The accessible chromatin landscape of the human genome

Published in final edited form as:
Nature. 2012 September 6; 489(7414): 75–82. doi:10.1038/nature11232.

The accessible chromatin landscape of the human genome

Robert E. Thurman[1,*], **Eric Rynes**[1,*], **Richard Humbert**[1,*], **Jeff Vierstra**[1], **Matthew T. Maurano**[1], **Eric Haugen**[1], **Nathan C. Sheffield**[2], **Andrew B. Stergachis**[1], **Hao Wang**[1], **Benjamin Vernot**[1], **Kavita Garg**[3], **Richard Sandstrom**[1], **Daniel Bates**[1], **Theresa K. Canfield**[1], **Morgan Diegel**[1], **Douglas Dunn**[1], **Abigail K. Ebersol**[4], **Tristan Frum**[4], **Erika Giste**[1], **Lisa Harding**[4], **Audra K. Johnson**[1], **Ericka M. Johnson**[4], **Tanya Kutyavin**[1], **Bryan Lajoie**[5], **Bum-Kyu Lee**[6], **Kristen Lee**[1], **Darin London**[2], **Dimitra Lotakis**[4], **Shane Neph**[1], **Fidencio Neri**[1], **Eric D. Nguyen**[4], **Alex P. Reynolds**[1], **Vaughn Roach**[1], **Alexias Safi**[2], **Minerva E. Sanchez**[4], **Amartya Sanyal**[5], **Anthony Shafer**[1], **Jeremy M. Simon**[7], **Lingyun Song**[2], **Shinny Vong**[1], **Molly Weaver**[1], **Zhancheng Zhang**[7], **Zhuzhu Zhang**[7], **Boris Lenhard**[8], **Muneesh Tewari**[3], **Michael O. Dorschner**[9], **R. Scott Hansen**[4], **Patrick A. Navas**[4], **George Stamatoyannopoulos**[4], **Vishwanath R. Iyer**[6], **Jason D. Lieb**[7], **Shamil R. Sunyaev**[10], **Joshua M. Akey**[1], **Peter J. Sabo**[1], **Rajinder Kaul**[4], **Terrence S. Furey**[7], **Job Dekker**[5], **Gregory E. Crawford**[2], and **John A. Stamatoyannopoulos**[1,11,†]

[1]Department of Genome Sciences, University of Washington, Seattle, WA

[2]Institute for Genome Sciences and Policy, Duke University, Durham, NC

[3]Division of Human Biology, Fred Hutchinson Cancer Research Center, Seattle, WA

[4]Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, WA

[5]Program in Gene Function, University of Massachusetts Medical School, Worcester, MA

[6]Institute for Cellular and Molecular Biology, University of Texas, Austin, TX

[7]Department of Biology, University of North Carolina, Chapel Hill, NC

[8]Bergen Center for Computational Science, University of Bergen, Bergen, Norway

[9]Dept. of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA

[†]correspondence: jstam@uw.edu.
[*]these authors contributed equally

Author Contributions

Generation of DNaseI data was supervised by J.A.S. and G.E.C., with data collection carried out by D.B., T.K.C., R.S.H., M.D., D.D., E.G., T.K., K.L., F.N., V.R., A.S. (UW), S.V., M.W., B-K.L., D.L., A.S., L.S., Z.Z, and Z.Z. (Duke). 5C experiments were supervised by J.D. and performed by A.S. (UMass). Primary DNaseI data processing was performed by R.S., T.S.F., A.K.J., and A.P.R. Hypersensitivity Southerns and Enhancer cloning and transfection experiments were performed by E.M.J., A.K.E., T.F., E.D.N., L.H., and M.S.and supervised by P.A.N. and G.S. H3K4me3 ChIP-seq experiments were performed by H.W. Primary analysis of DNaseI data was performed by R.E.T., R.S., and R.H. Joint analysis of DNaseI and transcription factor ChIP-seq data was performed by J.V. and A.B.S. Promoter prediction analysis was performed by R.E.T. DNaseI vs. DNA methylation analysis was performed by M.T.M. DHS-promoter connectivity analysis was performed by E.R. Integration of DNaseI and 5C data was performed by R.H. with assistance from B.L. (UMass). DHS stereotyping pattern analysis was performed by E.H. Self-organizing map analysis was performed by N.S. and B.L. (Bergen). Variation analysis was performed by B.V. and E.R. under direction of S.S., J.M.A., and J.A.S. Data interpretation and figure design were performed by J.A.S., R.E.T., J.D.L., V.R.I., G.E.C., and T.S.F. J.A.S., R.E.T., E.R., R.H., J.V., M.T.M., A.B.S., and N.S. wrote the paper.

Competing Interests

The authors declare no competing interests.

Data Availability

DNaseI-seq data are available through the UCSC browser, and through the NCBI Gene Expression Omnibus (GEO) data repository under accessions GSE29692, GSE32970. H3K4me3 and 5C data are available through the UCSC browser, and through the NCBI Gene Expression Omnibus (GEO) data repository. Gene expression data are available through the UCSC browser, and through the NCBI Gene Expression Omnibus (GEO) data repository under accessions GSE19090, GSE15805, GSE17778.

[10]Dept. of Medicine, Division of Genetics, Brigham & Women's Hospital and Harvard Medical School, Boston, MA

[11]Department of Medicine, Division of Oncology, University of Washington, Seattle, WA

Abstract

DNaseI hypersensitive sites (DHSs) are markers of regulatory DNA and have underpinned the discovery of all classes of *cis*-regulatory elements including enhancers, promoters, insulators, silencers, and locus control regions. Here we present the first extensive map of human DHSs identified through genome-wide profiling in 125 diverse cell and tissue types. We identify ~2.9 million DHSs that encompass virtually all known experimentally-validated *cis*-regulatory sequences and expose a vast trove of novel elements, most with highly cell-selective regulation. Annotating these elements using ENCODE data reveals novel relationships between chromatin accessibility, transcription, DNA methylation, and regulatory factor occupancy patterns. We connect ~580,000 distal DHSs with their target promoters, revealing systematic pairing of different classes of distal DHSs and specific promoter types. Patterning of chromatin accessibility at many regulatory regions is choreographed with dozens to hundreds of co-activated elements, and the trans-cellular DNaseI sensitivity pattern at a given region can predict cell type-specific functional behaviors. The DHS landscape shows signatures of recent functional evolutionary constraint. However, the DHS compartment in pluripotent and immortalized cells exhibits higher mutation rates than that in highly differentiated cells, exposing an unexpected link between chromatin accessibility, proliferative potential and patterns of human variation.

INTRODUCTION

Cell-selective activation of regulatory DNA drives the gene expression patterns that shape cell identity. Regulatory DNA is characterized by the cooperative binding of sequence-specific transcriptional regulatory factors in place of a canonical nucleosome, leading to a remodeled chromatin state characterized by markedly heightened accessibility to nucleases[1]. DNaseI hypersensitive sites (DHSs) in chromatin were first identified over 30 years ago, and have since been extensively leveraged to map regulatory DNA regions in diverse organisms[2]. DNaseI hypersensitivity is the *sine qua non* of all defined classes of active *cis*-regulatory elements including enhancers, promoters, silencers, insulators, and locus control regions [2-4]. Because DNaseI hypersensitivity overlies *cis*-regulatory elements directly and is maximal over the core region of regulatory factor occupancy, it enables precise delineation of the genomic *cis*-regulatory compartment. DHSs are flanked by nucleosomes, which may acquire histone modification patterns that reflect the functional role of the adjoining regulatory DNA, such as the association of histone H3 lysine 4 trimethylation (H3K4me3) with promoter elements[5]. Recent advances have enabled genome-scale mapping of DHSs in mammalian cells[6, 7], laying the foundations for comprehensive catalogues of human regulatory DNA.

General features of the accessible chromatin landscape

Two ENCODE production centers (University of Washington and Duke University) profiled DNaseI sensitivity genome-wide using massively parallel sequencing[7-9] in a total of 125 human cell and tissue types including normal differentiated primary cells (n=71), immortalized primary cells (n=16), malignancy-derived cell lines (n=30) and multipotent and pluripotent progenitor cells (n=8) (Supplementary Table 1). The density of mapped DNaseI cleavages as a function of genome position provides a continuous quantitative measure of chromatin accessibility, in which DNaseI hypersensitive sites (DHSs) appear as prominent peaks within the signal data from each cell type (Fig. 1a and Supplementary Figs. 1,2). Analysis using a common algorithm (see Methods) identified 2,890,742 distinct high-

confidence DHSs (false discovery rate of 1%; see Methods), each of which was active in one or more cell types. Of these DHSs, 970,100 were specific to a single cell type, 1,920,642 were active in 2 or more cell types, and a small minority (3,692) was detected in all cell types. The relative accessibility of DHSs along the genome varies by >100-fold and is highly consistent across cell types (Supplementary Figs. 1, 2). To estimate the sensitivity and accuracy of the sequencing-derived DHS maps, one ENCODE production center (UW) performed 7,478 classical DNaseI hypersensitivity experiments by the Southern hybridization method[2]. Using Southerns as the standard, the average sensitivity, per cell type, of DNaseI-seq (at a sequencing depth of 30M uniquely mapping reads) was 81.6%, with specificity of 99.5-99.9%. Of DHSs classified as false negatives within a particular cell type, an average of 92.4% were detected as a DHS in another cell type or upon deeper sequencing. As such, we estimate that the overall sensitivity for DHSs of the combined cell type maps is >98%.

Approximately 3% (n=75,575) of DHSs localize to transcriptional start sites (TSSs) defined by Gencode[10] and 5% (n=135,735, including the aforementioned) lie within 2.5 kb of a TSS. The remaining 95% of DHSs are positioned more distally, and are roughly evenly divided between intronic and intergenic regions (Fig. 1b). Promoters typically exhibit high accessibility across cell types, with the average promoter DHS detected in 29 cell types (Fig 1c, second column). By contrast, distal DHSs are largely cell selective (Fig. 1c, third column).

MicroRNAs comprise a major class of regulatory molecules and have been extensively studied, resulting in consensus annotation of hundreds of conserved microRNA genes[11], approximately one third of which are organized in polycistronic clusters[12]. However, most predicted promoters driving microRNA expression lack experimental evidence. Of 329 unique annotated miRNA TSSs (Supplementary Methods), 300 (91%) either coincided with or closely approximated (<500 bp) a DHS. Chromatin accessibility at microRNA promoters was highly promiscuous compared with Gencode TSSs (Fig. 1c, fourth column), and showed cell lineage organization, paralleling the known regulatory roles of well-annotated lineage-specific microRNAs (Supplementary Fig. 3).

The 20-50 bp read lengths from DNaseI-seq experiments enabled unique mapping to 86.9% of the genomic sequence, allowing us to interrogate a large fraction of transposon sequences. A surprising number contain highly regulated DHSs (Fig. 1c, fifth column and Supplementary Figs. 4, 5), compatible with cell-specific transcription of repetitive elements detected using ENCODE RNA sequencing data[13]. DHSs were most strongly enriched at LTR elements, which encode retroviral enhancer structures (Supplementary Table 2). Two such examples are shown in Supplementary Fig. 4, which also illustrates the strong cell-selectivity of chromatin accessibility seen for each major repeat class. We also documented numerous examples of transposeon DHSs that displayed enhancer activity in transient transfection assays (Supplementary Table 3).

Comparison with an extensive compilation of 1,046 experimentally validated distal, non-promoter *cis*-regulatory elements (enhancers, insulators, locus control regions, etc.) revealed the overwhelming majority (97.4%) to be encompassed within DNaseI hypersensitive chromatin (Supplementary Table 4), typically with strong cell selectivity (Supplementary Fig. 2b).

Transcription factor drivers of chromatin accessibility

DNaseI hypersensitive sites result from cooperative binding of transcriptional factors in place of a canonical nucleosome[1,2]. To quantify the relationship between chromatin accessibility and the occupancy of regulatory factors, we compared sequencing depth-

normalized DNaseI sensitivity in the ENCODE common cell line K562 to normalized ChIP-seq signals from all 42 transcription factors mapped by ENCODE ChIP-seq[14] in this cell type (Fig. 2). Simple summation of the ChIP-seq signals strikingly parallels quantitative DNaseI sensitivity at individual DHSs (Fig. 2a) and across the genome ($R = 0.79$, Fig. 2b). For example, the beta glob in locus control region contains a major enhancer element at hypersensitive site 2 (HS2), which appears to be occupied by dozens of TFs (Supplementary Fig. 6a). Such highly overlapping binding patterns have been interpreted to signify weak interactions with lower-affinity recognition sequences potentiated by an accessible DNA template[15]. However, HS2 is a compact element with a functional core spanning ~110bp that contains 5-8 sites of transcription factor-DNA interaction *in vivo* depending on the cell type[16-18]. The fact that the cumulative ChIP-seq signal closely parallels the degree of nuclease sensitivity at HS2 and elsewhere is thus most readily explained by interactions between DNA-bound factors and other interacting factors that collectively potentiate the accessible chromatin state (Supplementary Fig. 6b). Given the relatively limited number of factors studied, it may seem surprising that such a close correlation should be evident. However, most of the factors selected for ENCODE ChIP-seq studies have well-described or even fundamental roles in transcriptional regulation, and many were identified originally based on their high affinity for DNA. Alternatively, as originally proposed by Weintraub[19], a limited number of factors may be involved in establishment and maintenance of chromatin remodeling, while others may interact non-specifically with the remodeled state. 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), suggesting that regulators acting through these sequences are key drivers of the accessibility landscape.

Overall, 94.4% of a combined 1,108,081 ChIP-seq peaks from all ENCODE TFs fall within accessible chromatin (Fig. 2c and Supplementary Fig. 7a), with the median factor having 98.2% of its binding sites localized therein. Notably, a small number of factors diverged from this paradigm, including known chromatin repressors, such as the KRAB-associated factors KAP1, SETDB1 and ZNF274[20, 21] (Fig. 2c). We hypothesized that a proportion of the occupancy sites of these factors represented binding within compacted heterochromatin. To test this, we developed targeted mass spectrometry assays[22] for KAP1 and three factors localizing almost exclusively within accessible chromatin (GATA1, c-Jun, NRF1), and quantified their abundance in biochemically-defined heterochromatin[23] vs. a total chromatin fraction (Supplementary Fig. 7b). This analysis confirmed that factors such as KAP1 significantly occupy heterochromatin (Supplementary Fig. 7c).

An invariant directional chromatin signature at promoters

The annotation of sites of transcription origination continues to be an active and fundamental endeavor[15]. In addition to direct evidence of TSSs provided by RNA transcripts, H3K4me3 modifications are closely linked with TSSs[24]. We therefore explored systematically the relationship between chromatin accessibility and H3K4me3 patterns at well-annotated promoters, its relationship to transcription origination, and its variability across ENCODE cell types.

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

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

Chromatin accessibility and DNA methylation patterns

CpG methylation has been closely linked with gene regulation, based chiefly on its association with transcriptional silencing[25]. However, the relationship between DNA methylation and chromatin structure has not been clearly defined. We analyzed ENCODE reduced-representation bisulfite sequencing (RRBS) data, which provide quantitative methylation measurements for several million CpGs[26]. We focused on 243,037 CpGs falling within DHSs in 19 cell types for which both data types were available from the same sample. We observed two broad classes of sites: those with a strong inverse correlation across cell types between DNA methylation and chromatin accessibility (Fig. 4a, Supplementary Fig. 11a), and those with variable chromatin accessibility but constitutive hypomethylation (Fig. 4a, right). To quantify these trends globally, we performed a linear regression analysis between chromatin accessibility and DNA methylation at the 34,376 CpG-containing DHSs (see Supplementary Methods). Of these sites, 6,987 (20%) showed a significant association (1% FDR) between methylation and accessibility (Supplementary Fig. 11b). Increased methylation was almost uniformly negatively associated with chromatin accessibility (>97% of cases). The magnitude of the association between methylation and accessibility was strong, with the latter on average 95% lower in cell types with coinciding methylation vs. cell types lacking coinciding methylation (Supplementary Fig. 11c). Fully 40% of variable methylation was associated with a concomitant effect on accessibility.

The role of DNA methylation in causation of gene silencing is presently unclear. Does methylation reduce chromatin accessibility by evicting transcription factors? Or does DNA methylation passively 'fill in' the voids left by vacating TFs? Transcription factor expression is closely linked with the occupancy of its binding sites[27]. If the former of the two above hypotheses is correct, methylation of individual binding site sequences should be independent of TF gene expression. If the latter, methylation at TF recognition sequences should be inversely correlated with TF abundance (Fig. 4b).

Comparing TF transcript levels to average methylation at cognate recognition sites within DHSs revealed significant negative correlations between TF expression and binding site methylation for the majority (70%) of TFs with a significant association ($P < 0.05$). Representative examples are shown in Fig. 4c and Supplementary Fig. 12a. These data argue strongly that methylation patterning paralleling cell-selective chromatin accessibility results from passive deposition following the vacation of TFs from regulatory DNA, generalizing other recent reports [28]

Interestingly, a small number of factors showed positive correlations between expression and binding site methylation (Supplementary Fig. 12b), including MYB and LUN1. Both of

these TFs showed increased transcription and binding site methylation specifically within acute promyelocytic leukemia cells (NB4), and both interact with PML bodies[29, 30], a sub-nuclear structure disrupted in promyelocytic leukemia cells. The anomalous behavior of these two TFs with respect to chromatin structure and DNA methylation may thus be related to a specialized mechanism seen only in pathologically altered cells.

A genome-wide map of distal DHS-to-promoter connectivity

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 Figure 13). To generalize this, we analyzed the patterning of 1,454,901 distal (>2.5kb from TSS) DHSs 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 ±500kb (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, 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 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 3- to 10-fold over a promoter-only control, confirming enhancer function.

We next examined comprehensive promoter-vs-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 ($P < 10^{-13}$, Supplementary Fig. 14b). We also examined K562 promoter-DHS interactions detected by Pol II ChIA-PET [24], which quantify interactions between promoter-bound polymerase and distal sites. The ChIA-PET interactions were also markedly enriched for DHS-promoter pairings ($P < 10^{-15}$, 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, suggesting 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; 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 found that the most complexly regulated human genes were strikingly 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, 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 distal DHSs correlated with promoter DHSs containing the NANOG motif; enrichment of NANOG, SOX2, and OCT4 distal motifs co-occurring with promoter OCT4; and enrichment of distal SOX2 and OCT4 motifs with promoter SOX2 (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).

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, SPIB, and others) is present within a promoter DHS, motif PU.1 is significantly more likely to be observed in a correlated distal ($P < 10^{-5}$). These results suggest that a limited set of general rules may govern the pairing of co-regulated distal DHSs with particular promoters.

Stereotyped chromatin accessibility parallels function

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. 16,17). 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 same pattern of cell-selective chromatin accessibility shared between two DHSs can be achieved by distinct mechanisms, likely involving complex combinatorial tuning.

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 hypersensitive site 2 (HS2) of the beta-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 pattern, and tested these in transient transfection assays in K562 cells (Supplementary Methods). Seventy percent (14/20) 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 vs. control) than HS2, which is itself one of the most potent known enhancers[4]. Next we selected 3 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 15/25 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 suggests more generally that DHSs with stereotyped cellular patterning are likely to fulfill similar functions.

To visualize the qualities and prevalence of different stereotyped cross-cellular DHS patterns, we constructed a self-organizing map (SOM) 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, 19). Many of the stereotyped patterns discovered by the SOM encompass large numbers of DHSs, with some counting >1000 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 are likely to share similar functions.

Genetic variation in regulatory DNA linked to mutation rate

The DHS compartment as a whole is under evolutionary constraint, which varies between different classes and locations of elements[35], and may be heterogeneous within individual elements[36]. To understand the evolutionary forces shaping regulatory DNA sequences in humans, we estimated nucleotide diversity (π) in DHSs using publicly available whole-genome sequencing data from 53 unrelated individuals[37] (see Supplementary Methods). We restricted our analysis to nucleotides outside of exons and RepeatMasked regions. To provide a comparison with putatively neutral sites, we computed π in four-fold degenerate synonymous positions (third positions) of coding exons. This analysis showed that, taken together, DHSs exhibit lower π than four-fold degenerate sites, compatible with the action of purifying selection.

Fig. 7a shows π for the DHSs of all analyzed cell types, with color coding to indicate the origin of each cell type. Particularly striking is the distribution of diversity relative to proliferative potential. DHSs in cells with limited proliferative potential have uniformly lower average diversity than immortal cells, with the difference most pronounced in malignant and pluripotent lines. This ordering is identical when highly mutable CpG nucleotides are removed from the analysis.

If differences in π are due to mutation rate differences in different DHS compartments, the ratio of human polymorphism to human-chimpanzee divergence should remain constant across cell types. By contrast, differences in π due to selective constraint should result in pronounced differences. To distinguish between these alternatives, we first compared polymorphism and human-chimp divergence for DHSs from normal, malignant, and pluripotent cells (Fig. 7b). Differences in polymorphism and divergence between these three groups are nearly identical, compatible with a mutational cause. Second, raw mutation rate is expected to affect rare and common genetic variation equally, whereas selection is likely to have a larger impact on common variation. We consistently observe ~62% of SNPs in DHSs of each group to have derived-allele frequencies below 0.05. DHSs in different cell lines exhibit differences in SNP densities but not in allele frequency distribution (Fig. 7c). Collectively, these observations are consistent with increased relative mutation rates in the DHS compartment of immortal cells vs. cell types with limited proliferative potential, exposing an unexpected link between chromatin accessibility, proliferative potential, and patterns of human variation.

DISCUSSION

Since their discovery over 30 years ago, DNaseI hypersensitive sites have guided the discovery of diverse *cis*-regulatory elements in the human and other genomes. Here we have presented by far the most comprehensive map of human regulatory DNA, revealing novel relationships between chromatin accessibility, transcription, DNA methylation, and the

occupancy of sequence-specific factors. The wide spectrum of different cell and tissue types covered by our data greatly expands the horizons of cell-selective gene regulation analysis, enabling the recognition of systematic long-distance regulatory patterns, and previously undescribed phenomena such as stereotyping of DHS activation and mutation rate variation in normal vs. immortal cells. The extensive resources we have provided should greatly facilitate future analyses, and stimulate new areas of investigation into the organization and control of the human genome.

METHODS SUMMARY

DNaseI hypersensitivity mapping was performed using protocols developed by Duke[7] or UW[8] on a total of 125 cell-types (Supplementary Table 1). Datasets were sequenced to an average depth of 30 million uniquely mapping sequence tags (27-35 bp for UW and 20 bp for Duke) per replicate. For uniformity of analysis, some cell type data sets that exceeded 40M tag depth were randomly sub sampled to a depth of 30 million tags. Sequence reads were mapped using the Bowtie aligner, allowing a maximum of two mismatches. Only reads mapping uniquely to the genome were used in our analyses. Mappings were to male or female versions of hg19/GRCh37, depending on cell type, with random regions omitted. Data were analyzed jointly using a single algorithm[7] (Supplementary Methods) to localize DNaseI hypersensitive sites. H3K4me3 ChIP-seq was performed using antibody 9751 (Cell Signaling) on 1% formaldehyde crosslinked samples sheared by Diagenode bioruptor. Gene expression measurements for each cell type were performed on Affymetrix Human Exon microarrays. 5C experiments were performed as described[31, 32]. Transcription factor recognition motif occurrences within DHSs were defined with FIMO[38] at significance $P < 10^{-5}$ using motif models from the TRANSFAC database.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank our ENCODE colleagues for many helpful insights into the data types generated by different centers and for help with coordinated analyses. We thank Ian Stanaway for assistance with the variation analysis, and many colleagues, particularly Fyodor Urnov and Sam John, for their helpful critiques of the manuscript and figures. This work was funded by NIH grants HG004592 (J.A.S.) and HG004563 (G.E.C.). J.R.V. is supported by a National Science Foundation Graduate Research Fellowship. N.C.S. is supported by a National Science Foundation Graduate Research Fellowship and the Research Council of Norway.

References

1. Felsenfeld G, Boyes J, Chung J, Clark D, Studitsky V. Chromatin structure and gene expression. Proc Natl Acad Sci U S A. 1996; 93:9384–8. [PubMed: 8790338]
2. Gross DS, Garrard WT. Nuclease hypersensitive sites in chromatin. Annu Rev Biochem. 1988; 57:159–97. [PubMed: 3052270]
3. Gaszner M, Felsenfeld G. Insulators: exploiting transcriptional and epigenetic mechanisms. Nat Rev Genet. 2006; 7:703–13. [PubMed: 16909129]
4. Li Q, Harju S, Peterson KR. Locus control regions: coming of age at a decade plus. Trends Genet. 1999; 15:403–8. [PubMed: 10498936]
5. Heintzman ND, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007; 39:311–318. [PubMed: 17277777]
6. Hesselberth JR, et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Meth. 2009; 6:283–289.
7. Boyle AP, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008; 132:311–22. [PubMed: 18243105]

8. John S, et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nature Genetics. 2011; 43:264–268. [PubMed: 21258342]
9. Song L, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 2010; 21:1757–67. [PubMed: 21750106]
10. Harrow J, Frankish A, Gonsalez JM, et al. GENCODE: The reference human genome annotation for the ENCODE project. Genome Res. 2012 In Press.
11. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. Nucleic Acids Res. 2008; 36:D154–8. [PubMed: 17991681]
12. Farazi TA, Spitzer JI, Morozov P, Tuschl T. miRNAs in human cancer. J Pathol. 2011; 223:102–15. [PubMed: 21125669]
13. Djebali S, Davis CA, LaGarde J, et al. Landscape of transcription in human cell lines. Nature. 2012 In Press.
14. Wang J, et al. Genome-wide mapping of the binding sites of 119 transcription factors. Nature. 2012 In Press.
15. Biggin MD. Animal transcription networks as highly connected, quantitative continua. Dev Cell. 2011; 21:611–26. [PubMed: 22014521]
16. Reddy PM, Stamatoyannopoulos G, Papayannopoulou T, Shen CK. Genomic footprinting and sequencing of human beta-globin locus Tissue specificity and cell line artifact. J Biol Chem. 1994; 269:8287–95. [PubMed: 8132552]
17. Forsberg EC, Downs KM, Bresnick EH. Direct interaction of NF-E2 with hypersensitive site 2 of the beta-globin locus control region in living cells. Blood. 2000; 96:334–9. [PubMed: 10891470]
18. Talbot D, Grosveld F. The 5'HS2 of the globin locus control region enhances transcription through the interaction of a multimeric complex binding at two functionally distinct NF-E2 binding sites. Embo J. 1991; 10:1391–8. [PubMed: 1902783]
19. Weisbrod S, Weintraub H. Isolation of a subclass of nuclear proteins responsible for conferring a DNase I-sensitive structure on globin chromatin. Proceedings of the National Academy of Sciences of the United States of America. 1979; 76:630–634. [PubMed: 284387]
20. Schultz DC, Ayyanathan K, Negorev D, Maul GG, Rauscher FJ. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. Genes & development. 2002; 16:919–32. [PubMed: 11959841]
21. Frietze S, O'Geen H, Blahnik KR, Jin VX, Farnham PJ. ZNF274 recruits the histone methyltransferase SETDB1 to the 3' ends of ZNF genes. PloS One. 2010; 5:e15082–e15082. [PubMed: 21170338]
22. Stergachis AB, Maclean B, Lee K, Stamatoyannopoulos JA, Maccoss MJ. Rapid empirical discovery of optimal peptides for targeted proteomics. Nat Methods. 2011; 8:1041–3. [PubMed: 22056677]
23. Henikoff S, Henikoff JG, Sakai A, Loeb GB, Ahmad K. Genome-wide profiling of salt fractions maps physical properties of chromatin. Genome Research. 2009; 19:460–469. [PubMed: 19088306]
24. Li G, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 2012; 148:84–98. [PubMed: 22265404]
25. Siegfried Z, et al. DNA methylation represses transcription in vivo. Nat Genet. 1999; 22:203–6. [PubMed: 10369268]
26. Varley KE, Gertz J, Bowling KM, et al. Dynamic DNA methyation across diverse human cell lines and tissues. Nature. 2012 In Press.
27. O'Geen H, et al. Genome-wide analysis of KAP1 binding suggests autoregulation of KRAB-ZNFs. PLoS Genet. 2007; 3:e89. [PubMed: 17542650]
28. Bell O, Tiwari VK, Thoma NH, Schubeler D. Determinants and dynamics of genome accessibility. Nat Rev Genet. 2012; 12:554–64. [PubMed: 21747402]
29. Rasheed ZA, Saleem A, Ravee Y, Pandolfi PP, Rubin EH. The topoisomerase I-binding RING protein, topors, is associated with promyelocytic leukemia nuclear bodies. Exp Cell Res. 2002; 277:152–60. [PubMed: 12083797]

30. Dahle O, Bakke O, Gabrielsen OS. c-Myb associates with PML in nuclear bodies in hematopoietic cells. Exp Cell Res. 2004; 297:118–26. [PubMed: 15194430]

31. Dostie J, et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. 2006; 16:1299–309. [PubMed: 16954542]

32. Sanyal A, Lajoie B, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012 In Press.

33. Kim J, Chu J, Shen X, Wang J, Orkin SH. An extended transcriptional network for pluripotency of embryonic stem cells. Cell. 2008; 132:1049–61. [PubMed: 18358816]

34. Tuan D, Kong S, Hu K. Transcription of the hypersensitive site HS2 enhancer in erythroid cells. Proc Natl Acad Sci U S A. 1992; 89:11219–23. [PubMed: 1454801]

35. The_ENCODE_Consortium. Integrative Analysis of the Human Genome. Nature. 2012 In Press.

36. Neph S, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012 In Press.

37. Vernot B, et al. Personal and population genomics of human regulatory variation. Genome Res. 2012 In Press.

38. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011; 27:1017–8. [PubMed: 21330290]

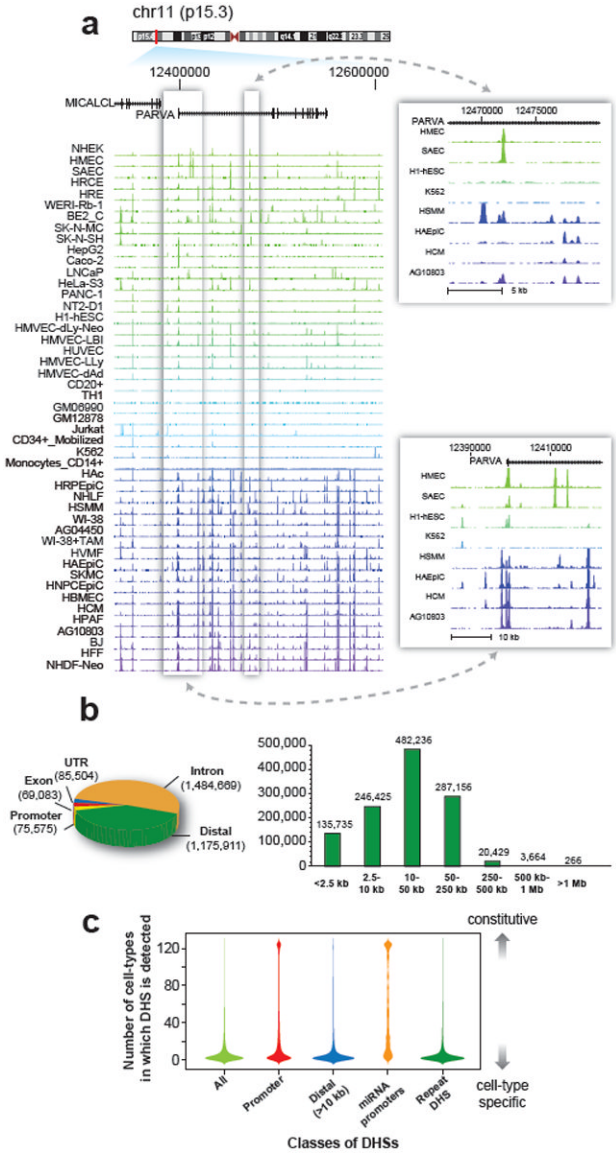

Figure 1. General features of the DHS landscape

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

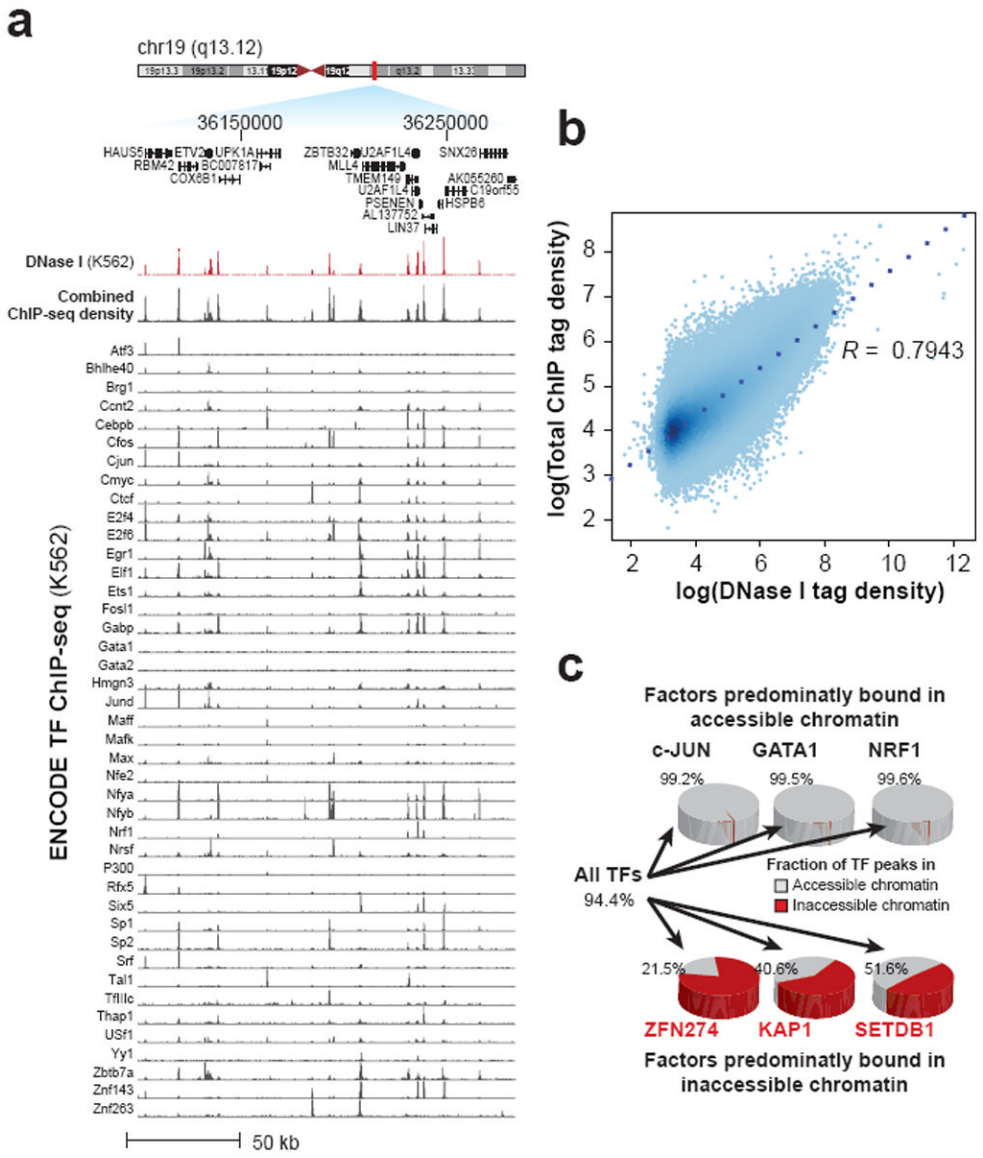

Figure 2. Transcription factor drivers of chromatin accessibility

a, DNaseI tag density is shown in red for a 175 kb region of Chr19. Below, normalized ChIP-seq tag density for 45 ENCODE ChIP-seq experiments from K562 cells, with a cumulative sum of the individual tag density tracks shown immediately below the K562 DNaseI data. **b**, Genome-wide correlation ($R = 0.7943$) between ChIP-seq and DNaseI tag densities (log10) in K562 cells. **c**, Left, 94.4% of a combined 1,108,081 ChIP-seq peaks from all TFs assayed in K562 cells fall within accessible chromatin (grey pie areas). Top, three examples of TFs localizing almost exclusively within accessible chromatin. Bottom, three factors from the KRAB-associated complex localizing partially or predominantly within inaccessible chromatin

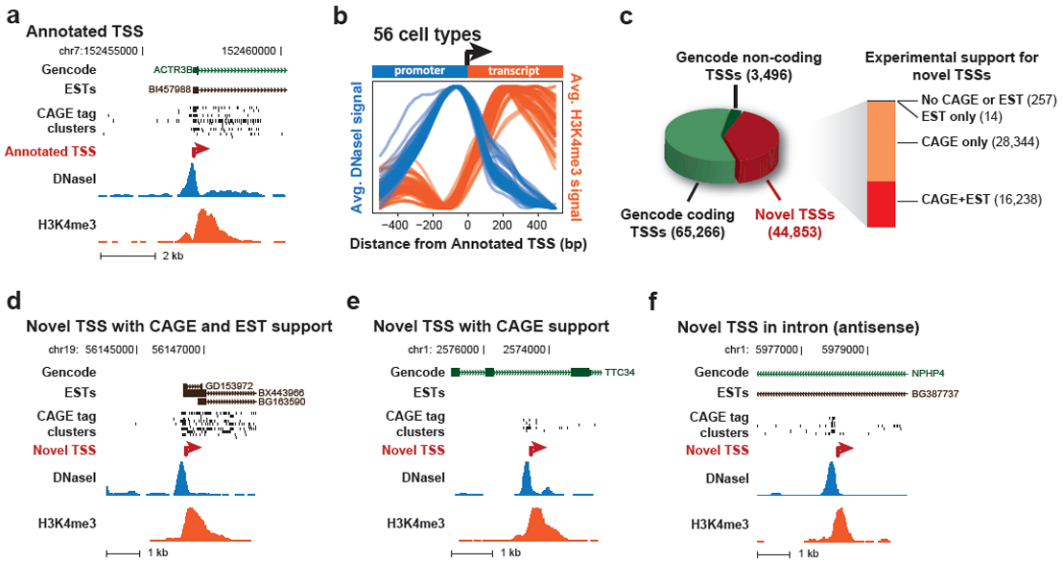

Figure 3. Identification and directional classification of novel promoters

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

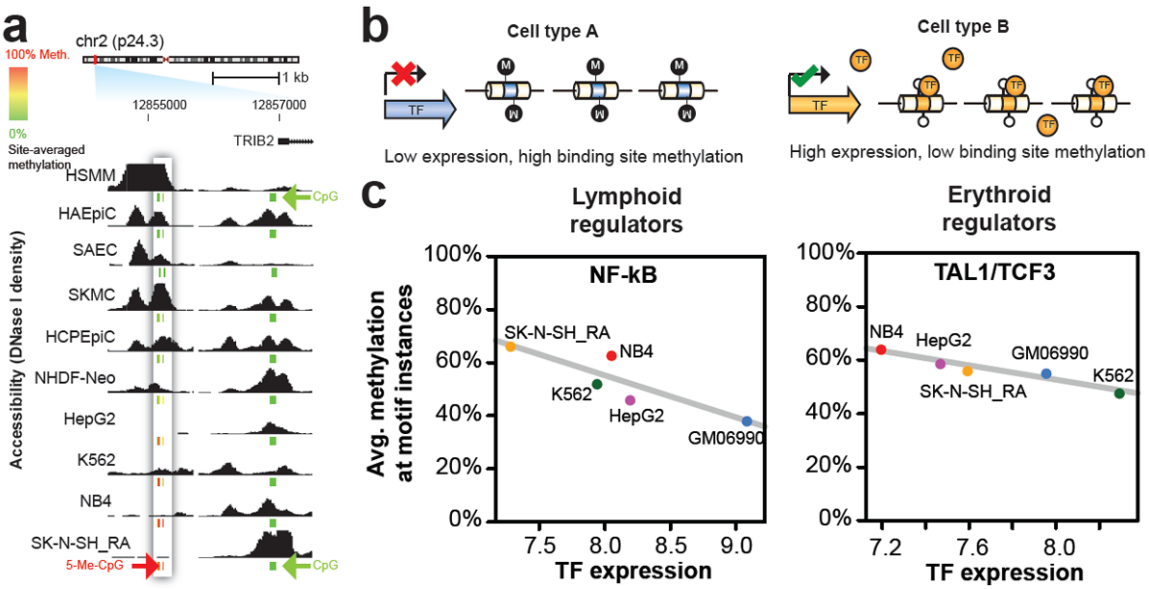

Figure 4. Chromatin accessibility and DNA methylation patterns

a, DNaseI sensitivity in 19 cell types with ENCODE Reduced Representation Bisulfite Sequencing data. Inset box: accessibility (y-axis) decreases quantitatively as methylation increases. Other DHSs (right) show low correlation between accessibility and methylation. CpG methylation scale: Green, 0%; yellow, 50%; red, 100%. **b**, Model of TF-driven methylation patterns in which methylation passively mirrors TF occupancy. **c**, Relationship between TF transcript levels and overall methylation at cognate recognition sequences of the same TFs. Lymphoid regulators in B-lymphoblastoid line GM06990 (left) and erythroid regulators in the erythroleukemia line K562 (right). Negative correlation indicates that site-specific DNA methylation follows TF vacation of differentially expressed TFs.

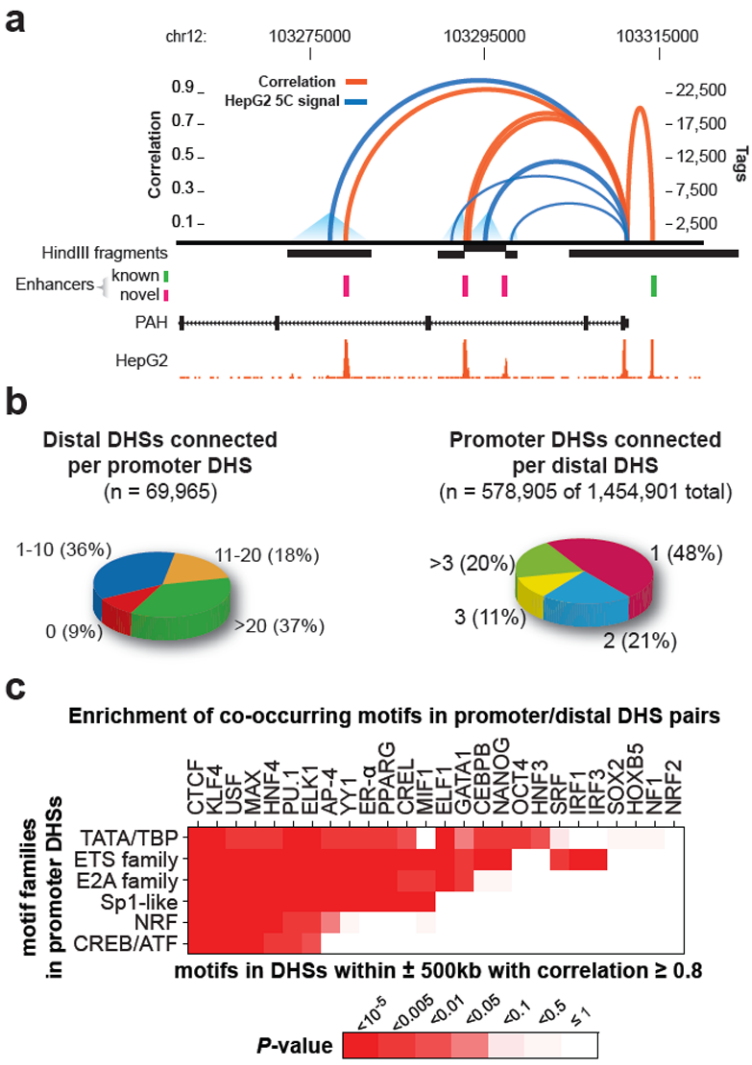

Figure 5. A genome-wide map of distal DHS-to-promoter connectivity

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 since 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 families with specific motifs in distal DHSs.

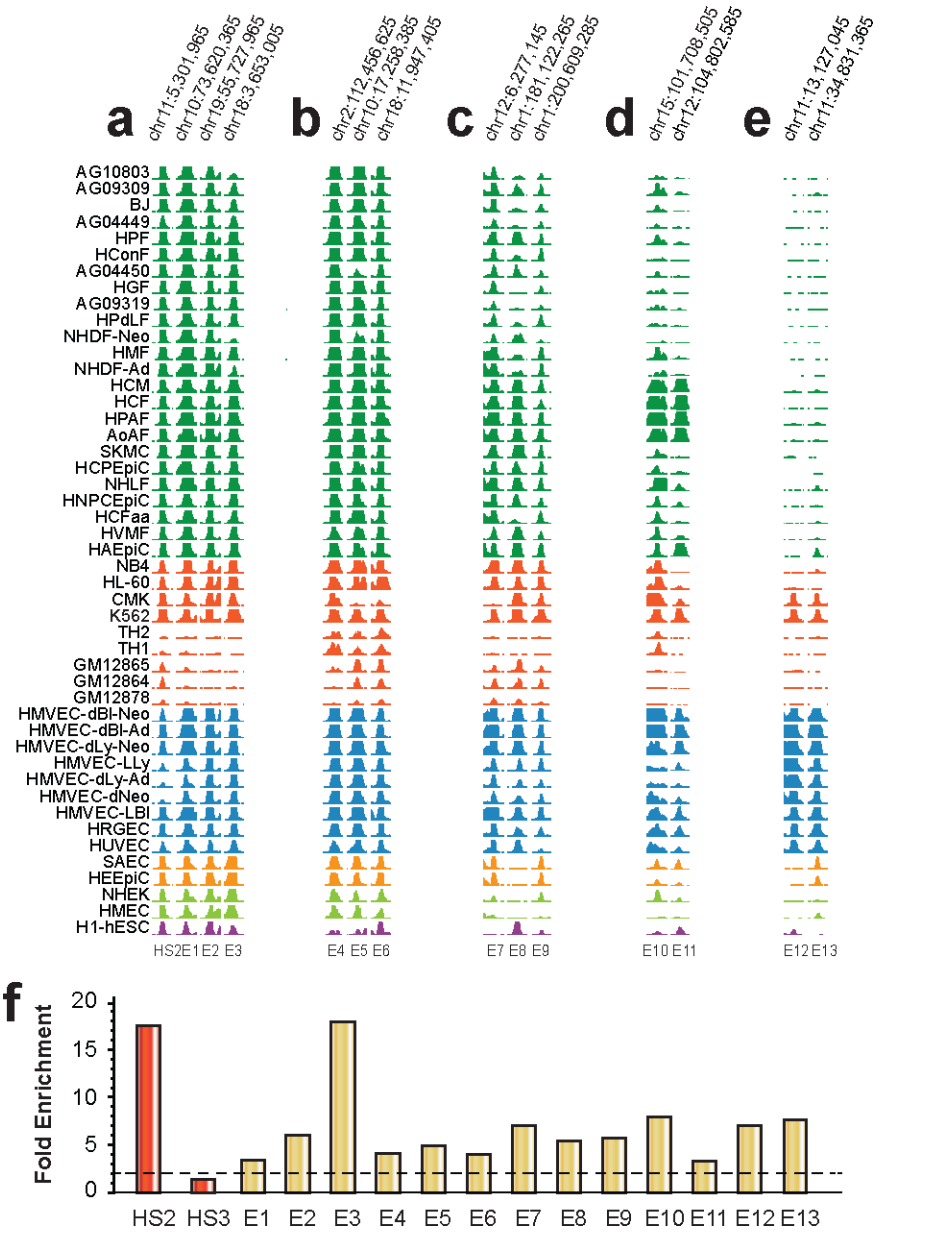

Figure 6. Stereotyped regulation of chromatin accessibility

(a)-(e), Enhancers grouped by similar chromatin stereotypes. HS2 from the beta-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 = data from two potent in vivo enhancers, beta-globin LCR HS2 and HS3; the latter requires chromatinization to function and is not active in transient assays. Gold bars = data from E1-E13 from (a)-(e) above.

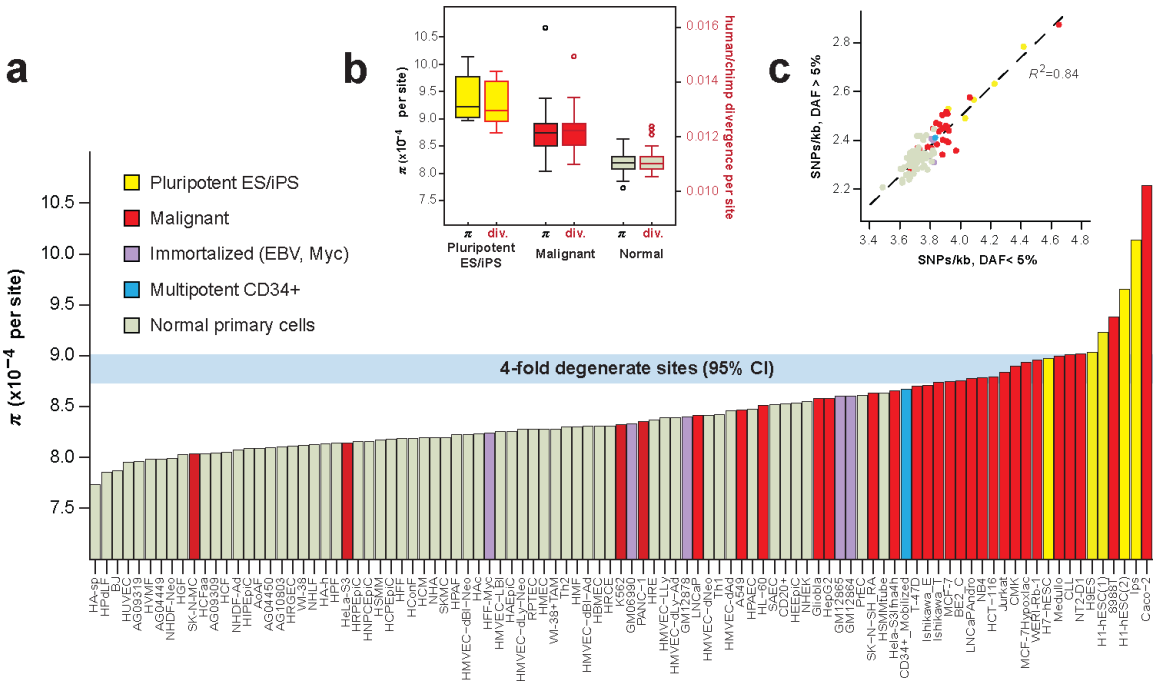

Figure 7. Genetic variation in regulatory DNA linked to mutation rate

a, Mean nucleotide diversity (π, y-axis) in DHSs of 97 diverse cell types (x-axis) estimated using whole-genome sequencing data from 53 unrelated individuals. Cell types are ordered left-to-right by increasing mean π. Horizontal blue bar shows 95% confidence intervals on mean π in a background model of four-fold degenerate coding sites. Note the enrichment of immortal cells at right. **b**, Mean π (left y-axis) for pluripotent (yellow) vs. malignancy-derived (red) vs. normal cells (light green), plotted side-by-side with human-chimp divergence (right y-axis) computed on the same groups. Boxes indicate 25-75%-iles, with medians highlighted. **c**, Both low- and high-frequency derived alleles show the same effect. Density of SNPs in DHSs with derived allele frequency (DAF) <5% (x-axis) is tightly correlated (R^2 = 0.84) with the same measure computed for higher-frequency derived alleles (y-axis). Color-coding is same as in panel (a).