8 Enhancer discovery and characterization

ENCODE has developed methods to discover enhancers, and characterized them using comparisons with other data sets and by molecular biology experiments

To identify functional regions genome-wide, we next integrated elements independent of genomic landmarks using either discriminative training methods, where a subset of known elements of a particular class were used to train a model that was then used to discover more instances of this class, or using methods in which only data from ENCODE assays were employed without explicit knowledge of any annotation.

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

We tested a subset of these predicted enhancers in both Mouse and Fish transgenic models (examples in Figure 6), with over half of the elements showing activity, often in the corresponding tissue type.

Characterization of enhancer RNA

It has recently been reported that RNA polymerase II binds some distal enhancer regions and can produce enhancer-associated transcripts named eRNA[34-36]. We used our RNA assays to detect and characterize transcriptional activity at enhancer loci predicted genome-wide from ENCODE ChIP-seq data[21,37].

Figure 5a shows the aggregate pattern of RNA-seq and CAGE signal in a strand-specific manner around the subset of predicted gene-distal enhancers containing DNaseI hypersensitive sites and centered on those sites. In these plots, as denoted by the accumulation of CAGE tags signifying TSSs, transcription initiation within the enhancer region is observed, and continues outwards for several kilobases (kb). This behaviour can be observed for the polyadenylated and non-polyadenylated RNA fractions mapping in both intronic and intergenic regions. As previously reported[34], we observe a large diversity of expression levels at each of the transcribed enhancers. Polyadenylated to non-polyadenylated RNA ratios, as well as nuclear to cytoplasmic ratios, vary at individual enhancers (Supplementary Fig. 21a, b). However, contrary to some previous reports, although most eRNAs are prevalent in the nuclear non-polyadenylated RNA fraction, some eRNAs seemed to be polyadenylated in the nucleus. This pattern was significantly different compared to transcripts from Gencode annotated and novel predicted[21] promoters (Fig. 5b).

Transcribed enhancers on average show a significantly different pattern of chromatin modification than non-transcribed ones[38-41]. The enhancer regions displayed stronger signals for H3K4 methylation, H3K27 acetylation and H3K79 dimethylation along with higher levels of RNA polymerase II binding, all associated with transcriptional initiation and elongation (Fig. 5c). Both the transcripts and the chromatin states are cell-type specific (Fig. 5d). Taking the GM12878 cell line as an example, the enhancer loci producing eRNA demonstrate enrichment of CAGE tag detection (Fig. 5d, top) and the presence of H3K27ac histone modification (Fig. 5d, bottom) in this cell line compared to five other analysed cell lines. This strongly suggests that the regulatory

regions governing the expression of enhancer transcripts are distinguished from regulatory regions located at the beginning of genic regions.

In order to identify general enhancers, we modified our prediction procedure to replace information specific to the mouse assay, such as the binding motifs of TRFs expressed in mouse embryos, by some general features of enhancers, such as signals of the histone modification H3K4me1. We developed two complementary methods, and took the intersection of them as our high-confidence predictions (Materials and methods). In total we identified 13,539 potential enhancers (full list available in the Additional files), among which 50 were randomly chosen. 20 of them were tested by the mouse assay, and an independent set of 27 were tested by the Medaka fish assay (Materials and methods).

The validation results for the mouse and fish assays are shown in Table 3 and Table 4, respectively. In the mouse experiments, 6 of the 20 (30%) tested sequences showed enhancer activities in various types of tissues in the nose, heart, limb and tail. In the fish experiments, 19 of the 27 (70%) tested sequences showed some enhancer activities, out of which 15 (56%) had strong activities.

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 19258 distal TF-TF edges (Fig. 2A).

Active enhancers often express enhancer RNAs[26]. We used a comprehensive enhancer RNA dataset generated by the ENCODE consortium to determine whether TSSs preferentially interact with active enhancer-like elements[27]. We find that E-elements that are looping to TSSs are significantly more likely to express enhancer RNAs than E-elements that are not looping ($P<5\times10\text{-}5$, hypergeometric test, Supplementary Figure 10; Enhancer RNAs taken from[27]). We conclude that looping interactions preferentially involve active enhancer-like elements.

In each cell line we identified large numbers of statistically significant TSS-distal fragment interactions, of which ~60% were observed in only one of the 3 cell lines (Figure 2a). These data point to intricate cell type specific three-dimensional folding of chromatin. 3C-based assays detect specific and functional interactions, e.g. TSSs with gene regulatory elements[8]. In addition the assay will detect "structural" interactions, e.g. close spatial proximity as a result of other nearby specific looping interactions (bystander interactions) or overall higher order folding of the chromatin fiber. To determine which looping interactions involved distal sites that displayed specific chromatin features associated with functional elements we compared our data with datasets generated by the ENCODE consortium (Figure 2b; Supplementary Table 7). We find that looping interactions in all cell lines are significantly enriched for distal fragments that are bound by CTCF, a protein known to mediate DNA looping[21], contain open chromatin (as determined by FAIRE[22] or DHS mapping[20]), and/or histones with modifications that are characteristic for active functional elements (H3K4me1, H3K4me2, H3K4me3). Long-range interactions are also enriched for H3K9ac and H3K27ac, but are not enriched or significantly depleted for H3K27me3, a mark typically found at inactive or closed chromatin.

To gain more insights into the types of elements present in the distal looping fragments we made use of genome-wide and cell line specific segmentation analyses that identified seven distinct chromatin states based on histone modifications, the presence of DHSs and the localization of proteins such as RNA polymerase II and CTCF ([23]; Figure 2b). These states are 1) "Enhancer" (E), 2) "Weak Enhancer" (WE), 3) "TSS", 4) "Predicted Promoter Flanking regions" (PF), 5) "Insulator element" (CTCF), 6) "Predicted Repressed region" (R) and 7) "Predicted Transcribed region" (T). The ENCODE consortium tested sets of the E elements in enhancer assays and confirmed that >50% display enhancer activity[4]. We find that looping interactions are significantly enriched

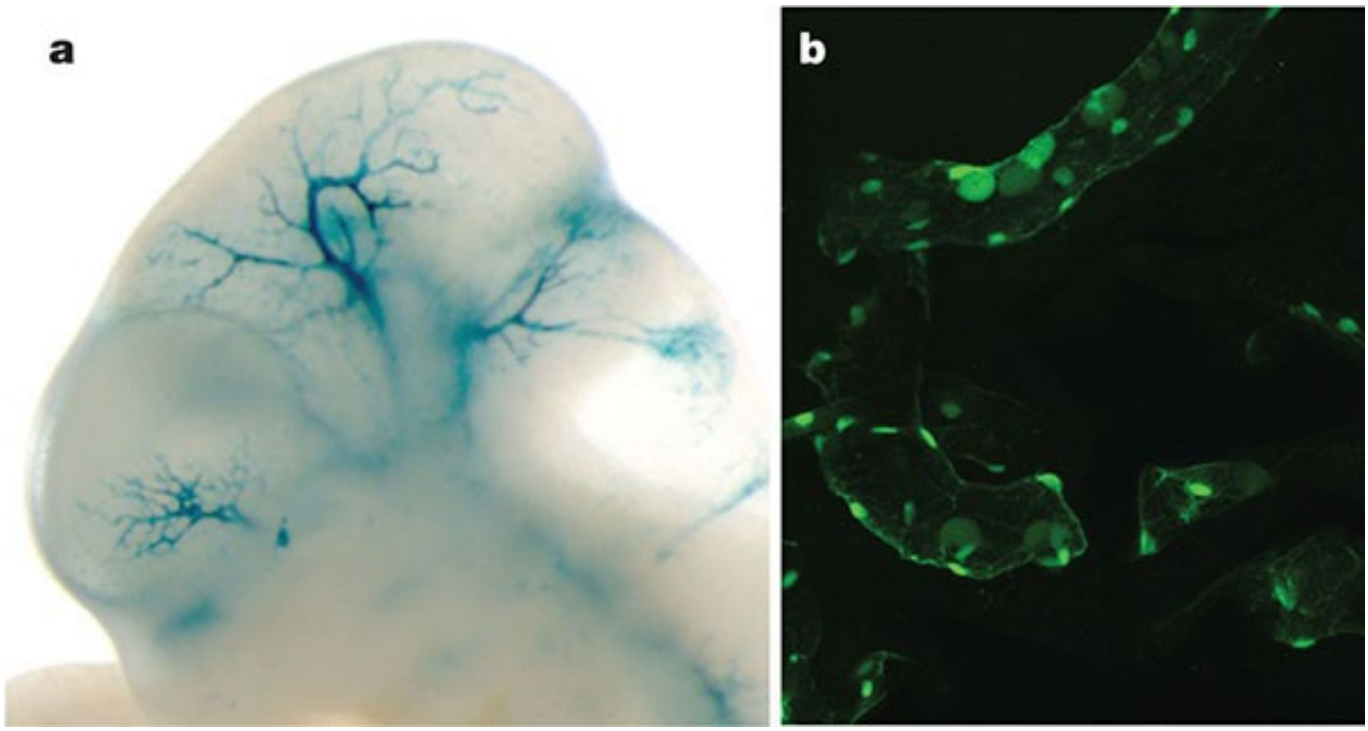
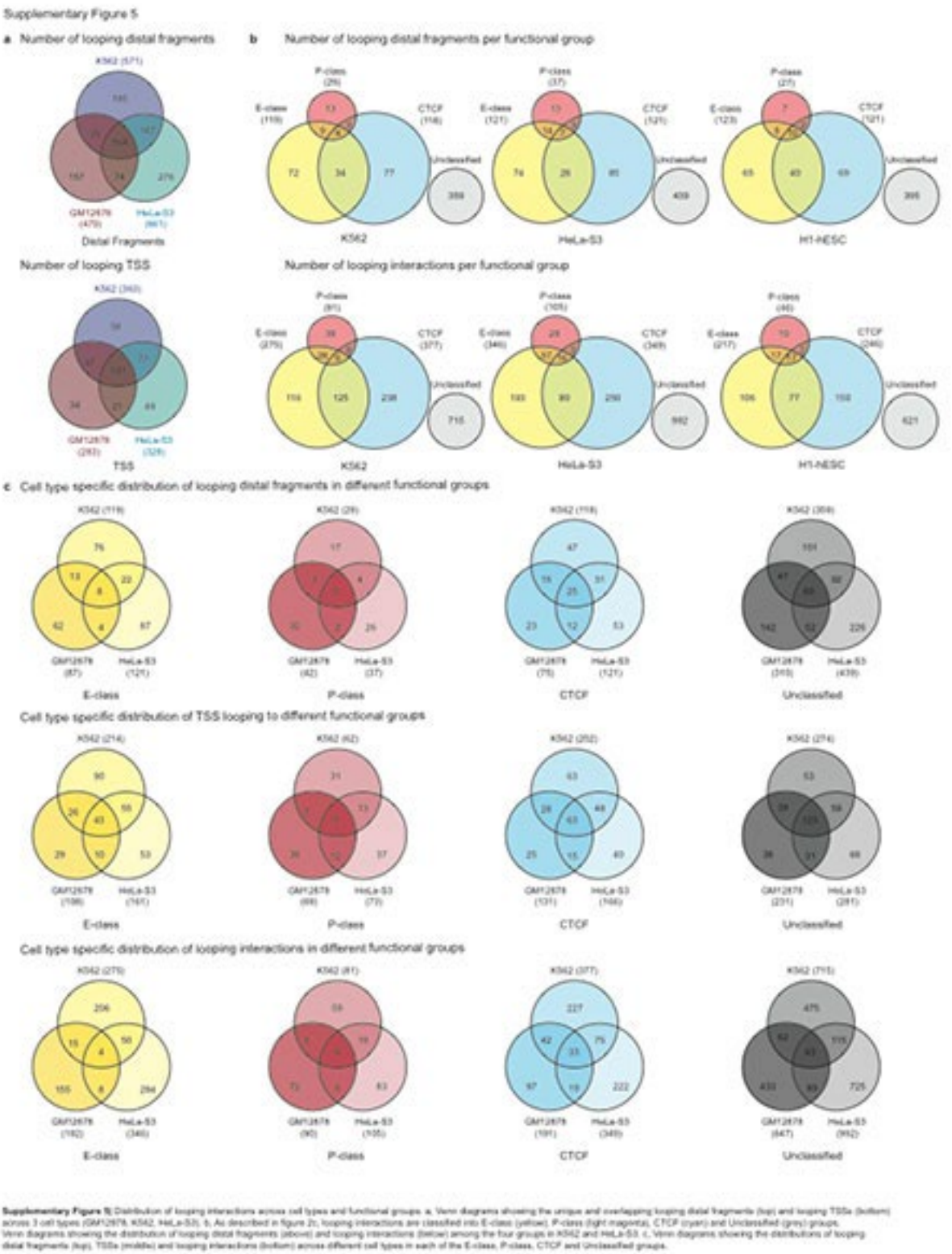


Figure 6 | Experimental characterization of segmentations. Randomly sampled E state segments (see Table 3) from the K562 segmentation were cloned for mouse- and fish-based transgenic enhancer assays. (a) Representative LacZ-stained transgenic embryonic day (E)11.5 mouse embryo obtained with construct hs2065 (EN167, chr10: 46052882-46055670, GRCh37). Highly reproducible staining in the blood vessels was observed in 9 out of 9 embryos resulting from independent transgenic integration events. (b) Representative green fluorescent protein reporter transgenic medaka fish obtained from a construct with a basal *hsp70* promoter on meganuclease-based transfection. Reproducible transgenic expression in the circulating nucleated blood cells and the endothelial cell walls was seen in 81 out of 100 transgenic tests of this construct.

for distal fragments that contain E, WE and CTCF elements, and the actively transcribed chromatin state ("T"), but are depleted for the repressed chromatin state ("R"). We note that some distal looping fragments contain elements classified as "TSS" or "PF", even though they do not contain TSSs as defined by the GENCODE v7 annotation[24]. Possibly, these are yet to be annotated TSSs.

We find that TSS-Enhancer and TSS-Promoter interactions are more cell type specific than TSS-CTCF interactions: in case of the TSS-Enhancer and TSS-Promoter categories the ratio of interactions that is seen in only one cell line vs more than one cell line is $\sim$4:1, whereas it is close to $\sim$1:1 for TSS-CTCF category (Supplementary figure 5).

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 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 $\pm$500 kb (Supplementary Fig. 14a). We identified a total of 578,905 DHSs that were highly correlated ($r > 0.7$) with at least one promoter ($P < 10-100$), 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 hepatic cells, and an enhancer has been defined upstream of its TSS (Fig. 5a). The correlation



Supplementary Figure 5 | **Distribution of looping interactions across cell types and functional groups.** (**a**) Venn diagrams showing the unique and overlapping looping distal fragments (top) and looping TSSs (bottom) across 3 cell types (GM12878, K562, HeLa-S3). (**b**) As described in figure 2c, looping interactions are classified into E-class (yellow), P-class (light magenta), CTCF (cyan) and Unclassified (grey) groups. Venn diagrams showing the distribution of looping distal fragments (above) and looping interactions (below) among the four groups in K562 and HeLa-S3. (**c**) Venn diagrams showing the distributions of looping distal fragments (top), TSSs (middle) and looping interactions (bottom) across different cell types in each of the E-class, P-class, CTCF and Unclassified groups.

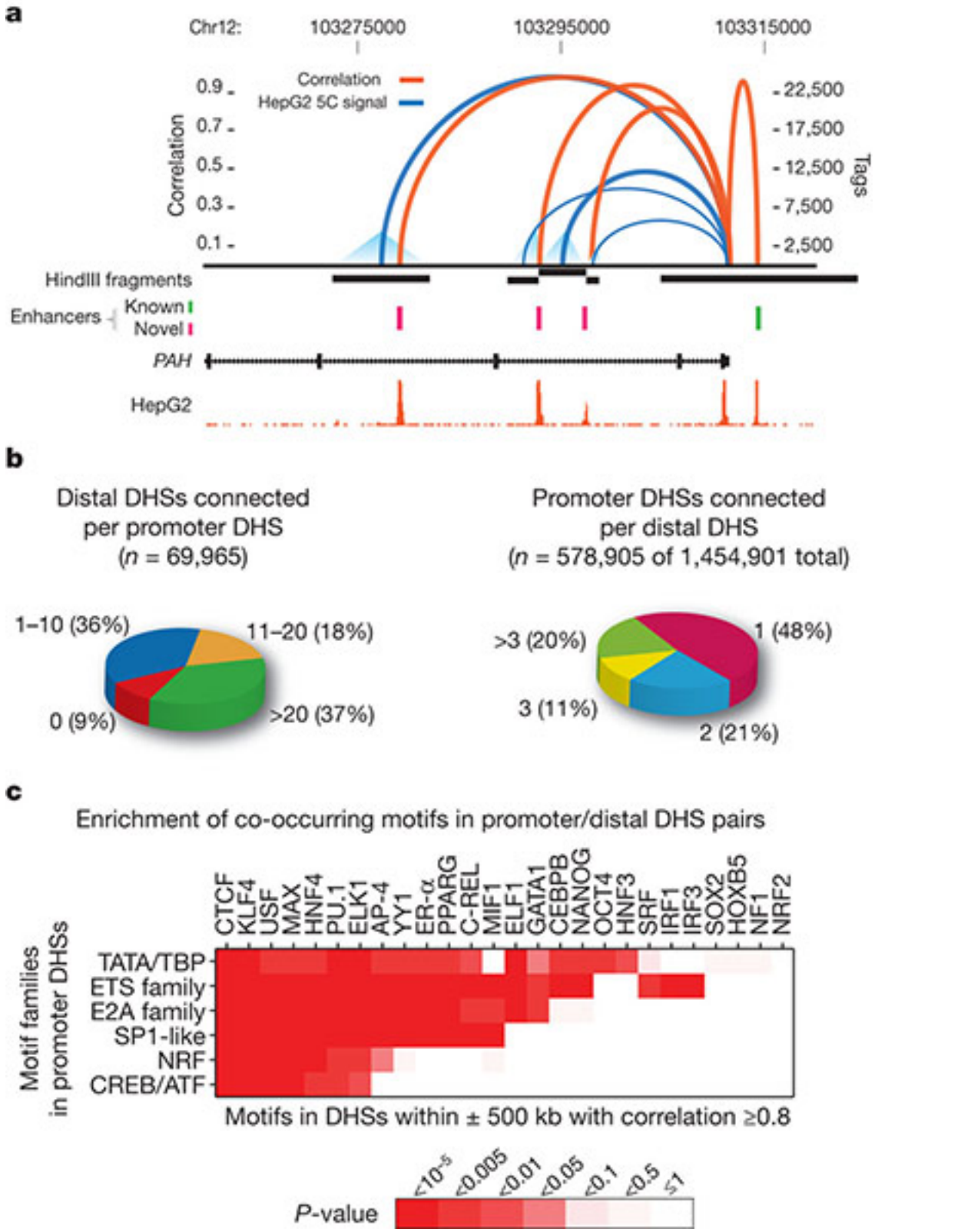
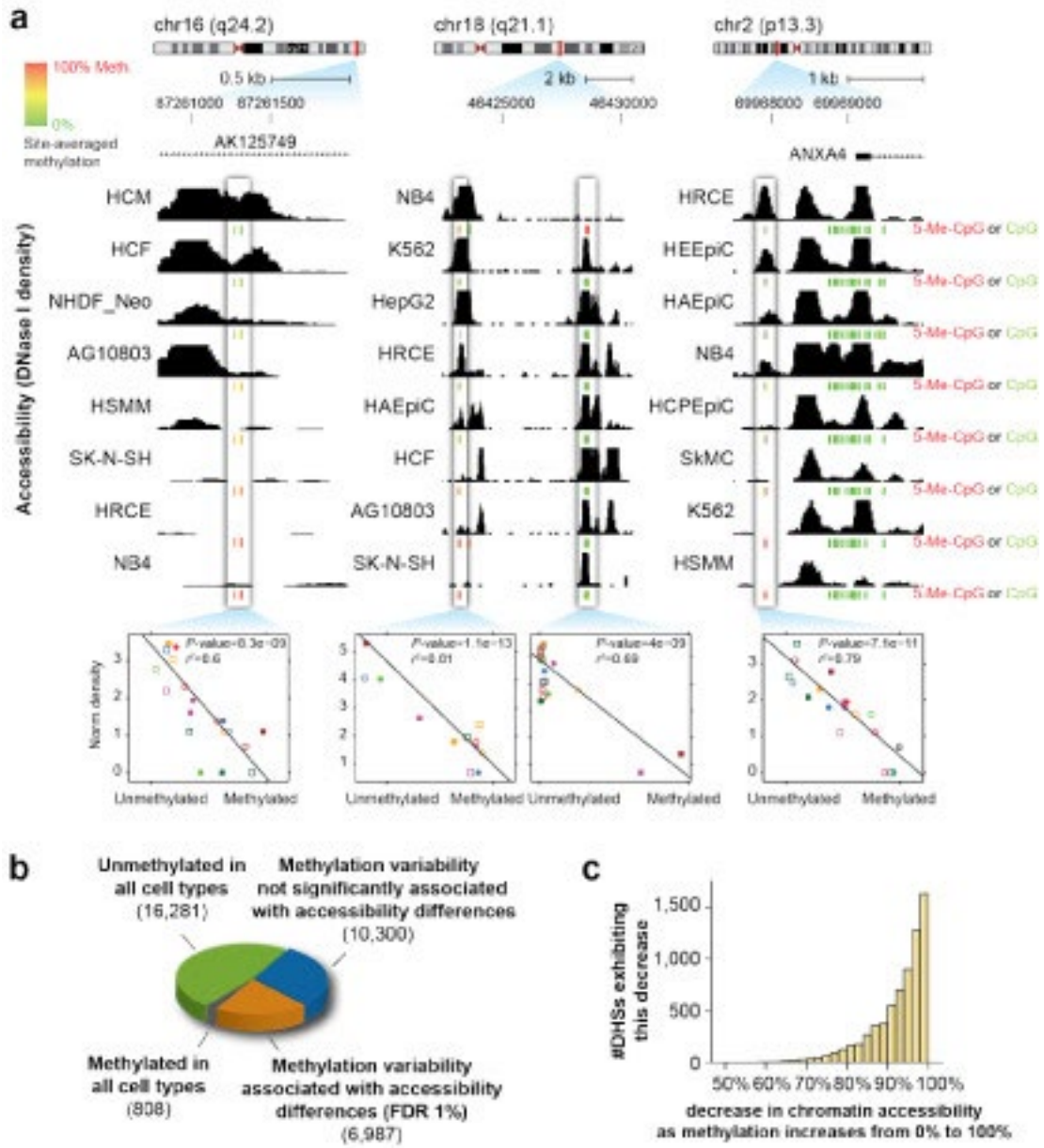


Figure 5 | A genome-wide map of distal DHS-to-promoter connectivity. (a) Cross-cell-type correlation (red arcs, left *y* axis) of distal DHSs and *PAH* promoter closely parallels chromatin interactions measured by 5C-seq (blue arcs, right *y* axis); black bars indicate HindIII fragments used in 5C assays.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

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

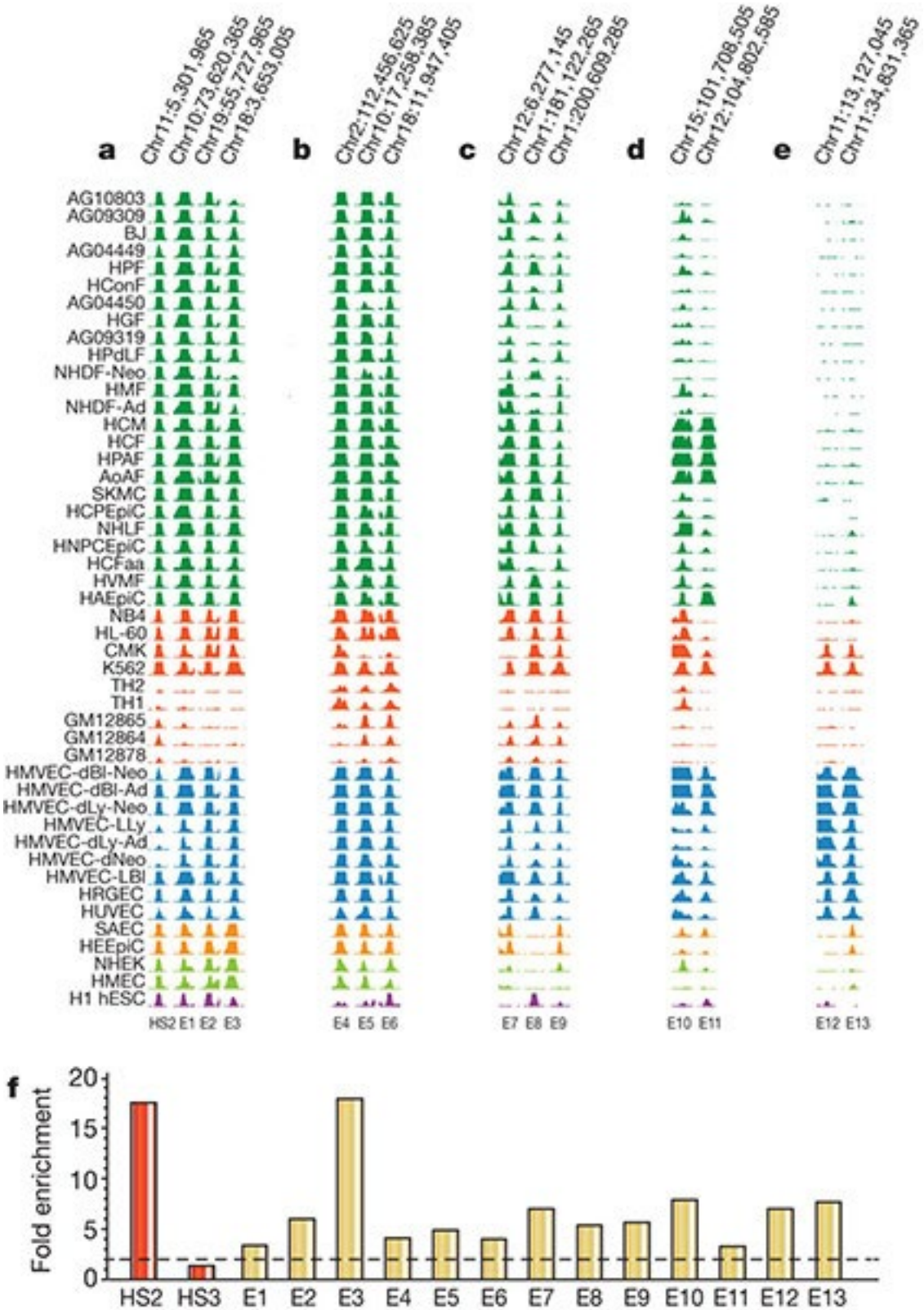


Figure 6 | **Stereotyped regulation of chromatin accessibility. (a-e)** Enhancers grouped by similar chromatin stereotypes. Related cell lines are colour matched. HS2 from the β-globin locus control region is at left. E1-E11 represent progressively weaker matches to the HS2 stereotype. E12-13 derive 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.

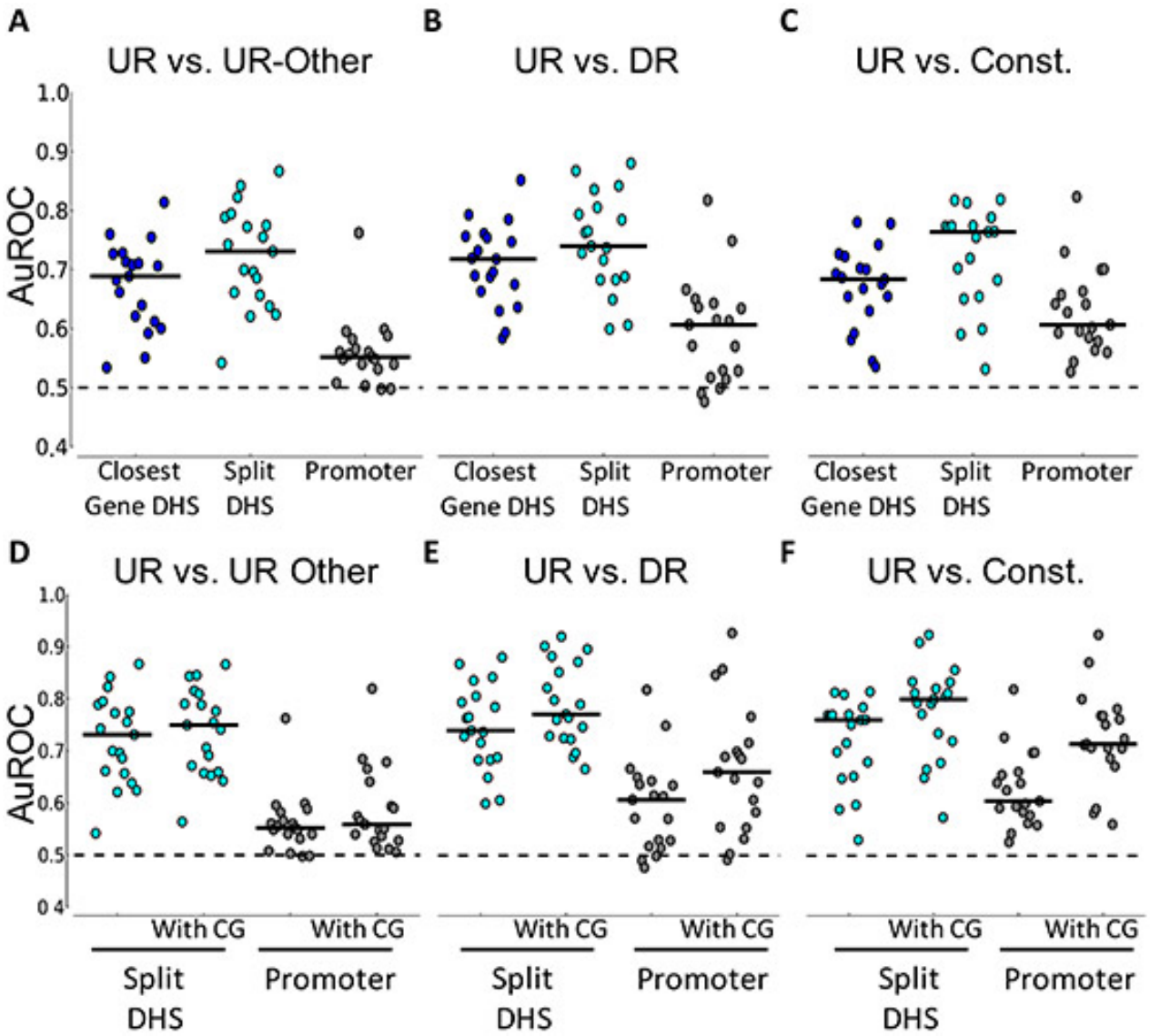


Figure 5 | **Classifier performance for various classification tasks.** (**a-c**) Performance of the classifier using all PWMs. Each figure compares the performance of two methods of associating DHSs to genes (Closest Gene DHS and Split DHS) with the proximal promoter. The solid black lines across the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

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

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

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 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.

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 β-globin locus control region[16-18]. HS2 is detected in many cell types, but exhibits potent enhancer activity only in erythroid cells[34]. 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 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

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 enhancers[4]. 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 fulfil similar functions.

DRMs have stronger binding signals of CTCF and the cohesin proteins RAD21 and SMC3 than PRMs, which in turn have stronger binding signals than the whole genome in general. The stronger signals at DRMs than PRMs is consistent with the known role of CTCF in binding insulators[51-52] and the frequent co-occurrence of the binding sites of CTCF and the cohesin complex[53-54]. On the other hand, the stronger signals at PRMs than the genomic background suggests that CTCF also binds some proximal regions, which may reflect the ability of it to act as transcriptional insulator, repressor or activator depending on the context of the binding site[55-56]. A recent study also found that, contrary to the enhancer blocking model, CTCF may actually promote communications between functional regulatory elements by connecting promoters and enhancers through long-range DNA interactions[57].

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

First round of validation: Human enhancers active in mouse embryos

We first predicted potential human enhancers that are active in mouse embryos on embryonic day 11.5. Specifically, from the list of BARs, we selected those that are far away from TSSs and exons, and scored them based on both their sequence conservation and the presence of motifs of TRFs known to be expressed in mouse embryos (Materials and methods). We then took the top 50 predictions, and randomly chose 6 of them for experimental validation (Additional file 1, Table S3). These 6 regions were extended according to some experimental requirements, and tested for enhancer activities in a mouse assay previously established[61]. These

experiments were performed by Dr. Len Pennacchio's group, for testing a larger cohort of in total 33 potential enhancers identified by several sub-groups of the ENCODE consortium using different prediction methods (Pennacchio and The ENCODE Project Consortium, unpublished data).

Among our 6 tested predictions, 5 (83%) were found to have enhancer activities in various tissues with good reproducibility (Table 2, data available at the VISTA database[6]). Interestingly, most predicted enhancers were found to be active in tissues related to neurodevelopment, which is likely due to the particular set of development-related TRFs we considered in our method.

We then examined the TRFs associated with the DRM-target transcript pairs. We found that DRMs potentially regulating Poly A+ transcripts have a higher fraction of EP300 binding than both the set of all DRMs and the whole genome (except in H1-hESC, which has too few DRMs to compute the fraction accurately) (Additional file 1, Table S4). This observation suggests that the correlation method for associating DRMs and target transcripts could help identify DRMs that have stronger activities.

The performance of the classifier using only proximal promoter information is close to that of a random classifier, across all tasks. All the classifiers using DHS sequences display strong improvements in performance over this baseline in discriminating genes that are up-regulated in different cell types (UR vs. UR-Other, Figure 5A), with a greater improvement in performance coming from the Split DHS approach with separate features for the TSS and Distal DHSs (median AuROC ~0.73). Similar results were obtained when training classifiers to distinguish between specifically up- and down-regulated genes from the same cell types (UR vs. DR, Figure 5B), and to distinguish up-regulated from constitutively expressed genes (UR vs. Const., 11 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.

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.

TCF7L2 binds to enhancer regions

The fact that TCF7L2 can bind to regions far from core promoters suggested that TCF7L2 might bind to enhancers. Recent studies have shown that enhancers can be identified by enrichment for both the H3K4me1 and H3K27Ac marks [25, 27]. To determine if the regions bound by TCF7L2 are also bound by these modified histones, we performed ChIP-seq experiments in PANC1, HEK293, HCT116 and MCF7 cells using antibodies that specifically recognize histone H3 only when it is monomethylated on lysine 4 or when it is acetylated on lysine 27; we also used H3K4me1 and H3K27Ac data for HeLa and HepG2 cells from the ENCODE Project. Duplicate ChIP-seq experiments were performed using two different cultures of cells for each cell type, peaks were called individually to demonstrate reproducibility (Additional file 4), the reads were merged and a final peak set for both H3K4me1 and H3K27Ac was obtained. We then identified predicted active enhancers as regions having both H3K4me1 and H3K27Ac and determined the percentage of the TCF7L2 sites that have

either or both of the modified histones (Table 2). We found that, for most cells, the majority of TCF7L2 sites co-localized with H3K4me1 and H3K27Ac. However, a smaller percentage of the TCF7L2 sites in MCF7 cells co-localized with active enhancers. Heatmaps of the tag density of the histone ChIP-seq experiments for each cell line relative to the center of the TCF7L2 peak locations are shown in Figure 3C. Although most TCF7L2 binding sites show robust levels of both marks, the TCF7L2 sites in MCF7 cells again show a smaller percentage of sites having high levels of the modified histones. To determine if the TCF7L2 binding sites in MCF7 cells correspond to sites bound by histone modifications associated with transcriptional repression, we performed duplicate ChIP-seq analysis using antibodies to H3K9me3 and H3K27me3; we also used H3K4me3 and RNA Polymerase II ChIP-seq data from the ENCODE Project. As shown in Figure 3D, neither the proximal or distal TCF7L2 binding sites show high levels of H3K9me3 or H3K27me3.

To further investigate the role of TCF7L2 in cell type-specific enhancers, we determined the percentage of active enhancers in each of the 6 cell types (i.e. genomic regions bound by both H3K4me1 and H3Kk27Ac) that are also bound by TCF7L2. We found that more than 40% of all enhancers in the different cell lines are occupied by TCF7L2 (Figure 3B). These results indicate that TCF7L2 ChIP-seq data identifies many of the active enhancers in a given cell type and suggests that TCF7L2 may play a critical role in specifying the transcriptome in a variety of cancer cells. An example of TCF7L2 binding to sites marked by H3K4me1 and H3K27Ac in HepG2 cells is shown in Additional file 12; TCF7L2 does not bind to this same site in HeLa cells and these sites are also not marked by the modified histones in HeLa cells.