ARTICLE

doi:10.1038/nature09906

Mapping and 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]* & Bradley E. Bernstein[1,3,4]*

Chromatin profiling has emerged as a powerful means of genome annotation and detection of regulatory activity. The approach is especially well suited to the characterization of non-coding portions of the genome, which critically contribute to cellular phenotypes yet remain largely uncharted. 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 multicell 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 the interpretation of genome-wide association studies. Top-scoring disease single nucleotide polymorphisms 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 suggesting a mechanism for the association. Our study presents a general framework for deciphering *cis*-regulatory connections and their roles in disease.

A major challenge in biology is understanding 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 of 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 nine chromatin marks and input controls across nine 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 the interpretation of genome-wide association studies (GWASs). 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 suggesting 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 data sets (Methods and 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 leukaemia 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 trimethylation (H3K4me3), a modification associated with promoters[4,5,9]; H3K4me2 (dimethylation), associated with promoters and enhancers[1,3,6,9]; H3K4me1 (methylation), 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, RNA polymerase II (RNAPII) and H2A.Z (also known as H2AFZ) in a subset of cells.

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [2]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA. [3]Howard Hughes Medical Institute, Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA. [4]Center for Systems Biology and Center for Cancer Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.
*These authors contributed equally to this work.

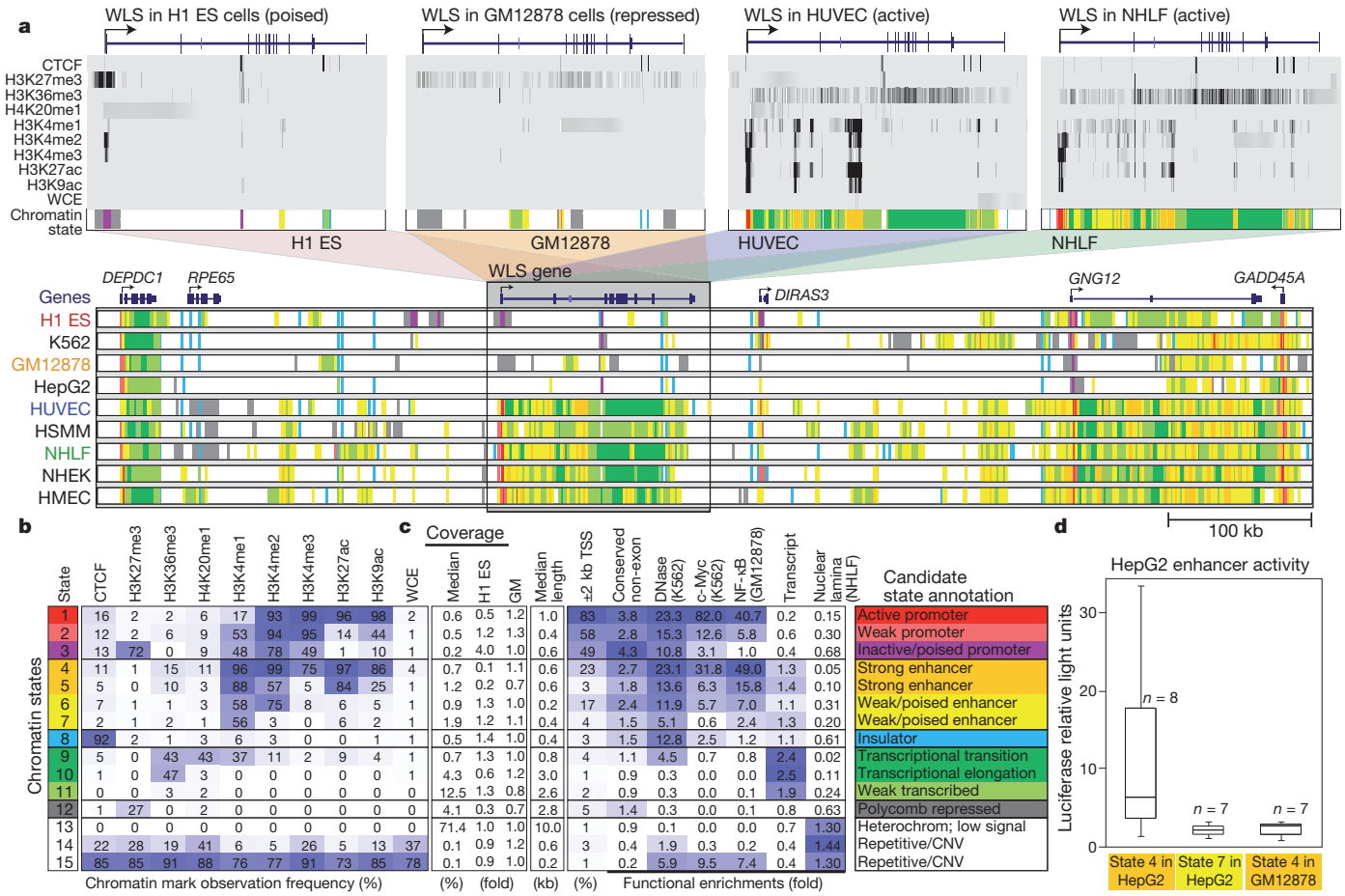

Figure 1 | Chromatin state discovery and characterization. **a**, Top: profiles for nine chromatin marks (greyscale) are shown across the WLS gene in four cell types, and summarized in a single chromatin state annotation track for each (coloured according to **b**). WLS is poised in ESCs, repressed in GM12878 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 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, whereas loss of H3K27me3 and gain of H3K27ac and H3K9ac mark promoter activation. WCE, whole-cell extract. Bottom: nine chromatin state tracks, one per cell type, in a 900-kb region centred at WLS, summarizing 90 chromatin tracks in directly interpretable dynamic annotations and showing activation and repression patterns for six genes and hundreds of regulatory regions, including enhancer states. **b**, Chromatin states learned jointly across cell types by a multivariate hidden Markov model. The table shows emission parameters learned *de novo* on the basis of 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. CNV, copy number variation; GM, GM12878. **d**, Box plots depicting enhancer activity for predicted regulatory elements. Sequences 250 bp long corresponding either to strong or weak/poised HepG2 enhancer elements or to GM12878-specific strong enhancer elements were inserted upstream of a luciferase gene and transfected into HepG2. 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. Boxes indicate 25th, 50th and 75th percentiles, and whiskers indicate 5th and 95th percentiles.

This resulted in 90 chromatin maps corresponding to ~2,400,000,000 reads covering ~100,000,000,000 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 data sets into nine readily interpretable annotations, one per cell type, we applied a multivariate hidden Markov model 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 neighbouring genomic segments in a set of 'transition' parameters (Methods). It has the advantage of capturing regulatory elements with greater reliability, robustness and precision than is possible by 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 that showed distinct biological enrichments and were consistently recovered (Fig. 1a, b and Supplementary Fig. 1). Even though states were learned *de novo* solely on the basis of 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] (MYC) 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 level, 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 Figs 2–4).

The states vary widely in their average segment length (~500 base pairs (bp) for promoter and enhancer states versus 10 kb for inactive

regions) and in the portion of the genome covered (<1% for promoter and enhancer states versus >70% for inactive state 13). For each state, coverage was relatively stable across cell types (Supplementary Fig. 5), with the exception of embryonic stem cells (ESCs) in which the poised promoter state is more abundant but 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 or 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 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. 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 Figs 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 Figs 11 and 12) suggest that regulatory regions vary drastically in activity level across cell types. Enhancer states show frequent interchange between strong and weak, and promoter states vary between active, weak and poised. Promoter states seem more stable than enhancers; they are eight times more likely to remain promoter states, controlling for coverage. Switching was also observed among promoter, enhancer and transcriptional transition states, but no preferential changes to other groups were found. 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 Figs 13 and 14). Cell-type-specific patterns were also evident when TSSs were simply assigned to the most prevalent chromatin state. Promoters active in skeletal muscle were associated with extracellular structure genes (8.5-fold enrichment), those active in lymphoblastoid cells were associated with immune response genes (7.2-fold enrichment) and those active in both were associated with metabolic housekeeping genes.

Clustering of promoter and enhancer states on the basis of their activity patterns

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

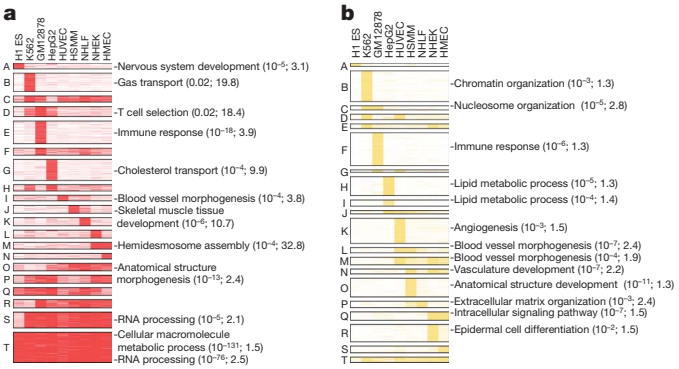

Figure 2 | Cell-type-specific promoter and enhancer states and associated functional enrichments. **a**, Clustering of genomic locations (rows) assigned to active promoter state 1 (red) across cell types (columns) reveals 20 common patterns of activity (A–T; Methods). For each cluster, enriched gene ontology terms are shown with hypergeometric *P* value and fold enrichment, based on the 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.

response (GM12878 specific, $P < 10^{-6}$), lipid metabolism (HepG2 specific, $P < 10^{-5}$) and angiogenesis (HUVEC specific, $P < 10^{-3}$).

Promoters and enhancers differed in their overall specificities. 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 seem to be 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 (for example immune genes and steroid metabolism genes) seem to be 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 vicinities.

Overall, this dynamic view of the chromatin landscape suggests that multicell 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 among 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 transcription factors recognizing each motif (Fig. 3e). We used correlations between these profiles to probabilistically link enhancers to their downstream targets and upstream regulators (Methods).

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

We validated our approach using quantitative trait locus mapping studies that use covariation between single nucleotide polymorphism (SNP) alleles and gene expression levels to link *cis*-regulatory regions

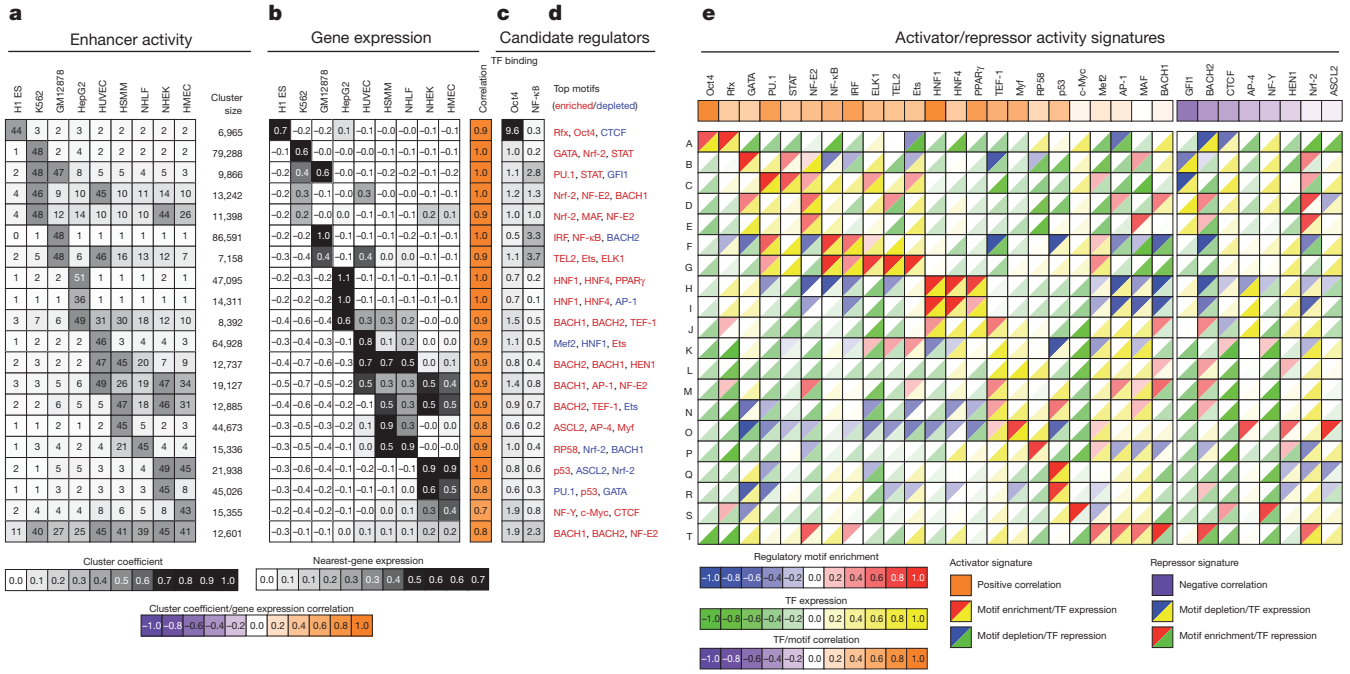

Figure 3 | Correlations in activity patterns link enhancers to gene targets and upstream regulators. **a**, Average enhancer activity across the cell types (columns) for each enhancer cluster (rows) defined in Fig. 2b (labelled A–T) and number of 200-bp windows in each cluster. **b**, Average messenger RNA 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 of linking enhancers to target genes. **c**, Enrichment for Oct4 binding in ESCs[24] and NF-κB binding in lymphoblastoid cells[14] for each cluster. TF, transcription factor. **d**, Strongly enriched (red) or depleted (blue) motifs for each cluster, from a catalogue of 323 consensus motifs. Rfx: Rfx family; Nrf-2: NFE2L2; STAT: STAT family; Ets: Ets family; Mef2: MEF2A and MYEF2; Myf: Myf family; NF-Y: NFYA, NFYB and NFYC. **e**, Predicted causal regulators for each cluster based on positive (activators) or negative (repressors) correlations between motif enrichment (top left triangles) and transcription factor expression (bottom right triangles). For example, the red–yellow combination indicates that Oct4 is a positive regulator of ESC-specific enhancers, as its motif-based predicted targets are enriched (red upper triangle) for enhancers active in ESCs (cluster A), and the Oct4 gene is expressed specifically in ESCs, resulting in a positive transcription factor expression correlation (yellow triangle). Overall correlations between motif enrichment and transcription factor expression across all clusters denote predicted activators (positive correlation, orange) and repressors (negative correlation, purple).

to target genes. Investigation of four recent quantitative trait locus 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 transcription factor expression and motif enrichment predict upstream regulators

We next predicted, on the basis of regulatory motif enrichments, sequence-specific transcription factors likely to target enhancers in a given cluster. This implicated a number of transcription factors whose known biological roles matched the respective cell types (Fig. 3d and Supplementary Fig. 18). When ChIP-seq data on the relevant cell type was available, we confirmed that enriched motifs were preferentially bound by the cognate factor (Fig. 3c). Oct4 (POU5F1) motif instances in cluster A (ESC-specific enhancers) were preferentially bound by Oct4 in ESCs[24], and NF-κB motif instances in cluster F (lymphoblastoid-specific enhancers) were preferentially bound by NF-κB in lymphoblastoid cells[14]. In both cases, motif instances in cell-type-specific enhancers showed a ~5-fold increase in binding in comparison with 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 transcription factors (Fig. 3e). We then correlated a 'motif score' based on motif enrichment in a given cluster, and a 'transcription factor expression score' based on the agreement between the transcription factor expression pattern and the cluster activity profile (Methods). A positive correlation between the two scores implies that the transcription factor may be establishing or reinforcing the chromatin state. A negative correlation would instead imply that the transcription factor may act as a repressor. For example, in addition to the enrichment of the Oct4 motif in the ESC-specific cluster A, Oct4 is specifically expressed in ESCs, leading to the prediction that it is a causal regulator of ESCs (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 (HNF1A), HNF4 (HNF4A) and PPARγ (PPARG) are predicted as activators of HepG2-specific enhancers (clusters H and I), PU.1 (SPI1) and NF-κB as activators of lymphoblastoid (GM12878) enhancers (clusters C, F and G), GATA1 as an activator of K562-specific enhancers (cluster B) and Myf family members as HSMM enhancers[14,25–27] (cluster O).

The analysis also revealed potentially novel regulatory interactions. ETS-related factors (ELK1, TEL2 (ETV7) and Ets family members) 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 (TP53) as an activator in HSMM, NHLF, NHEK and HMEC (clusters N, Q and R) probably reflects its maintained activity in these primary cells, as opposed to cell models in which it may be suppressed by mutation (K562)[29], viral inactivation (GM12878)[30] or cytoplasmic localization (ESCs)[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 (clusters B and C) and BACH2 in ESCs

(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 the silencing of satellite repeats[35]. Our regulatory inferences suggest that these regulators also modulate chromatin to inhibit enhancer activity, thus suggesting a new mechanism for distal gene regulation.

Validation of predicted binding events and regulatory outcomes

The regulatory inferences above imply transcription-factor-binding events at motif instances within enhancer regions in specific cellular contexts, and we sought to validate these inferences 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 these are indicative of transcription factor binding. We confirmed that dips were present in individual signal tracks for 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). Because dips only appear when the factor is expressed, they also support the identity of the *trans*-acting transcription factor.

To confirm 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 ten 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.

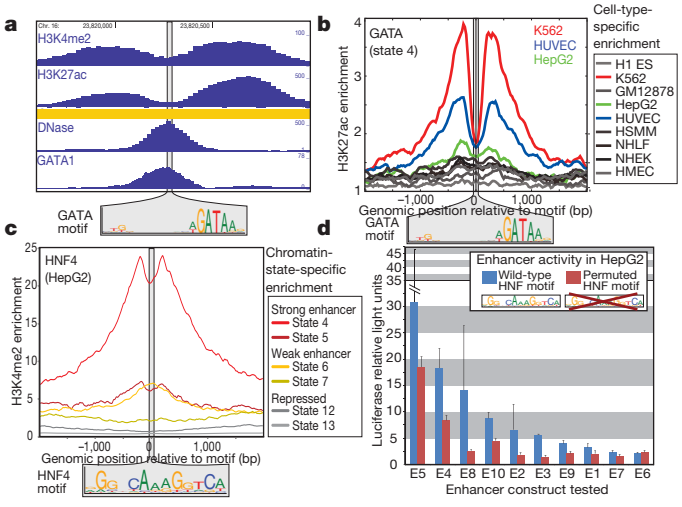

Figure 4 | Validation of regulatory predictions by nucleosome depletions and enhancer activity. **a**, Dips in chromatin intensity profiles in a K562-specific strong enhancer (orange) coincide with a predicted causal GATA motif instance (logo). The dips probably reflect nucleosome displacement associated with transcription factor binding, supported by DNase hypersensitivity[12] and GATA1 binding[25]. **b**, Superposition of H3K27ac signal across loci containing GATA motifs, centred on motif instances, shows dips in K562, as predicted. **c**, Superposition of H3K4me2 signal for HepG2 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 in comparison with wild type (blue), with an average twofold reduction. Data shown are mean luciferase relative light units over three replicates and 95% confidence intervals.

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 catalogues, an exceedingly small proportion of which are understood at present[37].

We found that disease-associated SNPs are significantly more likely to coincide with strong enhancers (states 4 and 5; twofold enrichment, $P < 10^{-10}$), despite the fact that no notable association with 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 data sets. We identified ten studies[38–47] whose variants showed significant correspondences to cell-type-specific strong enhancer states (Methods and 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 erythrocytic leukaemia cell (K562) enhancers, SNPs associated with systemic lupus erythematosus[39] were found in lymphoblastoid cell (GM12878) enhancers and 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 leukaemia[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 in which 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 of triaging variants within a haplotype block, a common problem faced in GWASs.

Lastly, we identified several cases in which a disease-associated SNP created or disrupted a regulatory motif instance for a predicted causal transcription factor 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 demonstrates the power of multicell chromatin profiles 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 predict *cis*-regulatory interactions that drive gene expression programs. By intersecting our predictions with non-coding SNPs from GWAS data sets, we proposed 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 drastically reduced the large combinatorial space of 90 chromatin data sets (2^{90} combinations) to a manageable set of biologically interpretable annotations, thus providing an efficient and robust way to track coordinated changes across cell types. This allowed the systematic identification and comparison of more than 100,000 promoter and enhancer elements. Both types of element are cell type specific, are associated with motif enrichments and assume

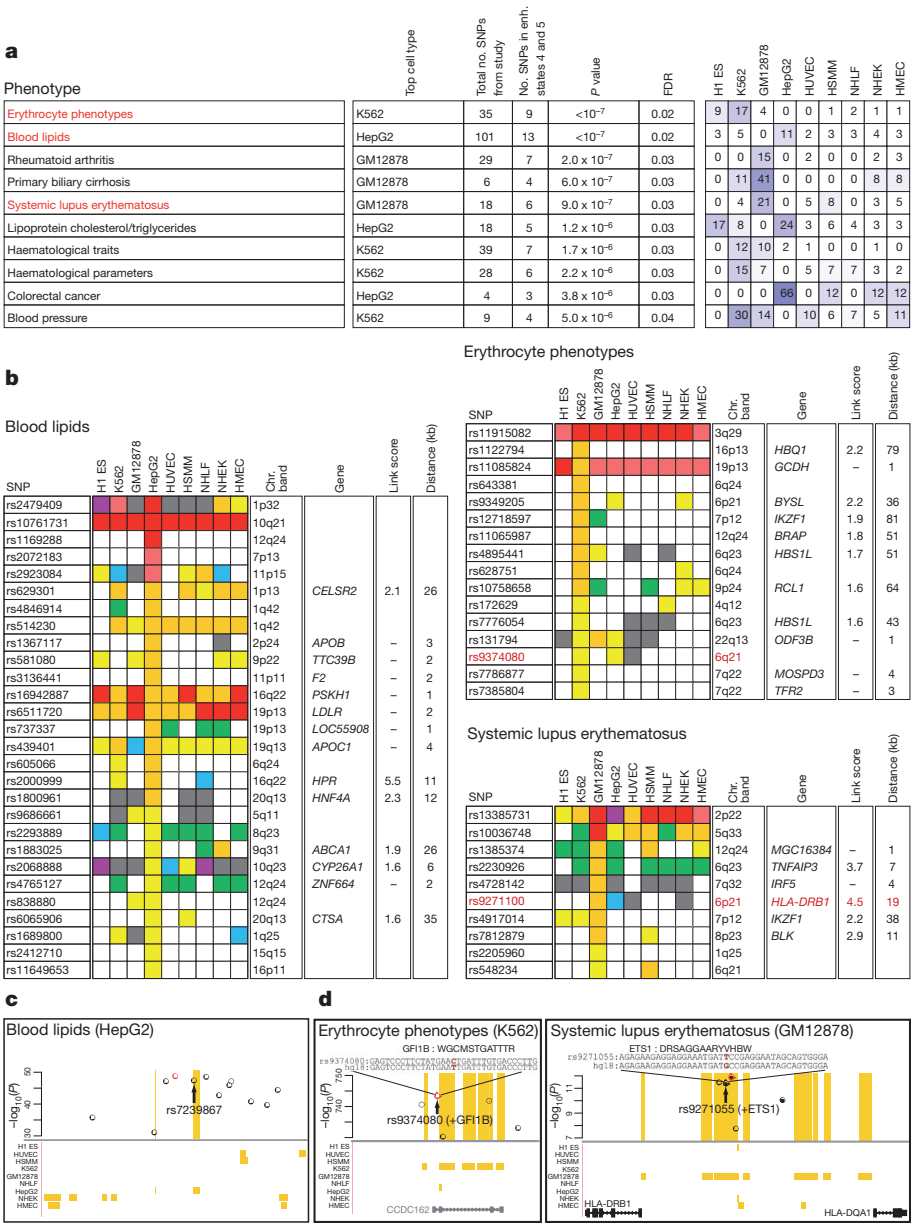

Figure 5 | Disease variants annotated by chromatin dynamics and regulatory predictions. **a**, Intersection of strong enhancer states (4 and 5) with disease-associated SNPs from GWASs shows significant enrichment (blue shading) in relevant cell types (Methods). Fold enrichments of the SNPs in strong enhancer states for each cell type are indicated. FDR, false-discovery rate. **b**, For three GWAS data sets[38–40], state annotations are shown for a subset of lead SNPs in the nine cell types (colours as in Fig. 1b, except state 11 is white). The 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 trait[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 a 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 and strengthens a motif for GFI1B, a predicted repressor in K562 (Fig. 3e). Right: SNP rs9271055 associated with lupus[39] coincides with a lymphoblastoid (GM12878) strong enhancer and strengthens a motif for ETS1, a predicted activator of lymphoblastoid enhancers (Fig. 3e). This factor is further implicated by lupus-associated variants that directly affect the *ETS1* locus[39].

strong, weak and poised states that correlate with neighbouring gene expression and function. Enhancers showed very high 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 should 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 quantitative trait locus mapping studies. Predicted transcription factor/motif interactions within cell-type-specific enhancers were confirmed in specific cases by transcription factor 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 multicell chromatin profiles are unique and highly complementary to data sets for transcription factor binding, expression, chromatin accessibility, nucleosome positioning and chromosome conformation[50]. For example, our regulatory predictions can help focus the spectrum of transcription-factor-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 data sets 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, suggest *cis*- and *trans*-acting interactions that may be modulated by the sequence variants. Although the present study represents only a first, small step in this direction, we expect that future iterations with a greater diversity of cell types and improved methodologies will help define the molecular underpinnings of human disease.

METHODS SUMMARY

We performed ChIP-seq analysis in biological replicate as previously described[4], using antibodies validated by western blots and peptide competitions. ChIP DNA and input controls were sequenced using the Illumina Genome Analyser. Expression profiles were acquired using Affymetrix GeneChip arrays. Chromatin states were learned jointly by applying a hidden Markov model[8] to ten data tracks for each of the nine 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 nine genome-wide chromatin state annotations, which were validated by additional ChIP experiments and reporter assays. Multicell 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. We predicted enhancer/target gene linkages 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 transcription factor motifs assembled from multiple sources. Motif instances were identified by sequence match and evolutionary conservation. We based *P* values for GWAS studies on randomizing the location of SNPs, and based the false-discovery rate on randomizing the assignment of SNPs across studies. Data sets 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).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 August 2010; accepted 4 February 2011.
Published online 23 March 2011.

1. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
2. Kim, H. D., Shay, T., O'Shea, E. K. & Regev, A. Transcriptional regulatory circuits: predicting numbers from alphabets. *Science* **325,** 429–432 (2009).
3. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837 (2007).
4. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).
5. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130,** 77–88 (2007).
6. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genet.* **39,** 311–318 (2007).
7. Hon, G., Wang, W. & Ren, B. Discovery and annotation of functional chromatin signatures in the human genome. *PLOS Comput. Biol.* **5,** e1000566 (2009).
8. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnol.* **28,** 817–825 (2010).
9. Bernstein, B. E. *et al.* Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120,** 169–181 (2005).
10. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–112 (2009).
11. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137,** 1194–1211 (2009).
12. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl Acad. Sci. USA* **107,** 139–144 (2010).
13. Raha, D. *et al.* Close association of RNA polymerase II and many transcription factors with Pol III genes. *Proc. Natl Acad. Sci. USA* **107,** 3639–3644 (2010).
14. Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science* **328,** 232–235 (2010).
15. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453,** 948–951 (2008).
16. Jaenisch, R. & Young, R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell* **132,** 567–582 (2008).
17. De Santa, F. *et al.* A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol.* **8,** e1000384 (2010).
18. Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465,** 182–187 (2010).
19. Talbert, P. B. & Henikoff, S. Histone variants — ancient wrap artists of the epigenome. *Nature Rev. Mol. Cell Biol.* **11,** 264–275 (2010).
20. Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6,** e107 (2008).
21. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).
22. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464,** 773–777 (2010).
23. Veyrieras, J. B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* **4,** e1000214 (2008).
24. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genet.* **42,** 631–634 (2010).
25. Fujiwara, T. *et al.* Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. *Mol. Cell* **36,** 667–681 (2009).
26. Lemaigre, F. & Zaret, K. S. Liver development update: new embryo models, cell lineage control, and morphogenesis. *Curr. Opin. Genet. Dev.* **14,** 582–590 (2004).
27. Sabourin, L. A. & Rudnicki, M. A. The molecular regulation of myogenesis. *Clin. Genet.* **57,** 16–25 (2000).
28. Bartel, F. O., Higuchi, T. & Spyropoulos, D. D. Mouse models in the study of the Ets family of transcription factors. *Oncogene* **19,** 6443–6454 (2000).
29. Law, J. C., Ritke, M. K., Yalowich, J. C., Leder, G. H. & Ferrell, R. E. Mutational inactivation of the p53 gene in the human erythroid leukemic K562 cell line. *Leuk. Res.* **17,** 1045–1050 (1993).
30. Forte, E. & Luftig, M. A. MDM2-dependent inhibition of p53 is required for Epstein-Barr virus B-cell growth transformation and infected-cell survival. *J. Virol.* **83,** 2491–2499 (2009).
31. Solozobova, V., Rolletschek, A. & Blattner, C. Nuclear accumulation and activation of p53 in embryonic stem cells after DNA damage. *BMC Cell Biol.* **10,** 46 (2009).
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* **116,** 499–509 (2004).
33. Wei, C. L. *et al.* A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124,** 207–219 (2006).
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.* **141,** 719–727 (2007).
35. Vassen, L., Fiolka, K. & Moroy, T. Gfi1b alters histone methylation at target gene promoters and sites of gamma-satellite containing heterochromatin. *EMBO J.* **25,** 2409–2419 (2006).
36. He, H. H. *et al.* Nucleosome dynamics define transcriptional enhancers. *Nature Genet.* **42,** 343–347 (2010).
37. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106,** 9362–9367 (2009).
38. Ganesh, S. K. *et al.* Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nature Genet.* **41,** 1191–1198 (2009).
39. Han, J. W. *et al.* Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nature Genet.* **41,** 1234–1237 (2009).
40. Kathiresan, S. *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature Genet.* **40,** 189–197 (2008).
41. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713 (2010).
42. Houlston, R. S. *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature Genet.* **40,** 1426–1435 (2008).
43. Newton-Cheh, C. *et al.* Genome-wide association study identifies eight loci associated with blood pressure. *Nature Genet.* **41,** 666–676 (2009).
44. Stahl, E. A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genet.* **42,** 508–514 (2010).
45. Liu, X. *et al.* Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis. *Nature Genet.* **42,** 658–660 (2010).
46. Kamatani, Y. *et al.* Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nature Genet.* **42,** 210–215 (2010).
47. Soranzo, N. *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nature Genet.* **41,** 1182–1190 (2009).
48. Papaemmanuil, E. *et al.* Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nature Genet.* **41,** 1006–1010 (2009).
49. Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461,** 199–205 (2009).
50. Naumova, N. & Dekker, J. Integrating one-dimensional and three-dimensional maps of genomes. *J. Cell Sci.* **123,** 1979–1988 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank members of the epigenomics community at the Broad Institute and the Bernstein and Kellis laboratories, and M. Daly, D. Altshuler and E. Lander for discussions and criticisms. We also thank M. Suva, E. Mendenhall and S. Gillespie for assistance with experiments, and L. Goff and A. 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; B.E.B.), R01 HG004037 (M. Kellis), RC1 HG005334 (M. Kellis), the Howard Hughes Medical Institute (B.E.B.), the National Science Foundation (awards 0644282 (M. Kellis) and 0905968 (J.E.)) and the Sloan Foundation (M. Kellis).

Author Contributions J.E. conducted chromatin state analysis. J.E. and P.K. conducted regulatory motif analysis. J.E. and L.W. conducted GWAS SNP analysis. T.S.M., N.S. and T.D. implemented the ChIP-seq data processing pipeline. C.B.E., X.Z., L.W., R.I., M.C. and M. Ku developed the experimental pipeline and conducted experiments. M. Kellis designed and directed the computational analysis. B.E.B. designed the experimental approach and oversaw the work. J.E., M. Kellis and B.E.B. wrote the paper.

Author Information Sequencing and expression data has been deposited into the Gene Expression Omnibus under accession number GSE26386. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M. Kellis (manoli@mit.edu).

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 collected at a passage number between 30 and 40. Before collection, cells were karyotyped and stained for Oct4 to confirm pluripotency. K562 erythrocytic leukaemia cells (ATCC CCL-243, lot no. 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 at between 3×10^5 and 7×10^5 cells ml^{-1}. 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^{-1} with minimal disruption, and maintained at between 3×10^5 and 7×10^5 cells ml^{-1}. HepG2 hepatocellular carcinoma cells (ATCC HB-8065, lot no. 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 of between 15 and 20% and then collected at ~75% confluence. NHEK normal human epidermal keratinocytes isolated from skin (Lonza CC-2501, lot no. 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 the recommended density (3,500 cells cm^{-2}), subjected to two or three passages on polystyrene tissue culture plates and collected at a confluence of 70 to 80%. HSMM primary human skeletal muscle myoblasts (Lonza CC-2580, lot no. 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 the recommended density (3,500 cells cm^{-2}), subjected to two or three passages and collected at a confluence of 50 to 70%. NHLF primary normal human lung fibroblasts (Lonza CC-2512, lot no. 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 the recommended density (2,500 cells cm^{-2}), subjected to two or three passages and collected at an approximate confluence of 80%. HUVEC primary human umbilical vein endothelial cells (Lonza CC-2517, lot no. 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 the recommended density (2,500–5,000 cells cm^{-2}), subjected to two or three passages and collected 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 at the recommended density (2,500 cells cm^{-2}), subjected to two or three passages and collected 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. Cells were harvested by crosslinking 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 $\sim 10^7$ 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). Immunoprecipitation was performed as previously described, retaining a fraction of input 'whole-cell extract' as a control[4]. Briefly, sonicated chromatin was diluted tenfold and incubated with ~5 μg antibody overnight. Antibody–chromatin complexes were pulled-down using protein A sepharose, washed and then eluted. After crosslink 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^{-1}, NEB), ligated to standard Illumina adapters (75 bp with a 'T' overhang) using DNA ligase (0.05 U μl^{-1}, 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[52] (NanoString nCounter) to confirm representation and specific enrichment of DNA species. Libraries were sequenced in one or two lanes on the Illumina Genome Analyser 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 data sets 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/maq-man.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 previously described[4], and were then combined into a single data set. For the hidden Markov model (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 200-bp resolution on the basis of a Poisson background model using a threshold of 10^{-4}.

Joint learning of HMM states across cell types. To handle data from the nine 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 as previously described[8]. Here we used a Euclidean distance for determining initial parameters for the nested initialization step. After the HMM had learned and evaluated 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 ESCs[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 on the basis of the hypergeometric distribution and corrected for multiple testing using Bonferroni 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 Fig. 2a reflect an aggregation across all 72 possible pairs of cell types.

Pairwise promoter clustering. Promoters for RefSeq genes were clustered on the basis of the most likely chromatin state assignment across a 2-kb region centred 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 in which each component corresponded to a position along the promoter, cell type and state. Clustering was performed on these vectors using the *k*-means algorithm in MATLAB. Gene expression values were calculated on the basis of the corresponding Affymetrix probe set closest to the TSS.

Multicell 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 whose components are the estimated probabilities that it be in the strong promoter state for each of the nine 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 using the *k*-means algorithm in 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 Fig. 2, the locations were randomly down-sampled. For the purpose of identifying enriched functional gene categories in Fig. 2b, enhancers were linked to the nearest TSS up to 50 kb distant excluding those within 5 kb. Enhancer–gene correspondences based on the nearest gene were used for the expression analysis of distance-based linked genes in Fig. 3b.

Linking enhancer locations to correlated genes. To predict linkages between enhancer states and target genes, we combined distance-based information with multicell type correlations between gene expression levels and normalized signal intensities for histone modifications associated with enhancer states (H3K4me1, H3K4me2 and H3K27ac). For each enhancer state (4–7), cell type, and 200-bp interval between 5 kb and 125 kb 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. So that the classifiers learned a smooth and robust function at each position, we included as part of the training all enhancer state assignments within a 10-kb window centred 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 threshold of 2.5. The expression QTL data was obtained from the University of Chicago QTL browser (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/). In the QTL evaluation, each SNP that overlapped a strong enhancer state (4 or 5), was within 125 kb of a TSS, excluding locations within 5 kb, 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 on the basis of 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 database, we used the same link score threshold and compared gene assignments against the distance-based assignments defined above. The base set of genes in the enrichment analysis here were all genes that could be linked in at least one cluster.

Motif and transcription factor analysis. A database of known transcription factor motifs was collated by combining motifs from TRANSFAC[59] (version 11.3), JASPAR[60] (2010-05-07) and protein-binding microarray data sets[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 (K. Lindblad-Toh *et al.*, submitted). These were filtered using a significance threshold of $P < 4^{-8}$ for the motifs[64], and a confidence level based on conservation. Motifs were linked to corresponding transcription factors using metadata provided by the source. Motif enrichments for chromatin state clusters were computed as ratios to the instances of shuffled motifs, 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 transcription factor, the one that showed the most variance in enrichment across clusters was selected.

For predicting causal activators and repressors, motif scores and transcription factor expression scores were correlated as follows. Motif scores were calculated as described above. Transcription factor expression scores were calculated for each cluster by correlating the expression of the transcription factor 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 transcription factor expression scores were then correlated against each other to identify positively and negatively correlated transcription factors.

Transcription factor/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 previously described[36]. Superposition plots show tag enrichments relative to a uniform background computed on the basis of 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 three 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, 250-bp sequences from the reference genome (hg18) corresponding to different chromatin states (eight from HepG2 state 4, seven from HepG2 state 7 and seven 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 per 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 h 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 transcription factor motifs, ten sequences of 250 bp corresponding to HepG2-specific strong enhancers (state 4) with dips and HNF motifs were tested as above, and compared with 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 catalogue and the UCSC browser[37,54] (October 30, 2010). This set was refined by extending the blood lipid GWAS[41] set to contain all reported SNPs, and by bifurcating the haematological and biochemical traits study[46] into a haematological traits set and a biochemical traits set. We limited our analysis to studies reporting two 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 by using 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 *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 on the basis of the lead cell type using the linking method described above.

Data access. Data sets 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).

51. Ludwig, T. E. *et al.* Feeder-independent culture of human embryonic stem cells. *Nature Methods* **3,** 637–646 (2006).
52. Geiss, G. K. *et al.* Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature Biotechnol.* **26,** 317–325 (2008).
53. Reich, M. *et al.* GenePattern 2.0. *Nature Genet.* **38,** 500–501 (2006).
54. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
55. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35,** D61–D65 (2007).
56. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
57. Ernst, J. & Bar-Joseph, Z. STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics* **7,** 191 (2006).
58. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
59. Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31,** 374–378 (2003).
60. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32,** D91–D94 (2004).
61. Berger, M. F. *et al.* Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133,** 1266–1276 (2008).
62. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324,** 1720–1723 (2009).
63. Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnol.* **24,** 1429–1435 (2006).
64. Touzet, H. & Varre, J. S. Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms Mol. Biol.* **2,** 15 (2007).
65. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–326 (2006).