DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Hidden Markov Models Predict Epigenetic Chromatin Domains

Citation

Larson, Jessica. 2012. Hidden Markov Models Predict Epigenetic Chromatin Domains. Doctoral dissertation, Harvard University.

Permanent link

http://nrs.harvard.edu/urn-3:HUL.InstRepos:10087396

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

©2012- Jessica Lynn Larson
All rights reserved

Dissertation Advisor: Professor Guo-Cheng Yuan Jessica Lynn Larson

Hidden Markov Models Predict Epigenetic Chromatin Domains

Abstract

Epigenetics is an important layer of transcriptional control necessary for cell-type specific gene regulation. We developed computational methods to analyze the combinatorial effect and large-scale organizations of genome-wide distributions of epigenetic marks. Throughout this dissertation, we show that regions containing multiple genes with similar epigenetic patterns are found throughout the genome, suggesting the presence of several chromatin domains.

In Chapter 1, we develop a hidden Markov model (HMM) for detecting the types and locations of epigenetic domains from multiple histone modifications. We use this method to analyze a published ChIP-seq dataset of five histone modification marks in mouse embryonic stem cells. We successfully detect domains of consistent epigenetic patterns from ChIP-seq data, providing new insights into the role of epigenetics in long-range gene regulation.

In Chapter 2, we expand our model to investigate the genome-wide patterns of histone modifications in multiple human cell lines. We find that chromatin states can be used to accurately classify cell differentiation stage, and that three cancer cell lines can be classified as differentiated cells. We also found that genes whose chromatin states change dynamically in accordance with differentiation stage are not randomly distributed

across the genome, but tend to be embedded in multi-gene chromatin domains. Moreover, many specialized gene clusters are associated with stably occupied domains.

In the last chapter, we develop a more sophisticated, tiered HMM to include a domain structure in our chromatin annotation. We find that a model with three domains and five sub-states per domain best fits our data. Each state has a unique epigenetic pattern, while still staying true to its domain's specific functional aspects and expression profiles. The majority of the genome (including most introns and intergenic regions) has low epigenetic signals and is assigned to the same domain. Our model outperforms current chromatin state models due to its increased domain coherency and interpretation.

Contents

Title Page ..... **i**

Abstract ..... **iii**

Table of Contents ..... **v**

List of Figures ..... **vii**

List of Tables ..... **x**

Acknowledgements ..... **xi**

Preface ..... **xii**

1 Epigenetic domains found in mouse embryonic stem cells via a hidden Markov model **1**

1.1 Introduction **2**

1.2 Results **4**

1.2.1 Data type, preliminary manipulation and clustering 4

1.2.2 Prediction of epigenetic domains 6

1.2.3 Functional coherence of the predicted domains 10

1.2.4 Validation against known domains 13

1.2.5 Domain changes in neural progenitor cells 16

1.3 Discussion **18**

1.4 Methods **20**

1.4.1 Gene-level summary score 20

1.4.2 Clustering 20

1.4.3 Hidden Markov model 21

1.4.4 Significance of detected domains 23

1.4.5 Numerical simulations 24

1.4.6 Gene Ontology analysis 24

1.4.7 Gene expression data analysis 25

1.4.8 Epigenetic domains in NP cells 25

1.5 References **26**

2 Chromatin states accurately classify cell differentiation stages **30**

2.1 Introduction **31**

2.2 Results **33**

2.2.1 An assembly of genome-wide data for 5 histone modifications in 27 human cell lines 33

2.2.2 Histone modification patterns accurately classify cell differentiation status 35

2.2.3 Numerous locations of epigenetic difference found between cell lines 36

2.2.4 Coordinated switches in chromatin domains 40

2.2.5 Histone modification patterns in cancer cells are similar to differentiated cell types 42
2.2.6 Intrinsic chromatin state variability is associated with multiple human diseases 45
2.3 Discussion 46
2.4 Methods 48
2.4.1 Raw data processing 48
2.4.2 Cell differentiation stage classification 50
2.4.3 A simple classification model based on three regulatory modules 51
2.4.4 Identification of differential regions 52
2.5 References 53

3 Tiered hidden Markov models predict epigenetic chromatin domains 57
3.1 Introduction 58
3.2 Results 59
3.2.1 Data pre-processing 59
3.2.2 Model selection 60
3.2.3 Tiered hidden Markov model 61
3.2.4 Domain properties 63
3.2.5 Differences between sub-states within each domain 69
3.2.6 Comparison with Ernst et al model 70
3.2.7 Genome-wide predictions 72
3.3 Discussion 75
3.4 Methods 76
3.4.1 Data collection and pre-processing 76
3.4.2 Model exploration and evaluation 77
3.4.3 Our tiered hidden Markov model 79
3.4.4 Ernst et al. hidden Markov model 80
3.4.5 Quantifying domain retention 80
3.4.6 Active exon agreement 81
3.5 References 82

Supplementary Figures 85

Supplementary Tables 88

List of Figures

Figure 1.1: Outline of methods **5**
Brief four-step depiction of methods used in this study to find chromatin domains.

Figure 1.2: Results from gap statistics analysis of a random sample of the Data **6**
The expected and observed *log* (W_K) values are shown in (a) for various levels of *K* (the number of clusters), where W_K is the pooled within cluster sum of squares around the cluster means. The number of clusters versus *Gap(K)*, the difference between the observed and expected values (mean value for 1000 random bootstrap permutations), is shown in (b). According to these results, three is the optional number of clusters for our data.

Figure 1.3: Histogram distribution of significant domains **8**
Distribution of significant domains for the (a) two-state HMM and (b) three-state HMM. X-axis corresponds to the domain size; y-axis corresponds to frequency of observation.

Figure 1.4: Average within-domain variance of a modification versus a random distribution **11**
Two-state model (K=2) results (and corresponding permutation test p-values) are shown in red, three-state in blue (K=3), random distribution shown in black. For four of the modifications, the HMM domains have a significantly lower average variance (H3K4me2, H3K4me3 & H3K27me3: p-values<0.001), suggesting that the HMM has produced coherent domain bounds.

Figure 1.5: Coherence within predicted chromatin domains **13**
The values (and corresponding permutation test p-values) for predicted chromatin domains (red for two state model, K=2, and blue for three, K=3) are compared with the distribution estimated from 1000 random permutations (black). Shown is the average within domain variance in gene expression (left) and the average level of Gene Ontology accordance as measured by Fleiss' Kappa (right) under the null hypothesis of incoherent domains. Note that our models result in significantly different values.

Figure 1.6: Heatmaps of known gene clusters **15**
The 35 gene region on Ch6 from *Npy* to *2410066E13Rik* (49,772,728 to 54,650,400) as depictured as a heatmap of histone modification and gene expression for (a) the ES cell line and (c) ES and NP cell lines. The 250 gene region on Ch7 from *Art2a* to *Insc* (108,701,290-121,993,728) as depictured as a heatmap of histone modification and gene expression for (b) the ES cell line and (d) ES and NP cell lines. For all figures on the left (a and c), cluster state assignment is given in the first and second tracks and HMM state assignments are in the third and fourth tracks (red for active state, blue for non-active state, yellow for null state). For figures on the right (b and d), NP HMM state assignment is given in the first and second tracks and ES HMM state assignments are in the third and

fourth tracks. Whether (black) or not (white) a gene is a respective gene cluster in shown in the bottom track in all figures.

Figure 1.7: DAVID analysis of genes within each ES significant domain 16
Genes in each state are described by the top three significant DAVID clusters. Red corresponds to genes in the active state, blue for those in the non-active state and yellow for null state.

Figure 2.1: Overview of the data analysis strategy 33
ChIP-seq data of 5 histone modifications in 27 human cell lines were obtained from the public domains and analyzed independently using four different representations (bin, gene score, and chromatin state level). For each representation, a support vector machine (SVM) model was used to classify cell differentiation status from histone modification data. "Hotspot" genes or bins were detected by ANOVA and further investigated by functional genomic tools. The SVM model obtained from normal cell lines was used to classify the differentiation status of cancer cell lines.

Figure 2.2. Differentiation related variation of histone modification patterns 38
(a) Heatmap showing the gene-level H3K4me1 scores for the 100 most significantly different genes. (b) Heatmap of the chromatin states for the 722 "hotspot" genes, whose chromatin states are significantly different across differentiation statuses. Red – active state; yellow – null state; blue – non-active state. The cell line information is shown at both sides of the heatmap and color-coded by the differentiation status (black – pluripotent cells (P); red – multipotent (M); green – unipotent/differentiated (U/D)).

Figure 2.3: Functional enrichment analysis of the "hotspot" genes 40
A total of 722 "hotspot" genes were identified by applying ANOVA analysis to the chromatin states. Enriched functions and pathways were identified by using DAVID. (a) Representative enriched functional categories. (b) The Hedgehog signaling pathway is significantly enriched (p-value = 0.008). The genes were color-coded based their corresponding adjusted p-values obtained from ANOVA analysis.

Figure 2.4: Representative chromatin domains identified by the hidden Markov model 42
Heatmap of the chromatin state distribution at the local genomic loci. (a) the HOXB gene cluster; (b) The histone gene cluster. Genes are ordered according to their genomic positions. Red – active state; yellow – null state; blue – non-active state. The cell line information is shown at both sides of the heatmap and color-coded by the differentiation status (black – plutipotent (P) cells; red – multipotent (M); green – unipotent/differentiated (U/D)).

Figure 2.5. Cancer cells display similar chromatin state patterns as fully differentiated cells 45
(a) A scatter plot of cell-type specific chromatin states associated with the three ES regulatory modules. The chromatin state of a module is summarized by the fraction of non-active-state genes. Each data point corresponds to one cell-type and is color-coded

according to the differentiation status. The three cancer cells are labeled. (b) Hierarchical clustering of the 27 cell lines based on the chromatin states.

Figure 3.1: The topology of our THMM **62**
Each circle represents a sub-state within a boxed domain. Non-active domain states are shown in blue, null in yellow and active domain states are in red. Note that only one state per domain can lead to an inter-domain change. The rest of the states can only transition within their respective domains.

Figure 3.2: The distribution of the modifications in each state plotted as a Z-score 64
Z-scores for each modification within each state are calculated as follows:

$$Z_{ik}=(X_{ik}-\mu_i)/\sigma_i,$$

where X_{ik} is the average count for modification $i=1,2,3,4,5$ within state $k=1,2,\ldots15$, μ_i is the average modification level in the truncated genome, and σ_i is the standard deviation of modification i in the truncated genome.

Figure 3.3: The distribution of domain assignment for the truncated genome **65**
The vast majority of bins are assigned to the null domain.

Figure 3.4: Location of domains and states **66**
The proportion of each (a) domain and (b) state found in promoter (green), introns (purple) and exons (black) regions.

Figure 3.5: Transcriptional activity for each state **67**
The average RNA-seq values for bins assigned to each state. Vertical lines represent the mean value for each domain. The bottom track is the truncated genome-wide average.

Figure 3.6: Functional analysis of each domain **68**
Representative enriched functions and pathways identified using DAVID. P-values are adjusted for multiple testing.

Figure 3.7: Comparison of our model to that of Ernst et al [3.14] **72**
(a) Auto-correlation of state or domain agreement between bins over genetic distance, for our THMM as 15 states and 3 domains and EK's model as 15 states and 3 clusters. X-axis is the lag in number of 200bp bins, y-axis is the corresponding kappa statistic, which is a measure of accordance. Note that our domain-based model (orange) has the highest level of concordance over any genetic distance. Even at a distance of 2KB, we retain a kappa statistic of 0.6, compared to the random model that has no agreement. (b) Agreement, as measured by kappa, between the status of our active domains and each of the EK clusters and the status of our active exons; the black line is the corresponding kappa of a random domain assignment.

Figure 3.8: The distribution of domain assignment for the intergenic genome **73**
98.8% of the truncated genome is assigned to the null domain, 0.83% to the non-active domain, and 0.36% to the active domain.

List of Tables

Table 1.1: **7**
Summary statistics for significant domains in the two- state HMM

Table 1.2: **8**
Summary statistics for significant domains in the three- state HMM

Table 1.3: **9**
Average (variance) histone modification activity within a state for the two-state HMM

Table 1.4: **9**
Average (variance) histone modification activity within a state for the three-state HMM

Table 2.1: **43**
Outcome of classifying the differentiation status of three cancer cell lines (K562, HeLa, and VCaP) by applying the support vector machine to histone modification data at different levels. The results are represented as three numbers, corresponding to the number of models for which the cell line classified as pluripotent (P), multipotent (M), or unipotent/differentiated (U/D), respectively.

Table 3.1: **63**
Posterior means for each state. The active domain is associated with active marks, the non-active domain is associated with the repressive mark H3K27me3, and the null domain is not associated with any mark examined here.

Acknowledgements

I would like to thank to my adivsor, Guo-Cheng Yuan, for his unceasing guidance and support. Without his encouragement and mentoring, this project would never have been successful.

Thank you to my committee members Curtis Huttenhower and John Quackenbush. I am grateful to them for their invaluable comments and questions regarding my research, and for constantly reminding me about the 'big picture.'

I will always be thankful for this experience for bringing Lauren Hund, Layla P. Bartroff and Jessica Minnier into my life. I would also like to thank Dave Zhao and Brian Claggett for consistently challenging and encouraging me.

Thank you to my parents Dan and Darlene and to my sister Sharon. I am thankful that you have constantly pushed and exposed me to the best educational opportunities imaginable. Finally, I would also like to thank my husband, Abraham Howland, for always supporting me in all my academic and personal endeavors.

Preface

Every cell of a multi-cellular organism shares the same genetic sequence, but is programmed with cell-type specific gene expression, resulting in distinct cell types with corresponding specialized functions. This precise expression is, in part, directed by epigenetics, including histone modifications. In eukaryotic cells, 147 bp of DNA is wound around an octomer of histone proteins before being further compacted to form higher-order chromatin structures. These histone proteins have tails that can be biochemically modified to affect the affinity of transcriptional machinery, thereby regulating gene expression. Numerous histone modifications modulate the interpretation of, but do not change, the primary genetic sequence, providing an epigenetic regulatory mechanism to maintain cellular identity and developmental changes.

Recent studies have shown significant epigenetic patterns associated with developmental stages and diseases. These studies have been mostly limited to focal epigenetic patterns, whereas methods for analyzing large-scale organizations are still incomplete. In each of the three chapters of this dissertation, we present a hidden Markov model (HMM) for characterizing the eukaryotic genome into higher-order chromatin domains. Our contributions are three-fold: First, in Chapter 1, we develop and explore an HMM to find chromatin domains in mice stem cells. Second, in Chapter 2, we extend this model to 27 human cell lines to find epigenetic differences between differentiated and undifferentiated cells. In the final chapter, we build a more sophisticated tiered HMM to further annotate the human genome into chromatin domains.

In the first chapter, we developed a HMM approach for detecting the types and locations of epigenetic domains from multiple histone modifications. We used this method to analyze a published ChIP-seq dataset of five histone modification marks (H3K4me2, H3K4me3, H3K27me3, H3K9me3, H3K36me3) in mouse embryonic stem (ES) cells. We identified three types of domains, corresponding to active, non-active, and null states. These domains were validated by a number of statistical criteria in Section 1.2.3. The largest domains correspond to olfactory receptor (OR) gene clusters. Each *Hox* gene cluster also forms a separate epigenetic domain. We found that each type of domain is associated with distinct biological functions and structural changes during early cell differentiation. We describe these features in detail in Section 1.2.5. The HMM approach successfully detects domains of consistent epigenetic patterns from ChIP-seq data, providing new insights into the role of epigenetics in long-range gene regulation.

While recent studies have identified global chromatin state changes across cell-types, it remains unclear to what extent these changes are co-regulated during cell-differentiation. In the second chapter, we present a comprehensive computational analysis by assembling a large dataset containing genome-wide occupancy information of 5 histone modifications in 27 human cell lines (including 24 normal and 3 cancer cell lines), followed by independent analysis at three different representations. We classified the differentiation stage of a cell-type based on its genome-wide pattern of chromatin states, and found that our method was able to identify normal cell lines with nearly 100% accuracy. In Section 2.2.4, we note that the "hotspot" genes, whose chromatin states change dynamically in accordance to the differentiation stage, are not randomly

distributed across the genome but tend to be embedded in multi-gene chromatin domains, and that specialized gene clusters tend to be embedded in stably occupied domains. In Section 2.2.5, we applied our model to classify the cancer cell lines and found that each can be unequivocally classified as differentiated cells. The differences can be in part explained by the differential activities of three regulatory modules associated with embryonic stem cells.

Several studies, including those presented in Chapters 1 and 2, have developed methods to characterize the combinatorial pattern of multiple histone modification marks; however, these methods have not taken into account the fact that the genomic distribution of histone modification marks is not uniform but varies over multiple length scales. In Chapter 3, we developed a new, tiered hidden Markov model (THMM) approach to identify higher-order chromatin state structures. By applying our approach to a set of ChIP-seq data in human embryonic stem cells, we have identified three chromatin domains each containing five sub-states. We again found that the active domain is associated with high levels of RNA transcription, whereas the null domain is associated with introns and intergenic regions. In Section 3.2.6**,** we show that our model outperforms existing chromatin state models in terms of spatial and functional coherence. Our THMM approach to partition chromatin states provides new mechanistic insights into epigenetic regulation.

Chapter 1

Epigenetic domains found in mouse embryonic stem cells via a hidden Markov model

Jessica L. Larson[1] and Guo-Cheng Yuan[1,2]

[1]Department of Biostatistics, Harvard School of Public Health,
Boston, Massachusetts, USA
[2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute,
Boston, Massachusetts, USA

1.1 Introduction

Well before the first eukaryotic genome was sequenced, the notion that chromatin is partitioned into larger than gene-size domains (such as the heterochromatin and euchromatin) was conceived [1.1, 1.2]. Genes that are functionally related and co-regulated are often located close to each other. These include the *Hox* and the *β-globin* gene clusters [1.3, 1.4]. More generally, unrelated spatially proximal genes can still be co-regulated, and such co-regulation can in part be explained by long-range chromosomal interactions [1.5]. However, a major barrier for understanding the mechanism for long-range gene regulation is the difficulty of generating high-resolution long-range chromosomal interaction data on a genomic scale [1.6].

The fundamental unit of chromatin is the nucleosome, which is a histone octamer wrapping up ~147 bp DNA. The amino acid residues on the N-terminal tails of the histone proteins can be covalently modified in a number of different ways, and different biochemical modification marks may have very different biological functions [1.7]. In recent years, genome-wide distributions of various histone modifications in several eukaryotic organisms have been mapped using chromatin immunoprecipitation followed by either microarrays (ChIP-chip) or DNA sequencing (ChIP-seq) [1.8-1.10]. It is thought that different combinations of histone modifications result in different functional specificities [1.11] and that several modifications form broad domains [1.12-1.14], which is helpful in stabilizing chromatin state and transmitting the states in cell division [1.15]. Integration of multiple histone modification marks has identified distinct epigenetic patterns, associated gene activities, and regulatory elements [1.16, 1.17] .

However, most of these earlier studies are done on a gene-by-gene basis and have ignored the spatial correlation of epigenetic patterns.

Because it is difficult to detect long-range chromosomal interactions through experimental methods, it is valuable to develop computational methods to do so, such as epigenetic domains based on histone modification data. This can be used to infer such interactions. Here we use the term "epigenetic domain" to refer to a large-scale region containing multiple genes with epigenetic patterns. For the rest of the chapter, we will use the terms "epigenetic domain" and "domain" interchangeably. Such epigenetic domain methods should provide mechanistic insights into coordinated gene regulation. Several approaches have already been developed in recent years [1.18-1.24]. However, two of the studies [1.18, 1.20] partition the genome without any constraint, making the results hard to interpret. Other studies consider only a single histone modification mark, which can only give partial epigenetic information [1.19, 1.20]. Or the researchers only examined a certain pattern of two modifications [1.21, 1.22], simply small regions [1.23], or were not looking at neighboring regions, only clusters as a whole [1.24]. Therefore, further improvement is still needed.

Here, we developed a novel method that uses genome-wide histone modification data to identify epigenetically consistent, multi-gene domains. We applied our method to analyze a recently published ChIP-seq dataset for mouse embryonic stem (ES) cells [1.10, 1.25] and found that we were able to identify a number of domains that are significantly large (i.e. not just due to chance). We validated our predictions by integrating a number of data sources. We also explored these histone modifications in the neural progenitor (NP) cell line to determine what, if any, changes in domain size,

structure, and/or function occur during early cell differentiation. Our method provides a useful tool to investigate the role of epigenetic domains in development.

1.2 Results

1.2.1 Data type, preliminary manipulation and clustering

Genome-wide location data for five different histone modification marks, H3K4me2, H3K4me3, H3K27me3, H3K9me3, and H3K36me3, was taken from two ChIP-seq datasets [1.10, 1.25]. Since our goal was to identify multi-gene regions with consistent histone modification patterns, we treated each gene as a unit and summarized the local distribution of each histone modification by a single score. We arrived at a five-dimensional summary score for each gene, corresponding to the average of the sequence tag counts over the promoter (for H3K4me2, H3K4me3, H3K9me3, and H3K27me3) or coding region (for H3K36me3) (Figure 1.1).

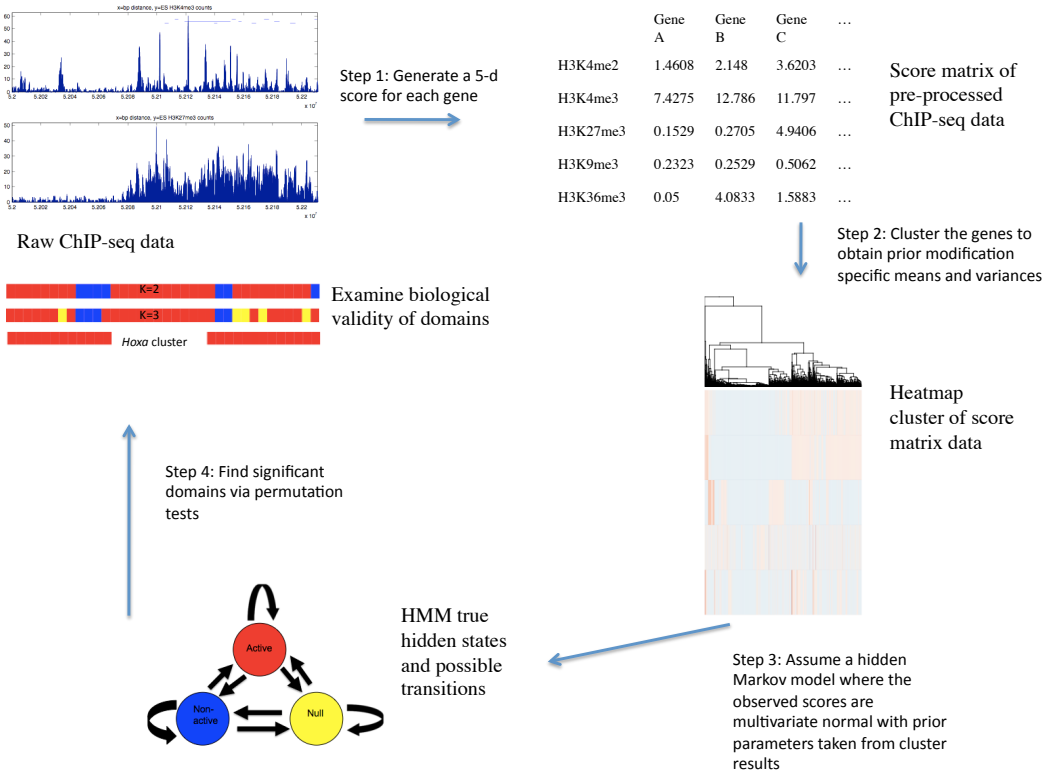

Figure 1.1: Outline of methods
Brief four-step depiction of methods used in this study to find chromatin domains.

We clustered the genes based on their histone modification patterns and determined the number of clusters via calculations of a "gap" statistic [1.26], which compares the observed within-cluster dispersion for one run with that expected by chance (i.e. averaging over 1000 permutations). The maximum gap value is achieved at K = 3, where K is the number of clusters, and the two-cluster partition also corresponds to a relatively high gap statistic (Figure 1.2). In the two-cluster scenario, there is a cluster of genes associated with high H3K4me2 and H3K4me3 levels, while the other is associated with moderate H3K27me3 level. In the three-cluster scenario, there is an additional cluster characterized by moderate level of the other modifications.

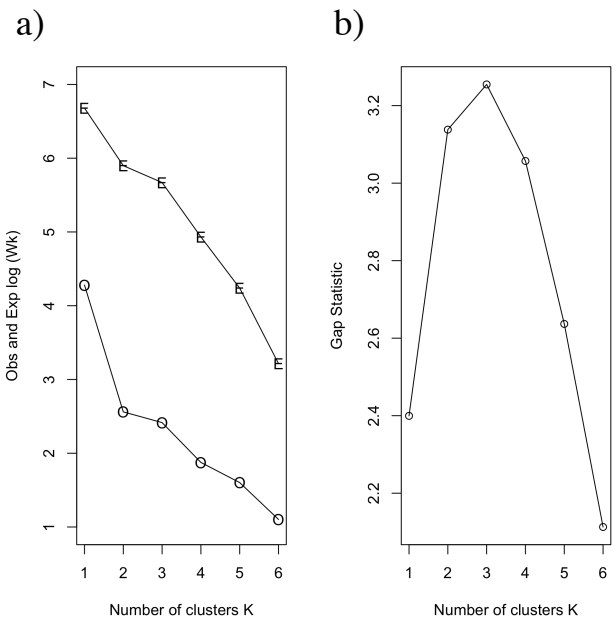

Figure 1.2: Results from gap statistics analysis of a random sample of the data
The expected and observed *log* (W_K) values are shown in (a) for various levels of K (the number of clusters), where W_K is the pooled within cluster sum of squares around the cluster means. The number of clusters versus *Gap(K)*, the difference between the observed and expected values (mean value for 1000 random bootstrap permutations), is shown in (b). According to these results, three is the optional number of clusters for our data.

1.2.2 Prediction of epigenetic domains

To constrain on spatial correlation of histone modification patterns, we applied a hidden Markov model (HMM) to partition the genome into epigenetic domains based upon the five-dimensional summary histone modification scores. In accordance with previous methods [1.18], we assumed that multivariate Gaussian distribution could approximate the emission probability. This gives a better fit than a Poisson distribution (Supplementary Figure 1.1).

For hidden-state assignment, we compared the results from two commonly used methods: the Viterbi and Forward-Backward algorithms, on Chromosome 19 (Ch 19).

The results from the methods were 99.8% identical (Supplementary Figure 1.2). Since the Viterbi method is more computationally efficient and essentially as accurate, it was used to assign our hidden gene states.

We considered two model configurations, corresponding to two and three hidden states. For the rest of this paper, we will further explore these two choices and compare the corresponding results.

We used a likelihood ratio test to evaluate the statistical significance of a detected domain. With a false discovery rate (FDR) cutoff of 0.01, we found 401 and 258 significant domains based on their likelihood ratio test statistics for the two- and three-state HMMs respectively (Tables 1.1 and 1.2 and Supplementary Tables 1.1-1.5). The average size of each domain was 11.31 genes for the two-state model and 9.66 genes for the three-state model (Figure 1.3). Both models found domains over the length of 155 genes and had a minimum domain size of 2.

Table 1.1: Summary statistics for significant domains in the two- state HMM

State	Number of significant domains	Min domain size	Max domain size	Average domain size	Variance in domain size
Non-active	230	2	157	14.5	186.14
Active	171	2	29	7.02	16.45
Total	**401**	**2**	**157**	**11.31**	**127.28**

Table 1.2: Summary statistics for significant domains in the three- state HMM

State	Number of significant domains	Min domain size	Max domain size	Average domain size	Variance in domain size
Non-active	25	2	17	9.48	12.59
Active	80	2	15	7.03	11.41
Null	153	2	155	11.06	227.02
Total	**258**	**2**	**155**	**9.66**	**142.28**

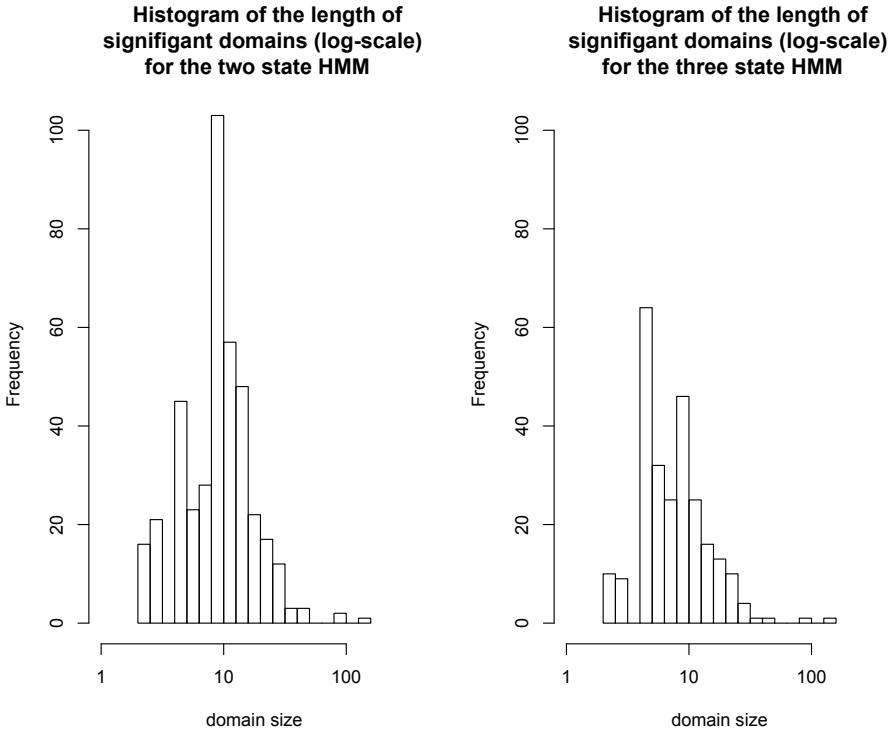

Figure 1.3: Histogram distribution of significant domains
Distribution of significant domains for the (a) two-state HMM and (b) three-state HMM. X-axis corresponds to the domain size; y-axis corresponds to frequency of observation.

As expected, the two-state HMM resulted in an active state (high H3K4me2/3 and low H3K27me3 activity) and a non-active state (low H3K4me2/3 and high H3K27me3

activity) (Table 1.3). For the three-state HMM, we also identified three distinct epigenetic patterns: an active cluster characterized by high H3K4me2/me3 and low H3K27me3 level, a non-active cluster characterized by high H3K27me3 and moderately low H3K4me2/me3 level, and a null cluster characterized by low level of all histone marks (Table 1.4). This null state cannot be captured by the two-state HMM.

Table 1.3: Average (variance) histone modification activity within a state for the two-state HMM

Modification	Non-active State	Active State
H3K4me2	1.453 (2.492)	4.653 (4.836)
H3K4me3	3.580 (18.538)	14.586 (60.923)
H3K27me3	2.089 (7.346)	0.740 (0.209)
H3K9me3	0.416 (0.112)	0.398 (0.0297)
H3K36me3	0.352 (0.176)	1.685 (5.266)

Table 1.4: Average (variance) histone modification activity within a state for the three-state HMM

Modification	Non-active State	Null State	Active State
H3K4me2	2.844 (2.106)	0.216 (0.083)	4.523 (4.934)
H3K4me3	7.020 (20.585)	0.499 (0.186)	14.205 (62.267)
H3K27me3	3.623 (10.534)	0.743 (0.412)	0.695 (0.130)
H3K9me3	0.458 (0.194)	0.383 (0.038)	0.394 (0.027)
H3K36me3	0.391(0.241)	0.367 (0.190)	1.615 (5.232)

1.2.3 Functional coherence of the predicted domains

It is intrinsically difficult to define a 'gold-standard' set of epigenetic domains. To test whether our predicted domains are biologically meaningful, we examined a number of properties that are associated with chromatin domains, as explained below.

First, we tested whether the histone modification patterns were indeed consistent within each predicted significant domain. To this end, we calculated the within-domain variance of the summary score for each modification in our significant domains, and tested whether this is significantly lower than expected by chance. Using permutation tests, we found that both the two- and three- state HMMs had significantly lower variances in the histone modifications considered in this study (p-values <0.05), with the exception of H3K9me3 (Figure 1.4). This suggests that H3K9me3 does not play a major role in determining our domain states.

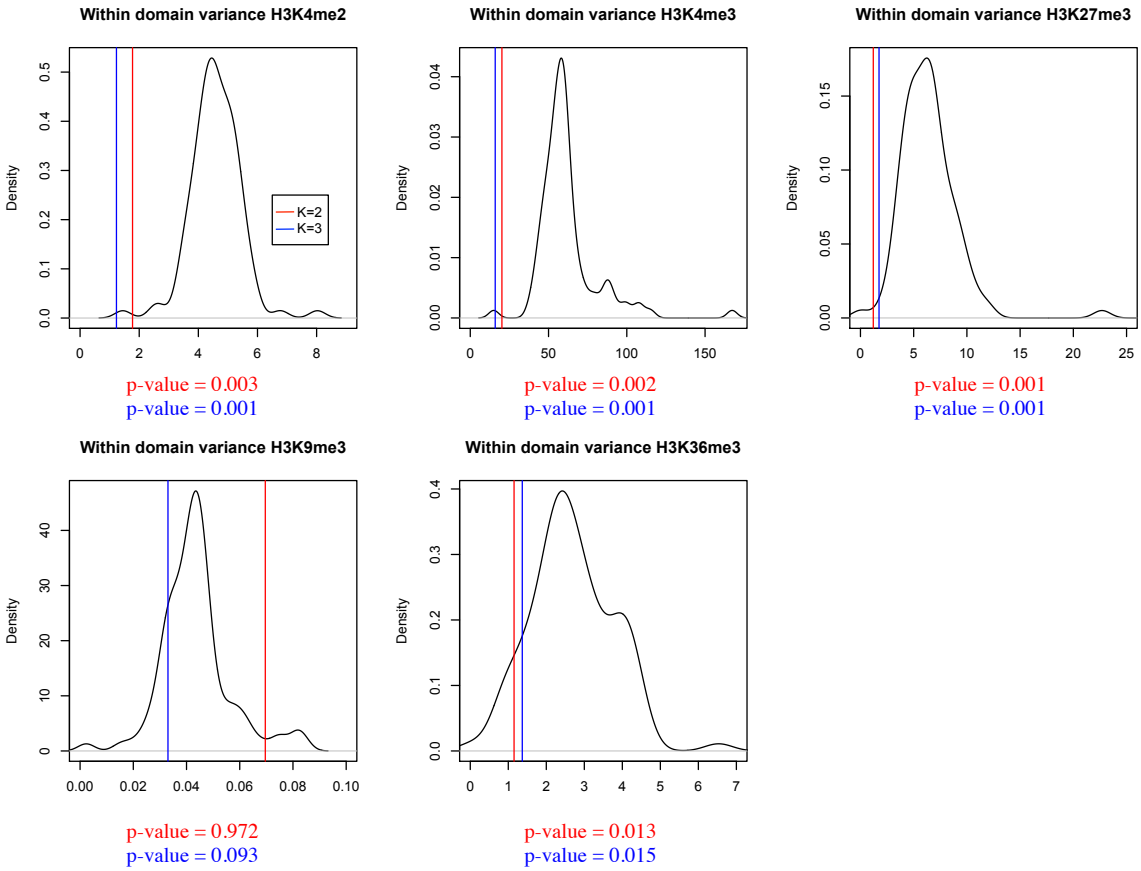

Figure 1.4: Average within-domain variance of a modification versus a random distribution
Two-state model (K=2) results (and corresponding permutation test p-values) are shown in red, three-state in blue (K=3), random distribution shown in black. For four of the modifications, the HMM domains have a significantly lower average variance (H3K4me2, H3K4me3 & H3K27me3: p-values<0.001), suggesting that the HMM has produced coherent domain bounds.

Second, we reasoned that genes that are embedded in the same epigenetic domain are likely to be activated or repressed together. Therefore, the gene expression levels should be more correlated within an epigenetic domain than what we would expect by chance. To test this property, we compared the within-domain variance of expression within a predicted domain with that for random neighboring gene sets of the same size (in terms of number of genes). For both models, the within-domain variance in gene

expression was lower than expected by chance (p-values <0.05) (Figure 1.5a). Also, the variance for the three-state HMM is lower than the two-state HMM, further supporting the notion that the three-state HMM is a more appropriate model.

Third, we asked whether the genes that are embedded in the same domain tend to have similar biological functions. To this end, we examined the Gene Ontology (GO) patterns within and between domains. We used Fleiss' Kappa [1.27], an accordance statistic, to measure the coherence of GO terms within our significant domains. Compared to a random selection of domain bounds, our predicted significant domains had a more concordant GO structure (p-values < 0.001) for the two- and three- state models (Figure 1.5b). Again, the three-state HMM had a higher level of accordance, suggesting that it is a better fit of the data.

Finally, we recognized that the histone modification patterns within a predicted domain are consistent, but substantial variance still remains. As a further validation, we asked whether our HMM was expected to provide reliable predictions under such circumstances. To this end, we designed numerical simulations mimicking the parameters for the real data (Methods). For the two-state HMM, we were 98% accurate in predicting the true states when using the observed variances, and 97% accurate when using twice the observed variances (Supplementary Figure 1.3a & Supplementary Figure 1.4a). For the three-state HMM, we were 99% accurate in our simulations (Supplementary Figure 1.3b & Supplementary Figure 1.4b). This suggests that if such domain patterns actually exist in the data, our model would be sufficiently able to detect them.

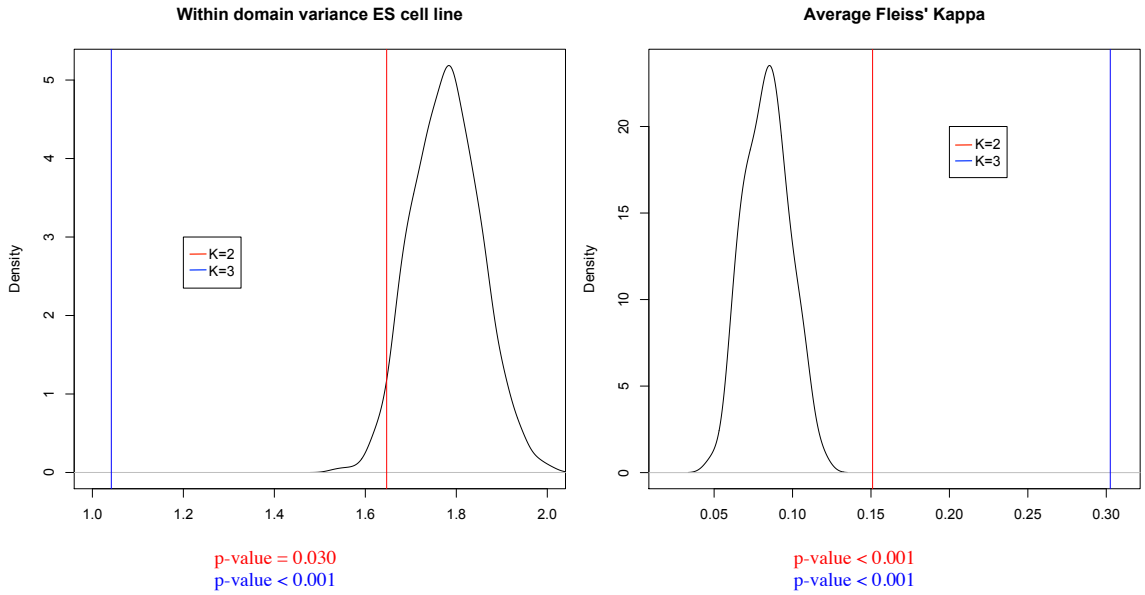

Figure 1.5: Coherence within predicted chromatin domains
The values (and corresponding permutation test p-values) for predicted chromatin domains (red for two state model, K=2, and blue for three, K=3) are compared with the distribution estimated from 1000 random permutations (black). Shown is the average within domain variance in gene expression (left) and the average level of Gene Ontology accordance as measured by Fleiss' Kappa (right) under the null hypothesis of incoherent domains. Note that our models result in significantly different values.

1.2.4 Validation against known domains

The *Hox* gene clusters are a well-described epigenetic domain family. These genes regulate the anterior-posterior axis of metazoan organisms and are expressed in a sequential order during cell differentiation. In ES cells, the *Hox* genes are targeted by the Polycomb group (PcG) proteins and associated with bivalent domains [1.12, 1.28]. Our method correctly detected each of the four *Hox* clusters to be in a non-active and significant domain. The results for the *Hoxa* cluster on Ch 6 are shown in Figure 1.6a. Simple K-means clustering failed to capture these state assignments.

Surprisingly, six of the ten largest predicted domains in each model are null domains enriched with olfactory receptor (OR) genes. For example, the largest OR gene cluster on Ch 7 (as described by [1.29]), also showed a clear distinction between HMM assignments and gene function (Figure 1.6b). That is, all the 208 OR genes in this 250 gene region are assigned to the same domain type, while their neighboring genes (and even non-OR genes within this cluster) are found to be in different domain types. The state assignments also correspond nicely with expression level changes (Figure 1.6b).

The OR genes are only expressed on in sensory neurons, and only a single gene (out of 1300) is activated in each cell [1.30]. We searched the literature for domain-level regulation of OR genes and noticed a recent paper showing that the selectivity of OR gene activation is established by the long-range chromosomal interaction between a single enhancer element and its target promoter [1.31]. Due to the lack of sequence specificity of such an interaction, it is reasonable to assume that the maintenance of an open chromatin environment over a large domain plays an important role in the regulation of OR genes. A comparison of the domain organization between sensory neurons and other cell types may provide further insights into the unique feature of OR gene regulation.

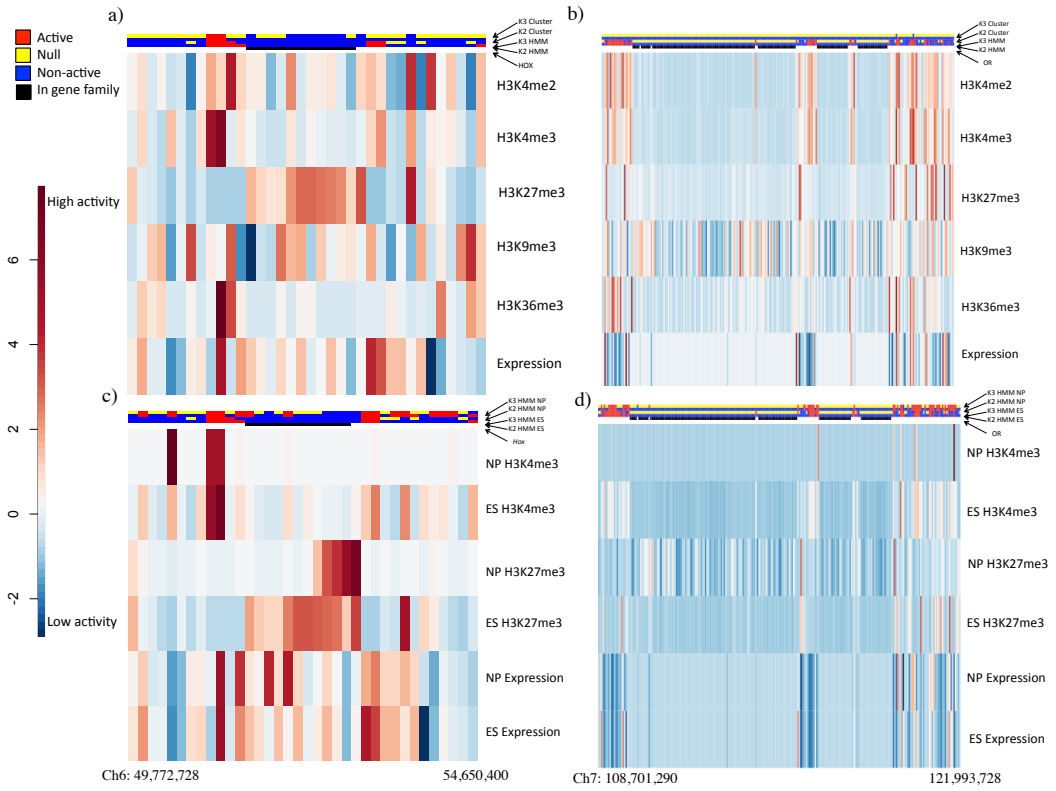

Figure 1.6: Heatmaps of known gene clusters
The 35 gene region on Ch6 from *Npy* to *2410066E13Rik* (49,772,728 to 54,650,400) as depictured as a heatmap of histone modification and gene expression for (a) the ES cell line and (c) ES and NP cell lines. The 250 gene region on Ch7 from *Art2a* to *Insc* (108,701,290-121,993,728) as depictured as a heatmap of histone modification and gene expression for (b) the ES cell line and (d) ES and NP cell lines. For all figures on the left (a and c), cluster state assignment is given in the first and second tracks and HMM state assignments are in the third and fourth tracks (red for active state, blue for non-active state, yellow for null state). For figures on the right (b and d), NP HMM state assignment is given in the first and second tracks and ES HMM state assignments are in the third and fourth tracks. Whether (black) or not (white) a gene is a respective gene cluster in shown in the bottom track in all figures.

We investigated the enriched biological functions associated with each domain type by exploring the top three significant DAVID clusters [1.32] (Figure 1.7) for genes in each domain. We found that the active domains are enriched with genes involved in key cellular processes, such as protein localization and transport. In contrast, the genes

embedded in null domains tend to be associated with functions of a terminally differentiated cell-type, e.g. keratin and olfactory receptors. Non-active genes tend to be involved with development, e.g. limb morphogenesis and homeobox. Thus, each domain type is characterized by a different function. Our analysis suggests that the epigenetic information provides useful insights into cell-type specific regulation.

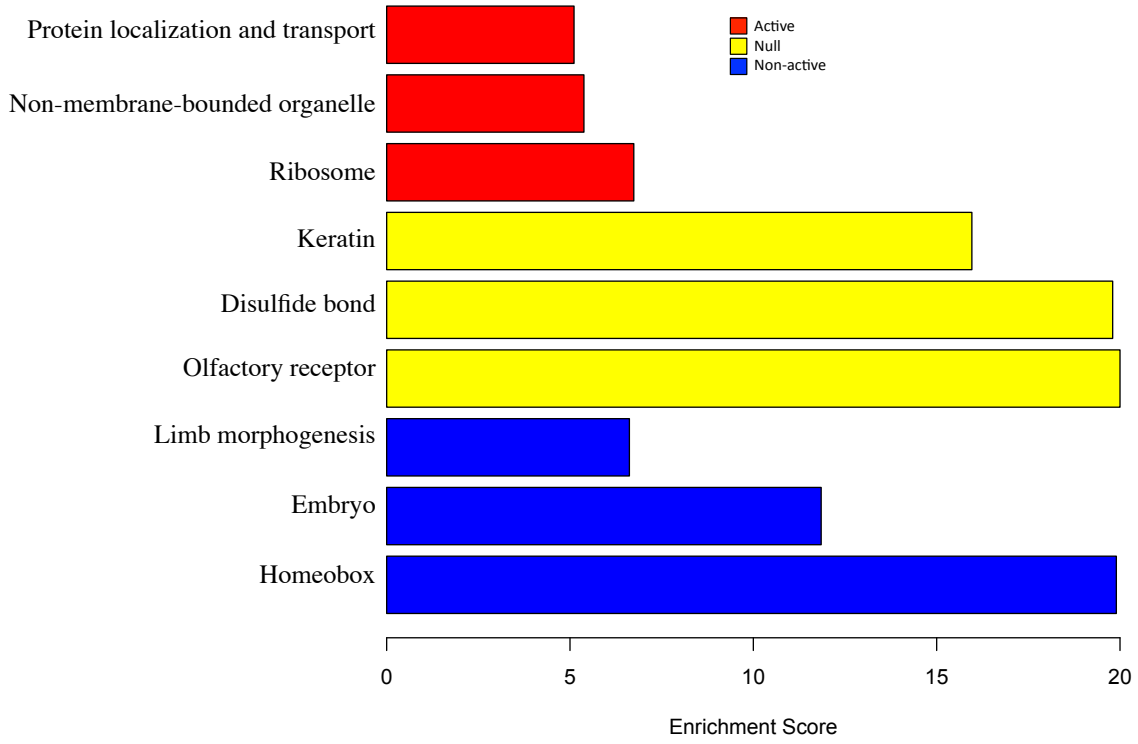

Figure 1.7: DAVID analysis of genes within each ES significant domain
Genes in each state are described by the top three significant DAVID clusters. Red corresponds to genes in the active state, blue for those in the non-active state and yellow for null state.

1.2.5 Domain changes in neural progenitor cells

Previous studies have identified dramatic epigenetic changes during cell differentiation [1.10, 1.14, 1.33]. To test whether the epigenetic changes also occur at the

domain level, we applied our three-state HMM to infer domain states in the NP cells and compared the results for ES cells (Supplementary Table 1.6). While the overall change is moderate, we noticed some important changes at specific loci. For example, we found that the non-active *Hoxa* domain shrank significantly in NP cells (Figure 1.6c), consistent with a previous time-course study [1.34]. Such change is accompanied by activation of certain *Hoxa* genes in NP cells (Supplementary Figure 1.5a). The loss of H3K27me3 is accompanied by a moderate increase of H3K4me3 and expression levels. Little changed for the OR genes between the ES and NP cell lines (Figure 1.6d and Supplementary Figure 1.5b).

In total, we found that 179 of the 258 domains in ES cells contain at least one gene that changes epigenetic state in the NP cell line. For each significant ES domain, if any genes within the domain have a new state in the NP cell line, then those genes would represent a domain change. Note that there may be smaller NP domains than their corresponding ES domains, and that one ES domain could be multiple NP domains. Thus our 258 significant ES domains were 450 NP domains (Supplementary Table 1.6). For these NP domains changes, we again used DAVID to analyze the functions of the genes that change state during this early stage of development by examining the top three significant DAVID clusters (Supplementary Figure 1.6) for each of our 6 types of change (and for genes that remained state unchanged). Eight domains are non-active in ES cells but become active in NP cells. These domains are enriched with developmental regulators. On the other hand, six domains, containing 108 genes, switch from null to active states during differentiation.

Some of the non-active domains in ES cells remain non-active in NP cells, and they may be important for further development. On the other hand, a number of domains switch from the active to the non-active state in NP cells, and the functions of these domains are typically related to early embryonic development. They are enriched with functions such as sex differentiation and apoptosis. Thus, the early developmental genes are epigenetically marked in ES cells rather than at a later developmental stage.

1.3 Discussion

As the nucleosome level chromatin states become increasingly well described [1.15], the next frontier becomes the characterization of higher-order chromatin structure. Numerous studies have suggested that epigenetic domains play important roles in gene regulation [1.5, 1.35], yet the detection of genome-wide long-range chromosomal correlations remains technically challenging [1.6]. On the other hand, genome-wide histone modification data provides important information about long-range gene regulation [1.12-1.14, 1.36]. Thus it is valuable to develop computational methods to detect large-scale domains based on histone modification data.

Here we developed an HMM-based method to predict epigenetic domains. A similar method has recently been used to characterize the epigenetic states associated with gene promoters [1.37]. However, we extend this approach to identify large-scale epigenetic patterns. Compared to previous domain detection methods [1.18-1.20], our model can easily accommodate additional histone modification marks and provide easily interpretable prediction outcomes.

Our model detects three distinct types of epigenetic domains, two of which are transcriptionally inactive, which we call non-active and null. These two domain types are also distinct functionally in terms of both activation potential and biological functions. For example, both the *Hox* and OR gene clusters form epigenetic domains that are transcriptionally inactive in ES cells. Yet the *Hox* genes are critical for the overall development and are found in non-active domains in the ES cell line, whereas OR genes are expressed only in specific adult cell types and are found in null domains in the ES and NP cell lines. Therefore the epigenetic patterns provide more regulatory information than can be appreciated by gene expression data alone, signifying the importance of characterizing domain types.

Recent studies have shown that spreading of histone modification marks is an important epigenetic signature of cell differentiation [1.13, 1.14]. Our work can be viewed as an extension in terms of considering the combinatorial patterns of multiple histone modifications instead of focusing on a single modification alone. Indeed, we found changes of epigenetic domains from ES to NP cells, which are accompanied by coordinated activation of neuron-specific genes. Our analysis suggests that epigenetic domain-level changes may play an important role in neuron differentiation and organismal development.

We recognize that our model still has a number of limitations. For example, the reduction of the spatial epigenetic patterns by gene-level summary scores precludes us from pinpointing the exact locations of domain boundaries. In addition, we have ignored the correlation between different histone modification marks, which may important if

data for a large number of marks is available. We plan to overcome these limitations in future studies.

1.4 Methods

1.4.1 Gene-level summary score

Gene annotation was based on Refseq; we obtained 17,772 genes in total. For four of the five modifications (H3K4me2, H3K4me3, H3K27me3 and H3K9me3), the tag counts peak near the TSS; therefore the average was taken over the regions from -2kb to +2kb with respect to the TSS. The tag counts for H3K36me3 are highest around the 3'-end of the coding region of a gene; thus, we took the average tag counts over this area as a gene's corresponding score. Genes with more than 50% repetitive sequences in either of these two regions were not used in further analysis. Our preliminary data manipulations led to the elimination of sites with poor ChIP-seq coverage, resulting 17,469 (98.3%) genes used in further analysis.

1.4.2 Clustering

With the gene-specific summary scores, each gene is associated with an m-dimensional vector, where m is the number of histone modification marks. As an initial guess for the number of domain types, we clustered the m-dimensional vectors using the k-means average agglomeration clustering method. The optimal cluster number k was selected using the gap statistic [1.26], defined as

where W_k^* is the observed within-cluster sum of squares around the clusters means for one run, and $E(\cdot)$ represents the mean value for 1000 random bootstrap permutations.

1.4.3 Hidden Markov model

We chose HMM to infer domain locations, where the hidden state at a given gene represented the associated domain type, and the emission variables are the *m*-dimensional vector summarizing the local histone modification pattern. We assumed that our emissions followed a multivariate Gaussian distribution.

When dealing with ChIP-seq data, researchers often like to assume a Poisson distribution for the counts mapped to each bin. This was not appropriate for our analysis for two main reasons: (1) we fit our model on non-integer summary scores to examine domain structure at the gene level and (2) the multi-dimensionality of our study. By the central limit theory, even if our raw counts followed a Poisson distribution, an average score of these counts (say over a promoter region) would follow a Gaussian distribution. To evaluate all five modifications simultaneously, we assumed that together they were from a multivariate Gaussian emission distribution with no covariance structure. We also checked the validity of our assumption by comparing distribution of the score data to its corresponding Poisson and Gaussian distributions. For these reasons, we assumed that our score data followed a multivariate Gaussian distribution.

We used the expectation-maximization (EM) algorithm to estimate the model parameters, and then used the Viterbi algorithm to infer the maximum likely state configuration [1.38]. One technical problem is that the EM algorithm can only achieve

local optimization, and the results are dependent on the initial condition. One possible approach to overcome this problem is to repeat the procedure many times, each with a randomly selected initial guess. However, we found that it is more efficient to choose a particular initial guess based on the clustering results; that is, using the cluster means and variances as initial guesses for the model parameters associated with the hidden states.

To test whether our clustering method-based prior led to the optimal model (in addition to being more efficient), we compared its resulting log-likelihood to that of 100 models where we randomly selected our prior parameters from the semi-conjugate hierarchical model:

$$\mu_{ks} | \sigma_{ks}^2 \sim N(\theta_k, s^2_k)$$

$$\sigma_{ks}^2 \sim Inv\text{-}\chi^2 (\nu_k),$$

where μ_{ks} is the prior mean for modification k ($k=1, 2, \ldots, 5$) and state s ($s=1, 2, 3$), σ_{ks}^2 is the initial variance for modification k and state s, θ_k is the sample mean for modification k, s^2_k is the sample variance for modification k, and ν_k is the degrees of freedom for modification k such that $E(\sigma_{ks}^2)=(\nu_k-2)^{-1} = s^2_k$. We found that using the results from k-means clustering to pick our prior results in a much higher final log-likelihood than this hierarchal prior (Supplementary Figure 1.7).

We also used the Viterbi algorithm, an approximation of the Forward-Backward algorithm, to assign a state to each gene. To determine the accuracy of the Viterbi method, we compared its corresponding hidden maximum likelihood estimate (MLE) state assignment to that obtained by the forward-backward algorithm, which assigns states based on posterior probabilities, also known as a maximum a posteriori (MAP) estimate. These two algorithms were compared on the 615 gene chromosome 19 and

produced similar hidden states. The Viterbi algorithm was used to assign gene state as it is more computationally efficient.

The number of hidden states was equal to the optimal number of clusters from our gap statistic results. For the mouse data considered in this paper, our mouse data, the optimal number was $K = 3$, where K is the number of clusters. However, the gap statistic for $K = 2$ was similar, so we compared both setups in our analysis.

1.4.4 Significance of detected domains

To determine statistical significance of a domain, we first calculated a likelihood ratio test statistic for each domain j ($j=1,2, \ldots, n$):

$$\lambda_j(\mathbf{X}) = -2\ln\frac{\sup_{\Theta_0} L(\theta|\mathbf{x_j})}{\sup_{\Theta} L(\theta|\mathbf{x_j})} = -2\ln\frac{P(\mathbf{x_j}|H_0)}{P(\mathbf{x_j}|HMM)} = -2\ln\frac{P(\mathbf{x_i}|H_0)*P(\mathbf{x_{i+1}}|H_0)\ldots*P(\mathbf{x_{n_j}}|H_0)}{P(\mathbf{x_i}|HMM)*P(\mathbf{x_{i+1}}|HMM)\ldots*P(\mathbf{x_{n_j}}|HMM)},$$

where $\boldsymbol{x}_i$ is the observed m-dimensional vector of histone modifications for gene i ($i=1, 2, n_j$), n_j is the number of genes in domain j, $\boldsymbol{X}_i \mid H_0 \sim N(\mu_0, \Sigma_0)$ and $\boldsymbol{X}_i \mid HMM \sim N(\mu_s, \Sigma_s)$. Note that μ_0 and Σ_0 are the m-dimensional mean vector and diagonal variance matrix of the entire dataset whereas μ_s and Σ_s are their individual state-based counterparts. To calculate $\lambda_j(\boldsymbol{X})$ for each domain, the μ_s and Σ_s of the corresponding maximum likelihood estimate state was used. Based on likelihood theory, $\lambda_j(\mathbf{X}) \sim \chi^2_{df}$ where the degrees of freedom (df) is the difference in parameters between models. Thus, $df = 2*S*m-2*m$, where S is the number of states in the model. To correct for multiple hypothesis testing, the significant domains were selected such that the FDR = 0.01.

1.4.5 Numerical simulations

We simulated histone modification data corresponding to prescribed domain configurations and assessed the accuracy of our model. We assumed that the histone modification data for each gene was normally distributed with the mean and variances estimated from the real data (Tables 1.3 & 1.4). To further explore the model robustness against data noise, we also repeated the above simulations with variances equal to twice the observed variances. We applied our HMM inference procedure to the simulated data. The accuracy of our model was quantified as the percentage of correctly assigned states.

1.4.6 Gene Ontology analysis

To test for functional coherence, we examined the accordance of GO memberships within a domain via a new statistic for each significant domain. For domain j, we calculated the level of GO accordance, Y_j, with Fleiss' Kappa [1.27], which is a generalization of the standard kappa for more than two raters (or in our case, ontologies). We then calculated the average accordance for these domains, and compared it to a distribution made under the null hypothesis (i.e., the hypothesis that domains are independent of GO memberships).

For the null distribution, we took into account the fact that neighboring genes often share GO annotations. To this end, we selected $1000*n$ random sections of the genome (of corresponding equal length in terms of number of genes to our significant domains), calculated their Y_j's and averaged them to get a null distribution for our average accordance. The p-values were evaluated as the proportion of permuted means

that are larger than the observed mean accordance. Thus the minimum possible p-value was 0.001.

1.4.7 Gene expression data analysis

To account for variability in gene activity, expression data in the ES cell line was normalized across data from 23 cell lines [1.10, 1.25] to get a Z-statistic for each gene. We then calculated the within-domain variance for each significant domain and compared the average of these values to that of a random permutation.

To determine the significance of our within domain variance, we randomly selected 1000*n sections of the genome (each with an equal number of genes as its corresponding significant domains) and calculated the average within domain variances.

1.4.8 Epigenetic domains in NP cells

The histone modification data in NP cells were obtained from a published dataset [1.10, 1.25]. The ChIP-seq data in NP cells were normalized against those in ES cells by a negative-binomial regression as recommended [1.39]. We assumed that the model parameters in the NP cells are identical to those in the ES cells, and inferred the hidden states in NP cells by using the Viterbi algorithm again. We determined domain changes by comparing the results in NP and ES cells. A domain is called changed if at least one gene changed state between the two cell lines.

1.5 References

1.1 Cooper KW (1959) Cytogenetic analysis of major heterochromatic elements (especially Xh and Y) in Drosophila melanogaster, and the theory of "heterochromatin". Chromosoma 10:535-588.

1.2 Elgin SC (1996) Heterochromatin and gene regulation in Drosophila. Current opinion in genetics & development 6(2):193-202.

1.3 de Laat W, Grosveld F (2003) Spatial organization of gene expression: the active chromatin hub. Chromosome Res 11(5):447-459.

1.4 Duboule D (2007) The rise and fall of Hox gene clusters. Development (Cambridge, England) 134(14):2549-2560.

1.5 Kosak ST, Groudine M (2004) Gene order and dynamic domains. Science 306(5696):644-647.

1.6 Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326(5950):289-293.

1.7 Kouzarides T (2007) Chromatin modifications and their function. Cell 128(4):693-705.

1.8 Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Rolfe PA, Herbolsheimer E et al. (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. Cell 122(4):517-527.

1.9 Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. Cell 129(4):823-837.

1.10 Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448(7153):553-560.

1.11 Jenuwein T, Allis CD (2001) Translating the histone code. Science 293(5532):1074-1080.

1.12 Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125(2):315-326.

1.13 Wen B, Wu H, Shinkai Y, Irizarry RA, Feinberg AP (2009) Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. Nat Genet 41(2):246-250.

1.14 Hawkins RD, Hon GC, Lee LK, Ngo Q, Lister R, Pelizzola M, Edsall LE, Kuan S, Luu Y, Klugman S et al. (2010) Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell 6(5):479-491.

1.15 Bernstein BE, Meissner A, Lander ES (2007) The mammalian epigenome. Cell 128(4):669-681.

1.16 Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature genetics 39(3):311-318.

1.17 Ernst J, Kellis M (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature biotechnology 28(8):817-825.

1.18 Thurman RE, Day N, Noble WS, Stamatoyannopoulos JA (2007) Identification of higher-order functional domains in the human ENCODE regions. Genome Res 17(6):917-927.

1.19 de Wit E, Braunschweig U, Greil F, Bussemaker HJ, van Steensel B (2008) Global chromatin domain organization of the Drosophila genome. PLoS Genet 4(3):e1000045.

1.20 Lian H, Thompson WA, Thurman R, Stamatoyannopoulos JA, Noble WS, Lawrence CE (2008) Automated mapping of large-scale chromatin structure in ENCODE. Bioinformatics (Oxford, England) 24(17):1911-1916.

1.21 Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458(7235):223-227.

1.22 Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics (Oxford, England) 25(15):1952-1958.

1.23 Regha K, Sloane MA, Huang R, Pauler FM, Warczok KE, Melikant B, Radolf M, Martens JH, Schotta G, Jenuwein T et al. (2007) Active and repressive chromatin are interspersed without spreading in an imprinted gene cluster in the mammalian genome. Molecular cell 27(3):353-366.

1.24 Hon G, Ren B, Wang W (2008) ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. PLoS computational biology 4(10):e1000201.

1.25 Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB et al. (2008) Genome-scale DNA

methylation maps of pluripotent and differentiated cells. Nature 454(7205):766-770.

1.26 Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a dataset via the gap statistic J Royal Statistic Soc B 32(2):411-423.

1.27 Fleiss J (1971) Measuring nominal scale agreement among many raters. Psychological Bulletin 76(5):378-382.

1.28 Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, Lee TI, Levine SS, Wernig M, Tajonar A, Ray MK et al. (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature 441(7091):349-353.

1.29 Zhang X, Firestein S (2002) The olfactory receptor gene superfamily of the mouse. Nature neuroscience 5(2):124-133.

1.30 Chess A, Simon I, Cedar H, Axel R (1994) Allelic inactivation regulates olfactory receptor gene expression. Cell 78(5):823-834.

1.31 Lomvardas S, Barnea G, Pisapia DJ, Mendelsohn M, Kirkland J, Axel R (2006) Interchromosomal interactions and olfactory receptor choice. Cell 126(2):403-413.

1.32 Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome biology 4(5):P3.

1.33 Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, Stadler MB, Bibel M, Schubeler D (2008) Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Molecular cell 30(6):755-766.

1.34 Soshnikova N, Duboule D (2009) Epigenetic temporal control of mouse Hox genes in vivo. Science 324(5932):1320-1323.

1.35 Sproul D, Gilbert N, Bickmore WA (2005) The role of chromatin structure in regulating the expression of clustered genes. Nat Rev Genet 6(10):775-781.

1.36 Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E et al (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129(7):1311-1323.

1.37 Schuettengruber B, Ganapathi M, Leblanc B, Portoso M, Jaschek R, Tolhuis B, van Lohuizen M, Tanay A, Cavalli G (2009) Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. PLoS Biol 7(1):e13.

1.38 Rabiner L (1995) A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of IEEE 77:257-286.

1.39 Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol 26(11):1293-1300.

Chapter 2

Chromatin states accurately classify cell differentiation stages

Jessica L. Larson[1] and Guo-Cheng Yuan[1,2]

[1]Department of Biostatistics, Harvard School of Public Health,
Boston, Massachusetts, USA
[2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute,
Boston, Massachusetts, USA

2.1 Introduction

Multi-cellular organisms are composed of diverse cell types that, despite sharing the same genome, are programmed with distinct gene expression patterns. How such diversity is regulated mechanistically is a fundamental biological question. Eukaryotic DNA is packaged in chromatin. The fundamental unit of which is the nucleosome, where 147bp DNA is wrapped around a histone octamer protein complex. The N-terminal tails of histone proteins are decorated by different marks resulting from covalent modifications. The combinatorial patterns of these marks, which we refer to as the chromatin states, may recruit specific regulatory proteins, which in turn control transcription [2.1, 2.2].

Recent genome-wide location studies have identified distinct chromatin states that demarcate regulatory elements [2.3-2.7]. Furthermore, the chromatin states changes significantly between different cell types, in accordance with gene expression level changes [2.3, 2.8-2.15], providing strong evidence that the chromatin states play an important role in development. On the other hand, these studies have been limited to comparing a small number of cell types. As a result, it is difficult to evaluate to what extent cell lineages are associated with chromatin states. Characterization of the molecular signatures associated with cell lineages will not only provide insights into the transcription control, but will also help identifying the cell-of-origin, which is an important task for many diseases. For example, an intensively investigated area of cancer research is whether a tumor is originated from cancer stem cells or normal differentiated cells. Understanding the origin of cancer cells has important implications in developing therapeutic methods.

The idea of using genomic data to classify cell lineages is not new. There have been extensive studies based on gene expression data (reviewed by [2.16]). However, one major limitation is that gene expression levels do not inform us of the underlying controlling mechanism, which is fundamental for understanding developmental processes and diseases. For example, gene expression analyses have discovered the intriguing phenomenon that tumors with poor clinical outcome often display a signature that is similar to stem cells [2.17]. However, the underlying mechanism remains incompletely understood. Recently, it has been shown that the similarity is mainly due to the activity of the MYC regulatory module rather than the core module targeted by pluripotent factors [2.18].

A large amount of genome-wide histone modification data have been generated and made publicly available, in part due to the effort of ENCODE and Roadmap Epigenomics consortia [2.15, 2.19]. These data have provided a great opportunity to identify general principles of chromatin regulation. In this chapter, we will focus on evaluating the association between chromatin states and cell differentiation stages. To this end, we assembled a large dataset from the public domain of genome-wide locations of 5 histone modifications in 27 human cell lines and analyzed the data independently using four different spatial representations (see Figure 2.1 for a schematic overview). We found that cell differentiation status can be classified with nearly 100% accuracy from chromatin states alone, that chromatin state switches are frequently associated with multi-gene domains, and that the cancer cell lines have similar chromatin states as differentiated cells.

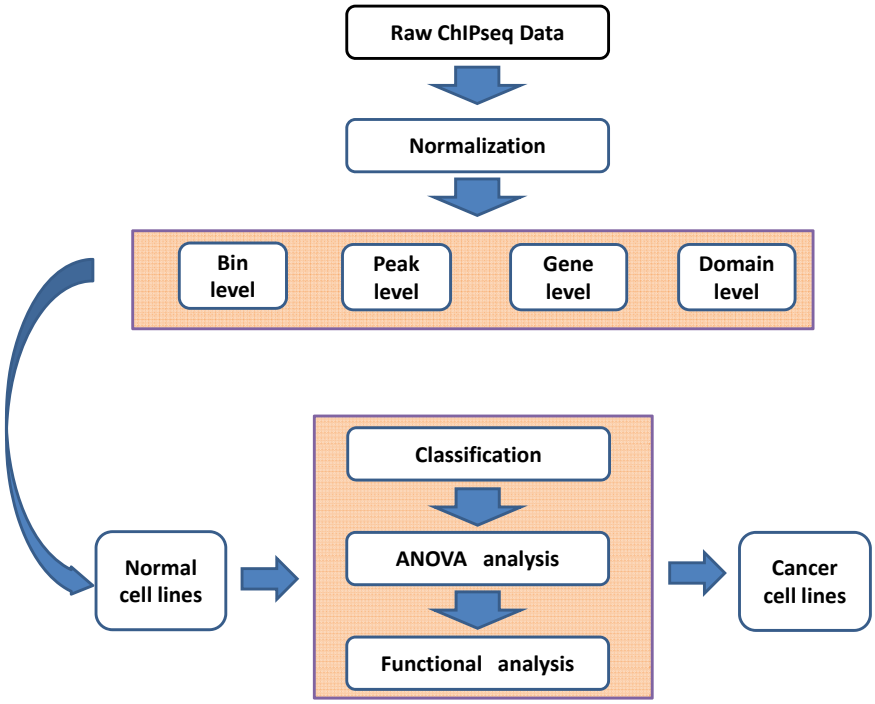

Figure 2.1: Overview of the data analysis strategy. ChIP-seq data of 5 histone modifications in 27 human cell lines were obtained from the public domains and analyzed independently using four different representations (bin, gene score, and chromatin state level). For each representation, a support vector machine (SVM) model was used to classify cell differentiation status from histone modification data. "Hotspot" genes or bins were detected by ANOVA and further investigated by functional genomic tools. The SVM model obtained from normal cell lines was used to classify the differentiation status of cancer cell lines.

2.2 Results

2.2.1 An assembly of genome-wide data for 5 histone modifications in 27 human cell lines

We collected genome-wide histone modification data from NIH Epigenome Roadmap [2.20] and other public domains [2.4, 2.11, 2.13, 2.21]. We focused on five

histone modifications that have been profiled extensively, including four associated with active genes (H3K4me1, H3K4me3, H3K36me3, H3K9ac) and one associated with gene silencing (H3K27me3). We examined 27 human cell lines for which data for all five modifications are available, including 24 normal and 3 cancer cell lines (Supplemental Table 2.1). Of the 24 normal cell lines, five are pluripotent cells (P), four are multipotent (M), which may further differentiate into multiple cell-types; and the others are either unipotent progenitors or terminally differentiated cells, which were grouped together (U/D). The three cancer cell lines are K562 (chronic myelogenous leukemia), HeLa (cervical cancer), and VCaP (prostate cancer).

To systematically compare different length scales, we analyzed the data independently based on three different representations, corresponding to increasing complex signatures: (1) the bin level sequence reads; (2) the gene-level summary scores associated with each histone mark; and (3) the combinatorial patterns of multiple histone marks referred to as the chromatin states (see Methods for more details).

In previous work [2.22], we developed a hidden Markov model (HMM) to identify chromatin states, treating each gene as a unit. Here we applied this approach to analyze the ChIP-seq dataset for the 27 human cell lines. As before, we determined the number of chromatin states based on the gap statistic [2.23] (see Methods for details), and found that the optimal number of clusters is 3 (Supplemental Figure 2.1). This is the same number of chromatin states we identified previously for mouse data [2.22]. We found that three HMM states described the genome-wide pattern: active (associated with active marks), non-active (associated with H3K27me3), and null (lack of both active and repressive marks) (Supplemental Table 2.2). We applied a common model to infer

genome-wide chromatin states in all cell lines (Supplemental Table 2.3). As the non-active state is associated with relative high density of both H3K27me3 and H3K4me3, we were also interested to test if there was a significant overlap with the bivalent domains. Indeed, we found that 25.8% of genes containing bivalent domains correspond to the non-active state (Supplemental Figure 2.2). On the other hand, we also found that 67.0% of bivalent domains correspond to the active state. These genes are typically associated with higher density of additional active marks such as H3K4me1 and H3K36me3. This observation is consistent with a recent study showing that a subset of genes marked by bivalent domains is actively transcribed [2.24], but it also suggests that the chromatin states may be further refined.

2.2.2 Histone modification patterns accurately classify cell differentiation status

We wanted to compare our three methods of analyzing ChIP-seq data (i.e., the bin, gene, and chromatin state levels) to determine which one is the 'best' at classification of cell lineages, which were grouped by the differentiation status: P, M, or U/D. For each representation, we built a support vector machine (SVM) model to classify the membership of a cell line based on the histone modification data (Methods). In order to avoid overfitting, we evaluated the classification accuracy using leave-one-out cross-validation. The classification accuracy was quantified by the percentage of agreement between the model predicted and true differentiation status. Surprisingly, we found that all representations led to 100% accuracy (compared to 62% obtained by using the null model, which classifies every cell line as the largest group, i.e., U/D). To see if that result is robust or depends on a particular approach, we repeated the analysis by using the radial

kernel function instead and again achieved 100% classification accuracy. These results strongly suggest that distinct chromatin states are associated with different differentiation statuses. This striking difference prompted us to further dissect biological features of these distinctive chromatin states and to pursue the utility of chromatin states for the classification of poorly characterized cell-types.

2.2.3 Numerous locations of epigenetic difference found between cell lines

To gain functional insights, we searched for regions that are most discriminative across different cell lineages: P, M, or U/D. To this end, we applied a permutation ANOVA F- test and selected those regions (bins or genes) that are differentially modified (FDR < 0.05) (Methods). Indeed we found extensive differences at each level.

Bin level. Our ANOVA analysis identified 249,705 differential bins for the H3K4me1 analysis, 21,224 bins for H3K4me3, 5,354 bins for H3K9ac, 25,385 bins for H3K27me3, and 69,373 bins for H3K36me3 (Supplemental Table 2.4). On the other hand, only 7 differential bins were common to all five modifications (Supplemental Figure 2.3a), consistent with the previous results that they each demarcate different regions [2.4, 2.9]. While the H3K27me3 and H3K4me1 bins are distributed quite uniformly across the genome, the other three modifications showed a strong bias toward coding regions, promoters, CpG islands and shores (Supplemental Figure 2.3b). While the mean occupancy levels of H3K4me3 and H3K9ac are highest in promoter regions, the variance can be high in coding regions as well. Conversely, while H3K36me3 mainly targeted toward coding regions, the variance can also be high in promoter regions.

Gene level. We found 2,501 genes that are differential based on their H3K4me1 gene-level score, 2,119 genes for H3K4me3, 368 genes for H3K9ac, 569 genes for H3K27me3, and 4,731 genes for H3K36me3 (Supplemental Table 2.5). For most of these genes, the gene-level scores are higher in pluripotent cells than multipotent and differentiated cells (Figure 2.2a**,** and Supplemental Figures 2.4b-e). Again, the overlap between different modifications is low: only 4 genes were common to all modifications (Supplemental Figure 2.4a).

Chromatin state level. By applying our ANOVA analysis to the chromatin state information obtained by our hidden Markov model, we found 722 differential genes, which were analyzed further. We call these genes the "hotspot" genes to highlight their role in chromatin state remodeling (Supplemental Table 2.6). We observed two main patterns of chromatin state switch during cell differentiation (Figure 2.2b): 1) most "hotspot" genes are in the non-active state in pluripotent cells and switch to the null state in differentiated cells; 2) a smaller subset of genes are in the active state in pluripotent cells and switch to another state during differentiation. Interestingly, most "hotspot" genes are in the null state in U/D cells. A closer examination suggested that a number of additional genes were also marked by H3K4me3 and H3K9ac, although the gene-level scores tended to be lower compared to the active state. 244 of these hotspot genes display distinct chromatin state pattern in ES cells compared with the other cell lines (Supplemental Table 2.7). Among these 244 genes, 209 are in the non-active state in ES cells.

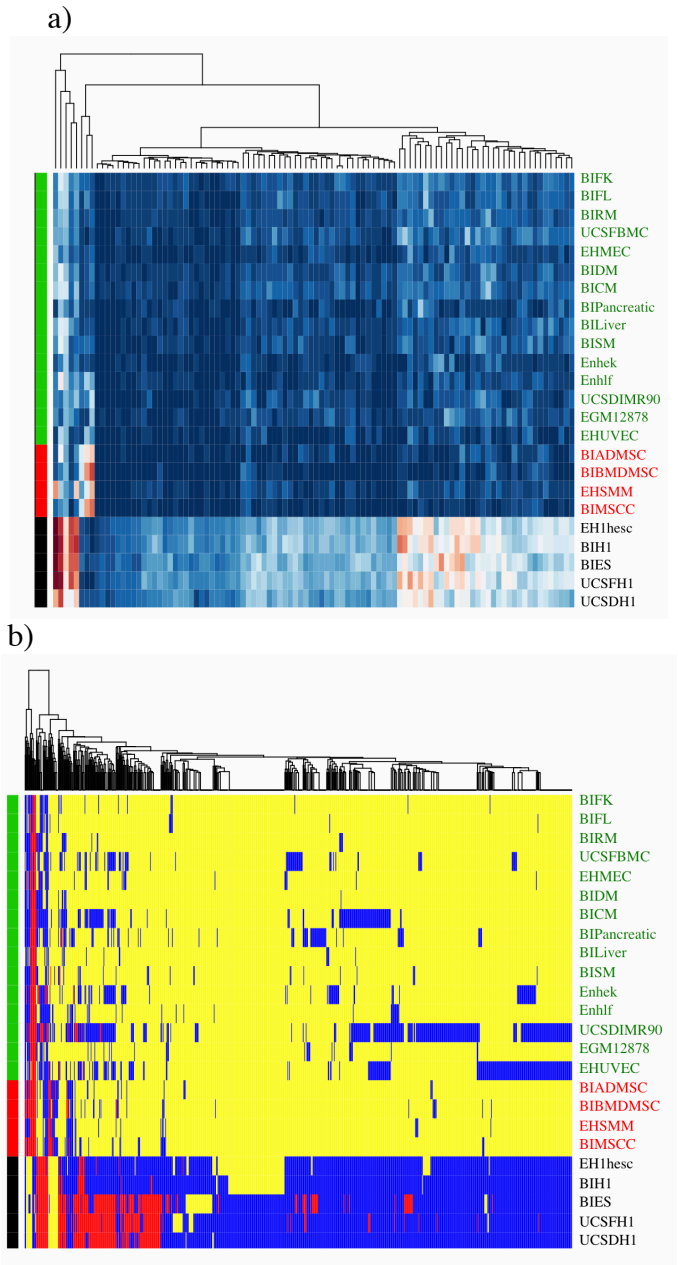

Figure 2.2. Differentiation related variation of histone modification patterns. (a) Heatmap showing the gene-level H3K4me1 scores for the 100 most significantly different genes. (b) Heatmap of the chromatin states for the 722 "hotspot" genes, whose chromatin states are significantly different across differentiation statuses. Red – active state; yellow – null state; blue – non-active state. The cell line information is shown at both sides of the heatmap and color-coded by the differentiation status (black – pluripotent cells (P); red – multipotent (M); green – unipotent/differentiated (U/D)).

To gain functional insights, we applied the Database for Annotation, Visualization and Integrated Discovery (DAVID) [2.25] to identify enriched functional categories that

are associated with the "hotspot" genes. We found significant enrichment of genes associated with the Homeobox, cell-cell signaling, or neuron development (Figure 2.3a), consistent with an important role of chromatin state remodeling during development.

In addition, those "hotspot" genes that undergo different remodeling paths during cell differentiation tend to be associated with different biological functions. Specifically, the genes that are active in pluripotent cells (such as HIST1H4F) tend to be associated with chromatin organization and methylation, the genes that are non-active in pluripotent cells (such as the Hox genes) are usually involved in organism development, while the genes that are in the null state in pluripotent cells (such as APOL6) seem to participate in diverse biological processes (Supplemental Figure 2.5).

Our functional analysis also identified four signaling pathways enriched in the "hotspot" genes: neuroactive ligand-receptor interaction (p-value= 0.0017), calcium-signaling (p-value = 0.006), hedgehog-signaling (p-value= 0.008), and TGF-Beta signaling pathway (p-value = 0.013). The hedgehog-signaling pathway is particularly interesting since it plays an essential role in embryo segmentation and is conserved from flies to humans [2.26]. There are four "hotspot" genes (adjusted p-value < 0.05) in this pathway, including SHH, SMO, WNT, and ZIC2 (Figure 2.3b), suggesting that chromatin state remodeling may play an important role in regulating the cell-type specific activity of this important pathway. Likewise the TGF-Beta signaling pathway is involved in embryo differentiation, left-right axis determination, apoptosis and mesoderm/endoderm development [2.27].

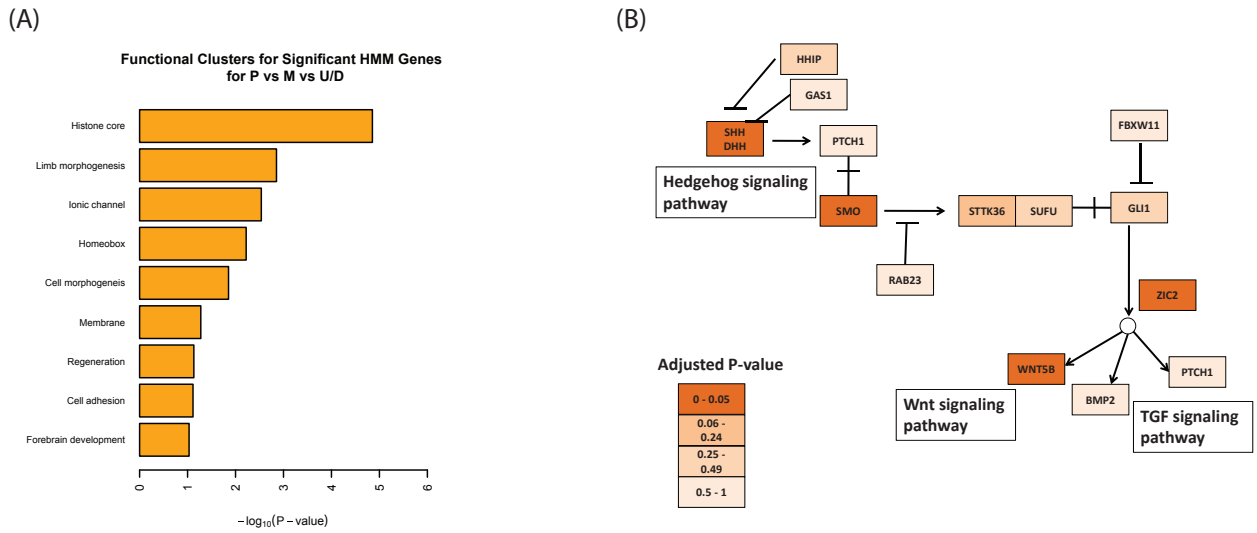

Figure 2.3: Functional enrichment analysis of the "hotspot" genes. A total of 722 "hotspot" genes were identified by applying ANOVA analysis to the chromatin states. Enriched functions and pathways were identified by using DAVID. (a) Representative enriched functional categories. (b) The Hedgehog signaling pathway is significantly enriched (p-value = 0.008). The genes were color-coded based their corresponding adjusted p-values obtained from ANOVA analysis.

2.2.4 Coordinated switches in chromatin domains

We were interested in finding the extent to which developmentally related chromatin state remodeling was spatially coordinated. As before [2.22], we merged neighboring genes of the same state into blocks, and identified chromatin domains as those blocks that were significantly larger than expected by chance (see Methods for detail). We found that 1,874 genes were contained in a significant domain in at least two cell lines. In the following we refer to these as the domain-associated genes.

1,874 genes are found in the domain-associated group. Interestingly, we found there is a significant overlap between this group and the "hotspot" genes identified by our ANOVA analysis. 11.2% of the "hotspot" genes fall into this category (p-value = 0.0346). In addition, the genes that are significant only in one modification are also

strongly associated (p-value < 0.01). These results suggest that chromatin state remodeling does not occur by random but is regulated in a spatially coordinated manner, perhaps through active maintenance of the domain boundaries.

One of the classical examples of chromatin domains is the HOXB gene cluster (Figure 2.4a). In particular, in the ES cell lines, the genes found in this domain are in the non-active state, which is characterized by high H3K27me3 occupancy. In differentiated cells, the HOXD genes switched to the null state. The histone gene cluster on chromosome 6 also undergoes a domain-level change during cellular differentiation (Figure 2.4b). These genes are mostly in the active state in ES cells, and then some of them (HIST1H14D to HIST1H3G) switch to a null state for most of the multipotent cell lines. There is then a switch back to a mostly non-active state for the unipotent/differentiated cell lines.

A small subset (containing 171 genes) of the domain-associated genes are persistent in the sense that they are embedded in a domain in almost every cell type (n > 21). This set of genes consists of specialized gene clusters, such as the olfactory receptor (OR) clusters, the keratin-associated protein cluster, and the Leukocyte Ig-like receptors (LIRs) cluster. These gene clusters tend to be silenced in almost every cell type except for few highly specialized cell types such as the olfactory receptors, keratin cells, or leukocytes. Consistent with gene silencing, these genes are not associated with active histone modifications.

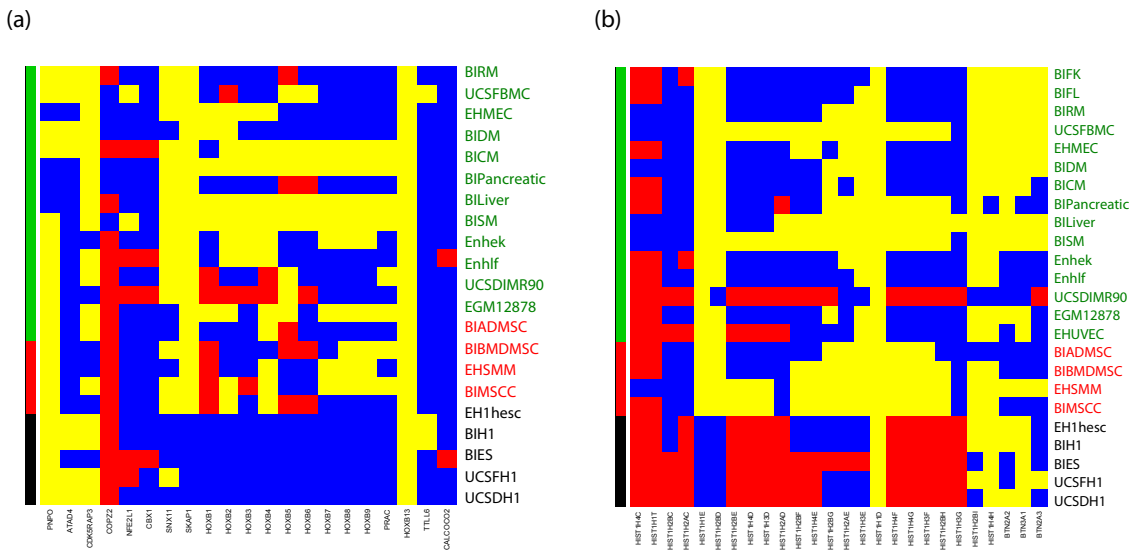

Figure 2.4: Representative chromatin domains identified by the hidden Markov model. Heatmap of the chromatin state distribution at the local genomic loci. (a) the HOXB gene cluster; (b) The histone gene cluster. Genes are ordered according to their genomic positions. Red – active state; yellow – null state; blue – non-active state. The cell line information is shown at both sides of the heatmap and color-coded by the differentiation status (black – plutipotent (P) cells; red – multipotent (M); green – unipotent/differentiated (U/D)).

2.2.5 Histone modification patterns in cancer cells are similar to differentiated cell types

The high accuracy of our chromatin-based classification models suggests that it may be useful for classification of poorly characterized cell-types. Cancer cells do not follow normal cell differentiation pathway and their lineages are poorly characterized. It has been noted that cancer cells often display characteristics similar to stem cells. Our above results suggested that chromatin states can be used to provide mechanistic insights into the relationship between cancer and stem cells. To this end, we applied our classification models to three cancer lines (K562, HeLa, VCaP), for which we were able to obtain the histone modification data. Strikingly, each cell line was unequivocally

classified as U/D (Table 2.1), suggesting significant and robust chromatin structural differences between cancer and stem cells (either pluripotent or multipotent cells).

Table 2.1: Outcome of classifying the differentiation status of three cancer cell lines (K562, HeLa, and VCaP) by applying the support vector machine to histone modification data at different levels. The results are represented as three numbers, corresponding to the number of models for which the cell line classified as pluripotent (P), multipotent (M), or unipotent/differentiated (U/D), respectively.

	K562	HeLa	VCaP
Bin (5 models)	0/0/5	1/0/4	0/0/5
Gene, single mark (5 models)	0/0/5	0/0/5	0/0/5
Chromatin state (1 model)	0/0/1	0/0/1	0/0/1

We further investigated the association between chromatin state changes and known regulatory modules. In a recent study, Kim et al. identified three ES regulatory modules based on protein-DNA interaction data [2.18]. These modules correspond to target genes of core ES cell regulators such as OCT4 and NANOG (called the Core module); of Polycomb group complexes (called the PRC module); and of the MYC related regulators (called the MYC module), respectively. These authors found that the gene expression patterns of cancer and stem cells are similar for the MYC module but significantly different for the other two modules. These observations led us to compare the chromatin state organization at these modules between different cell-types. As a simple quantitative metric, we evaluated the fraction of genes within each module falling into the non-active state (Figure 2.5a). The difference between different cell-lineage groups is apparent. Furthermore, the cancer cell lines seem to be distinct from the stem cells (either pluripotent or multipotent).

To test whether the activity of these modules is sufficient to explain the differences among differentiation stages, we built a multinomial logistic regression model to classify differentiation status solely based on the three module-specific chromatin signatures (Methods). This simple model already has 88% classification accuracy. We then applied this module-based model to classify the differentiation status of the three cancer cell lines. The results are similar to those obtained from the full SVM models. All three cancer cell lines (HeLa and VCaP) were classified as U/D. Therefore, the differences between cancer cells and stem cells can be interpreted simply by the differential activity of the three ES regulatory modules.

We divided the cell lines into groups of similar chromatin states by hierarchical clustering (Figure 2.5b). Both pluripotent cells and multipotent cells form distinct clusters, providing additional support to our classification results. Interestingly, two of the three U/D cell lines that were clustered together with multipotent cells may also be viewed as somewhat undifferentiated, since they are both fibroblasts and can undergo further differentiation and become more specialized. On the other hand, the three cancer cell lines are not only clustered with U/D cell lines, but also positioned next to the cell lines from the same germ-layer.

In addition, we found that 290 genes have different chromatin state patterns compared to the normal differentiated cells (Supplemental Table 2.8). 197 of these genes are in the non-active state in cancer cells. Only 10 genes from this list overlap with the "hotspot" genes that differentiating normal cell lines across differentiation stages, suggesting that these cancer cell lines contain additional chromatin signature that is distinct from normal differentiated cell lines. Functional analysis suggested that these

genes are enriched with genes associated with nucleosome organization (p-value = 0.0002).

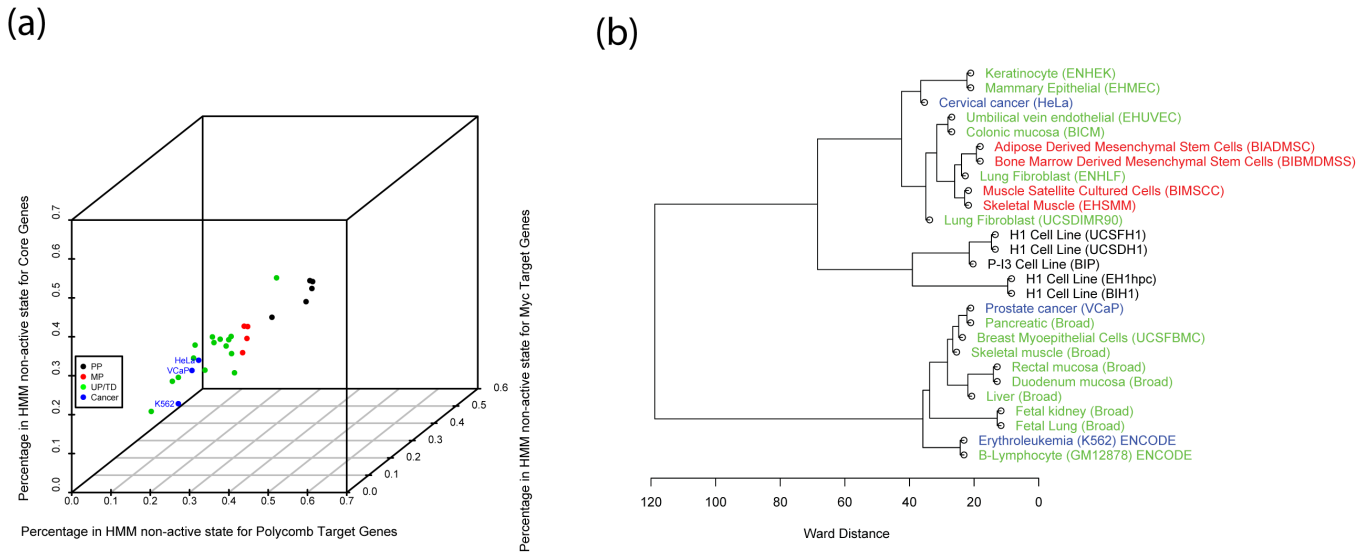

Figure 2.5. Cancer cells display similar chromatin state patterns as fully differentiated cells. (a) A scatter plot of cell-type specific chromatin states associated with the three ES regulatory modules. The chromatin state of a module is summarized by the fraction of non-active-state genes. Each data point corresponds to one cell-type and is color-coded according to the differentiation status. The three cancer cells are labeled. (b) Hierarchical clustering of the 27 cell lines based on the chromatin states.

2.2.6 Intrinsic chromatin state variability is associated with multiple human diseases

Aberrant epigenetic regulation has been linked to many diseases, including cancer, endocrine, and respiratory diseases [2.28, 2.29]. We hypothesized that such alteration may be partially contributed to intrinsic variability that occurs during normal differentiation and reasoned that, if so, the "hotspot" genes identified by our study should be significantly associated with various diseases.

By using the DAVID analysis tool again, we found that the "hotspot" genes are significantly associated with chemical dependency diseases (adjusted p-value = 0.0075;

associated genes include HTR2C, CHRNA4, and APOE), developmental disease (adjusted p-value = 0.0081; associated genes include NLGN3, and GLO1), and physiological diseases (adjusted p-value = 0.0092; associated genes include APOL2 and OXT). In contrast, we did not find any cancer type with significant association, with the lowest adjusted p-value at 0.92 (for prostate carcinoma). This lack of association further supports our view that chromatin states in cancer cells are fundamentally different from stem cells.

2.3 Discussion

Through a systematic analysis of a large dataset containing 5 histone modifications in 27 human cell lines at three different representations, we found that chromatin states can classify cell differentiation stages with nearly 100% accuracy. To our knowledge, this is the first study to classify cell differentiation stages based on chromatin information. Our results strongly suggest that the chromatin states are co-regulated at each developmental stage. We identified 722 "hotspot" genes, whose chromatin states are significantly associated with the differentiation stages. These genes are enriched with functions related to development, cell-cell signaling, and chromatin structure.

The success of our classification model led us to test if it can be used to gain insights into the origin of cancer cells. To this end, we applied our model to classify three cancer cell lines for which we obtained genome-wide histone modification data from the public domains, including K562 (chronic myelogenous leukemia), HeLa (cervical cancer), and VCaP (prostate cancer). We found that these cancer cells can be

unequivocally classified as differentiated cells based on the chromatin states, and that the differences between cancer and stem cells can be interpreted simply by using the three regulatory modules identified in ES cells [2.18]. Furthermore, all three cancer cell lines were clustered next to cell lines from the same germ layer, suggesting they may indeed originate from normal differentiated cells. Our analysis has provided new insights into the different regulatory mechanisms between cancer and stem cells. In future work, it will be very interesting to characterize the chromatin states in tumor samples and to investigate to what extent the chromatin states are associated with clinical outcome.

Recent studies have identified large-scale domains formed by various epigenetic marks [2.12, 2.13, 2.30]. A major difference between our and the aforementioned studies is that we treat each gene as a distinct unit, thereby ignoring the interruption of histone modification patterns at intergenic regions which may not be relevant for gene regulation. This allowed us to identify active domains despite the absence of active histone marks in intergenic regions. We also found that the "repressive" domains can be further divided into two types: "non-active" and "null". Their main difference is that, while the "non-active" domains are associated with high H3K27me3 activity, the null domains do not appear to be associated with any histone mark. It will be interesting to further investigate that the null domains may be associated with certain repressive histone marks that are not included here. These two domain types also differ functionally. The non-active domains are associated with poised gene activation, and the null domains seem to be able to achieve more stable gene silencing and therefore are desirable for the regulation of highly specialized gene clusters such as keratin and olfactory receptors.

The extensive presence of null domains was first discovered in mouse ES cells, where we found that many OR genes were associated with this pattern [2.22]. The functional relevance has been supported by a recent experimental study, which showed that the transition from the null state to a new state marked by H3K9me3 and H4K20me3 plays an important role in the development of olfactory neurons [2.31]. Our analysis here has extended these previous studies, suggesting that transition from null states may be a general mechanism for control of cell-type specific gene regulation. It will be very interesting to experimentally test this hypothesis in future studies.

2.4 Methods

2.4.1 Raw data processing

Bin level. Raw ChIP-seq data were divided into 100bp bins via BEDTools [2.32] and normalized to reads per million reads (RPM) to allow comparison across cell lines. Bins that overlapped 50% or more with known repetitive regions [2.33] were removed from further analysis. Remaining bins were merged into 1 kilobase (Kb) regions. The bins with no reads in any cell line were removed. Ultimately, we were left with 2,388,489 bins for further analysis.

Gene level. Gene annotations were based on Refseq [34]. Promoter regions were defined as the [-2Kb, +2Kb] region with respect to transcription start sites (TSS). For H3K4me1, H3K4me3, H3K27me3 and H3K9ac, the gene-level scores were defined by averaging normalized sequence reads over each gene promoter. For H3K36me3, it has been shown that the sequence reads are highest at 80-95% of the coding region of a gene

[2.4]; the gene-level scores were defined as by averaging over these regions. After removing the genes that substantially overlap with repetitive regions, we were left with 18,385 genes for further investigation.

Chromatin state level. The chromatin states were detected using a hidden Markov model (HMM) as previously described [2.22]. Briefly, the HMM combines gene-level scores for all five histone modifications (the emission variable) and classifies them as distinct chromatin states (the hidden variable). For simplicity, the emission probability is modeled by a multivariate Gaussian distribution with no covariance structure. To determine the optimal number of chromatin states, we clustered these five-dimensional vectors using the k-means average agglomeration clustering method. The optimal cluster number k was selected using the gap statistic [2.23] defined as

$$Gap(k) = E(\log(W_k)) - \log(W_k^*)$$

where W^*_k is the observed within-cluster sum of squares around the clusters means for one run, and E(·) represents the mean value for 1000 random bootstrap permutations. The gap statistic is maximized at k = 3 (Supplemental Figure 2.1).

We initially fit a three-state model separately for each cell line on chromosome 22, and used the expectation-maximization (EM) algorithm to estimate these model parameters. These cell-type specific models were averaged to obtain a single common model, which was then applied to determine the genome-wide chromatin states in the 27 cell lines via the Viterbi algorithm [2.35].

The bivalent genes were identified similar to the traditional definition [2.3], but with the modification necessary to map to the bin-level data. Specifically, we identified all the bivalent bins, that is, those 1Kb bins that overlap with both H3K4me3 and

H3K27me3 peak locations, where the peak locations were detected by adapting the CisGenome algorithm to bin-level data [2.36]. The bivalent genes were identified as those whose promoter overlaps with at least one bivalent bin.

Domain level. For each cell line, consecutive genes sharing the same chromatin states were merged as domains. As in previous work [2.22], we further used a likelihood-ratio test to identify significant domains in order to remove those domains that simply occur by chance. Specifically, for each domain, we calculated the ratio of the likelihood of observing its corresponding gene-level scores under the assumption that they were in the common chromatin state to the likelihood of observing the same data under the null hypothesis of no domain states. We estimated the null distribution based on 1,000 random permutations of all genes, and selected a cutoff domain size corresponding to the false discovery rate (FDR) at 0.05. Only those domains larger than the cutoff size were deemed significant and retained for further analysis. This analysis was repeated separately for each cell line; therefore the results are independent of the composition of cell lines in our assembly. A gene that is embedded in a significant domain in at least two cell lines is called domain-associated.

2.4.2 Cell differentiation stage classification

We classified the differentiation status (pluripotent (P), multipotent (M), or unipotent/differentiated (U/D)) from histone modification data by using support vector machines (SVM) [2.37] using either linear

or radial

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) \equiv \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) = exp(-\gamma \, ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2) ,$$

(where $\gamma>0$, and is estimated by cross-validation) kernel functions [2.38]. To determine which of the three ChIP-seq data representations (i.e., the bin, gene, and chromatin state levels) is most informative, we analyzed each representation independently. For the gene level analyses, all genes were used. For fair comparison, an equal number (i.e. equal to the number of genes) of most variable bins were used to construct the model. Calculations were done with the R package e1071 [2.39].

2.4.3 A simple classification model based on three regulatory modules

Using the three ES regulatory modules (described in the main text), we fit the following multinomial logistic regression model on the 24 normal cell lines:

$$\log\left(\frac{\pi_{ij}}{\pi_{ij}^*}\right) = \alpha + \mathbf{x}_i^T \beta_j, \qquad j \neq j^*$$

where π_{ij} is the probability that the i^{th} cell line *(i =1, 2, …, 24)* is a member of differentiation class *j (j = 1, 2, 3)* and $\boldsymbol{x}_i$ is the percentage of genes in one of the three modules assigned to the null state. The i^{th} cell line will be classified as belonging to the j^{th} class, if $\pi_{ij} > 0.5$. It is possible that all π_{ij} are less than 0.5; in such a case, the differentiation status will be classified as unknown.

2.4.4 Identification of differential regions

ANOVA analysis (F-test) was used to detect differential regions (bin/gene), where the histone modification patterns change significantly in accordance to the differentiation status. The null hypothesis was that there were no systematic differences across different cell-lineage groups. This test was conducted via the multtest package in R [2.40]. To correct for the multiple hypothesis testing bias, we calculated the FDR values by using 100,000 random permutations [2.41]. A cutoff value of FDR = 0.05 was used to select differential regions.

2.5 References

2.1. Strahl BD, Allis CD (2000) The language of covalent histone modifications. Nature 403: 41-45.

2.2. Kouzarides T (2007) Chromatin modifications and their function. Cell 128: 693-705.

2.3. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125: 315-326.

2.4. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823-837.

2.5. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39: 311-318.

2.6. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, et al. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459: 108-112.

2.7. Ernst J, Kellis M (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol 28: 817-825.

2.8. Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, et al. (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature 441: 349-353.

2.9. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553-560.

2.10. Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, et al. (2008) Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol Cell 30: 755-766.

2.11. Wei G, Wei L, Zhu J, Zang C, Hu-Li J, et al. (2009) Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. Immunity 30: 155-167.

2.12. Wen B, Wu H, Shinkai Y, Irizarry RA, Feinberg AP (2009) Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. Nat Genet 41: 246-250.

2.13. Hawkins RD, Hon GC, Lee LK, Ngo Q, Lister R, et al. (2010) Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell 6: 479-491.

2.14. Hebenstreit D, Gu M, Haider S, Turner DJ, Lio P, et al. (2010) EpiChIP: gene-by-gene quantification of epigenetic modification levels. Nucleic Acids Res 39: e27.

2.15. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473: 43-49.

2.16. Allison DB, Cui X, Page GP, Sabripour M (2006) Microarray data analysis: from disarray to consolidation and consensus. Nature reviews Genetics 7: 55-65.

2.17. Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, et al. (2008) An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nat Genet 40: 499-507.

2.18. Kim J, Woo AJ, Chu J, Snow JW, Fujiwara Y, et al. (2010) A Myc network accounts for similarities between embryonic stem and cancer cell transcription programs. Cell 143: 313-324.

2.19. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799-816.

2.20. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, et al. (2010) The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 28: 1045-1048.

2.21. Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, et al. (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature 471: 68-73.

2.22. Larson JL, Yuan GC (2010) Epigenetic domains found in mouse embryonic stem cells via a hidden Markov model. BMC Bioinformatics 11: 557.

2.23. Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society, Series B 63: 411–423.

2.24. Young MD, Willson TA, Wakefield MJ, Trounson E, Hilton DJ, et al. (2011) ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. Nucleic Acids Res 39: 7415-7427.

2.25. Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1-13.

2.26. Huangfu D, Anderson KV (2006) Signaling from Smo to Ci/Gli: conservation and divergence of Hedgehog pathways from Drosophila to vertebrates. Development 133: 3-14.

2.27. Massague J, Chen YG (2000) Controlling TGF-beta signaling. Genes & development 14: 627-644.

2.28. Jirtle RL, Skinner MK (2007) Environmental epigenomics and disease susceptibility. Nat Rev Genet 8: 253-262.

2.29. Jones PA, Baylin SB (2007) The epigenomics of cancer. Cell 128: 683-692.

2.30. Thurman RE, Day N, Noble WS, Stamatoyannopoulos JA (2007) Identification of higher-order functional domains in the human ENCODE regions. Genome Res 17: 917-927.

2.31. Magklara A, Yen A, Colquitt BM, Clowney EJ, Allen W, et al. (2011) An epigenetic signature for monoallelic olfactory receptor expression. Cell 145: 555-570.

2.32. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.

2.33. RepeatMasker website. Available: http://www.repeatmasker.org. Accessed 2012 January 12.

2.34. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35: D61-65.

2.35. Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis: Cambridge University Press. 356 p.

2.36. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, et al. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol 26: 1293-1300.

2.37. Cortes C, Vapnik V (1995) Support-Vector Networks. Machine Leaming 20: 273-297.

2.38. Keerthi S, Lin C-J (2003) Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel. Neural Computation 15: 1667-1689

2.39. Chang C, Lin C (2001) LIBSVM: a library for support vector machines.

2.40. Pollard K, Gilbert H, Ge Y, Taylor S, Dudoit S (2005) multtest: Resampling-Based Multiple Hypothesis Testing, 2005. R package version 1.23.5.

2.41. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B 57: 289-300.

Chapter 3

Tiered hidden Markov models predict epigenetic chromatin domains

Jessica L. Larson[1] and Guo-Cheng Yuan[1,2]

[1]Department of Biostatistics, Harvard School of Public Health,
Boston, Massachusetts, USA
[2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute,
Boston, Massachusetts, USA

3.1 Introduction

In Eukaryotic cells, the first level of DNA compaction is the wrapping of 147 bp of DNA around the nucleosome, an octomer of histones consisting of two copies of each of the core histone proteins: H2A, H2B, H3 and H4 [3.1, 3.2]. The nucleosomes are further compacted to form higher-order chromatin structures. Each core histone has an N-terminal tail that can be biochemically modified to affect the affinity of transcriptional machinery, thereby regulating gene expression. These modifications modulate the interpretation of the primary sequence, regulating cellular identity and development [3.3, 3.4]. Different histone modification marks play different roles in gene regulation. For example, H3K4me3 is associated with active transcription start sites; H3K36me3 is involved in transcriptional elongation, while H3K27me3 is associated with gene silencing [3.5-3.8]. The chromatin state, a combinatorial pattern of multiple histone modification markers as a whole, controls the epigenetic status of each gene [3.4, 3.9].

Since virtually all cells share the same genome, the cell-type specific gene activity can only be achieved by differential interpretation of the genome. A major layer of control is provided by the cell-type specific genomic organization of chromatin states. Several computational methods have been previously developed to systematically characterize the chromatin states. For example, Filion et al [3.10] applied principal component analysis and divided the *Drosphila* genome into five chromatin domain types based on genome-wide binding locations of by 53 chromatin proteins. Both we [3.11, 3.12] and Ernst et al [3.13, 3.14] developed a hidden Markov model (HMM) to characterize chromatin states. While Ernst and colleagues [3.13, 3.14] focused on the fine resolution, assigning each 200bp bin in the human genome to one of 15 chromatin states,

we treated each gene as a unit. These previous studies have shed light on the epigenetic patterns found in eukaryotes, but one limitation is that they all operate on a single length scale. It is well known that the genomic distribution of histone modification marks vary over multiple length scales, in part due to the hierarchical organization of the chromatin structure. A model that takes into account this higher-order structure would provide the best fit and interpretation of the epigenomic data.

In this study, we developed a tiered hidden Markov model (THMM) approach to simultaneously model the variation of chromatin states at two distinct length scales: bin and domain. We applied this approach to characterize the chromatin state in human embryonic stem cells based on a set of ChIP-seq data containing five histone modifications (H3K4me1, H3K4me3, H3K9me3, H3K27me3 and H3K36me3). Based on the data from a truncated genome (corresponding to annotated genes), we found that the overall chromatin state distribution can be well described using a model with three domains and five sub-states per domain. We compared the performance of our model and an existing model and found our model led to improved functional coherency. By applying our model to the whole genome, we find many long noncoding intergenic RNAs (lincRNAs) that are demarcated by active chromatin domains.

3.2 Results

3.2.1 Data pre-processing

ChIP-seq data from a human embryonic stem cell line (ESC) in five histone marks (H3K4me1, H3K4me3, H3K9me3, H3K27me3 and H3K36me3) was collected

from the UCSD Human Reference Epigenome Mapping Project [3.15]. After removing repetitive regions, we made a truncated genome containing only the promoter and coding bins. Our five modifications are most active near promoters and gene bodies and have a very low signal in the intergenic regions. Thus, the truncated genome data has more numeric variability and less skew toward regions with little or no ChIP-seq data than the whole genome data (i.e., there is an increased chance of observing high activity in our truncated genome, whereas there is almost no such chance in the whole genome due to the overwhelming size of the intergenic genome). The truncated genome was used for subsequent model estimation due to its simplicity and mathematical flexibility.

3.2.2 Model selection

In previous studies, others and we have applied hidden Markov models to identify chromatin states and proved to be successful [3.11-3.14]. Here we wanted to extend this approach to identify clusters of states that share similar properties. To this end, we developed a tiered hidden Markov model (THMM) approach, which contains two steps. In the first step, we apply a fully connected HMM to fit the data, as a way to assess the global variation patterns. In the next step, we refine the HMM topology to cluster the hidden states that are similar to each other. To determine the optimal number of states, we fit a series of HMMs, with the number of states ranging from 3 to 30 on the truncated chromosome 22 (Ch 22). We then compared the log likelihoods of these models to corresponding models fit on a permuted dataset. After 15 states, the difference between these two types of models begins to taper; thus adding more than 15 states will not drastically improve the model fit (Supplemental Figure 3.1). As an additional validation,

we also enumerated the number of combinatorial patterns in a simple manner, where each ChIP-seq data was transformed to a binary representation via thresholding the level of sequence reads. Out of the 32 possible combinations, we found that 15 were sufficient to explain the pattern of 99% of bins on Ch 22 (Supplemental Figure 3.2). This analysis has provided additional support that our choice of 15 hidden states is appropriate.

Our goal was to determine the structure of these states at the domain level and which prior would lead to the most probable model. For our priors, we explored two options: a random, non-informative prior and a prior where the means and variances of our states were based on hierarchal clustering. To this end, we fit each of three differently structured models (a three domain model with five sub-states per domain, a five domain model with three sub-states per domain and a four domain model with four sub-states per domain) with two types of priors, for a total of six THMMs. The model with the largest average likelihood was a three domain model with five states per domain and a completely random prior (Supplemental Figure 3.3).

3.2.3 Tiered hidden Markov model

A three domain model with five states each domain (Figure 3.1) was fit on the truncated genome. Our model allows for inter-domain transitions through only one of the states in each domain (Supplemental Table 3.1). We found that the states within each domain indeed share similar properties. Specifically, the states in Domain 1 (States 1-5) are associated with high levels of H3K27me3; the states in Domain 2 (States 6-10) have relatively little activity in any modification; and the states in Domain 3 (States 11-15) are characterized by high levels of H3K4me3, and H3K36me3, and low values of

H3K27me3 (Table 3.1, Supplemental Table 3.2 and Figure 3.2). Since these patterns are similar to what we found in previous studies [3.11, 3.12], we named the three domains as non-active, null, and active, respectively.

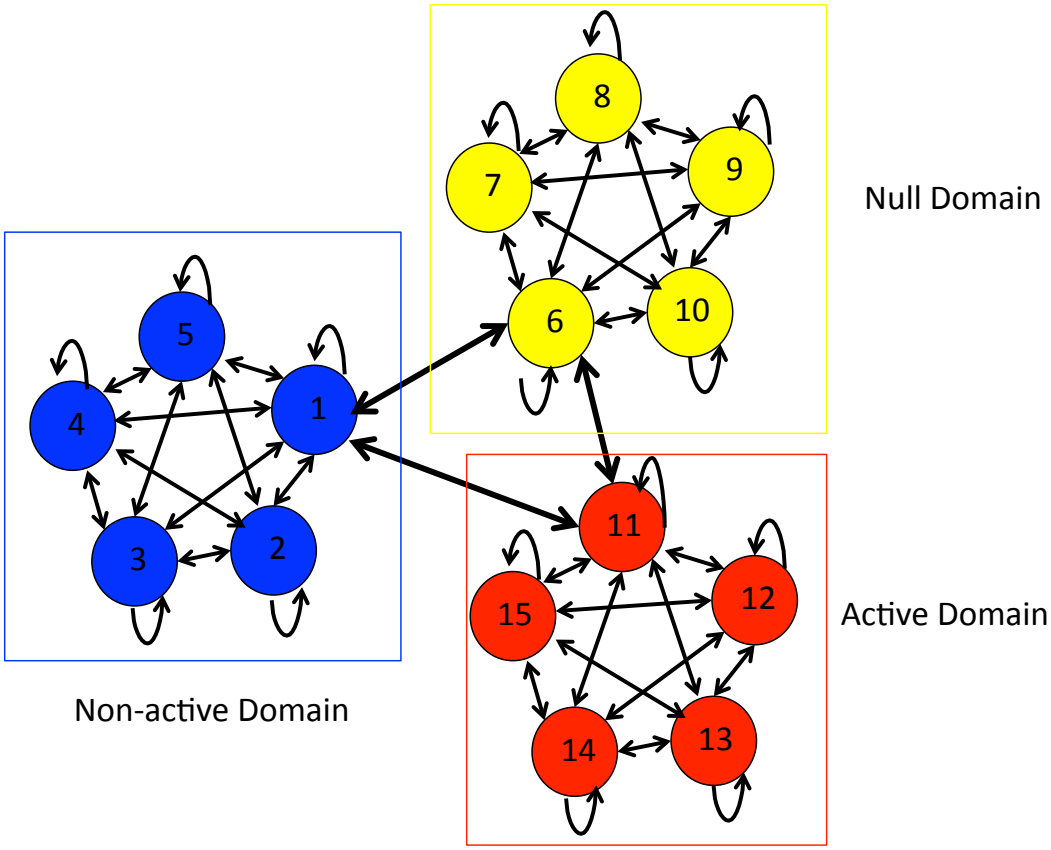

Figure 3.1: The topology of our THMM.
Each circle represents a sub-state within a boxed domain. Non-active domain states are shown in blue, null in yellow and active domain states are in red. Note that only one state per domain can lead to an inter-domain change. The rest of the states can only transition within their respective domains.

Table 3.1: Posterior means for each state. The active domain is associated with active marks, the non-active domain is associated with the repressive mark H3K27me3, and the null domain is not associated with any mark examined here.

	H3K4me1	H3K4me3	H3K9me3	H3K27me3	H3K36me3
1	15.896	4.5759	6.0922	11.25	4.3886
2	14.351	12.85	15.161	127.64	10.488
3	19.129	40.165	6.5304	51.966	4.5995
4	31.208	7.0683	5.5242	5.2304	5.0529
5	11.884	122.75	4.0895	8.5239	2.3941
6	10.93	3.1584	4.7478	3.9943	5.5029
7	3.1636	2.0858	4.3066	3.0364	11.356
8	2.0534	1.5287	2.6397	2.2567	2.491
9	3.3543	2.2208	7.2868	3.9244	5.4976
10	5.3693	2.3397	5.0364	5.371	3.2326
11	9.0342	3.9727	11.521	4.4245	15.042
12	3.2078	2.2509	4.8334	3.2469	25.815
13	7.4797	272.76	4.6026	4.7943	3.321
14	14.955	33.633	20.328	10.467	36.836
15	22.194	55.043	5.0483	4.6467	5.6511

3.2.4 Domain properties

The majority (95.6%) of the truncated genome is assigned to the null domain (Figure 3.3). More specifically, these bins are assigned to State 8, the state with the lowest histone modification density (Figure 3.2 and Supplemental Figure 3.4). This is in agreement with previous results.

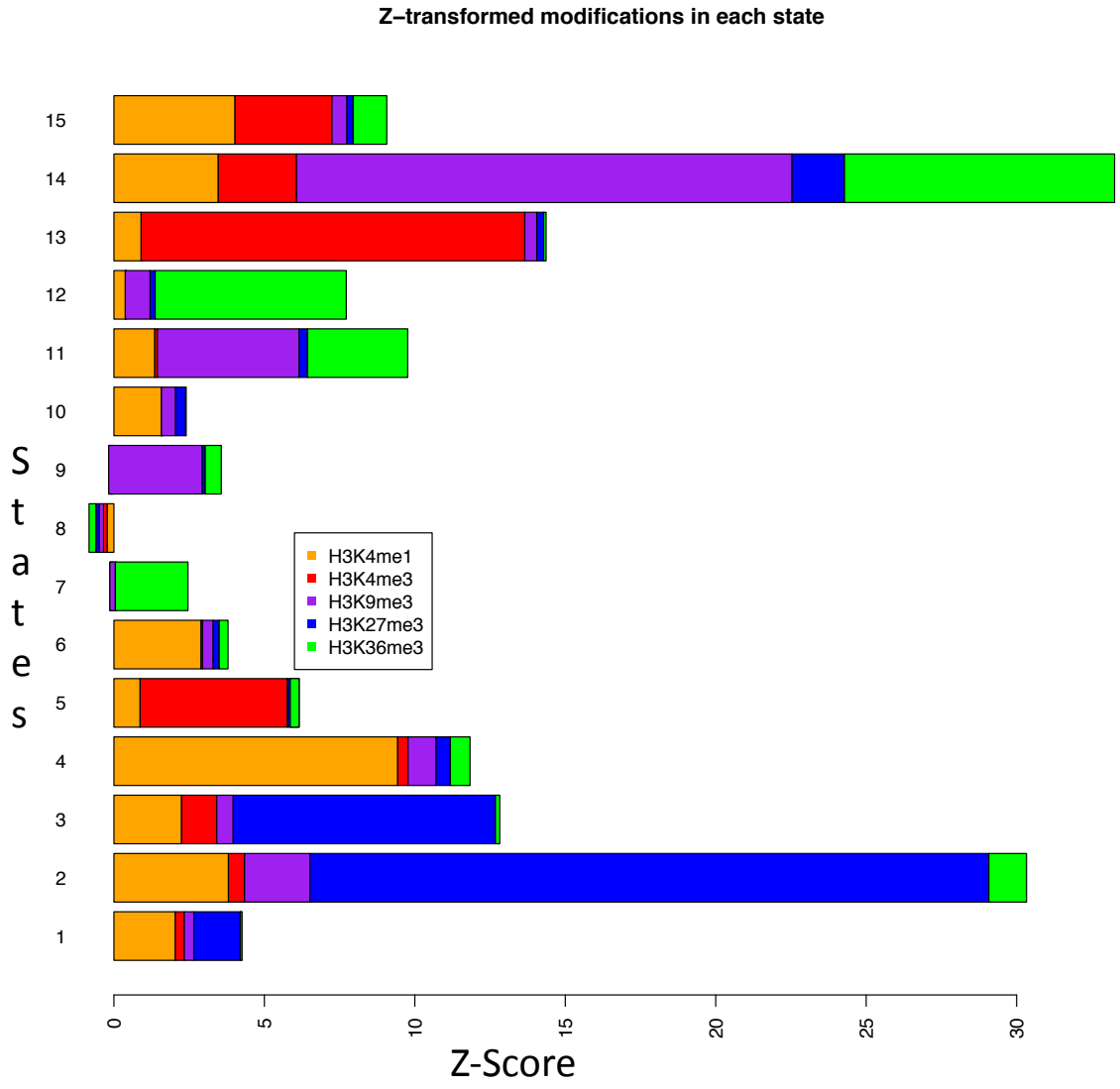

Figure 3.2: The distribution of the modifications in each state plotted as a Z-score. Z-scores for each modification within each state are calculated as follows:

$$Z_{ik}=(X_{ik}-\mu_i)/\sigma_i,$$

where X_{ik} is the average count for modification $i=1,2,3,4,5$ within state $k=1,2,\ldots15$, μ_i is the average modification level in the truncated genome, and σ_i is the standard deviation of modification i in the truncated genome.

The average domain in our truncated genome is 35.6 bins (~7KB) long. The null domains are the longest on average, with a mean length of 53.9 bins (~11 KB) long, and often covering multiple genes. On the other hand, both active and non-active domains

are relatively short (~5.4 KB and ~2.2KB, respectively) and are often contained within a single gene. Therefore, unlike in our previous work [3.11, 3.12], most domains do not span over multiple genes. However, these domains still capture certain large-scale features, as described below.

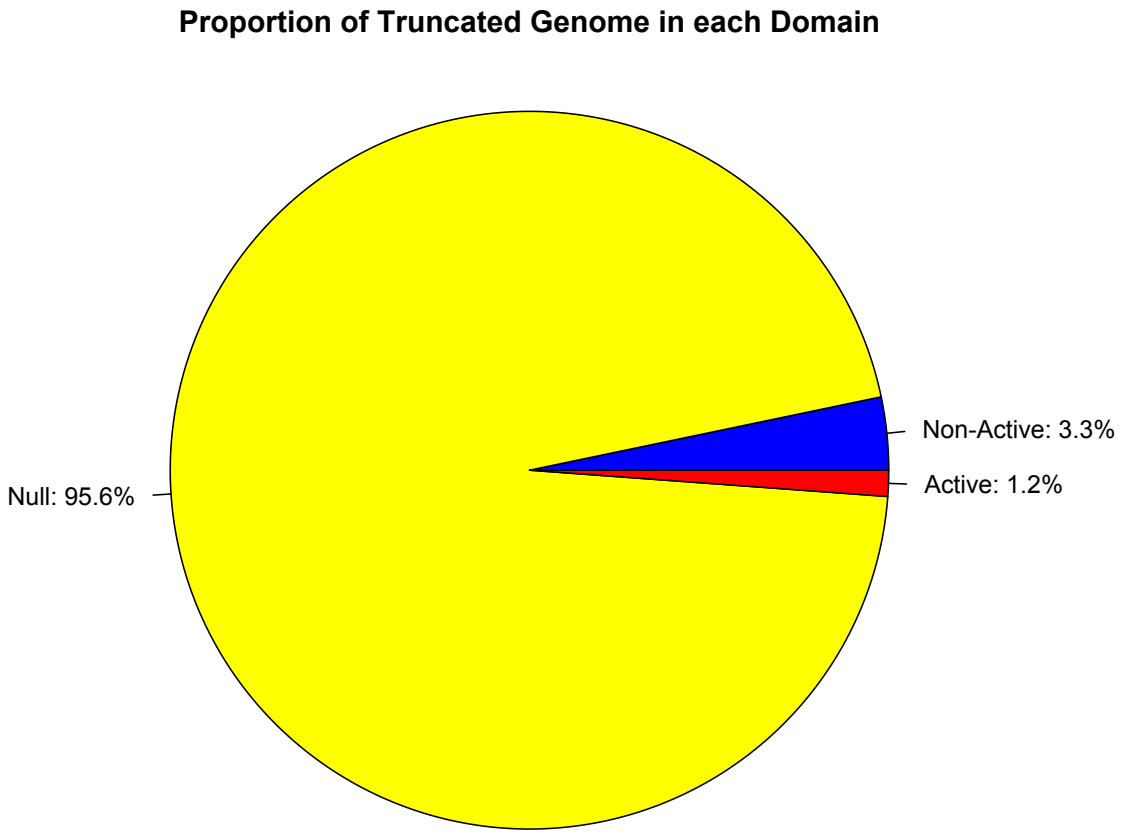

Figure 3.3: The distribution of domain assignment for the truncated genome. The vast majority of bins are assigned to the null domain.

Each domain has a unique distribution among the promoter, intron and exon bins (Figure 3.4a). Most null domain bins correspond to introns, whereas most non-active bins (such as States 1, 3, and 5) correspond to the promoter regions and most active bins (such as States 11, 12, and 14) correspond to the gene bodies (Figure 3.4b). This suggests that

there is a selection for coding and non-coding regions at the epigenetic level and that introns are often ignored by our marks, unless they are in active genes.

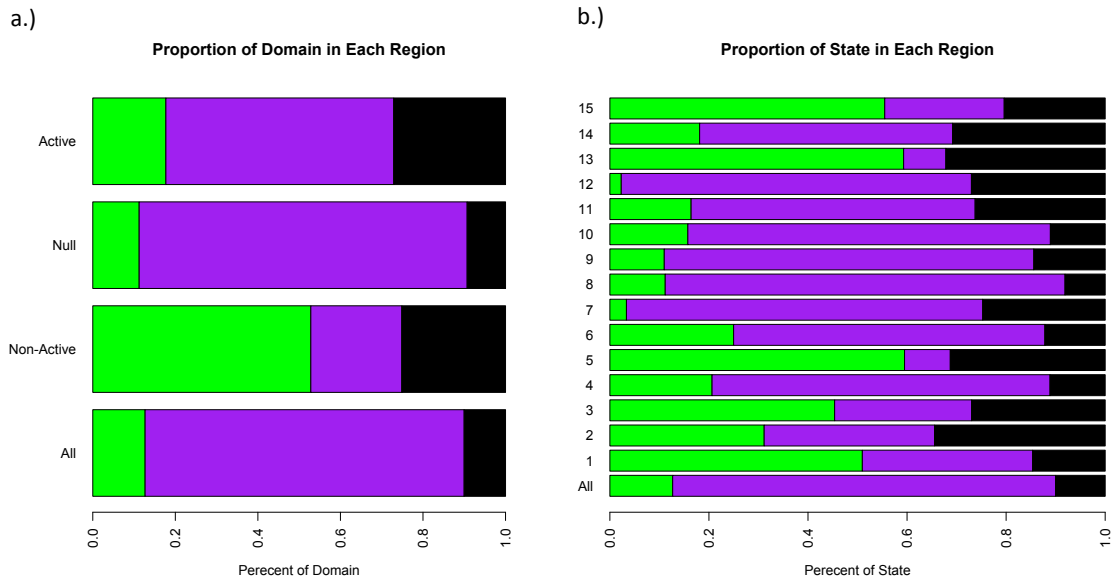

Figure 3.4: Location of domains and states.
The proportion of each (a) domain and (b) state found in promoter (green), introns (purple) and exons (black) regions.

If the active domains indeed mark transcriptionally active regions, they should be enriched for high level of mRNA transcripts. To test this prediction, we obtained RNA-seq data from the literature [3.18], and mapped the sequence reads to the bins as we did for the ChIP-seq data. As predicted, those bins in the active domain (States 11-15) correspond to the highest levels of RNA-seq compared the non-active and null domains (Figure 3.5). In addition, we also found that bins in the non-active domains have a slightly higher expression value than those in the null states, consistent with the previous finding that H3K27me3 mark does not prevent transcription initiation.

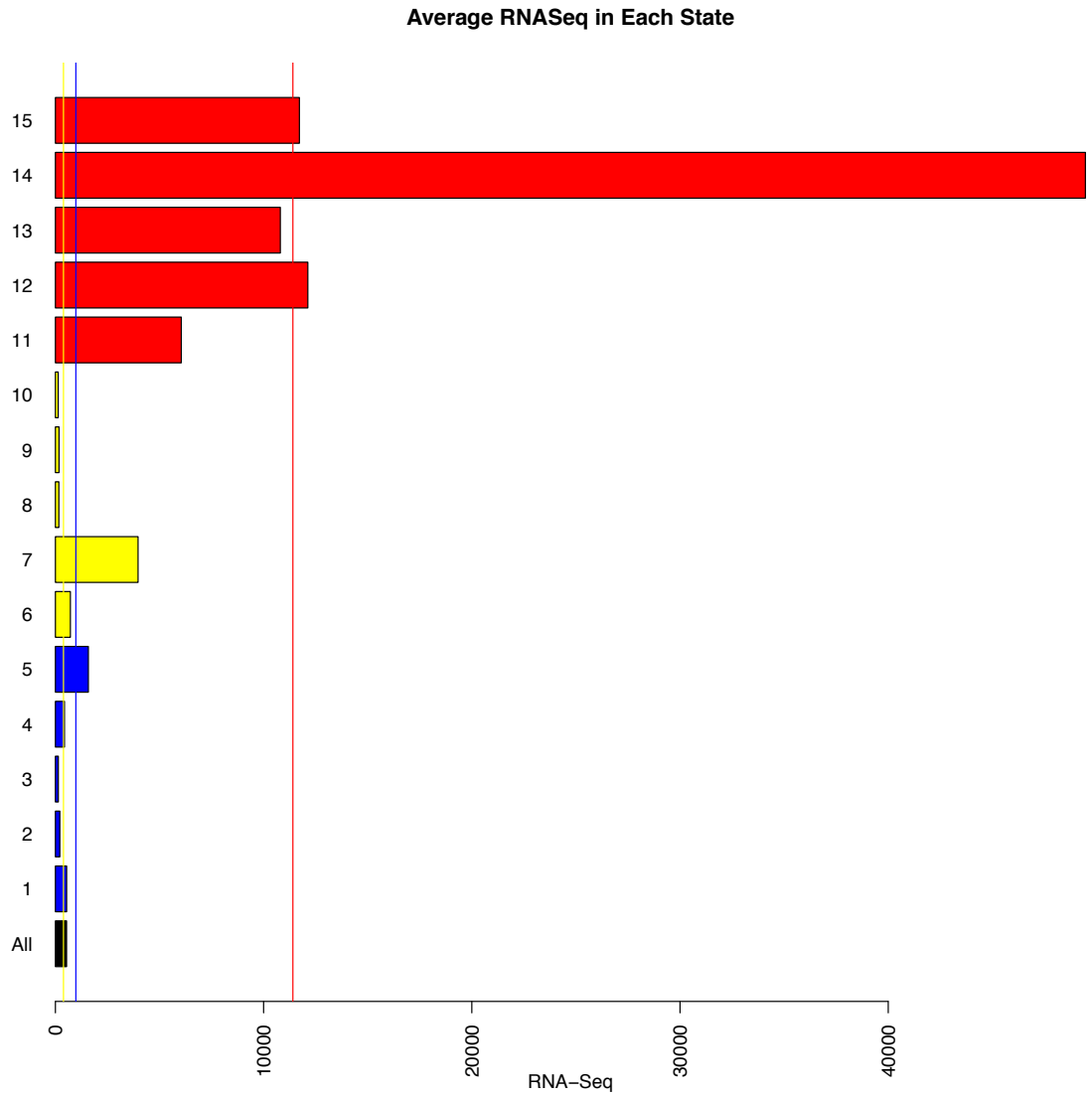

Figure 3.5: Transcriptional activity for each state
The average RNA-seq values for bins assigned to each state. Vertical lines represent the mean value for each domain. The bottom track is the truncated genome-wide average.

To gain functional insights, we applied the Database for Annotation, Visualization and Integrated Discovery (DAVID) to identify enriched functional categories that are associated with each domain type [3.23]. For each gene we calculated its most prominent state, after adjusting for the domain's genome wide distribution. We have 9,484 non-active genes, 11,703 null genes and 2,130 active genes for this cell line. Genes assigned

to the active domain are associated with cell maintenance groups such as the zinc finger region, ribosome, RNA binding, and cell cycle (Figure 3.6), which are all known to be active in embryonic stem cells. Null domain bins are involved in activities related to fully differentiated tissues such as olfactory receptor, defense response, taste and vision. Non-active bins are related to development and cell growth such as the regulation of transcription, homeobox gene clusters, mitochondrion and, the negative regulation of cell death. Thus, our ES cells are repressing developmental genes to maintain pluripotency, activating essential cell livelihood genes and ignoring genes only expressed in adult cells.

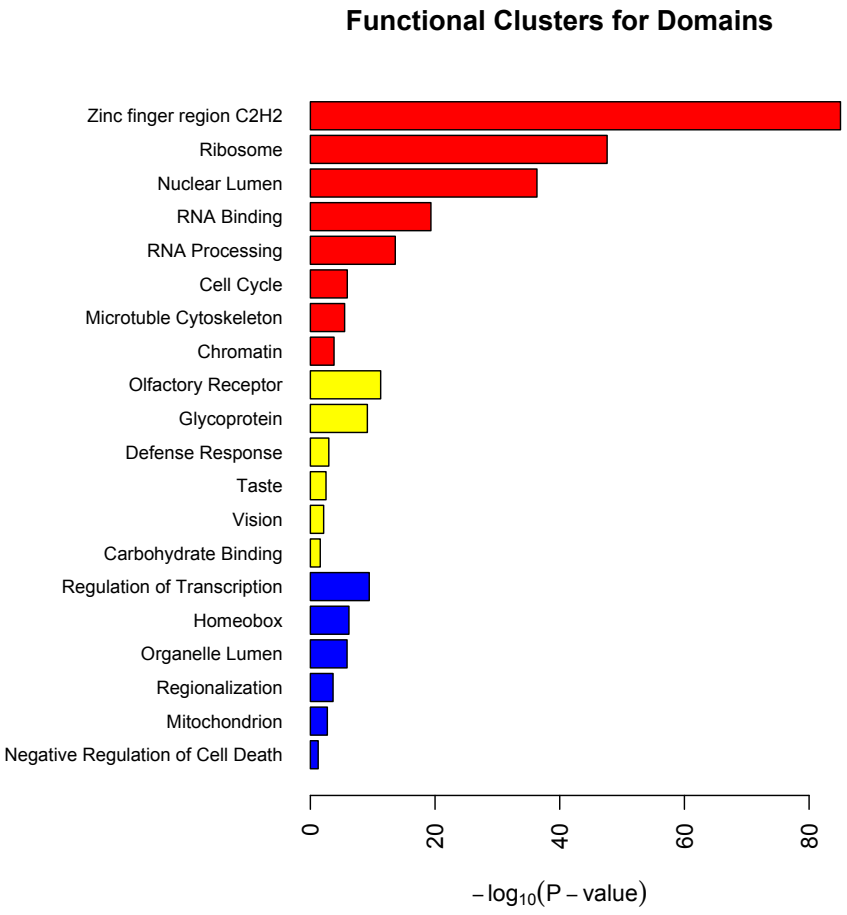

Figure 3.6: Functional analysis of each domain
Representative enriched functions and pathways identified using DAVID. P-values are adjusted for multiple testing.

3.2.5 Differences between sub-states within each domain

While the states within each domain share similar properties as described above, there still exist significant differences between them, especially within the active sub-states. The high percentage of active domain introns (Figure 3.4a) is due to the gene-body preference distribution of active states 11, 12 and 14 (Figure 3.4b). Note that these three states have the highest H3K36me3 signal (Figure 3.2). They are also most enriched for RNA transcript levels (Figure 3.5). In contrast, States 13 and 15 are more enriched in promoter regions of active genes (Figure 3.4b). Taken together, these active states are enriched in active genes.

Similarly, each state in the non-active domains also has its own property, with States 1, 3 and 5 more strongly enriched in promoter regions than the other two states (Figure 3.4). Interestingly, the combinatorial pattern of State 5 is similar to an active state. This is not surprising, since active and repressive histone marks often colocalize in ES cells, forming bivalent domains [3.24]. In contrast, State 2 has the highest H3K27me3 (Figure 3.2) and some of the lowest RNA-seq values (Figure 3.5) of all the states, but is not preferentially located (Figure 3.4b). These results suggest that non-active domains are enriched both in poised and in silent genes.

Our null states have less variability than those in the active and non-active domains. However, there also exist subtle differences among the different states. For example, States 6 and 10 have higher H3K4me1 density (Figure 3.2) and are more enriched in promoter regions than the other states (Figure 3.4b). Also, States 6 and 7 are associated with weak but recognizable signals of H3K36me3 and RNA-seq. However, in

general these null states can be simply described as marking the regions with low transcriptional activity.

3.2.6 Comparison with Ernst et al model

In order to test if our THMM approach is advantageous to detect coherent chromatin signature, we compared our tiered HMM results with a previous model developed by Ernst and colleagues [3.13, 3.14]. They fit a 15 state standard HMM on the genome in nine cell lines. For comparison, we used their software to fit a 15 state HMM on the five histone marks from our truncated genome. Hierarchal clustering of the corresponding emissions from the resulting model (denoted EK) shows that their 15 state model corresponds to three clusters, which allows for a parallel comparisons to our three domains (Supplemental Figure 3.5).

To determine which method leads to the most coherent (i.e. least variable) state, domain, or cluster assignment we evaluated our model, the EK model, and a purely random model in two ways: (a) with auto-correlation to determine which model preserves domain coherency over the largest genetic distance, and (b) with a kappa statistic to determine the level of agreement between active exons and domain or cluster status.

We wanted a measure of the correlation of state assignment between one bin and its neighbors to determine the level of spatial cohesion. We thus calculated the auto-correlation for our model and the EK model to determine which model is best at persevering state status over genetic distance. Because our domains are discrete events, we used Cohen's Kappa Statistic [3.21] to measure the agreement between the domain status in bin x and the domain status in bin $x+n$, where n is some genetic distance.

Kappa assesses the concordance between two assigners, and is more robust than a simple percentage of agreement measure because it takes into account the agreement occurring by chance. We found that our model (from the domain perspective) outperforms EK's model, even at the cluster level (Figure 3.7a). The THHM has a higher level of domain agreement over a longer distance than the EK model. There is thus evidence that our model retains domain status over quite a long distance; even at a distance of 2Kb, our model still retains a kappa of 0.6, compared to 0 expected by chance. This suggests that bins 2KB apart from each other still have a much higher level of domain coherency than in the EK model or than expected by chance.

Finally, because our active states are enriched in exons (Figure 3.4) and have high RNA-seq values (Figure 3.5), we calculated the agreement, measured by Cohen's Kappa, between active exons and our state assignments. Again, kappa is a measure of agreement of two or more raters. Here, we are comparing the active domain status of a bin with its corresponding active exon status. High levels of kappa indicate high levels of accordance. We compared our kappa to those from the EK model and 1000 random permutations of our states (Figure 3.7b). We see that our active domains have the highest agreement with the active exons (kappa = 0. 39, p-value < 0.001). We also noticed that the first (kappa = 0. 14, p-value < 0.001) and second (kappa = 0.034, p-value < 0.001) EK clusters also have significant but weaker agreement with the active exons. Note that because there is not the same number of active exons as active domain bins, the maximum kappa for this data is not 1, but 0.48 (see Methods). Therefore our model is close to the maximum possible level of agreement, suggesting that the functional coherence is maintained.

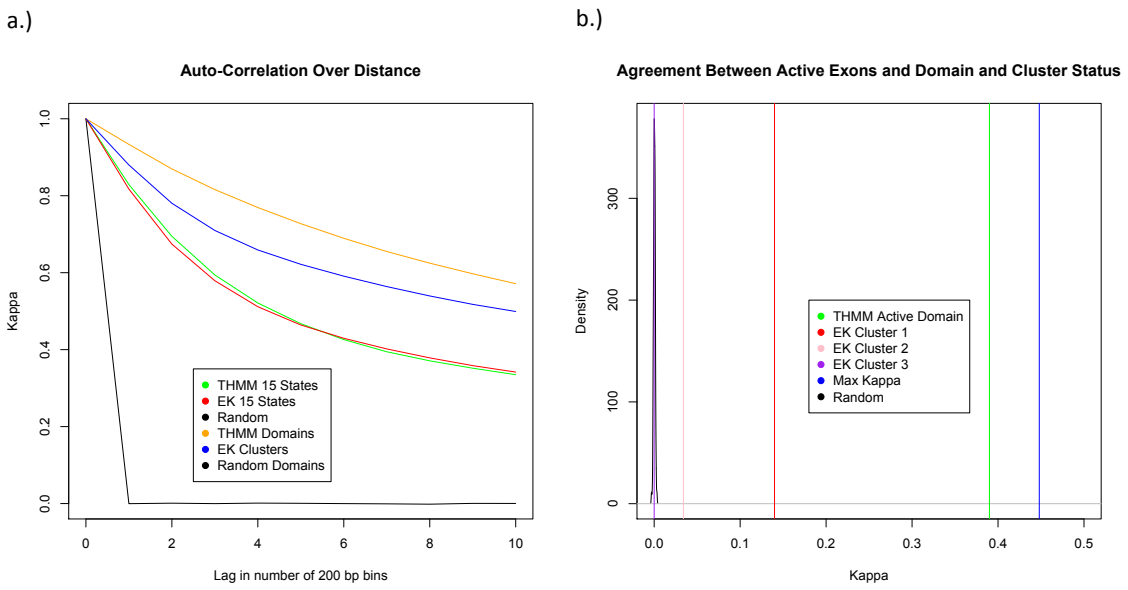

Figure 3.7: Comparison of our model to that of Ernst et al [3.14].
(a) Auto-correlation of state or domain agreement between bins over genetic distance, for our THMM as 15 states and 3 domains and EK's model as 15 states and 3 clusters. X-axis is the lag in number of 200bp bins, y-axis is the corresponding kappa statistic, which is a measure of accordance. Note that our domain-based model (orange) has the highest level of concordance over any genetic distance. Even at a distance of 2KB, we retain a kappa statistic of 0.6, compared to the random model that has no agreement. (b) Agreement, as measured by kappa, between the status of our active domains and each of the EK clusters and the status of our active exons; the black line is the corresponding kappa of a random domain assignment.

3.2.7 Genome-wide predictions

We applied the THMM, which was learned on the truncated genome, to make genome-wide predictions. As expected, the vast majority (98.8%) of intergenic bins were assigned to the null domain (Figure 3.8) and more specifically to the state with the least amount of histone mark activity, State 8 (Supplemental Figure 3.6). Our intergenic null domains (~ 26.4 KB) are much longer than in the truncated genome (10.8 KB). On the other hand, there are 34,024 intergenic bins that fall into the active domains. These domains are much shorter on average than in the truncated genome (~2.8KB compared to

~5.43 KB). We hypothesized that these active regions are actively transcribed. To test this hypothesis, we compared the locations with non-coding RNA transcripts [3.18]. Indeed, we found that the active bins have an average RNA-seq (1617.8) that is 25 times that of the other intergenic bins (66.38), although the transcription level is significantly weaker compared to the coding regions (11406.13, p-value<0.0001) (Supplemental Figure 3.7).

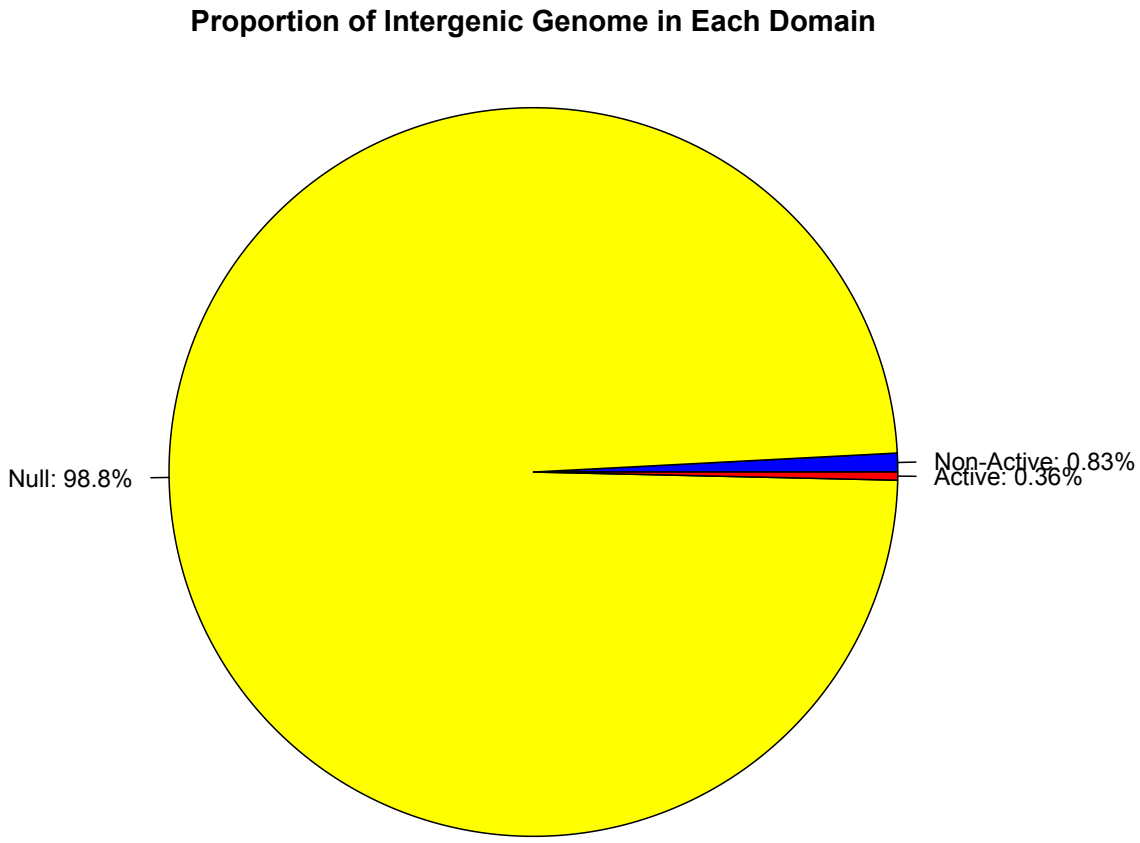

Figure 3.8: The distribution of domain assignment for the intergenic genome. 98.8% of the truncated genome is assigned to the null domain, 0.83% to the non-active domain, and 0.36% to the active domain.

One important class of ncRNA is long intergenic non-coding RNAs (lincRNAs), which have been increasingly recognized as key regulators of diverse cellular processes [3.25-3.27]. To test whether lincRNAs are enriched in active domains,

we aligned a human embryonic stem cell RNA-seq dataset [3.18], with a lincRNA annotation in human embryonic stem cell lines [3.28]. We evaluated the chromatin states of the bins overlapping with lincRNAs, and found that the lincRNA bins are significantly enriched with the active domain compared to other intergenic bins ($\chi^2 = 279.82$, df = 2, p-value < 2.2e-16). Moreover, lincRNA bins assigned to the active domains have average RNA-seq levels (4644) much higher than the rest of the active intergenome (1617.8, p-value = 0.01), and State 12 has an RNA-seq level almost 4 times greater than those bins assigned state 12 in the truncated genome (4.61×10^5 versus 1.20×10^5) (Supplemental Figure 3.8a). Only states 11 and 12 have a strong transcriptional signal among the lincRNAs, which suggests that lincRNA transcription is based on a different and exclusive combination of histone marks than in the coding regions, mainly high levels of H3K36me3 without H3K4me3 (Figure 3.2). Nevertheless, our model, due to its multiple sub-states, is excellent at picking out these active regions based solely on these epigentic patterns in the intergenic regions of the genome.

Another interesting class of features in intergenic regions is pseudogenes, which are dysfunctional fossils of coding genes that, through duplication, deletion, insertion or retrotransposition events, have lost their ability to be transcribed into operative products. However, it remains unclear whether there are major epigenetic differences between pseudogenes and their functional counterparts. We aligned the most recent annotation [3.29-3.31] with our intergenic bins and found that the majority of the pseudogenes are indeed embedded in the null domain. Nonetheless, twice as many such bins are assigned to the active domain (2.54%) than in the truncated genome (1.2%, p-value<0.0001), which is consistent with the fact that these pseudogenes are still transcribed at a higher

rate (3165.71) compared to the rest of the intergenome (66.38043, p-value<0.0001). Our most transcriptionally active bins that overlap with pseudogenes are assigned to the non-active domain (Supplemental Figure 3.8b), suggesting that there are fundamental epigenetic differences between transcribed pseudogenes and their functional counterparts in the truncated genome, which tend to be embedded in active domains as described above. Taken together, these results provide evidence that our model can lead to new biological insights.

3.3 Discussion

Chromatin plays an important role in diverse biological processes such as gene transcription, DNA replication and repair. As genome-wide location data of various histone modification marks are generated at a rapid speed, it has become increasingly clear that new computational methods are needed to comprehensively describe the overall chromatin state. The THMM method developed in this paper is aimed to build on previous studies to further take into account the hierarchical structure of chromatin, which is important for understanding the overall topology of chromatin structure and the epigenetic mechanism for gene regulation. We have shown that our method indeed provides a useful modeling framework and outperforms existing approaches [3.13, 3.14] in terms of functional coherence.

Through our exploration of epigenetic data in human embryonic stem cells, we have annotated the human genome into three domains and 15 chromatin sub-states. Each of our sub-states has a unique epigenetic pattern, location preference, and transcriptional signal, while still retaining the specific characteristics of its corresponding

domain. Our active domain is characterized by high levels of RNA-seq and active modifications such as H3K4me3 and H3K36me3. Our non-active domain has high levels of H3K27me3, while our null domain has very little histone modification presence.

We also found that our model can provide useful biological predictions. While our model was learned on a truncated genome contains genetic regions only, we applied it to the whole genome. We found that the majority of the intergenic regions share the same epigenic pattern as introns. More importantly, our model also predicted a number of active domains in intergenic regions that are enriched with lincRNAs and has shown that the epigenomic signature at transcribed pseudogenes is fundamentally different from their functional counterpart. In future work, it will be very interesting to characterize the chromatin domains in multiple cell lines and to investigate to what extent they are associated with cell identity.

3.4 Methods

3.4.1 Data collection and pre-processing

ChIP-seq data for the human embryonic stem cell line (ESC) was obtained from the UCSD Human Reference Epigenome Mapping Project [3.15]. Five modifications (H3K4me1, H3K4me3, H3K9me3, H3K27me3 and H3K36me3) with well-known biological functions were chosen for simplicity. Raw ChIP-seq data was processed and separated into 200bp bins via BEDTools [3.16]. The data was then normalized to reads per million reads (RPM). Bins that overlapped 50% or more with known repetitive regions [3.17] were removed from further analysis. RNA-seq expression data for human

ESC was collected from the ENCODE project via the UCSC Genome Browser [3.18] and aligned to our 200bp bins.

Gene annotations, including transcription start and end points and exon and intron locations, were taken from RefSeq [3.19]. Gene redundancy was limited by the removal of genes with the same transcription start or end site or the same gene name. In order to focus on the role of chromatin states on gene regulation and to simplify computational analysis, we considered a truncated genome consisting of only the promoter and coding regions of known RefSeq genes. That is, for each gene, we extracted the region from 2KB upstream from the transcription start site to the transcription end site. As a result, the intergenic regions, which consist of the majority of the genome, were excluded from our initial analysis.

3.4.2 Model exploration and evaluation

We used a two-tier approach to identify the best hidden Markov model (HMM) topology (hence the term THMM) that describes the data. First, we assumed a fully connected topology, with the aim to estimate the total number of states. Second, we refined the topology to have local clustered properties. In the first step, using the bin-level five-dimensional score as a multivariate normal emission vector, we fit a series of HMMs. We fit 28 HMMs, with the number of states ranging from 3 to 30 on truncated chromosome 22 (Ch 22). We randomly permuted this same dataset 10 times in both the genetic location and histone modification dimensions to get a permuted genome and subsequently fit each model on these random permutations to get a corresponding average log likelihood. For each possible number of states, we compared the log

likelihoods of the appropriate models. The model with the fewest number of states that led to a thresholding of the difference between the log likelihood of the structured and the average log likelihood of the corresponding random models was explored further.

In the second step, we introduced a tiered structure in the HMM topology so that the hidden states are divided into distinct locally-connected clusters (Figure 3.1). Each cluster corresponds to a distinct chromatin domain. The model topology can be specified by the number of domains (D) and the number of states for each domain (K). Since we determined the total number of states is 15, we compared three different THMM topology, corresponding to $D = 3$ and $K = 5$, $D = 5$ and $K = 3$, and $D = 4$ and $K = 4$ (4 states per domain), respectively. Model selection was done through a comparison of the likelihood of each model's fit on the truncated chromosome 22. Each of these THMMs was fit with one of two choices of priors: a random, non-informative prior and a prior where the means and variances of our states were based on hierarchal clustering.

We calculated the random priors by taking a random sample of 100 bins on our chromosome and determining the corresponding mean and variance vectors. This was repeated for each state, resulting in a *Sx5* matrix of means and a *Sx5* matrix of variances, where *S* is the total number of states (which is equal to *D* times *K*).

For the cluster-based prior, we first grouped the data into *D* clusters with kmeans clustering. Then, for each of our *D* clusters, we repeated the kmeans clustering to get *K* sub-clusters, giving us a total of *S* ($S = D \times K$) sub-clusters. For each sub-cluster, we calculated the mean and variance vectors to get a *Sx5* matrix of means and a *Sx5* matrix of variances.

The emission probability was modeled by a multivariate Gaussian distribution with no covariance structure. We fit each of our six THMMs, one with a random and one with a cluster-based prior for our three potential structures (*3X5*, *5X3*, and *4X4*), 100 times on truncated Ch 22. Parameters for each model were estimated with the expectation-maximization (EM) algorithm [3.20].

We evaluated the performance of each of our six models based on its corresponding log likelihoods and selected the final model based on model fit. That is, we compared the average likelihood from our six potential THMMs, and refit the model with the largest average log likelihood genome-wide.

3.4.3 Our tiered hidden Markov model

Using the Viterbi algorithm [3.20], the model parameters (means, variances and transitions) from the most successful model fit on truncated Ch 22, were applied to the entire truncated genome to get a corresponding state for each bin. We also applied our THMM to the full genome, including the intergenic bins with the Viterbi algorithm, and then removed the truncated bins to get an intergenic-only signal.

To determine a gene's most prominent domain, we calculated the following statistic:

$$T_{gd} = (P_{gd})/(P_d),$$

where P_{gd} is the proportion of gene, g, assigned to domain d, and P_d is the proportion of the genome assigned to domain d, for each domain type $d=1,2,3$. The domain with the highest T_{gd} was selected as that gene's domain for the purpose of our functional analysis.

3.4.4 Ernst et al. hidden Markov model

Ernst, et al. [3.13, 3.14] developed an HMM to characterize chromatin states in nine human cell lines. Briefly, their method assigns HMM states based on multidimensional binary chromatin marks. We applied ChromHMM to analyze the same dataset as described above. For this purpose, each ChIP-seq track was binarized by thresholding. A 15 state model was fit to the data on the truncated genome. For comparison with the domain level predictions from our THMM approach, we applied hierarchical clustering to group the output states into three clusters.

3.4.5 Quantifying domain retention

We evaluated the spatial coherence of a chromatin state and compared the EK model with ours. For our THMM and the EK model, we calculated the auto-correlation over the distance 200bp to 2Kb. Auto-correlation is a measure of the similarity between observations as a function of the distance separation between them. Since we were comparing state assignments (which are not normally distributed) over genetic distance, we used Cohen's Kappa [3.21] for our correlation measure:

$$\kappa = \frac{P_a - P_e}{1 - P_e},$$

where P_a is the proportion of agreement among bins and P_e is the expected agreement due to chance. Kappa can take values from 0 (no agreement) to 1 (perfect agreement). As a guideline, values over 0.75 are often considered excellent, while those below 0.40 are considered poor [3.22]. We also permuted the data and re-ran the kappa analysis over 2KB for comparison.

3.4.6 Active exon agreement

We assigned each exon in our truncated genome an active or non-active status based on its RNA-seq values. Exons with expression values above the median were considered active. We then compared the status of each exon to its domain and cluster status and calculated the agreement via the kappa statistic (see above). We then permuted our domain assignments and calculated to corresponding agreement between this permutation and our active exons 1000 times to get a distribution of kappa under the null hypothesis of no agreement. From this distribution, we are able to calculate p-values for our observed kappas; p-values are taken as the proportion of permutation-based kappas greater than or equal to our observed kappa.

Because kappa has a theoretical maximum value of 1 only when the observers distribute the same number of each code, and here we do not have the same number of active exons and active domain bins, we also calculated the maximum possible kappa value. Kappa max is:

$$\kappa_{max} = \frac{P_{max} - P_e}{1 - P_e},$$

where P_{max} is the maximum possible agreement based on the distribution of the states.

3.5 References

3.1. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389: 251-260.

3.2. Kornberg RD, Lorch Y (1999) Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. Cell 98: 285-294.

3.3. Jenuwein T, Allis CD (2001) Translating the histone code. Science 293: 1074-1080.

3.4. Kouzarides T (2007) Chromatin modifications and their function. Cell 128: 693-705.

3.5. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39: 311-318.

3.6. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553-560.

3.7. Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, et al. (2008) Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. PLoS Genet 4: e1000242.

3.8. Zhao XD, Han X, Chew JL, Liu J, Chiu KP, et al. (2007) Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. Cell Stem Cell 1: 286-298.

3.9. Campos EI, Reinberg D (2009) Histones: annotating chromatin. Annu Rev Genet 43: 559-599.

3.10. Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, et al. (2010) Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell 143: 212-224.

3.11. Larson JL, Yuan GC (2010) Epigenetic domains found in mouse embryonic stem cells via a hidden Markov model. BMC Bioinformatics 11: 557.

3.12. Larson J, Yuan G (2012) Chromatin states accurately classify cell differentiation stages. PLoS ONE.

3.13. Ernst J, Kellis M (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol 28: 817-825.

3.14. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473: 43-49.

3.15. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, et al. (2010) The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 28: 1045-1048.

3.16. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.

3.17. Smit A, Hubley R, Green P (1996-2010) RepeatMasker Open-3.0.

3.18. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511-515.

3.19. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35: D61-65.

3.20. Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis: Cambridge University Press. 356 p.

3.21. Cohen JAcoafnsEaPM (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20 37–46.

3.22. Fleiss J (1981) Statistical methods for rates and proportions. New York: John Wiley.

3.23. Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1-13.

3.24. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, et al. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125: 315-326.

3.25. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, et al. (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129: 1311-1323.

3.26. Guttman M, Amit I, Garber M, French C, Lin MF, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458: 223-227.

3.27. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, et al. (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A 106: 11667-11672.

3.28. Loewer S, Cabili MN, Guttman M, Loh YH, Thomas K, et al. (2010) Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nat Genet 42: 1113-1117.

3.29. Zhang Z, Harrison PM, Liu Y, Gerstein M (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res 13: 2541-2558.

3.30. Zheng D, Zhang Z, Harrison PM, Karro J, Carriero N, et al. (2005) Integrated pseudogene annotation for human chromosome 22: evidence for transcription. J Mol Biol 349: 27-45.

3.31. Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, et al. (2007) Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res 17: 839-851.

Supplementary Figures

Supplementary Figure 1.1:
Histogram plots of each of our five modifications and their corresponding Poisson (red) and Gaussian (blue) approximation distributions.

Supplementary Figure 1.2:
Posterior distributions and Viterbi path for the 615 genes on Ch19. Posterior probabilities for each of the three epigenetic-states are shown in the top three plots. The black horizontal line corresponds to 0.5 probability. The bottom plot is the state assignment for each gene, determined by the Viterbi path. Genes colored blue were assigned state 1 (non-active) by the Viterbi algorithm, yellow were assigned state 2 (null), and red were assigned state 3 (active).

Supplementary Figure 1.3:
Simulation results for the (a) two- and (b) three- state HMMs. The two-state HMM captures the truth 98% of the time, while the three-state HMM captures it 99% of the time. The top two tracks are a simulated H3K4me3 and H3k27me3 count, respectively. The third track is the true state (based on a random permutation), and the fourth track is the states as predicted by our model.

Supplementary Figure 1.4:
Simulation results for the (a) two- and (b) three- state HMMs where the simulated modification counts are based on high variance models. The two-state HMM captures the truth 97% of the time, while the three-state HMM captures it 99% of the time. The top two tracks are a simulated H3K4me3 and H3k27me3 count, respectively. The third track is the true state (based on a random permutation), and the fourth track is the states as predicted by our model.

Supplementary Figure 1.5:
Heatmaps for the NP cell line. (a) The 35 gene region on Ch6 from *Npy* to *2410066E13Rik* (49,772,728 to 54,650,400) as depictured as a heatmap of histone modification and gene expression. (b) The 250 gene region on Ch7 from *Art2a* to *Insc* (108,701,290-121,993,728) as depictured as a heatmap of histone modification and gene expression. NP HMM state assignments are in the first and second tracks (red for active state, blue for low state, yellow for null state). Whether (black) or not (white) a gene is a respective gene cluster in shown in the bottom track in all figures.

Supplementary Figure 1.6:
DAVID cluster analysis of genes within each ES significant domain in the NP cell line. Genes in each type of change are described by the top three significant DAVID clusters. Red corresponds to genes in the active state, blue for those in the non-active state and yellow for null state, in the NP cell line.

Supplementary Figure 1.7:
Log likelihood results for the 100 randomly chosen priors (black) versus a prior based on K-means clustering (red).

Supplementary Figure 2.1:
Determining number of chromatin states by using the gap statistic**.** The expected and observed log (WK) values are shown in (a) for various levels of K (the number of clusters), where WK is the pooled within cluster sum of squares around the cluster means. The number of clusters versus Gap(K), the difference between the observed and expected values (mean value for 1000 random bootstrap permutations), is shown in (b). According to these results, three is the optional number of clusters for our data.

Supplementary Figure 2.2:
Proportion of bivalent genes in different chromatin states identified by HMM. Red – active state; yellow – null state; blue – non-active state.

Supplementary Figure 2.3:
Overall distribution of the differential bins identified by the ANOVA analysis. (a) A Venn diagram showing the overlap among different histone modifications. (b) Enrichment of various functional elements in the differential bins, where the enrichment scores were computed by ratio between the frequency of differential bins falling into one functional element category and that expected by chance.

Supplementary Figure 2.4:
Overall distribution of the differential genes identified by applying the ANOVA analysis to gene-level scores. (a) A Venn diagram showing the overlap among different histone modifications. (b-e) Heatmap of gene-level scores for the 100 most differential genes: (b) H3K4me3. (c) H3K9ac. (d) H3K27me3. (e) H3K36me3. The cell line information is shown at both sides of the heatmap and color-coded by the differentiation status (black – pluripotent (P) cells; red – multipotent (M); green – unipotent/differentiated (U/D)).

Supplementary Figure 2.5:
Functional analysis of "hotspot" genes. "Hotspot" genes were identified by applying ANOVA analysis to the chromatin states inferred by the hidden Markov model. These genes were further divided into three categories based on their corresponding state in the ES cells: red – active; yellow – null; blue – non-active.

Supplementary Figure 3.1:
Comparison of the log likelihoods of 27 hidden Markov models. (a) The corresponding log likelihoods for HMMs fit on the true data (red) and the average log-likelihood of the models fit on the permuted data (blue) for models with 3 to 30 states. (b) The difference between the structured and permuted models is plotted for models containing 3 to 30 states. In both plots, the vertical line represents a model with 15 states, and is where the difference between the two models begins to taper.

Supplementary Figure 3.2:
Proportion of the genome assigned to each for a binary model. The top horizontal line represents 0.05, and the bottom one is 0.01. Thus, 99% of the genome is covered by 15 states.

Supplemental Figure 3.3:
Plot of the average -log-likelihoods for 6 different models. Error bars represent 1 standard deviation.

Supplemental Figure 3.4:
The frequency of state assignment for the truncated genome.

Supplemental Figure 3.5:
Clustering dendrogram of the posterior emissions from the EK model.

Supplemental Figure 3.6:
The frequency of state assignment for the intergenic genome.

Supplemental Figure 3.7:
The average RNA-seq values for bins assigned to each state in the intergenic genome. Vertical lines represent the corresponding value for each domain.

Supplemental Figure 3.8:
The average RNA-seq values for bins assigned to each state in for (a) lincRNAs and (b) pseudogenes. Vertical lines represent the corresponding value for each domain.

Supplementary Tables

Supplementary Table 1.1:
Significant domains for the two-state HMM.

Supplementary Table 1.2:
Significant domains for the three-state HMM.

Supplementary Table 1.3:
Genes in significant non-active domains for the three-state HMM.

Supplementary Table 1.4:
Genes in significant null domains for the three-state HMM.

Supplementary Table 1.5:
Genes in significant active domains for the three-state HMM.

Supplementary Table 1.6:
Corresponding NP domains for the significant ES domains fit by the three-state HMM.

Supplementary Table 2.1:
Description of the cell lines used in this study. Code used for cell-type group: P – pluripotent cell; M – multipotent cell; U/D – unipotent/differentiated cell; C – cancer cell.

Supplementary Table 2.2:
Mean (standard deviation) gene-level histone modification scores associated with each chromatin state identified by the hidden Markov model. The active state is associated with active marks, the non-active state is associated with the repressive mark H3K27me3, and the null state is not associated with any mark examined here.

Supplementary Table 2.3:
Cell-type specific chromatin states inferred by the hidden Markov model. (1 – null; 2 – non-active; 3 – active)

Supplementary Table 2.4:
List of differential bins for each histone modification mark.

Supplementary Table 2.5:
List of differential genes based on gene-level scores for each histone modification mark.

Supplementary Table 2.6:
List of "hotspot" genes whose chromatin states are significantly associated with differentiation status.

Supplementary Table 2.7:
List of genes whose chromatin states are significantly different between pluripotent cells and other normal cell types.

Supplementary Table 2.8:
List of genes whose chromatin states are significantly different between the normal and cancer cell lines.

Supplementary Table 3.1:
Transition probabilities for each state in our THMM, cell *i, j* is the probability of a transition from state *i* to state *j*.

Supplementary Table 3.2:
Histone modification variances for each state in our THMM.