nature genetics

Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome

Nathaniel D Heintzman[1,2], Rhona K Stuart[1], Gary Hon[1,3], Yutao Fu[4], Christina W Ching[1], R David Hawkins[1], Leah O Barrera[1,3], Sara Van Calcar[1], Chunxu Qu[1], Keith A Ching[1], Wei Wang[5], Zhiping Weng[4,6], Roland D Green[7], Gregory E Crawford[8] & Bing Ren[1,9]

Eukaryotic gene transcription is accompanied by acetylation and methylation of nucleosomes near promoters, but the locations and roles of histone modifications elsewhere in the genome remain unclear. We determined the chromatin modification states in high resolution along 30 Mb of the human genome and found that active promoters are marked by trimethylation of Lys4 of histone H3 (H3K4), whereas enhancers are marked by monomethylation, but not trimethylation, of H3K4. We developed computational algorithms using these distinct chromatin signatures to identify new regulatory elements, predicting over 200 promoters and 400 enhancers within the 30-Mb region. This approach accurately predicted the location and function of independently identified regulatory elements with high sensitivity and specificity and uncovered a novel functional enhancer for the carnitine transporter *SLC22A5* (*OCTN2*). Our results give insight into the connections between chromatin modifications and transcriptional regulatory activity and provide a new tool for the functional annotation of the human genome.

Activation of eukaryotic gene transcription involves the coordination of a multitude of transcription factors and cofactors on regulatory DNA sequences such as promoters and enhancers and on the chromatin structure containing these elements[1–3]. Promoters are located at the 5′ ends of genes immediately surrounding the transcriptional start site (TSS) and serve as the point of assembly of the transcriptional machinery and initiation of transcription[4]. Enhancers contribute to the activation of their target genes from positions upstream, downstream or within a target or neighboring gene[5,6]. Deciphering the regulatory information encoded in the genome will require a thorough understanding of the relationships between the transcriptional activities of these different types of *cis*-regulatory sequence elements and the epigenetic features of the chromatin surrounding them. Significant progress in the fields of epigenetics and chromatin biology suggests a histone code[7] of ever-increasing complexity with profound implications for chromatin as both a receptive substrate and a predictive signal in a variety of biological processes[3,8].

Recent investigations using chromatin immunoprecipitation (ChIP) and microarray (ChIP-chip) experiments have described the chromatin architecture of transcriptional promoters in yeast, fly and mammalian systems[9]. In a manner largely conserved across species, active promoters are marked by acetylation of various residues of histones H3 and H4 and methylation of H3K4, particularly trimethylation of this residue. Nucleosome depletion is also a general characteristic of active promoters in yeast and flies, although this feature remains to be thoroughly examined in mammalian systems. Although some studies suggest that distal regulatory elements like enhancers may be marked by similar histone modification patterns[10–13], the distinguishing chromatin features of promoters and enhancers have yet to be determined, hindering our understanding of a predictive histone code for different classes of regulatory elements. Here, we present high-resolution maps of multiple histone modifications and transcriptional regulators in 30 Mb of the human genome, demonstrating that active promoters and enhancers are associated with distinct chromatin signatures that can be used to predict these regulatory elements in the human genome.

RESULTS

Chromatin architecture and transcription factor localization

We performed ChIP-chip analysis[14] to determine the chromatin architecture along 44 human loci selected by the ENCODE consortium as common targets for genomic analysis[15], totaling 30 Mb.

[1]Ludwig Institute for Cancer Research, [2]Biomedical Sciences Graduate Program, [3]Program in Bioinformatics and [9]Department of Cellular and Molecular Medicine, University of California San Diego (UCSD) School of Medicine, 9500 Gilman Drive, La Jolla, California 92093-0653 USA. [4]Bioinformatics Program, Boston University, 24 Cummington Street, 1002, Boston, Massachusetts 02215 USA. [5]Department of Chemistry and Biochemistry, UCSD, 9500 Gilman Drive, La Jolla, California 92093 USA. [6]Biomedical Engineering Department, Boston University, 44 Cummington Street, Boston, MA 02215 [7]NimbleGen Systems, Inc., 1 Science Court, Madison, Wisconsin 53711 USA. [8]Institute for Genome Sciences & Policy and Department of Pediatrics, Duke University, 101 Science Drive, Durham, North Carolina 27708, USA. Correspondence should be addressed to B.R. (biren@ucsd.edu).

Received 27 September 2006; accepted 3 January 2007; published online 4 February 2007; doi:10.1038/ng1966

We investigated the patterns of core histone H3 and five histone modifications: acetylated H3K9/14, acetylated H4K5/8/12/16 and mono-, di-, and trimethylated histone H3K4. We also examined binding of two components of the basal transcriptional machinery, RNA polymerase II (RNAPII) and TBP-associated factor 1 (TAF1), and the transcriptional coactivator p300 to identify active promoters and enhancers, respectively. We carried out three biological replicate ChIP-chip experiments for each marker in HeLa cells before and after treatment with interferon-gamma (IFNγ), as p300 is known to be involved in the cellular response to this cytokine[16]. We amplified, labeled and hybridized ChIP samples to tiling oligonucleotide microarrays covering the nonrepetitive sequences of 30 Mb at 38-bp resolution. We analyzed the microarray data by standard methods to determine average enrichments for each marker at every probe, generating high-resolution maps of histone modifications and transcriptional regulator binding for 1% of the human genome. To validate our ChIP-chip results, we performed conventional ChIP against RNAPII and tested for enrichment at 121 sites in the ENCODE regions using quantitative real-time PCR, which indicated an accuracy of 97%, a specificity of near 100% and a sensitivity of 82% for our method (see **Supplementary Methods** and **Supplementary Table 1** online). These values are comparable to other ChIP-chip studies[12,17,18] and confirm that our ChIP-chip data are very reliable.

Chromatin signatures of promoters

To explore chromatin features at human promoters, we examined ChIP-chip profiles along 10-kb regions surrounding well-annotated promoters in the ENCODE regions and performed computational clustering to classify each promoter on the basis of histone modification patterns. We examined only TSSs corresponding to well-annotated RefSeq[19] transcripts for which we had collected expression data, and to prevent interference from neighboring genes, we excluded TSSs within 10 kb of each other, resulting in a pool of 208 TSSs for clustering; 104 TSSs are defined as active promoters and 104 as inactive promoters by gene expression profiling experiments. We observed four distinct classes of promoters (P1–P4) in untreated HeLa cells, arranged by the proportion of active promoters within each class (**Fig. 1a** and **Supplementary Table 2** online). Expression of transcripts within each class generally increased from class P1 to class P4, as most of the inactive promoters were found in class P1, whereas classes P2–P4 were increasingly composed of active promoters. Average enrichment profiles for each marker within each class (**Fig. 1b**) showed that occupancy by all five histone modifications, RNAPII and TAF1 increased at active promoters in a manner related to gene expression levels. Moderate p300 enrichment was also present at many active promoters (see **Supplementary Fig. 1** for a representative active promoter), whereas the largely inactive class P1 was devoid of any markers. The patterns observed in HeLa cells treated with IFNγ were almost identical (**Supplementary Fig. 2**). The transition from trimethylated to dimethylated to monomethylated H3K4 moving downstream from active promoters into coding regions echoes the pattern seen in small-scale studies in human cells[20] and globally in yeast[17,21]. These results confirm previous observations in other organisms that histone modifications are linked to promoter activity.

We were interested to find a bimodal distribution of all histone modifications centered around peak binding of RNAPII and TAF1 at the TSS, implying depletion of nucleosomes at this position. ChIP-chip data for histone H3 supported this conclusion (**Fig. 1a**,**b**). Our findings indicate that the nucleosome-free region (NFR) observed at promoters in yeast and fly is indeed characteristic of active human promoters, supporting an evolutionarily constrained role for this phenomenon in transcriptional regulation. The degree of nucleosome depletion seems to be related to the level of gene expression, as we did not observe depletion in class P1, suggesting that the formation and maintenance of NFRs at active promoters is a regulated process. Distribution around the NFR varied among the histone modifications and promoter classes, but most modifications were found on both sides of the NFR with an asymmetrical bias toward the region immediately downstream, particularly for trimethylated H3K4. Acetylated H4 was the only outlier to this trend. The observed histone acetylation and methylation at nucleosomes upstream of the TSS may represent signatures of chromatin architecture at promoters that are specific to mammals.

Chromatin signatures of enhancers

Next, we investigated the chromatin features of human transcriptional enhancers. As previous studies have demonstrated that p300 and related acetyltransferases are present at enhancers (as well as promoters)[10,11], we identified genomic regions in HeLa cells enriched in p300 binding (124 binding sites in untreated cells and 182 sites in treated cells, listed in **Supplementary Table 3**) and found that the p300 binding sites demonstrated several features of enhancers. First, the genomic distribution of p300 sites was consistent with the widespread location of enhancers relative to their target genes, as over 75% of p300 binding occurs more than 2.5 kb from known 5′ ends of genes (as determined by GENCODE[22]) (**Supplementary Fig. 3** online). Second, transcriptional regulatory elements such as enhancers have long been known to show increased nuclease sensitivity[23], so we mapped the DNaseI hypersensitive sites (DHSs) in HeLa cells along the ENCODE regions using a recently developed DNase-chip method[24] (**Supplementary Table 4** online) and found that a significant number of distal p300 sites (69.7%, $P < 1 \times 10^{-16}$) overlapped with DHSs, representing ~12% of the distal DHSs we identified (**Supplementary Fig. 3**). Third, most distal p300 sites were conserved across species; over 60% ($P < 1 \times 10^{-16}$) of these sites contained strongly conserved sequence (**Supplementary Methods**). Fourth, a significant number of the distal p300 sites (44.4%, $P = 4.6 \times 10^{-15}$) contained independently predicted regulatory modules (PReMods) identified based on clustering of conserved transcription factor binding motifs[25] (**Supplementary Methods**). These lines of evidence provide strong support that the distal p300 binding sites represent a subset of enhancers.

Using the distal p300 binding sites to anchor 10-kb regions surrounding each putative enhancer, we performed computational clustering as described above to generate three classes of enhancers (**Fig. 1c**,**d**; classes are arbitrarily named E1–E3 to simplify discussion). We were interested to find that monomethylated H3K4 was strongly enriched in a broad pattern at nearly all enhancers. Here we also found depletion of histone H3 at enhancers, suggesting that nucleosome depletion is a general feature of both promoters and enhancers, consistent with their DNaseI hypersensitivity. Although most active promoters are marked by substantial enrichment of trimethylated H3K4 at the TSS, enhancers generally lack this histone modification. Furthermore, active promoters showed a marked depletion of monomethylated H3K4 at the TSS and enrichment of this modification more than 1 kb downstream and upstream, whereas enhancers showed strong enrichment of monomethylated H3K4 at the peak of p300 binding. Acetylated H4, acetylated H3 and dimethylated H3K4 were present in varying degrees at both promoters and enhancers, although the bimodal distribution of these modifications observed at active promoters was less pronounced at enhancers. TAF1 and RNAPII were also present at some enhancers, albeit more weakly than at promoters (reminiscent of the weak p300 enrichment at promoters seen in **Fig. 1a**,**b**), suggesting docking of the transcriptional machinery at

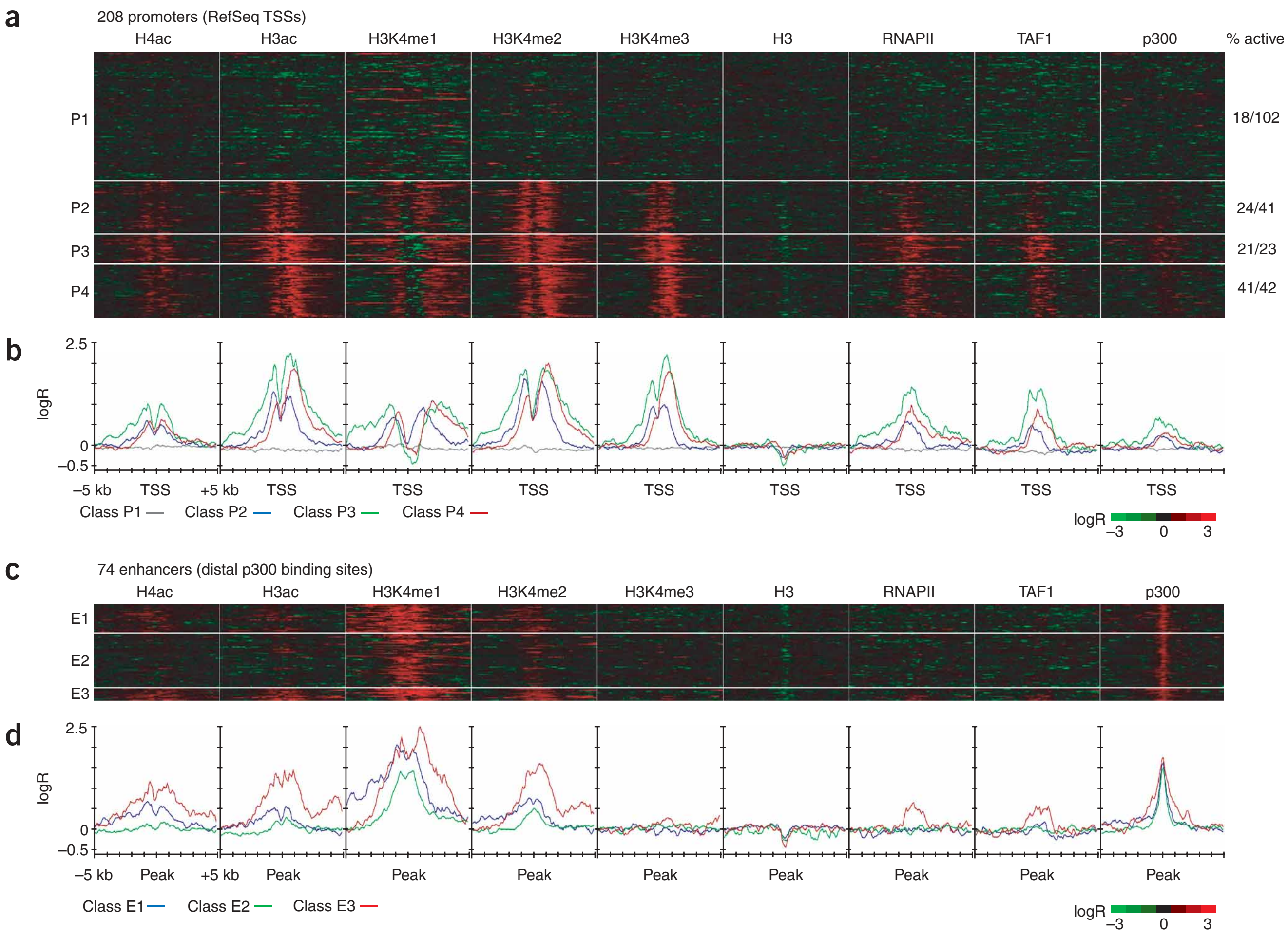

Figure 1 Features of human transcriptional promoters and enhancers. ChIP-chip was performed against six histone markers and three general transcription factors in the ENCODE regions, and the data were clustered to reveal patterns at annotated promoters (**a**) and distal p300 binding sites (**c**). Promoter clustering was performed with 10-kb windows centered on RefSeq TSSs; enhancer windows were centered on promoter-distal p300 binding peaks. Average profiles for each marker within each class are shown below the clusters (**b**,**d**); each class is represented by a different color. The proportion of expressed genes ('% active') in each promoter class is presented to the right of the cluster, illustrating the relationship between histone modification patterns and gene expression. Comparison of the clusters shows that active promoters and enhancers are similarly marked by nucleosome depletion (column H3) but distinctly marked by mono- and trimethylation of histone H3K4 (columns H3K4me1 and H3K4me3). Note the depletion of H3K4me1 and the peak of H3K4me3 at the TSS in promoters, compared with the enrichment of H3K4me1 and lack of H3K4me3 at enhancers. The presence of histone methylation and acetylation upstream and downstream of the TSS at promoters is distinct from the primarily downstream localization of these markers observed at yeast promoters. The same procedure was performed on data from treated HeLa cells, yielding similar results (**Supplementary Fig. 2**). H4ac, acetylated H4; H3ac, acetylated H3.

enhancers or physical interaction between enhancers and active promoters as proposed in various models of enhancer action[5,6]. In spite of some similarities between the histone modification profiles of active promoters and enhancers, the sharp contrasts of their monomethylated H3K4 and trimethylated H3K4 profiles represent distinct chromatin signatures for these different classes of regulatory elements.

Predicting promoters and enhancers via chromatin signatures

Next, we investigated the possibility that the different chromatin signatures of active promoters and enhancers could be used to predict these transcriptional regulatory elements in the human genome. We constructed training sets with histone modification profiles surrounding known TSSs and p300 binding sites in untreated HeLa cells and used them to develop a computational prediction algorithm to locate promoters and enhancers in the ENCODE regions based on similarity to the training set chromatin profiles (**Fig. 2a** and **Supplementary Methods**). Our two-stage method of regulatory element identification consists of a primary descriptive prediction followed by secondary discriminative filters (**Supplementary Methods**). To qualify as a high-confidence predicted regulatory element, a region of chromatin must unambiguously match one of the training set profiles.

A total of 198 active promoters (**Supplementary Table 5** online) were predicted in the ENCODE regions in untreated HeLa cells, clustered, as described previously, into four classes (named PI–PIV to distinguish them from the known promoters presented in **Fig. 1**) (**Fig. 2b**). In HeLa cells treated with IFNγ, we predicted 208 promoters (**Supplementary Table 5**), with greater than 90% overlap between the untreated and treated prediction sets (**Fig. 2c**), supporting the accuracy of our method in identifying promoters in an independent data set. The untreated prediction set contained 140 (79%) of the 177 active RefSeq promoters within the ENCODE regions and 32 (21%) of 155 inactive RefSeq promoters, and 180 predictions (91%) mapped to known GENCODE 5′ ends of genes (**Fig. 2d**), indicating a high degree of sensitivity and accuracy of promoter prediction. Promoter predictions in treated cells

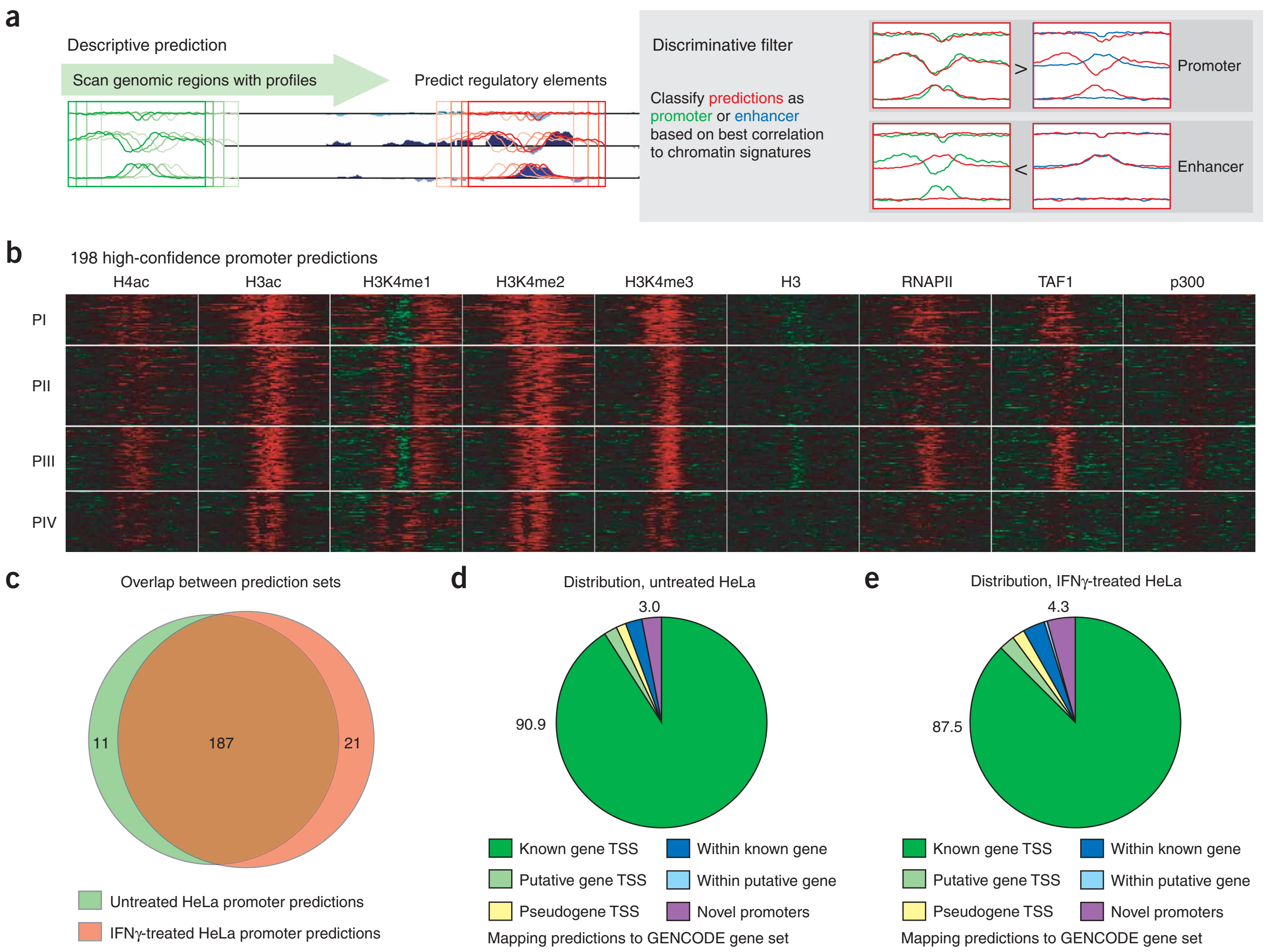

Figure 2 Prediction of promoters based on chromatin signatures. (**a**) A general scheme of the prediction method. Left: features of established transcriptional promoters (and enhancers) were analyzed to yield descriptive histone modification profiles used in scanning genomic regions for novel regulatory elements (**Supplementary Fig. 6** online). Right: predictions were filtered and classified as promoters or enhancers based on correlation with monomethylated H3K4 (H3K4me1) and trimethylated H3K4 (H3K4me3) chromatin signatures (**Supplementary Methods**). (**b**) 198 active promoters were predicted in the ENCODE regions in untreated HeLa cells and clustered into classes PI–PIV. The predictions contain 140 active RefSeq promoters and 32 inactive RefSeq promoters, indicating a sensitivity of 79.1% for active promoter detection. (**c**) The high degree of overlap between untreated and IFNγ-treated HeLa promoter prediction sets supports the applicability of our approach to independent data sets. The majority of predicted promoters map to known GENCODE 5′ ends in untreated (**d**) and treated cells (**e**), confirming the accuracy of our predictions.

were distributed very similarly (**Fig. 2e**). Comparison with the recent RIKEN human CAGE data set[26] showed that the vast majority of the predicted promoters were supported by multiple CAGE tags (**Supplementary Methods**). Even predicted promoters that did not map to a known GENCODE 5′ end were largely supported by multiple CAGE tags (50% in untreated cells, 27% in treated cells) or DHSs (83% in untreated cells, 73% in treated cells). It is possible that the inactive promoters identified in our analysis correspond to transcripts that are expressed below the detection threshold, or these promoters might retain some features of transcriptional competence in the absence of active transcription. Six promoter predictions in untreated HeLa cells (nine predictions in treated cells) did not correspond to any known or putative 5′ ends and probably represent previously unknown promoters; all of these predicted new promoters overlap with DHSs.

We also predicted 389 enhancers (**Supplementary Table 5**) in untreated HeLa cells (**Fig. 3a**; enhancer predictions are classified EI–EIV to distinguish them from the p300 binding sites presented in **Fig. 1**) and 324 enhancers in treated cells, with 89% overlap between prediction sets (**Fig. 3b**). Although the prediction algorithm was trained on the histone modification profiles of untreated cells, it accurately identified 77% of the distal p300 binding sites in IFNγ-treated cells, indicating a high degree of sensitivity for the prediction of enhancers in an independent data set. Several lines of evidence support the function of these predictions as enhancers. First, over 85% of predictions were located more than 2.5 kb from known gene 5′ ends (**Fig. 3c** and **Supplementary Fig. 4** online), consistent with their predicted function. Second, they were evolutionarily conserved, with 53.3% ($P < 1 \times 10^{-16}$) containing a strongly conserved sequence. Third, many overlapped with predicted transcriptional regulatory modules (36.3%, $P = 1.7 \times 10^{-4}$). Fourth, a significant proportion of the enhancer predictions (55.3%, $P < 1 \times 10^{-16}$) overlapped with DHSs, including the well-known HS2 enhancer in the β-globin locus[27] (**Supplementary Fig. 5** online). Of 587 distal DHSs in HeLa cells, we predict that 175 (29.8%) are enhancers; the other distal DHSs probably represent additional regulatory elements such as repressors or insulators or sequences that contribute to chromatin organization.

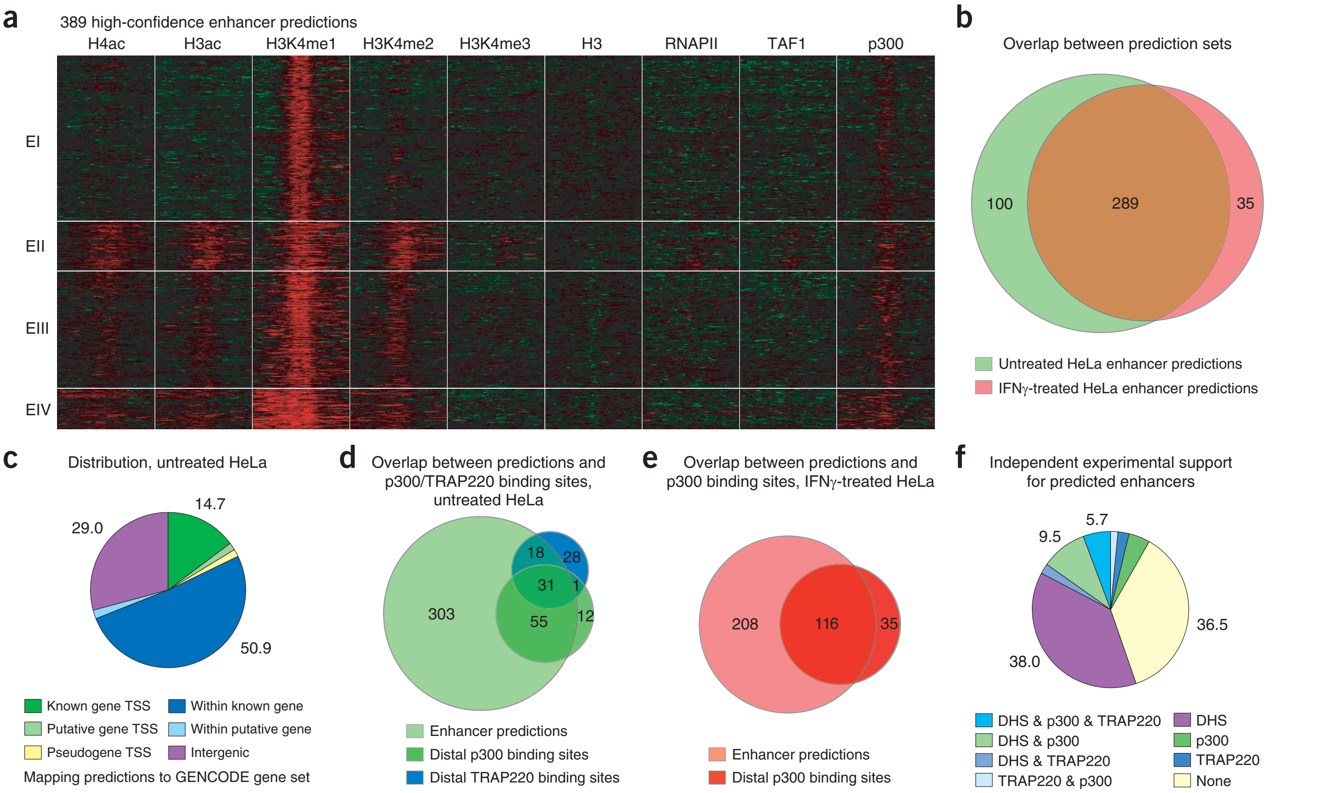

Figure 3 Prediction of enhancers based on chromatin signatures. (**a**) 389 enhancers were predicted in the ENCODE regions in untreated HeLa cells and clustered into classes EI–EIV. The predicted enhancers show the monomethylated H3K4 (H3K4me1) enrichment and lack of trimethylated H3K4 (H3K4me3) observed at distal p300 binding sites. H4ac, acetylated H4; H3ac, acetylated H3. (**b**) The high degree of overlap between untreated and IFNγ-treated HeLa enhancer prediction sets supports the applicability of our approach to independent data sets. The majority of enhancer predictions in untreated (**c**) and treated cells (see **Supplementary Fig. 4**) are found away from known GENCODE 5′ ends, similar to p300 binding site distribution. The enhancer prediction sets contain the majority of known distal p300 binding sites in untreated (**d**) and treated cells (**e**), confirming the sensitivity of our approach, even though the prediction algorithm was trained only on data from untreated cells. (**f**) Most enhancers predicted on the basis of their chromatin signatures are also supported by DNaseI hypersensitivity and/or binding of p300 and/or TRAP220.

Finally, 86 enhancer predictions in the untreated set (and 116 in the treated set) mapped to distal p300 binding sites (**Fig. 3d,e**) and many others seemed to be enriched in p300 binding but fell below the threshold of our target selection.

We also discovered that many predicted enhancers lacked p300 binding. To determine if these genomic regions were occupied by p300-independent transcriptional coactivator complexes, we performed additional ChIP-chip experiments to examine binding of TRAP220, a component of the Mediator complex that has been shown to occupy enhancers as well as promoters[10,11]. Of 162 TRAP220 binding sites we identified in the ENCODE regions (**Supplementary Table 6** online), 78 (48.1%) were located far from known 5′ ends of transcripts and might represent potential enhancers. Almost 63% of the distal TRAP220 sites were contained within our enhancer prediction set (**Fig. 3d**), and 18 of them were bound by TRAP220 but not p300, confirming the identity of these predicted enhancers. This result suggests that our chromatin-based prediction model is not limited only to enhancers marked by p300. Overall, the majority of predicted enhancers (63.5%) are supported by DNaseI hypersensitivity, binding of p300, binding of TRAP220 or a combination of these features (**Fig. 3f**).

Identification of a novel enhancer for *SLC22A5*

To confirm the potential of our approach to identify enhancers that regulate the activity of target human promoters, we next examined a predicted enhancer located 6 kb upstream of *SLC22A5* (also known as *OCTN2*) on chromosome 5 (**Fig. 4a**). *SLC22A5* is a widely expressed gene that encodes a carnitine transporter[28–31]. Mutations in this gene have been identified as a cause of systematic carnitine deficiency, a condition occurring mostly in children that prevents the body from using fats for energy and can result in symptoms including encephalopathy, cardiomyopathy, hypoglycemia and, in serious complications, heart failure, liver problems, coma and sudden unexpected death[32–35]. Although substantial research has been devoted to the role of *SLC22A5* in carnitine transport, fatty acid metabolism and related human diseases, very little is known about the transcriptional regulation of this gene. We cloned a region of the *SLC22A5* locus (L) containing the promoter and predicted enhancer (E) into a luciferase reporter construct and compared its activity to that of the locus without the predicted enhancer (LΔE) in transiently transfected HeLa cells. The deletion of the predicted enhancer caused a 2.5-fold reduction in reporter activity (**Fig. 4b**), supporting the necessity of this site for full activity of the *SLC22A5* promoter. We then cloned the predicted enhancer downstream of the luciferase gene in a construct containing the proximal *SLC22A5* promoter (P^S) and observed 4.2-fold greater reporter activity from the promoter-enhancer construct (P^SE) than the construct containing only the promoter (**Fig. 4b**), confirming that the predicted enhancer is sufficient to increase the activity of this promoter in a position-independent manner. The predicted enhancer

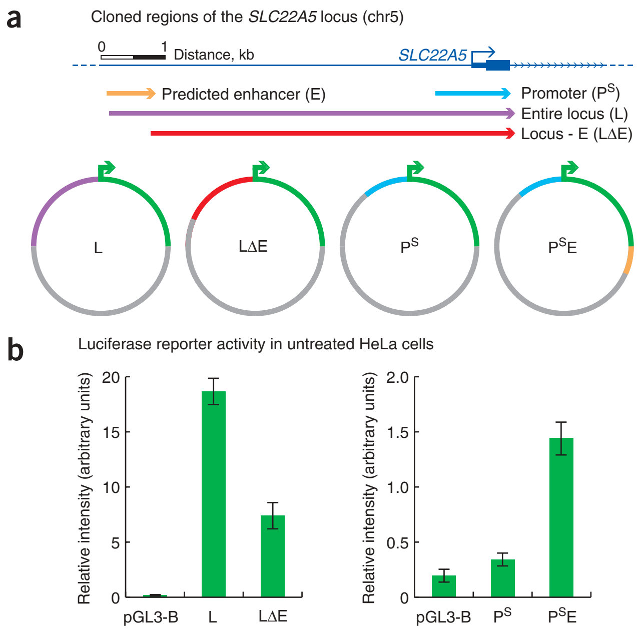

Figure 4 Identification of a putative novel enhancer for *SLC22A5*. (**a**) To test the effect of a predicted enhancer on a nearby promoter, regions of the *SLC22A5* locus were cloned into pGL3 reporter constructs in the direction indicated. (**b**) Luciferase activity of the entire 6.5-kb locus (L) was reduced 2.5-fold by the deletion of 700 bp containing the predicted enhancer (LΔE), and the presence of the predicted enhancer downstream of the luciferase gene in a construct containing the *SLC22A5* promoter (P^SE) caused a fourfold increase in activity compared with the promoter alone (P^S). Error bars represent s.d.

did not stimulate reporter activity in the absence of the *SLC22A5* promoter (data not shown). Our results suggest that the putative *SLC22A5* enhancer identified by our method is indeed critical for optimal transcriptional activation of this gene.

Functional validation of promoter and enhancer predictions

To further assess the accuracy of our predictions, we compared our high-confidence prediction sets to a list of *in vivo* STAT1 binding sites independently mapped in the ENCODE regions, hypothesizing that STAT1 sites are likely to occupy both promoters and enhancers. We performed ChIP-chip in HeLa cells before and after IFNγ treatment as described above and additionally on a PCR-product microarray platform (**Supplementary Methods**) and validated the results using quantitative real-time PCR, generating a list of 13 high-confidence STAT1 sites in IFNγ-treated cells (**Supplementary Table 7** online); as expected, we did not detect any STAT1 binding in cells before treatment. We compared the STAT1 sites with our prediction lists and found that seven STAT1 sites map to promoter predictions, four map to enhancer predictions, and two are not near any predictions, indicating that our prediction model is capable of detecting the majority (>84%) of an independently generated collection of putative regulatory elements. Four of the seven promoter predictions map to known TSSs: *IRF1* (a known STAT1 target), *RPS9*, *c21orf59* and *IFNAR2*. All of these genes are expressed in HeLa cells, supporting the accuracy of our active promoter predictions.

To validate the new promoter and enhancer predictions at STAT1 sites, we examined their functional properties using reporter assays. As two adjacent STAT1 sites on chromosome 5 (STAT1.02-.03) map to the same promoter prediction, we examined the closer of the two sites along with the other novel STAT1 promoter prediction (**Fig. 5a**), four STAT1 enhancer predictions (**Fig. 5b**), and the two non-predicted STAT1 sites (**Fig. 5c**). To test for promoter activity, regions containing the STAT1 sites were amplified from genomic DNA and cloned upstream of the luciferase gene in vectors lacking a promoter (**Fig. 5d**); to test for enhancer activity, the same fragments were cloned downstream of the luciferase gene into vectors containing the SV40 minimal promoter (**Fig. 5e**), as enhancers are thought to contribute to target gene activation regardless of their position relative to the gene promoter. Clones were transiently transfected into HeLa cells and assayed for reporter activity before and after treatment with IFNγ.

Both STAT1 promoter predictions stimulated reporter activity in the absence of the SV40 promoter when cloned in the upstream position (**Fig. 5d**), in accord with their predicted function. Three STAT1 enhancer predictions (STAT1.08-.10) stimulated strong reporter activity when cloned in the downstream position (**Fig. 5e**) but required the presence of the SV40 promoter (**Supplementary Methods**), consistent with the position-independence and promoter-dependence of enhancer activity. The fourth enhancer prediction (STAT1.11) showed only weak enhancer activity, though we noted that the STAT1 site in this region is further away from the prediction (710 bp) than any of the other STAT1 sites that we examined (average ~240 bp). The effect of IFNγ is variable among the different sites in both ChIP-chip binding profiles and reporter activity, though there seems to be a relationship between inducibility of p300 binding and reporter activity. Notably, one promoter prediction (STAT1.03) also showed enhancer activity (**Fig. 5e**). Examination of our prefiltering prediction lists (**Supplementary Methods**) uncovered a predicted enhancer within the STAT1.03 cloned region, explaining the apparent dual functional activity of this newly discovered promoter. The nonpredicted sites (STAT1.12-.13) did not show any functional activity and were not marked by either of the distinctive histone modification patterns, supporting the specificity of our model. It is still possible that these sites are actually regulatory elements that cannot be tested in our system owing to their function or a requirement for native chromatin context, but it is worth noting that these are the only two STAT1 sites that did not demonstrate DNaseI hypersensitivity.

These data provide functional validation for our model of distinct chromatin signatures at promoters and enhancers, confirm that our computational approach can accurately predict the position and function of these transcriptional regulatory elements on the basis of their chromatin signatures and suggest a direct connection between chromatin signatures and the regulatory potential of the DNA sequences that they denote.

DISCUSSION

In summary, we mapped five histone modifications, four general transcription factors and nucleosome density at high resolution in 30 Mb of the human genome, identifying chromatin features that distinguish promoters from enhancers. Although both kinds of regulatory elements share some features such as nucleosome depletion and enrichment of histone acetylation and dimethylated H3K4, the high-resolution profiles of these markers and the dichotomy of enrichment for trimethylated H3K4 and monomethylated H3K4 at active promoters and enhancers define chromatin signatures that can be used to locate novel regulatory elements in the human genome. The monomethylated H3K4 enhancer signature is present in HeLa cell chromatin at multiple loci whose enhancer activity was functionally validated, including a putative novel enhancer for *SLC22A5*.

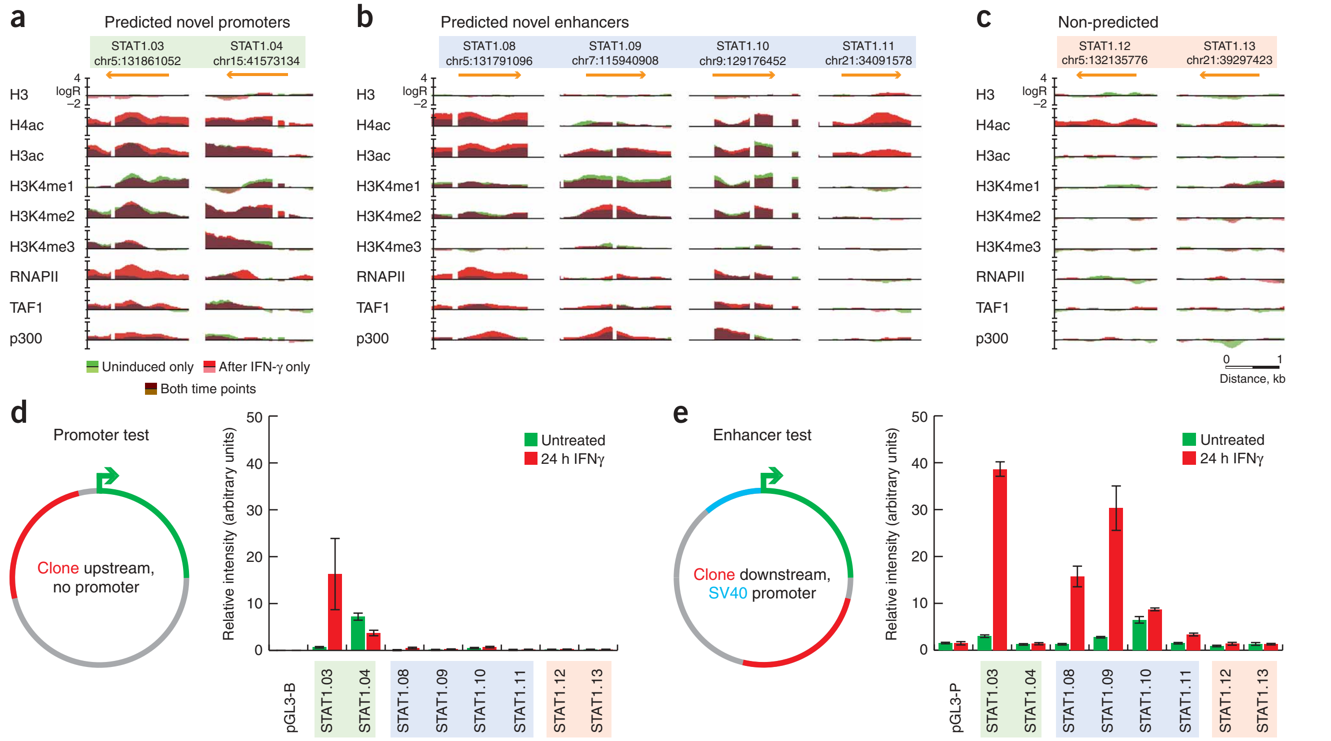

Figure 5 Validation of the prediction model by STAT1 binding and reporter assays. Of 13 high-confidence STAT1 binding sites, four are found at known promoters (not shown), two at predicted novel promoters (**a**) and four at predicted enhancers (**b**), and two are not predicted (**c**). The eight STAT1 sites in **a**–**c** were cloned into pGL3 reporter constructs to examine their regulatory potential as promoters (**d**) and enhancers (**e**). Error bars represent s.d. The coverage and direction of each clone are represented by orange arrows in **a**–**c**, and ChIP-chip profiles of each marker are shown at the eight STAT1 binding sites, before and after IFNγ treatment (green and red, respectively, where brown indicates enrichment at both time points; images generated in part at http://genome.ucsc.edu using hg17). The STAT1 binding sites in **a** and **b** function as predicted in the reporter assays, whereas the non-predicted sites in **c** show no reporter activity. H4ac, acetylated H4; H3ac, acetylated H3.

Previous studies have identified some histone modification patterns of promoters and heterochromatin, but our findings expand the current knowledge of chromatin architecture at human promoters and present evidence for previously unknown chromatin features of human enhancers, representing an effective new strategy for identifying and distinguishing promoters and enhancers. The presence of histone acetylation and methylation that we observe upstream of the TSSs of active human promoters has not been reported in yeast and suggests some transcriptional regulatory mechanisms specific to mammalian gene expression. Additionally, the discovery of monomethylated H3K4 at human enhancers may contribute to our understanding of how enhancers function in tissue-specific gene regulation.

In recent years, the genome sequences of a growing number of organisms have been obtained, but extracting functional information from these nucleotide sequences remains a great challenge, as our knowledge of transcription factor binding motifs is incomplete, and current sequence-based computational tools are limited in their ability to predict the regulatory function of genomic sequences. Here, we present a strategy to identify transcriptional regulatory elements on the basis of their epigenetic characteristics, independent of motifs or other sequence features. Our chromatin-based prediction model provides a means to locate and distinguish promoters and enhancers at high resolution and with high degrees of sensitivity and specificity. Although the prediction model was trained only on data from untreated HeLa cells, the sensitivity of the model in data from IFNγ-treated cells supports the utility of our approach in analyzing independent data sets. The results of the functional assays of predicted STAT1 binding sites confirm the ability of our prediction model for identifying the location and function of novel promoters and enhancers, even before their activation. Because we used the histone modification profiles at distal p300 binding sites as the basis for our enhancer prediction strategy, we were initially concerned that our predictions might be biased toward only the subset of enhancers bound by p300. Based on the overlap of our predictions with 63% of distal TRAP220 sites and 30% of distal DHSs, however, we conclude that our model is not biased toward p300 binding sites and that the chromatin signatures we observed are not limited to this subset of enhancers. Extension of our model to additional cell types and other components of chromatin architecture will be useful in determining the mechanisms of enhancer maintenance and function in regulating tissue-specific gene expression, findings that will be particularly important to our knowledge of how epigenetic factors and distal transcriptional regulatory elements contribute to human development and disease.

Our approach will also be valuable to the functional annotation of the human genome, as it provides a new, effective means to locate active transcriptional enhancers that have thus far eluded identification on a large scale. Given the degree of structural and functional conservation of chromatin and histone modifications from yeast to humans, these predictive chromatin signatures may be useful in annotating promoters and enhancers in the genomes of a variety of organisms.

METHODS

For detailed methods and materials, please refer to **Supplementary Methods.**

Briefly, HeLa cells were cultured under adherent conditions in DMEM + 10% FBS. Three biological replicates of IFNγ-treated and untreated cells were cross-linked and harvested as previously described[18], except that cells were cross-linked for 20 min at 37 °C. Chromatin preparation, ChIP-chip, DNA purification and LM-PCR were performed as previously described[18] using commercially available antibodies against the following proteins: histone H3 (Abcam ab1791), acetylated H4 (Upstate 06-866), acetylated H3 (Upstate 06-599), monomethylated H3K4 (Abcam ab8895), dimethylated H3K4 (Upstate 07-030), trimethylated H3K4 (Upstate 07-473), RNAPII (Covance MMS-126R), TAF1 (Santa Cruz sc-735), p300 (Santa Cruz sc-585) and TRAP220 (Santa Cruz sc-5334). ChIP-DNA samples were labeled and hybridized to NimbleGen ENCODE HG17 microarrays (NimbleGen Systems). Data were analyzed using standard methods, and ChIP-chip targets for RNAPII were selected with the Mpeak program and validated by quantitative real-time PCR using the iCycler and SYBR-green iQ Supermix (Bio-Rad Laboratories). Gene expression in treated and untreated HeLa cells was analyzed using HU133 Plus 2.0 microarrays (Affymetrix) as described[18]. DNase-chip was performed and the data analyzed as described[24]. Promoters (TSSs) and putative enhancers (p300 binding sites) were clustered on the basis of histone modification patterns using K-means clustering of 10-kb windows centered on each target. Average profiles were generated for each class of promoter and enhancer and were used to train a computational prediction model to identify promoters and enhancers on the basis of histone modification ChIP-chip profiles. Predictions were further filtered by correlation to chromatin signatures to remove false positives and ambiguous classifications. Predicted regulatory modules were obtained from http://genomequebec.mcgill.ca/PReMod/; phastCons and CAGE data were extracted from the UCSC Genome Browser and binding sites and predictions were mapped relative to the October 2005 hg17 GENCODE gene sets. ChIP-chip was performed against STAT1 (using anti-STAT1, Santa Cruz sc-345) as described above and using PCR microarrays, and the results were validated by quantitative real-time PCR. Predicted STAT1 sites were cloned into modified pGL3 reporter constructs (Promega), transiently transfected into HeLa cells and assayed for luciferase activity before and after IFNγ treatment using the Dual Luciferase Kit (Promega). Raw and processed data for the microarray experiments can be found at the UCSC genome browser (http://genome.ucsc.edu) and http://licr-renlab.ucsd.edu/download.html.

Accession codes. GEO: raw and processed data for the microarray experiments can be found under accession number GSE6273.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank J. Kadonaga, X. Fu and members of the Ren lab for comments. This work was supported by funding from the Ludwig Institute for Cancer Research (B.R.), the National Human Genome Research Institute (B.R., Z.W. and R.D.G.) and the National Cancer Institute (B.R.). Requests for materials should be addressed to B.R.

AUTHOR CONTRIBUTIONS

N.D.H., B.R. and R.D.G. designed the transcription factor and histone ChIP-chip experiments; G.E.C. designed and performed the DNase-chip experiments; N.D.H., R.K.S., C.W.C., R.D.H. and S.V.C. conducted the ChIP-chip experiments; N.D.H., G.H., L.O.B., K.A.C. and C.Q. analyzed the microarray data; G.H., N.D.H., B.R. and W.W. conceived and developed the promoter and enhancer prediction method. Independently, Y.F. and Z.W. discovered the promoter-associated chromatin signatures. N.D.H. and B.R. wrote the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Genetics* website for details).

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Lemon, B. & Tjian, R. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.* **14**, 2551–2569 (2000).
2. Orphanides, G. & Reinberg, D. A unified theory of gene expression. *Cell* **108**, 439–451 (2002).
3. Nightingale, K.P., O'Neill, L.P. & Turner, B.M. Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. *Curr. Opin. Genet. Dev.* **16**, 125–136 (2006).
4. Smale, S.T. & Kadonaga, J.T. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **72**, 449–479 (2003).
5. Blackwood, E.M. & Kadonaga, J.T. Going the distance: a current view of enhancer action. *Science* **281**, 60–63 (1998).
6. Bulger, M. & Groudine, M. Looping versus linking: toward a model for long-distance gene activation. *Genes Dev.* **13**, 2465–2477 (1999).
7. Strahl, B.D. & Allis, C.D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).
8. Margueron, R., Trojer, P. & Reinberg, D. The key to development: interpreting the histone code? *Curr. Opin. Genet. Dev.* **15**, 163–176 (2005).
9. Barrera, L.O. & Ren, B. The transcriptional regulatory code of eukaryotic cells - insights from genome-wide analysis of chromatin organization and transcription factor binding. *Curr. Opin. Cell. Biol.* **18**, 291–298 (2006).
10. Hatzis, P. & Talianidis, I. Dynamics of enhancer-promoter communication during differentiation-induced gene activation. *Mol. Cell* **10**, 1467–1477 (2002).
11. Wang, Q., Carroll, J.S. & Brown, M. Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking. *Mol. Cell* **19**, 631–642 (2005).
12. Bernstein, B.E. *et al.* Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**, 169–181 (2005).
13. Roh, T.Y., Cuddapah, S. & Zhao, K. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.* **19**, 542–552 (2005).
14. Kim, T.H. & Ren, B. Genome-wide analysis of protein-DNA interactions. *Annu. Rev. Genomics Hum. Genet.* **7**, 81–102 (2006).
15. The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).
16. Horvai, A.E. *et al.* Nuclear integration of JAK/STAT and Ras/AP-1 signaling by CBP and p300. *Proc. Natl. Acad. Sci. USA* **94**, 1074–1079 (1997).
17. Pokholok, D.K. *et al.* Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122**, 517–527 (2005).
18. Kim, T.H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–880 (2005).
19. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504 (2005).
20. Kouskouti, A. & Talianidis, I. Histone modifications defining active genes persist after transcriptional and mitotic inactivation. *EMBO J.* **24**, 347–357 (2005).
21. Liu, C.L. *et al.* Single-nucleosome mapping of histone modifications in *S. cerevisiae. PLoS Biol.* **3**, e328 (2005).
22. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* **7**(Suppl.), S4.1–S4.9 (2006).
23. Felsenfeld, G. Chromatin unfolds. *Cell* **86**, 13–19 (1996).
24. Crawford, G.E. *et al.* DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat. Methods* **3**, 503–509 (2006).
25. Blanchette, M. *et al.* Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.* **16**, 656–668 (2006).
26. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635 (2006).
27. Li, Q., Peterson, K.R., Fang, X. & Stamatoyannopoulos, G. Locus control regions. *Blood* **100**, 3077–3086 (2002).
28. Schomig, E. *et al.* Molecular cloning and characterization of two novel transport proteins from rat kidney. *FEBS Lett.* **425**, 79–86 (1998).
29. Sekine, T. *et al.* Molecular cloning and characterization of high-affinity carnitine transporter from rat intestine. *Biochem. Biophys. Res. Commun.* **251**, 586–591 (1998).
30. Tamai, I. *et al.* Molecular and functional identification of sodium ion-dependent, high affinity human carnitine transporter OCTN2. *J. Biol. Chem.* **273**, 20378–20382 (1998).
31. Wu, X., Prasad, P.D., Leibach, F.H. & Ganapathy, V. cDNA sequence, transport function, and genomic organization of human OCTN2, a new member of the organic cation transporter family. *Biochem. Biophys. Res. Commun.* **246**, 589–595 (1998).
32. Nezu, J. *et al.* Primary systemic carnitine deficiency is caused by mutations in a gene encoding sodium ion-dependent carnitine transporter. *Nat. Genet.* **21**, 91–94 (1999).
33. Shoji, Y. *et al.* Evidence for linkage of human primary systemic carnitine deficiency with D5S436: a novel gene locus on chromosome 5q. *Am. J. Hum. Genet.* **63**, 101–108 (1998).
34. Stanley, C.A. Carnitine deficiency disorders in children. *Ann. NY Acad. Sci.* **1033**, 42–51 (2004).
35. Wang, Y., Ye, J., Ganapathy, V. & Longo, N. Mutations in the organic cation/carnitine transporter OCTN2 in primary carnitine deficiency. *Proc. Natl. Acad. Sci. USA* **96**, 2356–2360 (1999).