1 Transcription factor motifs

ENCODE discovers many new transcription-factor-binding-site motifs and explores their properties

To directly identify regulatory regions, we mapped the binding locations of 119 different DNA-binding proteins and a number of RNA polymerase components in 72 cell types using ChIP-seq (Table 1, Supplementary Table N1, ref 19); 87 (73%) were sequence-specific TFs (TFSS). Overall, 636,336 binding regions covering 231Mb (8.1%) of the genome are enriched for regions bound by DNA-binding proteins across all cell types. We assessed each protein-binding site for enrichment of known DNA-binding motifs and the presence of novel motifs. Overall, 86% of the DNA segments occupied by TFSS contained a strong DNA-binding motif and in most (55%) cases, the known motif was most enriched (Pouya Kheradpour and Manolis Kellis, personal communication).

We organized all the information associated with each TF, including the ChIP-seq peaks, discovered motifs, and associated histone modification patterns, in FactorBook[26], a public resource which will be updated as the project proceeds.

We observed reduced levels of individual variation at functional binding sites compared to reshuffled motif matches and flanking regions for other *Drosophila* factors as well as human TFs (Figure 2A). Notably, the significance of this effect was similarly high in *Drosophila* and humans, despite the fact that the SNP frequency differed approximately 11-fold (2.9% vs 0.25%, respectively), as closely reflected by the 7.5-fold difference in the number of varying TFBS. This is consistent with the overall differences in the total number of SNPs detected in these two species, likely resulting from their different ancestral effective population sizes [39]. We also observed a significant anti-correlation between variation frequency at motif positions and their information content in both species (Figure 2B).

We proposed to express the deleterious effect of TFBS mutations in terms of *mutational load*, a known population genetics metric that combines the frequency of mutation with predicted phenotypic consequences that it causes [31,32] (see Materials and methods for details). We adapted this metric to use the reduction in PWM score associated with a mutation as a crude but computable measure of such phenotypic consequences.

We do not assume that TFBS load at a given site reduces an individual's biological fitness. Rather, we argue that binding sites that tolerate a higher load are less functionally constrained. This approach, although undoubtedly a crude one, makes it possible to consistently estimate TFBS constraints for different TFs and even different organisms and ask why TFBS mutations are tolerated differently in different contexts.

Among the TF binding sites that were ubiquitously functional, we compared the genomic footprints of sites where binding activated or repressed transcription in all four cell lines. Among the transcription factors we examined (see Table 1), YY1 had the most examples of each case (9 ubiquitously activating and 16 ubiquitously repressing sites). Fig. 2 shows the motifs derived from this analysis for YY1. The most striking difference between the YY1 motif for sites where binding is associated with activation (Fig. 2 (b)) and those where binding is associated with repression (Fig. 2 (c)) occurs at position 4, where the G has greater information content for repressing cases ($p < 0.012$ using a permutation test, see Additional file 1, Fig. S7). The repressive YY1 binding sites are closer to translational start sites than are the activating YY1 binding sites ($p=7.7 \times 10^{-4}$). Indeed, 12 of the repressing YY1 binding sites are located directly over the translational start site, whereas only a single activating YY1 binding sites is. The mutagenesis experiments reported here elucidate the functional distinction between the different classes of YY1 binding sites that were noted in a previous analysis of DNA binding (ChIP-chip) [75]: the class of YY1 binding sites localized around the translational start site are strongly associated with transcriptional repression,

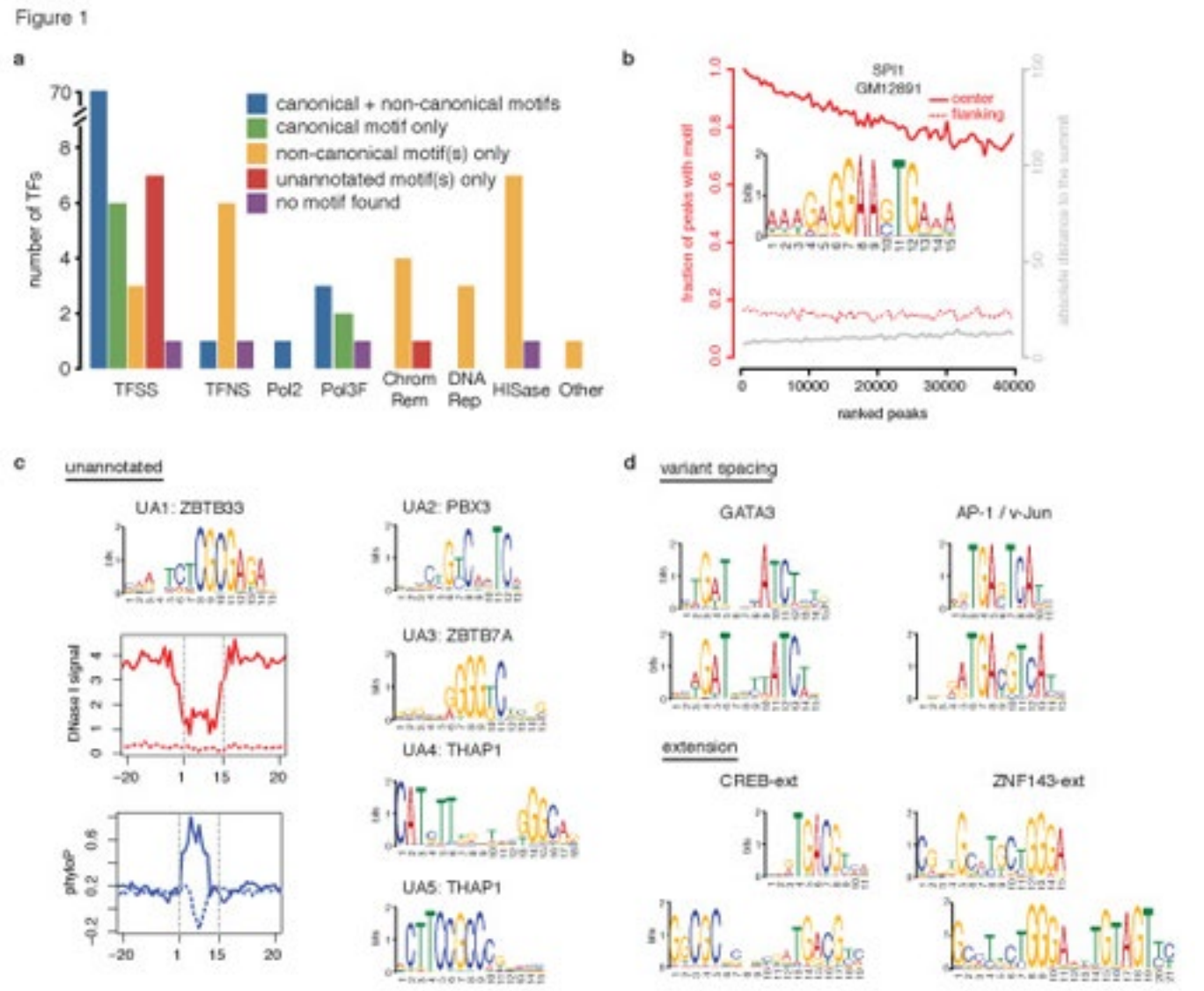


Figure 1 | **De novo discovery of sequence motifs.** (**a**) Statistics of motif discovery among 119 TFs, classified into 87 Pol II-associated sequence-specific TFs (TFSS), 8 General Pol II-associated, non-sequence-specific TFs (TFNS), Pol II (Pol2), 6 Pol III components and Pol III-associated TFs (Pol3F), 5 ATP-dependent chromatin complexes (ChromRem), 3 TFs involved in DNA repair (DNARep), 8 histone modification complexes (HISase) and 1 cyclin kinase associated with transcription (Other). The TATA box binding protein (TBP) is included in the TFNS category and its canonical motif is TATA, corresponding to the blue bar. (**b**) Example result for SPI1 in GM12891 cells illustrating the percentage of peaks with the motif (left y-axis in red) and distribution of absolute distances of the closer edge of motif sites relative to the peak summit (right y-axis in grey), plotted against ranks of peaks (ranked by ChIP-seq signal). (**c**) Five previously unannotated motifs that are likely to be canonical motifs of four sequence-specific TFs. Also shown are DNase I footprint and sequence conservation profiles around the sites of UA1 (likely the canonical motif of ZBTB33). Motif sites in ChIP-seq peaks (solid lines) were compared with motif sites outside peaks (dashed lines). DNase I and ChIP-seq data were both from K562 cells. Sequence conservation was computed using phyloP (Pollard *et al.* 2010). (**d**) Motifs with variant spacing and extensions.

while those localized closer to the TSS are associated with activation.

πIn Fig. 2 (d), we report the vertebrate phyloP score [87] for each nucleotide, averaged over sites where YY1 binding results in activation or repression of transcription, respectively. Error bars indicate the standard error of the mean. Conservation is generally high for YY1, relative to that for the other transcription factors in our study.

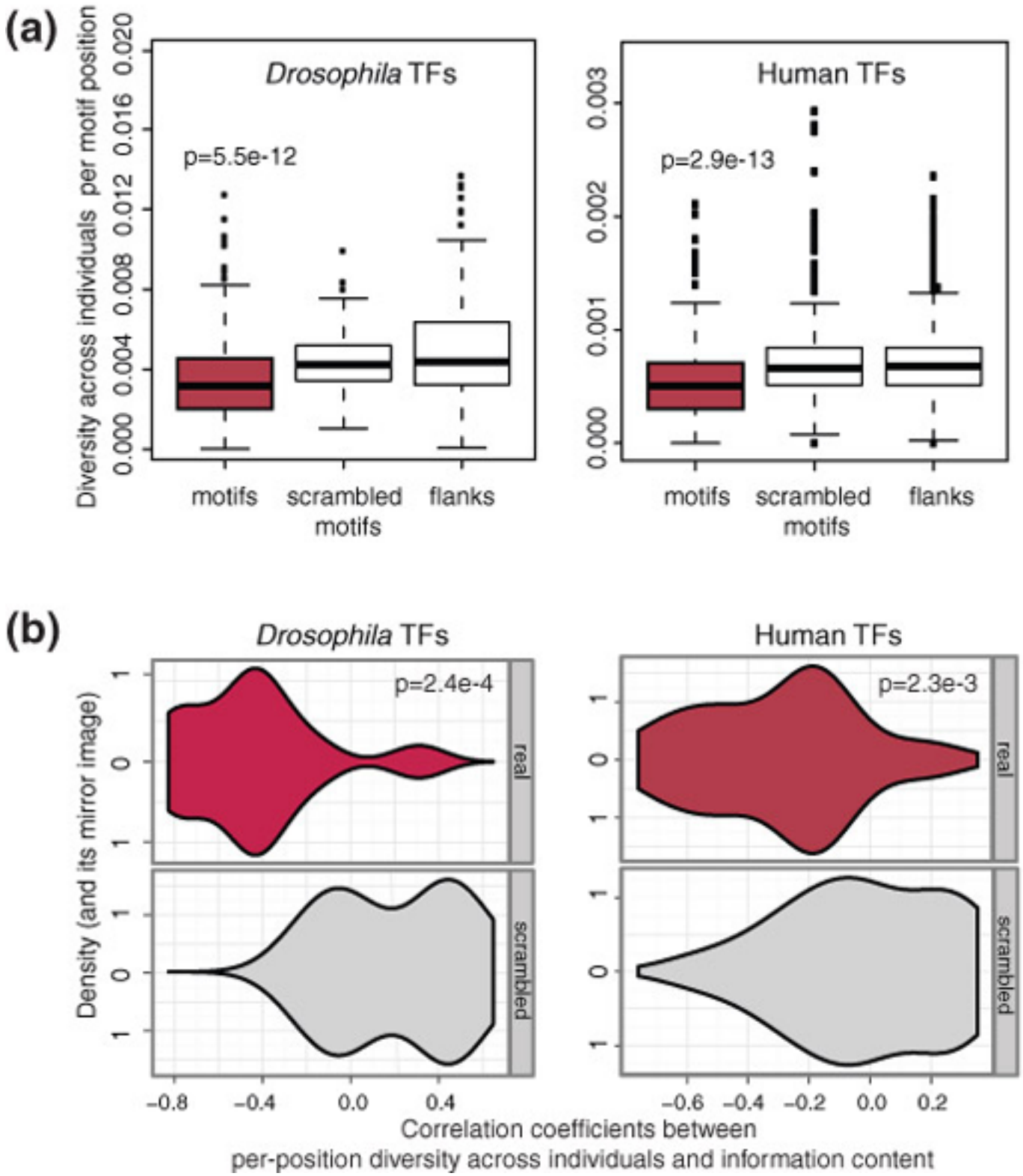


Figure 2 | **Individual variation of the binding sites for 15 *Drosophila* and 36 human TFs selected for this study.** (**a**) Distributions of position-wise diversity at motif positions (red), scrambled motifs and motif flanks at the TF-bound regions of *Drosophila* (left panel) and human (right) TFs; p-values are from Kruskal-Wallis non-parametric significance tests. (**b**) Violin plots (a combination of boxplots and two mirror-image kernel density plots) showing the correlation between individual variation and information content per motif position for the bound instances of *Drosophila* (left) and human (right) TFs included in this study (top, red) and their scrambled versions detected within the same bound regions (bottom, grey); p-values are from Wilcoxon two-sample non-parametric significance tests.

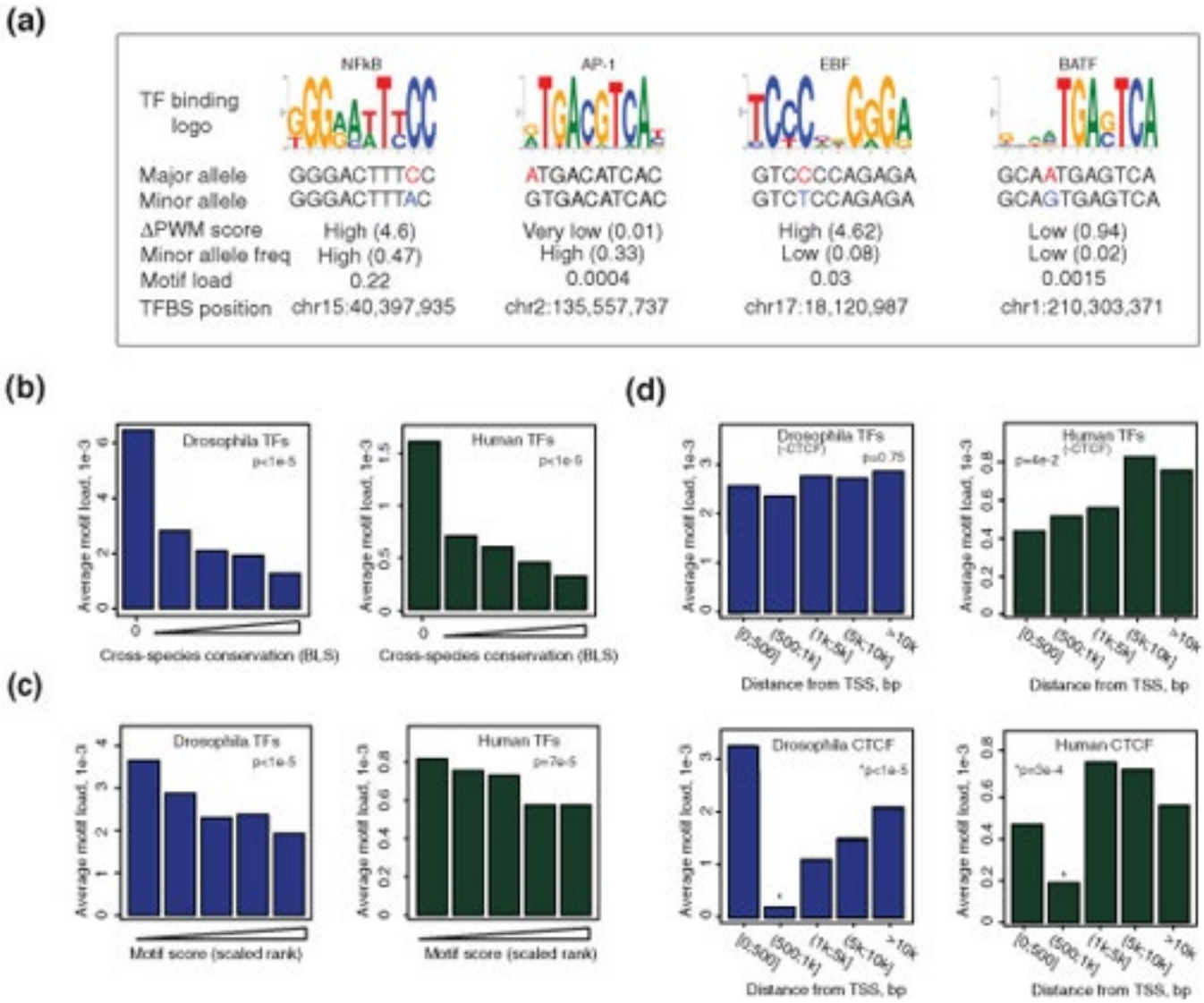


Figure 3 | Motif mutational load of *Drosophila* and human TFBSs located within different genomic contexts. (a) Examples of mutational load values for individual instances of four human TFs, (ranging from high to very low) showing different combinations of parameters that are combined in this metric: the reduction of Position Weight Matrix match scores at the minor allele ('△PWM score') and the number of genotypes within the mutation in the population (MAF; minor allele frequency). (b) Relationship between phylogenetic conservation and motif mutational load for *Drosophila melanogaster* (left) and human (right) TFs included in this study. Conservation is expressed as per-instance Branch Length Scores (BLS) for each instance computed against the phylogenetic tree of 12 *Drosophila* species. The average load for *D. melanogaster*-specific sites (BLS=0) is shown separately as these have an exceptionally high motif load. (c) Relationship between motif stringency and motif load in *Drosophila* (left) and humans (right). Motif stringency is expressed as scaled ranked PWM scores grouped into five incremental ranges of equal size (left to right), with average motif load shown for each range. (d) Relationship between distance from TSS and motif load in *Drosophila* (left) and humans (right) for all analysed TFs excluding CTCF (top) and for CTCF alone (bottom), with average motif load shown for each distance range. (b-d) Average motif load is computed excluding a single maximum value to reduce the impact of outliers. The p-values are from permutation tests, in which permutations are performed separately for each TF and combined into a single statistic as described in Materials and methods.

At position 4 of the YY1 motif, we observe that mean conservation is lower among the activating sites compared to the repressing sites ($p < 0.06$ using a Wilcoxon rank sum test). We also note that, while both activation- and repression-associated classes of YY1 binding sites show greater conservation over the binding site, relative to flanking regions, the conservation of the repression-associated class is greater than that of the activation-associated class, even beyond the 5' and 3' ends of the YY1 motif.

We also found that the recognition sequences for a small number of factors were consistently linked with elevated chromatin accessibility across all classes of sites and all cell types (Supplementary Fig. 6c), indicating that regulators acting through these sequences are key drivers of the accessibility landscape.

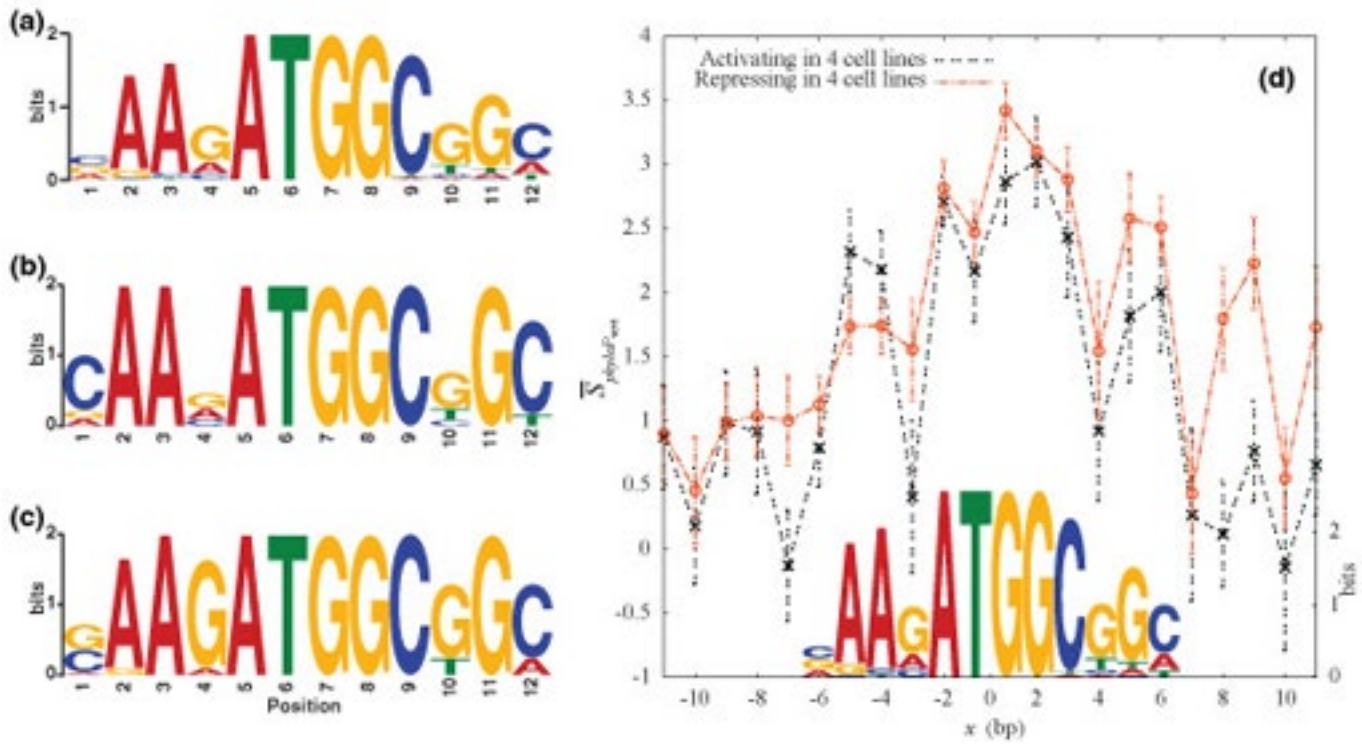


Figure 2 | **Characterization of functional YY1 binding sites. Sequence logo [102] for YY1 binding sites from (a) PWM and sites that are functionally (b) ubiquitously activating (9 BS) or (c) ubiquitously repressive (16 BS) in four human cell lines. In (d), we plot the mean vertebrate phyloP conservation score [90] around functional YY1 binding sites. The mean score, , was computed at each base for sites where the binding event ubiquitously activated (black line) or repressed (red line) transcription in all four cell lines. The position weight matrix that was used to predict YY1 binding sites is shown (scale on the right axis).**

Many eukaryotic genes are co-regulated by multiple TFs in a cell type-specific manner (Maston *et al.* 2006). For 70 of the 87 sequence-specific TFs, we discovered the canonical motifs as well as significant secondary motifs that were distinct from the canonical motifs of the TFs in question and that correspond to the canonical motifs of other TFs. Two scenarios may result in secondary motifs: two TFs bind to neighboring sites (co-binding), or one TF protein binds to another that in turn binds to DNA (tethered binding). To distinguish between these scenarios, we computed the percentages of peaks in a ChIP-seq dataset that contain sites for the canonical TF only, a non-canonical TF only, or both, and then we sorted the datasets by the percentages of peaks with only non-canonical motif sites (Fig. 2a; see Table S3 for the underlying data). We reasoned that if sites of a non-canonical motif were frequently found to be in the same ChIP-seq peaks as canonical motif sites (hence adjacent to them), the two TFs are likely to interact at the protein level and influence each other in binding to their DNA sites. Conversely, if the majority of the peaks contain only sites for non-canonical motifs, then tethered binding is a more plausible model. In this fashion, we identified 151 potential tethered binding and 104 co-binding sequence-specific TF pairs (255 in total). We then compared the pairs we discovered with experimentally detected pairs reported in a mammalian-two-hybrid study (Ravasi *et al.* 2010; Matys *et al.* 2003) and in the BIOGRID database (Badis *et al.* 2009; Stark *et al.* 2006) and found evidence for physical interaction for 27 (10.6%) of the pairs. Eighteen of the 151 tethered binding predictions were validated in the mammalian-two-hybrid data. We randomly picked 151 TF pairs for 5,000 trials and on average 4.19 pairs were validated in the mammalian-two-hybrid experiments (maximum 13 pairs), indicating that our predicted TF pairs were highly significant (p-value$<$2e-4). Thus our results both recapitulated previously reported observations and revealed novel potential interactions that can be tested by experimentation (see Table S3 for summary of all pairs).

Co-binding TFs bind to neighboring sites in the genome. For some TFs, multiple molecules of the same TF also can occupy neighboring sites. We asked whether these neighboring sites prefer to be on the same strand or opposite strands, and whether they prefer to be in a specific range of distances. In addition to the analysis presented in the previous section, which compared the canonical motif with each non-canonical motif

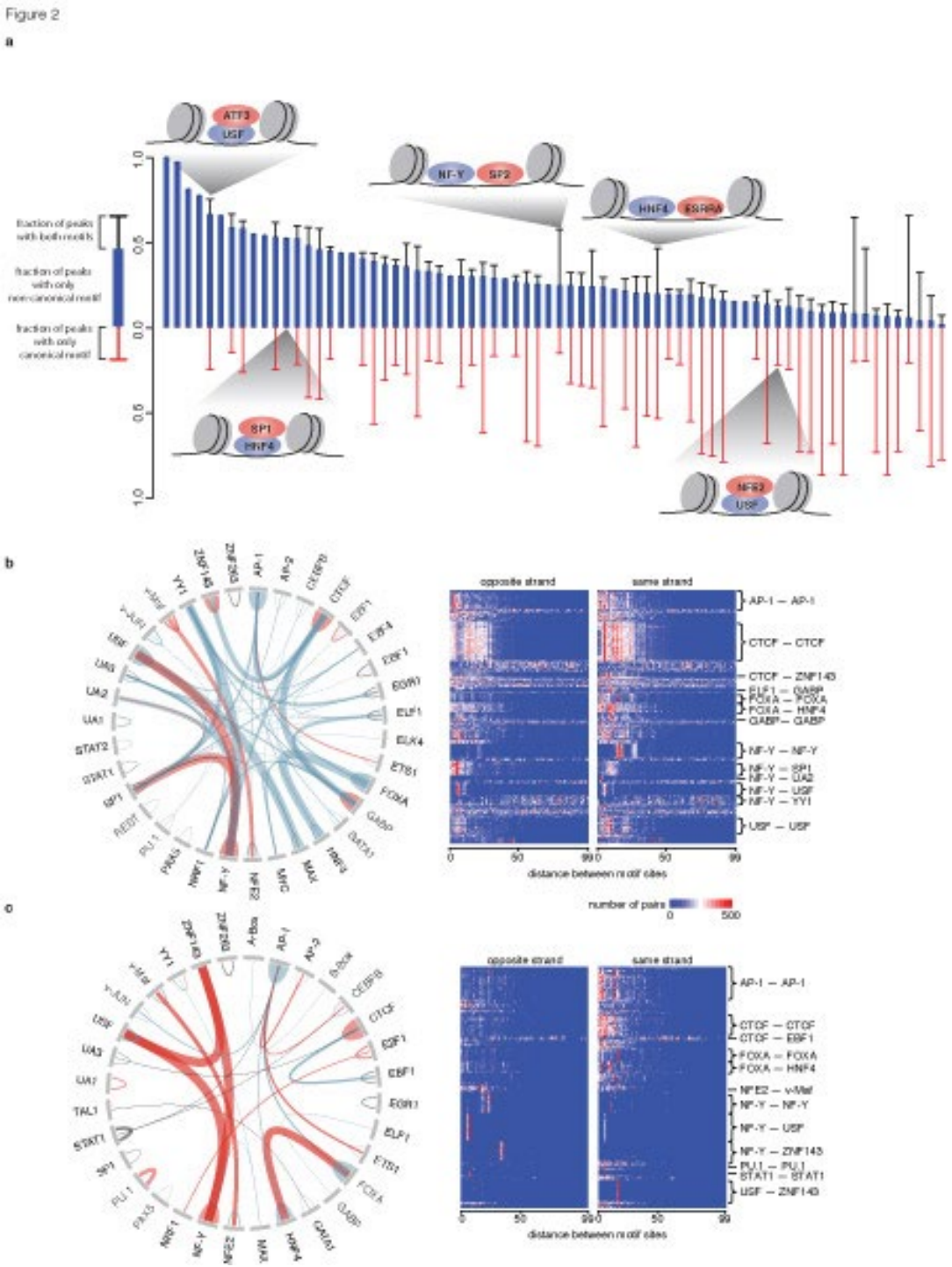


Figure 2 | **Interactions between TFs.** (**a**) Different modes of interaction between TFs are shown. Each bar indicates the canonical TF and one noncanonical TF whose motifs were identified in the same ChIP-seq data set, and the red, blue, and black segments of the bar indicate percentage of peaks in the ChIP-seq data set that contain only canonical motif sites, only noncanonical motif sites, or both. Cartoons depict examples of different models for TF-TF interactions. (**b**) Circos plot (Krzywinski *et al*. 2009) on the *left* depicts pairs of motifs (connected by an arch) with significant distance preferences between their sites. The thickness of a connection is proportional to the normalized frequency of the pair. A connection is depicted as blue, black, or red when the motif pair is discovered in different data sets, the same data set, or both, respectively. The heat map on the *right* shows the distributions of distances between motif pairs. Each row is a motif pair in a particular ChIP-seq data set, and each column represents an edge-to-edge distance (from 0 bp to 99 bp). (**c**) Similar to **b** except showing motif pairs discovered in repetitive regions.

Figure 3

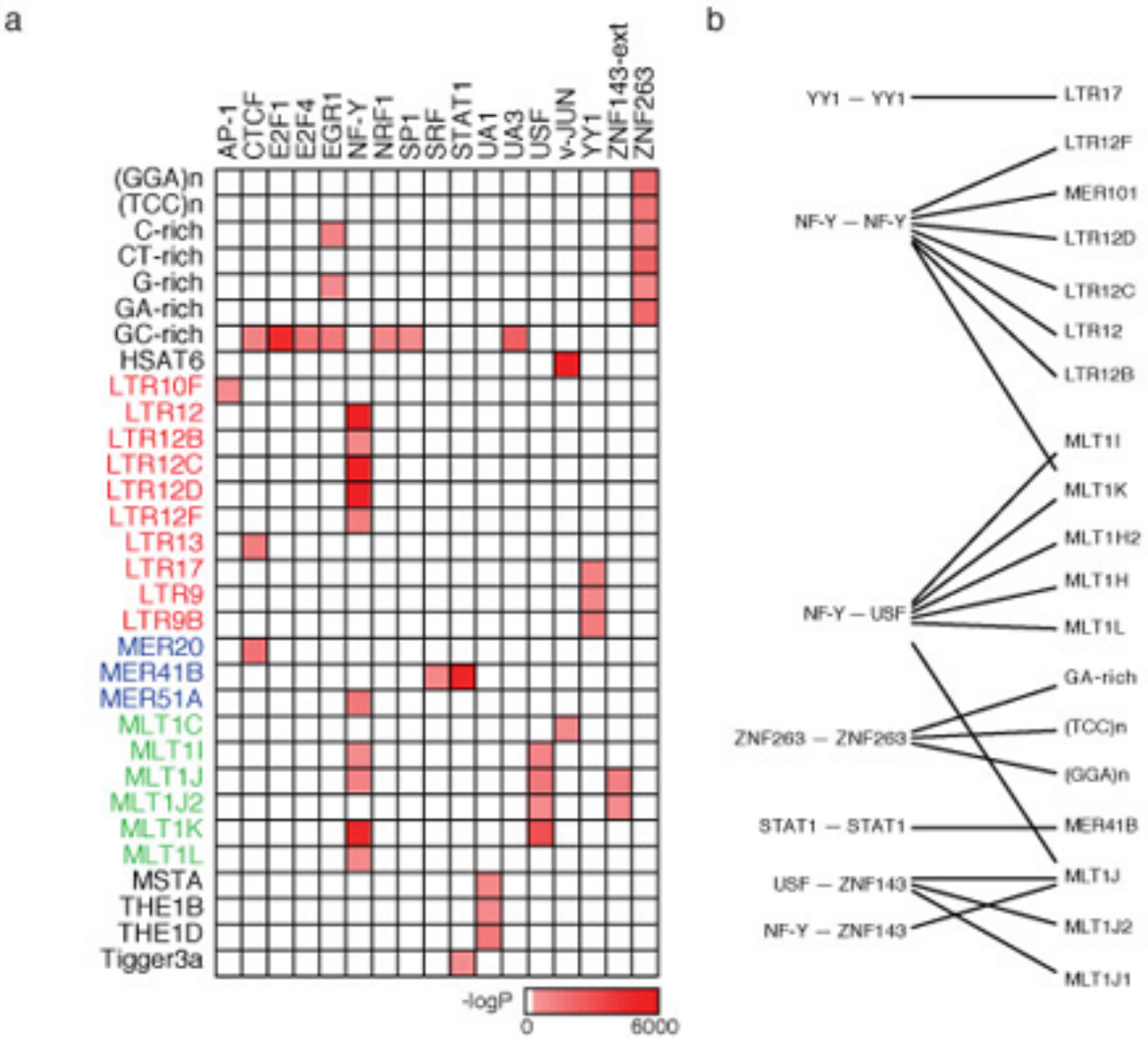


Figure 3 | **Binding sites of certain TFs or TF pairs are enriched in repeats.** (**a**) Enrichment of TF binding sites in repetitive elements. The redness of each grid point is proportional to the negative logarithm of enrichment *P*-value. Repetitive elements are color-coded by family. (**b**) Enrichment of motif pairs that strongly prefer a narrow distance range in various repetitive elements (Fig. 2C)

discovered in the same dataset, we also compared motifs discovered in different datasets collected using the same cell line. In Fig. 2b,c we summarize the heterotypic and homotypic TF pairs that show statistically significant orientation or distance preferences, separately in non-repetitive and repetitive regions of the genome (the underlying data are in Table S4). Out of the 78 motifs discovered from ChIP-seq datasets, 36 motifs (92 pairs; 62 heterotypic pairs and 30 homotypic pairs) are included in Fig. 2b, suggesting that preferred arrangements of nearby TF binding sites is a common phenomenon. The neighboring sites for many heterotypic TF pairs (e.g., CTCF-NF-Y, ELF1-GABP, and FOXA-HNF4) as well as the neighboring homotypic sites of many TFs (e.g., AP-1, CTCF, and USF) show a strong preference for an edge-to-edge distance of less than 30 bp and varying degrees of preference for one orientation over the other. For example, neighboring NF-Y sites prefer to be in the same orientation. NF-Y also prefers one orientation to the other when co-binding with SP1, PBX3 (its motif is UA2), and USF. We hypothesized that these 92 TF pairs are more likely to represent protein-protein interactions than the TF pairs we identified in the previous section without testing for position or orientation preferences.

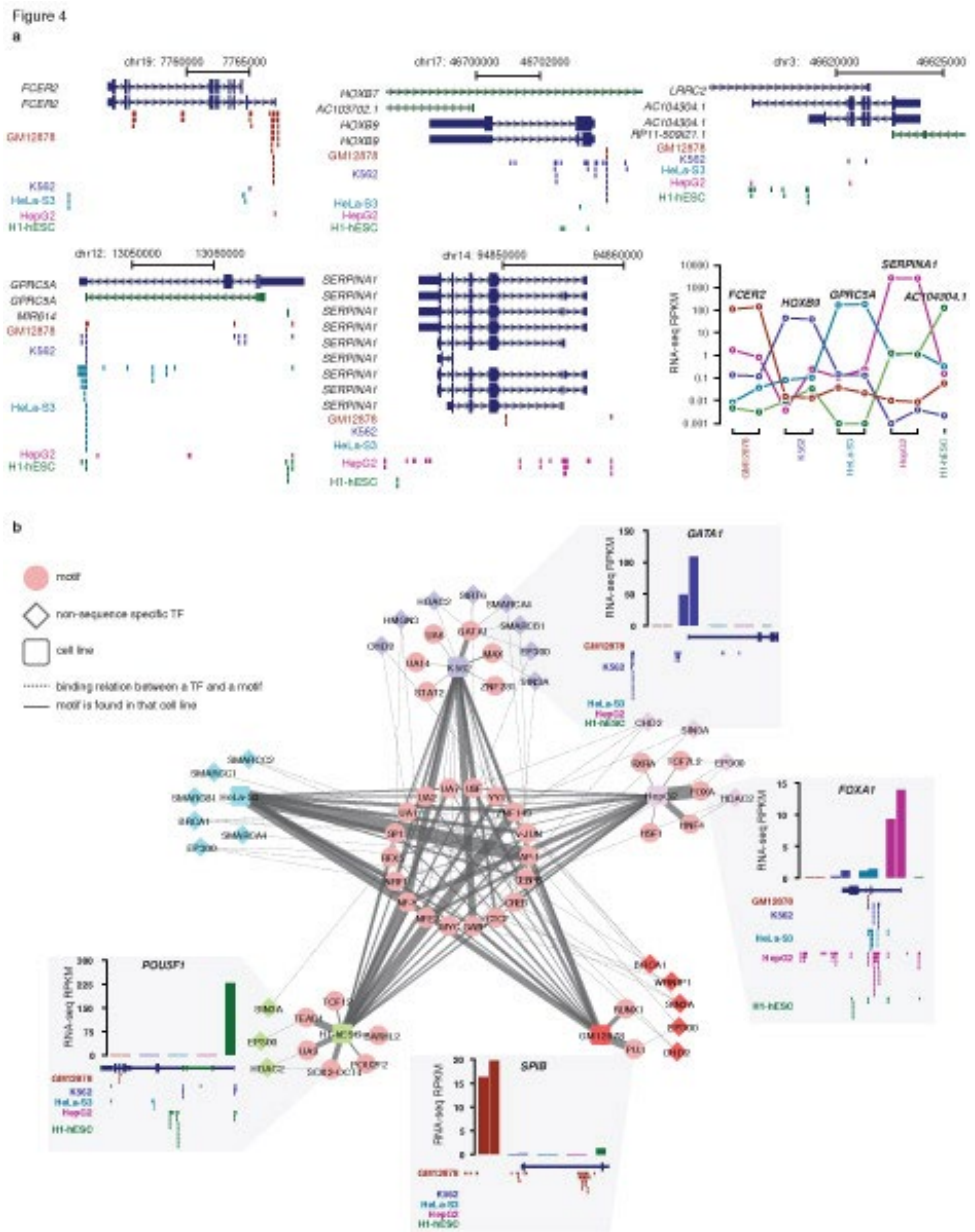


Figure 4 | **Cell-type-specific binding of sequence-specific and non-sequence-specific TFs.** (**a**) Abundant TF binding sites are observed near cell-line-specific transcripts. Binding sites are shown as vertical bars and colored by cell line (dark blue for K562, red for HepG2, brown for GM12878, green for H1-hESC, and cyan for HeLa-S3). (*Bottom*, *right*) Expression levels (in RPKM) for example cell-line-specific transcripts across the five cell lines with the most ChIP-seq data. (**b**) Secondary motifs identified in the ChIP-seq data sets of sequence-specific TFs and their enrichment in the ChIP-seq peaks of non-sequence-specific TFs. The five cell lines are indicated with color-coded squares, noncanonical motifs of sequence-specific TFs are shown in pink circles and a solid line connecting each motif to the respective cell line. The thickness of the solid line is proportional to the normalized frequency in which a noncanonical motif is discovered in a particular cell line. Non-sequence-specific TFs are shown in diamonds whose colors match the color of the cell line if there is a ChIP-seq data set of the TF in that cell line. Dashed lines connect non-sequence-specific TFs and noncanonical motifs, indicating that a noncanonical motif of a sequence-specific TF is enriched in the ChIP-seq peaks of the non-sequence-specific TF. (Four *insets*) Expression profiles of sequence-specific TFs whose canonical motifs are found to be specific to a cell type and the TF binding sites around the genes that encode these TFs in the appropriate cell line. The expression levels in each cell line are assigned a similar color as the cell line. For four cell lines, two biological replicates were available for RNA-seq data; hence, there are two bars for each of these cell lines. Only one biological replicate was available for H1-hESC.

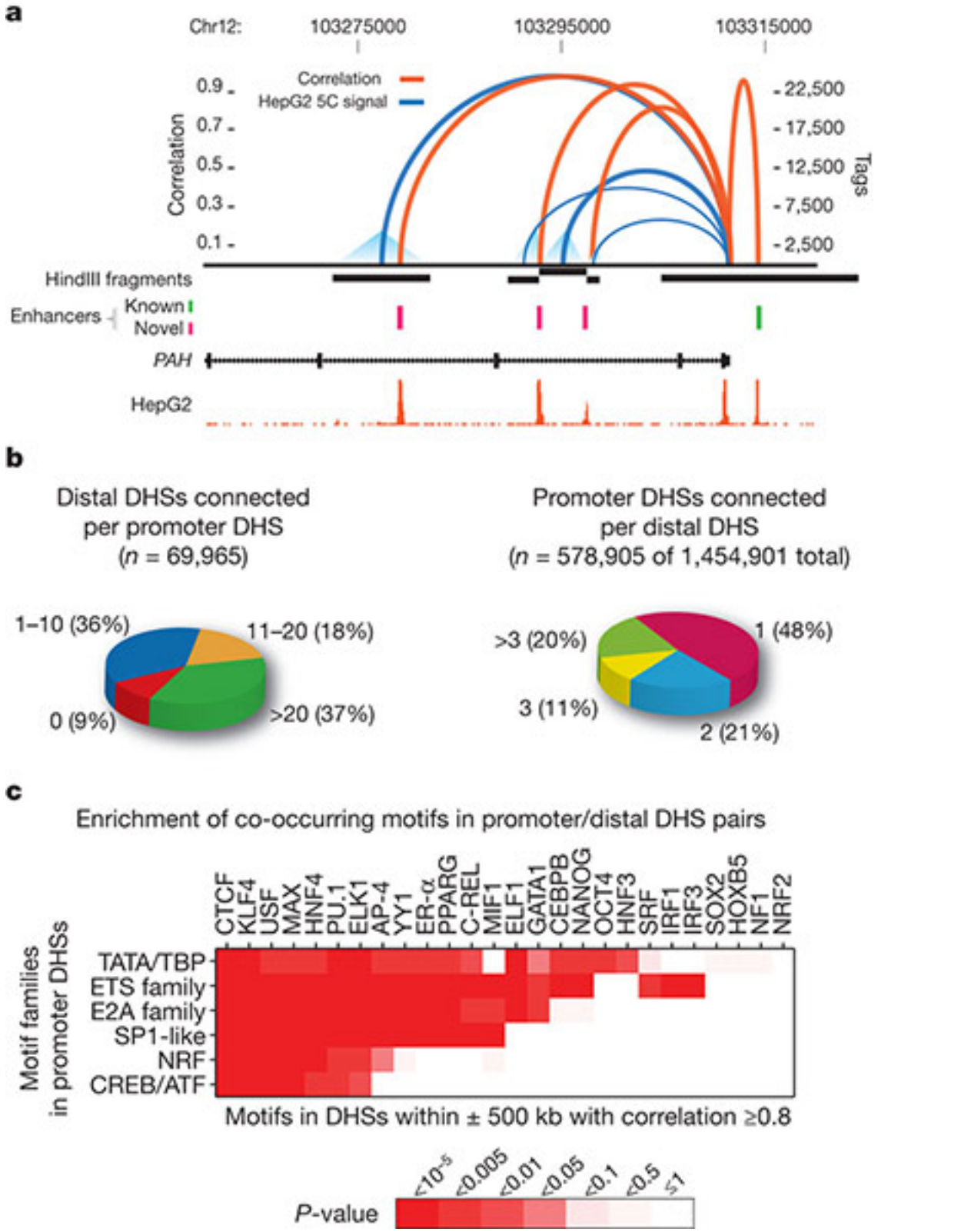


Figure 5 | A genome-wide map of distal DHS-to-promoter connectivity. (a) Cross-cell-type correlation (red arcs, left *y* axis) of distal DHSs and *PAH* promoter closely parallels chromatin interactions measured by 5C-seq (blue arcs, right *y* axis); black bars indicate HindIII fragments used in 5C assays. Known (green) and novel (magenta) enhancers confirmed in transfection assays are shown below. Enhancer at far right is not separable by 5C as it lies within the HindIII fragment containing the promoter. (b) Left: proportions of 69,965 promoters correlated ($r>0.7$) with 0 to >20 DHSs within 500 kb. Right: proportions of 578,905 non-promoter DHSs (out of 1,454,901) correlated with 1 to >3 promoters within 500 kb. (c) Pairing of canonical promoter motif families with specific motifs in distal DHSs.

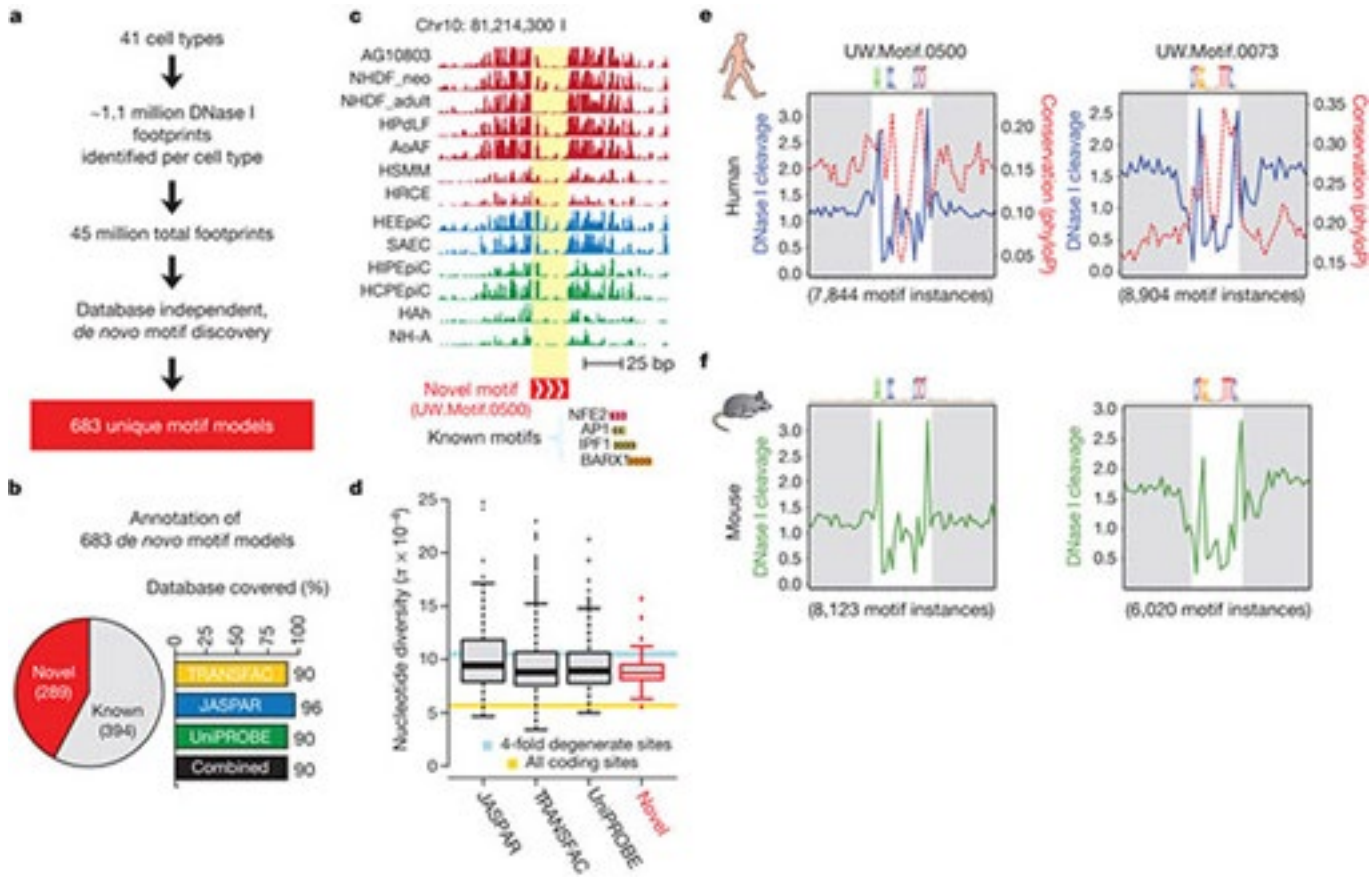


Figure 6 | ***De novo* motif discovery expands the human regulatory lexicon.** (**a**) Overview of *de novo* motif discovery using DNase I footprints. (**b**) Annotation of the 683 *de novo*-derived motif models using previously identified transcription factor motifs. A total of 394 of these *de novo*-derived motifs match a motif annotated within the TRANSFAC, JASPAR or UniPROBE databases, whereas 289 are novel motifs (pie chart). The *de novo* consensus matching TRANSFAC, JASPAR or UniPROBE sequences cover the majority of each database (bar chart). (**c**) Example of a DNase I footprint found in multiple cell types that is annotated solely by one of the novel *de novo*-derived motifs. (**d**) Box-and-whisker plot comparing the average nucleotide diversity at instances of the 289 novel *de novo*-derived motif models to instances of motifs present in databases of known specificities (*x* axis). The box defines the 25% and 75% percentiles and the whiskers display 1.5 times the inner quartile range of the distribution of π values in each respective database. The blue bar indicates the average nucleotide diversity (π) at fourfold degenerate coding sites (width is equal to 95% confidence interval); gold bar indicates π at all coding sites (width is equal to 95% confidence interval). (**e**) Phylogenetic conservation (red dashed) and per-base DNase I hypersensitivity (blue) for all DNase I footprints in dermal fibroblast cells matching two novel *de novo*-derived motifs. The white box indicates width of consensus motif. (**f**) Per-nucleotide mouse liver DNase I cleavage patterns at occurrences of the motifs in (**e**) at DNase I footprints identified in mouse liver.

Indeed, 14 heterotypic pairs and 17 homotypic pairs (33.7%) were detected in the aforementioned mammalian-two-hybrid study (Ravasi *et al.* 2010) or in the BIOGRID database (Stark *et al.* 2006).

TFs tend to bind gene-rich regions of the genome due to their role in regulating target gene expression (Carroll *et al.* 2006). Nonetheless, repetitive elements are known to harbor functional TF binding sites, especially when such elements occur near genes. We systematically compared our compilation of TF binding sites with all repeats annotated in the human genome and the results are summarized in Fig. 3a. We confirmed the previously reported enrichment of STAT1, NF-Y, and CTCF binding sites in various repetitive elements (Bourque *et al.* 2008; Schmid and Bucher 2010), and we uncovered many more TFs whose binding sites are enriched in certain repetitive elements, e.g., UA1 sites in THE1B and THE1D retrotransposons. It was shown that a long

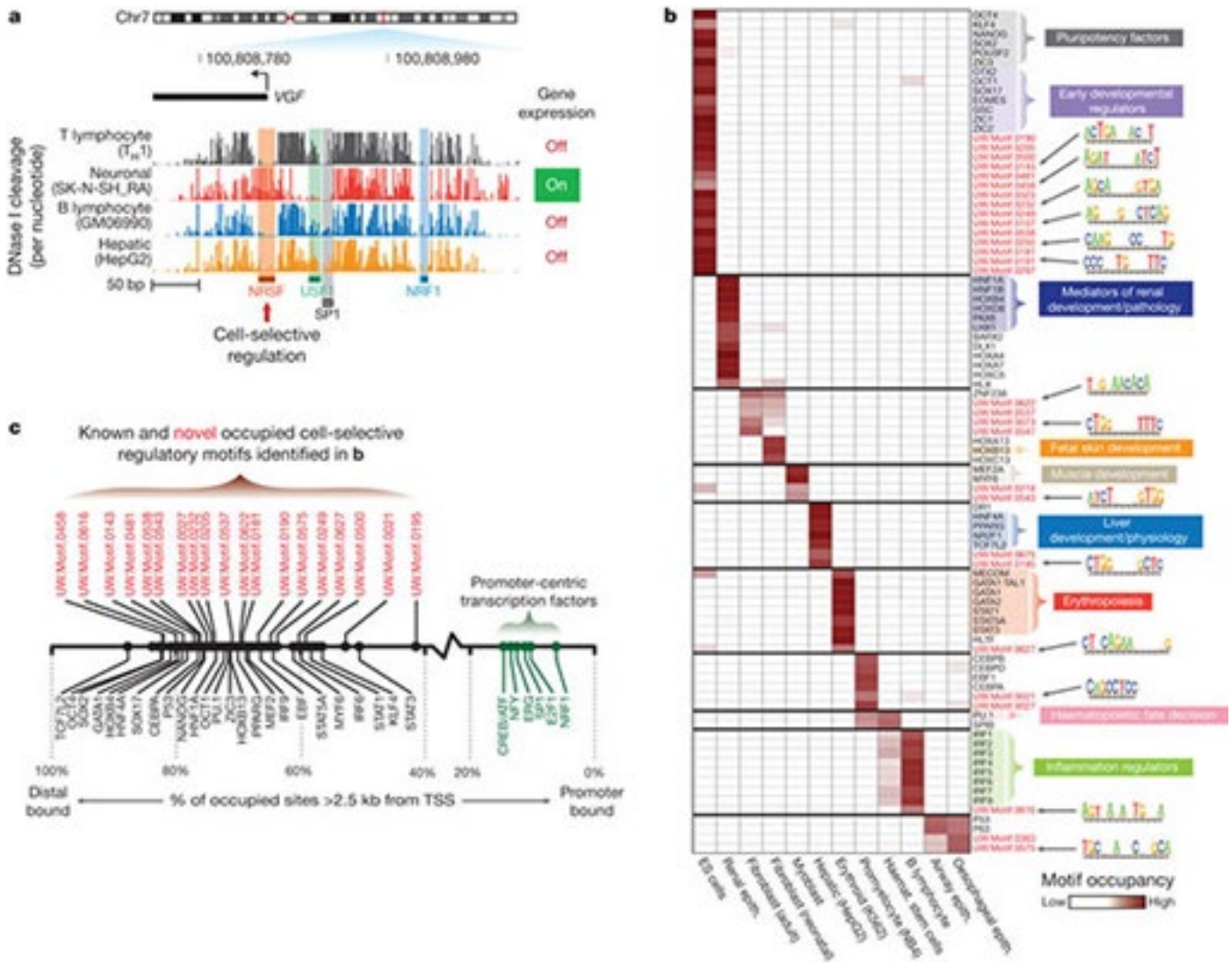


Figure 7 | **Multi-lineage DNase I footprinting reveals cell-selective gene regulators.** (**a**) Comparative footprinting of the nerve growth factor gene (*VGF*) promoter in multiple cell types reveals both conserved (NRF1, USF1 and SP1) and cell-selective (NRSF) DNase I footprints. (**b**) Shown is a heat map of footprint occupancy computed across 12 cell types (columns) for 89 motifs (rows), including well-characterized cell/tissue-selective regulators, and novel *de novo*-derived motifs (red text). The motif models for some of these novel de novo-derived motifs are indicated next to the heat map. (**c**) The proportion of motif instances in DNase I footprints within distal regulatory regions for known (black) and novel (red) cell-type-specific regulators in (**b**) is indicated. Also noted are these values for a small set of known promoter-proximal regulators (green). ES, embryonic stem.

terminal repeat (LTR) region of the THE1D retrotransposon was recruited as an alternative promoter for the human *IL2RB* gene and that the activity of this alternative promoter is regulated by DNA methylation (Cohen *et al.* 2011). The UA1 motif we identified in ZBTB33 peaks contains a prominent CGCG center (Fig. 1c) and ZBTB33 is known to bind methylated CpG dinucleotides (Yoon *et al.* 2003), raising the interesting possibility that the THE1B/D retrotransposons spread ZBTB33 binding sites across the genome, and that the regulation of the newly recruited target genes can be modulated by the DNA methylation mechanism. Fig. 2c and 3b summarizes all motif pairs that show statistically significant distance or orientation preference in repetitive regions of the genome. The NF-Y-USF site pairs that typically have an end-to-end distance of 5-6 bp are nearly all located in the MLT1 family of retrotransposons. Similarly, the NF-Y-NF-Y site pairs at a 9-bp distance are found most often in LTR12 retrotransposons. There are 181 copies of the MLT1J transposon in the genome that contain sites for the NF-Y, USF, and ZNF143 motifs simultaneously, bound directly by NF-Y, USF, and ZNF143 TFs, respectively. The relative distance among the sites are nearly invariant (Fig. 2c), indicating recent duplications of MLT1J. Our

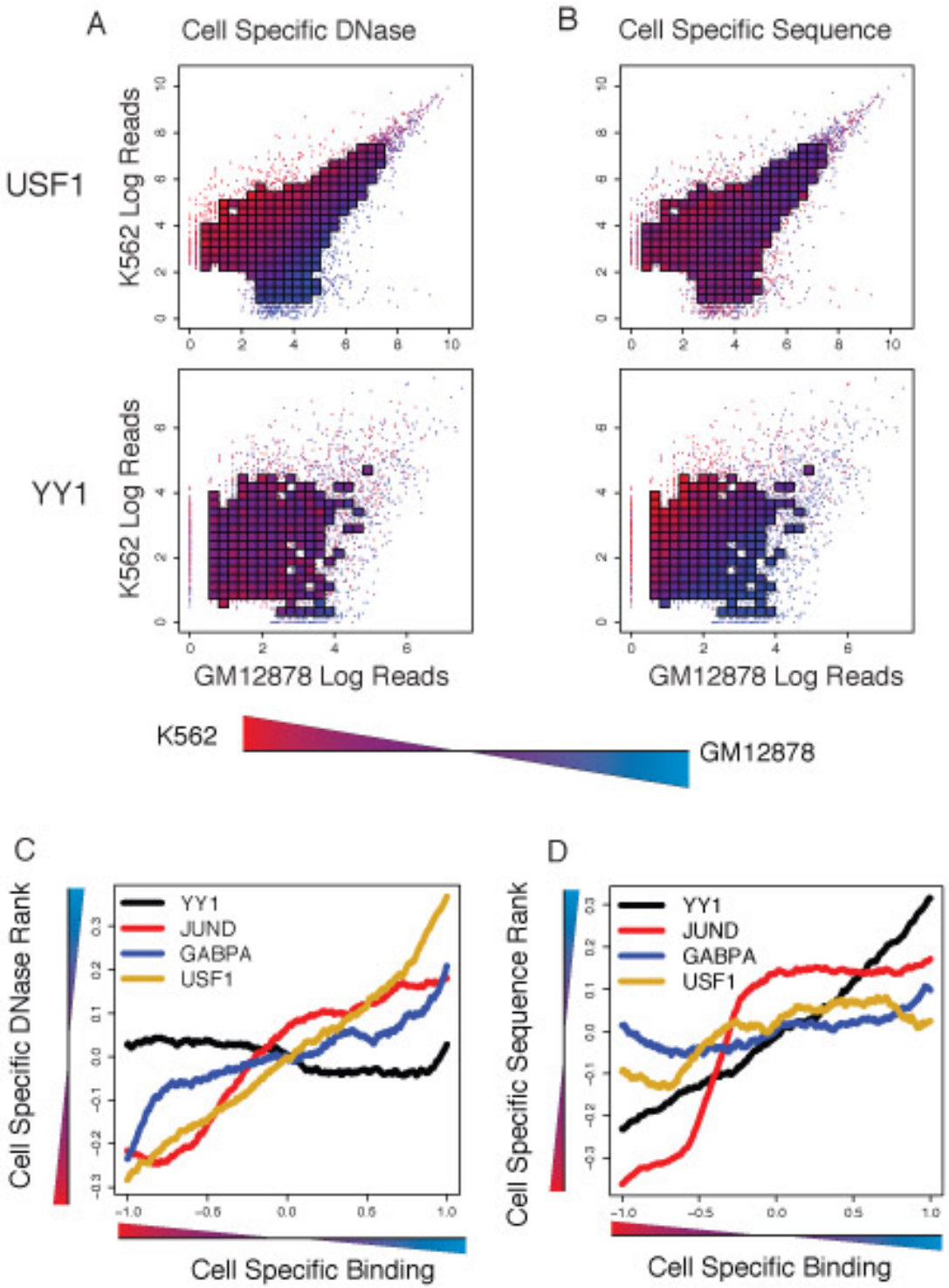


Figure 6 | Cell-type-specific TF binding is associated with differential DNase accessibility, sequence signal, or both. (a) Differential DNase accessibility (color) is shown for K562 versus GM12878 with respect to cell-type-specific binding (x-axis for GM12878; y-axis for K562). Each point represents a single binding site, and if there are a sufficient number of points in a region, their value is averaged and appears as a square. DNase accessibility, as measured by read-counts, for USF1 (top) correlates with cellspecific binding. This contrasts with YY1 (bottom), where DNase accessibility is evenly distributed across cell-type-specific and nonspecific peaks. (b) Differential sequence preference (color) is shown for K562 versus GM12878. *k*-mer SVM models are learned from K562 and GM12878 binding sites, and their differential scores are shown by color gradient. For YY1, but not USF1, we see that the differential *k*-mer SVM scores distinguish cell-type-specific binding sites. (c) Binding sites with differential TF occupancy also have differential DNase accessibility. Each line represents a TF that has been assayed in GM12878 and K562. The x-axis plots a ranking from the most K562-specific binding site to the most GM12878-specific binding sites, based on cell-to-cell log read count ratios, while the y-axis shows the difference in DNase-accessibility ranks in GM12878 and K562. The line plot is smoothed using the mean over a window of 500 binding sites. (d) For the same TFs, we plot the difference in K562- and GM12878-specific *k*-mer SVM score ranks (y-axis) as a function of the ranking of cell-to-cell log read count ratios, from the most K562-specific binding site to the most GM12878-specific binding sites. The line plot is smoothed using the mean over a window of 500 binding sites.

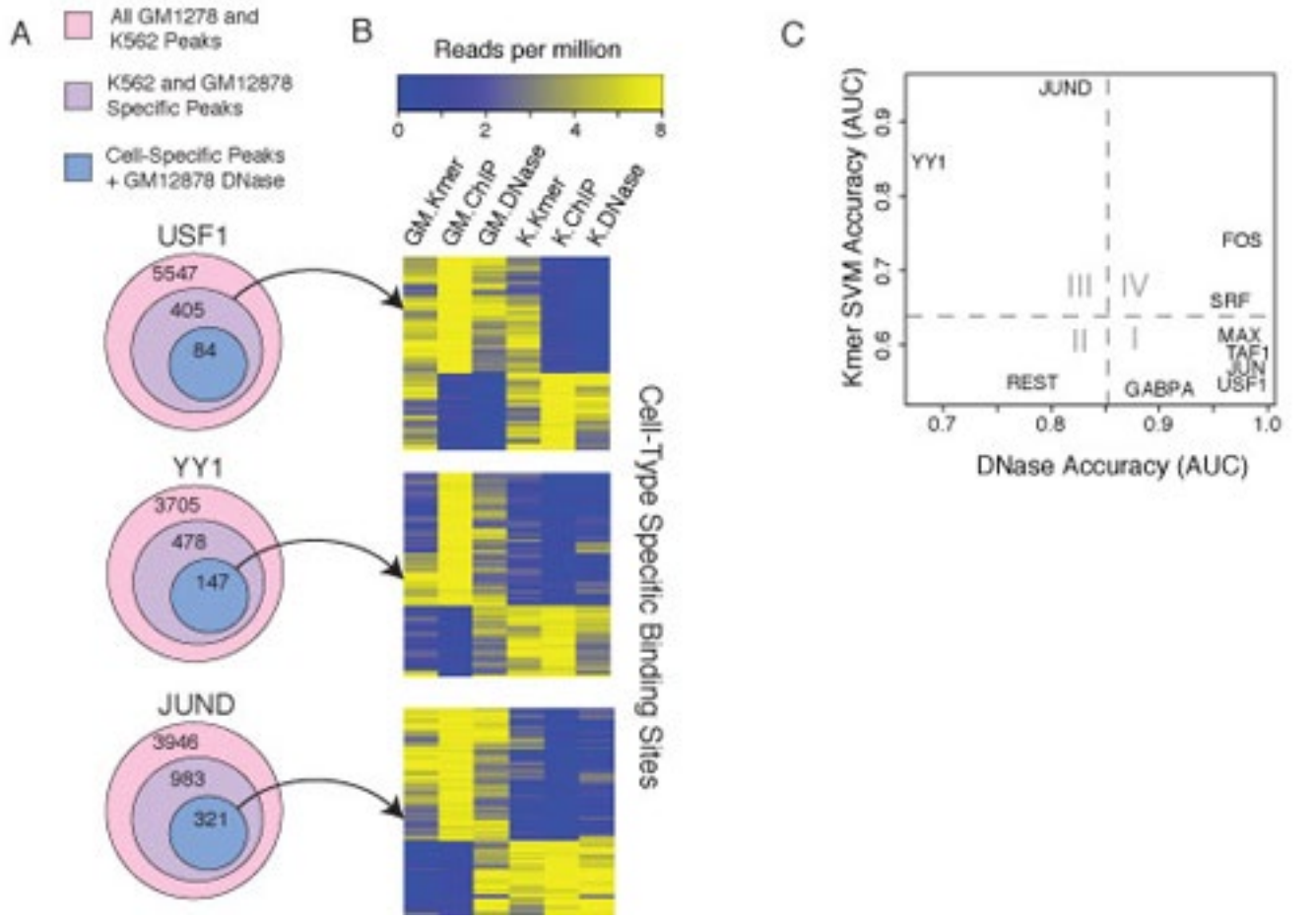


Figure 8 | Cell-type-specific sequence models can predict cell-type-specific binding at loci that are DNase accessible in both cell lines. (a) The number of binding sites, cell-type-exclusive binding sites, and exclusive binding sites that are DNase accessible in GM12878. (b) Cell-type-exclusive binding sites can be explained by cell-type-specific sequence preferences when a binding site is accessible in both cell lines. Cell-type-exclusive binding sites for USF1, YY1, and JUND are shown. For USF1, all GM12878- and K562-exclusive binding sites are shown, and DNase accessibility is able to explain cell-type-exclusive binding. In contrast, for JUND and YY1, there are cell-type-exclusive binding sites in GM12878 and K562 that are DNase accessible in both cell lines, and only these examples are plotted in the middle and bottom heatmaps. For these examples, the cell-type-specific SVM sequence scores can explain the celltype-specific binding. (c) AUC values for the task of discriminating between GM12878-exclusive peaks and K562-exclusive peaks by differential DNase reads (x-axis) or by cell-type-specific SVM sequence scores. For the SVM models, the GM12878- and K562-specific models were each used to discriminate between GM12878- and K562-exclusive binding sites, and the mean AUC over both models was reported. Binding site sequences used in training the models were held out of test sets for this evaluation. For most TFs, the cell-type-exclusive binding sites are well-predicted by differential DNase accessibility (I, IV). For REST, DNase is not predictive in general and the SVM models are consistent between the two cell lines (III). For JUND and YY1, DNase is not predictive of cell-type-exclusive binding, as many sites are DNase accessible in both cell lines; however, the cell-type-specific peaks tend to have different underlying *k*-mer sequences, enabling accurate discrimination by cell-type-specific SVM sequence models.

results suggest a mechanism whereby retrotransposons amplify functional TF site pairs across the genome through transposition, potentially bringing new genes under the regulation of those TFs.

The majority of the ENCODE ChIP-seq data were produced using five cell lines-K562, GM12878, HepG2, H1-hESC, and HeLa. Integrating ChIP-seq data with RNA-seq data for these five cell lines, we asked whether genes that are preferentially expressed in a given cell line (defined by the average expression level in one cell line being more than 10-fold higher than that in any of the remaining four cell lines) show enriched TF binding sites in the corresponding cell line. This is indeed the case for a large fraction of genes and Fig. 4a shows five examples, one per cell line.

We then asked whether the non-canonical motifs we discovered also reflect cell type specificity. Fig. 4b plots the non-canonical motifs (circles) detected in the ChIP-seq datasets of sequence- specific TFs for each of the five

cell lines (squares) with the most ENCODE ChIP-seq datasets. Cell line-specific non-canonical motifs are placed close to their respective cell lines in Fig. 4b. We defined cell line-specific motifs as those that were discovered three times more often in one cell line than in any other cell line. The remaining non-canonical motifs are placed in the center of the figure, and these motifs correspond to TFs that cooperate with other sequence-specific TFs across multiple cell lines. The thickness of the solid line connecting a non-canonical motif to a cell line indicates the proportion of datasets in that cell line that revealed the motif as a non- canonical motif.

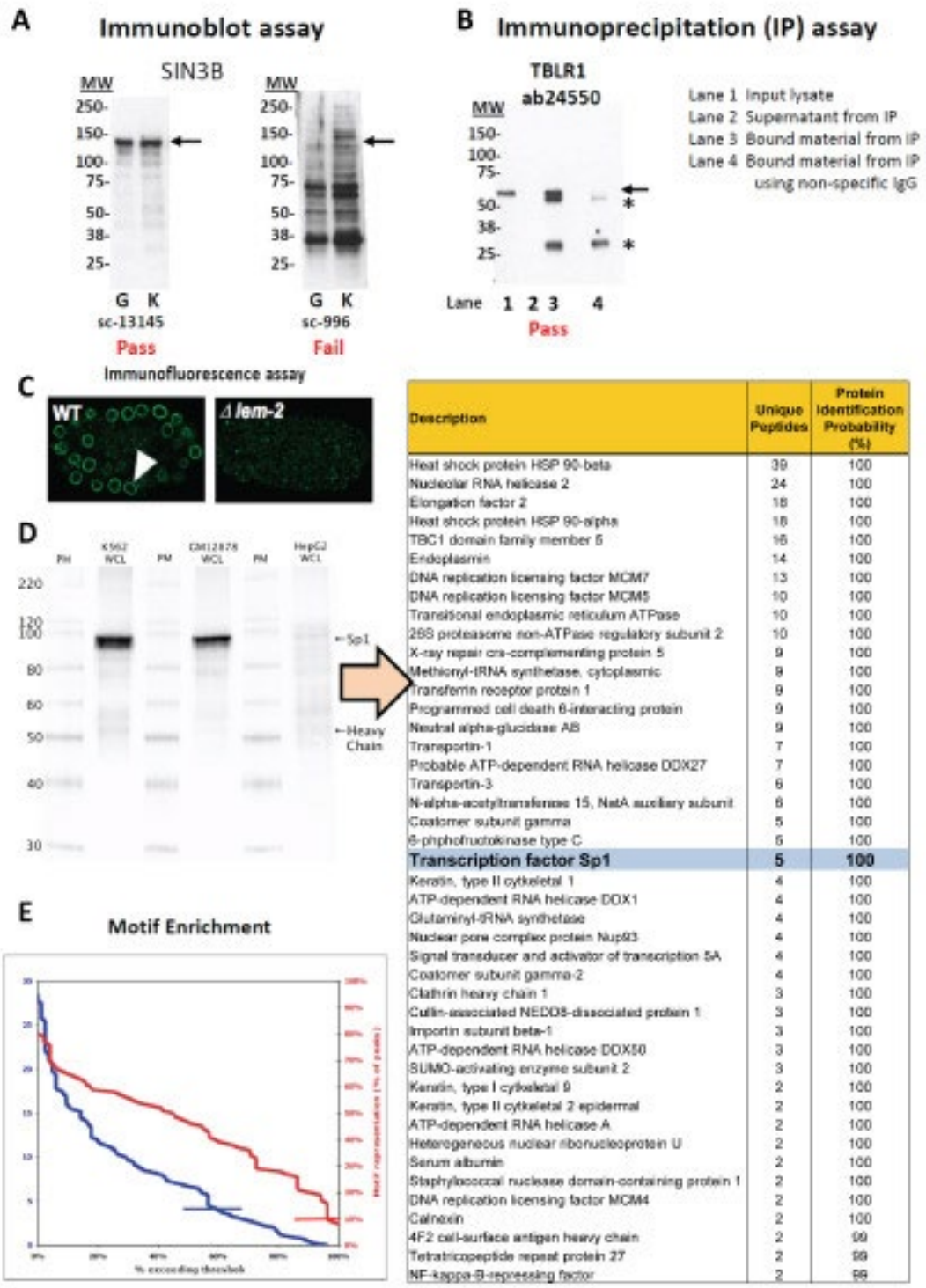
In Fig. 4b, we also included all non-sequence-specific TFs (diamonds) for which there are ChIP- seq data in these cell lines. Dashed lines connect non-sequence-specific TFs to the motifs discovered in their ChIP-seq peaks. Two non-sequence-specific TFs show cell line-specific enrichment in motifs: the enhancer-binding protein EP300 and the histone deacetylase HDAC2. There are seven datasets for EP300 in seven different cell lines, and three datasets for HDAC2 in three different cell lines. Distinct motifs were found in different cell lines: SPI1 for EP300 in GM12878 cells; GATA1 (and GATA1-ext) for both EP300 and HDAC2 in K562 cells; FOXA and HNF4 for HDAC2, and FOXA and TCF7L2 for P300 in HepG2 cells; SOX2-OCT4 and UA9 for HDAC2, and TEAD1 for EP300 in H1-hESC cells; and CEBPB, AP-1, and CREB for EP300 in HeLa cells. As described in the previous section, many of these motifs were most frequently and specifically observed as secondary motifs for sequence-specific TFs in the respective cell lines. Because non-sequence-specific TFs do not bind DNA directly, they tether onto sequence-specific TFs to bind target DNA. EP300 is known to interact with AP-1 and CEBPB (Chi-Chung Wang *et al.* 2007; Mink *et al.* 1997), and HDAC2 with TAL1-GATA (the motif is GATA1-ext) (Hu *et al.* 2009). Our results highlight that the interactions of EP300 and HDAC2 with sequence-specific TFs are highly cell type dependent.

We detect systematic relationships between specific combinations of regulatory factors between pair distal-promoter DNaseI hypersensitive sites. For example, KLF4, SOX2, OCT4 (also called POU5F1) and NANOG are known to form a well-characterized transcriptional network controlling the pluripotent state of embryonic stem cells33. We found significant enrichment ($P < 0.05$) of the KLF4, SOX2 and OCT4 motifs within distal DHSs correlated with promoter DHSs containing the NANOG motif; enrichment of NANOG, SOX2 and OCT4 distal motifs co-occurring with promoter motif OCT4; and enrichment of distal SOX2 and OCT4 motifs with promoter SOX2 motifs (Supplementary Fig. 15a).

We also find significant co-associations between promoter types (defined by the presence of cognate motif classes; see Supplementary Methods) and motifs in paired distal DHSs. For example, when a member of the ETS domain family (motifs ETS1, ETS2, ELF1, ELK1, NERF (also called ELF2), SPIB, and others) is present within a promoter DHS, motif PU.1 (also called SPI1) is significantly more likely to be observed in a correlated distal DHS ($P < 10\text{-}5$).

Comprehensive scans of DNaseI hypersensitive regions for high-confidence matches to all recognized transcription factor motifs in the TRANSFAC[10] and JASPAR[11] databases revealed striking enrichment of motifs within footprints ($P \sim 0$, Z-score $= 204.22$ for TRANSFAC; Z-score $= 169.88$ for JASPAR; Fig. 1b and Supplementary Fig. 3).

Given the enrichment of known sequence motifs within footprints, we sought to identify novel motifs within this genomic compartment. We performed *de novo* motif discovery within the ~45 million footprints identified in each of the 41 cell types resulting in 683 unique motif models (Fig. 6a). A total of 394 of the 683 (58%) *de novo* motifs matched distinct experimentally grounded motif models, accounting collectively for 90% of all unique entries across the three databases (Fig. 6b and Supplementary Fig. 14a-c). Notably, 289 of the footprint-derived motifs were absent from major databases (Fig. 6b and Supplementary Fig. 14d). These novel motifs populate millions of DNaseI footprints (Fig. 6c), and show features of *in vivo* occupancy and evolutionary constraint similar to motifs for known regulators, including marked anti-correlation with nucleotide-level vertebrate conservation (Figs 3b, 6e and Supplementary Figs. 8, 15a).



Description	Unique Peptides	Protein Identification Probability (%)
Heat shock protein HSP 90-beta	39	100
Nucleolar RNA helicase 2	24	100
Elongation factor 2	18	100
Heat shock protein HSP 90-alpha	18	100
TBC1 domain family member 5	16	100
Endoplasmin	14	100
DNA replication licensing factor MCM7	13	100
DNA replication licensing factor MCM5	10	100
Transitional endoplasmic reticulum ATPase	10	100
26S proteasome non-ATPase regulatory subunit 2	10	100
X-ray repair cro-complementing protein 5	9	100
Methionyl-tRNA synthetase, cytoplasmic	9	100
Transferrin receptor protein 1	9	100
Programmed cell death 6-interacting protein	9	100
Neutral alpha-glucidase AB	9	100
Transportin-1	7	100
Probable ATP-dependent RNA helicase DDX27	7	100
Transportin-3	6	100
N-alpha-acetyltransferase 15, NatA auxiliary subunit	6	100
Coatomer subunit gamma	5	100
6-phphofructokinase type C	5	100
Transcription factor Sp1	**5**	**100**
Keratin, type II cytkeletal 1	4	100
ATP-dependent RNA helicase DDX1	4	100
Glutaminyl-tRNA synthetase	4	100
Nuclear pore complex protein Nup93	4	100
Signal transducer and activator of transcription 5A	4	100
Coatomer subunit gamma-2	4	100
Clathrin heavy chain 1	3	100
Cullin-associated NEDD8-dissociated protein 1	3	100
Importin subunit beta-1	3	100
ATP-dependent RNA helicase DDX50	3	100
SUMO-activating enzyme subunit 2	3	100
Keratin, type I cytkeletal 9	2	100
Keratin, type II cytkeletal 2 epidermal	2	100
ATP-dependent RNA helicase A	2	100
Heterogeneous nuclear ribonucleoprotein U	2	100
Serum albumin	2	100
Staphylococcal nuclease domain-containing protein 1	2	100
DNA replication licensing factor MCM4	2	100
Calnexin	2	100
4F2 cell-surface antigen heavy chain	2	99
Tetratricopeptide repeat protein 27	2	99
NF-kappa-B-repressing factor	2	99

Fig. 2

Figure 2 | Motif Analysis. Deriving DNA sequence motifs for a ChIP-assayed factor is relatively simple and has been performed successfully for most ENCODE ChIP-seq datasets (Fig. 2E). Experiments that pass the thresholds we use for NRF, FRiP, and NSC/RSC typically produce thousands to tens of thousands of regions, a sub-sample of which can be readily used to deduce the recognition motif, although more than one motif subfamily is sometimes found by additional analysis (Johnson *et al.*, 2007). Causal motifs are typically centrally positioned and this can be used as a confirming diagnostic (Fig. 6F).

Cell-selective gene regulation is mediated by the differential occupancy of transcriptional regulatory factors at their cognate cis-acting elements. Figure 7b shows a heat-map representation of cell-selective occupancy at motifs for 60 known transcriptional regulators and for 29 novel motifs. This approach appropriately identified a number of

Table 1 | Summary of functional tests of 466 predicted TF binding sites in four human cell lines.Table 1 summarizes our data according to the TFs. The majority of the sites in our tests are bound by six TFs: CTCF, E2F family proteins, GABP, GATA2, STAT1 and YY1 (i.e. the TFBS sequences appear to be directly occupied by these factors, see Methods). These TFs have varying rates of being functional in at least one cell line, with CTCF, E2F family proteins, GABP, and GATA2 all exhibiting a functional verification rate of ~60%, while STAT1 and YY1 had their function verified at rates of 78% and 88%, respectively. However, compared with the other factors in our experiments CTCF and GATA2 had a much lower fraction of functional sites across all four cell lines. In the case of GATA2, this observed lower rate of ubiquitous function may be due to the varying expression levels of GATA family proteins in different cell lines. For example, it has been reported that HepG2 cells do not express GATA2 or GATA3 [76] but do express GATA4 [77] (these observations are broadly confirmed by the ENCODE Consortium RNA sequencing data reported in Additional file 1, Table S1). GATA6 is highly expressed in colon cancer epithelial cells [46], such as HCT-116. Since CTCF is broadly expressed, the relatively low rate of ubiquitous function across all four cell lines may be due to combinatorial interactions with other TFs. For example, in Table 3 we note that promoters with a functionally verified CTCF binding site were significantly enriched in AP-2, E2F4, GABP, NF-Y and Pax5 motifs.

TF	Func.	Tested	Ubq.Func.	Ubq. Act.	Ubq. Rep.	Func. K562	Func. HCT116	Func HT1080	Func HepG2	PWM	AUC
CTCF	104	168	9	9	0	62	52	49	53	Ref. [31]	0.84
E2F4	7	12	0	0	0	3	3	3	0	E2F:4 M00739	0.83
E2F6	2	3	0	0	0	1	0	1	0	E2F:1 M00938	0.78
EGR1	1	2	0	0	0	0	1	0	0	Egr:3 M00245	0.76
GABP	7	11	4	4	0	5	5	6	5	Ref. [103]	0.77
GATA1	4	4	1	1	0	4	4	1	1	GATA:1 M00128	0.69
GATA2	47	80	4	3	1	36	20	18	14	GATA:2 M00348	0.81
JUND	3	3	1	1	0	2	2	1	3	CREBP1 M00041	0.65
MAX	3	3	1	0	1	2	2	2	2	cMycMax M00118	0.77
STAT1	54	69	16	11	5	41	27	29	39	STAT1 M00224	0.74
USF1	2	2	1	1	0	2	2	2	1	USF M00121	0.86
YY1	86	98	26	9	16	63	56	53	58	Ref. [103]	0.82
Total	320	455	63	39	23	221	174	165	176		

known cell-selective transcriptional regulators including: (1) the pluripotency factors OCT4 (also called POU5F1), SOX2, KLF4 and NANOG in human embryonic stem cells37; (2) the myogenic factors MEF2A and MYF6 in skeletal myocytes[38]; and (3) the erythrogenic regulators GATA1, STAT1 and STAT5A in erythroid cells[39-41] (Fig. 7b).

Many of the novel, footprint-derived motifs displayed markedly cell-selective occupancy patterns highly similar with the aforementioned well-established regulators. This suggests that many novel motifs correspond to recognition sequences for important but uncharacterized regulators of fundamental biological processes. Notably, both known and novel motifs with high cell-selective occupancy predominantly localized to distal regulatory regions (Fig. 7c), further highlighting the role of distal regulation in developmental and cell-selective processes [42,43].

Using the whole set of binding peaks of all TRFs in each cell line as background, we found that motifless binding peaks have very significant overlaps with our HOT regions (Table 5). This is true no matter we consider all TRF peaks in the whole genome, or only those in intergenic regions. In all cases, the z-score is more than 25, which corresponds to a p-value of less than 3'10-138. A substantial portion of binding at HOT regions is thus attributed to non-sequence-specific binding. In our separate study, we found that motifless binding peaks have stronger DNase I hypersensitivity signals [20], which is also a signature of our HOT regions (Figure 4).

Our analysis also highlights the need of a more comprehensive catalog of sequence motifs of DNA binding proteins. If we instead define a TRF binding peak as motifless as long as it lacks either a previously-known motif or a newly discovered one, i.e., it could still have a motif from the other source, the overlap of the resulting "motifless" peaks with our HOT regions becomes statistically insignificant. Requiring a motifless binding peak to lack both types of motifs is likely more reliable.

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

Table 3 | Summary of genes regulated by ubiquitously functional TFBSs for five TFs: CTCF, GABP, GATA2, STAT1, and YY1.In Table 3, we list secondary TF motifs whose over-representation (or under-representation) on promoters containing binding sites for CTCF and STAT1, respectively, can be related to a functional outcome. The motifs listed in the "TF2" column of Table 3 are statistically over-represented (or under-represented) on promoters with a functional binding site for transcription factors listed in the "TF" column (i.e. CTCF and STAT1), relative to promoters with a predicted (CTCF or STAT1) binding site whose function was not verified.

TF	Ubiquitously activated	Ubiquitously repressed
CTCF	AL645504.2	
	ANKRD46	
	BICD2	
	C17orf81	
	CEP135	
	CRYAA	
	EGLN2	
	POMT2	
	TSFM	
GABP	GART°	
	PSMB4[a]	
	SYNJ1[a]	
	ZNF259[a]	
GATA2	CTSH	CCM2
	PLSCR2	
	TNFAIP8L1	
STAT1	ATG4C	HCFC1
	DCLRE1C	RPS24
	DIMT1L	TMED5
	ELP3	XXbac-BPG116M5.1
	GSTK1	ZNF367
	IRF7[b]	
	IRF9 [b]	
	KIF2A	
	MTMR9	
	NMI	
	SBNO2	
YY1	COQ5[cd]	AC091153.1
	CPNE1	ATP5O
	CPSF2 [cd]	BIRC6[d]
	CR613718	CAPZA2
	IP6K2[a]	CXorf26
	NARS[ac]	DKFZp434H247
	PAK4[d]	EFHA1
	PSMB4[ac]	MRPS10[c]
	UBR5	MRPS18B[acd]
		NUP160
		OXCT1
		PSMD8[ac]
		SNX27
		SNX3[ad]
		SRP68[ad]
		TNKS

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 cis-elements 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).

For HNF4A in HepG2 and GATA1 in K562 cells, ChIP data is available from the ENCODE project. To validate the predictions made by our model, we looked for overlap of these ChIP sites with DHS sites associated with different sets of genes. In HepG2 cells, 19% of all genes with an associated DHS overlapped a HNF4A binding site. Strikingly, 64.5% of the UR genes had a DHS overlapping an HNF4A ChIP peak (pvalue$<$ 1e-12, binomial test). Conversely, only 10.5% of DR genes had a DHS that overlapped an HNF4A site (p-value$<$1e-3). In K562 cells, we found that 6% of all genes had an associated DHS with a GATA1 ChIP peak. However, 31.5% (p-value$<$1e-12) of UR genes and only 3.5% (p-value$<$0.1) of DR genes had a DHS with a GATA ChIP peak. The ChIP binding data provided strong and independent evidence that our models identify relevant factors that regulate the transcriptional program in these cells.

To assess the presence of additional sequence motifs not accounted for by the sets of known PWMs, we used the discriminative version of MEME to perform motif finding (Bailey *et al.* 2010), identifying motifs differentially enriched between UR and UR-other respectively DR genes (Supplemental Table 8). While some of the identified motifs corroborated the importance of features from the set of top 10 TFs (FOXA2 [formerly HNF3B] in HepG2), others corresponded to TFs that were not in this list. These are candidate TFs that are not among the most differentially expressed, but still might be involved in the transcriptional program, potentially through other steps of activation. We note that we largely did not recover the motifs recently identified in a subset of 7 of the 19 cell lines (Song *et al.* 2011). In contrast to this study, which used the sequences from cell-type specific DHS as foreground and the subsets of cell-type specific DHS in other cell types as background, we analyzed the sequences from all DHS associated to a gene, and defined the background according to the classification tasks.

For several factors, we observed indicative footprints in the region of the motif (Figure 6). For example, CRX was predictive of UR genes in the medulloblastoma cell line, and it exhibited a protected region at the motif (Figure 6A). Importantly, in other cell lines such as GM12878, LnCAP and MCF7, the CRX motif did not display a similar level of protection. While CRX has been shown to be expressed in certain types of medulloblastoma sub-types (Kool *et al.* 2008), other factors such as OTX2 have nearly identical PWMs and are known to be important for transcriptional regulation in medulloblastomas (Bunt *et al.* 2011). This highlights a caveat in predicting expression from motifs; while we can identify biologically relevant motifs, this type of analysis only suggests a subset of factors that likely bind to a specific motif.

Motif analysis of genomic regions bound by TCF7L2

To investigate the predominant motifs enriched in TCF7L2 binding sites, we applied a *de novo* motif discovery program, ChIPMotifs [28, 29], to the sets of TCF7L2 peaks in each cell type. We retrieved 300 bp for each loci from the top 1,000 binding sites in each set of TCF7L2 peaks and identified the top represented 6-mer and 8-mer (Additional file 13). For all cell lines, the same 6-mer (CTTTGA) and 8-mer (CTTTGATC) motif was identified (except for HCT116 cells, for which the 8-mer was CCTTTGAT). These sites are almost identical to the Transfac binding motifs for TCF7L2 (TCF4-Q5:SCTTTGAW) and for the highly related family member LEF1 (LEF1-Q2:CTTTGA) and to experimentally discovered motifs in previous TCF7L2 ChIP-chip and ChIP-seq data [11, 30]. These motifs are present in a large percentage of the TCF7L2 binding sites. For example, more than 80% of the top 1,000 peaks in each dataset from each cell type contain the core TCF7L2 6-mer W1 motif, with the percentage gradually dropping to approximately ~50% of all peaks (Additional file 14).

Because the TCF7L2 motif is present in all the cell lines at the same genomic locations, but TCF7L2 binds to different subsets of the TCF7L2 motifs in the different cell lines, this suggests that a cell type-specific factor may help to recruit and/or stabilize TCF7L2 binding to specific sites in different cells. Also, as shown above, TCF7L2 binds to enhancer regions, which are typified by having binding sites for multiple factors. To test the hypothesis that TCF7L2 associates with different transcription factor partners in different cell types, we identified motifs for other known transcription factors using the program HOMER [31]. For these analyses, we used the subset of TCF7L2 binding sites that were specific to each of the 6 different cell types. The top 4 significantly enriched non-TCF7L2 motifs for each dataset are shown in Table 3; many of these motifs correspond to binding sites for factors that are expressed in a cell type-enriched pattern. To assess the specificity of the identified motifs with respect to TCF7L2 binding, we chose one motif specific to HepG2 TCF7L2 binding sites (HNF4α) and one motif specific to MCF7 TCF7L2 binding sites (GATA3) and plotted motif densities in the HepG2 cell type-specific TCF7L2 peaks (Figure 4A) and the MCF7 cell type-specific TCF7L2 peaks (Figure 4B). In HepG2 cells, the HNF4α motif, but not the GATA3 motif, is highly enriched at the center of TCF7L2 binding region. In contrast, in MCF7 cells the GATA3 motif, but not the HNF4α motif, is highly enriched at the center of TCF7L2 binding regions.

We find that transcription factors can have cell type specific primary motifs. This finding is in addition to the finding that TFs bind loci with cell-type specific cofactor motifs. We specifically show that YY1 and JunD can bind cell-type specific primary motifs that may correspond to cell-type specific oligomerization partners. This is particularly relevant for JunD and YY1 binding sites that are accessible in multiple cell types, but only bound in one, where the specificity seems solely provided by the differential primary motif. This is in contrast to factors that are differentially recruited by cofactors, where differential DNase-accessiblity has equal or greater capacity for predicting cell-type specific binding.