Published online 12 March 2013 *Nucleic Acids Research, 2013, Vol. 41, No. 8* ***4423–4432***
doi:10.1093/nar/gkt143

Comparative annotation of functional regions in the human genome using epigenomic data

Kyoung-Jae Won[1,2], Xian Zhang[1], Tao Wang[1], Bo Ding[1], Debasish Raha[3], Michael Snyder[3], Bing Ren[4,5] and Wei Wang[1,5,*]

[1]Department of Chemistry and Biochemistry, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0359, USA, [2]Department of Genetics, The Institute for Diabetes, Obesity and Metabolism, Perelman School of Medicine at the University of Pennsylvania, 3400 Civic Center Blvd, Philadelphia, PA 19104, USA, [3]Department of Genetics, Stanford University, 300 Pasteur Dr., M-344, Stanford, CA 94305-5120, USA, [4]Ludwig Institute for Cancer Research, UCSD School of Medicine, 9500 Gilman Drive, La Jolla, CA 92093, USA and [5]Department of Cellular and Molecular Medicine, UCSD School of Medicine, 9500 Gilman Drive, La Jolla, CA 92093, USA

Received September 30, 2012; Revised February 4, 2013; Accepted February 13, 2013

ABSTRACT

Epigenetic regulation is dynamic and cell-type dependent. The recently available epigenomic data in multiple cell types provide an unprecedented opportunity for a comparative study of epigenetic landscape. We developed a machine-learning method called ChroModule to annotate the epigenetic states in eight ENCyclopedia Of DNA Elements cell types. The trained model successfully captured the characteristic histone-modification patterns associated with regulatory elements, such as promoters and enhancers, and showed superior performance on identifying enhancers compared with the state-of-art methods. In addition, given the fixed number of epigenetic states in the model, ChroModule allows straightforward illustration of epigenetic variability in multiple cell types. Using this feature, we found that invariable and variable epigenetic states across cell types correspond to housekeeping functions and stimulus response, respectively. Especially, we observed that enhancers, but not the other regulatory elements, dictate cell specificity, as similar cell types share common enhancers, and cell-type–specific enhancers are often bound by transcription factors playing critical roles in that cell type. More interestingly, we found some genomic regions are dormant in cell type but primed to become active in other cell types. These observations highlight the usefulness of ChroModule in comparative analysis and interpretation of multiple epigenomes.

INTRODUCTION

Identifying cell-type–specific functional regions is an important step to understand the regulatory mechanisms underlying cell-type–specific gene expression. Histone modifications play critical roles in transcriptional regulation (1), and their patterns at enhancers often manifest the cell-type specificity (2,3). With the fast advancement of the next-generation sequencing technology, we have seen explosive accumulation of epigenomic data (3–8), particularly those generated in many different cell types by the ENCyclopedia Of DNA Elements (ENCODE) (9,10) and the NIH Epigenomics Roadmap consortium (11).

The availability of cell-type–specific epigenomic data provides a unique opportunity for genome-wide identification of regulatory regions (2,12–14) or transcription factor (TF)-binding sites (15–17), which in turn helped to predict cell-specific gene expression (18–20). Several computational methods have been developed to annotate epigenomic states using unsupervised learning methods (21–23). For example, Ernst and Kellis (21) used an unsupervised hidden Markov model (HMM) called ChromHMM to define 41 chromatin states using 49 histone marks. These 41 chromatin states were then annotated and grouped into five categories based on the enrichment of known functional sites, such as promoters or DNaseI hypersensitivity sites (DHSs). This approach provides a useful annotation of epigenomic states in the genome, but it also has limitations: for example, binary representation of the data from chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiment may not optimally capture spatial patterns of the epigenomic states, and the number of HMM states needs to be adjusted based on the number of histone marks. Given a deluge of epigenomic data becoming

[*]To whom correspondence should be addressed. Tel: +1 858 822 4240; Fax: +1 858 822 4236; Email: wei-wang@ucsd.edu

available in various cell types, an urgent need is to conduct comparative analysis to reveal the epigenomic landscape underlying the dynamics of transcriptional regulation.

We present here a novel supervised learning method called ChroModule that is based on an HMM with modular structure to annotate epigenomic states in the human genome, which is complementary to the unsupervised learning methods. Inspired by the study of Filion *et al*. (24) that only five major chromatin states were found in *Drosophila*, we chose to train the HMM using six modules to annotate genomic regions of five categories: promoter (forward and backward), enhancer, transcribed region, repressed region and background. Similar to the previous studies (12,15,22), ChroModule exploits mixture Gaussians to model the spatial patterns of the epigenomic data, which is crucial to capture open chromatin regions for potential TF binding. The probabilistic mixture Gaussians also flexibly represent the shapes of epigenomic signals without pre-selecting the number of HMM states. Once the model is trained, it can be applied to any other data set containing same epigenomic data, such as histone modification and chromatin accessibility data, and thus allows direct comparison of epigenomic states across cell types or cellular conditions. We illustrated this feature of ChroModule by training it in one cell type (Huvec) and annotating the other seven cell types without re-training. The predicted promoters/enhancers showed significant overlap with the ChIP-seq peaks of 58 TFs and p300-binding sites, which suggest that ChroModule captures functional regions of the genome.

The annotated regulatory regions in eight cell types provided an opportunity to comparatively analyse epigenomic states. We proposed an epigenomic variation score (EVS) to measure the variation of the epigenomic state across cell types. We observed that epigenetically invariable regions mark fundamental functions in a cell, whereas variable regions were enriched with genes related to cell signalling and response to stimuli. Comparison of active enhancers across the eight cell types identified cell-type–specific enhancers, which show distinct functions as well as putative binding sites of TFs crucial to the corresponding cell type. In addition, we also found that similar cell types share more common enhancers than the dissimilar ones, which is resonant to the concept that similar cell types reside close to each other in the epigenetic landscape. Interestingly, we identified cell-type–specific regulators by comparing the state representation across cell types, which was further validated by the ChIP-seq peaks of the TFs.

MATERIALS AND METHODS

Data

We used the data from the ENCODE project (http://genome.ucsc.edu/ENCODE/) and collected the common markers in eight cell types: GM12878 (lymphoblastoid), Hmec (human mammary epithelial cells), Hsmm (normal human skeletal muscle myoblasts), Huvec (human umbilical vein endothelial cells), K562 (leukaemia), Nhek (normal human epidermal keratinocytes) and Nhlf (normal human lung fibroblasts). The chromatin marks included H3K4me1/2/3, H3K9ac, H4K20me1, H3K27ac, H3K27me3 and H3K36me3. H3K9me1 was not included because it was not available in all the cell types. Additionally, we included chromatin accessibility data [DHSs or formaldehyde-assisted isolation of regulatory elements followed by sequencing (FAIRE-seq) data] (Supplementary Table S1).

The ChroModule model

ChroModule is composed of six modules: promoter (forward and backward), enhancer, transcribed, repressed and background module. ChroModule was trained on Huvec, and each module was trained independently. The HMM in each module has a left–right structure that is widely used in signal processing, such as speech recognition, and has been proved to be effective in capturing temporal patterns (25). We also chose mixture of Gaussians to characterize the shapes of histone modifications because it provides a flexible model to represent the variable profiles of the sequencing reads. Compared with methods that discretize the reads into a limited number of states (21), mixture of Gaussians is able to model a broad range of variability, which is crucial for handling the noise of the sequencing data (15,22).

In the previous work, we evaluated the impact of the number of HMM states and the mixture of Gaussians (the Gaussians are not tied) on the prediction performance of the model (12,15). We found that HMMs with ≥ 3 states and ≥ 2 Gaussians performed much better than HMMs with less number of states and Gaussians. This is because enough number of states/Gaussians can effectively capture the spatial pattern of histone data, such as the bimodal pattern of H3K4me3 at promoters and H3K4me1 at enhancers (15). In ChroModule, we chose to use a five-state HMM with five Gaussians for the promoter and enhancer module because the five-state model could detect not only the strong signals around the peaks but also the weak signals at the flanking regions of the peak (Supplementary Figure S1).

Supplementary Figure S1 shows the heatmap of histone modifications at promoters and enhancers, as well as the emission probabilities of the corresponding HMMs. The bimodal patterns of epigenomic signals in these regions were well represented by the emission probability distributions in the five HMM states. For example, the fourth state in the promoter HMM module modelled the open chromatin region, as shown by the high probability at small read count in H3K4me3 and H3K9ac, and it was flanked by two shoulder peaks represented by the third and fifth states; similarly, the HMM emission probabilities in the enhancer HMM module showed clear depletion of H3K4me1 and other histone marks on the third state, whereas the second and fourth states showed higher emission probabilities to represent shoulder peaks, and the first and fifth state represent the flanking weak signals. It is worth of noting that the fine resolution of the histone profiles captured by the five-state HMM can greatly facilitate further analyses.

We used one-state HMM to model the transcribed (marked by H3K36me3), repressed (marked by H3K27me3) regions and background, which is equivalent to those methods using a two-state HMM to identify the enrichment of these marks (26–28). Supplementary Table S2 summarizes the parameters we chose.

Bioinformatics tools used in the analyses

Homer (29) was used for *de novo* motif finding, and the found motifs were compared with the known motifs in human and mouse documented in the TRANSFAC (30), JASPAR (31) and Uniprobe (32) databases. We used DAVID (33) to perform gene ontology analysis.

RESULTS

ChroModule captures spatial pattern of the epigenomic data in the regulatory regions

ChroModule is composed of HMM modules, each of which has a left–right structure to capture the spatial patterns of the epigenomic signals using mixture of Gaussians to characterize the shapes of histone modifications because it provides a flexible model to represent the variable profiles of the sequencing reads [see detailed discussion in 'Materials and Methods' section and (12,15,22)]. The data used in this study included eight histone marks (H3K4me1/2/3, H3K9ac, H3K27ac, H3K27me3, H3K36me3 and H4K20me1), chromatin accessibility (DHSs) (Supplementary Table S1). As Filion *et al.* (24) only found five distinct chromatin states, we trained six modules on five categories of annotated regions (forward/backward promoter, enhancer, repressed region, transcribed region and background). Initially, each module was trained separately using the Baum–Welch algorithm (34). We then linked all modules to construct the final model in which the transition probabilities were learned from the data (Figure 1A). This modular design allows flexible representation of functional states and precise training of HMMs.

Supplementary Figure S1 shows the heatmap of histone modifications at promoters and enhancers, as well as the emission probabilities of the corresponding HMMs. The bimodal patterns of epigenomic signals in these regions are well represented by the emission probability distributions. The fourth and the third states in the promoter and enhancer HMM modules characterize the open chromatin regions as illustrated by the depletion of histone signals, such as H3K4me3 in promoter (Figure 1B) and H3K4me1 in enhancer, respectively (Supplementary Figure S1). Such a fine resolution of the histone profiles captured by ChroModule can greatly facilitate further analyses. For example, when checking enriched motifs in the enhancers, one can focus on the open chromatin regions decoded as the third state in the enhancer module, which would significantly narrow down the searching space for motif finding (Figure 4).

We used one-state HMM to model the transcribed (marked by H3K36me3) or repressed (marked by H3K27me3) regions, which is equivalent to those methods aiming to identify the enrichment of these marks (26–28). H3K36me3 and H3K27me3 are known marks of transcribed and repressed regions (35), respectively, and were used in the previous study for annotating these regions (36). Even though these two regions are not well defined by only one characteristic mark, it is important to distinguish them from promoters or enhancers, which is especially critical to calculate EVS for comparing epigenetic states across cell types.

ChroModule is a supervised learning model, and we selected the training data that represent the most probable loci belonging to each category (Supplementary Table S2). Because active promoters and enhancers are associated with strong H3K4me3 and H3K4me1/2 marks, respectively (14), we chose TSSs with the highest H3K4me3 to train the promoter module and strongest distal DHS peaks that are also associated with high H3K4me1/2 and low H3K4me3 to train the enhancer module. For the transcribed and regressed regions, we selected the top 1000 exons in chromosome 1 with a high H3K36me3 and H3K27me3 signals (>2 normalized read counts), respectively. We took the entire chromosome 1 to train the background module. We trained ChroModule in Huvec (see Supplementary Methods for details) and applied the trained model to annotating other cell types. We used the Viterbi algorithm (34) to assign HMM states to each 100-bp bin (Figure 1C shows the ChroModule annotation based on epigenome data in K562 and Figure 1D for all cell types).

Especially interesting, ChroModule showed flexibility to capture various types of spatial patterns, such as both uni- and bi-modal patterns: uni-modal enhancers were represented by a sequence of states without visiting third state (Supplementary Figure S2A) in contrast to bimodal enhancers represented with at least one third state, as well as second and fourth states. As uni-modal enhancers were observed at the binding loci of androgen receptor (17,37) because of the dynamic nucleosome positioning, a single model to capture divergent spatial patterns of histone modifications manifests the unique feature of ChroModule to represent diverse chromatin states in a general way. Indeed, when clustering predicted enhancers in Hmec, there are distinct sub-classes of enhancers with diverse combinations of histone modifications (Supplementary Figure S3). We investigated the expression levels of the nearest genes of the bi-modal and uni-modal enhancers (Supplementary Figure S2B) and did not observe statistically significant difference ($P = 0.4$). This observation is not unexpected, as a previous study showed that dynamic changes in nucleosome occupancy are not predictive of gene expression (38). There can be several possible reasons, including that genes are regulated by more than one enhancer, and their expressions are thus not tightly correlated with the dynamics of the nucleosome of one enhancer.

Genome-wide annotation using ChroModule

We observed that enhancers are distributed more broadly than promoters in the genome. ChroModule identified 38 214 (ranging from 21 140 in K562 to 27 237 in GM12878) non-overlapping promoter blocks in the eight

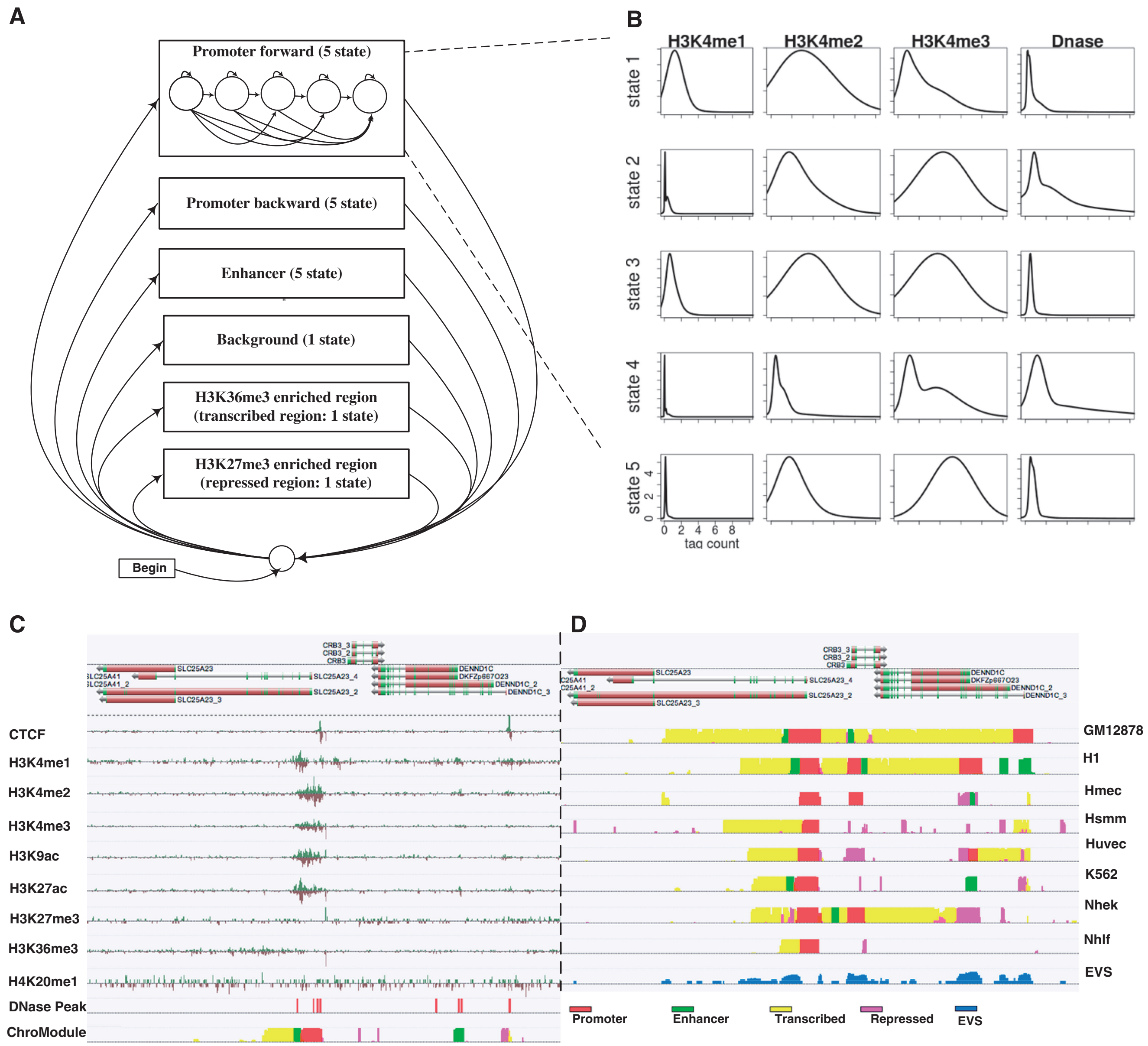

Figure 1. (**A**) The structure of ChroModule. There are six modules in ChroModule: forward promoter, backward promoter, enhancer, H3K36me3-enriched region (transcribed region), H3K27me3-enriched region (repressed region) and background. Each module has a left–right structure, i.e. each state transits to itself or the states located to its right (12,15). (**B**) Emission probabilities of the five-state HMM for promoter. The fourth state represents the open chromatin region of depleted H3K4me1/2/3 and enriched DNaseI signals. (**C**) Example ChroModule annotation and the epigenomic data in the K562 cells. (**D**) Example ChroModule annotation of the eight cell types. The probability of each HMM state and EVS are shown in STAR browser (http://wanglab.ucsd.edu/star/browser) for each cell type.

cell types spanning 89 Mb, compared with 199 200 (ranging from 37 328 in K562 to 66 419 in Nhlf) enhancer blocks spanning 260 Mb of the human genome (Supplementary Table S3). Indeed, similar number of predicted promoters was found across cell types, but the number of enhancers varied significantly. We also noticed that the majority of the human genome was unlabelled (Supplementary Figure S5) often because of insufficient sequencing reads available.

The annotation in the eight ENCODE cell types allowed comparative visualization of the epigenetic states. An example region is shown in Figure 1C and D: the promoter of *SLC25A23* shows invariable epigenetic states, in contrast to the variable *CRB3* promoter. SLC25A23 is a calcium-dependent mitochondrial solute carrier (39), and it is not surprising that it plays roles in many cell types. Crumbs protein homolog 3 (CRB3) functions in epithelial cell polarity (39), resonant to the active histone marks in epithelial (Hmec) and epidermal (Nhek) cells.

ChroModule annotating the genome with a high performance

To access the quality of ChroModule annotation, we evaluated the annotated promoters and enhancers. The promoter annotations in the eight cell types showed consistently satisfactory accuracy, as ~60% of the RefSeq genes were predicted to have active promoters with a false-positive rate <1% (Figure 2A). The remaining 40% RefSeq promoters are missed mainly because of lack of H3K4me3, an epigenetic mark for active promoter. This performance is similar to the previous studies using an unsupervised learning method ChromHMM (21).

To evaluate the accuracy of the predicted enhancers is challenging because of the lack of a gold standard of enhancers. As the binding of transcriptional co-factor p300 or any TF often indicates location of enhancers, we collected the p300-binding sites that are distal (>2.5 kb) from any annotated RefSeq transcription start site (TSS) in H1, K562 and GM12878 cells, as well as 58 TF ChIP-seq experiments (Supplementary Table S1) that determine the binding sites of these TFs (TF-binding sites) in GM12878 and K562 cells. We then evaluated the overlap between the predicted enhancers and p300- or TF-binding sites (Figure 2 and Table 1). As a comparison, we also evaluated the performance of another HMM-based method ChromHMM (21,36) using the same criteria. In contrast to ChroModule, ChromHMM is an unsupervised learning method and discretizes histone modification reads to binary states (presences or absence). We found that ChroModule consistently outperformed ChromHMM (Table 1, Figure 2 and Supplementary Figure S6) in predicting enhancers. For example, ~76% and 83% of all predicted enhancers in GM12878 and K562 cells, respectively, overlap with TF

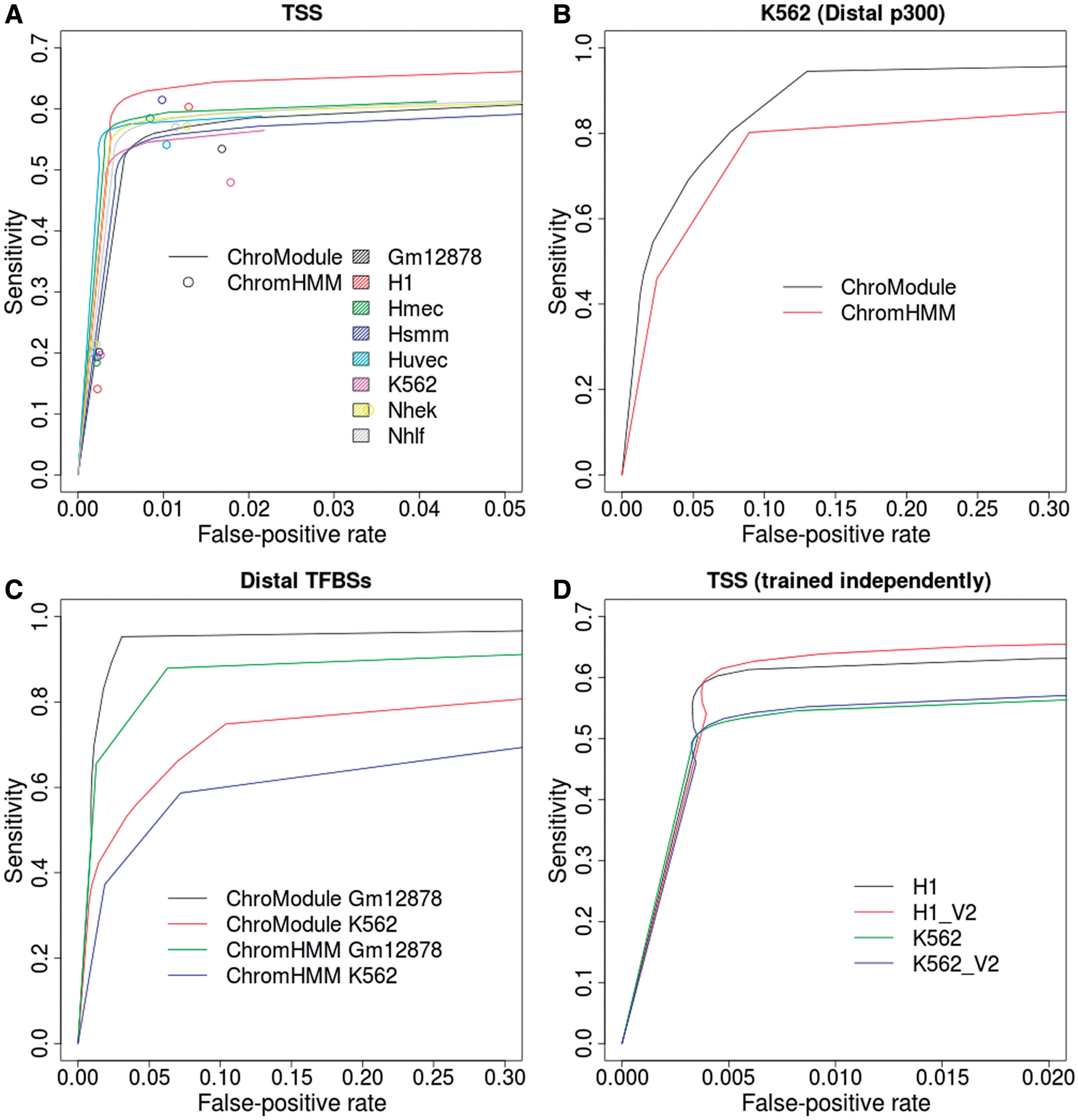

Figure 2. Evaluation of the ChroModule performance on (**A**) promoters (accessed using RefSeq TSSs). ChroModule results (promoters and strong promoters) were obtained from ENCODE (36). (**B**) Assessment of the enhancers predicted by ChroModule and ChromHMM using p300-binding sites that are distal (>2.5 kb) from Refseq TSSs. ChroModule outperformed ChromHMM in all the cell types. ChromHMM results (enhancer, strong enhancer) were downloaded from the study of Ernst *et al.* (36). Supplementary Figure S6 has comparison in H1, K562 and GM12878. (**C**) Assessment of the enhancers predicted by ChroModule and ChromHMM using TF-binding sites in Gm12878 and K562 cells. (**D**) The comparison of ChroModule models independently trained in Huvec and GM12878 (V2). Receiver operating characteristic curves (ROC) curves generated by using RefSeq promoters to evaluate the promoter prediction.

Table 1. Performance of ChroModule and ChromHMM on predicting promoters and enhancers evaluated using RefSeq promoters and distal p300-binding sites, respectively

Cell	Promoter predictions		Distal p300 (enhancer) predictions		
	ChroModule	ChromHMM	Cell	ChroModule	ChromHMM
H1	0.62	0.53	H1	0.77	0.62
GM12878	0.55	0.46	GM12878	0.71	0.65
K562	0.54	0.42	K562	0.84	0.62
Hmec	0.58	0.53			
Hsmm	0.54	0.56			
Huvec	0.57	0.50			
Nhek	0.57	0.52			
Nhlf	0.57	0.52			

Area under curve (AUC) of the ROC curve is shown. The values are scaled to the maximum value.

ChIP-seq peaks that are at least 2.5 kb away from promoters ($P < 10^{-130}$), compared with 68% (42%) and 68% (44%) for the strong (all) enhancers predicted by ChromHMM (Supplementary Table S4). It is also worth of noting that enhancers with relatively large open chromatin region (visiting the third states of the enhancer HMM at least twice) flanked by shoulder peaks (the second and fourth states of the enhancer HMM) have greater overlaps with DHSs (Supplementary Table S3), which further illustrates the advantage of capturing the fine spatial pattern of the chromatin modifications.

To investigate the robustness of ChroModule, we trained the model using the epigenomic data in GM12878 instead of Huvec and tested it on the other two cell lines: K562 and H1 (Figure 2D). Regardless of how the models were trained, ChroModule showed comparable performance. As the robustness of a model is crucial to annotating epigenomes, such a feature of ChroModule makes it a powerful tool for analysing epigenomic data in diverse cell types.

We also checked the portion of the predicted promoters and enhancers that overlap with DHSs measured in the same cell type and different cell types. We found that, although a majority of the predicted promoters and enhancers overlap with DHSs (FAIRE-seq data in Hsmm) in the same cell type, there is a significant increase of overlap percentage (from ~70 to >95% of predicted promoters/enhancers) when considering open chromatin regions in all eight-cell types (Supplementary Figure S4). Although the mechanism underlying this observation is unclear, these genomic regions may be dormant in one-cell type (no open chromatin) but primed to become active (open chromatin) in other cell types.

Epigenomic variation score

To quantitatively define the variation of epigenetic states across the eight cell types, we computed the entropy of ChroModule labels in each 100-bp bin as the EVS $EVS = -\sum P_i \ln(P_i)$, where P_i is the occurrence percentage of promoter, enhancer, transcribed region, repressed region or background labels in all cell types (Figure 1D). We observed 9504 bins with zero EVS (invariable) and 16 876 bins with an EVS of >1.1 (variable). Notably, enhancers and promoters, respectively, show consistently high and low average EVSs across the eight cell types, which indicate the intrinsic variability difference of their epigenetic states (Supplementary Table S5). Transcribed regions show consistently high EVSs, which may be due to alternative splicing in different cell types. It is not surprising that the average EVSs of repressed regions vary on cell types because repressed regions in one cell type might be active in other cell types.

Checking the genes associated with these annotated regions, we found that epigenetically invariable regions are related to housekeeping functions, such as promoters related to 'RNA processing' and 'cell cycle', transcribed regions to 'RNA splicing' and 'translation', enhancers to 'cell death' and 'actin cytoskeleton organization' and repressed regions to 'neuron differentiation'. In contrast, the epigenetically variable regions are related to stimulus response as shown by the enriched gene ontology (GO) terms found by DAVID (33), such as 'cell adhesion' and 'cell–cell signalling' (Table 2).

Enhancers dominate cell type specificity

To investigate which functional regions were critical to determine cell specificity, we calculated the number of mismatches of ChroModule states (promoter, enhancer, transcribed, repressed or background) between cell types and clustered the eight cells using this epigenomic distance as the metric. When using enhancers to compute the epigenomic distance, the pluripotent H1 cell is distinct from the remaining differentiated cells, and the epidermal (Hmec and Nhek) and lymphocytic cells (K562 and GM12878) are close to each other (Figure 3). This cluster of cells resembles the cell-type similarity much better than the clusters generated using promoter alone, promoter plus enhancer or all the annotated regions (Supplementary Figure S8). This observation confirmed that enhancers dictate cell-type specificity (2).

We then conducted GO term analysis (33) on the promoters and the closest genes of the enhancers predicted in all or only one cell type (Table 3 and Supplementary Table S6). The common enhancers across eight cell types are related to 'cell death', consistent with the enriched functions of invariable enhancers with low EVS. Checking the functions related to cell-type–specific enhancers, we observed a strong correlation between the enriched GO

Table 2. Functions of the genes in the epigenetically invariable and variable regions

	Number of blocks	Genes	GO terms (number of genes; *P*-value)
Invariable promoters	9422	8540	RNA processing (429; 1.4 e-65[a]) Cellular macromolecule catabolic process (517; 4.0e-56) Cell cycle (536; 2.0e-51)
Invariable transcribed region	983	668	RNA processing (78;3.3e-21[a]) Translation (54;6.8e-17)
Invariable enhancers	271	238	Cell death (25; 1.8e-3[a]) Regulation of apoptosis (24;1.4e-2)
Invariable repressed region	216	58	Neuron differentiation (14; 5.5e-8[a])
Variable region	16 876	1319	Cell adhesion (86;3.1e-5[a]) Cell–cell signalling (73; 1.8e-4)

DAVID (33) was used to perform GO analysis. We assigned enhancers to their closest gene, and multiple enhancers can be assigned to a single gene. Inside the parenthesis are the number of genes associated with each term and the Benjamini–Hochberg adjusted *P*-value.
[a]The most significant biological process.

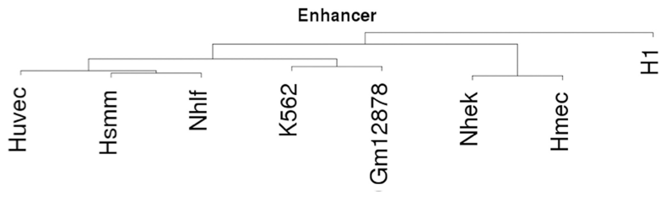

Figure 3. The epigenetic distance between cell types calculated based on the enhancer segmentation using the Pvclust R package (41). Clusters with unbiased $P > 0.95$ are indicated by the rectangles. See Supplementary Figure S8 for other clusters.

terms and the function of the cell. For example, ‘lymphocyte activation’ is the most significant GO term in the lymphoblastoid cell GM12878. The functions of common promoters, such as ‘RNA processing’ and ‘cellular macromolecule catabolic process’, are essential to the cell. The enriched GO terms associated with the cell-type–specific promoters are less well associated with cell specificity than enhancers.

Comparative methods found cell-type–specific master regulators

We searched for enriched sequence motifs in the enhancers using Homer (29) (Figure 4). We restricted the search in the open chromatin regions marked by the third state of the enhancer HMM. In the common enhancers, we found the enrichment of the motif recognized by Fos that regulates diverse biological processes from ‘proliferation and differentiation’ to ‘defence against invasion and cell damage’ (41). For cell-type–specific enhancers, we also found motifs recognized by TFs that specifically function in the corresponding cells. For example, a motif identified in H1 cell is similar to the motif of Oct4 (15), a master regulator of embryonic stem cell (Figure 4). Indeed, we observed much larger portion of Oct4 ChIP-seq–binding sites in the H1-specific enhancers with open chromatin region (78 of 504 H1-specific enhancers with bins marked as the third state in the enhancer HMM) compared with that in all the enhancers (564 of 24 280 enhancers) ($P < 10^{-130}$). Another example is the Pu.1 motif found in GM12878, which is consistent with the observed enrichment of ChIP-seq peaks of Pu.1 in the GM12878 enhancers (13 031 of 39 662 predicted enhancers) or those in the GM12878-specific enhancers with an open chromatin (2423 of 4998). The third example is the identification of the GATA1 motif in K562, whose functional roles in K562 were previously reported in the literature (42). Genome-wide ChIP-seq analysis of GATA1 also showed the enrichment of peaks in K562-specific enhancers (2632 of 6315 enhancers with an open chromatin, $P < 10^{-130}$).

DISCUSSION

With the fast accumulation of epigenomic data, there is an urgent need for global analysis of such data and annotation of the genome in a cell-type–specific manner. The ChroModule method developed in this study provides a useful tool to label functional regions based on epigenetic information. ChroModule has several unique features. First, the design of ChroModule allows separate training of individual modules and then linking these modules to build a full model, which significantly reduces the complexity of model tuning. The modular design of HMM has been successfully applied to biological sequence analysis (43–46). Modular HMM not only allows easy interpretation of the decoding results but also often achieves higher prediction accuracy than non-modular models especially in the biology domain because it is non-trivial to automatically learn the HMM structure from complicated and noisy biological data (47,48). In addition, modular design allows easy extension of the model to represent new biological observations by including additional modules.

Second, ChroModule models the sequencing data directly that avoids the arbitrariness of selecting the cut-off for discretization as done in ChromHMM. As shown in this and previous studies (12,15,22), mixture of Gaussians capture fine spatial patterns that can greatly facilitate follow-up analysis, such as searching for motifs recognized by TFs in the open chromatin states (e.g. the third state of the enhancer HMM module). In addition, as shown in the Supplementary Figure S3, diverse chromatin patterns can be represented by a single module in ChroModule.

Table 3. Cell-type–specific enhancers and the functions of the closest genes

Type	Number of cell-type–specific enhancers	Number of assigned genes	GO terms
Common enhancers	522	435	Cell death (43;2.2E-5[a]) Apoptosis (38;1.2E-5)
H1 specific	21 353	8274	Human embryonic stem cell Neuron differentiation (276;1.7E-23) ([a]) Cell morphogenesis involved in differentiation (169;7.0E-20)
GM12878 specific	19 430	7928	Lymphoblastoid Regulation of lymphocyte activation (107;9.5E-14) ([a]) Regulation of leucocyte activation (116;1.2E-13) Regulation of T cell activation (85;3.5E-11)
Hmec specific	10 224	5159	Human mammary epithelial cell Cell motion (173;3.9E-6) ([a]) Cell adhesion (236;3.9E-6)
Hsmm specific	10 934	5684	Normal human skeletal muscle myoblasts Skeletal system development (144;8.0E-11[a])
Huvec specific	11 383	5492	Human umbilical vein endothelial cell Enzyme-linked receptor protein signalling pathway (145;2.8E-8) ([a]) Blood vessel development (100;9.8E-5)
K562 specific	15 827	7287	Leukaemia Positive regulation of leucocyte proliferation (41;9.4E-5) ([a]) Positive regulation of lymphocyte proliferation (40;9.9E-5)
Nhek specific	8356	4959	Normal human epidermal keratinocytes Cell morphogenesis involved in differentiation (104;1.1E-7[a]) Neuron projection morphogenesis (94; 5.5E-8)
Nhlf specific	16 691	6377	Normal human lung fibroblasts Cell motion (219; 8.9E-11) ([a]) Lung development (53;1.5E-4)

Because multiple enhancers can be assigned to the same gene, the number of assigned genes is often smaller than that of enhancers. We used DAVID (33) for GO analysis. Inside the parenthesis are the numbers of genes in each term and the Benjamini–Hochberg adjusted *P*-value. We selected three biological processes from the significant categories.

[a] The most significantly enriched biological processes.

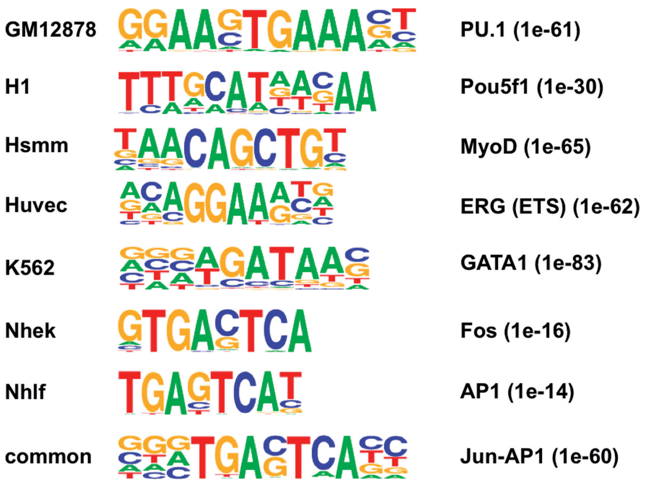

Figure 4. Enriched motifs found by Homer (29).

Third, ChroModule has a small number (five) of functional categories, which does not require the non-trivial process of determining the number of HMM states in the unsupervised learning approach. In addition, the output annotation of ChroModule is easy to interpret without relying on other data or knowledge to annotate the HMM states as in unsupervised learning methods (36).

Unsupervised learning methods have been applied to annotating epingenomes (23,36,49) and searching for combinatorial patterns of epigenetic modifications (50). We conducted a bona fide assessment of the performance of ChroModule, especially on the predicted promoters and enhancers, using p300- and TF-binding peaks. When compared with a state-of-art method ChromHMM, ChroModule consistently showed superior performance in identifying enhancers in different cell types. Unsupervised learning methods can uncover novel features of epigenetic modifications, whereas supervised learning methods can take advantages of the existing knowledge to extract information of interest more accurately. We thus believe ChroModule provides a powerful tool that is complementary to the unsupervised methods, such as ChromHMM and Segway (23,49) in annotating epigenetic states of the cell.

Given the fixed number of functional categories, the epigenetic annotations made by ChroModule can be compared directly across cell types. Taking advantage of such a comparative annotation, we defined the EVS to quantitatively measure the variability of epigenetic state. The functional analyses based on the EVS showed that EVS is a useful metric to define cell-type–specific regulation. Especially, enhancers showed relatively higher EVSs than the other functions regions and enhancers dominate the cell-type specificity. Furthermore, cell-type–specific enhancers are also enriched with motifs of TFs that play important roles in the corresponding cell.

Interestingly, we found some promoters and enhancers are dormant in one-cell type but become active in other cell types, as they do not overlap with DHS peaks in the same cell type but with DHS peaks in other cell types. TF binding or epigenetic marks priming for future gene activation in differentiation has been observed in embryonic stem (ES) cells (51,52). Our observation suggests epigenetic priming may exist more profoundly even in differentiated cells. This hypothesis is waiting for further experimental test.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–6, Supplementary Figures 1–8, Supplementary Methods and Supplementary Reference [53].

ACKNOWLEDGEMENTS

K.J.W. and W.W. conceived and designed the experiments. K.J.W. analysed the data. K.J.W. and X.Z. developed the HMM. T.W. developed the STAR browser. D.R. and M.S. conducted TF ChIP-seq experiments. W.W. analysed the data and supervised the whole study. K.J.W. and W.W. wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

Funding for open access charge: National Institutes of Health [R01GM096194 to W.W. and U01ES017166 to B.R. and W.W. in part].

Conflict of interest statement. None declared.

REFERENCES

1. Berger,S.L. (2002) Histone modifications in transcriptional regulation. *Curr. Opin. Genet. Dev.*, **12**, 142–148.
2. Heintzman,N.D., Hon,G.C., Hawkins,R.D., Kheradpour,P., Stark,A., Harp,L.F., Ye,Z., Lee,L.K., Stuart,R.K., Ching,C.W. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.
3. Hawkins,R.D., Hon,G.C., Lee,L.K., Ngo,Q., Lister,R., Pelizzola,M., Edsall,L.E., Kuan,S., Luu,Y., Klugman,S. *et al.* (2010) Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell*, **6**, 479–491.
4. Schones,D.E. and Zhao,K. (2008) Genome-wide approaches to studying chromatin modifications. *Nat. Rev. Genet.*, **9**, 179–191.
5. Pepke,S., Wold,B. and Mortazavi,A. (2009) Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*, **6**, S22–S32.
6. Hon,G.C., Hawkins,R.D. and Ren,B. (2009) Predictive chromatin signatures in the mammalian genome. *Hum. Mol. Genet.*, **18**, R195–R201.
7. Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
8. Alexander,R.P., Fang,G., Rozowsky,J., Snyder,M. and Gerstein,M.B. (2010) Annotating non-coding regions of the genome. *Nat. Rev. Genet.*, **11**, 559–571.
9. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., Thurman,R.E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
10. Bernstein,B.E., Birney,E., Dunham,I., Green,E.D., Gunter,C. and Snyder,M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
11. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
12. Won,K.J., Chepelev,I., Ren,B. and Wang,W. (2008) Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics*, **9**, 547.
13. Firpi,H.A., Ucar,D. and Tan,K. (2010) Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics*, **26**, 1579–1586.
14. Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Van Calcar,S., Qu,C., Ching,K.A. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.
15. Won,K.J., Ren,B. and Wang,W. (2010) Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biol.*, **11**, R7.
16. Ernst,J., Plasterer,H.L., Simon,I. and Bar-Joseph,Z. (2010) Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res.*, **20**, 526–536.
17. He,H.H., Meyer,C.A., Shin,H., Bailey,S.T., Wei,G., Wang,Q., Zhang,Y., Xu,K., Ni,M., Lupien,M. *et al.* (2010) Nucleosome dynamics define transcriptional enhancers. *Nat. Genet.*, **42**, 343–347.
18. Cheng,C. and Gerstein,M. (2012) Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res.*, **40**, 553–568.
19. Dong,X., Greven,M.C., Kundaje,A., Djebali,S., Brown,J.B., Cheng,C., Gingeras,T.R., Gerstein,M., Guigo,R., Birney,E. *et al.* (2012) Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.*, **13**, R53.
20. Karlic,R., Chung,H.R., Lasserre,J., Vlahovicek,K. and Vingron,M. (2010) Histone modification levels are predictive for gene expression. *Proc. Natl Acad. Sci. USA*, **107**, 2926–2931.
21. Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **28**, 817–825.
22. Kharchenko,P.V., Alekseyenko,A.A., Schwartz,Y.B., Minoda,A., Riddle,N.C., Ernst,J., Sabo,P.J., Larschan,E., Gorchakov,A.A., Gu,T. *et al.* (2011) Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature*, **471**, 480–485.
23. Hoffman,M.M., Buske,O.J., Wang,J., Weng,Z., Bilmes,J.A. and Noble,W.S. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473–476.
24. Filion,G.J., van Bemmel,J.G., Braunschweig,U., Talhout,W., Kind,J., Ward,L.D., Brugman,W., de Castro,I.J., Kerkhoven,R.M., Bussemaker,H.J. *et al.* (2010) Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell*, **143**, 212–224.
25. Rabiner,L.R. (1989) A tutorial on hidden Markov-models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.
26. Hah,N., Danko,C.G., Core,L., Waterfall,J.J., Siepel,A., Lis,J.T. and Kraus,W.L. (2011) A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell*, **145**, 622–634.
27. Li,W., Meyer,C.A. and Liu,X.S. (2005) A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, **21(Suppl. 1)**, i274–i282.
28. Qin,Z.S., Yu,J., Shen,J., Maher,C.A., Hu,M., Kalyana-Sundaram,S. and Chinnaiyan,A.M. (2010) HPeak: an

HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics*, **11**, 369.

29. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
30. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
31. Bryne,J.C., Valen,E., Tang,M.H., Marstrand,T., Winther,O., da Piedade,I., Krogh,A., Lenhard,B. and Sandelin,A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.
32. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
33. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
34. Juang,B.H. and Rabiner,L.R. (1991) Hidden Markov-models for speech recognition. *Technometrics*, **33**, 251–272.
35. Maunakea,A.K., Chepelev,I. and Zhao,K. (2010) Epigenome mapping in normal and disease States. *Circ. Res.*, **107**, 327–339.
36. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
37. Wang,D., Garcia-Bassets,I., Benner,C., Li,W., Su,X., Zhou,Y., Qiu,J., Liu,W., Kaikkonen,M.U., Ohgi,K.A. *et al.* (2011) Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*, **474**, 390–394.
38. Huebert,D.J., Kuan,P.F., Keles,S. and Gasch,A.P. (2012) Dynamic changes in nucleosome occupancy are not predictive of gene expression dynamics but are linked to transcription and chromatin regulators. *Mol. Cell. Biol.*, **32**, 1645–1653.
39. UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
40. Suzuki,R. and Shimodaira,H. (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, **22**, 1540–1542.
41. Shaulian,E. and Karin,M. (2002) AP-1 as a regulator of cell life and death. *Nat. Cell Biol.*, **4**, E131–E136.
42. Huang,D.Y., Kuo,Y.Y. and Chang,Z.F. (2005) GATA-1 mediates auto-regulation of Gfi-1B transcription in K562 cells. *Nucleic Acids Res.*, **33**, 5331–5342.
43. Petersen,L., Larsen,T.S., Ussery,D.W., On,S.L. and Krogh,A. (2003) RpoD promoters in Campylobacter jejuni exhibit a strong periodic signal instead of a -35 box. *J. Mol. Biol.*, **326**, 1361–1372.
44. Sonnhammer,E.L., von Heijne,G. and Krogh,A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 175–182.
45. Henderson,J., Salzberg,S. and Fasman,K.H. (1997) Finding genes in DNA with a hidden Markov model. *J. Comput. Biol.*, **4**, 127–141.
46. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
47. Fujiwara,Y., Asogawa,M. and Konagaya,A. (1995) Motif extraction using an improved iterative duplication method for HMM topology learning. *Pac. Symp. Biocumput*, **96**, 713–714.
48. Stolcke,A. (1994) Bayesian learning of probabilistic language models (doctoral dissertation). University of California at Berkeley.
49. Hoffman,M.M., Ernst,J., Wilder,S.P., Kundaje,A., Harris,R.S., Libbrecht,M., Giardine,B., Ellenbogen,P.M., Bilmes,J.A., Birney,E. *et al.* (2013) Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.*, **41**, 827–841.
50. Bonneville,R. and Jin,V.X. (2013) A hidden Markov model to identify combinatorial epigenetic regulation patterns for estrogen receptor alpha target genes. *Bioinformatics*, **29**, 22–28.
51. Liber,D., Domaschenz,R., Holmqvist,P.H., Mazzarella,L., Georgiou,A., Leleu,M., Fisher,A.G., Labosky,P.A. and Dillon,N. (2010) Epigenetic priming of a pre-B cell-specific enhancer through binding of Sox2 and Foxd3 at the ESC stage. *Cell Stem Cell*, **7**, 114–126.
52. Lin,C., Garruss,A.S., Luo,Z., Guo,F. and Shilatifard,A. (2012) The RNA Pol II elongation factor Ell3 marks enhancers in ES cells and primes future gene activation. *Cell*, **152**, 144–156.
53. Kunarso,G., Chia,N.Y., Jeyakani,J., Hwang,C., Lu,X., Chan,Y.S., Ng,H.H. and Bourque,G. (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.*, **42**, 631–634.