 1a shows such a map for the GATA1 context. Here, factors that consistently co-associate with each other and a substantial proportion of GATA1 peaks are termed 'primary partners' (e.g., group 6 TFs such as GATA2 and TAL1 in Fig. 1a). In addition to these factors, there are also groups of 'local partners' that co-associate with each other in the presence of GATA1, but only at specific subsets of GATA1 binding peaks (e.g., JUN in group 7 and MAX in 3; Fig. 1a and S2c-1). These biclusters, typically containing 2 to 5 TFs, can be mutually exclusive or partially overlapping.

To systematically identify all primary and local partners for each focus-factor context, we used a machine-learning approach. We learned non-linear, combinatorial models of each focus-factor's co-binding map relative



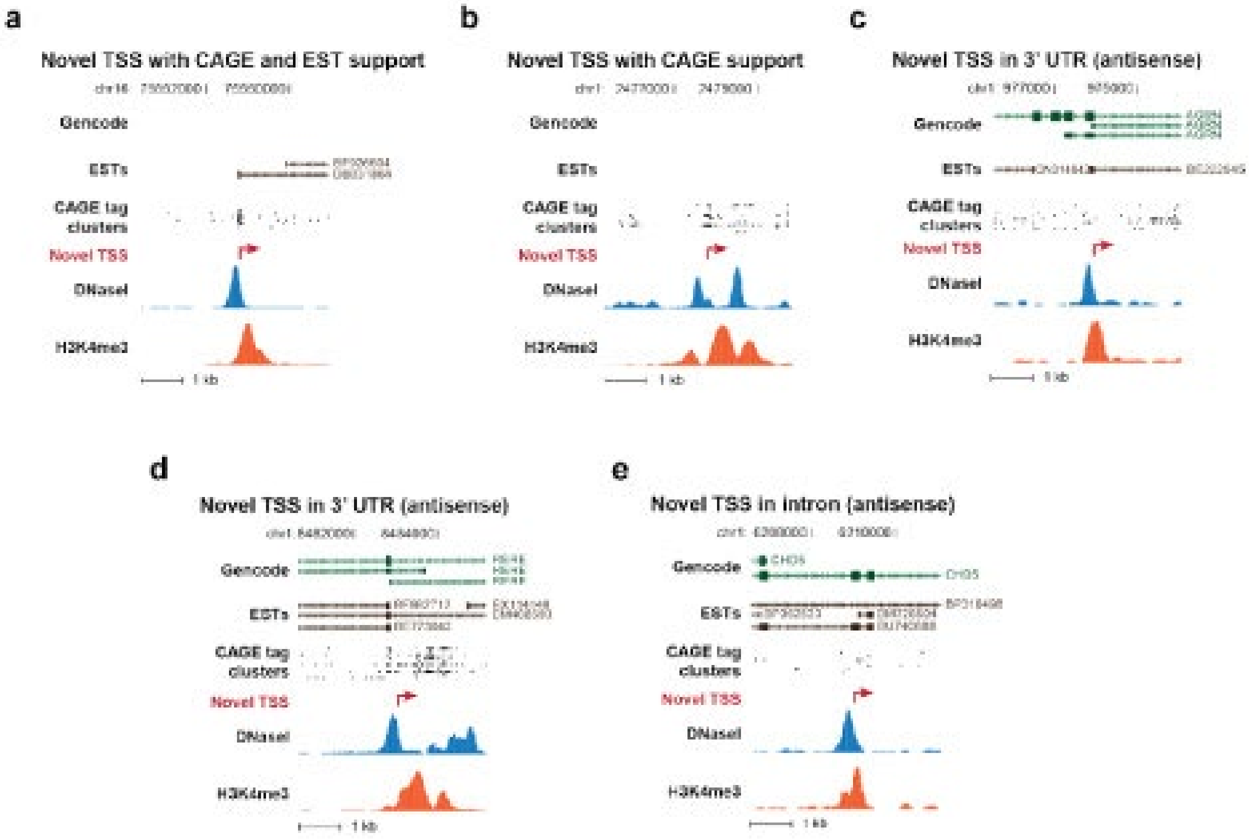
Supplementary Figure 9 | Overlaps between novel promoters, CAGE clusters, and ESTs. This is a refinement of Fig. 3d. The top pie charts are identical in both figures. The bottom two pie charts here show the breakdown of novel promoter predictions with regard to their overlap separately with Gencode CAGE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

to randomized control maps (SOM/C.2; Figs. S2a,b). Analysis of multivariate rules in these models, in turn, identified pairs and higher-order clusters of significantly co-associated TFs. Moreover, these co-associations are robust to peak overlap and calling thresholds (SOM/C.4).

We fit machine learning models to the GM12878 and K562 data independently, and on independent sets of replicates of RNA-seq data in order to assess the reproducibility of our conclusions (see Methods). Our classifiers distinguish between genes with at least one uniquely mapping peptide and those with no uniquely mapping peptides. We were able to construct models with misclassification rates of 21% in K562 and 23% in GM12878 computed on held-out test-sets in both cell lines and on either collection of independent replicates (see Methods). Furthermore, when the models are trained on one set of replicates and tested on the other, the average misclassification rate rises only slightly, to 22% in K562 and 25% in GM12878. Hence, our models are both biologically and technically robust.

The most important predictor in either cell line (in both the K562 full model and the model using only GM12878 available data), is the polyA- Cytosol RNA fraction, and the direction of dependence is positive: higher polyA- Cytosol RNA levels correspond to an increased likelihood of detectable translation (Fig. 2 and Supp. Fig. 3). Although there is some substantial re-ordering of covariate importance down the rank-list, this has only a moderate effect on model performance between the two cell lines, and indeed the precise order of covariates after polyA- Cytosol was unstable in K562 between biological replicates (see Supp. Fig. 3). The usual interpretation of this sort of effect is co-linearity between the variables: the various RNA fractions appear to provide some redundant information.

The marginal positive effect of increased polyA- Cytosol expression is actually greater than the marginal effect of increased polyA+ Cytosol expression (Fig. 2). Indeed, polyA- Cytosol RNA level is the single most important covariate for prediction, although there are minor differences between the two cell lines that may be due to underlying differences in RNA processing and degradation efficiency. We note that K562 is a chronic myeloid leukemia cell line, while GM12878 is a normal but EBV-immortalized LCL. We hypothesize that this fraction may be measuring post-translational RNA processing, by which we mean the degradation and metabolism



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.

of transcripts after translation, resulting in polyA- fragments localized in the cytosol. This illustrates the importance of considering the direction of causality in statistically predictive models: although the natural biological temptation is to think of RNA levels as 'causing' or 'influencing' protein levels (and therefore peptide detectability) the opposite may be true: abundant proteins may be translated from high-abundance transcripts with correspondingly abundant degradation products. That the presence of such degradation products, if our hypothesis is correct, is a better indicator of translational competence than the polyA+ Cytosol fraction remains an intriguing subject for future study. Future experiments should investigate whether these degraded polyA- sequences, derived from previously translated RNAs in the cytosol, are non-functional, or whether they are stable due to post-cleavage 5′-capping (Otsuka *et al.* 2009) and may carry out additional roles in the cell, attesting to multifunctionality and interrelatedness of long and short RNAs.

To visualize the qualities and prevalence of different stereotyped cross-cellular DHS patterns, we constructed a self-organizing map of a random 10% subsample of DHSs across all cell types and identified a total of 1,225 distinct stereotyped DHS patterns (Supplementary Figs 18 and 19). Many of the stereotyped patterns discovered by the self-organizing map encompass large numbers of DHSs, with some counting >1,000 elements (Supplementary Fig. 20).

For each of the five cell lines, we used the cell-line-specific TRF binding data to learn patterns in chromatin features and gene expression levels using machine learning methods. We then used the learned models to define six different types of genomic regions that form three pairs: 1) binding active regions (BARs) and binding inactive regions (BIRs), 2) promoter-proximal regulatory modules (PRMs) and gene-distal regulatory modules (DRMs), and 3) high occupancy of TRF (HOT) regions, and low occupancy of TRF (LOT) regions (Figure 1). In each pair, the two region types are mutually exclusive. On the other hand, region types from different pairs may overlap. For instance, DRMs are subsets of BARs, while some HOT regions overlap with PRMs and DRMs. Each of the six types of regions, however, exhibits some unique properties and we will discuss the six types separately. With the use of cell-line-specific data, we aimed at identifying regions that reflect the internal states of the particular cell types. For example, for PRMs and DRMs, our goal was to identify modules that have active regulatory roles in the particular cell line from which they were called, instead of modules that are only potentially active in some unknown cell types.[26]

However, while we are using one of the largest sets of ChIP-seq data currently available, it contains only a small portion of the estimated 1,700-1,900 human TFs[1]. We therefore took the regions covered by the TRF binding peaks as examples to learn a statistical model based on the observed chromatin features of these regions for each cell line using data produced by ENCODE (Materials and methods). We then applied the model to score all regions in the whole human genome. Cross-validation results show that our learned models can separate regions covered by TRF binding peaks from other random regions well (Additional file 2, Figure S1 and Additional file 2, Figure S2). Since some of the selected random regions may actually be bound by TRFs not in our dataset, we do not expect 100% accuracy, and the observed accuracy values are sufficiently high to indicate that our models have captured some general chromatin properties of regions with active binding.

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.