Genomics 102 (2013) 1–7

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

A tiered hidden Markov model characterizes multi-scale chromatin states

Jessica L. Larson [a], Curtis Huttenhower [a], John Quackenbush [a,b], Guo-Cheng Yuan [a,b,*]

[a] Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA
[b] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA

ARTICLE INFO

Article history:
Received 6 November 2012
Accepted 31 March 2013
Available online 6 April 2013

Keywords:
Hidden Markov model
Chromatin
Computational biology

ABSTRACT

Precise characterization of chromatin states is an important but difficult task for understanding the regulatory role of chromatin. A number of computational methods have been developed with varying levels of success. However, a remaining challenge is to model epigenomic patterns over multi-scales, as each histone mark is distributed with its own characteristic length scale. We developed a tiered hidden Markov model and applied it to analyze a ChIP-seq dataset in human embryonic stem cells. We identified a two-tier structure containing 15 distinct bin-level chromatin states grouped into three domain-level states. Whereas the bin-level states capture the local variation of histone marks, the domain-level states detect large-scale variations. Compared to bin-level states, the domain-level states are more robust and coherent. We also found active regions in intergenic regions that upon closer examination were expressed non-coding RNAs and pseudogenes. These results provide insights into an additional layer of complexity in chromatin organization.

1. Introduction

In a multi-cellular organism, virtually all the cells share the same genome, but each cell-type has a distinct gene expression pattern. Chromatin provides an important layer of cell-type specific transcriptional control [1,2]. The basic unit of chromatin is the nucleosome, which wraps a 147 bp sequence of the genome. The nucleosome contains two copies each of four core histone proteins: H2A, H2B, H3 and H4 [3]. Each histone has an N-terminal tail that can be covalently modified at multiple positions. Distinct combinatorial patterns (also known as chromatin states) play important roles in transcriptional regulation [1,2]. As genome-wide histone modification data are being generated in a rapid speed [4–13], there has been a growing interest in developing computational methods to precisely define chromatin states [10,14–18]. Previous methods have mainly focused on detecting local chromatin state variation, whereas large-scale patterns (also known as domains) remain poorly characterized. Nevertheless, epigenetic domains have been identified in various data-types [19–25]. To systematically identify domain patterns from multiple histone marks, we recently developed a hidden Markov model, treating each gene as a separate unit [26]. By applying this method to analyze a collection of ChIP-seq datasets in 27 human cell lines, we found that chromatin states can be used to classify cell-types with high accuracy [27].

Rather than focusing on each length scale separately, it is desirable to characterize multi-scale chromatin states in a single computational framework. To this end, we present a new approach called tiered hidden Markov model (THMM). We tested this approach by analyzing a publicly available ChIP-seq dataset from the Roadmap Epigenome Project [8]. Our analysis identified a two-tiered structure of chromatin states, which we call the bin- and domain-level states. Whereas bin-level states can effectively capture local (200 bp) variation of histone modification patterns, the domain-level state detects large-scale (>1 Kb) variations. We show that this two-tier characterization is useful for better understanding of the regulatory role of chromatin.

2. Results

2.1. Dataset collection and pre-processing

ChIP-seq data from the H1 human embryonic stem (ES) cell line was obtained from the Roadmap Epigenome Project [8] (http://www.epigenomebrowser.org/). Five modifications (H3K4me1, H3K4me3, H3K9me3, H3K27me3 and H3K36me3) with well-known biological functions were chosen for analysis. Raw sequence reads were mapped to non-overlapping 200 bp bins via BEDTools [28] and normalized to have the unit of reads per million reads (RPM). Bins that overlapped 50% or more with known repetitive regions [29] were removed due to possible alignment issues. After removing these highly repetitive regions, the remaining 99.97% bins were analyzed further.

* Corresponding author at: Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA.
E-mail addresses: larsonj5@gene.com (J.L. Larson), chuttenh@hsph.harvard.edu (C. Huttenhower), johnq@jimmy.harvard.edu (J. Quackenbush), gcyuan@jimmy.harvard.edu (G.-C. Yuan).

0888-7543/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2013.03.009

For simplicity, we focused on chromatin state organization around genic regions, and truncated the genome by keeping only the promoter and transcribed regions of protein-coding genes according to Refseq [30]. To be precise, in this study we use the term 'promoter' to include the region 2 Kb upstream from the transcription start site, whereas the 'gene body' is defined as the region from transcription start site to transcription end site. We thus excluded most intergenic regions, which consist of the majority of the genome, from our initial analysis. This truncated genome contains a total of 6,332,441 bins (1.27 Gb).

2.2. Tiered chromatin states in human ES cells

We applied our THMM approach to characterize the chromatin states on the truncated genome in human ES cells based on the five histone modification marks mentioned above. Because of its smaller size, we first determined the optimal number of bin-level states by using the data on chromosome 22. Since the log-likelihood of the model increases monotonically with model complexity, we used permuted data as a control, and evaluated the difference of log-likelihood for observed and permuted data, which was generated by randomly reordering all bin locations on chromosome 22 without changing the corresponding sequence reads. This strategy is similar to the gap-statistic commonly used for K-means clustering [31]. We varied the number of bin-level states from three to twenty eight, and found that the log-likelihood differences between the observed and permuted data plateaus around $K = 15$ (Supplemental Fig. 1), suggesting that the optimal number of bin-level states is around 15. As an additional validation, we found that 94% of truncated genome falls into one of the 15 most abundant combinatorial patterns (Supplemental Fig. 2). We compared three non-degenerative tiered structures that are consistent with this constraint, corresponding to a "3 × 5" model (that is, three domains with five bin-level states per domain), a "4 × 4 " model, and a "5 × 3 " model, respectively. The "3 × 5" model has the best performance but quite similar to the "5 × 3 " model (Supplemental Fig. 3). For simplicity and interpretability, we selected the "3 × 5" model as the final model. We then refined the parameter value estimate by fitting the entire truncated genome (Table 1) and used it for the rest of the analysis in this paper (see Materials and methods for details).

We found certain similarities among the bin-level states associated with a common domain-level state; most bin-level states within a domain share similar histone modification patterns. The bin-level states associated with Domain 1 (States 1–5) are generally associated with high levels of H3K27me3; Domain 2 (States 6–10) is generally absent of all histone marks; while Domain 3 (States 11–15) is enriched with H3K4me3 and H3K36me3 and depleted of H3K27me3 (Table 1 and Fig. 2). Following our previous work [26,27], we annotated Domains 1–3 as non-active, null, and active, respectively.

Next we examined the overall distribution of the domain-level states. The majority (95.6%) of the truncated genome is assigned to the null domain (Fig. 3, Supplemental Fig. 4), which is also the largest on average, with a mean length of 53.9 bins (10.8 Kb), but the domain size is highly variable with a standard deviation (SD) of 79.3 bins (15.9 Kb). In comparison, the active (average length ± SD: 5.4 ± 8.3 Kb) and non-active (average length ± SD: 2.3 ± 2.6 Kb) domains are smaller on average, and also have less absolute variability (though the relative variability is comparable). The null domains are primarily associated with introns, whereas the non-active domains are enriched in the promoter regions (Fig. 4).

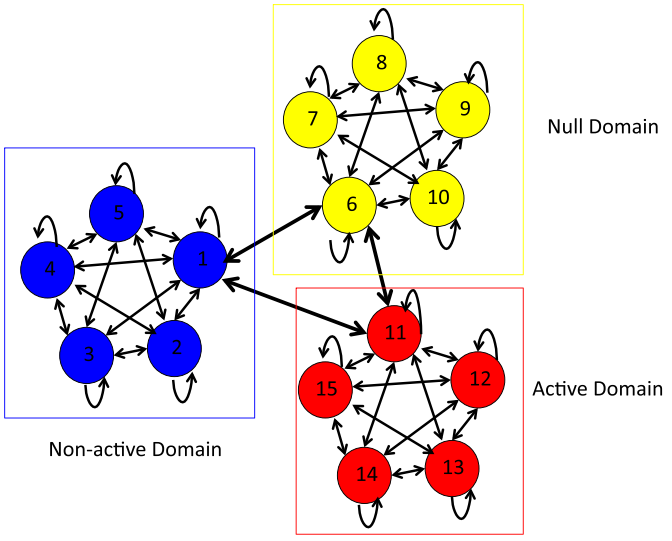

Fig. 1. The topology of our THMM. Each bin-level state is represented by a circled color-coded according to its corresponding domain-level state (represented by a box). Note that transitions between different domain-level states can only occur via a special bin-level state from each domain. States within the null domain are represented by the color light gray; states within the active domain are shown in medium gray; and states within the non-active domain are in dark gray.

While chromatin states are defined based on histone modification data alone, they are useful only if the resulting annotations are also functionally meaningful. It is well known that chromatin plays an important role in gene regulation, and previous studies have shown that active and inactive genes are associated with different sets of histone marks [5]. For example, while H3K36me3 is enriched in highly transcribed genes, the H3K27me3 mark is associated with transcriptionally inactive genes. To test whether our unsupervised chromatin state annotation methods can recapitulate such differences, we analyzed an ES RNA-seq dataset [32], focusing on domain-level states. Raw sequence reads were processed as for the ChIP-seq data and scaled to reads per million reads (RPM). The active domain (States 11–15) is indeed enriched with significantly higher expression levels (average RNA-seq level ± SD: 1.1E4 ± 8.3E4 RPM) compared to other domains (two sample *t*-test versus null and non-active domain p-values < 0.0001) (Fig. 5A), followed by the non-active domains (average RNA-seq level ± SD: 9.8E2 ± 1.7E4 RPM), and the null states have the lowest transcription level (average RNA-seq level ± SD: 3.9E3 ±7.8E3 RPM). These transcription associated changes are consistent with a role of H3K27me3 in gene silencing [33]. Taken together, these results have provided a functional validation of our method.

Table 1
Mean-level ChIPseq counts (RPM) for each chromatin state in the final THMM.

Domain-level	Non-active					Null					Active				
Bin-level	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
H3K4me1	88.3	137.1	131.9	103.4	71.0	24.8	4.0	2.6	3.7	6.6	53.9	4.3	27.3	244.7	102.3
H3K4me3	23.1	194.4	1921.4	35.7	7138.4	16.9	15.0	14.8	15.1	15.2	22.1	15.3	16797.0	5273.4	2017.1
H3K9me3	9.2	148.4	7.5	5.5	5.4	5.3	5.0	2.8	13.2	4.6	77.7	6.0	6.1	398.0	6.3
H3K27me3	51.9	9151.6	1044.3	7.6	36.1	6.4	4.7	4.4	6.0	9.7	7.7	4.8	7.0	173.1	7.6
H3K36me3	6.2	44.5	4.4	7.6	2.4	9.2	23.8	3.3	6.0	2.7	80.2	138.8	4.6	614.1	15.2