

Discovery and characterization of chromatin states for systematic annotation of the human genome

Jason Ernst^{1,2} & Manolis Kellis^{1,2}

A plethora of epigenetic modifications have been described in the human genome and shown to play diverse roles in gene regulation, cellular differentiation and the onset of disease. Although individual modifications have been linked to the activity levels of various genetic functional elements, their combinatorial patterns are still unresolved and their potential for systematic *de novo* genome annotation remains untapped. Here, we use a multivariate Hidden Markov Model to reveal 'chromatin states' in human T cells, based on recurrent and spatially coherent combinations of chromatin marks. We define 51 distinct chromatin states, including promoter-associated, transcription-associated, active intergenic, large-scale repressed and repeat-associated states. Each chromatin state shows specific enrichments in functional annotations, sequence motifs and specific experimentally observed characteristics, suggesting distinct biological roles. This approach provides a complementary functional annotation of the human genome that reveals the genome-wide locations of diverse classes of epigenetic function.

The primary DNA sequence of the human genome encodes the genetic information of each cell, but numerous epigenetic modifications can modulate the interpretation of the primary sequence. These modifications contribute to the diversity of phenotypes found across different human cell types, play key roles in the establishment and maintenance of cellular identity during development and have been associated with DNA repair, replication and human disease. Post-translational modifications in the tails of histone proteins that package DNA into chromatin constitute perhaps the most versatile type of such epigenetic information. More than a dozen positions of multiple histone proteins can undergo a number of modifications, such as acetylation and mono-, di- or tri-methylation^{1,2}.

More than 100 distinct histone modifications have been described, leading to the 'histone code hypothesis' that specific combinations of chromatin modifications would encode distinct biological functions³. Others, however, have instead proposed that individual epigenetic marks act in additive ways and the multitude of modifications simply contributes to stability and robustness⁴. The specific combinations of

epigenetic modifications that are biologically meaningful, and their corresponding functional roles, are still largely unknown.

To directly address these questions, we introduce an approach for the *de novo* discovery of 'chromatin states' (Fig. 1, Supplementary Table 1 and Supplementary Fig. 1), or biologically meaningful and spatially coherent combinations of chromatin marks, by performing a systematic genome-wide analysis based on a multivariate Hidden Markov Model (HMM). Multivariate HMMs are graphical probabilistic models that model multiple 'observed' inputs as generated by unobserved 'hidden' states, using transitions between hidden states to model spatial relationships (Online Methods).

Our model captures two types of chromatin information. The frequency with which different chromatin mark combinations are found with each other are captured by a vector of 'emission' probabilities associated with each chromatin state (Fig. 2 and Supplementary Figs. 2 and 3) and the frequency with which different chromatin states occur in spatial relationships of each other along the genome are encoded in a 'transition' probability vector associated with each state. These spatial relationships capture both the spreading of certain chromatin domains across the genome, as well as the functional ordering of different states such as from intergenic regions to promoter regions and transcribed regions (Supplementary Notes and Supplementary Figs. 4–6). Biologically the genomic locations associated with a given chromatin state may correspond to specific types of functional elements, such as transcription start sites (TSS), enhancers, active genes, repressed genes, exons or heterochromatin, which can be inferred solely from the corresponding combinations of chromatin marks in their spatial context, even though no information about these annotations is given to the model as input.

We applied our model to the largest data set of chromatin mark information available, consisting of the genome-wide occupancy data for a set of 38 different histone methylation and acetylation marks and for the histone variant H2AZ, RNA polymerase II (PolII) and CTCF in human CD4 T-cells. The maps were previously obtained using chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq) (Online Methods)^{5,6}. To understand the biological importance of the resulting chromatin states, we undertook a large-scale, systematic data-mining effort, bringing to bear dozens of genome-wide data sets including gene annotations, expression information, evolutionary conservation, regulatory motif instances, compositional biases, genome-wide association data, transcription-factor binding, DNaseI hypersensitivity and nuclear lamina maps.

This work provides an unbiased and systematic chromatin-driven annotation for every region of the genome at a 200 base pair resolution, refining previously described epigenetic states and introducing

¹MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts, USA. ²Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. Correspondence should be addressed to M.K. (manoli@mit.edu).

Published online 25 July 2010; doi:10.1038/nbt.1662

additional ones. Regardless of whether these chromatin states are causal in directing regulatory processes, or simply reinforcing independent regulatory decisions, these annotations should provide a resource for interpreting biological and medical data sets, such as genome-wide association studies for diverse phenotypes and could potentially help to identify new classes of functional elements.

RESULTS

Chromatin states model and comparison to previous work

Previous analyses have largely focused on characterizing the marks predictive of specific classes of genomic elements defined a priori such as transcribed regions, promoters or putative enhancers, and using the characterization to identify new instances of these classes^{5–12}.

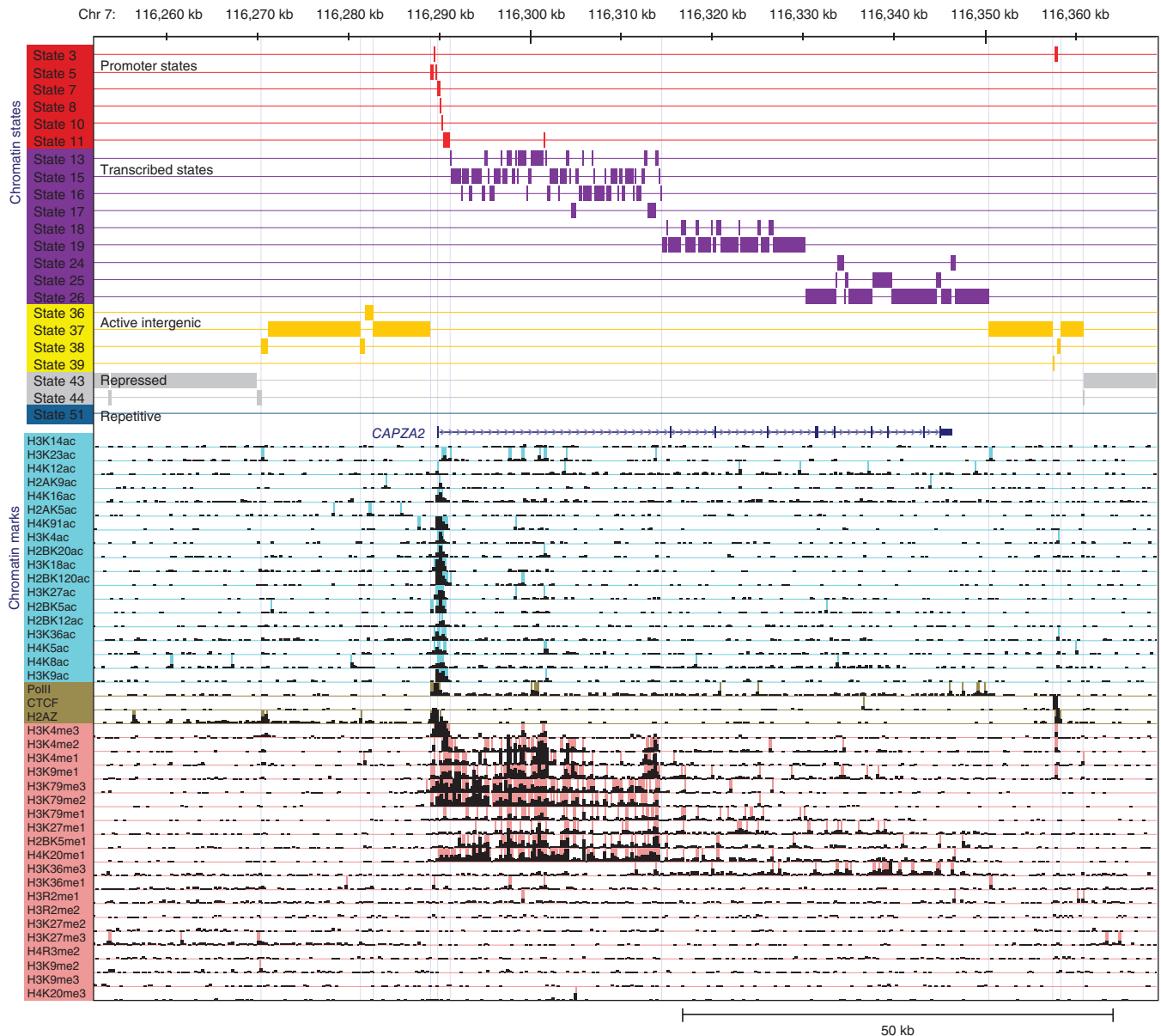


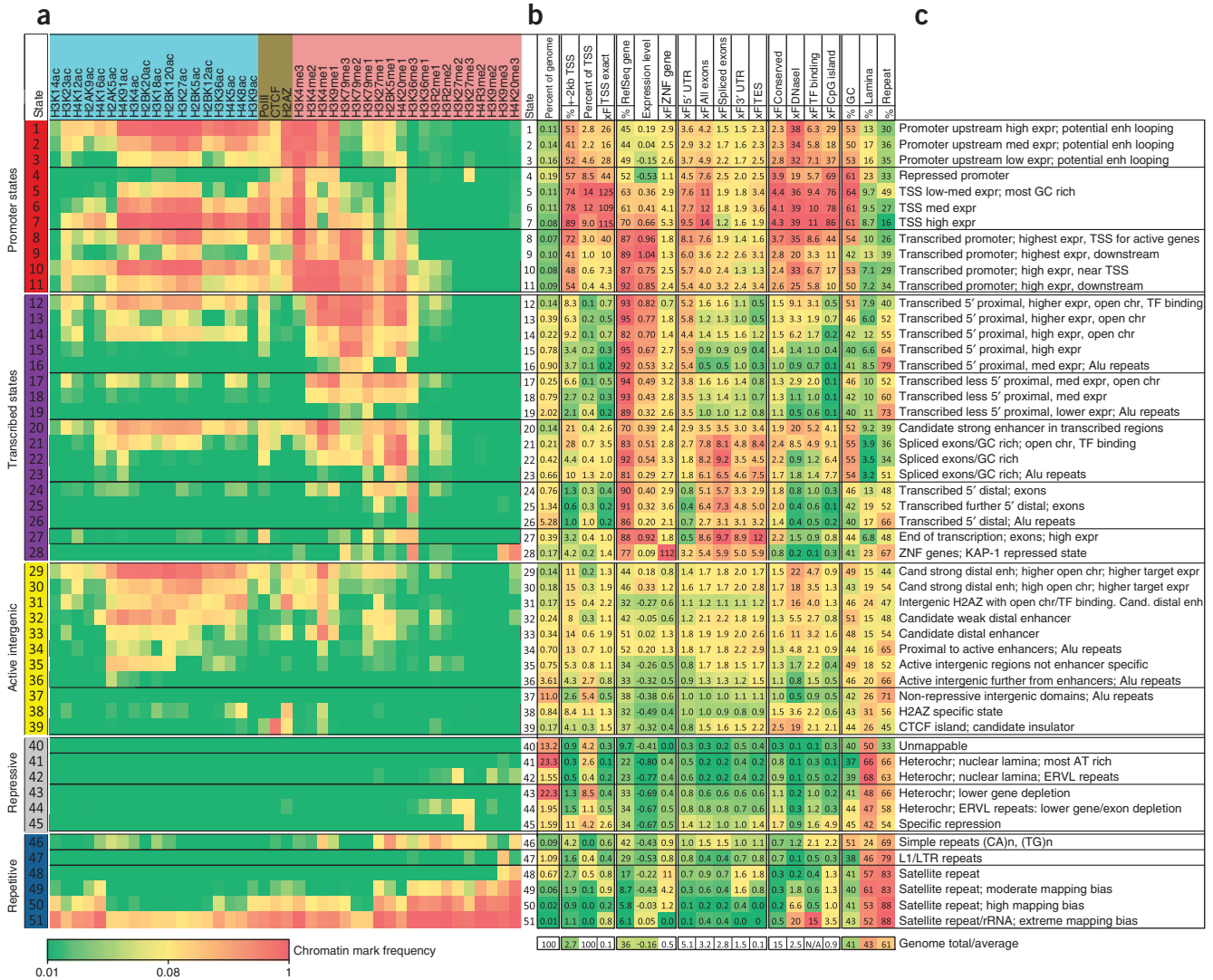
Figure 1 Example of chromatin state annotation. Input chromatin mark information and resulting chromatin state annotation for a 120-kb region of human chromosome 7 surrounding the *CAPZA2* gene. For each 200-bp interval, the input ChIP-Seq sequence tag count (black bars) is processed into a binary presence and/or absence call for each of 18 acetylation marks (light blue), 20 methylation marks (pink) and CTCF/Pol2/H2AZ (brown). The precise combination of these marks in each interval in their spatial context is used to infer the most probable chromatin state assignment (colored boxes). Although chromatin states were learned independently of any prior genome annotation, they correlate strongly with upstream and downstream promoters (red), 5'-proximal and distal transcribed regions (purple), active intergenic regions (yellow), repressed (gray) and repetitive (blue) regions (state descriptions shown in **Supplementary Table 1**). This example illustrates that even when the signal coming from chromatin marks is noisy, the resulting chromatin state annotation is very robust, directly interpretable and shows a strong correspondence with the gene annotation. Several spatially coherent transitions are seen from large-scale repressed to active intergenic regions near active genes, from upstream to downstream promoter states surrounding the TSS and from 5'-proximal to distal transcribed regions along the body of the gene. The frequent transitions to state 16 correlate with annotated Alu elements (57% overlap versus 4% and 25% for states 13 and 15, respectively). Transitions to state 13 are likely due to enhancer elements in the first intron of *CAPZA2*, a region where regulatory elements are commonly found and correlate with several enhancer marks. The maximum-probability state assignments are shown here, and the full posterior probability for each state in this region is shown in **Supplementary Figure 1**.

An unsupervised (without using prior knowledge) local chromatin pattern discovery method¹³ first demonstrated that many of the patterns previously associated with promoters and enhancers could be discovered *de novo*, but did not discover patterns associated with broader domains and left the vast majority of the genome unannotated (Supplementary Fig. 7).

Unsupervised HMM approaches that modeled chromatin mark signal intensity levels using multivariate normals or nonparametric histograms^{14–18} have been previously used, but in contrast we use a binarization approach that explicitly models the presence/absence

frequency of each mark. Specifically, we make a local call of whether a mark was present in each 200-bp interval, and use a Bernoulli random variable to model the probability of detection of each mark in isolation, and a product of independent probabilities to model the probability of each combination of marks (Online Methods). Our approach has the advantage that the model parameters are directly interpretable as the frequencies of each mark and each mark combination, in contrast to previous approaches for which the biological significance of the parameters corresponding to varying signal intensity levels for each mark is often unclear. Moreover, the binarization also makes our

© 2010 Nature America, Inc. All rights reserved.



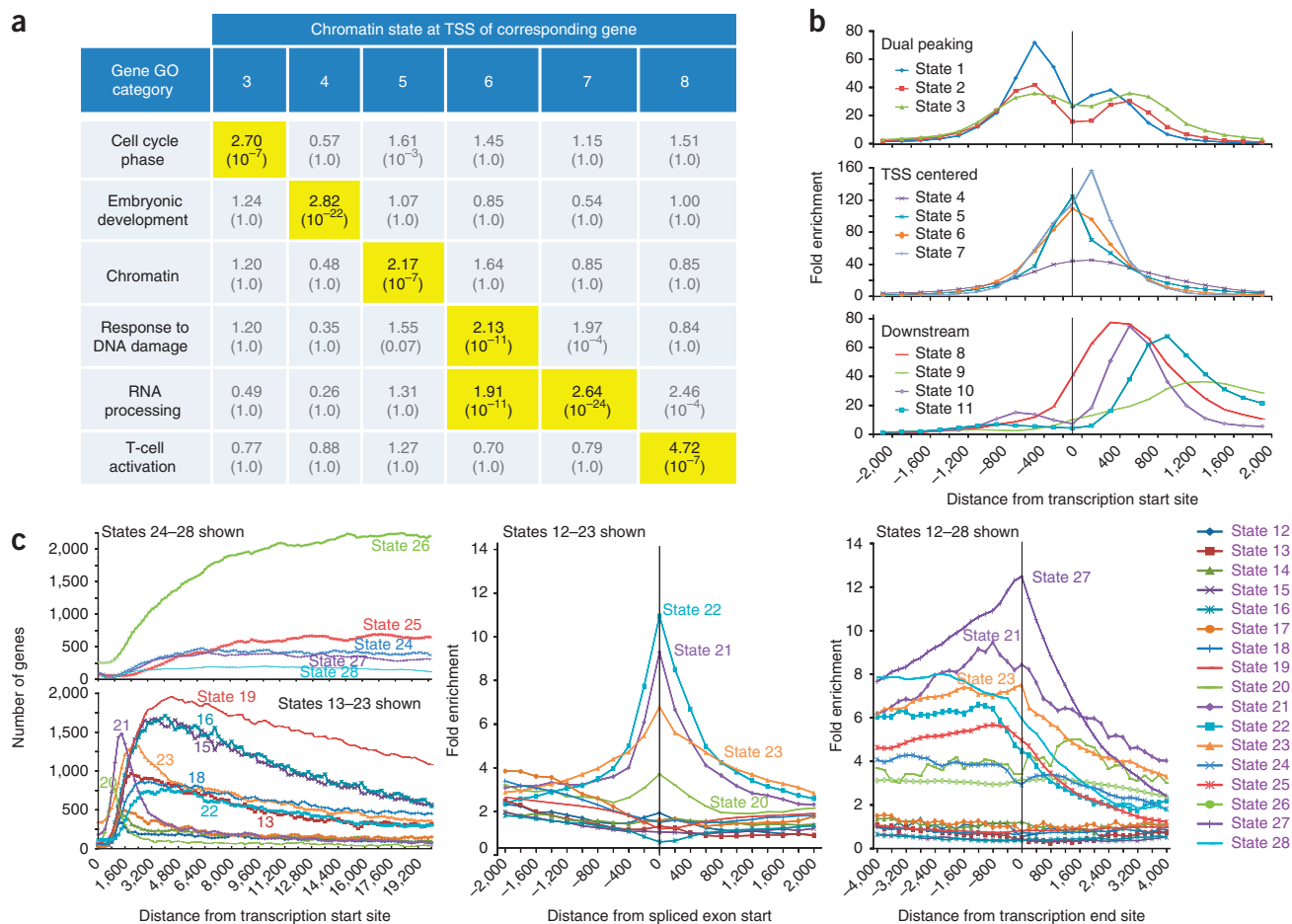


Figure 3 Promoter and transcribed chromatin states show distinct functional and positional enrichments. **(a)** Distinct Gene Ontology (GO) functional enrichments (fold and corrected P -values) found for genes associated with different promoter states at their TSS. For additional states and GO terms, see **Supplementary Figure 29**. **(b)** Distinct positional biases of promoter states with respect to nearest RefSeq TSS distinguish states peaking upstream, only downstream and centered at the TSS. **(c)** Positional biases of transcribed states with respect to TSS, nearest spliced exon start and transcription end sites (TES). These distinguish 5'-proximal states (12–23, left panel), 5'-distal states (24–28), states strongly enriched for spliced exons (middle panel, see also **Supplementary Fig. 24** for plot for states 24–28) and TES-associated states (with state 27 being particularly precisely positioned, right panel).

model less prone to forming states overfitting potentially insignificant variations in signal intensity levels. In contrast to models that use a multivariate normal distribution, our method avoids this strong parametric assumption, which is generally violated by the often relatively small discrete counts found in ChIP-seq experiments, enabling more robust models to be inferred. In comparison to the models previously inferred based on a nonparametric histogram strategy¹⁸, our binarization approach uses an order of magnitude fewer parameters per state, further increasing model robustness and interpretability.

We developed a procedure for learning sets of chromatin states across a range of model complexities. For a given number of states and from a set of initial parameters, standard expectation maximization based procedures enable simultaneous local optimization of the state definitions (emission and transition probabilities) and the corresponding genome annotation consistent with the observed data. However the model inferred and its quality can depend on the initial set of parameters, which can confound comparing models with different number of states learned from independent initializations. We therefore used a two-stage process that first selected a 79-state model which had the highest complexity-penalized likelihood score across a large compendium of randomly-initialized models of varying complexity.

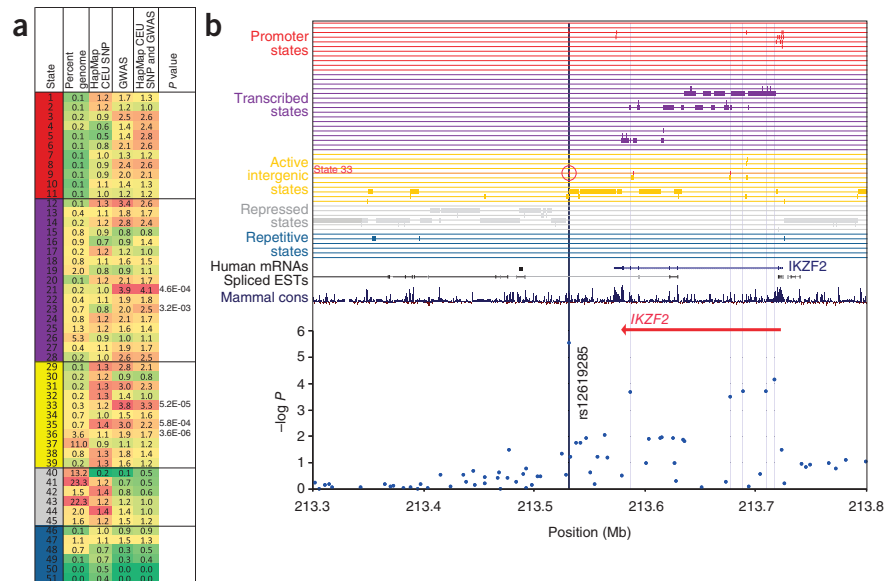
We then pruned and optimized this model down to smaller numbers of states, leading to a model with 51 states that were relatively consistently recovered across the compendium of models, and that sufficiently captured all states found in larger models for which we could give a distinct biological interpretation (see Online Methods). This enabled us to maintain a relatively small number of states while capturing most of the unique biology uncovered across our compendium of randomly-initialized models. Put in other words, this procedure enabled us to maximize biological interpretability, while minimizing model complexity. We further ensured that general properties of the resulting model validated our approach, including robustness to varying thresholds and different background models, and independence of marks given a chromatin state (**Supplementary Notes, Supplementary Figs. 8–21** and **Supplementary Table 2**).

We next describe the likely biological functions of the 51 discovered chromatin states, divided into five large groups.

Promoter-associated states

The first group of states, states 1–11, all had high enrichment for promoter regions: 40–89% of each state was within 2 kb of a RefSeq TSS, compared with 2.7% genome-wide ($P < 10^{-200}$, for all states).

Figure 4 SNP and GWAS enrichments for chromatin states. (a) Several chromatin states show enrichments for disease association data sets. For each state is shown: genome percentage; fold enrichment for SNPs from the HapMap CEU population; fold enrichment from a collection of 1,640 GWAS SNPs associated with a variety of diseases and traits from numerous studies²⁵; fold enrichment of GWAS SNPs relative to the HapMap CEU SNP enrichment; significance of GWAS SNPs relative to the underlying SNP frequency (when the corrected P -value < 0.01). (b) Example of intergenic SNP in GWAS-enriched state 33, found 40 kb downstream of the IKZF2 gene and associated with plasma eosinophil count levels²⁶. SNP significance as reported²⁶ is shown for each SNP in the region (blue circles) and associated chromatin state annotation (similar to Fig. 1). Red circle denotes top SNP and its overlap with state 33. In addition to top SNPs, secondary SNPs were also frequently found at or near GWAS-enriched states in several cases.



These states accounted for 59% of all RefSeq TSS although they covered only 1.3% of genome. These states all had a high frequency of H3K4me3 in common, as well as significant enrichments for DNaseI hypersensitive sites, CpG islands, evolutionarily conserved motifs and bound transcription factors (Fig. 2). They differed however in the presence and levels of other associated marks, primarily H3K79me2/3, H4K20me1, H3K4me1/2 and H3K9me1, and of numerous acetylations leading to varying strength of the aforementioned functional enrichments, and varying expression levels of the downstream genes (Supplementary Figs. 22 and 23).

Promoter states differed in the enrichment of Gene Ontology (GO) terms of associated genes including cell cycle, embryonic development, RNA processing and T-cell activation (Fig. 3a). For instance, the term 'embryonic development' is specifically enriched in state 4, whereas the term 'T-cell activation' is specifically enriched in state 8. Promoter states also differed in their preferentially enriched positions with respect to the TSS of associated genes (Fig. 3b). States 4–7 were most concentrated over the TSS (showing upwards of 100-fold enrichment), states 8–11 peaked between 400 bp and 1,200 bp downstream of the TSS and corresponded to transcribed promoter regions of expressed genes and states 1–3 peaked both upstream and downstream of the TSS.

Transcription-associated states

The second large group of chromatin states consisted of 17 transcription-associated states. These are 70–95% contained within RefSeq-annotated transcribed regions compared to 36% for the rest of the genome (Fig. 2b, $P < 10^{-200}$, for all states). This group was not predominantly associated with a single mark, but instead defined by combinations of seven marks, H3K79me3, H3K79me2, H3K79me1, H3K27me1, H2BK5me1, H4K20me1 and H3K36me3 (Fig. 2a). Inspection of the transition frequencies between these states revealed subgroups of states that are associated with 5'-proximal or 5'-distal locations and with different expression levels (Fig. 2c, Supplementary Notes, Supplementary Table 1 and Supplementary Fig. 4).

We observed several states strongly enriched for spliced exons (states 21–25 and 27–28 with 5.7- to 9.7-fold enrichments) (Figs. 2b and 3c and Supplementary Fig. 24). Spliced exons were previously reported to be enriched in several individual marks^{19–21}. In contrast to these previous studies, the combinatorial approach we have taken here shows that

individual marks in spliced exonic states are also frequently detected in several other states that show only a modest 1.3- to 1.6-fold enrichment for spliced exons (e.g., states 12, 13, 14 and 17). This suggests that the chromatin signature of spliced exons is not solely defined by the presence of the previously reported H3K36me3, H2BK5me1, H4K20me1 and H3K79me1 marks, but their specific combinations and the absence of H3K4me2, H3K9me1 and H3K79me2/3.

State 27 showed a 12.5-fold enrichment for transcription end sites (TES) with its enrichment peaking directly over these locations (Fig. 3c). It was characterized both by the presence of H3K36me3, PolII and H4K20me1 and the absence of H3K4me1, H3K4me2 and H3K4me3, distinguishing it from other transcribed states with higher PolII or H3K36me3 frequencies. This suggests a distinct signature for 3' ends of genes for which, to our knowledge, no specific chromatin signature had been described before. This was further validated by a 3.4-fold signal enrichment for the elongating form of PolII surveyed in an independent study²² (Supplementary Fig. 25), even though our input data did not distinguish between the elongating and non-elongating form.

State 28 showed a 112-fold enrichment in zinc-finger genes, which comprise 58% of the state. This state was characterized by the high frequency for H3K9me3, H4K20me3 and H3K36me3 and relatively low frequency of other marks. This specific combination has been independently reported as marking regions of KAP1 binding, a zinc-finger-specific co-repressor, which also shows a specific 44-fold enrichment for state 28 (refs. 23,24). Although the association of H3K9me3 and H4K20me3 with zinc-finger genes has been previously reported⁵, the *de novo* discovery of this highly specific signature of zinc-finger genes illustrates the utility of the methodology and also reveals the additional presence of H3K36me3 and lower frequency of other marks as complementing the signature of zinc-finger genes.

Active intergenic states

The third broad class of chromatin states consisted of 11 active intergenic states (states 29–39), including several classes of candidate enhancer regions, insulator regions and other regions proximal to expressed genes (Supplementary Notes). These states were associated with higher frequencies for H3K4me1, H2AZ, numerous acetylation marks and/or CTCF and with lower frequencies for other methylation marks (Fig. 2a and Supplementary Figs. 2

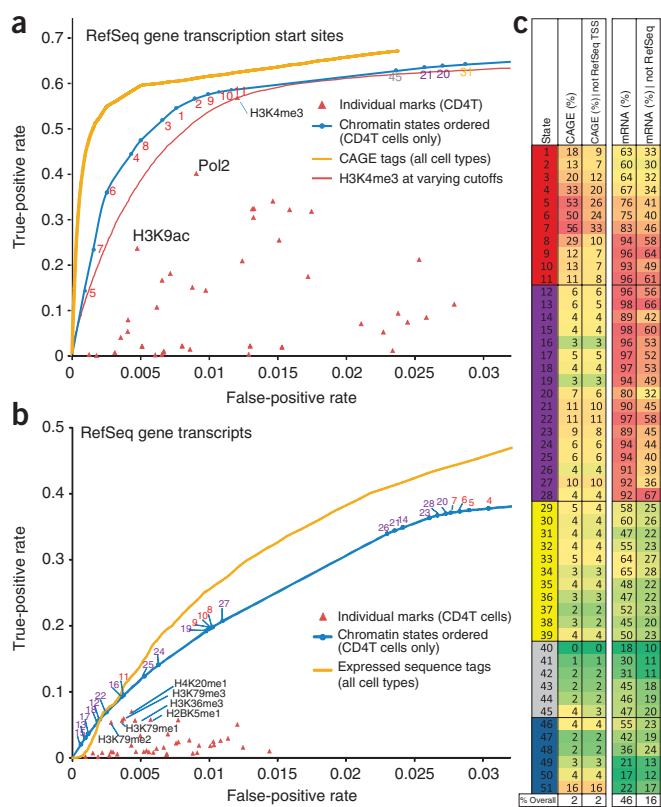


Figure 5 Discovery power of chromatin states for genome annotation. (a) Comparison of the power to discover TSS for individual chromatin marks (red), chromatin states (blue) ordered by their TSS enrichment and a directed experimental approach based on CAGE sequence tag data read counts from all available cell types³⁶ (gold), whereas the chromatin states and marks use only data from CD4 T-cells. Both chromatin states and CAGE tags are compared using a receiver operating characteristic (ROC) curve that shows the false-positive (x axis) and true-positive (y axis) rates at varying prediction thresholds or increasing numbers of states in the task of predicting if a 200-bp interval intersects a RefSeq TSS. Thin red curve compares performance of H3K4me3 mark at varying intensity thresholds. (b) Comparison of the power to detect RefSeq transcribed regions for chromatin states and marks as in a, and directed experimental information coming from EST data (gold) based on sequence counts from all available cell types^{37,38}. (c) Independent experimental information provides support that a significant fraction of false positives in a and b are genuine unannotated TSS and transcribed regions currently missing from RefSeq. Percentage of each state supported by a CAGE tag (column 1), and the same percentage for locations at least 2 kb away from a RefSeq TSS (column 2), suggests that many promoter-associated state assignments outside RefSeq promoters are supported by CAGE tag evidence. Similarly, percentage of each state overlapping a GenBank mRNA (column 3), and the same percentage specifically outside RefSeq genes (column 4), suggest that transcription-associated state assignments outside RefSeq genes are supported by mRNA evidence. Similar support is found by GenBank ESTs and evolutionarily conserved, predicted new exons (Supplementary Fig. 33).

for GWAS hits. In contrast, the surrounding region of the genome is assigned to other active or repressed intergenic states with no significant GWAS association.

Large-scale repressed states

The next group of states (40–45) marked large-scale repressed and heterochromatic regions, representing 64% of the genome. The two most frequently detected modifications in total for all the states in this group were H3K27me3 and H3K9me3. State 40, covering 13% of the genome, was essentially devoid of any detected modifications, states 41–42 (25% of the genome) had a higher frequency for H3K9me3 than H3K27me3, whereas states 43–45 (26% of the genome) had a higher frequency for H3K27me3. States 41–42 as compared to states 43–45 showed a stronger depletion for genes, promoters and conserved elements and stronger association with nuclear lamina regions²⁷ and the darkest-staining chromosomal bands²⁸. It also had a higher frequency of A/T nucleotides (Fig. 2b and Supplementary Figs. 26–28).

State 45 likely corresponds to targeted gene repression. It showed the highest frequency for H3K27me3 and was unique among repressed states to show enrichment for TSS. The corresponding genes were enriched for development-related GO categories (Supplementary Fig. 29), similar to the repressed promoter state 4 marked by H3K4me3. However, in contrast to state 4, state 45 showed almost no change in acetylation levels in response to histone deacetylase inhibitor (HDACi) treatment (Supplementary Fig. 30), suggesting that state 4 is poised for activation whereas state 45 is stably repressed²⁹.

Repetitive states

The final group of six states (46–51) showed strong and distinct enrichments for specific repetitive elements (Supplementary Fig. 31). State 46 had a strong enrichment of simple repeats, specifically (CA)_n, (TG)_n or (CATG)_n (44, 45 and 302-fold, respectively), possibly due to sequence biases in ChIP-based experiments³⁰. State 47 was characterized specifically by H3K9me3 and enriched for L1 and LTR repeats. State 48–51 all had higher frequencies of H4K20me3 and H3K9me3 and were heavily enriched for satellite repeat elements.

and 3). They occurred primarily away from promoter regions (85–97% outside 2 kb of a TSS) and outside of transcribed genes (48–64% outside of RefSeq annotations, Fig. 2b). When they overlapped gene annotations, it was mainly in regions that were repressed or not highly expressed (see expression column in Fig. 2b).

States 29–33 were notable as they corresponded to smaller fractions of the genome specifically associated with greater DNaseI hypersensitivity, transcription factor binding and regulatory motif instances and are likely to represent enhancer regions (Fig. 2 and Supplementary Fig. 23). Although these candidate enhancer states all shared higher H3K4me1 frequencies, they showed differences in the expression levels of downstream genes associated with subtle differences in their specific mark combinations (Supplementary Fig. 22). For instance, genes downstream of state 30 had a consistently higher average expression level than genes downstream of state 31 ($P < 0.001$ at 10 kb, two-sided t -test). The two states differed in the frequency of several acetylation marks (state 30 relative to 31 showed higher frequency for H2BK120ac, H3K27ac and H2BK5ac and lower frequency for H4K5ac, H4K8ac) and also in the level of H2AZ (higher in state 31 than 30), suggesting that these marks may be playing a more complex role than previously thought in enhancer regions.

Several active intergenic states showed significant enrichments for genome-wide association study (GWAS) hits (e.g., 3.3-fold for candidate enhancer state 33, Fig. 4a), based on a curated database of top-scoring single-nucleotide polymorphisms (SNPs) in a range of diseases and traits²⁵. These states thus provide a likely common functional role and means of refining many intergenic SNPs even in the absence of other annotations. For example (Fig. 4b), a SNP reported to be strongly associated with plasma eosinophil count levels in inflammatory diseases (rs12619285)²⁶ and located 40 kb downstream of IKZF2 in an intergenic region devoid of annotations is in a section of the genome in the chromatin state 33, which is enriched

mark combinations and spatial genomic information (**Supplementary Notes** and **Supplementary Fig. 32** for a comparison to *k*-means clustering and a supervised classifier). The prediction performance of chromatin states based on just CD4 T-cells was similar to that of cap analysis of gene expression (CAGE) tags and expressed sequence tags (ESTs) data, even though these were obtained across many diverse cell types. This was possible because active and inactive states together capture the information about genetic elements across cell type boundaries (**Fig. 5** and **Supplementary Figs. 33–35**). Moreover, based on our 51-state model, we could predict TSS and transcribed regions when applied to occupancy data obtained for a subset of ten chromatin marks in CD36 erythrocyte precursors and CD133 hematopoietic stem cells³³ (**Supplementary Fig. 36**).

We also found that chromatin states revealed candidate promoter and transcribed regions not in RefSeq, but further supported by independent experimental evidence. Candidate promoters overlapped with CAGE tags (**Fig. 5c**) and intergenic PolII (**Supplementary Fig. 37**), and candidate transcribed regions overlapped GenBank mRNAs (**Fig. 5c**) and EST data (**Supplementary Fig. 33**). A number of promoter and transcribed states outside known genes were also strongly enriched for not previously described protein-coding exons predicted using evolutionary comparisons of 29 mammals (Lin and M.K., unpublished data) (**Supplementary Fig. 33**). We note that some candidate promoters may represent distal enhancers, sharing promoter-associated marks potentially due to looping of enhancer to promoter regions⁷.

Recovery of chromatin states using subsets of marks

As the large majority of chromatin states were defined by multiple marks, we next sought to specifically study the contribution of each mark in defining chromatin states. First, we found several notable examples of both additive relationships, such as acetylation marks in promoter regions, and combinatorial relationships, such as methylation marks associated with repressive and repetitive elements (**Supplementary Notes** and **Supplementary Fig. 38**). We also evaluated varying subsets of chromatin marks in their ability to distinguish between chromatin states (**Supplementary Notes** and **Supplementary Figs. 39–41**). More generally, we sought to provide guidelines for selecting subsets of chromatin marks to survey in new cell types that would be maximally informative.

As a proof of principle, we evaluated the recovery power of increasing numbers of marks in a greedy way, that is, selecting the best mark given all previous selected marks, weighing each state equally and penalizing mismatches uniformly (see Online Methods), which provided an initial unbiased recommendation of marks to survey for a new cell type (**Fig. 6**). We find that increasing subsets of marks rapidly converge to a fairly accurate annotation of chromatin states (**Fig. 6c**), providing cost-efficient recommendations for new cell types. In addition to an overall error score, this analysis provides information on the proportion of each state accurately recovered, and specific pairwise state misassignments. Such information could be incorporated in a modified scoring function to provide chromatin mark recommendations targeted to the subset of chromatin states that are of particular biological interest, or the particular state distinctions that are most important to each study.

DISCUSSION

The discovery and systematic characterization of chromatin states presented here reveals a diverse epigenomic landscape with 51 functionally distinct chromatin states. Although the exact number of chromatin states can vary based on the number of chromatin marks surveyed and the desired resolution at which state differences are studied, our results suggest that the genome annotation resulting from these states can extend the interpretable part of the human genome, especially outside protein-coding

genes. The definition of the states themselves revealed numerous insights into the combinatorial and additive roles of chromatin marks, sometimes hinting at combinations of chromatin marks that were not previously described, and the genome-wide annotation of these states exposed many previously unannotated candidate functional elements.

We expect the usefulness of the methods presented here will increase as additional genome-wide epigenetic data sets become available, and as additional cell types are surveyed systematically. Chromatin states can be inferred with virtually any type of epigenetic and related information, including histone variants, DNA methylation, DNaseI hypersensitivity and binding of chromatin-associated and sequence-specific transcription factors. Although we focused on a single human cell type, the methods are generally applicable to any species and any number of cell types and even whole embryos, albeit in mixed cell populations mutually exclusive marks found in different subsets of cells could potentially be interpreted as co-occurring.

Specifically for understanding epigenomic dynamics, chromatin states can play a central role going forward, as they provide a uniform language for interpreting and comparing diverse epigenetic data sets, for selecting and prioritizing chromatin marks for additional cell types and for summarizing complex relationships of dozens of marks in directly-interpretable chromatin states. As several large-scale data production efforts are currently underway to map the epigenomes of many more cell types, exemplified by the ENCODE³⁴, modENCODE³⁵ and Epigenome Roadmap projects (<http://www.roadmapepigenomics.org/>), chromatin states will likely play a key role in the understanding of the human epigenome and its role in development, health and disease.

METHODS

Methods and any associated references are available in the online version of the paper at <http://www.nature.com/naturebiotechnology/>.

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

ACKNOWLEDGMENTS

We thank P. Kheradpour for regulatory motif instances and M.F. Lin for predicted new exons. We thank M. Garber, A. Siepel, K. Lindblad-Toh, and E. Lander for use of comparative information on 29 mammals. We thank B. Bernstein, N. Shores, C. Epstein and T. Mikkelsen for helpful discussions. We thank L. Goff, C. Bristow, R. Sealton and all members of the MIT CompBio Group for comments, feedback and support. This material is based upon work supported by the National Science Foundation under award no. 0905968 and funding from the US National Human Genome Research Institute (NHGRI) under awards U54-HG004570 and RC1-HG005334.

AUTHOR CONTRIBUTIONS

J.E. and M.K. developed the method, analyzed results and wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at <http://www.nature.com/naturebiotechnology/>.

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>.

- Bernstein, B.E., Meissner, A. & Lander, E.S. The mammalian epigenome. *Cell* **128**, 669–681 (2007).
- Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).
- Strahl, B.D. & Allis, C.D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).
- Schreiber, S.L. & Bernstein, B.E. Signaling network model of chromatin. *Cell* **111**, 771–778 (2002).
- Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
- Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.* **40**, 897–903 (2008).

7. Heintzman, N.D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
8. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
9. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
10. Hon, G., Wang, W. & Ren, B. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput. Biol.* **5**, e1000566 (2009).
11. Wang, X., Xuan, Z., Zhao, X., Li, Y. & Zhang, M.Q. High-resolution human core-promoter prediction with CoreBoost_HM. *Genome Res.* **19**, 266–275 (2009).
12. Won, K.J., Chepelev, I., Ren, B. & Wang, W. Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics* **9**, 547 (2008).
13. Hon, G., Ren, B. & Wang, W. ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. *PLOS Comput. Biol.* **4**, e1000201 (2008).
14. Day, N., Hemmaplardh, A., Thurman, R.E., Stamatoyannopoulos, J.A. & Noble, W.S. Unsupervised segmentation of continuous genomic data. *Bioinformatics* **23**, 1424–1426 (2007).
15. Jia, L. *et al.* Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS Genet.* **5**, e1000597 (2009).
16. Thurman, R.E., Day, N., Noble, W.S. & Stamatoyannopoulos, J.A. Identification of higher-order functional domains in the human ENCODE regions. *Genome Res.* **17**, 917 (2007).
17. Schuettengruber, B. *et al.* Functional anatomy of polycomb and trithorax chromatin landscapes in *Drosophila* embryos. *PLoS Biol.* **7**, e13 (2009).
18. Jaschek, R. & Tanay, A. Spatial clustering of multivariate genomic and epigenomic information. in *Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology* (ed. Batzoglou, S.) 170–183 (Springer, 2009).
19. Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.* **16**, 990–995 (2009).
20. Kolasinska-Zwierz, P. *et al.* Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat. Genet.* **41**, 376–381 (2009).
21. Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C. & Komorowski, J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res.* **19**, 1732–1741 (2009).
22. Schones, D.E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell.* **132**, 878–898 (2008).
23. Sripathy, S.P., Stevens, J. & Schultz, D.C. The KAP1 corepressor functions to coordinate the assembly of de novo HP1-demarcated microenvironments of heterochromatin required for KRAB zinc finger protein-mediated transcriptional repression. *Mol. Cell. Biol.* **26**, 8623–8638 (2006).
24. O'Geen, H. *et al.* Genome-wide analysis of KAP1 binding suggests autoregulation of KRAB-ZNFs. *PLoS Genet.* **3**, e89 (2007).
25. Hindorff, L.A., Junkins, H.A., Mehta, J.P. & Manolio, T.A. A catalog of published genome-wide association studies. <<http://www.genome.gov/gwastudies>> accessed July 22, 2009.
26. Gudbjartsson, D.F. *et al.* Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. *Nat. Genet.* **41**, 342–347 (2009).
27. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948–951 (2008).
28. Furey, T.S. & Haussler, D. Integration of the cytogenetic map with the draft human genome sequence. *Hum. Mol. Genet.* **12**, 1037–1044 (2003).
29. Wang, Z. *et al.* Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell* **138**, 1019–1031 (2009).
30. Johnson, D.S. *et al.* Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res.* **18**, 393–403 (2008).
31. Zang, C. *et al.* A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**, 1952–1958 (2009).
32. Zhang, Y., Shin, H., Song, J.S., Lei, Y. & Liu, X.S. Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. *BMC Genomics* **9**, 537 (2008).
33. Cui, K. *et al.* Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell* **4**, 80–93 (2009).
34. ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
35. Celnikier, S.E. *et al.* Unlocking the secrets of the genome. *Nature* **459**, 927–930 (2009).
36. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635 (2006).
37. Karolchik, D. *et al.* The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.* **36**, D773–D779 (2008).
38. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Wheeler, D.L. GenBank update. *Nucleic Acids Res.* **32**, D23–D26 (2004).

ONLINE METHODS

Input data for modeling. The initial unprocessed data were bed files containing the genomic coordinates and strand orientation of mapped sequence reads from ChIP-seq experiments^{5,6}. There was a separate bed file for each of the 18 acetylations, 20 methylations, H2AZ, CTCF and PolII in CD4 T cells. We used the updated version of the H3K79me1/2/3 data, as reported⁶, which differs from the version first reported⁵.

To apply the model we first divided the genome into 200-base-pair non-overlapping intervals within which we independently made a call as to whether each of the 41 marks was detected as being present or not based on the count of tags mapping to the interval. Each tag was uniquely assigned to one interval based on the location of the 5' end of the tag after applying a shift of 100 bases in the 5' to 3' direction of the tag. The threshold, t , for each mark was based on the total number of mapped reads for the mark (**Supplementary Table 2**), and was set to be the smallest integer t such that $P(X > t) < 10^{-4}$ where X is a random variable with a Poisson distribution with mean parameter set to the empirical mean of the number of tags per interval.

The probabilistic model. The probabilistic model is based on a multivariate instance of a Hidden Markov Model (HMM)³⁹. The model assumes a fixed number of hidden states K . In each hidden state, the emission distribution, that is the probability distribution over each combination of marks, is modeled with a product of independent Bernoulli random variables. Formally, for each of the K states, and $M = 41$ input marks, there is an emission parameter $p_{k,m}$ denoting the probability in state k ($k = 1, \dots, K$) that input mark m ($m = 1, \dots, M$) has a present call. Let $c \in C$ denote a chromosome where C is the set of all chromosomes. Let c_t denote an interval on chromosome c where $t = 1, \dots, T_c$ corresponds sequentially to the 200 bp intervals on chromosome c . c_1 is the interval corresponding to base pairs 1–200 on chromosome c and T_c is the number of nonoverlapping 200 bp intervals on chromosome c . Let $v_{c_t,m}$ be '1' if there is a present call for input mark m and '0' otherwise at location c_t . Denote the specific combination of marks at interval c_t as $v_{c_t} = (v_{c_t,1}, \dots, v_{c_t,m})$. Let b_{ij} denote the probability of transitioning from state i to j where $i = 1, \dots, K$ and $j = 1, \dots, K$. We also have parameters a_i ($i = 1, \dots, K$), which denote the probability that the state of the first interval on the chromosome is i . Let $s_c \in S_C$ be an unobserved state sequence through chromosome c and S_C be the set of all possible state sequences. Let s_{c_t} denote the unobserved state on chromosome c at location t for state sequence s_c . The full likelihood of all of the observed data v for the parameters a , b and p can then be expressed as:

$$P(v|a, b, p) = \prod_{c \in C} \sum_{s_c \in S_C} a_{s_{c_1}} \left(\prod_{t=2}^{T_c} b_{s_{c_{t-1}}, s_{c_t}} \right) \prod_{t=1}^{T_c} \prod_{m=1}^M p_{s_{c_t}, m}^{v_{c_t, m}} (1 - p_{s_{c_t}, m})^{(1 - v_{c_t, m})}$$

Model learning. We first used an iterative learning expectation-maximization approach to infer state emission and transition parameters that best summarize the observed genome-wide chromatin mark information using a fixed number of randomly-initialized hidden states, varying from 2 up to 80 states. To minimize the number of states and facilitate recovery of a robust and comparable set of states across models of varying complexity, we then used a nested initialization procedure that seeded parameters of lower-complexity models with states of higher-complexity models.

From an initial set of parameters we found a local optimum of the parameter values using a variant of the standard expectation-maximization based Baum-Welch algorithm for training HMMs³⁹. Our variant after the first full iteration over all the chromosomes, used an incremental expectation-maximization procedure⁴⁰, which would update the parameters through a maximization step after performing an expectation over any chromosome. This allowed improved parameter estimates from the maximization step to be more quickly incorporated in the more time consuming expectation step. Also for computational considerations, if a transition parameter fell below 10^{-10} during training we set the parameter value to 0, which allowed faster training with virtually no

impact on the final model learned. The transitions were initialized to be fully connected, and except for the 10^{-10} criterion there was no regularization forcing them closer to 0. We would terminate the training after 300 passes over all the chromosomes, which was sufficient for the likelihood to demonstrate convergence (**Supplementary Fig. 8**).

The procedure for determining the initial parameters used to learn the final set of HMMs was to first learn in parallel for every number of states from 2 to 80 three HMM models based on three different random initializations of the parameters. Each model was scored based on the log likelihood of the model minus a penalization on the model complexity determined by the Bayesian Information Criterion (BIC) of one-half the number of parameters times the natural log of the number of intervals. We then selected the model with the best BIC score among these 237 models, which had 79 states (**Supplementary Figs. 8 and 12**). We then iteratively removed states from this 79-state model. When removing a state the emission probabilities would be removed entirely, and any state that transitioned to it would have that transition probability uniformly redistributed to all the remaining states. This resulting set of models was then used as the initial parameters of the HMMs in the final model learning. During this final model learning, one HMM was learned for every number of states between 2 and 79 in parallel (**Supplementary Fig. 13**).

The criterion for selecting a state to remove from a model was based on first forming a set E containing all the emission vectors from all the 237 models learned from the random initializations. The procedure would then remove a state such that the elements in E had in total the least distance from their closest emission vector among the remaining states. Formally for a set of emission vectors C_n corresponding to states in a model the method would form a set C_{n-1} and corresponding model by removing r defined by

$$\operatorname{argmin}_r \sum_{e \in E} \min_{c \in C_n \setminus r} d(e, c)$$

where here we used $(1 - \rho)$ where ρ is the standard correlation coefficient as the distance d , though the method is general and could be used with other distance measures.

The entire procedure discovered models with comparable or superior likelihood scores to randomly initialized models, while also having sets of parameters that would be more directly comparable (**Supplementary Figs. 8 and 13**). The number of states for a model to analyze can then be selected by choosing the model trained from a nested initialization with the smallest number of states that sufficiently captures all states of offering distinct biological interpretations, which in our case was a 51-state model based on the recovery of the end of transcription state (State 27) (**Supplementary Notes**).

Associating genomic locations with states. After a model is learned, a posterior probability distribution over the state of each interval is computed using a forward-backward algorithm³⁹. Unless otherwise noted, the analysis was based on the 'soft' state assignments of the posterior distribution. We also formed hard assignments of states to locations by using the maximum-posterior state assignment at a location. Both the full posterior and hard assignments are available on the website <http://compbio.mit.edu/ChromatinStates/>.

External fold enrichments and percentage overlaps. For a state the sum of the posterior probability over all 200 bp intervals was computed, denoted by a . For an external data source the total number of 200 bp intervals that it intersects in at least one base was computed, denoted by b . For the state and the external data source the total sum of the posterior for the state in intervals intersecting the external data source were computed, denoted by c . Also the total number of 200 bp intervals is denoted by d . The percentage of a state's overlap with an external data source is defined as $(c/a * 100)$ whereas the fold enrichment is $(c/a)/(b/d)$. P -values of the overlap were computed based on the hypergeometric distribution.

The gene annotations used were the RefSeq annotations⁴¹ as of December 14, 2008 obtained from the UCSC genome browser³⁷ and are based on hg18.



The sequence data for computed nucleotide frequencies, CpG islands, repeats⁴² and conservation data were also obtained from the UCSC genome browser. The conservation data were based on PhastCon conserved elements using the 44-way vertebrate alignment^{43,44} (Lindblad-Toh, K. *et al.*, Broad Institute, unpublished data). Transcription factor binding enrichments were computed for 18 experiments from numerous publications (Supplementary Fig. 23), the median enrichment over all these experiments is reported in Figure 2b. The DNaseI hypersensitivity data was as described⁴⁵ obtained from the UCSC genome browser. The nuclear lamina data of human fibroblasts was obtained from ref. 27. The zinc-finger genes were defined as those that had 'ZNF' at the beginning of the gene symbol in the RefSeq gene table. For published coordinates that were in hg17 we converted them to hg18 using the liftOver tool from the UCSC genome browser⁴⁶.

Expression, motif and gene ontology analyses. We obtained the processed CD4 T expression data from ref. 47 for both replicates. We then averaged the two replicates. After averaging the two replicates we performed a natural log transform of the average values. We then standardized all values by subtracting the mean log transformed value and then dividing by the s.d. of the log transform values. The genome coordinates of each probe set were obtained from the UCSC genome browser. Each 200 bp interval that overlapped a probe set obtained the transformed expression score. If multiple probe sets overlapped the same 200 bp then the average of the expression values associated with these were taken.

We generated transcription factor motif enrichments as described⁴⁸, extended for position-weight matrices (PWMs) (Kheradpour, P., MIT, and M.K., unpublished data) based on the hard state assignments.

Gene ontology enrichments were based on the hard state assignment of the interval containing the RefSeq annotated TSS of the gene. Enrichments were computed using the STEM software (v.1.3.4) and the Bonferroni corrected *P*-values are reported⁴⁹.

SNP and GWAS analysis. The HapMap CEU⁵⁰ data were downloaded from the UCSC genome browser. Significant GWAS hits were taken from ref. 25. SNPs listed as occurring multiple times were only counted once, and for the SNP set listed as a 17-marker haplotype only the first SNP was used giving 1,640 SNPs. In computing enrichment for HapMap and GWAS SNPs, if two SNPs mapped to the same interval, we counted them multiple times. To determine if the number of GWAS SNPs in a chromatin state was more significant than would be expected based on the general SNP frequency in the state we used a binomial distribution where $n = 1,640$ and p is the proportion of HapMap CEU SNPs assigned to the state. We applied a Bonferroni correction for testing multiple states and only reported those *P*-values significantly enriched with $P < 0.01$.

RefSeq TSS and gene transcripts discovery. The ROC curve for the CAGE data was based on the number of CAGE tags mapping to a 200 bp interval retrieved from the Fantom database and converted from hg17 to hg18 using the UCSC genome browser liftOver tool³⁶. The overlap with EST was based on those EST listed in the UCSC genome browser all_est table as of November, 29, 2009 (refs. 37,38). The overlap with GenBank mRNA is based on the overlap with the UCSC genome browser mRNA listed in the table as of October 31, 2009 (refs. 37,38). The novel exon predictions are from (Lin, M.F., MIT, and M.K., unpublished data).

Mark subset evaluation and selection. When evaluating the coverage of a specified subset of marks, first a posterior distribution over the states at

each interval is computed using the model learned on the full set of marks, except that the marks not in the subset are omitted when computing emission probabilities. For an interval t we define here $s_{t,k}$ and $f_{t,k}$ to be the posterior assignment to state k at interval t based on the subset and full set of marks, respectively. The proportion of state k recovered with a subset of marks is defined as:

$$c_k = \frac{\sum_t \min(f_{t,k}, s_{t,k})}{\sum_t f_{t,k}}$$

where the sum is over all intervals t in the genome. The ordering of marks presented without any prior biological knowledge was based on a greedy forward selection algorithm designed to select marks that would minimize this function:

$$\sum_k (1 - c_k)^2$$

where the sum is over all states. At each step the algorithm would then choose the one additional mark, conditioned on all the other previously selected marks that would cause this function to be minimized. We note that this target function considers all nonidentical state assignments to have equal loss. An extension of this approach would be to apply target functions that weigh different misassignments differently. The proportion of state k with the full set of marks that is misassigned to state i using a subset of marks, $m_{k,i}$, as is presented in Supplementary Figures 39 and 40, is defined as:

$$m_{k,i} = \frac{\sum_t \left(\max(f_{t,k} - s_{t,k}, 0) \left(\frac{\max(s_{t,i} - f_{t,i}, 0)}{\sum_j \max(s_{t,j} - f_{t,j}, 0)} \right) \right)}{\sum_t f_{t,k}}$$

The first term in the sum in the numerator represents for an interval t the amount of posterior probability assigned to state k using the full set of marks not assigned using the subset of marks. The second term represents the portion of this posterior probability that will be credited to state i . The portion credited to state i is the proportion of the surplus posterior state i received with the subset of marks in the interval relative to the total surplus posterior all states received in the interval.

39. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. *Biological Sequence Analysis* (Cambridge Univ. Press, 1998).
40. Neal, R.M. & Hinton, G.E. A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learn. Graph. Models* **89**, 355–368 (1998).
41. 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).
42. Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-3.0* 1996–2010 <<http://www.repeatmasker.org>>.
43. Miller, W. *et al.* 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.* **17**, 1797–1808 (2007).
44. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
45. Boyle, A.P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
46. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
47. Su, A.I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).
48. Kheradpour, P., Stark, A., Roy, S. & Kellis, M. Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Res.* **17**, 1919–1931 (2007).
49. Ernst, J. & Bar-Joseph, Z. STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics* **7**, 191 (2006).
50. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).