Title

Mapping and analysis of chromatin state dynamics in nine human cell types

Permalink

https://escholarship.org/uc/item/47r1n67b

Journal

Nature, 473(7345)

ISSN

0028-0836

Authors

Ernst, Jason
Kheradpour, Pouya
Mikkelsen, Tarjei S
et al.

Publication Date

2011-05-01

DOI

10.1038/nature09906

Published in final edited form as:
Nature. 2011 May 5; 473(7345): 43–49. doi:10.1038/nature09906.

Systematic analysis of chromatin state dynamics in nine human cell types

Jason Ernst[1,2], **Pouya Kheradpour**[1,2], **Tarjei S. Mikkelsen**[1], **Noam Shoresh**[1], **Lucas D. Ward**[1,2], **Charles B. Epstein**[1], **Xiaolan Zhang**[1], **Li Wang**[1], **Robbyn Issner**[1], **Michael Coyne**[1], **Manching Ku**[1,3,4], **Timothy Durham**[1], **Manolis Kellis**[1,2,*], and **Bradley E. Bernstein**[1,3,4]

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[2]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts, USA

[3]Howard Hughes Medical Institute, Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA

[4]Center for Systems Biology and Center for Cancer Research, Massachusetts General Hospital, Boston, Massachusetts, USA

Abstract

Chromatin profiling has emerged as a powerful means for genome annotation and detection of regulatory activity. Here we map nine chromatin marks across nine cell types to systematically characterize regulatory elements, their cell type-specificities, and their functional interactions. Focusing on cell type-specific patterns of promoters and enhancers, we define multi-cell activity profiles for chromatin state, gene expression, regulatory motif enrichment, and regulator expression. We use correlations between these profiles to link enhancers to putative target genes, and predict the cell type-specific activators and repressors that modulate them. The resulting annotations and regulatory predictions have implications for interpreting genome-wide association studies. Top-scoring disease SNPs are frequently positioned within enhancer elements specifically active in relevant cell types, and in some cases affect a motif instance for a predicted regulator, thus proposing a mechanism for the association. Our study presents a general framework for deciphering cis-regulatory connections and their roles in disease.

[*]Correspondence should be addressed to M.K. (manoli@mit.edu). .

Author contributions

JE conducted chromatin state analysis. JE and PK conducted regulatory motif analysis. JE and LW conducted GWAS SNP analysis. TM, NS and TD implemented the ChIP-seq data processing pipeline. CE, XZ, LW, RI, MC and MK(1) developed the experimental pipeline and conducted experiments. MK(2) envisioned and directed the computational analysis. BB envisioned the experimental approach and oversaw the work. JE, MK and BB and wrote the paper.

Author information

Sequencing and expression data has been deposited into the Gene Expression Omnibus under accession GSE26386. Reprints and permissions information is available at www.nature.com/reprints.

The authors declare no competing financial interests.

Introduction

A major challenge in biology is to understand how a single genome can give rise to an organism comprising hundreds of distinct cell types. Much emphasis has been placed on the application of high-throughput tools to study interacting cellular components[1]. The field of systems biology has exploited dynamic gene expression patterns to reveal functional modules, pathways and networks[2]. Yet cis-regulatory elements, which may be equally dynamic, remain largely uncharted across cellular conditions.

Chromatin profiling provides a systematic means for detecting cis-regulatory elements, given the central role of chromatin in mediating regulatory signals and controlling DNA access, and the paucity of recognizable sequence signals. Specific histone modifications correlate with regulator binding, transcriptional initiation and elongation, enhancer activity and repression[1],[3]-[6]. Combinations of modifications can provide even more precise insight into chromatin state[7],[8].

Here, we apply a high-throughput pipeline to map 9 chromatin marks and input controls across 9 cell types. We use recurrent combinations of marks to define 15 chromatin states corresponding to repressed, poised, and active promoters, strong and weak enhancers, putative insulators, transcribed regions, and large-scale repressed and inactive domains. We use directed experiments to validate biochemical and functional distinctions between states.

The resulting chromatin state maps portray a highly dynamic landscape, with the specific patterns of change across cell types revealing strong correlations between interacting functional elements. We use correlated patterns of activity between chromatin state, gene expression and regulator activity to connect enhancers to likely target genes, to predict cell type-specific activators and repressors, and to identify individual binding motifs responsible for these interactions.

Our results have implications for interpreting genome-wide association studies. We find that disease variants frequently coincide with enhancer elements specific to a relevant cell type. In several cases, we can predict upstream regulators whose regulatory motif instances are affected or target genes whose expression may be altered, thereby proposing specific mechanistic hypotheses for how disease-associated genotypes lead to the observed disease phenotypes.

Results

Systematic mapping of chromatin marks in multiple cell types

To explore chromatin state in a uniform way across multiple cell types, we applied a production pipeline for chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) to generate genome-wide chromatin datasets (see **Methods**, Fig. 1a). We profiled nine human cell types, including common lines designated by the ENCODE consortium[1] and primary cell types. These consist of embryonic stem cells (H1 ES), erythrocytic leukemia cells (K562), B-lymphoblastoid cells (GM12878), hepatocellular carcinoma cells (HepG2), umbilical vein endothelial cells (HUVEC), skeletal muscle myoblasts (HSMM), normal lung fibroblasts (NHLF), normal epidermal keratinocytes (NHEK), and mammary epithelial cells (HMEC).

We used antibodies for histone H3 lysine 4 tri-methylation (H3K4me3), a modification associated with promoters[4,5,9]; H3K4me2, associated with promoters and enhancers[1,3,6,9]; H3K4me1, preferentially associated with enhancers[1,6]; lysine 9 acetylation (H3K9ac) and H3K27ac, associated with active regulatory regions[9,10]; H3K36me3 and H4K20me1, associated with transcribed regions[3-5]; H3K27me3, associated with Polycomb-repressed regions[3,4]; and CTCF, a sequence-specific insulator protein with diverse functions[11]. We validated each antibody by Western blots and peptide competitions, and sequenced input controls for each cell type. We also collected data for H3K9me3, RNAPII, and H2A.Z in a subset of cells.

This resulted in 90 chromatin maps corresponding to ~2.4 billion reads covering ~100 billion bases across nine cell types, which we set out to interpret computationally.

Learning a common set of chromatin states across cell types

To summarize these datasets into nine readily interpretable annotations, one per cell type, we applied a multivariate Hidden Markov Model (HMM) that uses combinatorial patterns of chromatin marks to distinguish chromatin states[8]. The approach explicitly models mark combinations in a set of ‘emission’ parameters and spatial relationships between neighboring genomic segments in a set of ‘transition’ parameters (see **Methods**). It has the advantage of capturing regulatory elements with greater reliability, robustness and precision relative to studying individual marks[8].

We learned chromatin states jointly by creating a virtual concatenation of all chromosomes from all cell types. We selected 15 states which showed distinct biological enrichments and were consistently recovered (Fig. 1a,b; Supplementary Fig. 1). Even though states were learned *de novo* based solely on the patterns of chromatin marks and their spatial relationships, they showed distinct associations with transcriptional start sites (TSSs), transcripts, evolutionarily-conserved non-coding regions, DNase hypersensitive sites[12], binding sites for the regulators, c-Myc[13] and NF-κB[14], and inactive genomic regions associated with the nuclear lamina[15] (Fig. 1c).

We distinguished six broad classes of chromatin states, which we refer to as promoter, enhancer, insulator, transcribed, repressed, and inactive states (Fig. 1c). Within them, active, weak and poised[4] promoters (states 1-3) differ in expression levels, strong and weak candidate enhancers (states 4-7) differ in expression of proximal genes, and strongly and weakly transcribed regions (states 9-11) also differ in their positional enrichments along transcripts. Similarly, Polycomb-repressed regions (state 12) differ from heterochromatic and repetitive states (states 13-15), which are also enriched for H3K9me3 (Supplementary Fig. 2-4).

The states vary widely in their average segment length (~500bp for promoter and enhancer states vs. 10 kb for inactive regions), and in the portion of the genome covered (<1% for promoter and enhancer states vs. >70% for inactive state 13). For each state, coverage was relatively stable across cell types (Supplementary Fig. 5), with the exception of ES cells in which the poised promoter state is more abundant while strong enhancer and Polycomb-repressed states are depleted, consistent with the unique biology of pluripotent cells[4],[16].

We confirmed that promoter and enhancer states showed distinct biochemical properties (Supplementary Fig. 6). RNAPII was highly enriched at strong promoters, weakly enriched at strong enhancers, and nearly undetectable at weak/poised enhancers, consistent with strong transcription at promoters, and reports of weak transcription at active enhancers[17],[18]. H2A.Z, a histone variant associated with nucleosome free regions[19], was enriched in active promoters and strong enhancers, consistent with nucleosome displacement at TSSs and sites of abundant transcription factor (TF) binding in active enhancers.

We also used luciferase reporter assays to validate the functionality of predicted enhancers, the distinction between strong and weak enhancer states, and their predicted cell type-specificity. We tested strong enhancers, weak enhancers, and strong enhancers specific to an unmatched cell type by transfection in HepG2 cells. We observed strong luciferase activity only for strong enhancer elements from the matched cell type (Fig. 1d).

These results and additional properties of the model (Supplementary Fig. 7-10) suggest that chromatin states are an inherent, biologically-informative feature of the genome. The framework enables us to reason about coordinated differences in marks by directly studying chromatin state changes between cell types (which we refer to as ‘changes’ or ‘dynamics’ without implying any temporal relationship).

Extent and significance of chromatin state changes across cell types

We next explored the extent to which chromatin states vary between pairs of cell types. The overall patterns of variability (Supplementary Fig. 11,12) suggest that regulatory regions vary dramatically in activity levels across cell types. Enhancer states show frequent interchange between strong and weak enhancers, and promoter states vary between active, weak and poised. Promoter states appear more stable than enhancers; they are eight times more likely to remain promoter states, controlling for coverage. Switching was also observed between promoter, enhancer, and transcriptional transition states, but no preferential changes were found to other groups. These general patterns suggest that despite varying activity levels, enhancer and promoter regions tend to preserve their chromatin identity as regions of regulatory potential.

Chromatin state differences between cell types relate to cell type-specific gene functions. An unbiased clustering of chromatin state profiles across annotated TSSs in lymphoblastoid and skeletal muscle cells distinguished informative patterns predictive of downstream gene expression and functional gene classes (Supplementary Fig. 13,14). Cell type-specific patterns were also evident when TSSs were simply assigned to the most prevalent chromatin state. Promoters activate in skeletal muscle were associated with extracellular structure genes (8.5-fold enrichment), those activate in lymphoblastoid cells with immune response genes (7.2-fold enrichment), and those active in both with metabolic housekeeping genes.

Clustering of promoter and enhancer states based on their activity patterns

Extending our pair-wise promoter analysis, we clustered strong promoter and strong enhancer regions across all cell types (see **Methods**). This revealed clusters showing common activity and associated with highly coherent functions (Fig. 2a,b). For promoter clusters, these include immune response (GM12878-specific clusters, $p<10^{-18}$), cholesterol transport (HepG2-specific, 10^{-4}), and metabolic processes (all cells, 10^{-131}). Remarkably, genes assigned to enhancer clusters by proximity also showed strong functional enrichments, including immune response (GM12878-specific, 10^{-6}), lipid metabolism (HepG2-specific, 10^{-5}) and angiogenesis (HUVEC-specific, 10^{-4}).

Promoters and enhancers differed in their overall specificity. The majority of promoter clusters showed activity in multiple cell types, consistent with previous work[5],[10] (Fig. 2a). Enhancer clusters are significantly more cell type-specific, with few regions showing activity in more than two cell types and a majority being specific to a single cell type (Fig. 2b).

We also found differences in the relative contributions of enhancer-based and promoter-based regulation among gene classes. Developmental genes appear strongly regulated by both, showing the highest number of proximal enhancers and diverse promoter states, including poised and Polycomb-repressed (Supplementary Fig. 15). Tissue-specific genes (e.g. immune genes, steroid metabolism genes) appear more dependent on enhancer regulation, showing multiple tissue-specific enhancers but less diverse promoter states. Lastly, housekeeping genes are primarily promoter-regulated with few enhancers in their vicinity.

Overall, this dynamic view of the chromatin landscape suggests that multi-cell chromatin profiles can be as productive for systems biology as expression analysis has traditionally been, and may hold additional information on genome regulatory programs, which we explore next.

Correlations in activity profiles link enhancers to target genes

We next investigated functional interconnections between enhancers, the factors that activate or repress them, and the genes whose expression they regulate, by defining 'activity profiles' for each across the cell types (Fig. 3). We complemented these enhancer activity profiles (Fig. 3a) with profiles for gene expression (Fig. 3b), sequence motif enrichment (Fig. 3d), and the expression of TFs recognizing each motif (Fig. 3e). We used correlations between these profiles to probabilistically link enhancers to their downstream targets and upstream regulators (see **Methods**).

We found that patterns of enhancer activity (**Fig.** 2b,3a) correlated strongly with patterns of nearest-gene expression (Fig. 3b, correlation >0.9 in 16 of 20 clusters). Since this correlation remained high even for large distances (>50kb), we used activity correlation as a complement to genomic distance for linking enhancers to target genes (see **Methods**). Activity-based linking yielded increased functional gene class enrichments for several clusters (Supplementary Fig. 16).

We validated our approach using quantitative trait locus (QTL) mapping studies which use co-variation between SNP alleles and gene expression levels to link cis-regulatory regions to target genes. Investigation of four recent QTL studies in liver[20] and lymphoblastoid cells[21]-[23] revealed remarkable agreement with our enhancer predictions. Enhancers linked to a given target gene by our method were significantly enriched for SNPs correlated with the gene's expression level (Supplementary Fig. 17), thus confirming our enhancer-gene linkages with orthogonal data.

Correlations with TF expression and motif enrichment predict upstream regulators

We next predicted sequence-specific TFs likely to target enhancers in a given cluster based on regulatory motif enrichments. This implicated a number of TFs whose known biological roles matched the respective cell types (Fig. 3d, Supplementary Fig. 18). When ChIP-seq data was available in the relevant cell type, we confirmed that enriched motifs were preferentially bound by the cognate factor (Fig. 3c). Oct4 motif instances in cluster A (ES-specific enhancers) were preferentially bound by Oct4 in ES cells[24], and NF-kB motif instances in cluster F (lymphoblastoid-specific enhancers) were preferentially bound by NF-kB in lymphoblastoid cells[14]. In both cases, motif instances in cell type-specific enhancers showed a ~5-fold increase in binding compared to other enhancers.

However, sequence-based motif enrichments do not distinguish causality. Enrichment could reflect a parallel binding event that does not affect the chromatin state, or the motif could actually be antagonistic to the enhancer state through specific repression in orthogonal cell types. To distinguish between these possibilities, we complemented the observed motif enrichments with cell type-specific expression for the corresponding TFs (Fig. 3e). We then correlated a 'motif score' based on motif enrichment in a given cluster, and a 'TF-expression score' based on the agreement between the TF expression pattern and the cluster activity profile (see **Methods**). A positive correlation between the two scores implies that the TF may be establishing or reinforcing the chromatin state. A negative correlation would instead imply that the TF may act as a repressor. For example, in addition to the enrichment of the Oct4 motif in the ES-specific cluster A, Oct4 is specifically expressed in ES cells, leading to its prediction as a causal regulator of ES cells (Fig. 3e), consistent with known biology[16].

For 18 of the 20 clusters, this analysis revealed one or more candidate regulators. Recovery of known roles for well-studied regulators validated our approach. For example, HNF1, HNF4, and PPARγ are predicted as activators of HepG2-specific enhancers (clusters H,I), PU.1 and NF-κB as activators of lymphoblastoid (GM12878) enhancers (clusters C,F,G), Gata1 as an activator of K562-specific enhancers (cluster B) and Myf as an activator of skeletal muscle (HSMM) enhancers (cluster O)[14],[25]-[27].

The analysis also revealed potentially novel regulatory interactions. ETS factors (Elk1,Tel2,Ets) are predicted activators of enhancers active in both GM12878 and HUVEC (cluster G), but not of GM12878-specific or HUVEC-specific clusters emphasizing the value of unbiased clustering. These connections are consistent with reported roles for ETS factors in lymphopoiesis and endothelium[28]. The prediction of p53 as an activator in HSMM, NHLF, NHEK and HMEC (clusters N,Q,R) likely reflects its maintained activity in these primary cells as opposed to other cell models where it may be suppressed by mutation

(K562)[29], viral inactivation (GM12878)[30] or cytoplasmic localization (ES cells)[31]. A widespread role for p53 in regulating distal elements is consistent with its known binding to distal regions[32],[33].

Our analysis also revealed several repressor signatures, including Gfi1 in K562 and GM12878 cells (clusters B,C) and Bach2 in ES cells (cluster A). Both regulators are known to repress transcription by recruiting histone deacetylases and methyltransferases to proximal promoters[34],[35], and Gfi1 has also been implicated in silencing of satellite repeats[35]. Our regulatory inferences suggest that they also modulate chromatin to inhibit enhancer activity, thus proposing a new mechanism for distal gene regulation.

Validation of predicted binding events and regulatory outcomes

The regulatory inferences above imply TF binding events at motif instances within enhancer regions in specific cellular contexts, which we sought to validate using a general molecular signature. Binding events are associated with nucleosome displacement, a structural change evident in ChIP-seq data for histones[36]. We thus studied local depletions in the chromatin intensity profiles (‘dips’) as indicative of TF binding. We confirmed that dips were present in individual signal tracks at active enhancers, and were associated with preferential sequence conservation and regulatory motif instances (Fig. 4a).

To test our specific predictions, we superimposed chromatin profiles of coordinately regulated enhancer regions, anchoring them on the implied motif instances. Striking dips precisely coincide with regulatory motifs, and are both cell type-specific and region-specific, exactly as predicted (Fig. 4b,c). As dips only appear when the factor is expressed, they also support the identity of the trans-acting TF.

To validate that predicted causal motifs contribute to enhancer activity, we used luciferase reporters. Our model implicated HNF regulators as activators of HepG2-specific enhancers (Fig. 3), and context-specific dips supported binding interactions (Fig. 4c). We thus selected for functional analysis 10 sites with HNF motifs showing dips in strong HepG2-specific enhancers, and evaluated them with and without the HNF motif. We found that permutation of the motif consistently led to a reduction in enhancer activity (Fig. 4d), supporting its predicted causal role.

Assigning candidate regulatory functions to disease-associated variants

Finally, we explored whether our chromatin annotations and regulatory predictions can provide insight into sequence variants associated with disease phenotypes. To that effect, we gathered a large set of non-coding SNPs from GWAS catalogs, an exceedingly small proportion of which are currently understood[37].

We found that disease-associated SNPs are significantly more likely to coincide with strong enhancers (states 4,5; 2-fold enrichment, $p<10^{-10}$), despite the fact that no notable association to these states are seen for SNPs in general or for those SNPs tested in the studies. To test whether SNPs associated with a particular disease might have even more specific correspondences, we examined 426 GWAS datasets. We identified 10 studies[38]-[47]

whose variants showed significant correspondences to cell type-specific strong enhancer states (see **Methods**; Fig. 5a).

Individual variants from these studies were strongly enriched in enhancer states specifically active in relevant cell types (Fig. 5a,b). For example, SNPs associated with erythrocyte phenotypes[38] were found in erythroleukemia cell (K562) enhancers, SNPs associated with systemic lupus erythematosus[39] were found in lymphoblastoid cell (GM12878) enhancers, while SNPs associated with triglyceride[40] phenotypes or blood lipid phenotypes[41] were found in hepatocellular carcinoma cell (HepG2) enhancers. We also applied our model to chromatin data for T-cells[3] (Supplementary Fig. 19), for which strong enhancer states correlated to variants associated with risk of childhood acute lymphoblastic leukemia[48], further validating our approach.

We also used our predicted enhancer-target gene associations to find candidate downstream genes whose expression might be affected by cis-changes occurring in the enhancer region. Although most of the predicted target genes are proximal to the enhancer, a subset of more distal predicted targets could reflect novel candidates for the disease phenotypes (Fig. 5b).

In addition, we identified several instances where a lead GWAS variant does not correspond to a particular chromatin element but a linked variant coincides with an enhancer with the predicted cell type-specificity (Fig. 5c). Thus, chromatin profiles may provide a general means to triage variants within a haplotype block, a common problem faced in GWAS.

Lastly, we identified several cases in which a disease-associated SNP created or disrupted a regulatory motif instance for a predicted causal TF in the relevant cell type (Fig. 5d), suggesting a specific molecular mechanism by which the disease-associated genotype could lead to the observed disease phenotype consistent with our regulatory predictions.

Discussion

Our work provides a systematic view of many chromatin marks across many cell types, demonstrating the power of chromatin profiling as an additional and dynamic layer of genome annotation. We presented methods to distinguish different classes of functional elements, elucidate their cell type-specificities, and reveal cis-regulatory interactions that govern them and ultimately drive target gene expression. By intersecting our predictions with non-coding SNPs from GWAS datasets, we propose potential mechanistic explanations for disease variants, either through their presence within cell type-specific enhancer states, or by their effect on binding motifs for predicted regulators.

Chromatin states dramatically reduced the large combinatorial space of 90 chromatin datasets (2^90 combinations) into a manageable set of biologically-interpretable annotations, thus providing an efficient and robust way to track coordinated changes across cell types. This enabled the systematic identification and comparison of >100 thousand promoter and enhancer elements. Both types of elements are cell type-specific, associated with motif enrichments, and assume strong, weak and poised states that correlate with neighboring gene expression and function. Enhancers showed exquisite tissue-specificity, enrichment in the vicinity of developmental and cell type-specific genes, and predictive power for proximal gene expression, reinforcing their roles as sentinels of tissue-specific gene expression[49]. By elucidating enhancers systematically, and linking them to upstream regulators and downstream genes, our analysis can help provide a missing link between regulators and target genes. The power of the approach should increase considerably as additional phenotypically-distinct cell types are surveyed, and enable a greater proportion of enhancer elements to be incorporated into the connectivity network.

The inferred cis-regulatory interactions make specific testable predictions, many of which were confirmed through additional experiments and analyses. Our enhancer-target gene linkages are supported by cis-regulatory inferences from QTL mapping studies. Predicted TF-motif interactions within cell type-specific enhancers were confirmed in specific cases by TF binding and more generally by depletions in the chromatin profiles at causal motifs in appropriate cellular contexts. Motifs predicted as causal regulators of cell type-specific enhancers were also confirmed in enhancer assays.

The regulatory inferences afforded by multi-cell chromatin profiles are unique and highly complementary to datasets for TF binding, expression, chromatin accessibility, nucleosome positioning, and chromosome conformation[50]. For example, our regulatory predictions can help focus the spectrum of TF binding events to a smaller number of functional interactions. The chromatin-centric approach also complements the extensive body of work on biological network inference from expression data with the potential to introduce enhancers and other genomic elements into connectivity networks.

Our study has important implications for the understanding of disease. Our detailed and dynamic functional annotations of the relatively uncharted non-coding genome can facilitate the interpretation of GWAS datasets by predicting specific cell types and regulators related to specific diseases and phenotypes. Furthermore, the connections derived for enhancer regions, to upstream regulators and downstream genes, propose cis- and trans-acting interactions that may be modulated by the sequence variants. While the current study represents only a first small step in this direction, we expect that future iterations with greater diversity of cell types and improved methodologies will help define the molecular underpinnings of human disease.

Methods Summary

ChIP-seq analysis was performed in biological replicate as described[4] using antibodies validated by Western blots and peptide competitions. ChIP DNA and input controls were sequenced using the Illumina Genome Analyzer. Expression profiles were acquired using Affymetrix GeneChip arrays. Chromatin states were learned jointly by applying an HMM[8] to 10 data tracks for each of the 9 cell types. We focused on a 15 state model that provides sufficient resolution to resolve biologically-meaningful patterns yet is reproducible across cell types when independently processed. We used this model to produce 9 genome-wide chromatin state annotations, which were validated by additional ChIP experiments and reporter assays. Multi-cell type clustering was conducted on locations assigned to strong promoter state 1 (or strong enhancer state 4) in at least one cell type using the k-means algorithm. Enhancer-target gene linkages were predicted by correlating normalized signal intensities of H3K27ac, H3K4me1 and H3K4me2 with gene expression across cell types as a function of distance to the TSS. Upstream regulators were predicted using a set of known TF motifs assembled from multiple sources. Motif instances were identified by sequence match and evolutionary conservation. P-values for GWAS studies were based on randomizing the location of SNPs, and the FDR based on randomizing the assignment of SNPs across studies. Datasets are available from the ENCODE website (http://genome.ucsc.edu/ENCODE), the supporting website for this paper (http://compbio.mit.edu/ENCODE_chromatin_states), and the Gene Expression Omnibus (GSE26386).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank members of the Epigenomics community at the Broad Institute and the Bernstein and Kellis laboratories for helpful discussions and criticisms. We also thank Mario Suva, Eric Mendenhall and Shawn Gillespie for assistance with experiments, and Loyal Goff and Andrew Chess for critical reading of the manuscript. We acknowledge the Broad Institute Genome Sequencing Platform for their expertise and assistance with data production. This research was supported by the National Human Genome Research Institute under an ENCODE grant (U54 HG004570), R01 HG004037, RC1 HG005334, the Howard Hughes Medical Institute, and the National Science Foundation (awards 0644282 and 0905968), and the Sloan Foundation.

Appendix

Methods

Cell culture

Human **H1 ES** cells were cultured in TeSR media[51] on Matrigel by Cellular Dynamics International. Cells were split with dispase and harvested at a passage number between 30 and 40. Prior to harvest, cells were karyotyped and stained for Oct4 to confirm pluripotency. **K562** erythrocytic leukemia cells (ATCC CCL-243, lot # 4607240) were grown in suspension in RPMI medium (HyClone SH30022.02) with 10% fetal bovine serum (FBS) and 1% Antibiotic - Antimycotic (GIBCO 15240-062). Cell density was maintained between 3×10^5 and 7×10^5 cells/ml. **GM12878** B-lymphoblastoid cells (Coriell Cell Repositories, “expansion A”) were grown in suspension in RPMI 1640 medium with 15% FBS (not heat inactivated), 2 mM L-glutamine and 1% penicillin/streptomycin. Cells were seeded at a concentration of $\sim 2 \times 10^5$ viable cells/ml with minimal disruption, and maintained between 3×10^5 and 7×10^5 cells/ml. **HepG2** hepatocellular carcinoma cells (ATCC HB-8065, lot# 4968519) were cultured in DMEM (HyClone SH30022.02) with 10% FBS and 1% penicillin/streptomycin. Cells were trypsinized, resuspended to single cell suspension, split to a confluence between 15 and 20% and then harvested at ~75% confluence. **NHEK** normal human epidermal keratinocytes isolated from skin (Lonza CC-2501, lot# 4F1155J, passage 1) were grown in keratinocyte basal medium 2 (KGM-2 BulletKit, Lonza) supplemented with BPE, hEGF, hydrocortisone, GA-1000, transferrin, epinephrine and insulin. Cells were seeded at recommended density (3500 cells/cm^2), subjected to two to three passages on polystyrene tissue culture plates and harvested at a confluence of 70 to 80%. **HSMM** primary human skeletal muscle myoblasts (Lonza CC-2580, lot#6F4444, passage 2) were cultured in Smooth Muscle Growth Medium-2 (SkGM-2 BulletKit, Lonza) supplemented with rhEGF, dexamethasone, L-glutamine, FBS, and GA-1000. Cells were seeded at recommended density (3500 cells/cm^2), subjected to two to three passages, and harvested at a confluence of 50 to 70%. **NHLF** primary normal human lung fibroblasts (Lonza CC-2512, lot#4F0758, passage 2) were grown in Fibroblast Cell Basal Medium 2 (FGM-2 BulletKit, Lonza) supplemented with hFGF-β, insulin, FBS and GA-100. Cells were seeded at recommended density (2500 cells/cm^2), subjected to two to three passages, and harvested at an approximate confluence of 80%. **HUVEC** primary human umbilical vein endothelial cells (Lonza CC-2517, lot# 7F3239, passage 1) were grown in endothelial basal medium 2 (EGM-2 BulletKit, Lonza) supplemented with hFGF-β, hydrocortisone, VEGF, R3-IGF-1, ascorbic acid, heparin, FBS, hEGF and GA-1000. Cells were seeded at recommended density (2500 – 5000 cells/cm2), subjected to two to three passages, and harvested at a confluence of 70 to 80%. **HMEC** primary human mammary epithelial cells from mammary reduction tissue (Lonza CC-2551, passage 7) were grown in mammary epithelia basal medium (MEGM BulletKit, Lonza) supplemented with hEGF-β, hydrocortisone, BPE, GA-1000 and insulin. Cells were seeded as recommended density (2500 cells/cm2), subjected two to three passages and harvested at 60 to 80% confluence.

Antibodies

ChIP assays were performed using the following antibody reagents: H3K4me1 (Abcam ab8895, lot 38311/659352), H3K4me2 (Abcam ab7766, lot 56293), H3K4me3 (Abcam ab8580, lot 331024; Milipore 04-473, lot DAM1623866), H3K9ac (Abcam ab44441, lot 455103/550799), H3K27ac (Abcam ab4729, lot 31456), H3K36me3 (Abcam ab9050, lot 136353), H4K20me1 (Abcam ab9051, lot 104513/519198), H3K27me3 (Millipore 07-449, lot DAM1387952/DAM1514011), CTCF (Millipore 07-729, lot 1350637), H3K9me3 (Abcam ab8898, lot 484088), H2A.Z (Millipore 07-594, lot DAM1504736) and RNAPII N-terminus (Santa Cruz sc-899X, lot H0510). All antibody lots were extensively validated for specificity and efficacy in ChIP-seq. Western blots were used to confirm specific recognition of histone protein (or CTCF). Dot plots performed using arrayed histone tail peptides representing various modification states were used to confirm specificity for the appropriate modification. ChIP-seq assays performed on a common cell reagent were used to confirm consistency between different lots of the same antibody.

Chromatin immunoprecipitation (ChIP)

Cells were harvested by cross-linking with 1% formaldehyde in cell culture medium for 10 min at 37 °C. After quenching with the addition of 125 mM glycine for 5 min at 37 °C, the cells were washed twice with cold PBS containing protease inhibitor (Roche). After aspiration of all liquid, pellets consisting of ~10^7 cells were flash frozen and stored at −80°C. Fixed cells were thawed and sonicated to obtain chromatin fragments of ~200 to 700 bp using a Bioruptor (Diagenode). Immunoprecipitations were performed as described, retaining a fraction of input ‘whole cell extract’ (WCE) as a control[4]. Briefly, sonicated chromatin was diluted 10-fold and incubated with ~5 μg antibody overnight. Antibody–chromatin complexes were pulled-down using protein A-sepharose, washed and then eluted. After cross-link reversal and proteinase K treatment, immunoprecipitated DNA was extracted with phenol, precipitated in ethanol, and treated with RNase. ChIP DNA was quantified by fluorometry using the Qubit assay (Invitrogen).

Next-generation sequencing

For each ChIP or control sample, ~5 ng of DNA was used to generate a standard Illumina sequencing library. Briefly, DNA fragments were end-repaired using the End-It DNA End-Repair Kit (Epicentre), extended with a 3′ ‘A’ base using Klenow (3′→5′ exo-, 0.3 U/μl, NEB), ligated to standard Illumina adapters (75 bp with a ‘T’ overhang) using DNA ligase (0.05 U/μl, NEB), gel purified on 2% agarose, retaining products between 275 and 700 bp, and subjected to 18 PCR cycles. These libraries were quantified by fluorometry and evaluated by quantitative PCR or a multiplexed digital hybridization-based analysis (NanoString nCounter)[52] to confirm representation and specific enrichment of DNA species. Libraries were sequenced in one or two lanes on the Illumina Genome Analyzer using standard procedures for cluster amplification and sequencing-by-synthesis.

Expression profiling

Cytosolic RNA was isolated using RNeasy Columns (Qiagen) from the same cell lots as above. Gene expression profiles were acquired using Affymetrix GeneChip arrays. The data were normalized using the GenePattern expression data analysis package[53]. CEL files were processed by RMA, quantile normalization and background correction. Two replicate expression datasets for each cell type were averaged and $\log_2$ transformed. Gene level normalization across cell types was computed by mean normalization.

Primary processing of sequencing reads

ChIP-seq reads were aligned to human genome build HG18 with Maq (http://maq.sourceforge.net/maqman.shtml) using default parameters. All reads were truncated to 36 bases before alignment. Signal density maps for visualization were derived by extending sequencing reads by 200 bp in the 3′ direction (the estimated median size of ChIP fragments), and then counting the total number of overlapping reads at 25 bp intervals. Replicate ChIP-seq experiments were verified by comparing enriched intervals as described[4], and then combined into a single dataset. For the HMM, density maps were derived by extending sequencing reads by 200 bp in the 3′ direction and then assigning them to a single 200 bp window based on the midpoint of the extended read. These maps were then binarized at a 200 bp resolution based on a Poisson background model using a threshold of 10^{-4}.

Joint learning of HMM states across cell types

To handle data from the 9 cell types, we concatenated their genomes to create an extended virtual genome that we used to train the HMM. We applied the model to ten tracks corresponding to the different chromatin marks and input using a multivariate HMM-model as described[8]. Here, we used a Euclidean distance for determining initial parameters for the nested initialization step. After learning and evaluating a set of roughly nested models, considering up to 25 states, we focused on a 15 state model that provides sufficient resolution to resolve biologically-meaningful chromatin patterns and yet is highly reproducible across cell types when independently processed (Supplementary Fig. 7). We used this model to compute the probability that each location is in a given state, and then assigned each 200 bp interval to its most likely state for each cell type. Even though our model focuses on presence/absence frequencies of marks, we found that our states also capture signal intensity differences between high-frequency and low-frequency marks (Supplementary Fig. 9).

Enrichment analysis

For each state, enrichments for different annotations were computed at 200 bp resolution, with the exception of conservation which was computed at nucleotide resolution. We used annotations obtained through the UCSC Genome Browser[54] for RefSeq TSSs and transcribed regions[55], PhastCons[56], DNase-seq for K562 cells[12], c-Myc ChIP-seq for K562 cells[13], NF-κB ChIP-seq for GM12878[14], Oct4 in ES cells[24] and nuclear lamina[15]. Gene functional group enrichments were determined using STEM[57] and biological process annotations in the Gene Ontology database[58]. P-values were calculated based on hypergeometric distribution and corrected for multiple testing using Bonferonni correction.

Comparing chromatin state assignments between cell types

For each pair of cell types, the chromatin state assignments at each genomic position were compared. We calculated the frequency with which each pair of states occurred, and normalized this against the expected frequency based on the amount of genome covered by each state. The fold-enrichments in Figure 2a reflect an aggregation across all 72 possible pairs of cell types.

Pair-wise promoter clustering

Promoters for RefSeq genes were clustered based on the most likely chromatin state assignment across a 2 kb region centered on the TSS. Clustering was performed jointly across GM12878 and HSMM, and was restricted to genes with corresponding Affymetrix expression. Briefly, each promoter was treated as a 330 element binary vector where each position of the vector corresponded to a position along the promoter, cell type and state. Clustering was performed on these vectors by k-means in Matlab. Gene expression values were calculated based on the corresponding Affymetrix probe set closest to the TSS.

Multi-cell type promoter and enhancer clustering

Promoter state clustering was performed for all 200 bp intervals assigned to the strong promoter state (state 1) in at least one cell type. Each interval was represented by a single vector corresponding to the estimated probabilities that it is in the strong promoter state for each of the 9 cell types, accounting for model assignment uncertainty and biological noise. These were determined from the model posterior probabilities of state assignments and a comparison of state assignments in replicate experimental data. Clustering was performed by k-means using Matlab. We found that 20 clusters provided sufficient resolution to distinguish major cell type-specific patterns. Enhancer state clustering was performed for all 200 bp intervals assigned to strong enhancer state 4 in at least one cell type using identical procedures. For the purposes of display in Figure 2, the locations were randomly downsampled. For the purpose of identifying enriched functional gene categories in Figure 2b, enhancers were linked to the nearest TSS up to 50 kb distant excluding those within 5kb. Enhancer-gene correspondences based on the nearest gene were used for the expression analysis of distance based linked genes in Figure 3b.

Linking enhancer locations to correlated genes

To predict linkages between enhancer states and target genes, we combined distance-based information with correlation across cell types of gene expression levels with read-depth normalized signal intensity of three histone modifications associated with enhancer states: H3K4me1, H3K4me2 and H3K27ac. For each enhancer state (4-7), cell type, and 200bp interval between 5kb and 125kb from the TSS, we trained logistic regression classifiers. The classifiers were trained to use mark intensity-expression correlation values to distinguish real instances of pairs of enhancer states and gene expression values from control pairs based on randomly re-assigning expression values to different genes. To learn a smooth and robust function at each position we included as part of the training all enhancer state assignments within a 10kb window centered at the position. The link score for a specific enhancer-gene linkage was defined as the ratio of the corresponding logistic regression classifier probability score to that for the randomized data.

For the evaluation of the expression quantitative trait loci (QTL) analysis we used a link score of 2.5. The eQTL data was obtained from the University of Chicago QTL browser (http://eqtl.uchicago.edu/cgibin/gbrowse/eqtl/). In the QTL evaluation for each SNP that overlapped a strong enhancer state (4 or 5) and was within 125kb of a TSS excluding locations within 5kb and was associated with a gene for which we had expression data was considered eligible to be supported by our linked predictions. We computed the fraction we observed linked based on our linked predictions relative to the fraction that would be expected to be linked conditioned on knowing the distance distributions of the SNPs relative to the gene TSS.

For the evaluation of linked predictions using the Gene Ontology we use the same link score and the same distance relative to the TSS used to define genes associated with clusters. The base set of genes in the enrichment analysis here were all genes which could be linked in at least one cluster.

Motif and TF analysis

A database of known TF motifs was collated by combining motifs from TRANSFAC (version 11.3)[59], Jaspar (2010-05-07)[60] and protein binding microarray datasets[61]-[63]. Motif instances in non-coding and non-repetitive regions of the genome were identified using these motifs and sequence conservation using a 29-way alignment of Eutherian mammal genomes (Lindblad-Toh et al, in preparation). These were filtered using a significance threshold of $p<4^{-8}$ for the motifs[64], and confidence level based on conservation. Motifs were linked to corresponding TFs using metadata provided by the source. Motif enrichments for chromatin state clusters were computed as ratios to the instances of shuffled motifs, in order to correct for non-specific conservation and composition. A confidence interval was calculated for each ratio using Wilson score intervals (z=1.5), selecting the most conservative value within the confidence interval. In cases where multiple motif variants were available for the same TF, the one that showed the most variance in enrichment across clusters was selected.

For predicting causal activators and repressors, motif scores and TF expression scores were correlated as follows. Motif scores were calculated as described above. TF expression scores were calculated for each cluster by correlating the expression of the TF across the cell types with the activity profile of the enhancers in that cluster (defined by the cluster means from the k-means clustering). The motif scores and the TF expression scores were then correlated against each other to identify positively- and negatively-correlated TFs.

TF-motif interactions predicted for strong enhancer states in specific cell types were validated by using the raw ChIP-seq tag enrichments as proxy for nucleosome positioning. For this purpose, sequencing reads were processed as above, except that the middle 75 bp of inferred ChIP fragments were used to derive signal density informative of nucleosome-depletion (dips), as described[36]. Superposition plots show tag enrichments relative to a uniform background computed based on sequencing depth.

Quantitative real-time PCR

Enrichment ratios for RNAPII and H2A.Z ChIPs were determined relative to input chromatin by quantitative real-time PCR using an ABI 7900 detection system, in biological replicate as described previously[65]. Regions used for validation correspond to 3 different chromatin states, including 13 for state 1 (arbitrarily selected), 11 for state 4 (arbitrarily selected but excluding regions within 2 kb of a state 1 annotation) and 11 for state 7 (arbitrarily selected but excluding regions within 2 kb of a state 1 or state 4 annotation). PCR primers are listed in **Supplementary Data 1**.

Functional enhancer assays

The SV40 promoter was first inserted between the HindIII and NcoI sites of pGL4.10 (Promega). Next, 250bp sequences from the reference genome (hg18) corresponding to different chromatin states (8 from HepG2 state 4, 7 from HepG2 state 7, and 7 from GM12878 state 4) were synthesized (GenScript) and then inserted between the two SfiI sites upstream of the SV40 promoter. HepG2 cells were seeded into 96 well plates at a density of 5×10^4 cells/well and expanded overnight to ~50% confluency. The cells were then transfected with 400 ng of a pGL4.10-derived plasmid and 100 ng of pGL4.73 (Promega) using Lipofectamine LTX. Firefly and Renilla luciferase activities were measured 24 hr post transfection using Dual-Glow (Promega) and an EnVision 2103 multilabel reader (PerkinElmer) from triplicate experiments. Data are reported as light units relative to a control plasmid. For validation of causal TF motifs, 10 sequences of 250 bp corresponding to HepG2-specific strong enhancers (state 4) with dips and HNF motifs were tested as above, and compared to identical sequences except with the HNF motif permuted. Tested enhancer elements are listed in **Supplementary Data 1**.

GWAS SNP Analysis

The GWAS variants and SNP coordinates were obtained from the NHGRI catalog and the UCSC browser (October 30, 2010)[37],[54]. This set was refined by extending the blood lipid GWAS[41] set to contain all reported SNPs, and by bifurcating the hematological and biochemical traits study[46] into a hematological traits set and a biochemical traits set. We limited our analysis to studies reporting 2 or more associated SNPs. The variants from each study were intersected with chromatin states from each of the cell types. The reported p-values were based on the overlap of associated SNPs with strong enhancer states 4 and 5. We controlled for non-independence between proximal SNPs based on a randomization test where SNPs were randomly shifted while preserving relative distance. We then defined an estimated false discovery rate based on permutations in which SNPs were randomly re-assigned to different studies, and recomputed corrected p-values. Estimates of false discovery rates based on these permutations control for multiple testing of studies and cell types and for general non-specific enrichments for states 4 and 5 with GWAS hits. Candidate gene targets were predicted for a subset of variants associated with enhancer states based on the lead cell type using the linking method described above.

References

51. Ludwig TE, et al. Feeder-independent culture of human embryonic stem cells. Nat Methods. 2006; 3:637–646. [PubMed: 16862139]
52. Geiss GK, et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol. 2008; 26:317–325. [PubMed: 18278033]
53. Reich M, et al. GenePattern 2.0. Nat Genet. 2006; 38:500–501. [PubMed: 16642009]
54. Kent WJ, et al. The human genome browser at UCSC. Genome Res. 2002; 12:996–1006. [PubMed: 12045153]
55. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007; 35:D61–65. [PubMed: 17130148]
56. Siepel A, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005; 15:1034–1050. [PubMed: 16024819]
57. Ernst J, Bar-Joseph Z. STEM: a tool for the analysis of short time series gene expression data. BMC Bioinformatics. 2006; 7:191. [PubMed: 16597342]
58. Ashburner M, et al. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat Genet. 2000; 25:25–29. [PubMed: 10802651]
59. Matys V, et al. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 2003; 31:374–378. [PubMed: 12520026]
60. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004; 32:D91–94. [PubMed: 14681366]
61. Berger MF, et al. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell. 2008; 133:1266–1276. [PubMed: 18585359]
62. Rosenkranz HS, Klopman G. A re-examination of the genotoxicity and carcinogenicity of azathioprine. Mutat Res. 1991; 251:157–161. discussion 163-154. [PubMed: 1944373]
63. Berger MF, et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol. 2006; 24:1429–1435. [PubMed: 16998473]
64. Touzet H, Varre JS. Efficient and accurate P-value computation for Position Weight Matrices. Algorithms Mol Biol. 2007; 2:15. [PubMed: 18072973]

65. Bernstein BE, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006; 125:315–326. [PubMed: 16630819]

References

1. Birney E, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. [PubMed: 17571346]
2. Kim HD, Shay T, O'Shea EK, Regev A. Transcriptional regulatory circuits: predicting numbers from alphabets. Science. 2009; 325:429–432. [PubMed: 19628860]
3. Barski A, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007; 129:823–837. [PubMed: 17512414]
4. Mikkelsen TS, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. 2007; 448:553–560. [PubMed: 17603471]
5. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. A chromatin landmark and transcription initiation at most promoters in human cells. Cell. 2007; 130:77–88. [PubMed: 17632057]
6. 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]
7. Hon G, Wang W, Ren B. Discovery and annotation of functional chromatin signatures in the human genome. PLoS Comput Biol. 2009; 5:e1000566. [PubMed: 19918365]
8. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010; 28:817–825. [PubMed: 20657582]
9. Bernstein BE, et al. Genomic maps and comparative analysis of histone modifications in human and mouse. Cell. 2005; 120:169–181. [PubMed: 15680324]
10. Heintzman ND, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009
11. Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell. 2009; 137:1194–1211. [PubMed: 19563753]
12. Hansen RS, et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci U S A. 2010; 107:139–144. [PubMed: 19966280]
13. Raha D, et al. Close association of RNA polymerase II and many transcription factors with Pol III genes. Proc Natl Acad Sci U S A. 2010; 107:3639–3644. [PubMed: 20139302]
14. Kasowski M, et al. Variation in transcription factor binding among humans. Science. 2010; 328:232–235. [PubMed: 20299548]
15. Guelen L, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. 2008; 453:948–951. [PubMed: 18463634]
16. Jaenisch R, Young R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell. 2008; 132:567–582. [PubMed: 18295576]
17. De Santa F, et al. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. 2010; 8:e1000384. [PubMed: 20485488]
18. Kim TK, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010; 465:182–187. [PubMed: 20393465]
19. Talbert PB, Henikoff S. Histone variants--ancient wrap artists of the epigenome. Nat Rev Mol Cell Biol. 2010; 11:264–275. [PubMed: 20197778]
20. Schadt EE, et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 2008; 6:e107. [PubMed: 18462017]
21. Pickrell JK, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464:768–772. [PubMed: 20220758]
22. Montgomery SB, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010; 464:773–777. [PubMed: 20220756]
23. Veyrieras JB, et al. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. 2008; 4:e1000214. [PubMed: 18846210]

24. Kunarso G, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 2010; 42:631–634. [PubMed: 20526341]

25. Fujiwara T, et al. Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. Mol Cell. 2009; 36:667–681. [PubMed: 19941826]

26. Lemaigre F, Zaret KS. Liver development update: new embryo models, cell lineage control, and morphogenesis. Curr Opin Genet Dev. 2004; 14:582–590. [PubMed: 15380251]

27. Sabourin LA, Rudnicki MA. The molecular regulation of myogenesis. Clin Genet. 2000; 57:16–25. [PubMed: 10733231]

28. Bartel FO, Higuchi T, Spyropoulos DD. Mouse models in the study of the Ets family of transcription factors. Oncogene. 2000; 19:6443–6454. [PubMed: 11175360]

29. Law JC, Ritke MK, Yalowich JC, Leder GH, Ferrell RE. Mutational inactivation of the p53 gene in the human erythroid leukemic K562 cell line. Leuk Res. 1993; 17:1045–1050. [PubMed: 8246608]

30. Forte E, Luftig MA. MDM2-dependent inhibition of p53 is required for Epstein-Barr virus B-cell growth transformation and infected-cell survival. J Virol. 2009; 83:2491–2499. [PubMed: 19144715]

31. Solozobova V, Rolletschek A, Blattner C. Nuclear accumulation and activation of p53 in embryonic stem cells after DNA damage. BMC Cell Biol. 2009; 10:46. [PubMed: 19534768]

32. Cawley S, et al. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell. 2004; 116:499–509. [PubMed: 14980218]

33. Wei CL, et al. A global map of p53 transcription-factor binding sites in the human genome. Cell. 2006; 124:207–219. [PubMed: 16413492]

34. Hoshino H, et al. Co-repressor SMRT and class II histone deacetylases promote Bach2 nuclear retention and formation of nuclear foci that are responsible for local transcriptional repression. J Biochem. 2007; 141:719–727. [PubMed: 17383980]

35. Vassen L, Fiolka K, Moroy T. Gfi1b alters histone methylation at target gene promoters and sites of gamma-satellite containing heterochromatin. EMBO J. 2006; 25:2409–2419. [PubMed: 16688220]

36. He HH, et al. Nucleosome dynamics define transcriptional enhancers. Nat Genet. 2010; 42:343–347. [PubMed: 20208536]

37. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009; 106:9362–9367. [PubMed: 19474294]

38. Ganesh SK, et al. Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nat Genet. 2009; 41:1191–1198. [PubMed: 19862010]

39. Han JW, et al. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nat Genet. 2009; 41:1234–1237. [PubMed: 19838193]

40. Kathiresan S, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet. 2008; 40:189–197. [PubMed: 18193044]

41. Teslovich TM, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010; 466:707–713. [PubMed: 20686565]

42. Houlston RS, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet. 2008; 40:1426–1435. [PubMed: 19011631]

43. Newton-Cheh C, et al. Genome-wide association study identifies eight loci associated with blood pressure. Nat Genet. 2009

44. Stahl EA, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet. 2010; 42:508–514. [PubMed: 20453842]

45. Liu X, et al. Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis. Nat Genet. 2010; 42:658–660. [PubMed: 20639880]

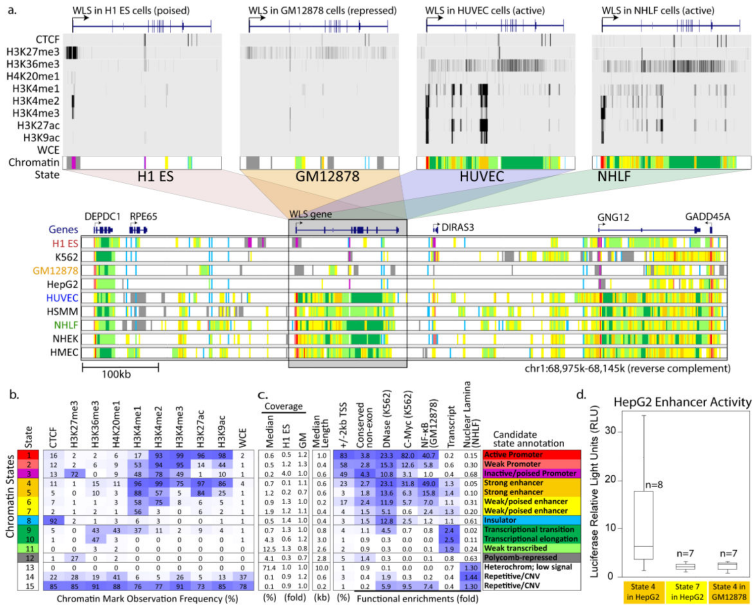
46. Kamatani Y, et al. Genome-wide association study of hematological and biochemical traits in a Japanese population. Nat Genet. 2010; 42:210–215. [PubMed: 20139978]

47. Soranzo N, et al. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat Genet. 2009; 41:1182–1190. [PubMed: 19820697]

48. Papaemmanuil E, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. Nat Genet. 2009; 41:1006–1010. [PubMed: 19684604]

49. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009; 461:199–205. [PubMed: 19741700]

50. Naumova N, Dekker J. Integrating one-dimensional and three-dimensional maps of genomes. J Cell Sci. 2010; 123:1979–1988. [PubMed: 20519580]



State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE	Coverage Median	Coverage H1 ES	Coverage GM	Median Length	+/-2kb TSS	Conserved non-exon	DNase (K562)	C-Myc (K562)	NF-κB (GM12878)	Transcript	Nuclear Lamina (NHLF)	Candidate state annotation
1	16	2	2	6	17	93	99	96	98	2	0.6	0.5	1.2	1.0	83	3.8	23.3	82.0	40.7	0.2	0.15	Active Promoter
2	12	2	6	9	53	94	95	14	44	1	0.5	1.2	1.3	0.4	58	2.8	15.3	12.6	5.8	0.6	0.30	Weak Promoter
3	13	72	0	9	48	78	49	1	10	1	0.2	4.0	1.0	0.6	49	4.3	10.8	3.1	1.0	0.4	0.68	Inactive/poised Promoter
4	11	1	15	11	96	99	75	97	86	4	0.7	0.1	1.1	0.6	23	2.7	23.1	31.8	49.0	1.3	0.05	Strong enhancer
5	5	0	10	3	88	57	5	84	25	1	1.2	0.2	0.7	0.6	3	1.8	13.6	6.3	15.8	1.4	0.10	Strong enhancer
6	7	1	1	3	58	75	8	6	5	1	0.9	1.3	1.0	0.2	17	2.4	11.9	5.7	7.0	1.1	0.31	Weak/poised enhancer
7	2	1	2	1	56	3	0	6	2	1	1.9	1.2	1.1	0.4	4	1.5	5.1	0.6	2.4	1.3	0.20	Weak/poised enhancer
8	92	2	1	3	6	3	0	0	1	1	0.5	1.4	1.0	0.4	3	1.5	12.8	2.5	1.2	1.1	0.61	Insulator
9	5	0	43	43	37	11	2	9	4	1	0.7	1.3	1.0	0.8	4	1.1	4.5	0.7	0.8	2.4	0.02	Transcriptional transition
10	1	0	47	3	0	0	0	0	0	1	4.3	0.6	1.2	3.0	1	0.9	0.3	0.0	0.0	2.5	0.11	Transcriptional elongation
11	0	0	3	2	0	0	0	0	0	0	12.5	1.3	0.8	2.6	2	0.9	0.3	0.0	0.1	1.9	0.24	Weak transcribed
12	1	27	0	2	0	0	0	0	0	0	4.1	0.3	0.7	2.8	5	1.4	0.3	0.0	0.1	0.8	0.63	Polycomb-repressed
13	0	0	0	0	0	0	0	0	0	0	71.4	1.0	1.0	10.0	1	0.9	0.1	0.0	0.0	0.7	1.30	Heterochrom; low signal
14	22	28	19	41	6	5	26	5	13	37	0.1	0.9	1.2	0.6	3	0.4	1.9	0.3	0.2	0.4	1.44	Repetitive/CNV
15	85	85	91	88	76	77	91	73	85	78	0.1	0.9	1.0	0.2	1	0.2	5.9	9.5	7.4	0.4	1.30	Repetitive/CNV
	Chromatin Mark Observation Frequency (%)										(%)	(fold)		(kb)	(%)	Functional enrichments (fold)						

Figure 1. Chromatin state discovery and characterization

a, Top: Profiles for nine chromatin marks (grayscale) are shown across the wntless (WLS) gene in four cell types, and summarized in a single chromatin state annotation track for each (colored according to **b**). WLS is poised in ES cells, repressed in GM12878 cells, and transcribed in HUVEC and NHLF. Its TSS switches accordingly between poised (purple), repressed (grey) and active (red) promoter states; enhancer regions within the gene body become strongly activated (orange, yellow); and its gene body changes from low signal (white) to transcribed (green). These chromatin state changes summarize coordinated changes in many chromatin marks; for example, H3K27me3, H3K4me3 and H3K4me2 jointly mark a poised promoter, while loss of H3K27me3 and gain of H3K27ac and H3K9ac mark promoter activation. Bottom: Nine chromatin state tracks, one per cell type, in a 900kb region centered at WLS summarize 90 chromatin tracks in directly-interpretable dynamic annotations, showing activation and repression patterns for 6 genes and hundreds of regulatory regions, including enhancer states. **b,** Chromatin states learned jointly across cell types by a multivariate HMM. Table shows emission parameters learned *de novo* based on genome-wide recurrent combinations of chromatin marks. Each entry denotes the frequency with which a given mark is found at genomic positions corresponding to the chromatin state. **c,** Genome coverage, functional enrichments, and candidate annotations for each chromatin state. Blue shading indicates intensity, scaled by column. **d,** Box plot depicts enhancer activity for predicted regulatory elements. 250bp-long sequences corresponding to strong or weak/poised HepG2 enhancer elements, or GM12878-specific strong enhancer elements were inserted upstream of a luciferase gene and transfected into HepG2 cells. Reporter activity was measured in relative light units. Robust activity is seen for strong enhancers in the matched cell type, but not for weak/poised enhancers or for strong enhancers specific to a different cell type. Box-and-whiskers indicate 5th, 25th, 50th, 75th and 95th percentiles.

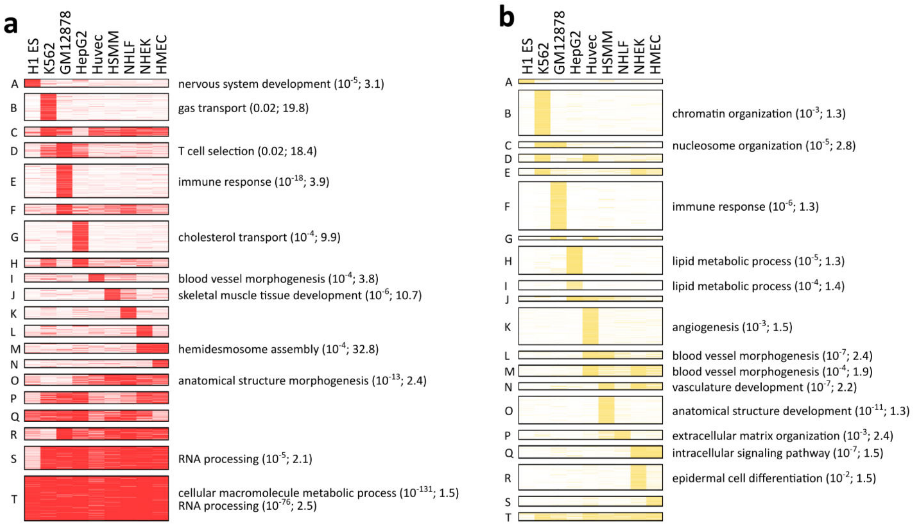


Figure 2. Cell type-specific promoter and enhancer states and associated functional enrichments

a, Clustering of genomic locations (rows) assigned to active promoter state 1 (red) across cell types (columns) reveals 20 common patterns of activity (A-T) (see **Methods**). For each cluster, enriched gene ontology (GO) terms are shown with hypergeometric P-value and fold-enrichment, based on nearest TSS. For most clusters, several cell types show strong (dark red) or moderate (light red) activity. **b,** Analogous clustering and functional enrichments for strong enhancer state 4 (yellow). Enhancer states show greater cell type-specificity, with most clusters active in only one cell type.

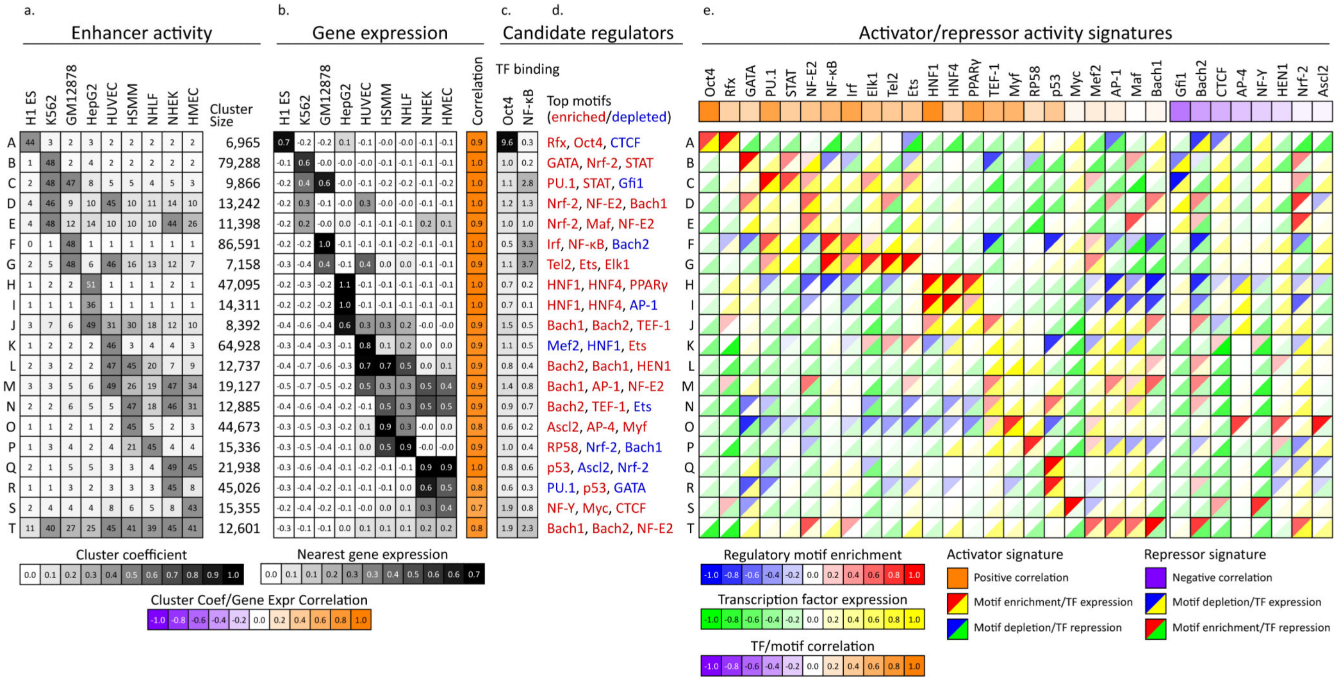


Figure 3. Correlations in activity patterns link enhancers to gene targets and upstream regulators

a, Average enhancer activity across the cell types (columns) for each enhancer cluster (rows) defined in Figure **2b** (labeled A-T) and number of 200bp windows in each cluster. **b,** Average mRNA expression of nearest gene across the cell types and correlation with enhancer activity profile from **a**. High correlations between enhancer activity and gene expression provide a means for linking enhancers to target genes. **c,** Enrichment for Oct4 binding in ES cells[24] and NF-κB binding in lymphoblastoid cells[14] for each cluster. **d,** Strongly enriched (red) or depleted (blue) motifs for each cluster, from a catalog of 323 consensus motifs. **e,** Predicted causal regulators for each cluster based on positive (activators) or negative (repressors) correlations between motif enrichment (top left triangles) and TF expression (bottom right triangles). For example, a red/yellow combination predicts Oct4 as a positive regulator of ES-specific enhancers, as its motif-based predicted targets are enriched (red upper triangle) for enhancers active in ES (cluster A), and the Oct4 gene is expressed specifically in ES cells, resulting in a positive TF expression correlation (yellow triangle). Overall correlations between motif and TF expression across all clusters denote predicted activators (positive correlation, orange) and repressors (negative correlation, purple).

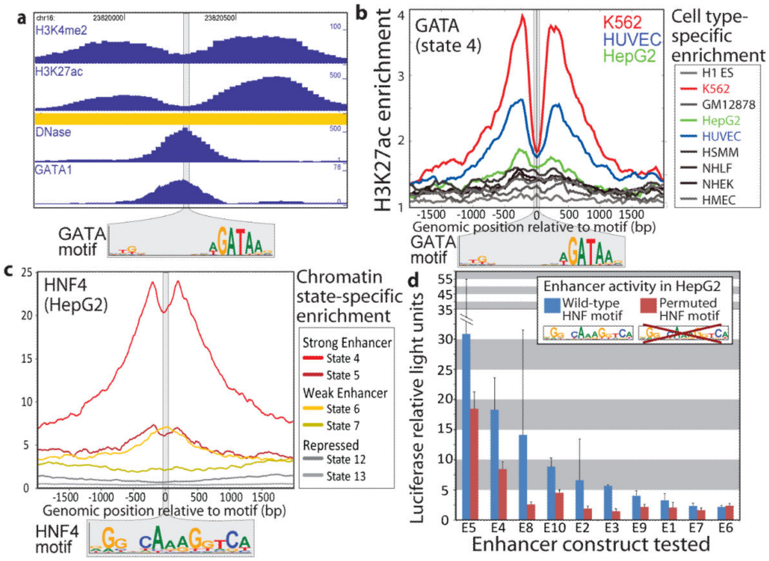
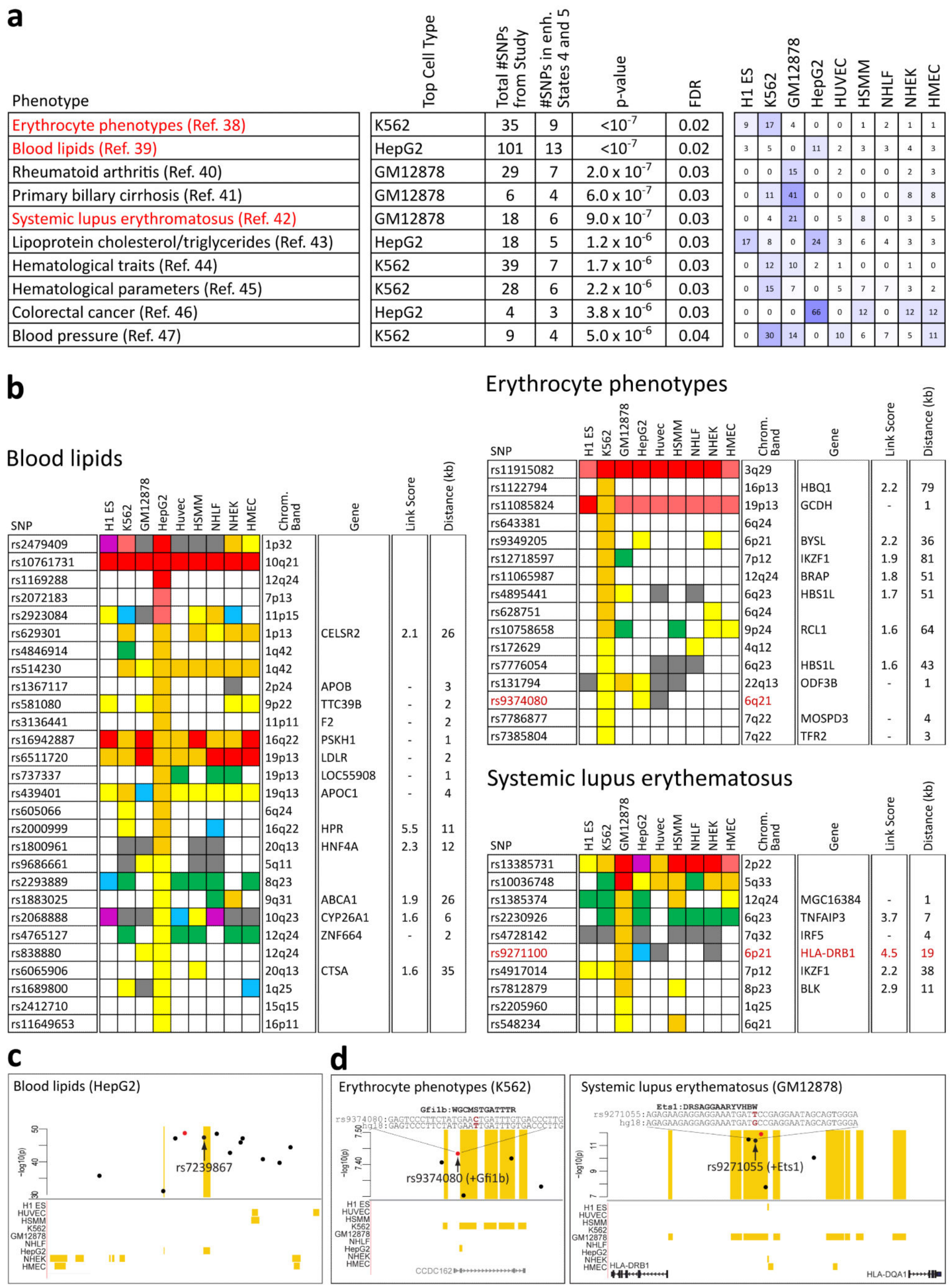


Figure 4. Validation of regulatory predictions by nucleosome depletions and enhancer activity

a, Dips in chromatin intensity profiles in a K562-specific strong enhancer (orange) coincide with a predicted causal GATA motif instance (logo). The dips likely reflect nucleosome displacement associated with TF binding, supported by DNase hypersensitivity[12] and GATA1 binding[25]. **b,** Superposition of H3K27ac signal across loci containing GATA motifs, centered on motif instances, shows dips in K562 cells, as predicted. **c,** Superposition of H3K4me2 signal for HepG2 cells shows dips over HNF4 motifs in strong enhancer states, as predicted. **d,** HepG2-specific strong enhancers with predicted causal HNF motifs were tested in reporter assays. Constructs with permuted HNF motifs (red) led to significantly reduced luciferase activity compared to wild type (blue), with an average 2-fold reduction. Mean luciferase relative light units over three replicates and 95% confidence intervals are indicated.



Phenotype	Top Cell Type	Total #SNPs from Study	#SNPs in enh. States 4 and 5	p-value	FDR	H1 ES	K562	GM12878	HepG2	HUVEC	HSMM	NHLF	NHEK	HMEC
Erythrocyte phenotypes (Ref. 38)	K562	35	9	$<10^{-7}$	0.02	9	17	4	0	0	1	2	1	1
Blood lipids (Ref. 39)	HepG2	101	13	$<10^{-7}$	0.02	3	5	0	11	2	3	3	4	3
Rheumatoid arthritis (Ref. 40)	GM12878	29	7	2.0×10^{-7}	0.03	0	0	15	0	2	0	0	2	3
Primary billary cirrhosis (Ref. 41)	GM12878	6	4	6.0×10^{-7}	0.03	0	11	41	0	0	0	0	8	8
Systemic lupus erythromatosus (Ref. 42)	GM12878	18	6	9.0×10^{-7}	0.03	0	4	21	0	5	8	0	3	5
Lipoprotein cholesterol/triglycerides (Ref. 43)	HepG2	18	5	1.2×10^{-6}	0.03	17	8	0	24	3	6	4	3	3
Hematological traits (Ref. 44)	K562	39	7	1.7×10^{-6}	0.03	0	12	10	2	1	0	0	1	0
Hematological parameters (Ref. 45)	K562	28	6	2.2×10^{-6}	0.03	0	15	7	0	5	7	7	3	2
Colorectal cancer (Ref. 46)	HepG2	4	3	3.8×10^{-6}	0.03	0	0	0	66	0	12	0	12	12
Blood pressure (Ref. 47)	K562	9	4	5.0×10^{-6}	0.04	0	30	14	0	10	6	7	5	11

Blood lipids

SNP	Chrom. Band	Gene	Link Score	Distance (kb)
rs2479409	1p32			
rs10761731	10q21			
rs1169288	12q24			
rs2072183	7p13			
rs2923084	11p15			
rs629301	1p13	CELSR2	2.1	26
rs4846914	1q42			
rs514230	1q42			
rs1367117	2p24	APOB	-	3
rs581080	9p22	TTC39B	-	2
rs3136441	11p11	F2	-	2
rs16942887	16q22	PSKH1	-	1
rs6511720	19p13	LDLR	-	2
rs737337	19p13	LOC55908	-	1
rs439401	19q13	APOC1	-	4
rs605066	6q24			
rs2000999	16q22	HPR	5.5	11
rs1800961	20q13	HNF4A	2.3	12
rs9686661	5q11			
rs2293889	8q23			
rs1883025	9q31	ABCA1	1.9	26
rs2068888	10q23	CYP26A1	1.6	6
rs4765127	12q24	ZNF664	-	2
rs838880	12q24			
rs6065906	20q13	CTSA	1.6	35
rs1689800	1q25			
rs2412710	15q15			
rs11649653	16p11			

Erythrocyte phenotypes

SNP	Chrom. Band	Gene	Link Score	Distance (kb)
rs11915082	3q29			
rs1122794	16p13	HBQ1	2.2	79
rs11085824	19p13	GCDH	-	1
rs643381	6q24			
rs9349205	6p21	BYSL	2.2	36
rs12718597	7p12	IKZF1	1.9	81
rs11065987	12q24	BRAP	1.8	51
rs4895441	6q23	HBS1L	1.7	51
rs628751	6q24			
rs10758658	9p24	RCL1	1.6	64
rs172629	4q12			
rs7776054	6q23	HBS1L	1.6	43
rs131794	22q13	ODF3B	-	1
rs9374080	6q21			
rs7786877	7q22	MOSPD3	-	4
rs7385804	7q22	TFR2	-	3

Systemic lupus erythematosus

SNP	Chrom. Band	Gene	Link Score	Distance (kb)
rs13385731	2p22			
rs10036748	5q33			
rs1385374	12q24	MGC16384	-	1
rs2230926	6q23	TNFAIP3	3.7	7
rs4728142	7q32	IRF5	-	4
rs9271100	6p21	HLA-DRB1	4.5	19
rs4917014	7p12	IKZF1	2.2	38
rs7812879	8p23	BLK	2.9	11
rs2205960	1q25			
rs548234	6q21			

Figure 5. Disease variants annotated by chromatin dynamics and regulatory predictions

a, Intersection of strong enhancer states (4,5) with disease-associated SNPs from GWAS studies shows significant enrichment (blue shading) in relevant cell types (see **Methods**). Fold-enrichments of the SNPs in strong enhancer states for each cell type are indicated. **b,** For three GWAS datasets[38]-[40], state annotations are shown for a subset of lead SNPs in the 9 cell types (colors as in Figure **1b**, except state 11 is white). Strong enhancer state (orange) is most prevalent in cell types related to the phenotype. For SNPs overlapping strong enhancers, proximal genes with correlated expression are indicated, with linking score and distance. **c,** Example GWAS locus with blood lipid traits[41] association, where the lead variant (red circle) has no functional annotation but a linked SNP (arrow) coincides with a HepG2-specific strong enhancer (orange), and may represent a causal variant. Strong enhancer annotations are shown for all cell types. **d,** Example GWAS loci where disease SNP affects a conserved instance of a predicted causal motif. Left: Lead SNP rs9374080 in the erythrocyte phenotype GWAS[38] is <100 bp from a strong enhancer in K562 erythroleukemia cells and strengthens a motif for Gfi1b, a predicted repressor in K562 (Fig. 3d). Right: SNP rs9271055 associated with lupus[39] coincides with a lymphoblastoid (GM12878)-specific strong enhancer and strengthens a motif for Ets1, a predicted activator of lymphoblastoid enhancers (Fig. 3d). This factor is further implicated by lupus-associated variants that directly affect the Ets1 locus[39].