BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 17 2008, pages 1911–1916
doi:10.1093/bioinformatics/btn335

Systems biology

Automated mapping of large-scale chromatin structure in ENCODE

Heng Lian[1], William A. Thompson[2], Robert Thurman[3], John A. Stamatoyannopoulos[4,*], William Stafford Noble[4,5,*] and Charles E. Lawrence[2,*]

[1]Division of Mathematical Sciences, SPMS, Nanyang Technological University, Singapore, [2]Center for Computational Molecular Biology, Division of Applied Mathematics, Brown University, Providence, RI, [3]Division of Medical Genetics, [4]Department of Genome Sciences and [5]Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA

Received on May 15, 2008; revised and accepted on 28 June, 2008

Advance Access publication June 30, 2008

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: A recently developed DNaseI assay has given us our first genome-wide view of chromatin structure. In addition to cataloging DNaseI hypersensitive sites, these data allows us to more completely characterize overall features of chromatin accessibility. We employed a Bayesian hierarchical change-point model (CPM), a generalization of a hidden Markov Model (HMM), to characterize tiled microarray DNaseI sensitivity data available from the ENCODE project.

Results: Our analysis shows that the accessibility of chromatin to cleavage by DNaseI is well described by a four state model of local segments with each state described by a continuous mixture of Gaussian variables. The CPM produces a better fit to the observed data than the HMM. The large posterior probability for the four-state CPM suggests that the data falls naturally into four classes of regions, which we call major and minor DNaseI hypersensitive sites (DHSs), regions of intermediate sensitivity, and insensitive regions. These classes agree well with a model of chromatin in which local disruptions (DHSs) are concentrated within larger domains of intermediate sensitivity, the accessibility islands. The CPM assigns 92% of the bases within the ENCODE regions to the insensitive regions. The 5.8% of the bases that are in regions of intermediate sensitivity are clearly enriched in functional elements, including genes and activating histone modifications, while the remaining 2.2% of the bases in hypersensitive regions are very strongly enriched in these elements.

Availability: The CPM software is available upon request from the authors.

Contact: jstam@stamlab.org; noble@gs.washington.edu; Charles_Lawrence@brown.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online. Source code is available at http://noble.gs.washington.edu/proj/segment.

1 INTRODUCTION

In the nucleus of a living cell, genomic DNA is packaged into a complex molecular structure known as chromatin. This structure mediates the interaction between the genome and all types of regulatory and transcriptional molecules. Consequently, understanding chromatin structure is essential if we are to understand how the cell accesses the information stored in the genome.

Perhaps the best understood features of chromatin structure are local disruptions, which correspond to the displacement of a single nucleosome by a collection of transcription factors. These are called DNaseI hypersensitive sites (DHSs) because they are identified using assays that depend upon the small, non-specific endonuclease DNaseI (Crawford *et al*., 2006; Dorschner *et al*., 2004; Keene *et al*., 1981; Sabo *et al*., 2004a, b; Wu, 1980).

We recently developed an array-based assay for measuring chromatin accessibility in a high-throughput fashion, and we have applied it to the ENCODE regions of the human genome (Sabo *et al*., 2006). In addition to identifying classical DHSs, this data gives us the first opportunity to measure experimentally chromatin structure on a large scale.

We undertook this study to characterize DNaseI accessibility via a multi-state model comprised of local regions of similar accessibility. While a hidden Markov model (HMM) makes a natural starting point for this modeling effort, we found that the lengths of segments were not well described by geometric distributions and, as Figure 4 indicates, that the emissions of this HMM could not be effectively captured using Gaussian models. To address these shortcomings, we employed a generalization of the HMM. Specifically, we employed a hierarchical change-point model (CPM) with a continuous mixture of Gaussians at each state and Gamma distributions for the lengths of segments. Because of the large number of observations in this study we employed an empirical Bayesian approach and obtained maximum *a posteriori* (MAP) estimates for the parameters of the hierarchical models and those of the length distribution.

Another important feature of these data is that there are a large number of regions of the genome for which data are missing. These missing blocks of observation arise because the human genome contains a large number of repeated elements (Smit *et al*., 2004). Because the high similarity of the sequences in these repeats would render localization of observations to genome coordinates nearly impossible, probes for these repeat regions were not included in the array design, leading to gaps in the observed data. As shown

*To whom correspondence should be addressed.

subsequently, by treating these gaps as missing data, the resulting algorithm enjoys the advantage of having a linear time and space complexity in the length of the dataset.

The resulting model fits the observed data much better than a HMM. Furthermore, we find that the posterior probability for a four-state CPM approaches 1, and thus the data strongly recommmends a four-state model.

2 METHODS AND RESULTS

2.1 Tiling array data

Data for this study was collected using the ENCODE Nimblegen tiled arrays, using a previously described protocol (Sabo *et al*., 2006). On these arrays, each probe is 50 bp long, and probes are spaced 38 bp apart (i.e. with a 12-bp overlap). Probes that overlap repetitive elements, as identified by RepeatMasker (Smit *et al*., 2004), are not included on the array. The ENCODE Nimblegen array contains 382 884 probes.

To measure DNaseI sensitivity, we treat with DNaseI a sample of intact nuclei and a control sample of bare genomic DNA. We then label the resulting collections of fragments with fluorescent tags (Cy5 and Cy3, respectively) and hybridize the mixture to the Nimblegen array. For a given oligonucleotide probe, a small Cy5 intensity relative to Cy3 intensity indicates that the DNaseI failed to cleave the intact nuclear DNA at that genomic position. Hence, the ratio of Cy5/Cy3 is inversely proportional to DNaseI sensitivity.

For the data used in this study, the assay was performed using the primary Epstein-Barr virus-transformed B-lymphoblastoid cell line GM06990, which is designated as a common reagent by the ENCODE consortium. The resulting collection of 382 884 fluorescence log-ratios, spanning the entire set of ENCODE regions, is available via the UCSC Genome Browser (http://genome.ucsc.edu). Our algorithm models a sequence of fluorescent log-ratios, and the only information about the sequence taken into account by the model is the distance between the probes (for the purpose of spanning the gaps).

2.2 The change point segmentation model

To segment our data, we developed a CPM that uses a recursive algorithm similar to that described by Liu and Lawrence (1999). Here we only give a brief description of the model; details can be found in the Supplementary Material.

Previously, multi-state single Gaussian HMMs have been successfully applied to tiled array data (ENCODE Consortium, 2007; Li *et al*., 2005; Thurman *et al*., 2007). However, we found that this approach produces a poor fit to the DNaseI array data, with many outlier log ratios (see Fig. 4). Such outliers have an adverse impact on state predictions of HMMs in two important ways: first, the outliers from one state can be incorrectly predicted to be members of adjacent states, and second, for many distributions, including the Gaussian, outlying observations have an unduly large impact on parameter estimates characterizing individual states.

The CPM and the HMM both assume that the observed data was generated by a hidden process consisting of a fixed number of hidden states. However, in the HMM, it is common to characterize each hidden state by a single Gaussian distribution. The CPM, in contrast, employs a hierarchical model that uses a continuous mixture of Gaussians at each state. This type of model has been used previously to model oligonucleotide array data at the individual probe level (Ji and Wong, 2005). In a hierarchical model, instead of having a single mean and variance for each state, these two parameters are themselves taken as random variables following a probability distribution. Thus, within each CPM state, there can be different means and variances for different segments (or substrings) of the data, with each substring having its own mean and variance. The hyperparameters of the hierarchial model are global, while those describing individual substrings are local. This structure permits integration over the substring parameters—the means and variances

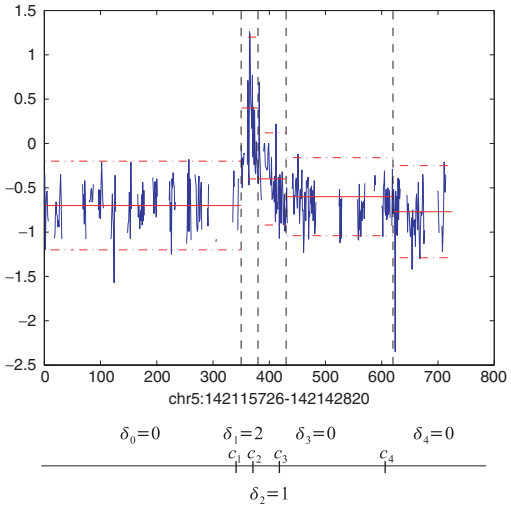

Fig. 1. The figure shows a sample segmentation for a 27 kb region within ENr212. The x-axis is the probe number and two probes with no gaps in between are connected by solid line so that the gaps becomes visually obvious in this figure. The figure contains four change points $(c_1,\ldots,c_4)$ and five regions $(\delta_0,\ldots,\delta_4)$, with each region assigned to a particular model state. Within each region, the mean and SD is shown by horizontal red dashed lines. Note that δ_3 and δ_4 are both in state 0, even though there is a change point at c_4.

of the substring intensities—for given values of the hyperparameters. In this way we marginalize over these high-dimensional missing data. We take an empirical Bayes approach and iteratively find the MAP estimates of the low-dimensional hyperparameters. However, all the means and variances of the substrings of a state are drawn from a common model, described by a specific hierarchical probability distribution. The states differ from one another because the parameters of their hierarchical models differ. In essence, the CPM assumes that within-state sequences are not homogeneous, but instead are characterized by substrings of unknown length, each of which is homogeneous.

For example, consider the data shown in Figure 1. The figure contains three segments in state 0 (δ_0, δ_3 and δ_4), each with its own mean and variance. While the means of these segments differ from one another, all three means are lower than those of the substrings from higher states (δ_1 and δ_2), because the means of segments in state 0 are drawn from a distribution that is shifted toward lower values. Because each state of the hierarchical CPM is described by a family of normal distributions, rather than by a single distribution, the model gains flexibility to capture complex variability in the data.

Figure 1 also illustrates a second important feature of the CPM. Note that the two segments δ_3 and δ_4 are generated by model state 0, even though there is a change point at c_4. Thus, segments of the data from a single state are comprised of subsegments, each following its own Gaussian distribution. This feature, and the fact that distances between change points are not geometrically distributed, are the distinguishing features of the CPM.

Formally, the CPM can be described as follows. Assume for now that we know the distribution on the length of each segment, for each of the states, and we know the transition probabilities between the states. Suppose the maximum number of change points is k_{max}. Denote a segmentation by $A=\{\delta_0,c_1,\delta_1,c_2,\ldots,\delta_{k-1},c_k,\delta_k\}$, $1\le c_1<c_2<\cdots<c_k\le n$, c_i is the i-th change point, and δ_i is the state between its neighboring change points taking value in the set $S=\{0,1,\ldots,D\}$. The prior probability of A before observing the data is

$$p(A)=\prod_{i=0}^{k}p_{\delta_i}(c_{i+1}-c_i+1)\cdot\pi(\delta_0)\prod_{i=0}^{k-1}K(\delta_i,\delta_{i+1}),$$

where $p_\delta(l)$ is the probability of the length of one ungapped data segment being l, given the segment state is δ, and K is the transition probability between states with initial distribution π. Note that we put transition probabilities between segments, instead of between each data point as in the traditional HMM. Also, transitions between the same state are allowed. We set $c_0 = 0$ and c_{k+1} to be the last data point.

Given the segmentation A, the mean and variance for each segment is generated from a normal-inverse-χ^2 distribution:

$$\mu_i|\sigma_i^2, A \sim N(\mu_{\delta_i}^{(h)}, \frac{\sigma_i^2}{k_{\delta_i}^{(h)}})$$

$$\sigma_i^2|A \sim Inv-\chi^2(\nu_{\delta_i}^{(h)}, \sigma_{\delta_i}^{2\,(h)})$$

Those parameters with superscript (h) are hyperparameters that need to be specified.

With segmentation A and μ_i, σ_i given for each segment, the observations are naturally modeled as normal with given mean and variance:

$$y_{c_i+1:c_{i+1}}|\mu_i, \sigma_i^2, A \overset{iid}{\sim} N(\mu_i, \sigma_i^2)$$

where we use the notation $y_{i:j} = \{y_k | k = i, i+1, \ldots, j\}$

In general, the lengths of the segments generated from each model state are not known ahead of time. Consequently, we employ a Bayesian recursive algorithm similar to that described by Liu *et al.* (1995) to infer the change points from the data. The recursive character of this model allows us to sample directly from the posterior distribution of the change points and the changes in state. This property is beneficial because, in discrete high-dimensional settings like the one we face here, there is no assurance that a maximum likelihood or MAP estimate will characterize the posterior space well (Carvalho and Lawrence, 2008). Here our main interest is in a high-dimensional discrete unknown, the assignment of states to all of the probes. In order to make better inferences about these variables we employ a centroid estimator (Carvalho and Lawrence, 2008).

To address the lower dimensional variables of these models, we employ an empirical Bayes approach. In this approach, the low-dimensional parameters of the prior hierarchical models are estimated from the data, rather than being set *a priori*. We estimate these parameters using the expectation maximization algorithm, as detailed in Section 2.

2.3 Spanning the gaps

The analysis of tiled array data is complicated by the absence from the array of many oligonucleotide probes, which correspond to repetitive DNA elements. To handle this missing data, the CPM includes terms that describe the probability of changing state as a function of distance along the chromosome. In general, when a gap is very long, the state probability distribution at one end of the gap is effectively independent of the state probability distribution at the other end of the gap. However, when these gaps are short, we expect adjacent data fragments to affect one another, with an increased likelihood that the bases at either end of the gap will be in the same state. For the CPM we can calculate the magnitude of this effect analytically. Figure 2 illustrates, for our four-state model, the probability of being in a given state at the end of a gap of a given length, given the state at the beginning of the gap. This is calculated using the transition probabilities as well as the distribution of segment lengths estimated with empirical Bayes. Using distributions like those in Figure 2, the CPM accounts for the effects of adjacent data substrings across gaps while retaining linear time complexity of the algorithm with respect to the size of the full dataset.

2.4 The four-state model

Consider first the determination of the number of states. Because of the large number of observations, from over 800 000 probes on the microarray, we needed to employ only a fraction (15%) of the observed probe intensity values to estimate the population parameters of this model. Here 'population parameters' refers to those hyperparameters associated with each state as well

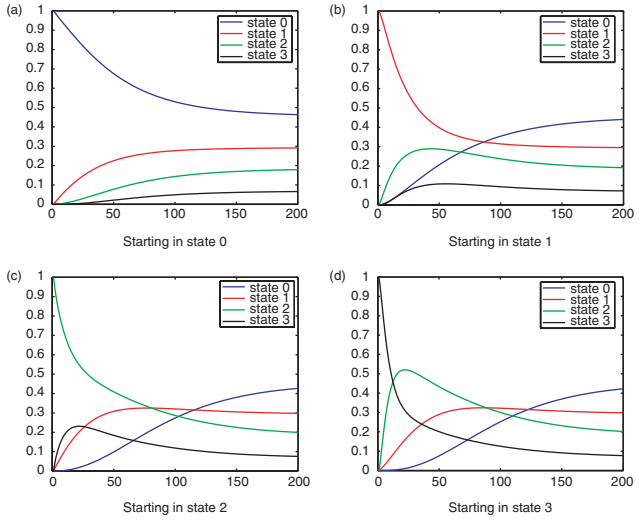

Fig. 2. Each figure plots the probability of being in a given state at the end of a gap of a given length. (**a**) assumes that the data prior to the gap is generated from state 0, and (**b**) assumes state 1, similarly for (**c**) and (**d**). After 200 probes (i.e. 7.6 kb), all probability distributions approach the same equilibrium.

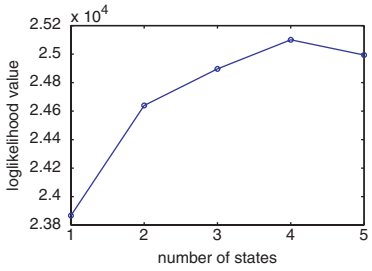

Fig. 3. Selecting the number of states using the likelihood value on the data. The figure plots the log-likelihood value as a function of the number of states in the CPM. The four-state model is optimal by this metric.

as the parameters for the underlying Markov chain since these parameters are not specific to a segment. Using the resulting point estimates with the remaining 85% of the data we marginalize over all the remaining unknowns using recursions of the CPM. Note that the hyperparameters do not need to be marginalized over as in traditional Bayesian model selection, since separate data are used for model comparison in our application. Figure 3 plots the log of the marginal probabilities of the data conditional on these population parameter estimates for models with 1 to 5 states. Assuming that these five models were equally likely *a priori*, the posterior probability of a four state model is nearly 1, with a five state model being the next most probable with a probability $< 10^{-15}$. This indicates, perhaps not surprisingly, that this large amount of data renders a strong preference for just one of these models, the one with four states.

2.5 Quality of the fit to the data

Next, we investigated the extent to which the four-state CPM and a four-state HMM Viterbi segmentation with normal distributions for each state fit the observed data. Both of these models assert that the residuals—the difference between the observed values and the local mean—follow a Gaussian distribution. Figure 4 shows quantile/quantile plots for each of the four states. A quantile gives the value of the residual corresponding to a

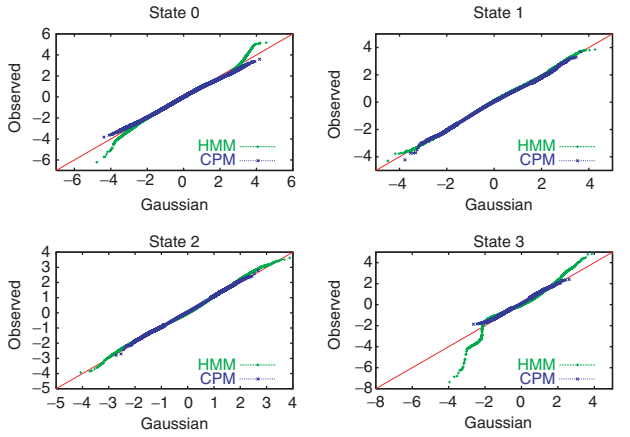

Fig. 4. Each figure plots, for a given model state, the quantiles of the theoretical Gaussian distribution versus the quantiles of the observed residuals. A perfect fit would follow the line $y=x$. In each plot, the two series correspond to the residuals from a four-state, single-Gaussian HMM and from the four-state CPM.

Table 1. Base-level coverage and segment lengths for the four states in the CPM segmentation

	Major DHS (State 3)	Minor DHS (State 2)	Intermediate (State 1)	Insensitive (State 0)
Number of segs	492	521	1085	1062
Number of bases	162 kb	518 kb	1772 kb	28 007 kb
Percent of bases	0.5%	1.7%	5.8%	92.0%
Mean seg length	329	994	1633	26 371
Median seg length	228	646	950	13 718
Hyperprior mean	1.7	0.66	0.0	−0.65
Length parameter	(0.8 475)	(1.4 480)	(1.0 950)	(1.9 1266)

specified cumulative frequency of the residual. For example, the 0.75 quantile of a random variable X gives the value x such that $P(X \leq x) = 0.75$. In these figures, the quantiles of the assumed theoretical Gaussian distribution are plotted against the quantiles of the observed data, whence the axes have been transformed to the Gaussian scale. If the observed data reasonably follow the Gaussian distribution, then the plot will follow a straight line. The HMM parameters were learned using unsupervised expectation maximization. As shown in Figure 4, while the residuals for States 1 and 2 are reasonably modeled by normal distributions for both the HMM and CPM, the residuals for States 0 and 3 depart substantially from this assumption for the HMM but not the CPM.

2.6 Properties of chromatin domains

Table 1 shows the base-level coverage of the four states, as well as the mean and median segment lengths. The median size of the major DHSs agrees well with experimentally observed sizes of DHSs, which generally range from 225–250 bp (Sabo *et al.*, 2004b).

Figure 5 shows histograms of the distribution of segment lengths, and Figure 6 shows histograms of probe intensities and segment mean intensities for the four states. These two figures show that there is considerable overlap between the probe intensities of adjacent states, but little overlap in the mean values of the four states. Thus, Figure 6 supports the use of subinterval averaging of the CPM as an effective means of averaging across noisy probes for these data. A key feature of the CPM is that we average over adjacent probes by using the hierarchical model and do not require that the means and

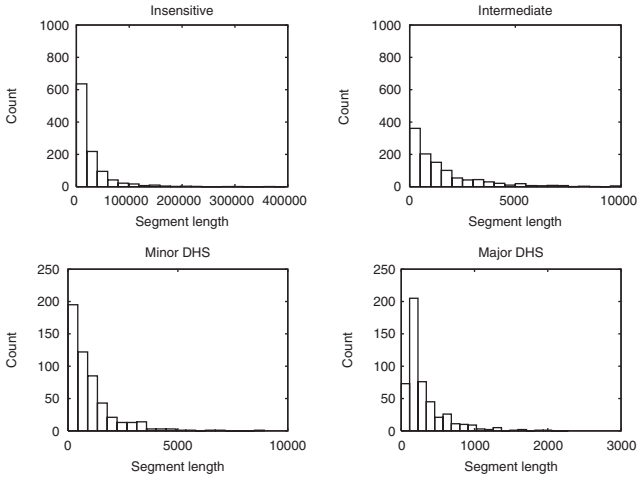

Fig. 5. Length distributions of segments. Each panel plots, for a given CPM state, the distribution of segment lengths.

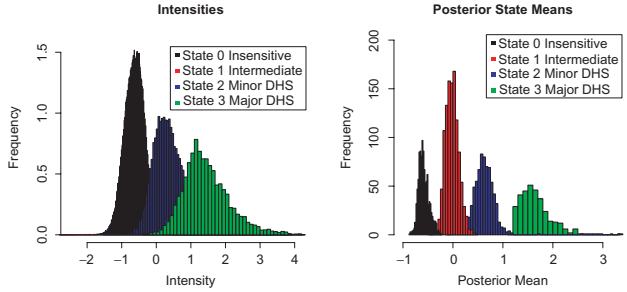

Fig. 6. Intensity distributions of segments. The two panels plot histograms of probe intensities and segment means for the four states, repectively.

Table 2. Transition probabilities estimated from the data, where the rows represent the states being transitioned from

	Insensitive (State 0)	Intermediate (State 1)	Minor DHS (State 2)	Major DHS (State 3)
State 0	0.34	0.66	$1e{-}6$	$1e{-}6$
State 1	0.67	$1e{-}6$	0.33	$1e{-}6$
State 2	0.08	0.45	1e-6	0.47
State 3	$1e{-}6$	$1e{-}6$	0.999	$1e{-}6$

variances of the four classes are fixed, but rather are drawn from state specific models. The strong evidence for four classes also likely stems from this strong separation. The CPM assigns 92% of the bases to the insensitive state, 5.8% to the intermediate sensitivity states and 2.2% to the hypersensitive state.

Table 2 shows the transition probabilities between states. Estimates of transition probabilities were obtained with the training dataset, and these parameters are estimated using an empirical Bayes method, detailed in the Supplementary Material. As the table indicates, there is a strong preference for transitions between adjacent states. Specifically, transitions into a hypersensitive state are almost never permitted from the insensitive state. Thus, hypersensitive sites almost always occur within regions of intermediate sensitivity. The resulting regions of intermediate sensitivity chromatin punctuated by hypersensitivity sites we call accessibility islands.

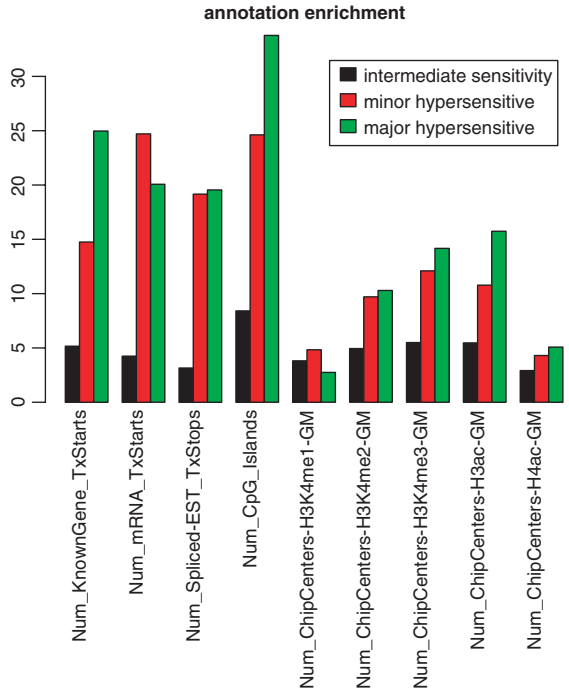

Fig. 7. Enrichment of annotated functional elements in sensitive regions (states 1–3) versus insensitive regions (state 0), as identified by the CPM.

Also notice that about two-thirds of the intermediate regions contain no hypersensitive sites, and that probabilities of returning to the current state are small for all except the insensitive state. We applied a Kolmogorov–Smirnov test for the intensity ratios of each pair of adjacent insensitive segments and find the average P-value to be very small (with more than 90% of the P-values $< 1e-3$). Thus we judge that the statistical properties of such adjacent regions are truly different.

To understand how the CPM segmentation relates to known functional elements in the human genome, we performed an enrichment analysis. We computed the relative enrichment/depletion relative to the insensitive state for each of the other states with respect to functional annotation from the UCSC Genome Browser. These included gene-like elements (KnownGenes, mRNAs and spliced ESTs), CpG islands and a variety of activating histone modifications (H3K4me1–3, H3ac and H4ac). The details of the calculations are explained in the Supplementary Material. As shown in Figure 7, all these functional tracts are moderately enriched in the intermediate state and highly enriched in the hypersensitive states. Furthermore, the major hypersensitive state is more highly enriched than the minor hypersensitive states for all except two of these functional elements.

3 DISCUSSION

We have described a Bayesian model-based method for segmenting DNaseI array data. The method relies upon a hierarchical CPM with hyperparameters estimated using an expectation maximization MAP estimation procedure. We have shown that the model produces a good fit to the observed data, and that the model divides the observed data into four classes. The model assigns 5.8% of the bases to regions of intermediate sensitivity, and these regions are significantly enriched in functional elements, including genes and activating histone modifications. The remaining 2.2% of the bases in hypersensitive regions are even more strongly enriched in these elements. Finally, our finding that transitions to hypersensitive states are very uncommon from the insensitive state supports a model of chromatin in which local hypersensitive disruptions are concentrated within larger domains of intermediate sensitivity, forming accessibility islands.

A recently published analysis of ENCODE datasets used a simple two-state HMM, coupled with wavelet analysis, to perform domain-level segmentation of the ENCODE regions (Thurman *et al*., 2007). That analysis differs from ours in two important respects. First, Thurman *et al*. (2007) analyze multiple data types simultaneously, whereas our analysis focuses on DNaseI. Second, by using wavelet smoothing, the HMM focuses on a single scale at a time, whereas our model simultaneously captures larger- and smaller-scale phenomena. As shown, for example, in Figure 4, the simple HMM does not fit the DNaseI data as well as the CPM. If the same observation holds for other types of data, then a multi-data set analysis using the CPM would likely yield an accurate picture of both large and small genomic domains, without requiring wavelet analysis.

Finally, some caveats are called for with respect to the results presented here. First, while the hierarchical model we employed does permit greater variability in sensitivity among the regions belonging to one state and an associated improvement in the fit of the model, it does come at a cost of two additional free parameters per state. However, this cost seems worthwhile, given the large size of genomic datasets. While the ratio of data to unknowns is favorable for the population parameters describing the states and transitions between states, the dimension of the space describing the unknown locations of change points is extraordinarily large. Thus we cannot expect that the favorable asymptotic characteristics of MAP estimates will be enjoyed in the inference of change points. This is why we employ centroid estimates of these unknowns. Second, although this model is not specific to DNaseI array data and could be applied in a similar fashion to other types of tiling array data, we expect it to be effective only when features of interest span local regions of the genome coordinates. While our findings support this view for accessibility, further study will be required to determine its appropriateness for other data. If this assumption is borne out for other features, then one obvious direction for future work is to simultaneously segment multiple genome-wide assays using a CPM.

ACKNOWLEDGEMENT

Funding: This work was funded by National Institutes of Health award U01 HG003161 and R01 GM071923.

Conflict of Interest: none declared.

REFERENCES

Carvalho,L.E. and Lawrence,C.E. (2008) Centroid estimation in discrete high-dimensional spaces with applications in biology, *Proc. Natl Acad. Sci. USA*, **105**, 3209–3214.

Crawford,G.E. *et al*. (2006) DNase-chip: a high-resolution method to identify DNaseI hypersensitive sites using tiled microarrays. *Nat. Methods*, **3**, 503–509.

Dorschner,M.O. *et al*. (2004) High-throughput localization of functional elements by quantitative chromatin profiling. *Nat. Methods*, **1**, 219–225.

ENCODE Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Ji,H. and Wong,W. (2005) TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics*, **21**, 3629–3636.

Keene,M.A. *et al*. (1981) DNaseI hypersensitive sites in Drosophila chromatin occur at the 5' ends of regions of transcription. *Proc. Natl Acad. Sci. USA*, **78**, 143–146.

Liu,J.S. and Lawrence,C.E. (1999) Bayesian inference on biopolymer models. *Bioinformatics*, **15**, 38–52.

Liu,J.S. (1995) Bayesian models for multiple local sequence alignment and gibbs sampling strategies. *J. Am. Stat. Assoc.*, **90**.

Li,W. *et al*. (2005) A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, **21**, 274–282.

Sabo,P.J. *et al*. (2004) Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc. Natl Acad. Sci. USA*, **101**, 16837–16842.

Sabo,P.J. *et al*. (2004) Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proc. Natl Acad. Sci. USA*, **101**, 4537–4542.

Sabo,P.J. (2006) Genome-scale mapping of DNaseI sensitivity *in vivo* using tiling DNA microarrays. *Nat. Methods*, **3**, 511–518.

Smit,A.F.A. *et al*. (2004) Repeatmasker open-3.0. Available at http://www.repeat masker.org.

Thurman,R.E. *et al*. (2007) Identification of higher-order functional domains in the human ENCODE regions. *Genome Res.*, **17**, 917–927.

Wu,C. (1980) The 5′ ends of *Drosophila* heat shock genes in chromatin are hypersensitive to DNaseI. *Nature*, **286**, 854–860.