10 Characterization of network topology

ENCODE data analysis helps to describe the various types of regulatory "wiring" implicit in the genome

The filtered transcription factor (TF) hierarchy consists of the strongest promoter-associated interactions. Building upon this skeleton, we added additional types of connections.

Interactions involving distal regulatory elements (e.g., enhancers) are more difficult to identify than those 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 19528 distal edges (Fig. 2A).

The regulatory interactions between TFs and ncRNAs constitute an additional layer of information to add to the meta-network. We used TF peaks proximal to ncRNAs to identify TF-to-ncRNA regulation. Next, we incorporated miRNA-to-TF regulatory interactions from TargetScan[36] (SOM/E.2). Finally, we incorporated physical protein-protein interactions[26], as well as predicted phosphorylations (SOM/F.3 and Fig. S7a). Overall, these different interactions form a dense and complex meta-network that was further analyzed for interesting biological properties.

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

Among the TRF binding sites, one subset of particular interest is the ones close to the TSSs of active genes, as they are likely actively involved in the regulation of these genes in the corresponding cell lines. Depending on the distance from a transcription start site, these regions may contain core promoters and proximal promoter elements[2]. We call these regions promoter-proximal regulatory modules (PRMs) in general. To define PRMs, instead of using an arbitrary distance threshold from TSSs, we determined distance cutoffs according to chromatin feature patterns using a machine learning framework. Specifically, for each cell line, we took TSSs of genes expressed in the cell line as positive examples, and random non-TRF binding sites and distal TRF binding sites as negative examples (Materials and methods). Expression of TSSs was determined by ENCODE data from Cap-Analysis of Gene Expression (CAGE)[27], Paired-End diTag (PET)[28], and RNA sequencing (RNA-seq)[29-30]. Based on the examples, a discriminative model was learned using chromatin features and TRF binding data of the cell line as explanatory variables. The resulting models separated positive and negative examples well in all cell lines (Additional file 2, Figure S3 and Additional file 2, Figure S4). Finally we used the learned models to give PRM scores to all regions in the whole genome. Since in this case we have a relatively complete set of positive examples from annotated genes, we used a more stringent threshold to call PRMs (Materials and methods).

In contrast to PRMs, there are also regulatory modules that are more distal to promoters. For example, enhancers are frequently thousands of bases pairs upstream or downstream of a promoter, and they can be within an intron of a gene[2]. To study properties unique to this type of DNA elements, we focused on BARs at least 10Kbp from any annotated coding and non-coding transcript (Materials and methods) and removed from this list any identified PRMs, to eliminate properties superimposed from annotated and potentially unannotated genes.

We find that ~50% of TSSs display one or more long-range interaction with some interacting with as many as 20 distal fragments (Figure 4a). Expressed TSSs interact with slightly more elements as compared to non-expressed TSSs (for GM12878 mean is 1.84 vs 1.35; or 3.79 vs 3.20 when including only those TSS with at least one

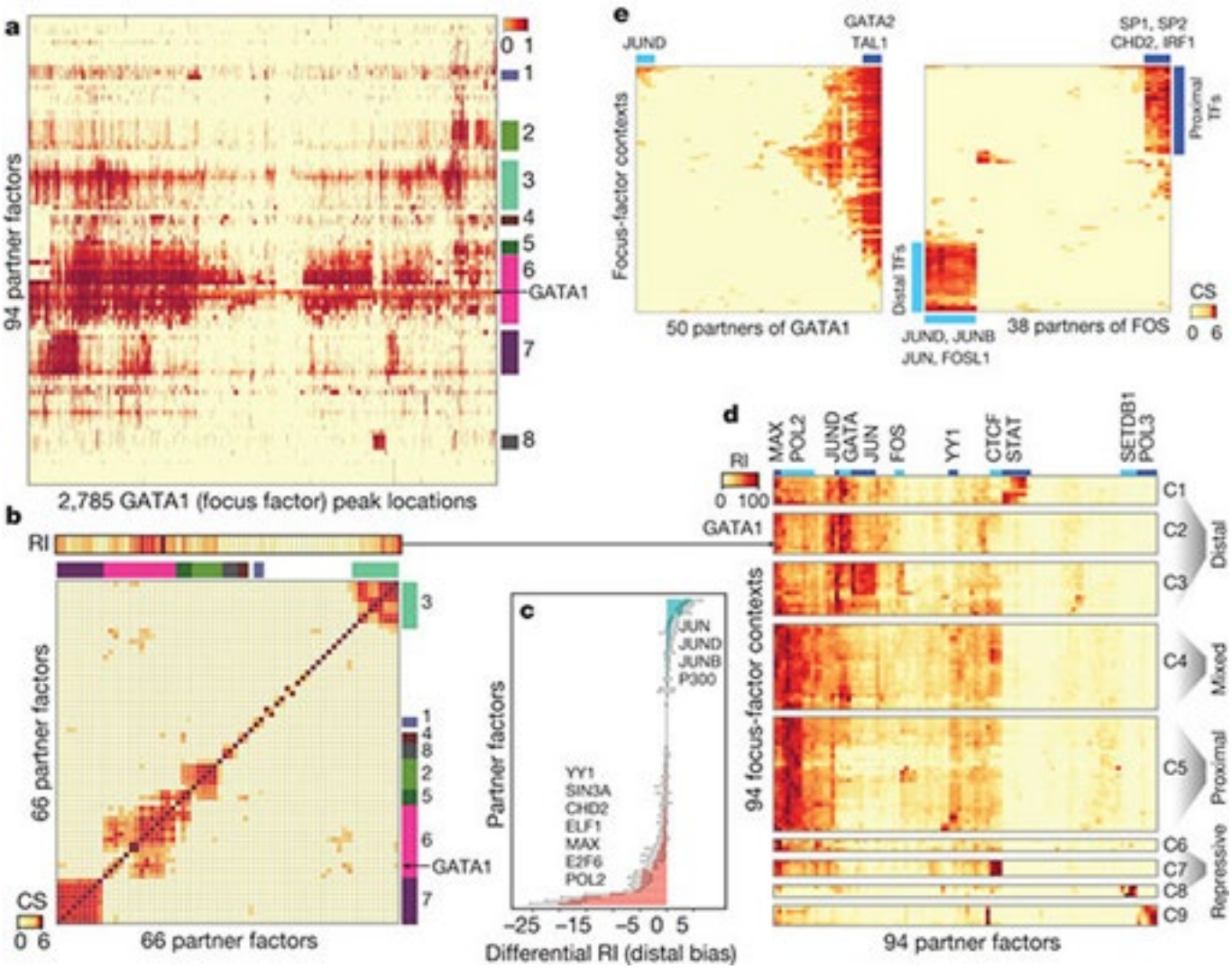


Figure 1 | Transcription factor co-association. (a) The co-binding map for the GATA1 focus-factor context in K562 shows the binding intensity of peaks of all TFs in K562 (rows) that overlap each GATA1 peak (columns). The colored rectangles represent 8 key clusters consisting of different combinations 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, S2I-2.)

interaction). Out of all distal fragments interrogated, ~10% interact with one or more TSS, with some interacting with more than 10 (mean 2.14 when including only those distal fragments with at least one interaction). […]

The degree distribution of the four categories of distal elements was very similar (Supplementary Figure 9).

Transcriptional regulatory circuitry of PPARGC1A and its networks partners

We next sought to explore the regulatory circuitry formed by *PPARGC1A* and its network partners by investigating their binding relationships with each other and with other transcription factors. We first examined the genes encoding each transcriptional regulator and determined which regulators were bound. As shown

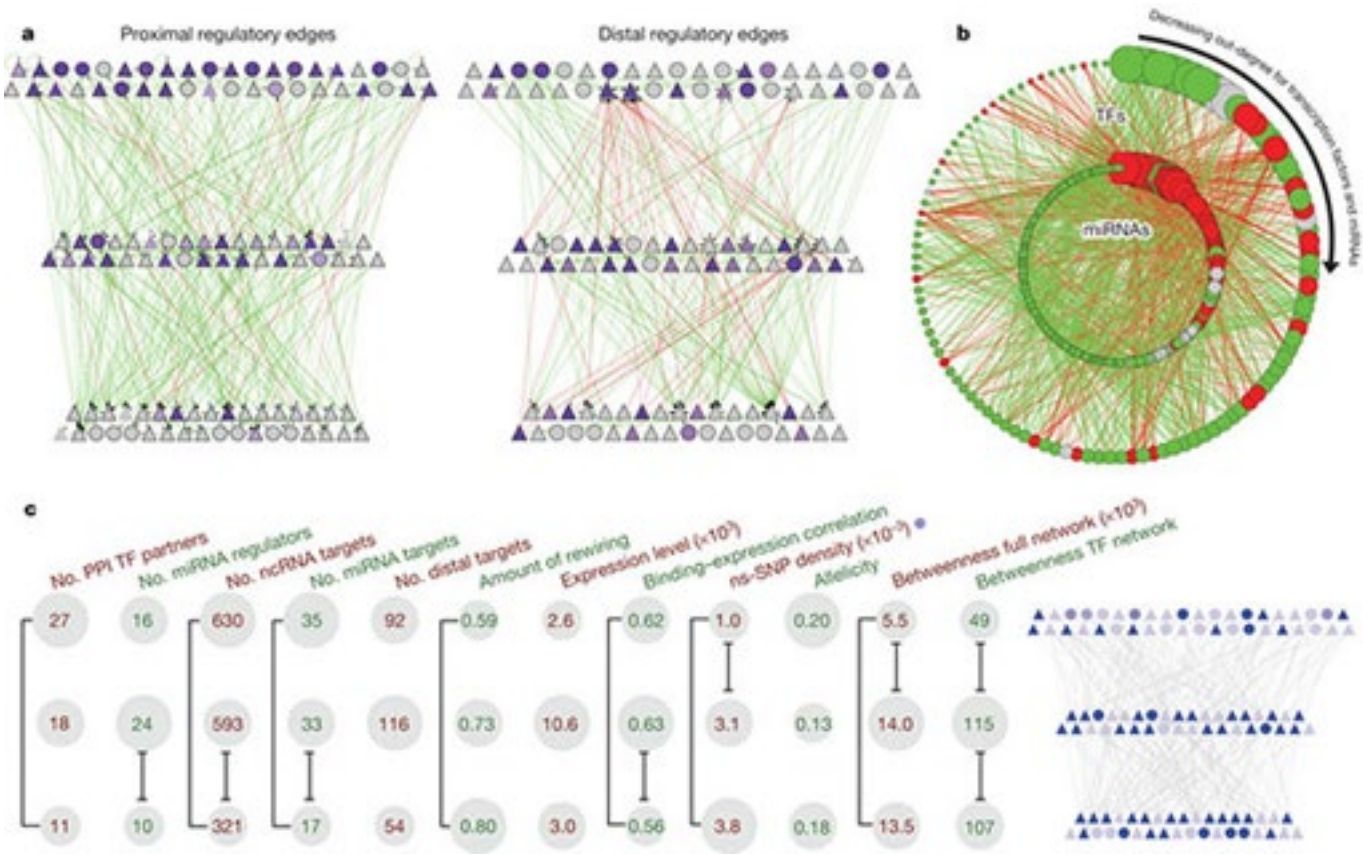


Figure 2 | Overall Network (a) Close-up of the TF hierarchy. The nodes depict the TFs: TFSSs are triangles, and non-TFSSs are circles. At the left we show the proximal-edge hierarchy with downward pointing edges colored in green, and upward pointing ones colored in red. The nodes are shaded according to their out-degree in the full network (as described in Table 1). The right part shows the TFs placed in the same proximal hierarchy but now with edges corresponding to distal regulation colored green and red, and nodes recolored according to out-degree in the distal network. We see that the distal edges do not follow the proximal-edge hierarchy. (b) Close-up of TF-miRNA regulation. The outer circle contains the 119 TFs, while the inner circle contains miRNAs. Red edges correspond to miRNAs regulating TFs; green ones, TFs regulating miRNAs. TFs and miRNAs each are arranged by their out-degree, beginning at 12 o'clock and decreasing in order clock-wise. Node sizes are proportional to out-degree. For TFs, the out-degree is as described in Table 1; for miRNAs, it is according to the out-degree in this network. Red nodes are enriched for miRNA-TF edges and green nodes are enriched for TF-miRNA edges. Gray nodes have a balanced number of edges (within ± 1). (b) Average values of various properties (topological, dynamic, expression-related, and selection-related - ordered consistently with Table 1) for each level are shown for the proximal-edge hierarchy. The top, middle, and bottom rows correspond to the top, middle, and bottom of the hierarchy, respectively. The sizing of the grey circles indicate the relative ordering of the values for the three levels. Significantly different values ($P<0.05$) using the Wilcoxon-rank-sum test are indicated by black brackets. The proximal-edge hierarchy depicted on the right shows non-synonymous SNP density, where the shading corresponds to the density for the associated TF. (More details in Fig S4.)

for the *PPARGC1A* gene in Figure 6A, we frequently observed binding of different combinations of regulators not only at promoter regions, but also at multiple sites within the gene body and adjacent to the 3′-end. Thus, we considered any such binding event to be a potential regulatory connection and used this information to construct a transcriptional regulatory network. In this network, 'nodes' represent each transcriptional regulator and 'edges' connecting nodes represent binding of one regulator to the gene encoding another regulator. As diagrammed in Figure 6B, the seven regulators form a highly connected network suggesting complex patterns of interdependent regulation. Various network motifs (the smallest units of network structure) are abundant in this network, including feed-forward loops, multi-component loops, and multi-input motifs (Blais and Dynlacht 2005). One particularly striking feature of this network is that every transcriptional regulator exhibits autoregulation, a motif that characterizes master regulators of important cellular pathways in various cell types (Boyer *et al.* 2005; Odom *et al.* 2006; Reed *et al.* 2008). Notably, an autoregulatory interaction occurs in intron 2 of the *PPARGC1A* gene (Fig. 6A), a region in which single nucleotide polymorphisms (SNPs) have been associated with the age at onset of Huntington's disease (Taherzadeh-Fard *et al.* 2009; Weydt *et al.* 2009).

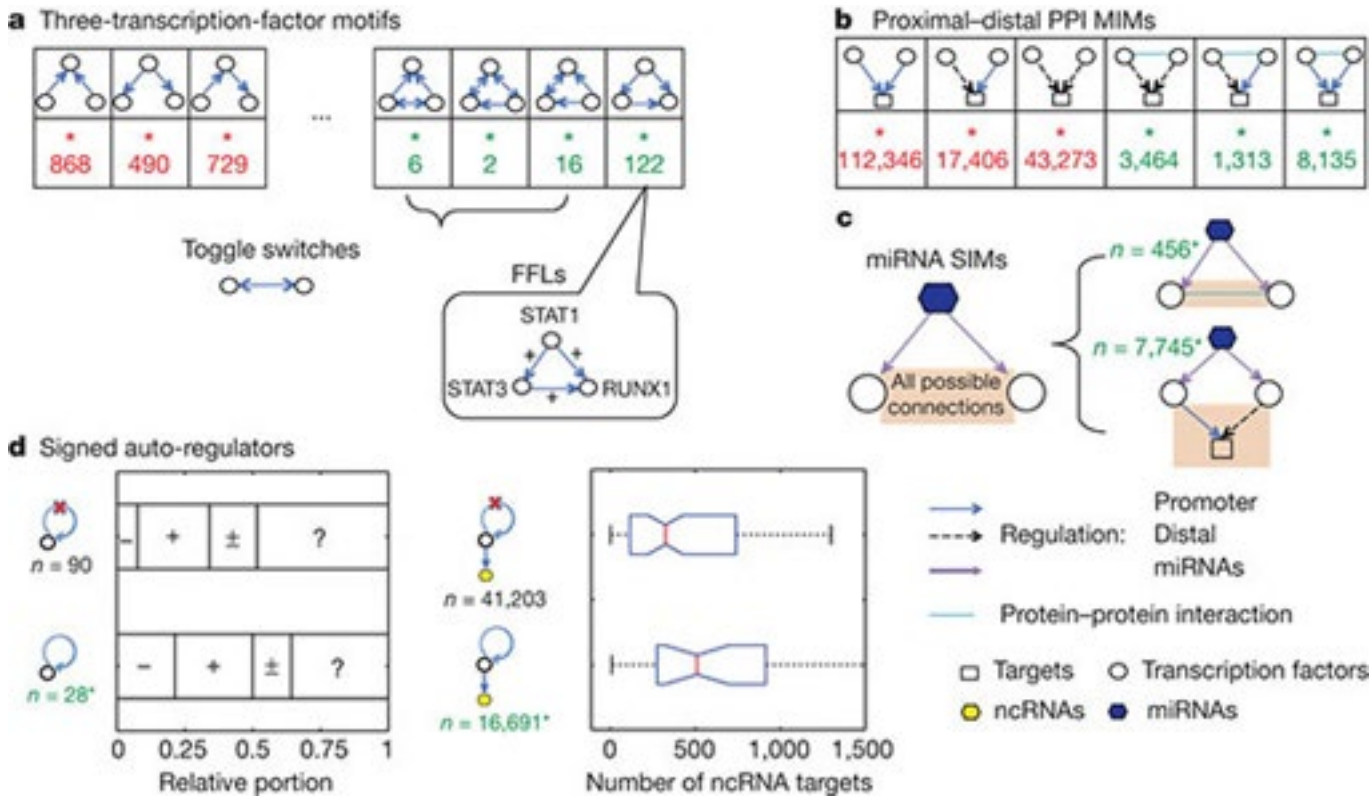


Figure 4 | Motif Analysis Motifs are accompanied by the occurrence frequency, N. Enriched motifs are highlighted in green, and depleted ones, in red. An occurrence frequency with a star means that the corresponding enrichment/depletion is statistically significant (P=1e-5). The motifs are sorted such that those at the ends have more significant p-values. (More details in Fig. S9h.) (a) Systematic search of 3-TF motifs. The most enriched motif is the FFL. A particular example formed by STAT1, STAT3 and RUNX1 is highlighted. Here, the "+" sign on an edge indicates that the correlation between the gene expression of the source and the target across tissues is positive. Other motifs containing a toggle-switch regulation on top of the basic FFL design are also indicated. (b) Proximal-Distal-PPI MIMs. Here we searched all motifs involving the co-regulation of two TFs (which could be either proximal or distal) with (or without) a protein-protein interaction between them. We found the motifs containing the protein-protein interaction tended to be enriched. (c) miRNA-SIMs. This figure shows the 2 enriched motifs resulting from enumerating all motifs in which a miRNA targets two TFs that are connected in various ways. These 2 motifs contain a protein complex of 2 TFs and a cooperative pair of promoter and distal regulatory TFs. (d) The auto-regulator motif is enriched in the TF-TF network: 28 of all TFs are auto-regulators. Moreover, auto-regulators are more likely to be repressors (-) relative to non-auto regulators, and they tend to have more ncRNAs as their targets.

The factors in this network also exhibit a hierarchical relationship in terms of the number of transcriptional regulators that they occupy, represented by the number of outgoing edges of each node (Fig. 6C). CEBPB and HNF4A, for example, are highly connected nodes, forming direct regulatory connections with all seven regulators that we examined. These findings indicate that CEBPB and HNF4A can influence the expression of a large assortment of genes both by direct interactions and by interactions mediated by multiple downstream regulators, and are consistent with the observed roles of HNF4A and CEBP family members in specifying and maintaining the hepatocyte transcriptional program (Lekstrom-Himes and Xanthopoulos 1998; Odom *et al.* 2004; Kyrmizi *et al.* 2006). ESRRA, GABP, HSF1, and PPARGC1A, on the other hand, occupy lower positions in the hierarchy, as each of these factors binds to a more limited subset of 3-4 regulators, suggesting that they influence the expression of genes that are involved in more focused biological pathways. One basic property of transcriptional regulatory networks is that edges are directional; thus, each node can contain differing numbers of incoming and outgoing connections (Barabasi and Oltvai 2004; Borneman *et al.* 2006). In the network formed by PPARGC1A and its partner TFs, both ESRRA and HNF4A contain 6 incoming connections, compared to only 3-4 for the remaining regulators (Fig. 6C). These findings indicate that the *ESRRA* and HNF4A genes are subject to greater regulatory input and potentially responsive to a wider array of biological stimuli. Previous work in yeast suggests that TFs that are downstream targets of many regulators can serve as 'target hubs' in regulatory

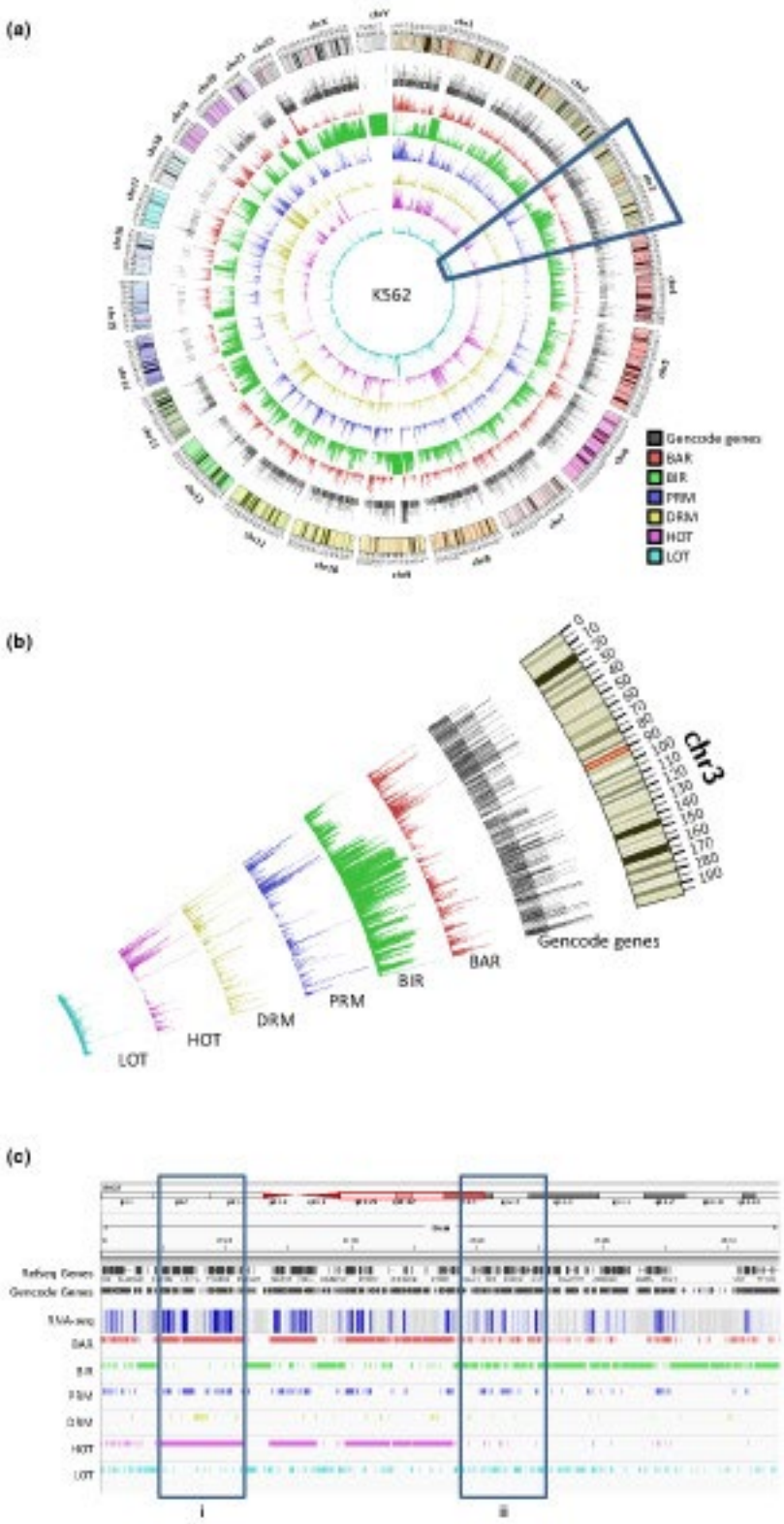


Figure 2 | Distribution of the six types of regions in the genome in K562. (a) Densities of the regions in the whole genome, defined as the running fractions of bases covered by the regions. The tracks are respectively, from outermost one to the innermost one, the ideogram for the human karyotype (genome build hg19), Gencode version 7 level 1 and level 2 genes, BARs, BIRs, PRMs, DRMs, HOT regions and LOT regions. The tracks are scaled separately to show density fluctuations. The highlighted segment corresponds to the area in panel B. (b) Zoom-in of chromosome 3 to show the correlated fluctuations of the different types of regions. (c) Locations of the six types of regions at the beginning of the q-arm of chromosome 22 in K562. Due to the high density of genes, only a subset of the gene names is shown. Expression values were measured by long poly-A+ RNA-seq of whole-cell RNA extract. A darker color indicates a higher average expression level in the local region. Box i marks a broad area with significant active TF binding and co-binding. Box ii marks an area with many small interspersed active and inactive TF binding regions.

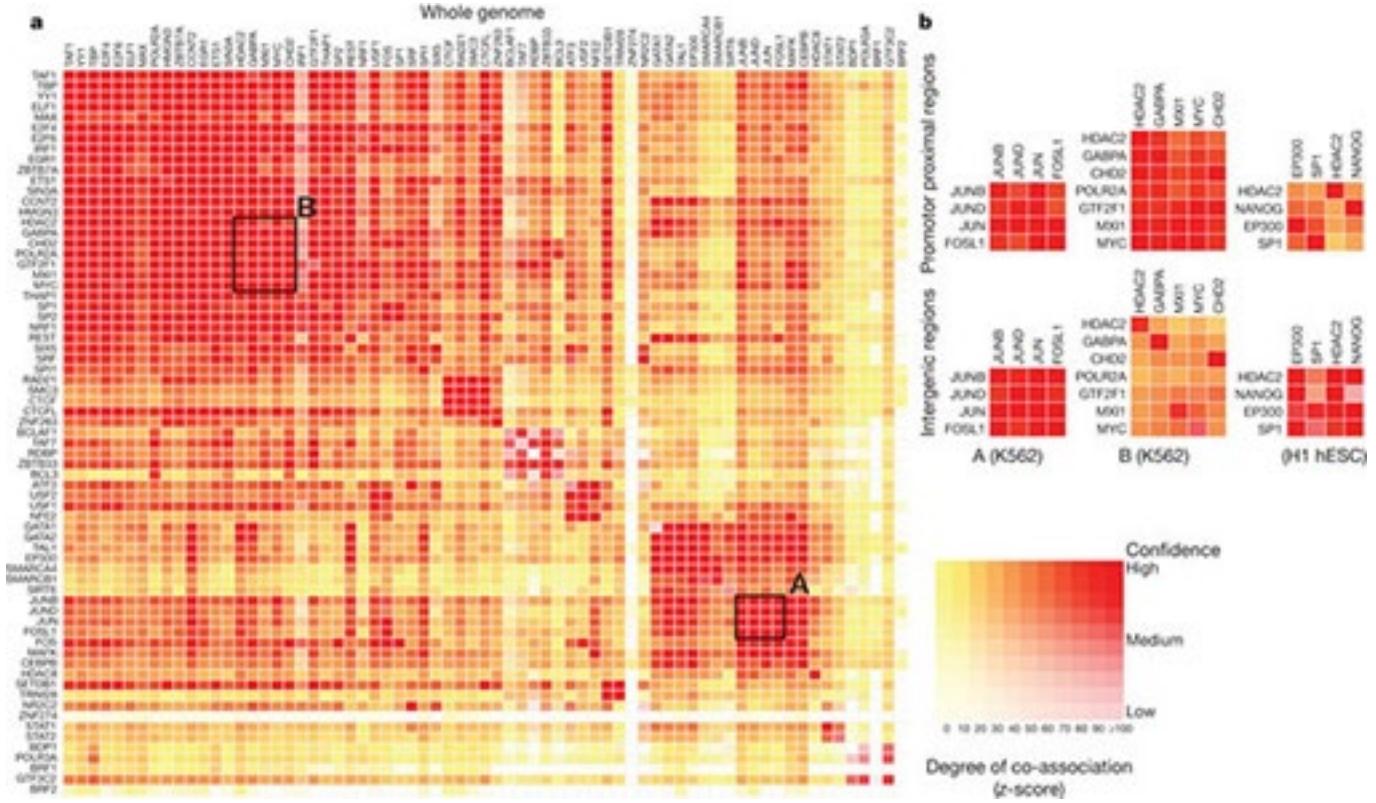


Figure 4 | **Co-association between Transcription Factors.** (**a**), Significant co-associations of transcription factor pairs using the GSC statistic across the entire genome in K562 cells. The colour strength represents the extent of association (from red (strongest), orange, to yellow (weakest)), whereas the depth of colour represents the fit to the GSC[20] model (where white indicates that the statistical model is not appropriate) as indicated by the key. Most transcription factors have a nonrandom association to other transcription factors, and these associations are dependent on the genomic context, meaning that once the genome is separated into promoter proximal and distal regions, the overall levels of co-association decrease, but more specific relationships are uncovered. (**b**), Three classes of behaviour are shown. The first column shows a set of associations for which strength is independent of location in promoter and distal regions, whereas the second column shows a set of transcription factors that have stronger associations in promoter-proximal regions. Both of these examples are from data in K562 cells and are highlighted on the genome-wide co-association matrix (**a**) by the labelled boxes A and B, respectively. The third column shows a set of transcription factors that show stronger association in distal regions (in the H1 hESC line). An interactive version of this figure is available in the online version of the paper.

networks whose activity represents the combined output of multiple upstream signals (Borneman *et al.* 2006; Zhu *et al.* 2007). These factors often serve as master regulators of important cellular pathways (Borneman *et al.* 2006; Zhu *et al.* 2007).

To identify additional regulators that might form important connections with the seven factors in our 'core' network, we expanded our network by including target genes encoding other transcriptional regulators that were occupied by at least 4 factors (Fig. 6D; Supplemental Table S10). The expanded network contains 155 additional regulators (Supplemental Table S10), many of which have demonstrated roles in critical cellular processes related to the function of PPARGC1A. Indeed, several of these regulators are known targets of PPARGC1A coactivation, including RXRA, RXRB, YY1, PPARD, and NR1H3 (also known as LXRA) (Fig. 6D) (Delerive *et al.* 2002; Oberkofler *et al.* 2003; Wang *et al.* 2003; Cunningham *et al.* 2007). Intriguingly, YY1, one of the TFs with the most connections to the core network, has been shown to bind directly to the *PPARGC1A* promoter in skeletal muscle and YY1 DNA-binding motifs are highly enriched in the promoters of mitochondrial genes regulated by PPARGC1A (Cunningham *et al.* 2007). Another factor in the expanded network, NR0B2 (also known as SHP), negatively regulates the expression of *PPARGC1A* in brown adipocytes (Wang *et al.* 2005). The

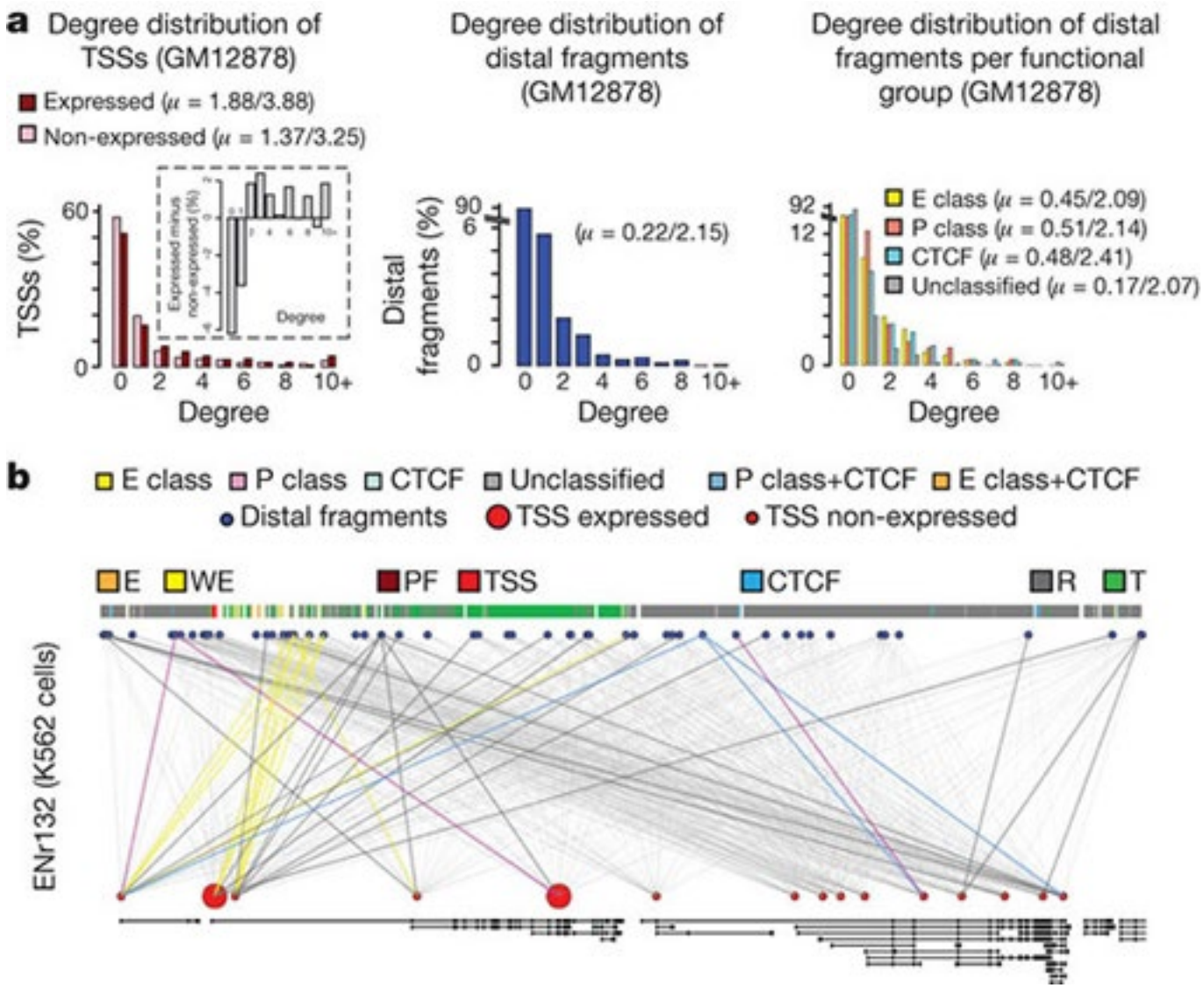
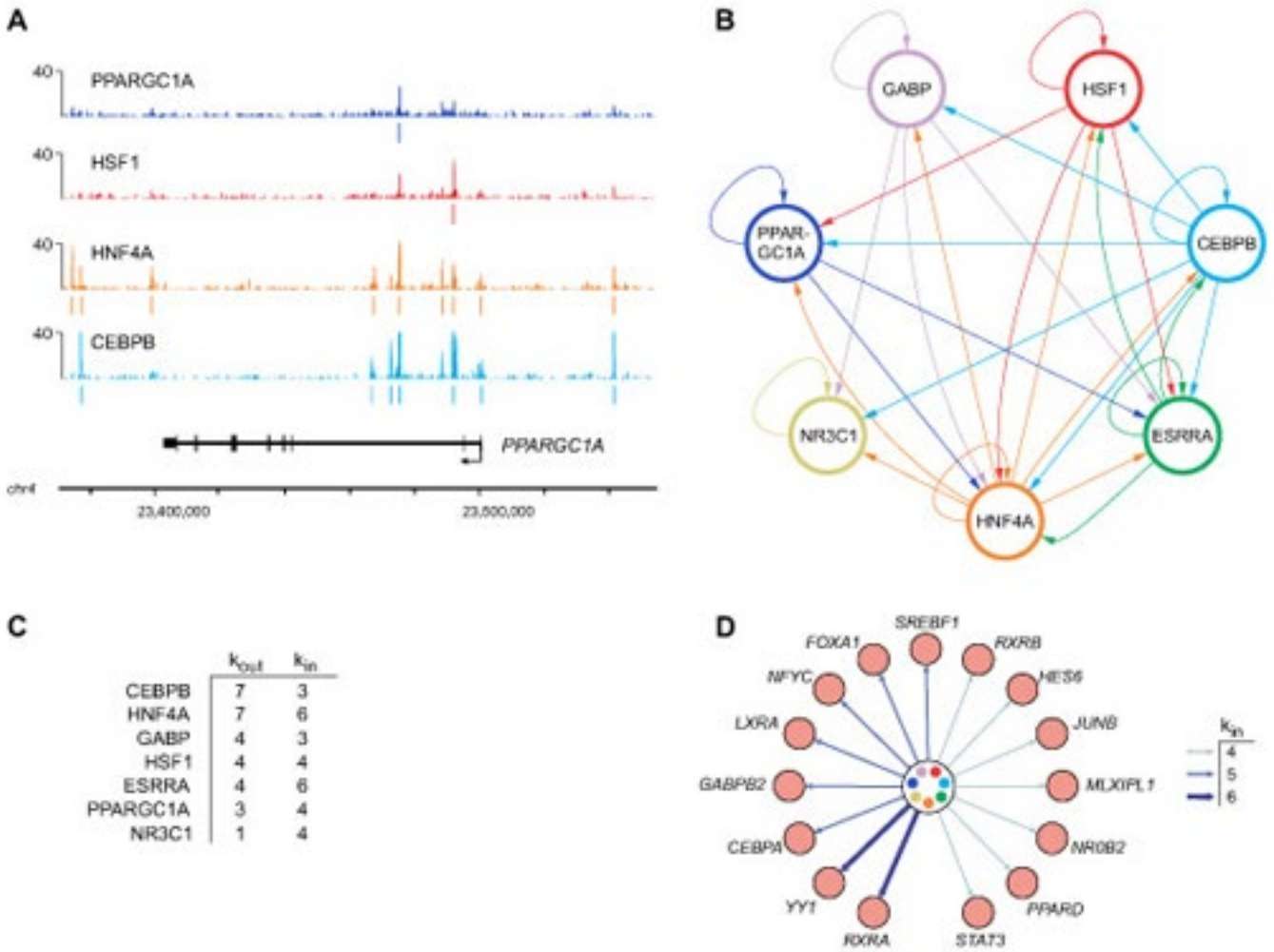


Figure 4 | Networks of looping interactions. (a) Histogram showing the number of TSSs (left, red) or distal fragments (middle, blue) in percentages that are involved in 0, 1, 2,….10 (and above) looping interactions (degree, *x* axis) in GM12878 cells. All of the values for degrees that are >9 are grouped under degree 10+. The dark red bars represent the percentages of looping TSSs that are expressed whereas light red bars represent the percentages of looping TSSs that are not expressed. Inset: the difference in percentage between looping TSSs that are expressed and not expressed for each degree is shown. The right panel shows the degree distribution for each functional group of distal fragments. The average degrees (mean, μ) for TSSs and distal fragments are indicated. The first value is the mean degree considering all the TSS/distal fragments (looping plus non-looping), whereas the second value is the mean degree of looping TSS/distal fragments (excluding degree = 0). (b) Web plot showing the long-range looping interactions in the ENr132 region in K562 cells. The interrogated distal fragments (blue circles) and the TSSs (red circles) are positioned according to genomic coordinates and the GENCODE v7 gene annotation is indicated. The size of the red circles indicates whether that TSS is expressed (large circles) or not expressed (small circles). The thin grey lines show all the interactions that were interrogated. The coloured lines show significant looping interactions between TSSs and distal fragments of a particular group.

presence of regulators such as YY1 and SHP and other TFs in our expanded network not only implicates these factors as players in the PPARGC1A regulatory network in the liver but also suggests a broad transcriptional network is influenced by PPARGC1A and its associated factors.

From examination of DNaseI profiles across many cell types we observed that many known cell-selective enhancers become DHSs synchronously with the appearance of hypersensitivity at the promoter of their target gene (Supplementary Fig. 13). To generalize this, we analysed the patterning of 1,454,901 distal DHSs (DHSs



	k_{out}	k_{in}
CEBPB	7	3
HNF4A	7	6
GABP	4	3
HSF1	4	4
ESRRA	4	6
PPARGC1A	3	4
NR3C1	1	4

Figure 6 | **Transcriptional regulatory circuitry of PPARGC1A and its network partners.** (**a**) Regulator binding at the *PPARGC1A* locus. Signal maps are displayed for four factors that occupy *PPARGC1A*. Significant peaks ($q < 0.01$) are indicated by colored lines under each signal map. Chromosomal positions are indicated on the x-axis. Gene structure is shown to scale below the signal maps. (**b**) Transcriptional regulatory network diagram displaying interactions among CEBPB, ESRRA, GABP, NR3C1, HNF4A, HSF1, and PPARGC1A. Arrows indicate direct binding of one regulator to the 5′-proximal, 3′-proximal, or intragenic region of the gene encoding another regulator (or the same regulator in the case of autoregulatory loops). (**c**) Regulatory hierarchy among the seven regulators depicted in (**b**). Factors are ranked first by the number of incoming network connections ("kin") then by the number of outgoing network connections ("kout") (Borneman *et al.* 2006). (**d**) Expanded transcriptional regulatory network. Additional TFs were added to the core network shown in (**b**), represented here by the circle in the center of the diagram, if they contained four or more incoming network connections. The number of incoming connections is indicated by arrow types according to the legend. Fifteen representative TFs are shown; the complete list of 155 TFs in the expanded network and their bound regulators is available in Supplemental Table S10.

separated from a TSS by at least one other DHS) across 79 diverse cell types (Supplementary Methods and Supplementary Table 6), and correlated the cross-cell-type DNaseI signal at each DHS position with that at all promoters within ±500 kb (Supplementary Fig. 14a). We identified a total of 578,905 DHSs that were highly correlated ($r > 0.7$) with at least one promoter ($P < 10\text{-}100$), providing an extensive map of candidate enhancers controlling specific genes (Supplementary Methods and Supplementary Table 7). To validate the distal DHS/enhancer-promoter connections, we profiled chromatin interactions using the chromosome conformation capture carbon copy (5C) technique[31]. For example, the phenylalanine hydroxylase (*PAH*) gene is expressed in

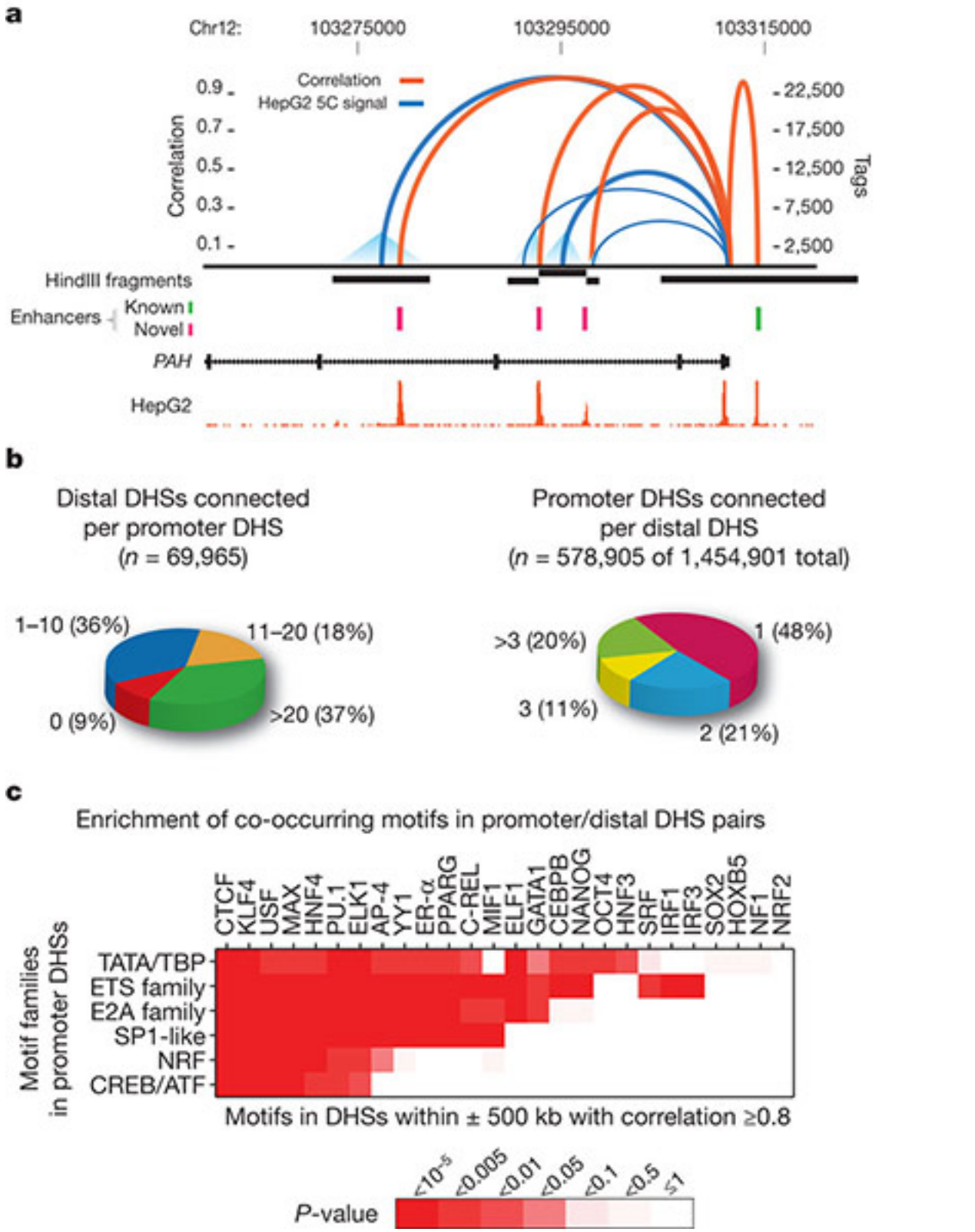
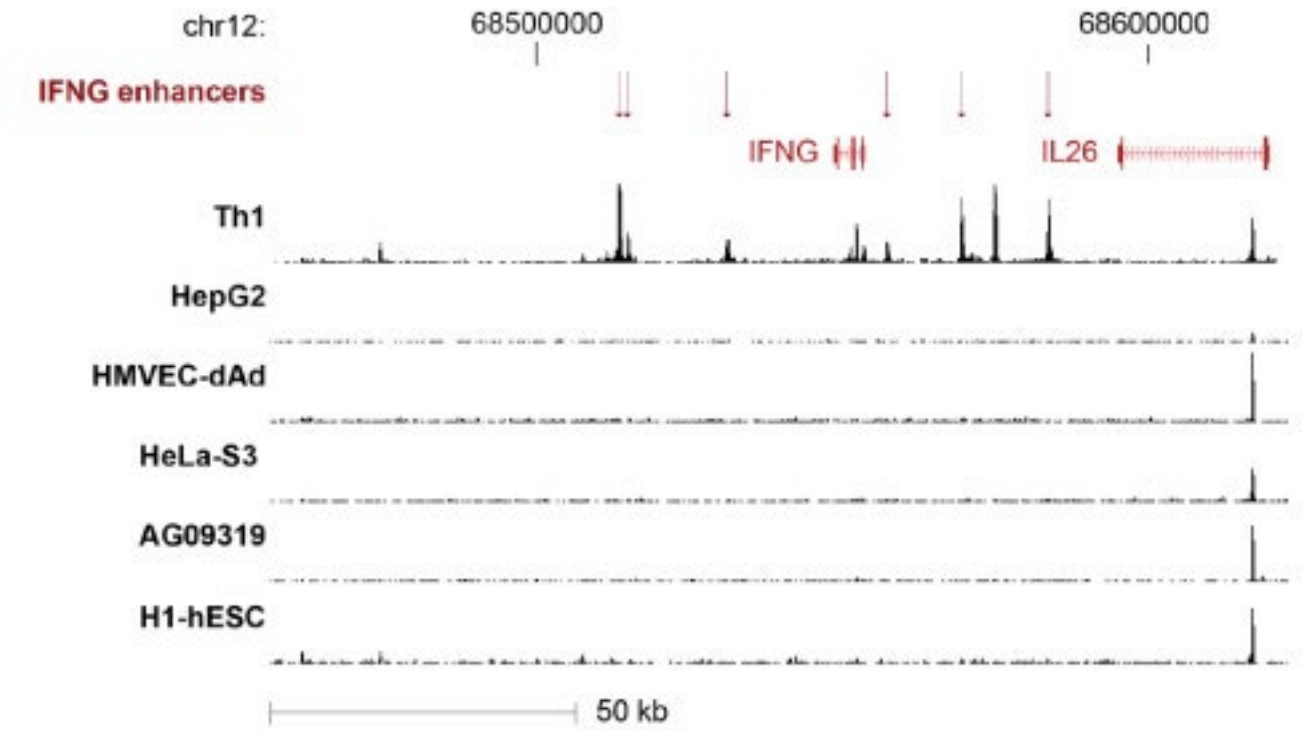


Figure 5 | A genome-wide map of distal DHS-to-promoter connectivity. (a) Cross-cell-type correlation (red arcs, left *y* axis) of distal DHSs and *PAH* promoter closely parallels chromatin interactions measured by 5C-seq (blue arcs, right *y* axis); black bars indicate HindIII fragments used in 5C assays. 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.



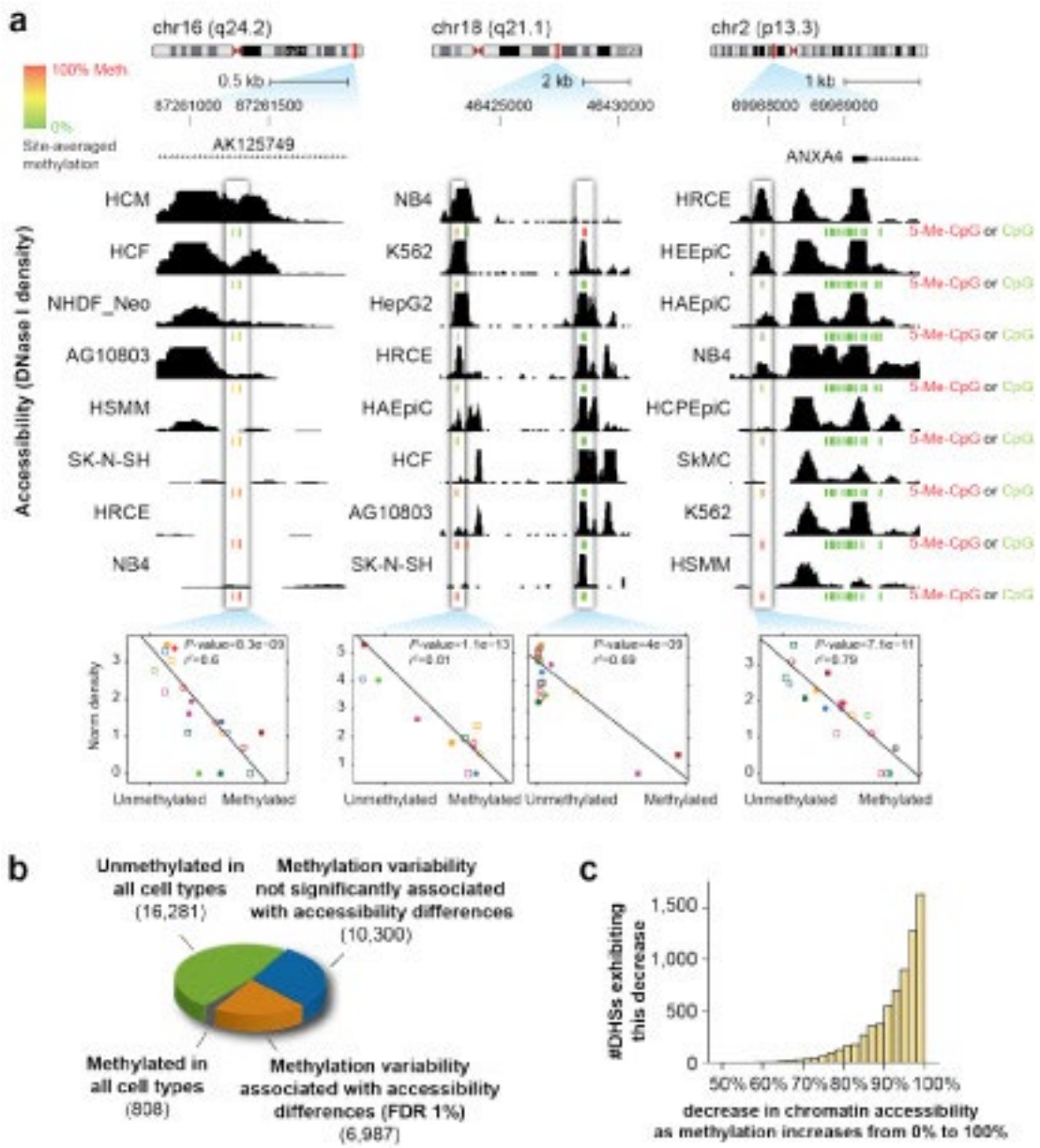
Supplementary Figure 13 | Cell-type-specific enhancers at the IFNG locus. Cell-specific enhancers (red arrows) in the IFNG locus. Enhancers of the IFNG gene[5] are marked by DHSs in the hTH1 (T lymphocyte) cell-type, consistent with the functioning of lymphocytes in producing the gene product interferon gamma. The enhancer loci are lacking in DHSs in other cell-types. Shown are DNaseI tag densities for six cell-types, including hTH1. See Supplementary Table 4 for IFNG enhancer coordinates and references.

hepatic cells, and an enhancer has been defined upstream of its TSS (Fig. 5a). The correlation values for three DHSs within the gene body closely parallel the frequency of long-range chromatin interactions measured by 5C. The three interacting intronic DHSs cloned downstream of a reporter gene driven by the *PAH* promoter all showed increased expression ranging from three- to tenfold over a promoter-only control, confirming enhancer function.

We next examined comprehensive promoter-versus-all 5C experiments performed over 1% of the human genome[32] in K562 cells. DHS-promoter pairings were markedly enriched in the specific cognate chromatin interaction ($P < 10\text{-}13$, Supplementary Fig. 14b). We also examined K562 promoter-DHS interactions detected by polymerase II chromatin interaction analysis with paired-end tag sequencing (ChIA-PET)[24], which quantifies interactions between promoter-bound polymerase and distal sites. The ChIA-PET interactions were also markedly enriched for DHS-promoter pairings ($P < 10\text{-}15$, Supplementary Fig. 14c). Together, the large-scale interaction analyses affirm the fidelity of DHS-promoter pairings based on correlated DNaseI sensitivity signals at distal and promoter DHSs.

Most promoters were assigned to more than one distal DHS, indicating the existence of combinatorial distal regulatory inputs for most genes (Fig. 5b and Supplementary Table 7). A similar result is forthcoming from large-scale 5C interaction data[32]. Surprisingly, roughly half of the promoter-paired distal DHSs were assigned to more than one promoter (Fig. 5b and Supplementary Methods), indicating that human cis-regulatory circuitry is significantly more complicated than previously anticipated, and may serve to reinforce the robustness of cellular transcriptional programs.

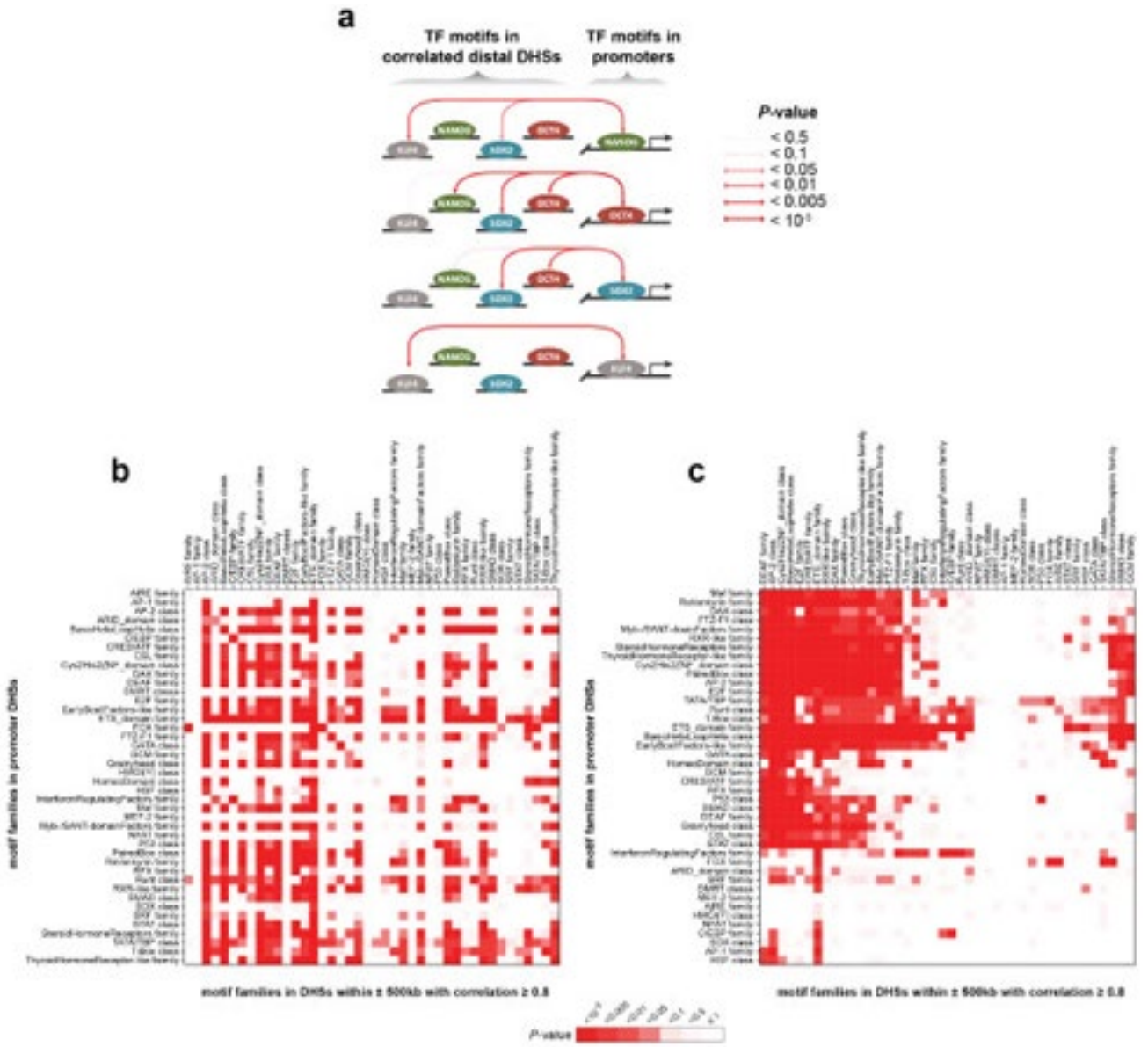
The number of distal DHSs connected with a particular promoter provides, for the first time, a quantitative measure of the overall regulatory complexity of that gene. We asked whether there are any systematic functional features of genes with highly complex regulation. We ranked all human genes by the number of distal DHSs paired with the promoter of each gene, then performed a Gene Ontology analysis on the rank-ordered list. We



Supplementary Figure 14 | Interaction and GO class enrichments via signal-vector correlation. (a) Further examples of association between methylation and accessibility. Data tracks show DNase I sensitivity in selected cell types. Green bars, CpG is 0% methylated; yellow, 50% methylated; red, 100% methylated. Association is quantified in the plots below the tracks. Each point in the graph represents one of 19 cell-types (a susbset of which is represented in the tracks). *X*-axis is the percent methylation of the site in that cell-type; *y*-axis is the normalised DNaseI tag density at the site in that cell type. In each example, accessibility (*y*-axis) quantitatively decreases as methylation increases (left to right). (b) Global characterisation of the effect of methylation on chromatin accessibility, surveyed at 34,376 DHSs with RRBS data. 40% of sites with variable methylation across cell-types were associated with differences in chromatin accessibility. (c) In cell lines with methylated DHSs, site accessibility was reduced on average by 95%. Shown are sites where increased methylation was significantly associated with decreased accessibility (= 97% of all sites in the orange slice shown in (b)).

found that the most complexly regulated human genes were markedly enriched in immune system functions (Supplementary Fig. 14d), indicating that the complexity of cellular and environmental signals processed by the immune system is directly encoded in the *cis*-regulatory architecture of its constituent genes.

Next, we asked whether DHS-promoter pairings reflected systematic relationships between specific combinations of regulatory factors (Supplementary Methods). 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 cells[33]. We found significant enrichment ($P < 0.05$) of the KLF4, SOX2 and OCT4 motifs within



Supplementary Figure 15 | Statistical significance of co-occurences of motif families. Statistical significances of co-occurrences of motifs and families and classes of motifs within connected ($r > 0.8$) distal/promoter DHS pairs genome-wide. (**a**), Co-occurrences among motifs for pluripotency factors KLF4, SOX2, OCT4, and NANOG. Enriched co-occurrences are denoted by arrows shaded by *P*-value. (**b-c**), Co-occurrences of families and classes of motifs. Family and class definitions are given in Supplementary Table 9. In (**b**), the motif families and classes are shown in alphabetical order. The matrix is clearly not symmetric; for example, within co-occurrences, TATA/TBP is enriched in several cases when it appears in a promoter DHS, but in only a few cases when it appears in a correlated distal DHS. Panel (**c**) shows the data from (**b**), hierarchically clustered by column and row. The DAX, FTZ-F1, RXR-like, Steroid Hormone Receptors, and Thyroid Hormone Receptor-like families, which all belong to the same class, cluster tightly together by rows (presence within promoter DHSs).

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). By contrast, promoters containing KLF4 motifs were associated with KLF4-containing distal DHSs, but not with DHSs containing NANOG, SOX2 or OCT4 motifs (Supplementary Fig. 15a, bottom).

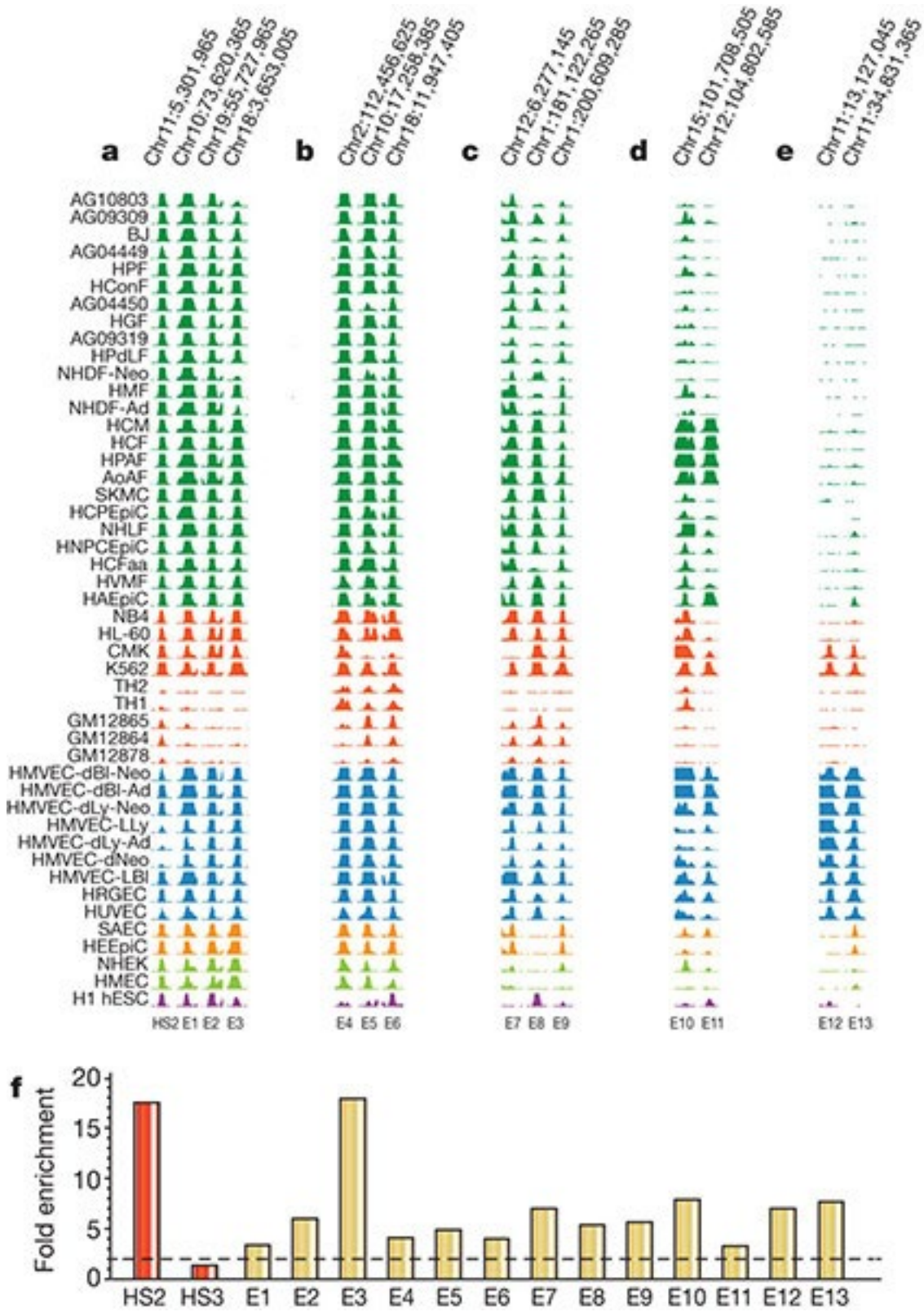
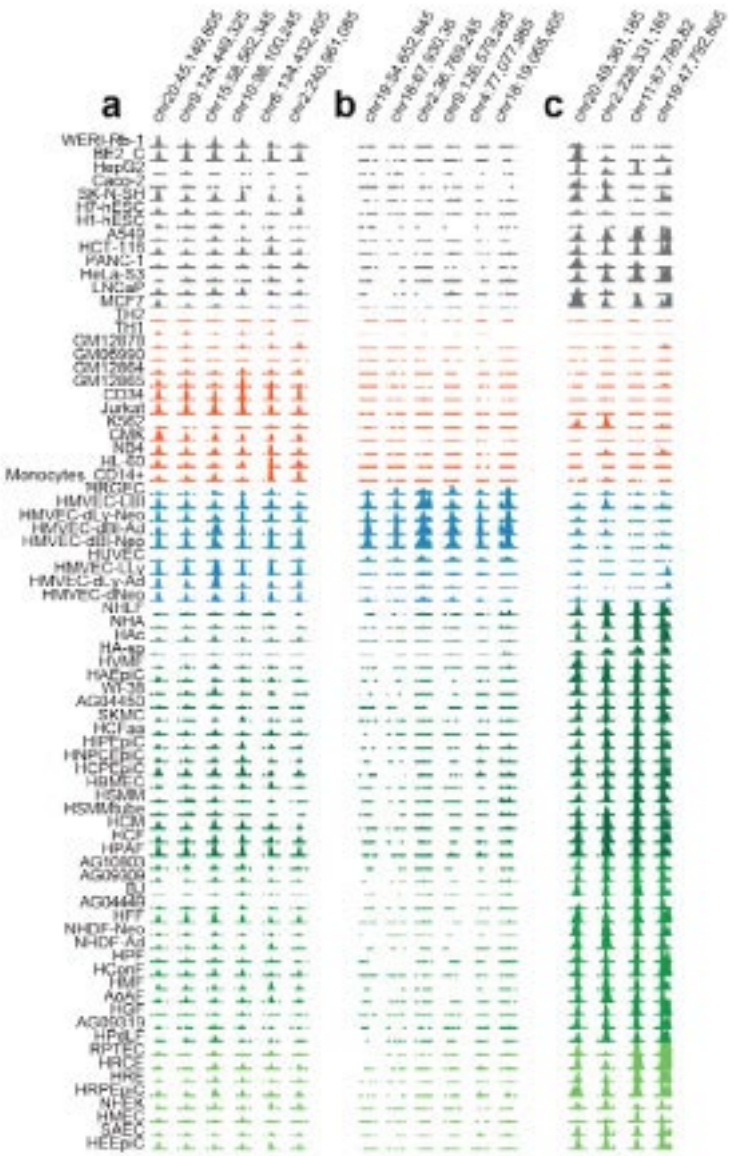


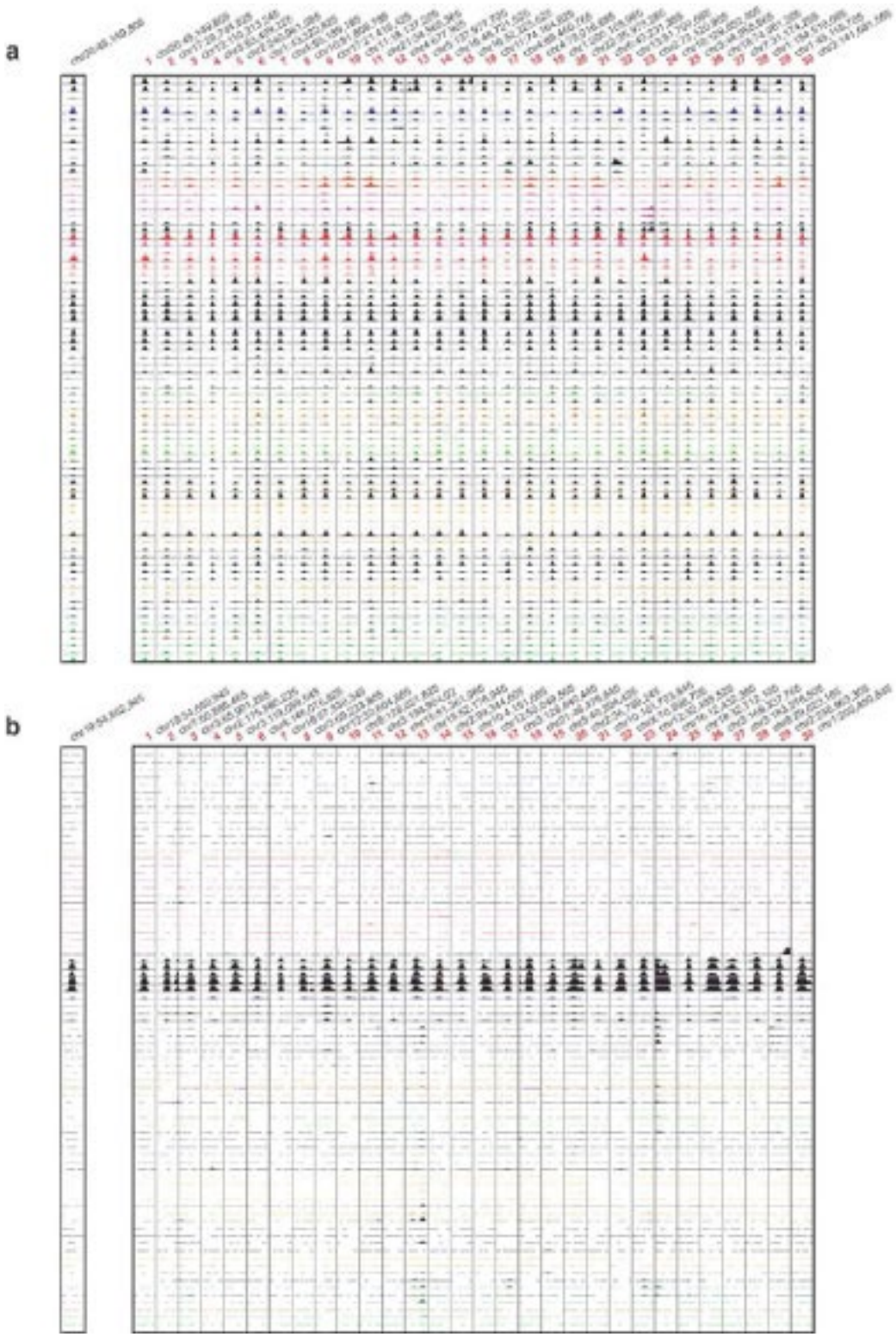
Figure 6 | Stereotyped regulation of chromatin accessibility. (a-e) Enhancers grouped by similar chromatin stereotypes. Related cell lines are colour matched. HS2 from the β-globin locus control region is at left. E1-E11 represent progressively weaker matches to the HS2 stereotype. E12-13 derive from matches to a different stereotype based on another K562 enhancer. (f) Experimental validation of enhancers detected by pattern matching. Bars indicate fold enrichment observed in transient assays in K562 relative to promoter-only control; mean of testing in both orientations is shown. Red bars indicate data from two potent *in vivo* enhancers, β-globin LCR HS2 and HS3; the latter requires chromatinization to function and is not active in transient assays. Gold bars indicate data from E1-E13 from a-e above.



Supplementary Figure 16 | Examples of stereotyped DNaseI patterns across cell lines. (a-c), Examples of stereotyping of DHSs. In each case, a nearly identical cross-cell-type pattern of chromatin accessibility at DHS positions is observed for groups of DHSs widely separated *in trans*. Grey = immortal cells (pluripotent cells and cancer cell lines). Red = hematopoietic cells. Blue = endothelial cells. Green = epithelial, stromal cells, and visceral cells, with shading to denote different pattern groups.

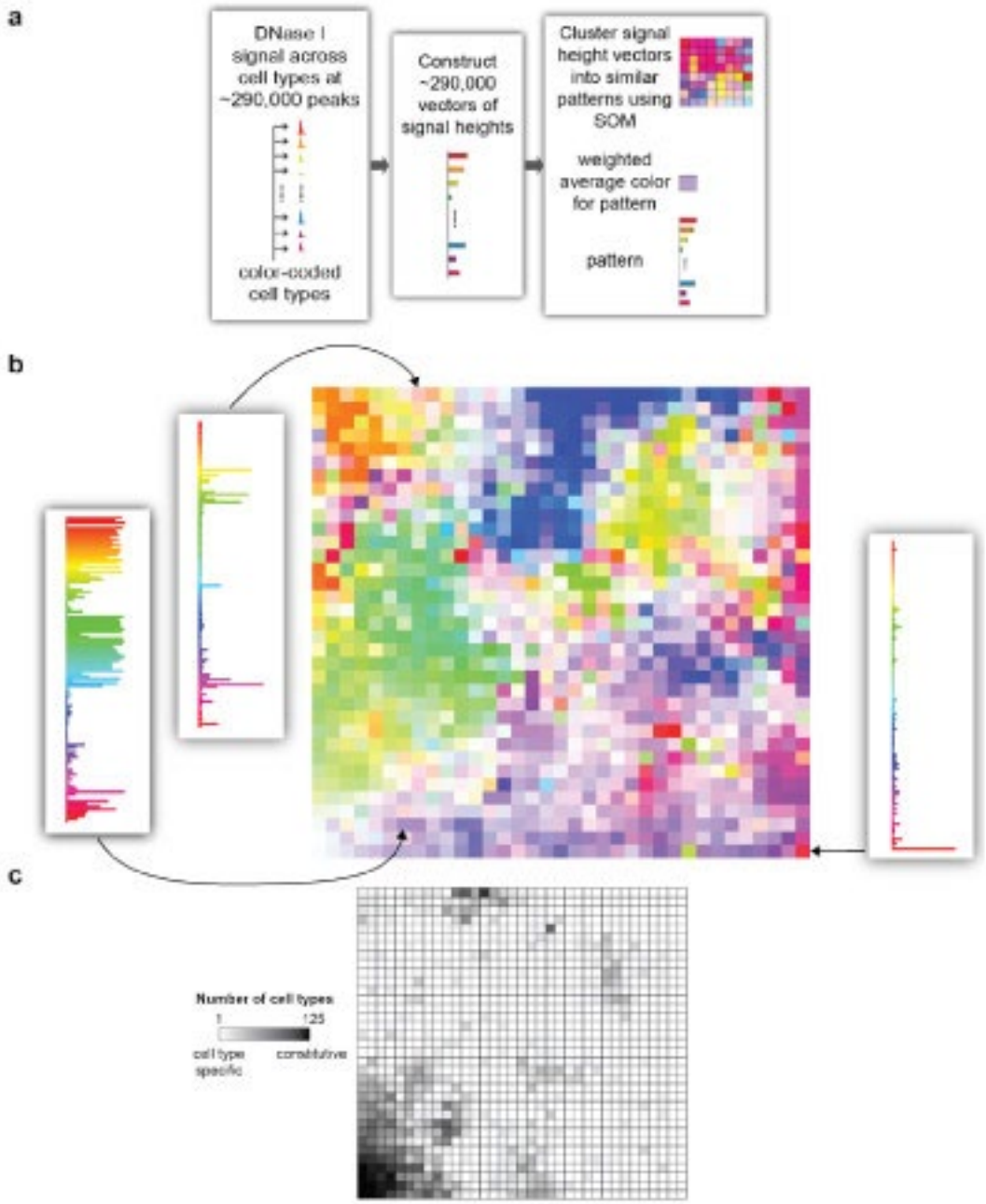
We also tested for significant co-associations between promoter types (defined by the presence of cognate motif classes; see Supplementary Methods) and motifs in paired distal DHSs (Fig. 5c and Supplementary Fig. 15b, c). For example, when a member of the ETS domain family (motifs ETS1, ETS2, ELF1, ELK1, 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 ($P < 10-5$). These results suggest that a limited set of general rules may govern the pairing of co-regulated distal DHSs with particular promoters.

In addition to the synchronized activation of distal DHSs and promoters described above, we observed a surprising degree of patterned co-activation among distal DHSs, with nearly identical cross-cell-type patterns of chromatin accessibility at groups of DHSs widely separated *in trans* (Supplementary Figs 16 and 17). For many patterns, we observed tens or even hundreds of like elements around the genome. The simplest explanation is that such co-activated sites share recognition motifs for the same set of regulatory factors. We found, however, that the underlying sequence features for a given pattern were surprisingly plastic. This suggests that the



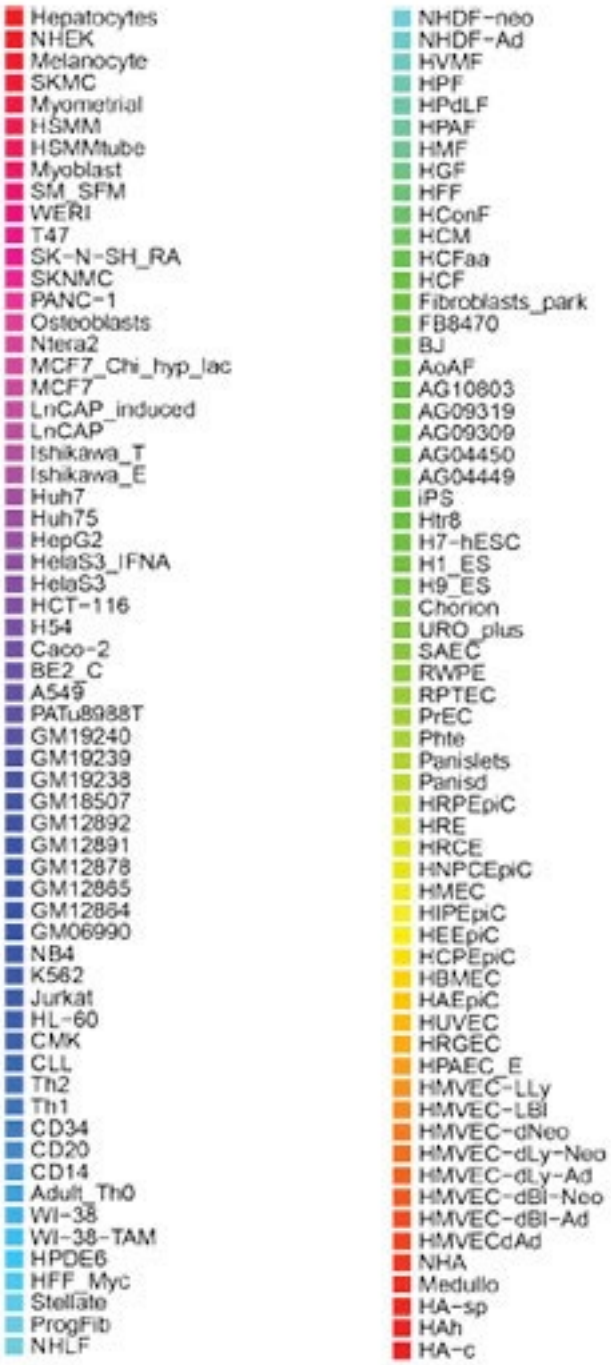
Supplementary Figure 17 | **Top-ranked matches of stereotyped DNaseI patterns across cell lines.** (**a**) Top 30 ranked matches using our DNaseI pattern-matching algorithm (see Supplementary Methods) for the pattern in Supplementary Fig. 16a. (**b**) Top 30 matches for the pattern in Supplementary Fig. 16b. (Cell-type colouring in (**a-b**) does not match that in Supplementary Fig. 16.)

same pattern of cell-selective chromatin accessibility shared between two DHSs can be achieved by distinct mechanisms, probably involving complex combinatorial tuning.



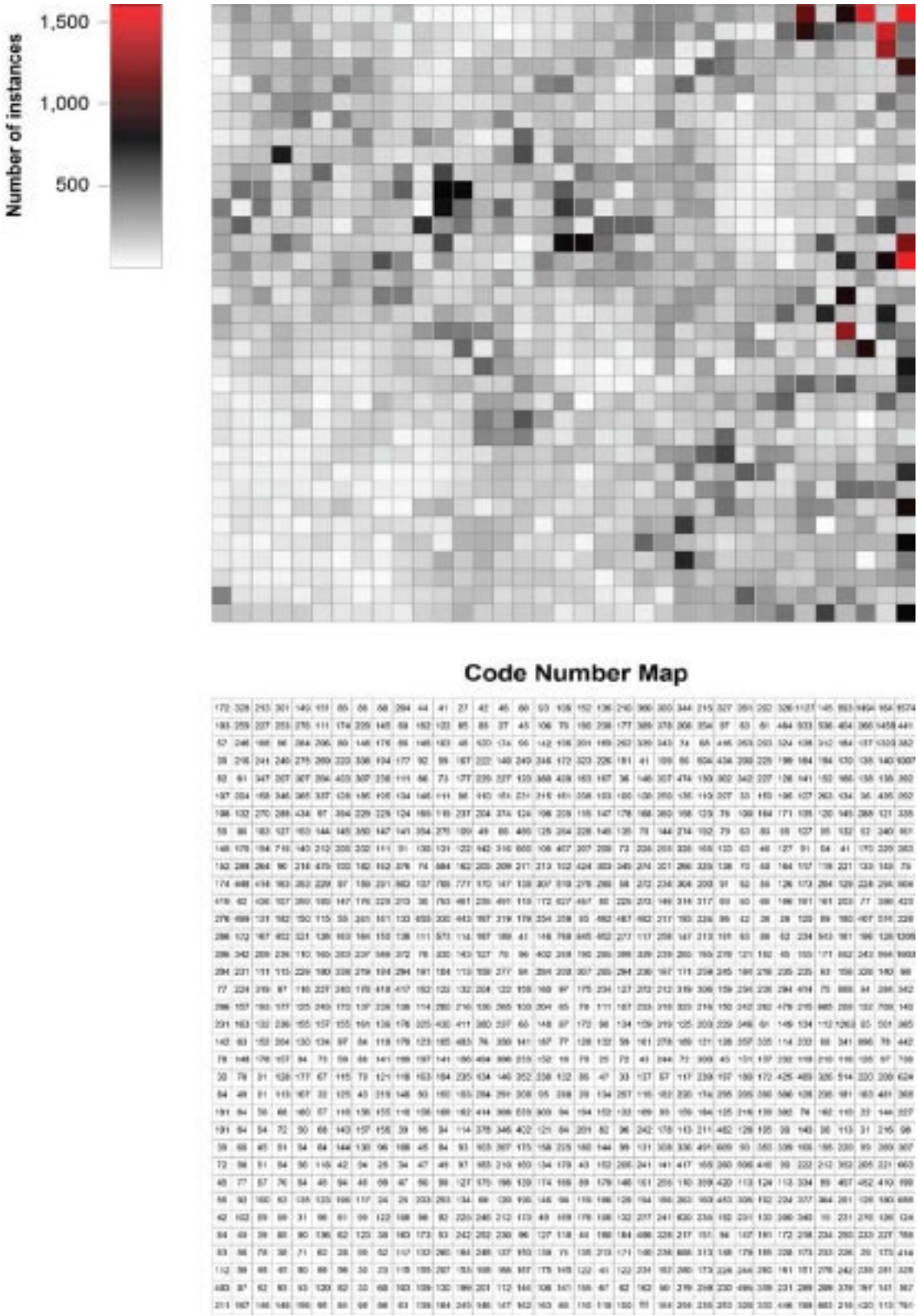
Supplementary Figure 18 | Using a self-organizing map to cluster DHSs by cross-cell-type pattern. Clustering of ~290,000 DHSs by cross-cell-type patterns using a self-organizing map (SOM), which learns patterns in the data and organizes DHSs into stereotyped groups analogous to those shown in Fig. 6a-e. (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.

We next asked whether distal DHSs with specific functions such as enhancers exhibited stereotypical patterning, and whether such patterning could highlight other elements with the same function. We examined one of the best-characterized human enhancers, DNaseI HS2 of the b-globin locus control region[16-18]. HS2 is detected in many cell types, but exhibits potent enhancer activity only in erythroid cells34. Using a pattern-matching algorithm (see Supplementary Methods) we identified additional DHSs with nearly identical cross-cell-type accessibility patterns (Fig. 6a). We selected 20 elements across the spectrum of the top 200 matches to the HS2



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

pattern, and tested these in transient transfection assays in K562 cells (Supplementary Methods). Seventy per cent (14 of 20) of these displayed enhancer activity (mean 8.4-fold over control) (Fig. 6a, f). Of note, one (E3) showed a greater magnitude of enhancement (18-fold versus control) than HS2, which is itself one of the most potent known enhancers4. Next we selected three elements from the 14 HS2-like enhancers, applied pattern matching (Methods) to each to identify stereotyped elements, and tested samples of each pattern for enhancer activity, revealing additional K562 enhancers (total 15 of 25 positive) (Fig. 6b-d, f). In each case, therefore, we were able to discover enhancers by simply anchoring on the cross-cell-type DHS pattern of an element with enhancer activity. Collectively, these results show that co-activation of DHSs reflected in cross-cell-type patterning of chromatin accessibility is predictive of functional activity within a specific cell type, and suggest more generally that DHSs with stereotyped cellular patterning are likely to fulfill similar functions.



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.

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

Supplementary Table 6 | Merging of DHSs from 79 cell types into 32 categories.Grouping of 79 cell types into 32 cell-type categories, for exploration of *cis*-connectivity among DHSs. The grouping was obtained by hierarchically clustering the cell types by their DHS locations across the genome. Descriptions of the cell types are given in Supplementary Table 1.

Category number	Cell types assigned to category
1	Cell types assigned to category WERI_Rb1
2	BE_2_C
3	CACO2, HEPG2, SKNSH
4	HESC, hESCT0
5	A549, HCT116, Hela, PANC1
6	LNCap, MCF7
7	CD56, CD4, hTH1, hTH2
8	GM06990, GM12864, GM12865, GM12878
9	CD34, Jurkat
10	K562, CMK
11	NB4, HL60, CD14
12	HRGEC, HMVEC_LBl, HMVEC_dLyNeo, HMVEC_dBlAd, HMVEC_dBlNeo, HUVEC
13	HMVEC_LLy, HMVEC_dLyAd, HMVEC_dNeo
14	NHLF, NHA
15	HAc
16	HAsp
17	HVMF
18	HAEpiC
19	WI_38, AG04450, IMR90
20	SkMC
21	HCFaa
22	HIPEpiC, HNPCEpiC, HCPEpiC, HBMEC
23	HSMM, HSMM_D
24	HCM, HCF, HPAF
25	AG10803, AG09309, BJ, AG04449, HFF
26	NHDF_Neo, NHDF_Ad
27	HPF, HConF, HMF, AoAF
28	HGF, AG09319, HPdLF
29	RPTEC, HRCE, HRE
30	HRPEpiC
31	HMEC, NHEK
32	SAEC, HEEpiC

discovered by the self-organizing map encompass large numbers of DHSs, with some counting $>$1,000 elements (Supplementary Fig. 20).

Taken together, the above results show that chromatin accessibility at regulatory DNA is highly choreographed across large sets of co-activated elements distributed throughout the genome, and that DHSs with similar cross-cell-type activation profiles probably share similar functions.

Many transcriptional regulators are posited to interact indirectly with the DNA sequence of some target sites through mechanisms such as tethering[25]. Approaches such as ChIP-seq detect chromatin occupancy, but cannot by themselves distinguish sites of direct DNA binding from non-canonical indirect binding. We therefore asked whether DNaseI footprint data could illuminate ChIP-seq-derived occupancy profiles by differentiating directly bound factors from indirect binding events. We first partitioned ChIP-seq peaks from each of 38 ENCODE transcription factors[26] mapped in K562 cells into three categories of predicted sites: ChIP-seq peaks containing a compatible footprinted motif (directly bound sites); ChIP-seq peaks lacking a compatible motif

Supplementary Table 7 | Promotor/distal DHS pairs with correlation ≥7Genomic coordinates of all promoter DHSs and distal, non-promoter DHSs within ±500 kb correlated with them at threshold 0.7. Due to the size of this file, we are making it available through the EBI ftp server.This compressed, tab-delimited text file contains 1,672,958 lines of data, for 63,318 distinct promoter DHSs that each have at least one distal DHS connected to it. Each promoter DHS overlaps a TSS, or is the nearest DHS to the TSS in the 5′ direction; columns 1-3 contain each promoter DHS's genomic coordinates (hg19). The Gencode gene names are given in column 4. Because distinct gene names can be given to the same TSS, and because distinct TSSs can have the same nearby DHS called as their promoter DHS, data for each promoter DHS is repeated in this file roughly three times on average, with a different gene name for each repetition (there are 207,878 distinct combinations of promoter DHS + gene name in this file). Columns 5-7 contain the genomic coordinates for each distal, non-promoter DHS within 500kb of the promoter DHS given in columns 1-3 that achieves correlation ≥0.7 with it; the correlation between the promoter/distal DHS pair is given in column 8. Distal DHSs appear multiple times in the file when they achieve correlation ≤0.7 with multiple promoter DHSs. Using program sort-bed from the BEDOPS genomic data analysis software suite, from the command line within a Unix system, the set of 578,905 distal DHSs connected with at least one promoter DHS can be extracted into a file named "outfile" by executing the commandcut -f5-7 infile | sort-bed - | uniq > outfilewhere "infile" represents the file genomewideCorrs_above0.7_promoterPlusMinus500kb_withGeneNames_32celltypeCategories.bed8.The first five lines of data are shown below.

chr1	66660	66810	AL627309.1	chr1	87640	87790	0.87171
chr1	66660	66810	AL627309.1	chr1	118840	118990	0.908176
chr1	66660	66810	AL627309.1	chr1	136960	137110	0.915177
chr1	66660	66810	AL627309.1	chr1	566760	566910	0.731457
chr1	96520	96670	RP11-34P13.8	chr1	237020	237170	0.786171

or footprint (indirectly bound sites); and ChIP-seq peaks overlying a compatible motif lacking a footprint (indeterminate sites). Predicted indirect sites showed significantly reduced ChIP-seq signal compared with predicted directly bound sites (Supplementary Fig. 10), consistent with lack of direct crosslinking to DNA (and therefore reduced ChIP efficiency). Indeterminate sites exhibited low ChIP-seq signal and were therefore excluded from further analysis (Supplementary Fig. 10).

The fraction of ChIP-seq peaks predicted to represent direct versus indirect binding varied widely between different factors, ranging from nearly complete direct sequence-specific binding (for example, CTCF), to nearly complete indirect binding (for example, TBP; Supplementary Fig. 11). In many cases factors that preferentially engage in direct DNA binding at distal sites show predominantly indirect occupancy in promoter regions and vice versa (Supplementary Fig. 12a, b).

Next, we analysed the frequency with which indirectly bound sites of one transcription factor coincided with directly bound sites of a second factor, indicative of protein-protein interactions (for example, tethering). This analysis recovered many known protein-protein interactions, such as CTCF-YY1 and TAL1-GATA1 [27], as well as many novel associations (Fig. 5). We observed enrichment for NFE2 indirect interactions at promoter-bound USF2 sites, compatible with their known interaction[28]. At distal sites, we observed the opposite, with NFE2 predominantly directly bound accompanied by USF2 indirect peaks (Supplementary Fig. 12a, b), indicating the possibility of a reciprocal or looping mechanism. Notably, directly bound promoter-predominant transcription factors were enriched for co-localization with indirect peaks compared to distal regions (Supplementary Fig. 13a, b). These results suggest that combining DNaseI footprinting with ChIP-seq has the potential to expose a previously unappreciated landscape of complex transcription factor occupancy modes.