COMPUTATIONAL ALGORITHM DEVELOPMENT FOR EPIGENOMIC ANALYSIS

A Dissertation
Presented to
The Academic Faculty

By

Jianrong Wang

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics

Georgia Institute of Technology

August, 2012

COMPUTATIONAL ALGORITHM DEVELOPMENT FOR EPIGENOMIC ANALYSIS

Approved by:

Dr. I. King Jordan, Advisor
School of Biology
Georgia Institute of Technology

Dr. Soojin Yi
School of Biology
Georgia Institute of Technology

Dr. Mark Borodovsky
Department of Biomedical Engineering
Georgia Institute of Technology

Dr. John McDonald
School of Biology
Georgia Institute of Technology

Dr. Yuhong Fan
School of Biology
Georgia Institute of Technology

Date Approved: June 28, 2012

ACKNOWLEDGEMENTS

I am very grateful to my advisor Dr. I. King Jordan for his guidance throughout my research and study as a Ph.D student. He introduced me to the exciting research field of epigenetics and next-generation sequencing data analysis, and is always helping me to learn the correct way of scientific research and critical thinking. His advice will continue to influence me for my whole academic career in the future.

I am also thankful to my committee members, John McDonald, Soojin Yi, Mark Borodovsky and Yuhong Fan for their valuable support and suggestions during my PhD study. Discussions with them have helped me to understand my research topic more thoroughly.

I am especially grateful to my colleagues, Lee Katz, Ahsan Huda, Andrew Conley and Daudi Jjingo for their friendship, which made the years of Ph.D study as an important memory in my life. I am also grateful to the other members of my lab, Kevin Lee, Aswathy Sebastian, Lu Wang and Angela Pena for interesting discussions for my research.

It is beyond words to express my gratitude for my wife Yu Zhang, for her love and support that has encouraged me to keep my interests in scientific discoveries.

Last but not the least, I am grateful to my parents, my mother Yingfan Chen, and my father Shu Wang, for their deep love that has filled my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

$CD4^+$	$CD4^+$ T cell
ChIP-seq	Chromatin Immunoprecipitation and Sequencing
CNE	Conserved non-coding elements
CTCF	CCCTC-binding factor
DNase I	DNase1- Hypersensitive site
DNA	Deoxyribonucleic Acid
EBA	Enhancer-blocking assay
EST	Expressed Sequence Tag
FE	Fold enrichment
GO	Gene Ontology
H2AZ	Histone variant H2A.Z
H2AK5ac	Histone H2A Lysine 5 acetylation
H2AK9ac	Histone H2A Lysine 9 acetylation
H2BK120ac	Histone H2B Lysine 120 acetylation
H2BK12ac	Histone H2B Lysine 12 acetylation
H2BK20ac	Histone H2B Lysine20 acetylation
H2BK5ac	Histone H2B Lysine 5 acetylation
H2BK5me1	Histone H2B Lysine 5 mono-metylation
H3K14ac	Histone H3 Lysine 14 acetylation
H3K18ac	Histone H3 Lysine 18 acetylation

H3K23ac	Histone H3 Lysine 23 acetylation
H3K27ac	Histone H3 Lysine 27 acetylation
H3K27me1	Histone H3 Lysine 27 mono-methylation
H3K27me2	Histone H3 Lysine 27 di-methylation
H3K27me3	Histone H3 Lysine 27 tri-methylation
H3K36ac	Histone H3 Lysine 36 acetylation
H3K36me1	Histone H3 Lysine 36 mono-methylation
H3K36me3	Histone H3 Lysine 36 di-methylation
H3K4ac	Histone H3 Lysine 4 acetylation
H3K4me1	Histone H3 Lysine 4 mono-methylation
H3K4me2	Histone H3 Lysine 4 di-methylation
H3K4me3	Histone H3 Lysine 4 tri-methylation
H3K79me1	Histone H3 Lysine 79 mono-methylation
H3K79me2	Histone H3 Lysine 79 di-methylation
H3K79me3	Histone H3 Lysine 79 tri-methylation
H3K9ac	Histone H3 Lysine 9 acetylation
H3K9me1	Histone H3 Lysine 9 mono-methylation
H3K9me2	Histone H3 Lysine 9 di-methylation
H3K9me3	Histone H3 Lysine 9 tri-methylation
H3R2me1	Histone H3 Arginine 2 mono-methylation
H3R2me2	Histone H3 Arginine 2 mono-methylation
H4K12ac	Histone H4 Lysine 12 acetylation
H4K16ac	Histone H4 Lysine 16 acetylation

H4K20me1	Histone H4 Lysine 20 mono-methylation
H4K20me3	Histone H4 Lysine 20 mono-methylation
H4K5ac	Histone H4 Lysine 5 acetylation
H4K8ac	Histone H4 Lysine 8 acetylation
H4K91ac	Histone H4 Lysine 91 acetylation
HMM	hidden Markov model
Kb	Kilo base-pair
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
Mb	Mega base-pair
MIR	Mammalian Interspersed Repeat
miRNA	microRNA
mRNA	messenger RNA
PCR	Polymerase Chain Reaction
Pol II	RNA polymerase II
Pol III	RNA polymerase III
Refseq	Reference Sequence Database
RIT	Regions in transition
RNA	Ribonucleic Acid
RNAi	RNA interference
SINE	Short Interspersed Nuclear Element
TCR	T-cell receptor pathway
TE	Transposable Element

tRNA	Transporter RNA genes
TSS	Transcription start site
TTS	Transcription termination site
YY1	ying-yang protein

SUMMARY

Epigenetics mainly refers to the biological phenomena that can alter gene expression and cellular differentiation without changing the underlying DNA sequences in the genome. The two main components of epigenetic regulation are DNA methylation and histone modifications (1,2). Both have been shown to be highly related with gene expression, and histone modifications, compared to DNA methylation, are more complex given that the tails of histone proteins can be modified by a variety of histone modifying enzymes. Different histone modifications are related with or participate in distinct regulatory processes and mark distinct regulatory elements, including transcription initiation, elongation, enhancers, insulators, imprinting and three-dimensional chromatin structures (3-11). Thus, a thorough understanding of the specific patterns of histone modification profiles and their associations with various genomic features is very important in functional genome biology research. Some general discoveries have been drawn from a number of studies on the dynamics of histone modifications in different cell types (3,4,7,12-18). But detailed analyses of the complex and diverse associations of histone modifications with genome regulatory systems, along with the underlying mechanisms, are still currently lacking.

Due to recent advances in next-generation high-throughput sequencing technology, experimental biologists have combined the chromatin immunoprecipitation technique with high-throughput sequencing (ChIP-seq) to obtain the genome-wide maps of various functional factors, including suites of histone modifications and transcription factors, in a number of cell types of different species (3,18,19). These large-scale datasets

provide rich sources of information for important regulatory functional systems or pathways of gene expression and cellular differentiation. In order to globally characterize the epigenomic modification landscape and its interaction with other transcription factors, effective and efficient computational algorithms and pipelines are needed. The development of such tools and approaches will play a critical role for the progress of research in epigenetics and gene regulation (19).

The algorithms and analyses that result from my own Ph.D. dissertation research provide both novel methods for basic data processing and advanced tools of information mining and pattern recognition for important biological questions. The specific advances of my research in the field of computational epigenomics are summarized as follows:

Research advance 1: A Gibbs sampling algorithm was developed to accurately map short ambiguous sequence tags back to the reference genome. Employing the information of neighboring tag-mapping profile information, the algorithm is the first which models the ambiguous read mapping problem in a unified Bayesian inference framework. It achieves better performances compared with existing methods, with respect to the higher fractions of correctly mapped ambiguous tags and higher accuracies of recovered real genomic sites measured by the recall, precision and *F* scores. The applications of this algorithm are shown to be able to discover more important biological signals in repetitive genomic regions, including transposable elements, simple repeats, peri-centromeric regions and segmental duplications.

Research advance 2: A broad peak calling algorithm was developed to identify enriched contiguous regions of diffuse ChIP-seq signals. Combining a Gibbs sampling procedure for parameter estimations of non-homogeneous Poisson processes and the maximal scoring segment algorithm, the method is capable of identifying broad peaks of various sizes, especially for certain histone modifications which can have peaks over 1 Mb. Compared with existing algorithms, this method requires many fewer parameters and is more data-adaptive. Applications of the method on simulated datasets prove that it has better recall and *F* scores, and it has better coverage for larger broad peaks. Application of this method to real ChIP-seq datasets from human cells shows that this algorithm is also useful for finding biologically meaningful patterns, such as large-scale chromatin states with particular regulatory meanings.

Research advance 3: A hypothesis-driven computational pipeline was designed to search for MIR retrotransposon derived insulators (MIR-insulators) and a list of such predicted insulators is found. This work presents one of the first reports of CTCF-independent insulators in the human genome. Several of the predicted MIR-insulators are experimentally validated by enhancer-blocking assays (EBA) in both human kidney cell lines and zebrafish embryos. Chromatin signatures, including histone modification profiles and RNA polymerase II and III bindings, are characterized for those putative MIR-insulators. Functional analysis of genes proximal to those MIR-insulators uncover that the T-cell receptor (TCR) pathway is enriched, and an interesting example of three adjacent TCR genes with pairs of MIR-insulators encompassing them is found. Comparative analysis of chromatin environments in different cell types classifies the MIR-insulators into cell type invariant and cell type specific groups.

Research advance 4: An unbiased (hypothesis-free) algorithm was developed to search for chromatin boundary elements in the human genome with possibly novel features and mechanisms. A list of boundary elements is predicted and a subset of them is

CTCF-independent. The classical boundary element BEAD1 (also the only one which is experimentally validated in human CD4$^+$ T cells) is successfully found by this algorithm. Specific combinatorial chromatin signatures are identified for those boundary elements and a set of interesting protein factors are predicted to be associated with boundary activity, including EVI1, USF and YY1. A subset of this list contains non-coding RNAs that are actively transcribed and bound by RNA polymerase III (Pol III). This is the first report of non-coding RNA, specifically tRNA, derived boundary elements in the human genome.

Research advance 5: An unsupervised algorithm was developed to predict novel combinatorial chromatin signatures without being restricted to the annotated genomic features or training datasets. As a high-dimensional pattern recognition method, it can do exploratory data analysis of genomic ChIP-seq datasets of various histone modifications and an inherent statistical criterion is derived for the final pattern identifications. The resulting combinatorial signatures are found to be related with distinct genomic features, such as transcriptional start sites (TSS), transcription termination sites (TTS), enhancers, conserved non-coding elements (CNE) and L1 retrotransposons. Bivalent signatures are also found and associated with cell type specific gene expression silencing. An additional advantage of this algorithm is that it is able to find both small and large signatures with very complex combinatorial profiles, *e.g.* spatially shifted enrichments of different histone modifications. Several large signatures are found to be highly related with gene bodies and have the potential to discover novel gene body annotations.

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Epigenetics and gene expression regulation

Variations of cellular phenotypes can exist even if the cells have identical genomic DNA sequences. These phenomena, along with their underlying mechanisms, are collectively called epigenetics (1,2). The two major topics of epigenetics research are DNA methylation and histone modifications. DNA methylation refers to the addition of methyl group to the 5 position of cytosine pyrimidine ring, and clear negative associations between DNA methylation states in promoter regions and gene expression levels have been observed in a number of studies (20). Compared to DNA methylation, histone modifications bear much more varieties, and their relationships with gene expression and cell differentiation are complex. Within each nucleosome, *i.e.* the basic units of chromatin structure, two copies of four histone proteins (H2A, H2B, H3 and H4) are wrapped around by DNA sequences, and the amino terminal tails of those histone proteins can be modified by different enzymes. The main types of histone modifications include acetylations, methylations, phosphorylations and ubiquitylations (1,2). Beyond the different types of histone modifications, the complexity comes mainly from the distinct locations on the tails of different histone proteins where the modifications occur. For example, the di-methylation of lysine 9 of H3 (H3K9me2) has very different effects compared with the di-methylation of lysine 4 of H3 (H3K4me2).

The basic models for the associations of histone modifications and gene expression can be classified into the following hypotheses (1,21). First, different histone modifications can change the local chromatin structure and thus affect the accessibility of the DNA sequences to transcription factors and/or RNA polymerase. Second, specific histone modifications and their associated chromatin enzymes can possibly be the recognition targets and recruit specific transcription factors that are necessary for transcription activation or repression. Third, specific histone modifications can be consequences, instead of causes, of different gene expression states.

Regardless of the detailed mechanistic hypotheses, a general picture of the associations between different histone modifications and gene expression activation/repression has emerged based on recent genome-wide analyses (3,7,8,11,16,18,22-24). Globally, histone modifications associated with active transcription and open chromatin are called active modifications and the ones associated with repressed transcription and closed chromatin are called repressive modifications. In promoter proximal regions, a set of histone modifications that are characteristic of open chromatin structures are strongly associated with active transcription initiation, including H3K4me3, H3K27ac and H3K9me1. In gene bodies, several broadly distributed histone modifications are associated transcription elongation, including H3K36me3, H3K79me2 and H3K79me3. In distal intergenic regions, some discrete locations that are marked by specific active modifications, including H3K4me1 and H3K27ac, are associated with active expression of genes and have been suggested as regulatory elements *in trans*. And a few specific repressive modifications are widely enriched within the repressive chromatin states, *e.g.* heterochromatin domains.

Besides the basic associations with specific gene expression patterns, the dynamics of histone modifications are largely related with cellular differentiation and reprogramming (7,12,14-16,21,25). Comparative analyses of the genome-wide landscapes of different histone modifications in various cell types have suggested that specific cell types are characterized by distinct histone modification profiles. One of the interesting observations is that some gene promoters modified by bivalent patterns (*i.e.* co-occurrence of active an repressive modifications) are paused for transcription in stem cells but turn to active transcriptions in differentiated cell types where the promoters become marked by active modifications alone (12). Most recently, the chromatin modifying enzymes are further shown to be modulators of cell type reprogramming which will help for mechanistic explanations of the relationship between histone modification landscape dynamics and cellular differentiation (26).

The biological importance of histone modifications is not restricted to their associations with gene expression. They are related with diverse biological pathways and the histone modification signatures contain abundant information to predict specific regulatory elements and distinct functional activities, including enhancers, insulators, replication timing, alternative splicing and three-dimensional chromatin looping (3,5,6,8-11,17,23-25,27-34). The most thoroughly studied case is enhancers. Distinct histone modification signatures are found to be informative to identify the cell type specific locations of enhancers (9,28,29). Thus, analyzing the genomic landscapes of histone modifications can provide valuable tools to capture cell-type specific regulatory systems.

ChIP-seq data analysis for epigenomics

The development of next-generation high-throughput sequencing technologies has enabled genome-wide analyses of histone modifications. By combining the chromatin immunoprecipitation technique with next-generation sequencing (ChIP-seq), researchers can effectively identify genomic locations of different histone modifications (using distinct antibodies). Compared with traditional experimental approaches, ChIP-seq has the following advantages: 1) it is more high-throughput and easier to scale up, 2) it has better signal to noise ratios, and 3) it has higher resolution for the locations of modifications (19). Due to these advantages, ChIP-seq has been employed by a number of genome-wide studies to investigate the distribution patterns of a suite of histone modifications in different species and cell types. As the experimental technique is now mature, more challenges come from the subsequent computational data analysis steps. In the last few years, many computational algorithms have been developed for effective and efficient data mining and analysis of large-scale ChIP-seq datasets. This new field is often referred to as computational epigenomics.

Generally speaking, computational epigenomic data analysis is composed of two different fields: 1) basic data processing, and 2) biological question driven data mining (Figure 1.1). The basic data processing deals directly with the short sequence tags (or reads) produced by ChIP-seq and generates genomic mapping profiles of tags that are noise reduced and can be analyzed for specific biological questions. Next, depending on the specific questions of interest, biological question driven data mining develops corresponding advanced algorithms which transform the genomic profiles of ChIP-seq tags to useful biological discoveries.

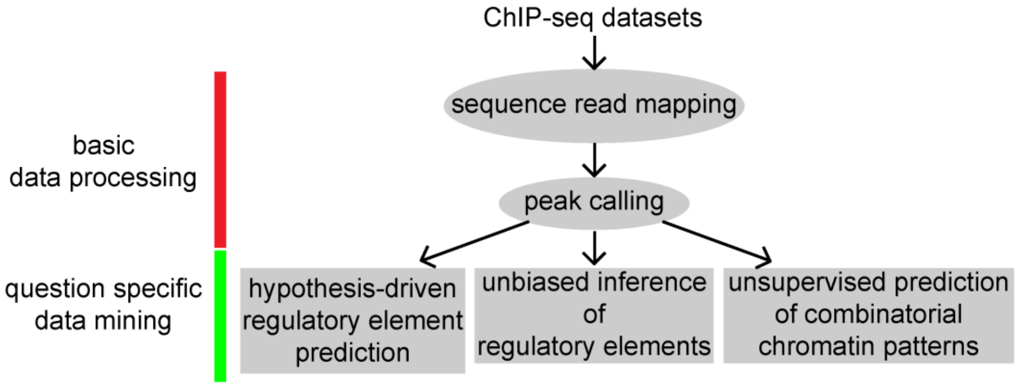


Figure 1.1: Steps of computational epigenomic data analysis.

There are two major steps in basic data processing, *i.e.* sequence read mapping and peak calling (Figure 1.1). Read mapping is the first and most important step of ChIP-seq data analysis since all subsequent analyses rely on accurate mappings of ChIP-seq reads. The main challenge is to make time-efficient methods for short sequence alignments that are applicable for large amounts of reads. There have been several successful algorithms available for read mapping (35-37), and one remaining problem is about ambiguous reads (also called as multi-mapping reads) which can be aligned to multiple genomic locations with almost identical sequence similarities. One solution for the ambiguous read mapping problem will be discussed in details in this dissertation. After accurate read mapping, the next critical data processing step is peak calling, namely identifying discrete genomic regions that have significantly higher tag-counts compared to the background noise levels (38-42). It can also be considered as noise reduction processing. The resulted discrete locations with significant tag-counts represent the real sites of the histone modifications or transcription factors, and characterize the basic landscapes of ChIP-seq profiles. The current methods mostly restricted their focuses on sharp and abrupt peaks, *i.e.* densely located small genomic bins with high tag-counts (38,39,42). Although this kind of peaks fit well with the observed features of some histone modifications and most transcription factors, certain histone modifications are widely distributed along large genomic regions without characteristic size ranges and accordingly their ChIP-seq datasets are diffuse (19). The corresponding peaks are called broad peaks and a novel algorithm for broad peak calling will be discussed later in this dissertation.

After the basic data processing, specific data mining algorithms are needed to solve different interesting biological questions (Figure 1.1). Depending on the specific biological question, descriptive features of the problem of interest are usually formulated into quantitative signatures or patterns. Based on the diverse characteristics of the signatures or patterns, appropriate pattern recognition algorithms need to be carefully selected and designed. Several interesting biological questions will be described in the following sections (Figure 1.1) and the corresponding analytical methods are detailed in subsequent chapters.

Enhancers and Insulators

Gene expression regulation is a complex process. Besides *cis* regulation through transcription factors, other types of regulations *in trans* also play critical roles, including enhancers and insulators. Enhancers can be located distal from their target gene promoters and are able to activate transcriptions in a cell-type specific manner. Chromatin looping is proposed to form via specific to connect the enhancer with its target promoter (28). Insulators are located between enhancers and their target promoters

(43,44). In the appropriate cell types, functional insulators can block the interactions of enhancers and promoters such that transcription will not be activated from the promoter. In this sense, insulators are classified as negative regulators. Another interesting feature of insulators is that they are not able to block enhancer-promoter interactions if they are located outside of the regions enclosed by enhancer-promoter pairs. Based on these features, different hypothetical models are proposed for insulator mechanisms, such as disturbances of local chromatin structures that block enhancer-promoter interactions, formations of three dimensional loops that partition enhancers and promoters into distinct domains and competing with promoters for the preferential interactions of enhancers (43,44).

As an important class of regulatory elements, insulators have been investigated for many years and there are several experimentally validated insulators in different species. In drosophila, a *gypsy* transposable element located between an enhancer and a promoter can be bound by proteins Su(Hw) and mdg4, and block the transcription activation caused by the enhancer (45,46). Another insulator, called 5'HS4 element, is found in the β-globin locus in the chicken genome (47-49). This insulator is the most investigated one in vertebrate species and its associated protein, CTCF, has been shown to be widely associated with insulator functions (50,51).

Among all the experimentally validated insulators, there is no unified mechanistic model to explain their functions. But for a subset of insulators, RNA polymerase III (Pol III) is shown to be a critical part of the system. In yeast, some tRNA genes (which are transcribed by Pol III) function as insulators and furthermore (52,53), many more locations that are bound by TFIIIC (a subunit of Pol III) can also function as insulators

(52,54). Similar results of tRNA-derived insulators are recently found in mouse (55). Most interestingly, a SINE B2 transposable element in the mouse genome, which evolved from tRNA, can be transcribed by Pol III and function as an insulator in a developmental stage dependent manner (56). Collecting these observations from different species together, Pol III machinery is highly probable to be related with insulator functions, at least for a subset of them. Actually, an evolutionary hypothesis has been proposed to suggest that some insulators originally evolved from Pol III promoters (57).

Chromatin boundary elements

Another related class of regulatory elements is called chromatin boundary elements or chromatin barriers. Chromatin boundary elements can block the spread of repressive chromatin domains, which are enriched with repressive histone modifications, and thus protect the active transcription within open chromatin domains (43). The ability of boundary elements to partition the chromatin into repressive domains (heterochromatin) and active domains (euchromatin) makes them to be related with insulators, because they can potentially block enhancer-promoter interactions by demarcating them into different domains (44,45,48). Actually boundary elements are sometimes classified as a subgroup of insulators due to this overlap of phenotypic outcomes, although the conceptual distinction is also clear.

Since chromatin boundary elements can change and restrict the landscape of large-scale chromatin domains, they can be viewed as higher-order regulators that function in long distance and influence groups of gene expression. The importance of boundary elements is further underscored by the fact that the linear configurations of active and repressive domains along chromosomes are also related with three dimensional chromatin structures (58). Some boundary elements have been shown to interact with each other to form hubs of three dimensional chromatin interactions (50,52,59-61).

Despite their importance, there is no clear unified model of the boundary element mechanisms and the number of experimentally validated barriers is limited. A straightforward algorithm based on one aspect of the observed features derived from a few known boundary element examples will just detect putative elements with the same features and unable to find barriers with different, or even novel, features. A carefully designed unbiased algorithm is needed to explore candidate boundary element locations with various features that are indicative of the underlying mechanisms.

Combinatorial chromatin signatures

At this time, the relationships between individual histone modifications and gene expression patterns have been studied a great deal and a general picture for the role of individual histone modifications has begun to emerge. Nevertheless, the much more complex relationships between different combinations of histone modifications with various biological activities and regulatory elements, such as transcriptional initiation, transcriptional termination, cell-type specific expression, enhancers and imprinting, have been under investigated. Although systematic analysis of this question is difficult, a hypothesis called the “histone code”, which proposes the specific relations between combinatorial histone modification signatures and different biological activities, has been raised based on some biochemical observations (21). Several canonical histone code has been found in the last few years (9,28,29,32,62). For instance, active promoters are associated with combinations of H3K4me3, H3K9me1 and several histone acetylations. Cell type specific enhancers are characterized by H3K4me1and H3K27ac.

As large-scale ChIP-seq datasets of various histone modifications in different cell types are accumulating, computational algorithms to search for distinct novel combinatorial histone modifications patterns and relating them with diverse genomic features will be valuable for the understanding of the regulatory functions of histone modifications. Furthermore, given the detailed locations of specific combinatorial chromatin signatures and their associations, comparative analysis among different cell types will reflect the dynamics of epigenetic regulation.

Overview of the dissertation

This dissertation focuses exclusively on the development and application of computational algorithm for ChIP-seq data analysis that are related with epigenomics. It contains both basic data processing methods and advanced data mining algorithms aimed at specific biological questions.

CHAPTER 2 presents a novel algorithm to accurately map ambiguous short ChIP-seq tags to reference genome sequences. Systematic performance comparisons with previous methods are reported for a set of simulated ChIP-seq data libraries. The utilities of this algorithm for the discoveries of biological signals within repetitive genomic regions are discussed.

CHAPTER 3 presents a method to identify broad peaks of diffuse ChIP-seq datasets. Besides the use of the maximal scoring segment algorithm, a detailed discussion on parameter estimations via Gibbs sampling on non-homogeneous Poisson processes is reported. Evaluations of the algorithmic performance on both real ChIP-seq datasets and simulated datasets are shown.

CHAPTER 4 presents a hypothesis-driven computational pipeline that predicts a specific subset of insulators: MIR-insulators. Both genomic and epigenomic features are integrated into this pipeline and a list of putative MIR-insulators are predicted. Several selected putative MIR-insulators are further experimentally validated. Functional annotations of genes proximal to those putative MIR-insulators, investigations on the local chromatin signatures, and the analysis of cell type specificity, are carried out for the predicted insulators.

CHAPTER 5 presents an unbiased algorithm to predict the locations of chromatin boundary elements in the human genome. The successful prediction of BEAD1 element is emphasized, and also the potential capabilities to discover elements with novel features are explored. The associations of the predicted boundary elements with CTCF binding and a set of chromatin features are analyzed. Some novel transcription factor binding motifs are shown to be enriched within those boundaries. A subset of boundaries containing non-coding RNAs genes are further analyzed for the binding of Pol III and the transcription states.

CHAPTER 6 presents a new unsupervised algorithm to search for recurrent combinatorial histone modification signatures. Applications of this algorithm resulted in a set of chromatin patterns and their relationships with diverse genomic features are systematically analyzed. The computational advantages of this algorithm are discussed in both algorithm descriptions and analyses of the resulted chromatin signatures.

CHAPTER 2

A GIBBS SAMPLING STRATEGY APPLIED TO THE MAPPING OF AMBIGUOUS SHORT SEQUENCE TAGS

Abstract

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) is widely used in biological research. ChIP-seq experiments yield many ambiguous tags that can be mapped with equal probability to multiple genomic sites. Such ambiguous tags are typically eliminated from consideration resulting in a potential loss of important biological information. We have developed a Gibbs sampling based algorithm for the genomic mapping of ambiguous sequence tags. Our algorithm relies on the local genomic tag context to guide the mapping of ambiguous tags. The Gibbs sampling procedure we use simultaneously maps ambiguous tags and updates the probabilities used to infer correct tag map positions. We show that our algorithm is able to correctly map more ambiguous tags than existing mapping methods. Our approach is also able to uncover mapped genomic sites from highly repetitive sequences that can not be detected based on unique tags alone, including transposable elements, segmental duplications and peri-centromeric regions. This mapping approach should prove to be useful for increasing biological knowledge on the too often neglected repetitive genomic regions.

Introduction

Genome-wide chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq) experiments are increasingly used in biological and medical research (3,19). ChIP-seq experiments produce a large amount of short sequence tags which need to be faithfully mapped back to the genome and processed to reveal biologically relevant signal. A number of algorithms have been recently developed to process ChIP-seq data (63). These include algorithms for genomic mapping of sequence tags (35,37), smoothing of ChIP-seq tag distribution signals (64) and detection of statistically significant tag peaks (65). One remaining challenge for the processing of ChIP-seq data is the mapping of ambiguous tags. Ambiguous tags are those that can be mapped to equally to multiple genomic sites, each of which has significant sequence similarity with the tag, and thus it is difficult to distinguish the real site from all the possible sites. Usually, researchers simply disregard ambiguous tags and only make use of uniquely mapped tags. This often results in a substantial loss of information and may bias conclusions based on the analysis of unique tags alone. This is particularly true for mammalian genomes, such as the human genome, which have numerous interspersed repeat sequences. Repeat sequences that are highly similar may produce a large amount of ambiguous tags, which if not mapped will be disregarded in subsequent analyses. Research has shown that interspersed repeat sequences provide a wide variety of functional elements to eukaryotic genomes (66). Therefore, disregarding ambiguous tags will cause an underestimate of the biological significance and functional roles of interspersed repeated DNA.

Two different approaches have been developed for the mapping of ambiguous sequence tags. The mapping software MAQ randomly selects a possible site and assigns it to the ambiguous tag (35). Each possible site has the same probability of being selected. In other words, there is no way to know if this approach yields a correct mapping of ambiguous tags. The second approach takes advantage of the local context of mapped tags to more accurately assign genomic locations for ambiguous tags. This approach rests on the assumption that real ambiguous tag sites are expected to have more sequence tags in the local vicinity, whereas the incorrect sites for the same ambiguous tags are expected to have fewer numbers of co-located tags (36,67). To apply this method for any ambiguous tag, the number of overlapping mapped tags at each of the possible ambiguous tag mapped positions are counted and used to assign fractional weights to each possible position. The ambiguous tag is then fractionally mapped to each possible position with the fractions weighted by the local mapped tag context. In other words, possible sites with more tags already mapped are deemed to deserve higher confidence and are accordingly assigned greater fractions of ambiguous tags. The fractional mapping method makes important contribution to the ambiguous tag mapping problem. But as the use of ChIP-seq in scientific research is increasing, it will be important to further refine the accuracy of mapping ambiguous tags. First, the fraction method is heuristic as the fractions assigned to the possible map sites are directly proportional to the number of tags mapped to each site. While this approach is consistent with biological intuition, it lacks statistical support. A more sensitive probabilistic method could be used to better represent and measure the confidence level of each possible site. Second, the fraction method deterministically fractionates the ambiguous tags without guarantee that the result is optimal. In other words, it doesn’t search the possible space of assignments of ambiguous tags and lacks information on the accuracy of the final results. Third, the fraction method is not realistic enough since it splits tags by assigning fractions of ambiguous tags to each possible site. In reality, each sequence tag is only derived from a single genomic site. Thus, fractioning sequence tags inevitably results in wasting signal on incorrect sites and weakening the signal level on real sites.

To address the outstanding issues with ambiguous tag mapping, we have developed a probabilistic Gibbs sampling based algorithm to map more ambiguous tags with greater accuracy. Our approach assigns ambiguous tags to single genomic sites, without fractionating tags, and iteratively samples within the space of the possible mappings of ambiguous tags. The Gibbs sampling strategy (68,69) guides the algorithm to achieve accurate unique mappings of ambiguous tags. The algorithm also provides statistical support for ambiguous tag mapping via the use of likelihood ratios that measure the confidence levels of possible genomic map sites. We evaluated the performance of our algorithm compared to existing approaches using sequence tag data from the highly repetitive human genome. We demonstrate that our probabilistic approach to mapping ambiguous tags yields superior results as measured by 1) the fraction of correctly mapped ambiguous tags, 2) the precision and recall of correctly recovered repetitive genomic sites and 3) the level of signal found at repetitive sites.

Methods

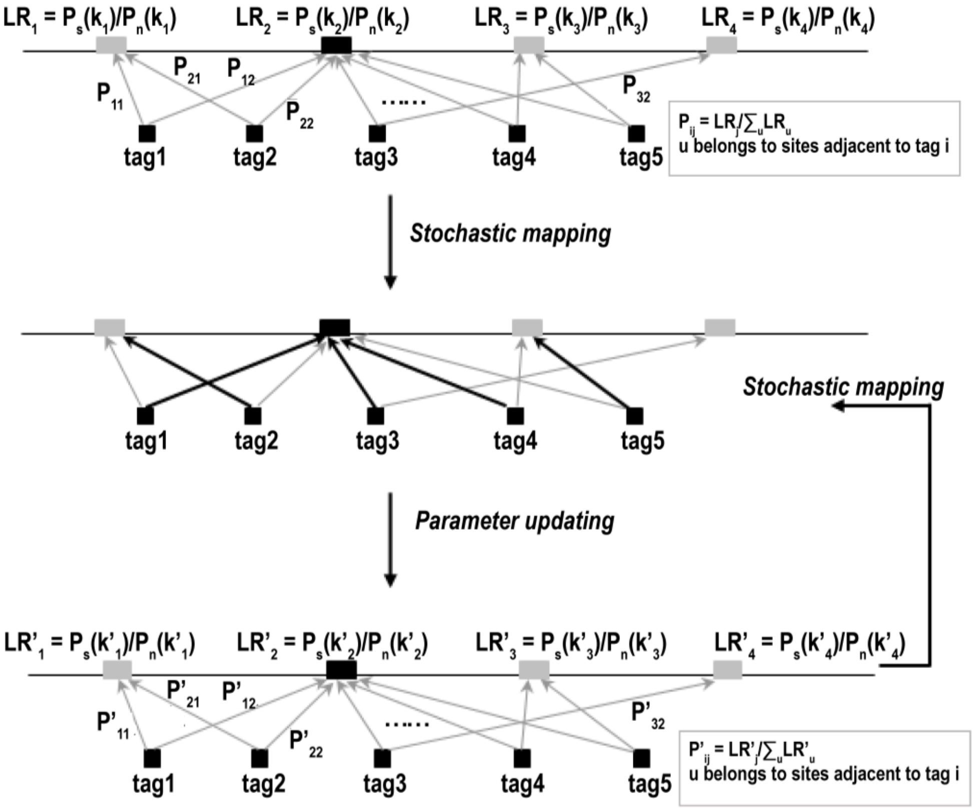
Overview of the algorithm

Our algorithm maps ambiguous tags to individual genomic sites by taking advantage of the local genomic context provided by co-located tags. For each possible map site of an ambiguous tag, the number of co-located tags are counted and used to calculate a normalized likelihood ratio that represents its probability of being the real map site. Map sites are randomly selected based on the underlying probability distributions from the likelihood ratios. Likelihood ratio scores then updated based on the new mapping, and this procedure iterates until convergence when there is little or no change in the map positions between iterations.

A Gibbs sampling strategy is used to iteratively map ambiguous tags to possible genomic sites while updating the probability that each tag is mapped to its most likely site. Gibbs sampling was chosen because it allows for a simultaneous updating of the map positions and the parameters for these positions. Through the updating iterations, the algorithm searches in the space of all possible mapping configurations, where each mapping configuration can be considered as a bipartite graph with edges connecting tags and sites (Figure 2.1). Intuitively, once an ambiguous tag is correctly mapped to the real site, it will guide the algorithm to map those tags derived from the same site to it with higher probability.

Problem formulation

For each ambiguous tag, there are multiple possible genomic sites to which it could be assigned. It is not possible to assign a specific site to an ambiguous tag with 100 percent confidence, and so we need to calculate the confidence for each probable site by some measurement and then select a reasonable site for each ambiguous tag based on those confidences. By “reasonable”, we mean a selection of sites that will minimize the number of incorrect mappings of ambiguous tags. Suppose there are T genomic sites associated with ambiguous tags and the set of ambiguous tags is $A = \{a_1, a_2.....a_N\}$, where a_i represents ambiguous tag i. We use $S_i = \{s_{i1}, s_{i2}.....s_{in_i}\}$ to denote the set of probable sites for a_i, where n_i is the total number of probable sites for a_i.



$LR_1 = P_s(k_1)/P_n(k_1)$
$LR_2 = P_s(k_2)/P_n(k_2)$
$LR_3 = P_s(k_3)/P_n(k_3)$
$LR_4 = P_s(k_4)/P_n(k_4)$
P_{11}
P_{21}
P_{12}
P_{22}
……
P_{32}
$P_{ij} = LR_j/\sum_u LR_u$
u belongs to sites adjacent to tag i
tag1
tag2
tag3
tag4
tag5
Stochastic mapping
tag1
tag2
tag3
tag4
tag5
Stochastic mapping
Parameter updating
$LR'_1 = P_s(k'_1)/P_n(k'_1)$
$LR'_2 = P_s(k'_2)/P_n(k'_2)$
$LR'_3 = P_s(k'_3)/P_n(k'_3)$
$LR'_4 = P_s(k'_4)/P_n(k'_4)$
P'_{11}
P'_{21}
P'_{12}
P'_{22}
……
P'_{32}
$P'_{ij} = LR'_j/\sum_u LR'_u$
u belongs to sites adjacent to tag i
tag1
tag2
tag3
tag4
tag5

Figure 2.1: Scheme of our Gibbs sampling algorithm. Possible tag map sites along with their likelihood ratios are shown prior to stochastic mapping. Gray boxes reprsent incorrect sites, and the black box represents the correct site. An arrow between a tag and a site means the tag could possibly be mapped to that site. One iterative cycle of joint stochastic mapping and parameter updating is shown. The black arrows point to selected sites for each tag after stochastic mapping.

There are two aspects of this problem. One is the measurement of confidence for each probable site, and the other one is the algorithm used to select reasonable sites for ambiguous tags. An applicable measurement of confidences of probable sites needs to be monotonic with the number of tags that are mapped to each specific site and should reflect both the information of the distribution of tag numbers of real sites and the information of the distribution of tag numbers of background. We use likelihood ratio as the confidence measurement based on both intuitive clues and theoretical analysis.

Intuitively, likelihood ratio is monotonic with tag counts and is also computationally tractable. Furthermore, it takes both the background distribution of tag counts and the estimated target distribution under consideration. Higher likelihood ratios correspond to higher confidences and increase non-linearly with tag counts. Likelihood ratios will increase sharply for large tag counts and be relatively low for sites with few tags. This property will help to avoid the problem of wasting fractions of mapped tags on sites that contain few tags; a problem that could be particularly vexing if many such low confidence sites exist for a single ambiguous tag. The likelihood ratio for s_{ij} is denoted as

$$\mathrm{LR}_{\mathrm{j}} = \frac{\mathrm{P}_{\mathrm{s}}(\mathrm{k}_{\mathrm{j}})}{\mathrm{P}_{\mathrm{n}}(\mathrm{k}_{\mathrm{j}})}.$$

P_s is the estimated target distribution of tag counts in real sites and P_n is the background distribution of tag counts. k_j is the tag count at site j . The details of these two distributions will be discussed in the next section. Given the calculated likelihood ratios, it is possible for us to reasonably map ambiguous tags.

Furthermore, from a theoretical point of view, normalized likelihood ratio is the measurement we will automatically derive from the calculation of the conditional probability of assigning ambiguous tags to a specific site given the assignments of all the other tags. We use D to denote the original data, which essentially represent the associations of tags with possible sites, and M to denote the whole assignment of tags to sites. $M_{[-i]}$ represents the assignments of tags to sites, except the assignment of tag i . $P(a_i \sim s_{ij} \mid M_{[-i]}, D)$ represents the conditional probability of assigning tag i to the j th probable site of i , given the original data and the assignment of all tags except tag i . We use U to represent the whole set of sites.

Below we show that this conditional probability is equal to the normalized likelihood ratio, as derived from Bayes rules:

$$P(a_i \sim s_{ij} \mid M_{[-i]}, D) = \frac{P(a_i \sim s_{ij}, M_{[-i]} \mid D)}{P(M_{[-i]} \mid D)} =$$

$$\frac{\{P_s(k_j+1)\prod_{m \in S_i \setminus j} P_n(k_m)\} \times P(U \setminus S_i)}{\sum_{\tau \in S_i} \{P_s(k_\tau+1)\prod_{m \in S_i \setminus \tau} P_n(k_m)\} \times P(U \setminus S_i)} = \frac{\left(\dfrac{P_s(k_j+1)}{P_n(k_j)}\right)}{\sum_{\tau \in S_i} \dfrac{P_s(k_\tau+1)}{P_n(k_\tau)}}$$

So the normalized likelihood ratio represents the conditional probability for the j th probable site given the assignment of other tags. Equivalently, this conditional probability serves as our predictive update formula for the Gibbs sampling procedure described below.

In order to calculate likelihood ratios for genomic sites, we need to first map those ambiguous tags to get the number of tags mapped to each specific site. In other words, mapping of ambiguous tags and calculating the likelihood ratios for each site are circular. This circularity led us to adopt Gibbs sampling strategy, which is a stochastic version of EM algorithms, to select reasonable sites for ambiguous tags. To do this, we first initialize the likelihood ratios for genomic sites using the total number of tags that can be probably mapped. Then we map each ambiguous tag to a specific site based on the initial likelihood ratios. To be more specific, we stochastically map each ambiguous tag to a genomic site with the probability equal to the normalized likelihood ratio of the site. Then we update the likelihood ratios given the current mapping of ambiguous tags. We continue the update on the mapping and the calculation of likelihood ratios until there is no significant change. Through the iterative updates (stochastic mapping and parameter updating), the overall likelihood ratios are expected to be optimized, and so we achieve an accurate mapping of ambiguous tags. Since the complete normalized likelihood ratio for a configuration of mapping is proportional to $\prod_{i\in U}(\frac{P_s(k_i)}{P_n(k_i)})$, where i is the index of genomic sites with tags mapped, we can rewrite this formula based on tag counts and obtain the formula as $\prod_{\tau\in\sigma}\left(\frac{P_s(\tau)}{P_n(\tau)}\right)^{n(\tau)}$, where $n(\tau)$ represents the number of sites with τ tags mapped. Here, σ represents the set of tag counts for all sites. For instance, if σ consists of large numbers, it means that most sites are mapped with large number of tags and the mapping is a reasonable one. Otherwise, most sites are mapped with a small number of tags and the set of tags are scattered into diverse sites. Taking the logarithm of this formula and dividing by Z, the total number of tags, we get $\sum_{\tau\in\sigma}\left(\frac{n(\tau)}{Z}\right)\log\left(\frac{P_s(\tau)}{P_n(\tau)}\right)$. When Z is sufficiently large, it approaches the relative entropy between P_s and P_n on the subset of σ. So essentially, the Gibbs sampling procedure described above searches a certain subset σ to maximize the relative entropy. When σ consists of only large numbers, the relative entropy is larger. This analysis further demonstrates that our algorithmic design is reasonable. The equation above shows that by using normalized likelihood ratios, our objective function is equivalent to the relative entropy.

In theory, Gibbs sampling will have good performance given a sufficient number of iterations. Thus, there may be concerns about the time necessary for the algorithm to converge. However, since unique tags count for the majority of the whole set of tags, and these help to guide the mapping of ambiguous tags, this has the effect of shortening the algorithm time significantly. In our experience, about 5 iterations are sufficient for convergence.

Algorithm

Next we describe each step of the algorithm in detail along with the definitions of necessary concepts. The scheme of the method is shown in Figure 2.1.

Phase 1. Initialization

Step 0. The program Bowtie (37) is used to map all sequence tags to the genome and only genomic loci with significant sequence similarities are used for the following steps. Sequence tags are classified into unique tags and ambiguous tags by the Bowtie mapping algorithm.

Step 1. To calculate the likelihood ratios, we need to model the distributions of tag counts for real modified sites (P_s) and for background (P_n). For real modified sites, we use the Normal distribution to approximate the real distribution of tag number

$P_s \sim N(\mu, \sigma^2)$.

To identify genomic sites that are most likely to actually be modified (i.e. real modified sites), we use sites with large numbers of mapped unique tags. We then use the numbers of unique tags associated with those sites to calculate the average tag count and standard deviation for each site genome-wide. Note that the average tag count calculated here is corrected by a factor which takes into consideration that the real average tag count will be greater once ambiguous tags are included. For background, we use the Poisson distribution to approximate the background distribution of tag counts: $P_n \sim Poisson(\lambda)$.

The Poisson distribution is an appropriate model for counting processes that produce rare random events and thus can be applied here to describe the background tag count distribution. We count the total number of tags (both unique and ambiguous tags) and calculate the average tag number for each site. The average tag number serves as the parameter (λ) of Poisson distribution. After getting all the parameters, we calculate the likelihood ratios for various tag counts: $LR(k) = \frac{P_s(k)}{P_n(k)}$, and get a table of likelihood ratios which will be used in subsequent steps.

Step 2. In order to obtain the initial settings of likelihood ratios for all the probable genomic loci, we use the number of tags of each site (both unique and ambiguous tags) to calculate the likelihood ratios. Since the ambiguous tags have not been assigned to a specific genomic site, here we assign each ambiguous tag to all the probable sites to initialize the likelihood ratios. The calculation of likelihood ratios for various tag numbers has already been done in Step 1 and the algorithm only needs to search the table of likelihood ratios. A special notion here is that we introduce the information content factor ($0 < f < 1$) of ambiguous tags compared to unique tags. Since the nature of uncertainty of ambiguous tags, the information content of ambiguous tags is smaller than unique tags. Thus, the effective number of ambiguous tags (k_e) is corrected by f and the number of tags used to calculate likelihood ratio is: $k = k_u + k_e = k_u + k_a f$, where k_u is the number of unique tags and k_a is the number of ambiguous tags. f can be set by the user based on their confidence of ambiguous tags and provide flexibility of the method. The suggested value of f is the inverse of the mean number of associated sites of ambiguous tags. If the mean number of associated sites of ambiguous tags is larger, then f should be made smaller to weight unique tags more heavily for the mapping.

Phase 2. Iterative weighted mapping

Step 3. Given the likelihood ratio (LR_j) of probable site j ($j = 1,2...n_i$) for ambiguous tag a_i , the algorithm stochastically selects a probable site and assigns it as the site of the corresponding ambiguous tag. The probability (P_{ij}) of probable site j to be selected for a_i is proportional to the likelihood ratio of site j : $P_{ij} = \frac{LR_j}{\sum_{k \in S_i} LR_k}$, where $k = 1,2...n_i$. Thus, probable sites with higher likelihood ratios will have a greater chance of being assigned.

Step 4. Based on the current assignments of sites for ambiguous tags obtained from Step 3, the likelihood ratios of all the probable sites are updated. The new likelihood ratio of each probable site is obtained accordingly to the current number of tags assigned to the site.

Step 5. Iterate through Step 3 and Step 4 until no significant changes occur, i.e. until convergence. For a given threshold, if the number of reassignments of ambiguous tags is smaller than the threshold, then the iterations will stop and output the final mapping of tags.

<u>Results</u>

Sequence tag data sets

In order to test the performance of our algorithm, we randomly selected ~50,000 sites of the human genome as a benchmark. Each site is 147bp in length (i.e. mono-nucleosomal) and the set of sites contains transposable elements and simple repeats in the same fractions as the human genome. Then we generate short sequence tags from these sites under a range of set of parameters. These parameters include sequence tag length

(L), signal-to-noise ratio (SNR) and sequencing error level (SE). In theory, shorter sequence tags are expected to have more ambiguous tags. To test the performance of our algorithm on different sequence tag lengths, we generate libraries with 20bp tags and libraries with 35bp tags. SNR corresponds to the specificity of the ChIP experiments. Noise here means the fraction of sequence tags derived from sites which are not the real modified sites. In experiments with high specificity, the majority of sequence tags are derived from the real modified sites, while in experiments with high level of noise, there are increased number of sequence tags derived from other sites. And we define the SNR as the ratio of the probability that a sequence tag is derived from the real modified sites over the probability that a sequence tag is derived from other sites. To test our algorithm's performance under different SNRs, we generate libraries with SNR set as 99 (corresponds to 99% tags derived from real modified sites) and libraries with SNR set as 9 (corresponds to 90% tags derived from real modified sites). The sequencing error level corresponds to the probability of errors in high-throughput sequencing. We generate libraries with sequencing error levels as $2/(5L)$ and $4/(5L)$. The reason to set SE this way is as follows. We assume that the sequencing errors on different sites are independent from each other. This is not completely true in reality but is acceptable as a first-order approximation. Then the total number of errors for each sequence tag with length L would follow binomial distribution. So under $SE = 2/(5L)$, the fraction of sequence tags without errors is about 60% and under $SE = 4/(5L)$, the fraction is about 50%. It means that the quality of the simulated sequencing is not very good. Under such conditions, some sequence tags might be mis-mapped or become ambiguous tags. The purpose of this setting is to make sure that our algorithm test results are conservative.

Since each of these three parameters only has two optional values, there are 8 combinations of different values of those parameters and so we generate one sequence tag library for each combination of the parameter values. The parameters for each library are listed in Table A.1.

We also used a second larger benchmark set consisting of 173,877 sites of the human genome. These sites were obtained from a ChIP-seq study of histone modifications based on ABI SOLiD sequencing platform (unpublished data) that only used unique sequence tags, and each site has significant number of tags. This dataset was used because it mimics conditions one would expect for real sites: a larger number of total sites and a realistic distribution of sites along the human genome. In order to test our algorithm, we generated sequence tags for these sites the same way as described above under one set of parameters (Table A.1).

After preparing sequence tags, we ran the program Bowtie (37) to map the sequence tags to the human genome. The fractions of ambiguous tags in the 9 libraries range from 9.7% to 37.6%. The fraction of sites undetected using unique tags alone are influenced by the tag threshold used. Higher threshold cause more undetected sites. For the lowest threshold (4 tags) used in our analyses, the fractions of undetected sites range from 16.4% to 28.4%. These values underscore the importance of accurately mapping ambiguous tags to recover undetected sites.

Fraction of correctly mapped ambiguous tags

The first and most direct measurement of the algorithm performance is the fraction of correctly mapped ambiguous tags. Since the fraction method does not assign the ambiguous tags to a specific site, this measurement is not applicable. So we compared our algorithm against the MAQ software method, which randomly selects a site for each ambiguous tag. The comparison on the 8 sequence tag libraries shows that our algorithm correctly maps 49% to 71% of ambiguous tags, while the MAQ method correctly maps 8% to 23% of ambiguous tags (Figure 2.2). Over all eight sequence tag libraries evaluated, our algorithm maps 38% to 51% more tags than MAQ. In the best case, our algorithm maps the majority of ambiguous tags (71%) and only a small fraction of information is lost.

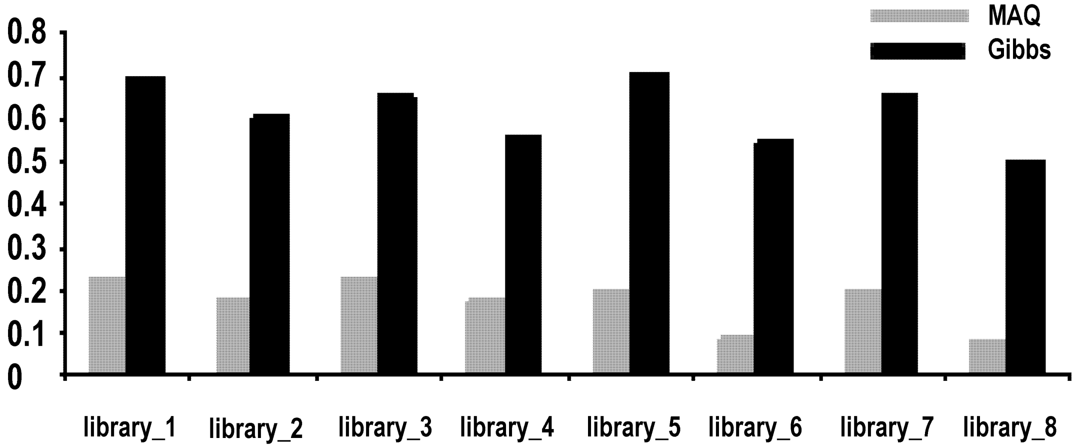


Figure 2.2: Fractions of correctly mapped ambiguous tags for each library. Library descriptions are given in Table A.1. Gray bars show results based on MAQ, and black bars show results based on our Gibbs sampling algorithm.

Comparison of rescued sites

The other measurement of the algorithm's performance is the numbers and fractions of correctly 'rescued' genomic sites, which can not be observed by unique tags alone. An important issue regarding the rescued sites is the tag number threshold, above which a site is called rescued with a certain number of tags (Figure 2.3.A). Different thresholds will result in different sets of true positives, false positives and false negatives. Since there are various methods to decide the threshold and different users usually set different thresholds, we tested our algorithm's performance on a set of three different thresholds (4 tags, 6 tags and 8 tags). Together with the previously described the 9 sequence tag libraries we use, this results in a set of 27 conditions for analysis.

The first thing we did was to compare the numbers of genomic sites identified using unique tags alone to the numbers of genomic sites identified by including ambiguous tags with our method (Table A.2). Over the 27 conditions, the inclusion of ambiguous tags yields an average increase of 11.46% in the fraction of genomic sites accurately identified. The use of ambiguous tags resulted in the identification of 2,602-51,508 sites missed with unique tags alone.

Next we compared our method for including ambiguous tags to the MAQ and fraction methods. To do this, after excluding sites that can be found by unique tags alone, we divide the set of sites rescued by ambiguous tags into two subsets by comparing the set with the benchmark. The correctly rescued sites are true positives (TP) and other sites are false positives (FP). The sites in the benchmark which remain undiscovered are false negatives (FN) (Figure 2.3.B). In order to test the performances, we employ recall $RE = TP/(TP + FN)$ and precision $PE = TP/(TP + FP)$ as measurements.

For the four libraries with 35bp tags and the four libraries with 20bp tags, our algorithm shows the highest recall over all conditions (6 tag thresholds shown in Figure 2.3.C, 4 and 8 tag thresholds shown in Figure A.1 and numbers of sites shown in Table A.3). Our algorithm also has the highest precision for these libraries over 14 of the 24 conditions evaluated (Figure 2.3.C, Figure A.1). For the 10 cases where our algorithm did not show the highest precision, the difference from the fractional method was marginal (Table A.3). In general, when recall increases precision may be expected to decrease. The simultaneous increase in both recall and precision in 14 cases evaluated here supports the improved performance of our algorithm. To more quantitatively evaluate the improvement in the performance of our algorithm for both recall and precision together, we used the harmonic mean (F) of the recall and precision values for each condition (i.e. each library and threshold combination). The F-values are higher for our algorithm over all conditions, indicating an improvement in performance when recall and precision considered together (Table A.4). Similar results can be seen when the larger tag library is evaluated with our algorithm over the three thresholds. Recall improves substantially in all cases, and precision decreases marginally for thresholds 6 and 8 (Figure 2.3.D and Table A.5). The F-values showing the combined recall and precision performance are higher for our method over all three thresholds (Table A.4).

In Figure 2.4, we provide two examples of our mapping results with the comparison against the benchmark and the result of fraction method. It can be seen that our algorithm rescues more sites than fraction method, and that the average number of tags at rescued sites is higher than seen for the fraction method. This can be attributed to the fact that the fraction method assigns a fraction of ambiguous tags on each site and wastes information on other sites. The greater number of tags per rescued site can help to ensure that these sites are robust to different user thresholds that are employed to distinguish signal from noise.

It should be noted that the two examples shown here represent segmental duplications (Figure 2.4.A) and satellite regions (Figure 2.4.B) respectively. It is expected that such highly repetitive regions will produce many ambiguous tags and thus would be difficult to uncover with ChIP-seq. However, our method achieves good performance in such repetitive regions. Furthermore, the second example is located very near to the centromere of chromosome 7. Centromeric regions are important in various cellular processes, such as cell division, and correct mapping of ambiguous tags to centromeric regions could help to uncover specific biological roles for such regions.

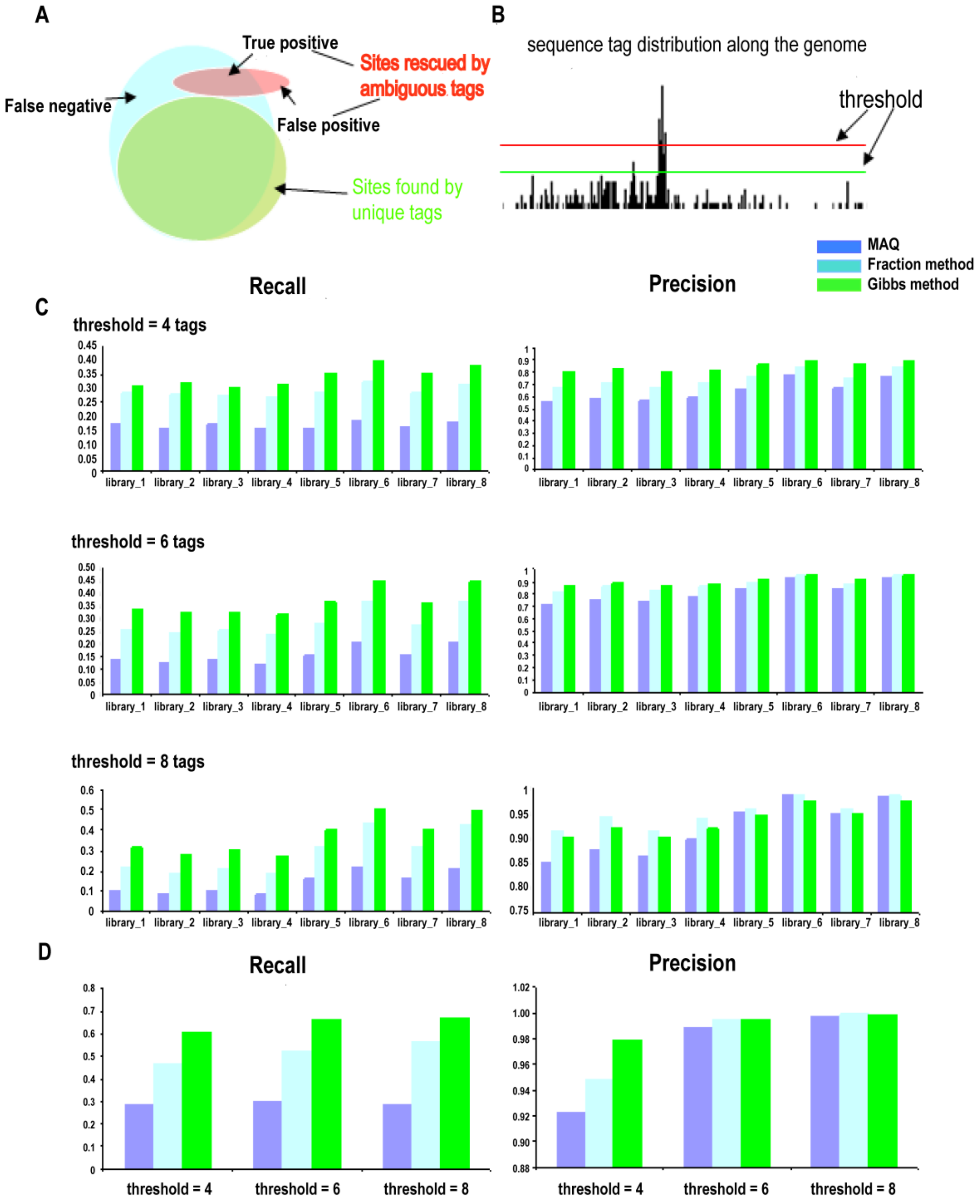


Figure 2.3: Comparison of algorithm performances. A. Illustration of data used to test algorithm performances. B. Variant tag thresholds could cause differences in the performance test. The lines (red and green) are two tag thresholds. C. Barplots of recall and precision for the three methods (MAQ-dark blue, fraction method-light blue, Gibbs method-green) on 8 libraries under 3 different tag thresholds. D. Barplots of recall and precision for the three methods (MAQ-dark blue, fraction method-light blue, Gibbs method-green) on the bigger library under 3 tag thresholds.

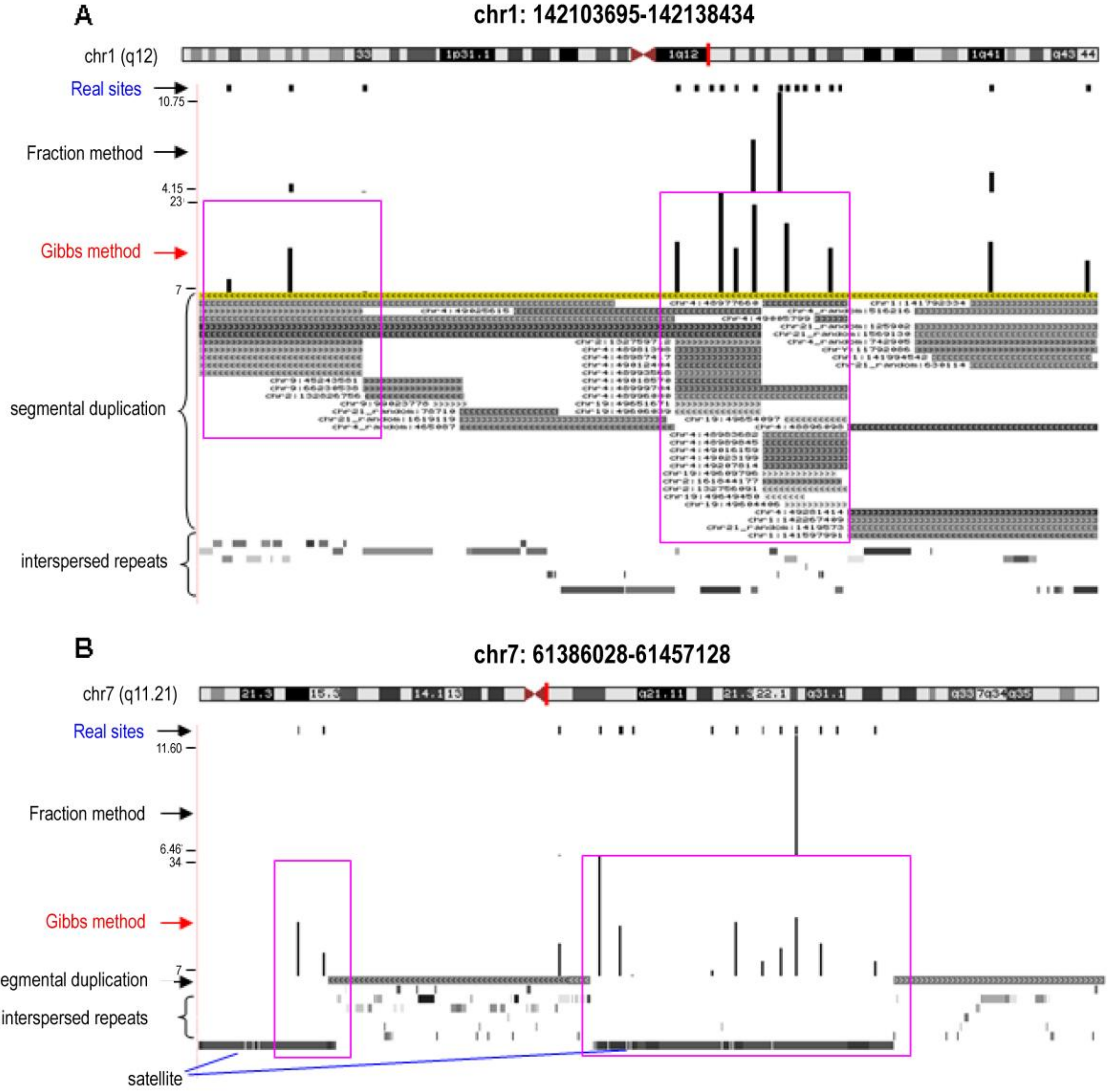


Figure 2.4: Examples of ambiguous tag mapping results. Tracks are shown through UCSC Genome Browser. The track of real sites shows the sites in the benchmark libraries. The track of Fraction method shows the mapping result by fraction method and the track of Gibbs method shows the mapping result by our Gibbs method. The heights of data represent the number of tags mapped to those sites. The tracks of repetitive genomic regions (segmental duplications, interspersed repeats and simple repeats) are also shown.

Biological relevance

Transposable elements, simple repeats, micro-satellites, segmental duplications and pericentromeric regions are genomic regions rich in repeat sequences. These regions could produce large numbers of ambiguous tags and will be difficult to uncover due to the technical problem of mapping ambiguous tags. The ability to correctly map ambiguous tags may facilitate novel discoveries regarding the biological significance of such repeat regions, many of which have been ignored in past chromatin immunoprecipitation studies. For instance, we show that our method is able to detect previously uncharacterized segmental duplications and satellite regions in Figure 2.4. In addition, our method uncovered a previously undetected modified histone site in the proximal promoter region of the CWF19-like 1 cell cycle control protein.

To further investigate whether our algorithm really helps us to find more sites in genomic repeats, we used the UCSC genome browser (70,71) to count the numbers and fractions of rescued sites in those regions and compared them against using unique tags alone (Figure 2.5). This analysis demonstrates that our algorithm is able to rescue substantial numbers of sites in genomic repeat regions, especially for segmental duplications and pericentromeric regions. Unique tags can only uncover around half of the sites in segmental duplications and pericentromeric regions, while our algorithm could uncover the majority of those sites (Figure 2.5.B). It is evident that our method has the potential to generate additional biological knowledge from ChIP-seq experiments.

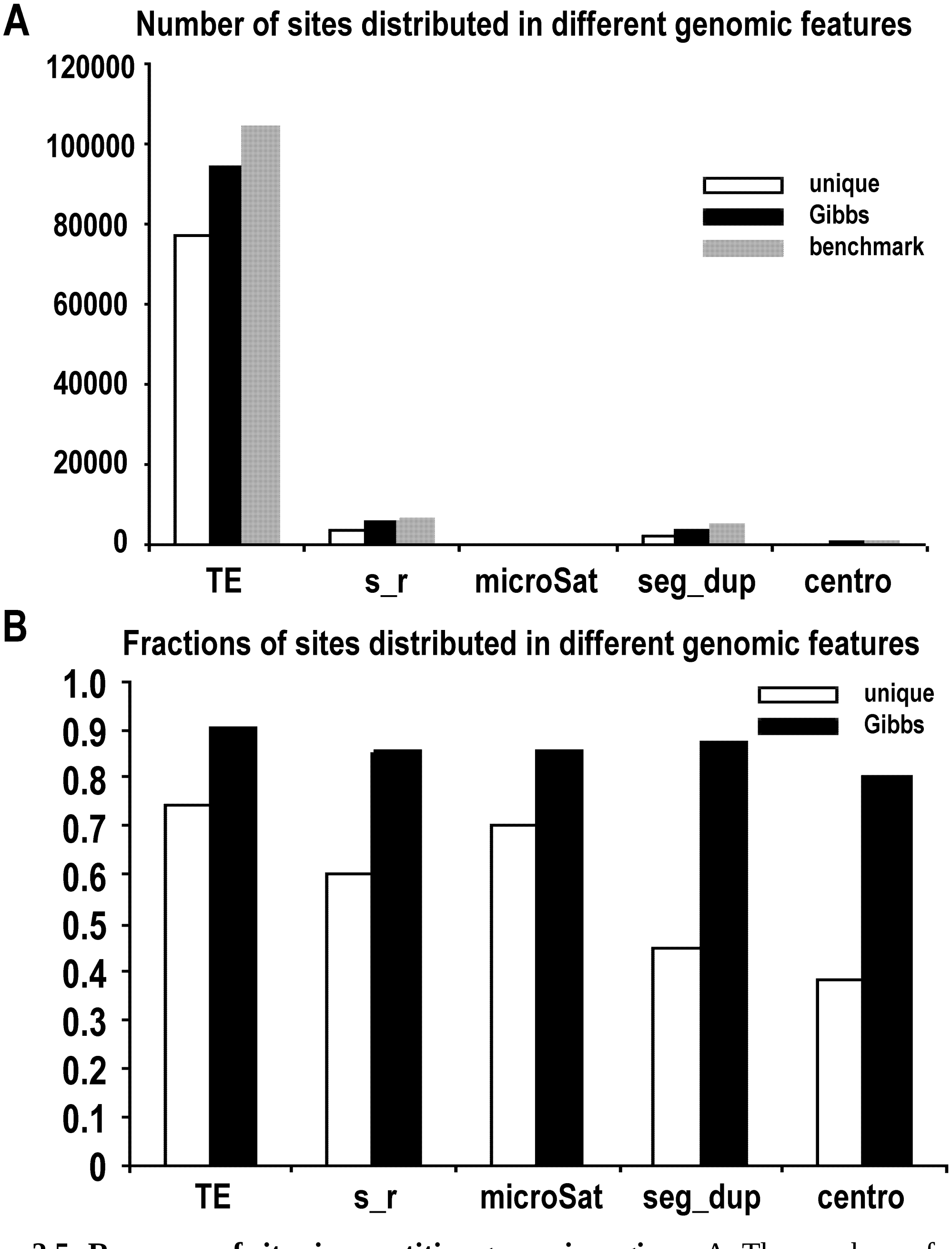


Figure 2.5: Recovery of sites in repetitive genomic regions. A. The numbers of correctly discovered sites in various genomic features by unique tags alone (white) and our Gibbs method (black) compared with the corresponding numbers in the benchmark library. B. The fractions of correctly discovered sites in various genomic features by unique tag alone (white) and our Gibbs method (black). [TE: transposable elements; s_r: simple repeats; microSat: microsatellites; seg_dup: segmental duplications; centro: peri-centromeric regions]

Discussion

Based on the results described above, we have shown that our algorithm significantly improves the accuracy of mapping ambiguous tags. The essential information used by the algorithm is the association between co-located sequence tags, which was originally utilized by Faulkner and co-workers (36) in the fraction method. Our contribution to this class of approach is to employ iterative probabilistic methods to achieve better performance. The use of likelihood ratios not only reflects the information on sequence tag associations but also the background distribution information. Furthermore, likelihood ratios are not linear to tag counts, but increase sharply for large tag counts and thus efficiently avoid wasting signal on sites with small tag counts. The Gibbs sampling procedure enables us to sample in the space of mapping and achieve a reasonable assignment of sites to sequence tags. For most experiments, unique tags are the majority of tags and they can guide the sampling efficiently. Thus, Gibbs sampling doesn't require too much time to reach the final result. We have also shown that correct mapping of ambiguous tags can facilitate our understanding of biology by recovering repeated genomic sites which are prone to produce ambiguous tags.

Although the length of sequence tags is increasing, there will still be a certain amount of ambiguous tags. As shown in Figure 2.4, genomic sites, such as segmental duplications and microsatellites will always produce ambiguous tags by their nature: with multiple copies in the genome. So the task of mapping ambiguous tags will not disappear due to the experimental technique advancements in short term, and our algorithm provides an efficient way to solve this problem.

Acknowledgements

Jianrong Wang was supported by the School of Biology at the Georgia Institute of Technology. I. King Jordan and graduate student Ahsan Huda were supported by an Alfred P Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839). Victoria V. Lunyak was supported by the Buck Trust Foundation.

CHAPTER 3

BROAD PEAK IDENTIFICATIONS FOR DIFFUSE CHIP-SEQ DATASETS

Abstract

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP--seq) has been widely used to characterize the genomic distributions of a variety of functional features, especially histone modifications. While some histone modifications show abrupt enrichment peaks at narrow and specific genomic locations, others have diffuse distributions along chromosomes and their large contiguous enrichment landscapes are better modeled as broad peaks. Here we present BroadPeak, a broad peak calling algorithm for diffuse ChIP-seq datasets. BroadPeak is able to find peaks of very different sizes, and its utility is expected to be helpful to the analysis of chromatin states and related biological questions.

Introduction

Histone modification landscapes are highly related with cell differentiations and the analyses of genomic distribution profiles of different histone modifications can help to understand the complex regulatory mechanism of the cell (1,2). ChIP-seq technology has been used to produce genome-wide maps of a suite of histone modifications in a number of cell types (3,7,18,25). To retrieve information from those ChIP-seq datasets, one of the critical data processing steps is peak calling, *i.e.* identifying the contiguous genomic regions that are significantly enriched with ChIP-seq tags compared with the genomic tag distribution as background (19). Some histone modifications, as well as transcription factors, are usually located to specific regions (*e.g.* promoters) and thus their ChIP-seq peaks are narrow and sharp. Computational methods have been developed to identify such peaks and their applications are very successful (38,39,42). But some histone modification's distribution profiles spread out along large contiguous genomic regions, *e.g.* chromatin domains, and accordingly their enriched peaks do not have a characteristic size range. For such diffuse datasets, the large regions with enriched ChIP-seq tags are modeled as broad peaks. Compared to the sharp peaks, which is featured by high tag counts of closely adjacent sites, the most important descriptive feature of broad peaks is that the spatial densities of sites with high tag counts (high-tag sites) within broad peaks are significantly higher than the genomic background. Gaps (*i.e.* low-tag sites) are allowed within broad peaks, and the broad peak sizes can grow to include more high-tag sites as far as the spatial densities are significantly high. There have been several algorithms designed to solve the broad peak calling problem (40,41). Compared with those methods, BroadPeak require fewer parameters and the number of gaps are adaptively determined from the data. Applications of BroadPeak on both real and simulated datasets support its good performance for broad peak calling.

Methods

Algorithm overview

The basic idea of BroadPeak is to assign appropriate positive scores to high-tag sites and negative scores to low-tag sites (gaps), and model the broad peaks as segments with maximal cumulative scores (maximal scoring segments) along chromosomes. Considering the cumulative score curve as a random walk along the chromosome, maximal scoring segments represent contiguous regions with significantly higher spatial densities of high-tag sites which, by our definition, are broad peaks. The input file for BroadPeak is the sorted tag-count profiles along chromosomes in bedGraph format. The output file is the list of broad peak locations in BED format (Figure 3.1.A).

Problem formulation

In the input file, the genome under consideration has been divided into small non-overlapping genomic bins with equal sizes (*e.g.* 200bp) and each bin is assigned with a tag-count. The bins are first classified into high-tag and low-tag bins based on a tag-count threshold derived from the standard tag-count *Poisson* distribution (which is parameterized by the genomic average bin tag-count λ). The regional spatial distributions of high-tag bins along chromosomes are expected to be non-homogeneous if broad peaks exist. Each high-tag bin is then assigned with a positive score s_1, and each low-tag bin is assigned with a negative score s_2. The cumulative score from bin *i* to bin *j* is $c_{ij} = \sum_{k=i}^{j} s_k$.

Maximal scoring segments are segments with maximal cumulative scores, *i.e.* the cumulative scores will decrease if the segments extend to longer segments or shrink to shorter segments. Thus, identifications of maximal scoring segments are equivalent to setting the boundaries of broad peaks with regional highest spatial densities of high-tag bins.

Scoring scheme and parameter estimations

The scores (s_1 and s_2) need to be carefully designed in order to obtain reasonable peaks. Based on the theorems proved by Karlin and Altschul (72), the optimal scoring scheme is log likelihood ratios: $s_1=ln(p/q)$ and $s_2=ln((1-p)/(1-q))$, where p is the estimated spatial density of high-tag bins in real broad peaks and q is the genomic background spatial density. Thus p and q are the only parameters needed for BroadPeak. One important feature of this scoring scheme is that, as the segment lengths are large, the spatial densities within the resulted maximal scoring segments will approximate the real target density p (72). This feature theoretically supports the validity of the final identified broad peaks since their compositions of high-tag bins will resemble real peaks and it also suggests that the gaps will be adaptively allowed based on the data, namely the target and background densities.

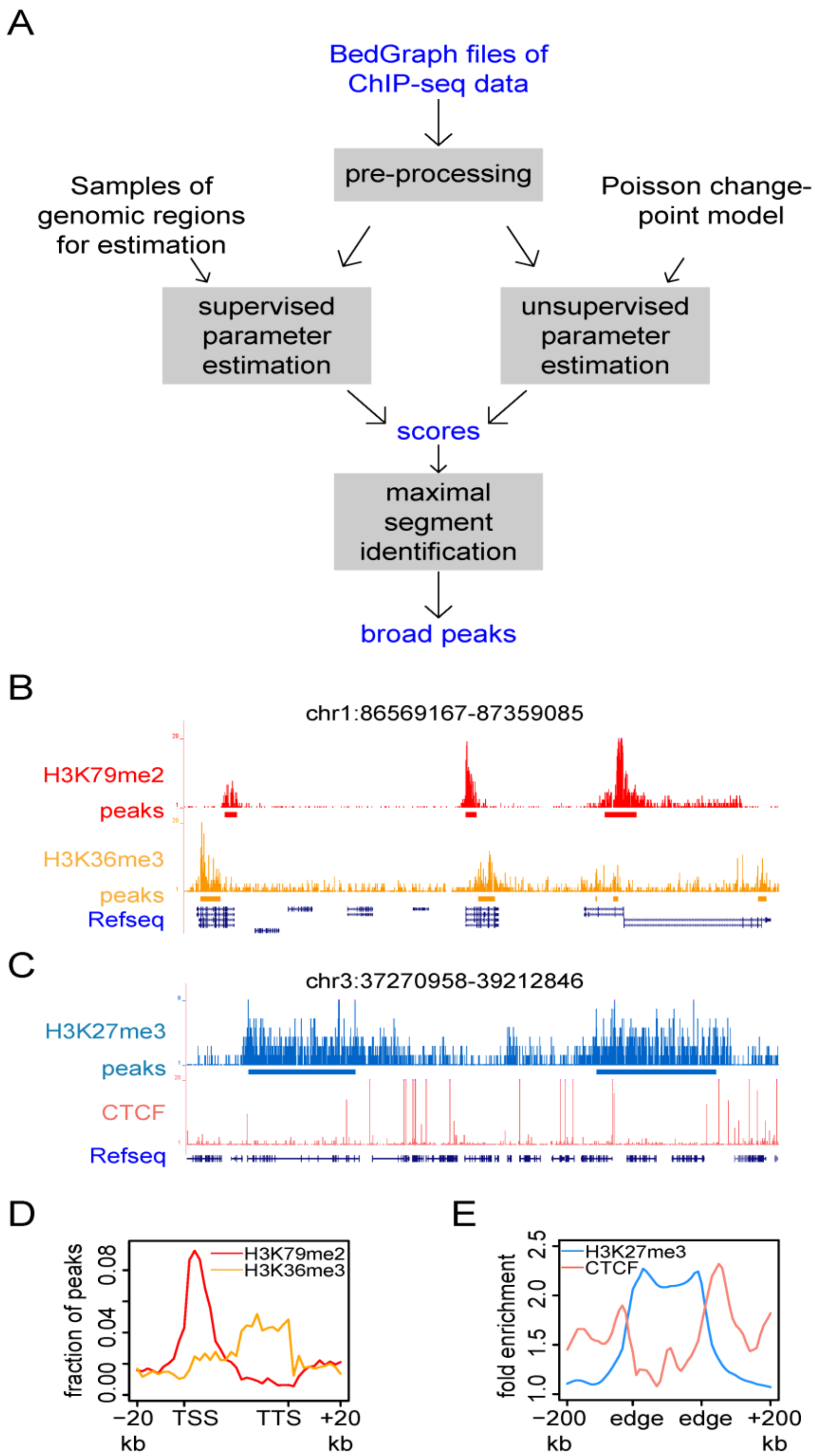


Figure 3.1: Scheme and evaluation of BroadPeak. (A) The algorithmic scheme and of BroadPeak. (B) Examples of broad peaks of H3K79me2 and H3K36me3. (C) Examples of broad peaks of H3K27me3. (D) Preferential distributions of broad peaks of H3K79me2 and H3K36me3. (E) Enrichments of CTCF bindings around the edges of large H3K27me3 broad peaks (>200kb).

In order to accurately estimate the target density *p*, BroadPeak provides two options: supervised and unsupervised estimations (Figure 3.1.A). For supervised estimation, user need to provide a list of regions that are enriched with broad peaks based on *a priori* knowledge (*e.g.* highly transcribed genes can be used for H3K36me3 parameter estimation). For unsupervised estimation, BroadPeak first uses a sliding window approach to obtain an initial set of regions showing spatial density changes and model the occurrence of high-tag bins in those regions as *non-homogeneous Poisson* processes with change-points. Conjugate *gamma* prior distributions are built and a *Gibbs sampling* algorithm is applied to estimate *p* and *q*. BroadPeak first uses a sliding window approach to scan the genome to sample a list of genomic regions that contain *change-points* of spatial densities, *i.e.* the spatial densities change, at one unknown location within the region, from background densities to significantly high densities that are only observed in broad peaks. These regions can be used to simultaneously estimate the target density *p* and background density *q*. Due to the resolution problem of sliding window approaches and the noisy fluctuations of ChIP-seq data, we also need to accurately predict the position of the *change-point*, in order to accurately estimate *p* and *q*. It leads us to adopt the *Gibbs sampling* method to iteratively estimate the location of *change-points*, *p* and *q*.

We used a 10kb sliding window (each step is bin-size) to scan the genome and calculated the high-tag bin densities for each sliding window. If the high-tag bin density is higher than twice of the genomic background density, the corresponding window is assigned as a putative region containing broad peaks or part of broad peaks. If we observe a large number of consecutive sliding windows with background densities followed by a large number of consecutive sliding windows with putative broad peak densities, then the whole region will be used later as a sample for parameter estimation.

Assume we get *N* such *change-point* containing regions after the sliding window scan as described above and each region contain *L* genomic bins, we will divide the *L* bins (for each sample region) into *n* super-bins and each super-bin is consisted of *m* consecutive bins. Finally, we obtain *N* data series with length *n* and they are denoted as: $D_i = (d_{i1}, d_{i2}, \ldots d_{in})$, where d_{ij} corresponds to the number of high-tag bins in the *j*th super-bin. Due to the way they are sampled, for each data series D_i, there exists a super-bin *k* such that $d_{ij} \sim Poisson(\lambda_1)$ for $j \le k$ and $d_{ij} \sim Poisson(\lambda_2)$ for $j > k$. So *k* is the unknown *change-point* and λ_1 is the rate for background spatial density of high-tag bins and λ_2 is the rate for target spatial density of high-tag bins ($\lambda_1 < \lambda_2$).

The whole data series is thus modeled as a *non-homogeneous Poisson process* with two distinct rates. *Gibbs sampling* has been previously used for parameter estimations of *non-homogeneous Poisson processes* and here we applied this strategy. We assume λ_1 and λ_2 follow the conjugate prior distributions: $\lambda_1 \sim \lambda_1^{\alpha_1 - 1} e^{-\beta_1 \lambda_1}$ and $\lambda_2 \sim \lambda_2^{\alpha_2 - 1} e^{-\beta_2 \lambda_2}$. The prior distributions for the hyperparameters β_1 and β_2 are: $\beta_1 \sim \beta_1^{\sigma_1 - 1} e^{-\varepsilon_1 \beta_1}$ and $\beta_2 \sim \beta_2^{\sigma_2 - 1} e^{-\varepsilon_2 \beta_2}$. And a series of conditional probabilities are as follows:

$$P(\lambda_1 \mid D_i, k, \alpha_1, \beta_1) \sim \lambda_1^{\alpha_1 + \sum_{j=1}^{k} d_{ij} - 1} e^{-(\beta_1 + k)\lambda_1}$$

$$P(\lambda_2 \mid D_i, k, \alpha_2, \beta_2) \sim \lambda_2^{\alpha_2 + \sum_{j=k+1}^{n} d_{ij} - 1} e^{-(\beta_2 + n - k)\lambda_2}$$

$$P(\beta_1 \mid D_i, k, \lambda_1, \alpha_1, \sigma_1, \varepsilon_1) \sim \beta_1^{\alpha_1+\sigma_1-1} e^{-(\lambda_1+\varepsilon_1)\beta_1}$$

$$P(\beta_2 \mid D_i, k, \lambda_2, \alpha_2, \sigma_2, \varepsilon_2) \sim \beta_2^{\alpha_2+\sigma_2-1} e^{-(\lambda_2+\varepsilon_2)\beta_2}$$

$$P(k \mid D_i, \lambda_1, \lambda_2) \sim \frac{P(D_i \mid k, \lambda_1, \lambda_2)}{\sum_{c=1}^{n} P(D_i \mid c, \lambda_1, \lambda_2)}.$$

Before *Gibbs sampling*, λ_1 is initialized as $\hat{\lambda}_1 = \frac{\sum_{j=1}^{\tau} d_{ij}}{\tau}$ and λ_2 is initialized as $\hat{\lambda}_2 = \frac{\sum_{j=n-\tau+1}^{n} d_{ij}}{\tau}$ because the *change-point* is not likely to occur in the first and last a few super-bins. Because the prior distributions for λ_1 and λ_2 are gamma, and we consider $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are good estimates of the means of the gamma distributions, then β_1 is initialized as $\hat{\beta}_1 = \hat{\lambda}_1 / \mathrm{var}_\tau$ and β_2 is initialized as $\hat{\beta}_2 = \hat{\lambda}_2 / \mathrm{var}_{n-\tau}$, where var_τ is the variance of the first few super-bins and $\mathrm{var}_{n-\tau}$ is the variance of the last few super-bins. α_1 is estimated as $\hat{\alpha}_1 = \hat{\lambda}_1 \times \hat{\beta}_1$, and α_2 is estimated as $\hat{\alpha}_2 = \hat{\lambda}_2 \times \hat{\beta}_2$. ε_1 and ε_2 are set as 0.5 and σ_1 is estimated as $\hat{\sigma}_1 = \hat{\beta}_1 \times 2$, and σ_2 is estimated as $\hat{\sigma}_2 = \hat{\beta}_2 \times 2$.

After initializations, we use *Gibbs sampling* on those conditional probabilities to iteratively estimate *k*, λ_1 and λ_2. Finally, the target spatial density of high-tag bins is $p = \lambda_1 / m$ and $q = \lambda_2 / m$. The estimated densities will then be used to calculate the log likelihood ratios as the scores for maximal scoring segment identifications.

Broad peak identifications

After estimating parameters and setting scores, BroadPeak applies the linear-time Ruzzo-Tompa algorithm (73) to search for all maximal scoring segments (Figure 3.1.A).

For each maximal scoring segment, the observed spatial density of high-tag bins is compared with the background using *z-test* and only the ones with significantly higher densities (*P*<0.05) are added to the final broad peak list. Finally a BED format file of broad peak locations is generated as the output.

Results

Evaluations on real ChIP-seq datasets

In order to evaluate the performance of BroadPeak, we applied it on the genomic ChIP-seq datasets of several diffuse histone modifications in human CD4$^+$ T cells (3). Investigations of the resulted broad peak examples shows that they are consistent with the diffuse ChIP-seq tag count distributions (Figure 3.1.B and 3.1.C). As a global check of the performance, we found that the peaks of H3K79me2 and H3K36me3 concentrate around transcriptional start sites (TSS) and transcriptional termination sites (TTS) respectively (Figure 3.1.D). Also, we found that the edges of the resulted broad peaks of H3K27me3 are more enriched with CTCF binding (Figure 3.1.E). Since CTCF is thought to be related with chromatin barriers, its enrichments around the H3K27me3 broad peak edges support the performance of BroadPeak to identify repressive chromatin domains. Similar results are also observed for broad peaks of H3K9me3 (Figure B.1). We also compared the results based on supervised and unsupervised parameter estimations for H3K36me3 and they are very similar with each other (Figure B.2).

Performance comparisons on simulated datasets

We also generated simulated tag libraries for a list of pre-set broad peaks and applied BroadPeak, along with two existing broad peak calling methods: SICER and

RSEG. We first simulated 3 libraries for tests by selecting 5,000 non-overlapping human genes with different sizes as the real broad peaks. The human genome is divided into 200bp bins. For non-broad-peak regions, the background spatial density of high-tag bins is set as $1x10^{-4}$. For real broad peak regions, the density is set as 20, 50 and 100 fold of the background density respectively for different libraries. The tag count distribution of high-tag bins is a Gaussian distribution with mean of 8 and standard deviation of 2. The spatial density of low-tag bins (noise) is the same throughout the whole genome and is set as 0.5, namely about half of the genome have noise. The tag count distribution of low-tag bins is a Poisson distribution with the average rate as 0.7, which is similar to the H3K36me3 library. Similar to the comparison procedure of RSEG, we run BroadPeak, SICER and RSEG on the three simulated libraries and compared the identified broad peaks with the real peaks. A real broad peak is considered as correctly identified if a certain fraction of it is covered by predicted peaks. Similarly, the predicted broad peak is considered as true if a certain fraction of it is covered by real peaks. The three thresholds of fractions are 20%, 50% and 80%. Based on these basic counts, recall and precision are used to measure the performance and the *F* score is used as the final measurement of the overall performance of the algorithms.

Among all the tested simulated datasets, BroadPeak achieves substantial improvements on recall (Table B.1 and Figure B.3), while maintaining good precision (slightly lower than SICER). The *F* scores of BroadPeak are the best for all the datasets tested. BroadPeak is especially better for larger peaks (Figure B.3). Globally, the size distribution of the resulted broad peaks is much wider for BroadPeak, compared with SICER and RSEG (Figure B.4).

Acknowledgements

We would like to thank members of the Jordan lab for helpful discussions. This work was supported by the Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology [BR-4839]; and the Georgia Tech Integrative Biosystems Institute pilot program grant.

CHAPTER 4

PREDICTION AND VALIDATION OF MIR RETROTRANSPOSON DERIVED INSULATORS IN THE HUMAN GENOME

Abstract

Insulators are regulatory sequence elements that help to organize eukaryotic chromatin via enhancer-blocking and chromatin barrier activity. While there are several examples of transposable element (TE)-derived insulators, there are no known human insulators provided by TEs. Mammalian-wide interspersed repeats (MIRs) are a conserved family of human TEs that have substantial regulatory capacity and share sequence characteristics with tRNA-related insulators. We sought to evaluate whether MIRs can serve as insulators in the human genome. To do this, we applied a bioinformatic screen using genome sequence and functional genomic data from CD4$^+$ T cells to identify a set of 1,178 predicted MIR-insulators genome-wide. These predicted MIR-insulators were computationally validated to serve as chromatin barriers and regulators of gene expression in CD4$^+$ T cells. The activity of predicted MIR-insulators was experimentally validated using enhancer-blocking assays. MIR-insulators are enriched around genes of the T cell receptor pathway and protect these genes from repressive chromatin to facilitate their cell-type specific expression and function. Overall, 58% of the MIR-insulators predicted here show evidence of T cell specific chromatin barrier and gene regulatory activity. MIR-insulators show a distinct local chromatin environment with marked peaks for RNA Pol III and a number of histone

modifications, suggesting that MIR-insulators recruit transcriptional complexes and chromatin modifying enzymes *in situ* to help establish chromatin and regulatory domains in the human genome. The provisioning of insulators by MIRs across the human genome suggests a specific mechanism by which TE sequences can be used to establish gene regulatory networks.

Introduction

Insulators are regulatory sequence elements that help to organize eukaryotic chromatin into functionally distinct domains (44,74). Insulators can encode two different functions: enhancer-blocking activity and chromatin barrier activity. Enhancer-blocking insulators prevent the interaction of enhancer and promoter elements located in distinct domains, and chromatin barrier insulators, also known as boundary elements (43,75), protect active chromatin domains by blocking the spread of repressive chromatin. These two functional roles are not mutually exclusive; compound insulators may encode both enhancer-blocking and chromatin barrier activities (48).

Transposable element sequences are known to provide a variety of regulatory sequences to eukaryotic genomes (66), and there are several examples of TE-derived insulators. The best studied TE-insulator comes from the Drosophila *gypsy* element (46,76-78). *Gypsy* is a long terminal repeat retrotransposon that contains an insulator sequence in its 5’ untranslated region. The *gypsy* insulator interacts with the suppressor of hairy wing [su(Hw)] and modifier of mdg4 [mod(mdg4)] proteins to block regulatory interactions between distal enhancer and proximal promoter sequences. This same insulator can also protect transgenes from position effects indicating that it encodes chromatin barrier activity as well.

More recently, TE-derived insulator sequences have been discovered in mammalian genomes. The short interspersed nuclear element (SINE) B1 has insulator activity that is mediated by the binding of specific transcription factors along with the insulator associated protein CTCF (79). Another mouse TE, the SINE B2 element, serves as a developmentally regulated compound insulator, encoding both enhancer-blocking and chromatin barrier activity, at the growth hormone locus (56). B2 is a tRNA-derived SINE, and the connection to tRNAs is intriguing given the fact that tRNA gene sequences have been shown to encode insulators in yeast (52,53,80,81), mouse (55) and human (33,82). A survey of six mammalian species revealed that lineage-specific expansions of retrotransposons have contributed numerous CTCF binding sites to their genomes (83). A number of these TE-derived CTCF binding sites in the mouse and rat genomes are capable of segregating domains enriched or depleted for acetylation of histone 2A lysine 5 (H2AK5ac), suggesting that they may encode insulator function. Interestingly, this same analysis did not detect retrotransposon driven expansion of CTCF binding sites in the human genome.

Despite the fact that human TEs have yet to be implicated as insulators, the genome is made up of a substantial fraction of TE sequences including numerous tRNA-derived SINE retrotransposons with the potential to encode insulator function (84). Mammalian-wide interspersed repeats (MIRs) are an ancient family of TEs (85) that bear several features suggesting that they may serve as genome regulators in general and insulators in particular. First of all, a number of non-coding MIR sequences were found to be highly conserved, indicative of some functional, presumably regulatory, role (86). Later, it was shown that MIRs are enriched for open chromatin sites (87), encode regulatory RNAs (88), host gene promoters (89) and enhancers (90) and are also associated with tissue-specific expressed genes (91). Finally, MIRs are tRNA-derived SINEs (92) and their sequences include recognizable regulatory motifs, such as the promoter B-box element, that are thought to be important for insulator activity.

In light of these known MIR regulatory sequence characteristics, we sought to evaluate whether MIR elements can encode insulator activity in the human genome. To do this, we employed a bioinformatics screen of genome sequence and functional genomic data to identify a subset of MIR sequences that possess insulator-like features. These features include the presence of intact B-box sequences, occupancy by RNA Pol III and the partitioning of active and repressive chromatin domains (Figure 4.1.A). This procedure resulted in the identification of >1,000 putative MIR-derived insulator sequences, which were first validated computationally and experimentally and then evaluated with respect to a number of functional properties.

Results

Bioinformatic screen and validation

We developed and applied a bioinformatic screen to search for human MIR sequences that may encode insulator activity (Figure 4.1.A). To do this, we evaluated human genome sequence data along with functional genomic data from CD4$^+$ T cells. CD4$^+$ T cells were chosen owing to their importance as a model system for immunology and for the abundance of available functional genomic data that exist for this cell-type. The genome sequence data analyzed consisted of TE and gene annotations, and the functional genomic data included RNA-seq and microarray expression data along with ChIP-seq data for RNA Pol III binding and 39 histone modifications.

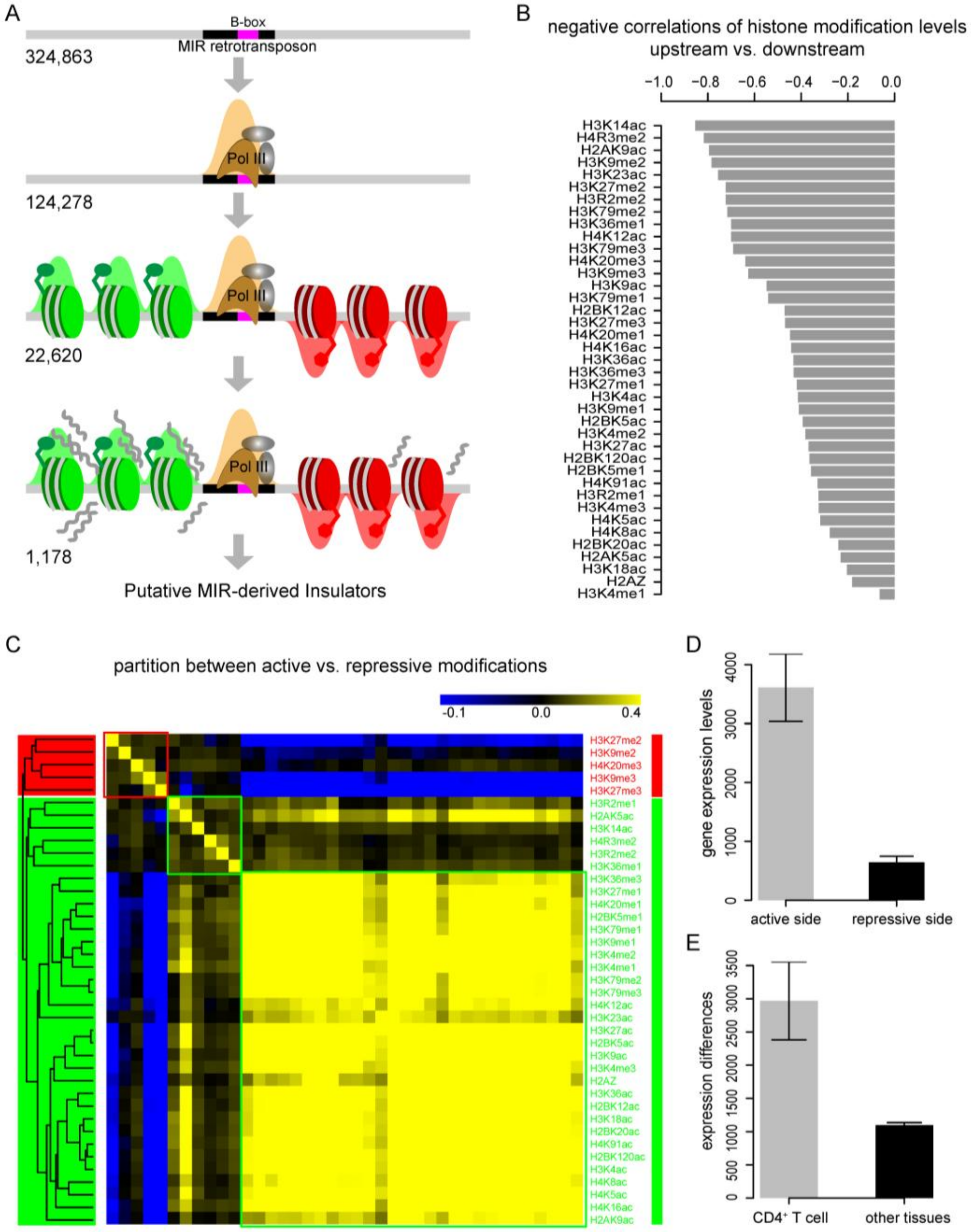


Figure 4.1: Bioinformatic screen and validation of MIR-insulators. (A) Scheme of bioinformatic screen used to predict MIR-insulators. (B) Spearman correlations for individual histone modification profiles upstream versus downstream of predicted MIR-insulators. (C) Heatmap showing Spearman correlations for pairs of histone modification profiles. (D) Average (± standard error) CD4+ T cell expression levels of proximal genes from the active (grey) and repressive (black) sides of predicted MIR-insulators. (E) Average (± standard error) differences in gene expression levels for genes located on the opposite sides of individual predicted MIR-insulators.

First, all MIR sequences in the human genome that contain intact B-boxes and are bound by RNA Pol III in CD4$^+$ T cells were identified. Then, these MIRs were evaluated for their ability to partition active versus repressive chromatin using a previously described approach (33) that segregates histone modifications associated with expressed (active) versus silent (repressive) genomic regions. To do this, broad genomic distributions of 39 histone modifications, with 34 characterized as active and 5 characterized as repressive, were evaluated in order to detect large contiguous regions (domains) of active and repressive chromatin. The B-box containing and RNA Pol III bound MIR elements found to be located between adjacent active versus repressive were then selected for further analysis. Finally, RNA-seq was used to further reduce the list of putative MIR-insulators to those that delineate high versus low expressed genomic regions. This procedure resulted in the identification of 1,178 putative MIR-derived insulators across the human genome (Figure 4.1.A).

The putative MIR-derived insulators were computationally validated with respect to their affects on chromatin and gene expression. For chromatin, the putative insulators were evaluated for their ability to partition individual histone modifications and to delineate sets of modifications that have been previously characterized (18) as active versus repressive (Figure C.1). ChIP-seq tag counts for all 39 histone modifications analyzed here are negatively correlated for the regions upstream versus downstream of the putative MIR-insulators (Figure 4.1.B and Table C.1), indicating that this set of MIRs partitions specific histone modification sites in the local chromatin environment. In addition, when the histone modifications are considered as an ensemble, by clustering their joint upstream versus downstream ChIP-seq profiles, active and repressive modifications can be seen to group together (Figure 4.1.C and Figure C.2). This result indicates that the putative MIR-insulators identified here also delineate active versus repressive chromatin marks. Consistent with this result, genes located proximal to the MIR-insulators in the active chromatin environment are expressed at higher levels than the genes located on the repressive side of the insulators (Figure 4.1.D and Figure C.3).

We also evaluated the role that the putative MIR-insulators play in regulating tissue-specific expression by measuring the differences in expression levels, across 79 human tissues, for genes that flank the insulators. Genes that flank MIR-insulators show greater differences in expression, between the active and repressive sides of the insulators, in CD4$^+$ T cells than seen for the other human tissues (Figure 4.1.E and Figure C.4), consistent with a role for the insulators in establishing tissue-specific chromatin domains. Taken together, the results of the bioinformatic analyses support the notion that human MIRs can serve as insulators and suggest that the putative MIR-insulators identified here encode chromatin barriers that function as tissue-specific regulators.

Experimental validation

We sought to experimentally validate the enhancer-blocking activity for a subset of the MIR-insulators predicted here using previously described human and zebrafish enhancer-blocking assays (EBAs) (56,79,93,94). For the human EBA, a luciferase reporter construct transfected in human HEK 293 cells was used to evaluate both short (200 - 400 bp) and long (1000 – 1200 bp) sequences centered on three predicted MIR-insulators (Table C.2). All three MIR-insulators tested here showed enhancer-blocking activity comparable to the 5’ HS4 positive control (Figure 4.2.A). For the most case, both short and long sequences show similar levels of enhancer-blocking activity. The only exception was the long sequence from the chromosome 11 MIR-insulator, which showed slightly lower enhancer-blocking activity than both the positive control and the short sequence from the same locus. This suggests the possibility of interference from adjacent sequences for this insulator.

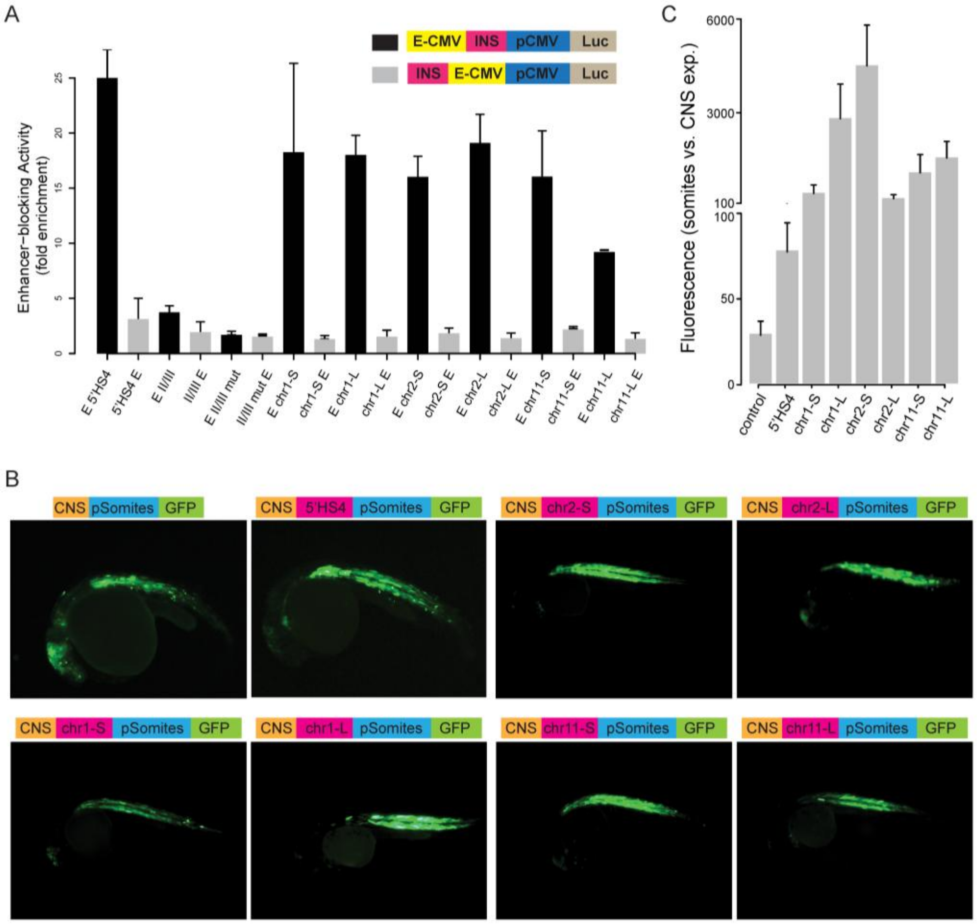


Figure 4.2: Enhancer-blocking assays (EBAs) for predicted MIR-insulators. (A) Human EBA. Enhancer-blocking activity levels (fold-enrichment) are normalized relative to the empty vector. Average enhancer-blocking activity levels (± standard error) for positive (5'HS4 and II/III) and negative (II/III mutated) controls along with results for short and long sequences surrounding predicted MIR-insulators from chromosomes 1, 2 and 11 are shown. For each sequence analyzed, inserts were cloned upstream of the enhancer (negative control site) and between the enhancer and promoter (test site). (B) Zebrafish EBA. Positive (5' HS4) and negative (empty vector) control sequences along with short and long sequences surrounding predicted MIR-insulators from chromosomes 1, 2 and 11 were inserted between the CNS enhancer and the somite promoter. GFP expression in somites versus enhancers indicates relative enhancer-blocking activity. (C) Enhancer-blocking activity in zebrafish is quantified as the average (± standard error) ratio of somite over CNS expression.

The same short and long MIR-insulator sequences were tested in a zebrafish EBA using a GFP reporter construct transiently transfected in embryos. This EBA tests the ability of putative insulator sequences to block interaction of a central nervous system

(CNS) enhancer with a somite GFP promoter (Figure 4.2.B). All MIR-insulator sequences tested show enhancer-blocking activity greater than seen for the 5’ HS4 positive control (Figure 4.2.C). There are some differences between the short and long sequences, but they are not consistent across the tested sequences. Taken together, the results of in these EBAs demonstrate that selected MIR-insulator sequences encode strong enhancer-blocking activity, which is conserved between cell-types and across species.

MIR insulator chromatin features

Having established the chromatin barrier and enhancer-blocking activity of predicted MIR-insulators, we performed a series of enrichment analyses to characterize the local chromatin environment at-and-around these insulators. RNA Pol III occupancy levels peak at MIR-insulator sequences (Figure 4.3.A), which is consistent with the initial bioinformatic screen used for their identification. Nevertheless, the distinct RNA Pol III peak at MIR-insulators differs from the previously observed broad genomic distribution of RNA Pol III binding (95) suggesting the possibility that MIR-insulators are activated via specific recruitment of RNA Pol III and possibly transcriptional activation. In addition, the negative control, performed on a randomly selected set of B-box containing MIRs, shows that specific RNA Pol III binding is not a generic feature of MIRs across the genome. RNA Pol II levels, on the other hand, increase steadily from the MIR-insulator region into the flanking active chromatin environment (Figure 4.3.B), consistent with their role as barriers against the spread of repressive chromatin.

Binding of the insulator-associated protein CTCF also peaks around MIR-insulator sequences relative to flanking genomic regions and shows a strong enrichment compared to the genomic background (Figure 4.3.C). However, CTCF binding levels are slightly depleted right at the locations of the MIR-insulators, raising the possibility of co-operative action between CTCF-independent MIR-insulator mechanisms and the CTCF binding at adjacent genomic loci.

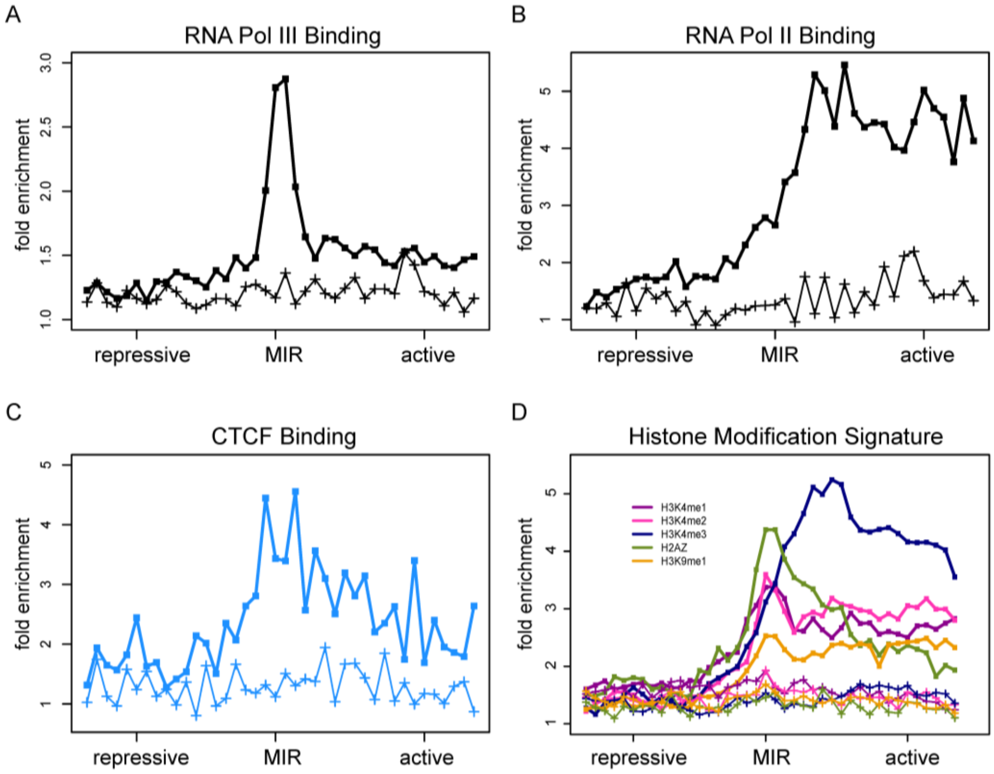


Figure 4.3: Enrichment of chromatin features around predicted MIR-insulators. 8kb windows centered on predicted MIR-insulators were evaluated for the fold-enrichment (compared to genomic background) of (A) RNA Pol III binding, (B) RNA Pol II binding, (C) CTCF binding and (D) levels of five histone modifications. For each enrichment curve, a corresponding negative control (lower lines marked with crosses) is shown based on a randomly selected set of B-box containing MIR sequences of the same size.

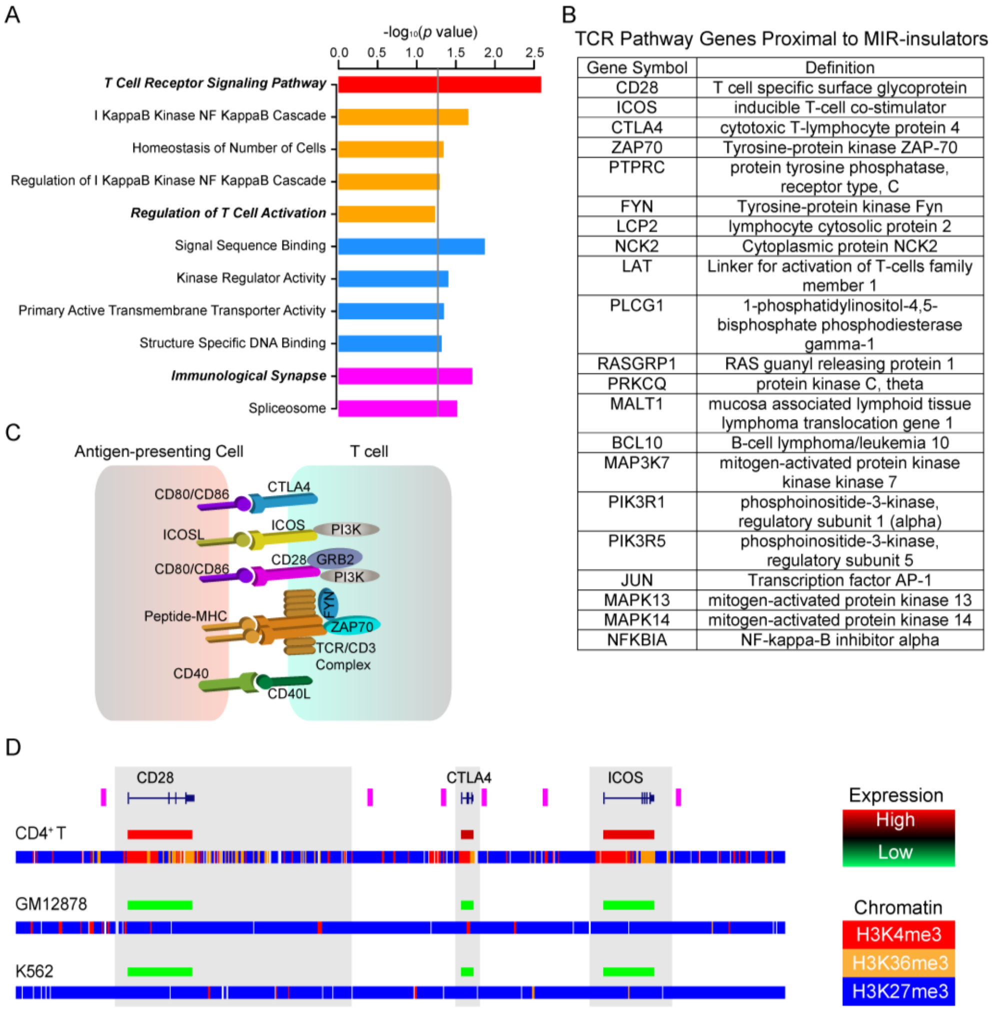
MIR-insulators show a characteristic histone modification signature with distinctive peaks of the H2AZ histone variant, H3K4me1, H3K4me2 and H3K9me1

(Figure 4.3.D). Such peaked patterns can not be expected based on the approach used to detect putative MIR-insulators since the algorithm evaluates broad distributions of active versus repressive histone modifications over 100kb windows surrounding the MIRs. H3K4me3 levels peak adjacent to the locations of the MIR-insulators on the active chromatin side and remain high across the local active chromatin domain. Most of these marks are associated with active chromatin and transcriptional initiation, suggestive of the recruitment of chromatin modifying complexes to MIR-insulators resulting in the local opening of chromatin and priming for gene expression. Consistent with this possibility, MIR-insulators are much closer to the nearest gene transcription start site (TSS) on the active chromatin side than on the repressive side (Figure C.5). H3K4me1 modifications are often associated with enhancer sequences, raising the possibility of some mechanistic overlap between MIR-insulators and enhancers, as has been previously suggested (44).

Tissue-specific chromatin barrier functions of MIR-insulators

Human and zebrafish EBAs indicate that MIR sequences are likely to encode enhancer-blocking activity via conserved mechanisms (Figure 4.2), whereas computational validation of the MIR-insulator predictions suggest that MIR-insulator chromatin barrier activity is tissue-specific (Figure 4.1.E). We sought to further evaluate the possible tissue-specific functional roles played by the MIR-insulators predicted here. To do this, we performed an analysis of the gene ontology (GO) and pathway (KEGG) annotations of the genes located on the active chromatin sides of the MIR-insulators. These genes are enriched for a number of GO functional categories related to T cell function including cell-cell interactions and immune signaling cascades (Figure 4.4.A).

Perhaps most strikingly, this analysis revealed that 21 genes found in the T cell receptor signaling pathway (KEGG: hsa04660) are located adjacent to MIR insulators on the active chromatin side (Figure 4.4.A, Figure 4.4.B and Figure C.6). Among this list, there are several transmembrane receptor proteins, which mediate interactions with antigen-presenting cells, including a co-located genomic cluster of two T cell co-stimulators (CD28 and ICOS) and the co-inhibitor CTLA4 (Figure 4.4.C). The chromatin environment at this genomic cluster, along with the cell type-specific expression patterns of these three genes, exemplifies the T cell-specific regulatory function of the MIR-insulator encoded barrier activity (Figure 4.4.D). In $CD4^+$ T cells, these three genes are flanked by pairs of MIR-insulators that surround an open and active chromatin environment (H3K4me3 and H3K36me3) to the exclusion of repressive chromatin marks (H3K27me3) in the adjacent regions. This pattern stands in contrast to what is seen for GM12878 and K562 cells where the entire locus is marked by repressive chromatin. Accordingly, CD28, ICOS and CTLA4 are highly expressed in $CD4^+$ T cells compared to GM12878 and K562 cells (Figure 4.4.D). Similar cell type-specific distributions of chromatin and gene expression for MIR-insulators and their adjacent genomic regions are observed when the same histone marks and expression levels are compared for all 21 MIR-insulator proximal genes found in the T cell receptor pathway (Figure C.7).



TCR Pathway Genes Proximal to MIR-insulators

Gene Symbol	Definition
CD28	T cell specific surface glycoprotein
ICOS	inducible T-cell co-stimulator
CTLA4	cytotoxic T-lymphocyte protein 4
ZAP70	Tyrosine-protein kinase ZAP-70
PTPRC	protein tyrosine phosphatase, receptor type, C
FYN	Tyrosine-protein kinase Fyn
LCP2	lymphocyte cytosolic protein 2
NCK2	Cytoplasmic protein NCK2
LAT	Linker for activation of T-cells family member 1
PLCG1	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase gamma-1
RASGRP1	RAS guanyl releasing protein 1
PRKCQ	protein kinase C, theta
MALT1	mucosa associated lymphoid tissue lymphoma translocation gene 1
BCL10	B-cell lymphoma/leukemia 10
MAP3K7	mitogen-activated protein kinase kinase kinase 7
PIK3R1	phosphoinositide-3-kinase, regulatory subunit 1 (alpha)
PIK3R5	phosphoinositide-3-kinase, regulatory subunit 5
JUN	Transcription factor AP-1
MAPK13	mitogen-activated protein kinase 13
MAPK14	mitogen-activated protein kinase 14
NFKBIA	NF-kappa-B inhibitor alpha

Figure 4.4: T cell-specific functions of predicted MIR-insulators. (A) Results of a gene ontology (GO) and pathway (KEGG) analysis of proximal genes on the active domain side of MIR-insulators. P-values (-log10 normalized) are shown for the KEGG (red), GO biological process (orange), GO molecular function (blue) and GO cellular component (purple) analyses; the grey line corresponds to P=0.05. (B) List of 21 T cell receptor signaling pathway genes located on the active domain side proximal to MIR-insulators. (C) Portion of the T cell receptor pathway showing membrane receptors that mediate T cell stimulation via antigen presenting cells. (D) Expression levels and the chromatin environment across a genomic cluster of three T cell receptor genes – CD28, CTLA4 and ICOS (blue gene models) – and their co-located MIR-insulators (purple bars) are shown for CD4+ T cells, GM12878 and K562. Relative gene expression levels (high-red to low-green) are shown coincident with the gene models. Genomic distributions of three histone modifications are shown as H3K4me3 (red), H3K36me3 (orange) and H3K27me3 (blue).

We expanded the tissue-specific chromatin and expression analysis to include all MIR-insulators predicted here. To do this, we first classified MIR-insulators as cell-type specific based on the relative distributions of chromatin marks across MIR-insulators in CD4$^+$ T cells versus GM12878 and K562 cells. 681 out 1,178 (58%) of predicted MIR-insulators show skewed distributions of active versus repressive marks in CD4$^+$ T cells, with divergent peaks on opposing sides of the MIR-insulators, compared to relatively flat distributions of the same histone marks in GM12878 and K562 cells (Figure 4.5A-C). Accordingly, these tissue-specific MIR-insulators have proximal genes on the active domain side that are expressed at higher levels in CD4$^+$ T cells than the same genes in GM12878 and K562 (Figure 4.5.D). Furthermore, these MIR-insulators separate pairs of genes, on the active versus repressive chromatin sides of the insulators, that have greater differences in their levels of expression in CD4$^+$ T cells than seen for the same pairs of genes in GM12878 and K562 (Figure 4.5.E). The 42% of MIR-insulators that do not show evidence of tissue-specific function may have broader activity reflecting chromatin boundary establishment earlier in development. It is also possible that additional MIRs not detected in our bioinformatic screen, *e.g.* those that lack intact B-boxes or those do not bind RNA Pol III, may also serve as insulators in CD4$^+$ T cells and/or in other tissues.

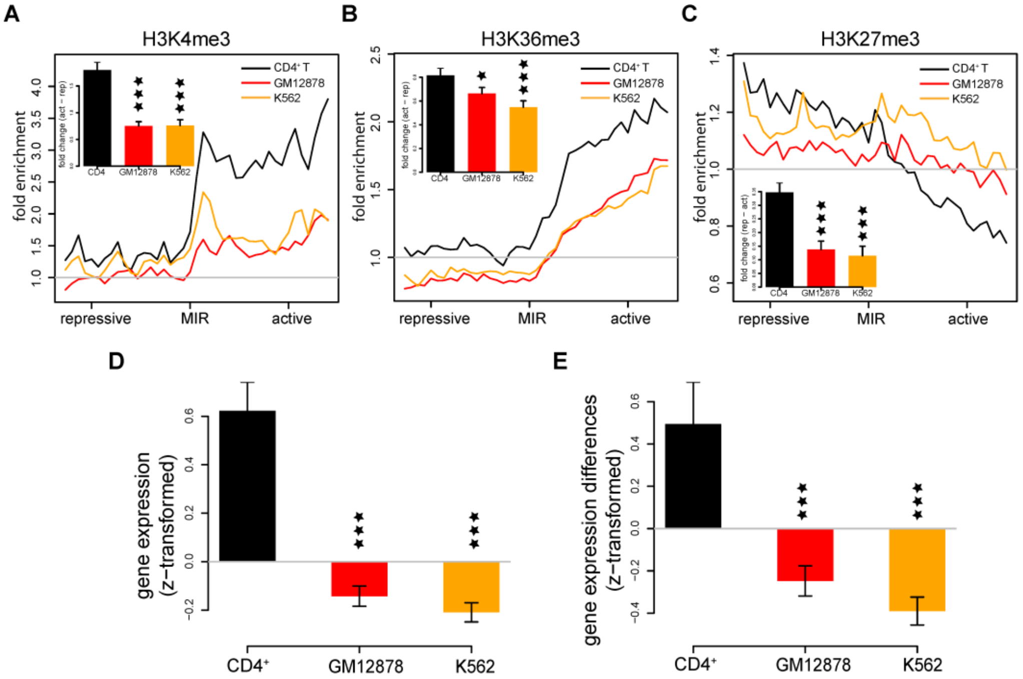


Figure 4.5: Cell-type specific chromatin barrier activity and gene regulation by MIR-insulators. ChIP-seq fold enrichment levels around tissue-specific MIR-insulators are shown for (A) H3K4me3, (B) H3K36me3 and (C) H3K27me3 in CD4+ T cells (black), GM12878 cells (red) and K562 (orange) cells. Insets show the average differences (± standard error) between the active versus repressive domains surrounding MIR-insulators for the marks and cells. (D) Average gene expression levels (± standard error) are shown for genes located in the active domain side proximal to MIR-insulators. (E) Average (± standard error) differences in the gene expression levels for genes located on the opposite sides of individual MIR-insulators. For all bar plots, significance of the differences between CD4+ T cells and other cells are indicated as * P<0.05 ** P<0.01 *** P<0.001.

Discussion

MIRs are relatively ancient and conserved TEs, *i.e.* formerly selfish genetic elements, that have been co-opted to provide a variety of regulatory sequences to their host genomes. Together with their conservation and regulatory capacity, the tRNA-derived sequence features of MIRs suggested to us that they might help to organize human chromatin via the provisioning of insulator elements. Therefore, we screened the human genome for putative MIR-insulators and attempted to validate their activity using a combined computational and experimental approach. The results of our analysis suggest that numerous MIR sequences serve as insulators across the human genome. These predicted MIR-insulators show evidence of both chromatin barrier and enhancer-blocking activity. Interestingly, while the chromatin barrier activity of the MIR-insulators appears to be cell type-specific (Figure 4.1.E, Figure 4.4 and Figure 4.5), the mechanisms underlie MIR's enhancer-blocking activity are seemingly conserved between cell-types and between species (Figure 4.2). This may be attributed to the fact that MIR sequences in isolation possess an innate capacity to provide enhancer-blocking activity via the interaction with conserved protein factors, but *in situ* MIRs interact with cell-type restricted factors to yield a more narrow and specific range of activity. Given that the EBAs were performed with minimal (<1200 bp) constructs, it may be the case that synergistic binding of sites outside the MIR-insulators help to provide cell-type specific barrier activity.

The MIR-insulators identified here have a distinct local chromatin environment (Figure 4.3) that may yield some clues as to their mechanisms of action. For example, while RNA Pol II and RNA Pol III CD4$^+$ T cell binding profiles are highly correlated across the human genome (95), their patterns at-and-around MIR-insulators are quite distinct. RNA Pol III occupancy levels peak right at the MIR-insulators, whereas RNA Pol II levels steadily increase from the MIR-insulators into the adjacent active chromatin domains. This suggests the possibility that RNA Pol III is specifically recruited to MIR-insulators to help establish their activity, thus priming the adjacent chromatin for opening and transcriptional activity as reflected by the increasing RNA Pol II levels. The histone modification profiles around MIR-insulators are consistent with this model. There are clear local peaks of modifications right at the MIR-insulators, such as seen for H3K4me1 and H3K4me2, but these same marks of open chromatin are also maintained at relatively higher levels in the adjacent active domains. H3K4me3 shows a similar pattern, but its peak is shifted further into the active domain and it is maintained at higher levels through this domain. Thus, there may be a wave of progressive methylation of the H3K4 position starting at the MIR-insulator locations and continuing with the addition of methyl groups into the active domain, similar to what we observed previously for human chromatin barriers(33).

The location of MIR-insulators relative to proximal gene promoters also sheds some light on their mechanism of action. MIR-insulators are located much closer to the promoters of the genes that are located on the active side of the insulator compared to the genes located on the repressive side (Figure C.5). This suggests that MIR-insulators are not only located in such a way to protect proximal promoters from the encroachment of repressive chromatin, but also restrict interactions with promoters to only those enhancers that are located nearby or within genes. This scenario can be illustrated by cluster the co-located T cell receptors – CD28, CTLA4 and ICOS – each of which is flanked by a pair of MIR-insulators (Figure 4.4.D). This apparent restriction to local enhancers would seem to be odds with the textbook definition of enhancers as regulatory elements that exert their effects over long ranges. However, recent genome-wide analyses of chromatin reveal that gene bodies are enriched for enhancer elements (7,25,28) and these local regulatory sequences may be largely responsible for cell-type specific expression.

TE-derived insulators have previously been associated with CTCF binding events (83). The dependence of MIR-insulators on the vertebrate insulator protein CTCF is far from clear based on the results of our analysis. While there is a clear enrichment of CTCF binding in the local proximity of MIR-insulators (Figure 4.3.C), only 52 of 1,178 (4.5%) MIR-insulator sequences predicted here are actually bound by CTCF in CD4$^+$ T cells. In fact, for many of the MIR-insulators, CTCF binding appears to peak in the genomic regions just adjacent to the elements. This suggests the possibility of cooperativity between MIR sequences and the local genomic context in establishing insulator activity. However, if this were indeed the case, one would expect that the longer insulator sequence inserts used in the EBA constructs would invariably yield higher enhancer-blocking activity and this was clearly not the case (Figure 4.2). These results raise the possibility that MIR-insulators function in a largely CTCF independent manner.

Many questions as to the specific mechanisms underlying MIR-insulator activity remain to be answered. For example, while the compound insulator activity of the mouse tRNA-derived SINE B2 is related to the transcriptional activity of the element (56), it is not clear if the same can be said for MIR-insulators. Furthermore, many of the protein factors that interact with MIR-insulators remain to be elucidated. Nevertheless, the finding that numerous MIRs across the human genome can provide insulator activity raises intriguing possibilities. In particular, when their repetitive nature is considered together with their role in organizing chromatin, it suggests a possible mechanism for the establishment of cell-type specific regulatory networks by TEs as long ago envisioned by McClintock (96) and Britten and Davidson (97).

Materials and methods

Genomic and functional genomic data sets

The human genome reference sequence (NCBI build 36.1, UCSC version hg18) was analyzed with respect to the locations of MIR TE sequences and NCIB RefSeq gene locations using the UCSC Genome Browser 'RepeatMasker' and 'RefSeq Genes' tracks respectively. ChIP-seq data (3,18) were used to characterize the genomic locations of 38 histone modifications and one histone variant in CD4$^+$ T cells. ChIP-seq data were used to characterize the genomic locations of RNA Pol II, CTCF (3) and RNA Pol III (95) binding sites in CD4$^+$ T cells. ChIP-seq data from the ENCODE consortium were used to characterized the locations of three histone modifications in GM12878 and K562 cells (7,98). Microrray data were used to characterized gene expression levels across 79 human tissues (99), including CD4$^+$ T cells, along with GM12878 and K562 (100,101). Microarray signal intensity values were normalized using the z-transformation in order to compare relative expression levels across tissues and microarray platforms. RNA-seq data from CD4$^+$ T cells (95) were used to characterize genome expression levels.

Bioinformatic prediction and validation of MIR-insulators

Human genome MIR sequences (candidate insulators) were screened through a series of filters to identify a final set of predicted MIR-derived insulators (Figure 4.1.A). The final set of predicted MIR-insulators (*n*=1,178) contains the following set of properties: intact B-box promoter sequences, occupancy by RNA Pol III, segregation of active versus repressive chromatin domains and segregation of expressed versus silent genomic regions. The ability of the predicted set of MIR-insulators to segregate individual histone modifications and to group active and repressive modifications together were computationally validated using correlation analysis of ChIP-seq data for the 39 CD4$^+$ T cell histone modifications. Details of the MIR-insulator computational validation procedure can be found in the Figure C.1.

Enhancer blocking assays (EBAs)

Human EBAs were performed as previously described (56,102) using the pELuc vector and transient transfection HEK 293 cells. Selected MIR-insulator sequences were cloned upstream (negative control) or between (test) enhancer and promoter sequences and enhancer-blocking activity was measured based on relative levels of luciferase expression. The 5’ HS4 insulator from the chicken beta-globin locus and the minimal insulator sequence motifs (II/III) from this same element were used as positive controls in this assay. Mutated II/III sequence motifs, incapable of binding CTCF, were used as negative controls. Three replicates were performed for each EBA.

Zebrafish EBAs were performed as previously described (103) using a Tol2 transposon-based vector and transient transfection of zebrafish embryos. Selected MIR-insulator sequences were cloned between a central nervous system (CNS) enhancer and a promoter that drives somite expression, and enhancer-blocking activity was measured based on relative levels of somite/CNS GFP expression. The 5’ HS4 insulator from the chicken beta-globin locus was used as a positive control in this assay; an empty vector was used as a negative control. For each putative MIR-insulator sequence tested, 41-46 replicates were assayed to control for chromatin position effects.

Acknowledgements

This work was supported by an Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839 to J.W. and I.K.J.);

Georgia Tech Integrative BioSystems Institute pilot program grant (to J.W. and I.K.J.); the Buck Institute Trust Fund (to V.V.L.); the Spanish Ministry of Science and Innovation (MICINN) (BIO2009-1297 to L.M., BFU2010-14839 to J.L.G.-S and CSD2007-00008 to J.L.G.-S) and by Junta de Andalucía (CVI-3488 to J.L.G.); S.E.M. was supported by CIBERER (ISCIII).

CHAPTER 5

GENOME-WIDE PREDICTION AND ANALYSIS OF HUMAN CHROMATIN BOUNDARY ELEMENTS

Abstract

Boundary elements partition eukaryotic chromatin into active and repressive domains, and can block regulatory interactions between domains. Boundary elements act via diverse mechanisms making accurate feature-based computational predictions difficult. Therefore, we developed an unbiased algorithm that predicts the locations of human boundary elements based on the genomic distributions of chromatin and transcriptional states, as opposed to any intrinsic characteristics that they may possess. Application of our algorithm to ChIP-seq data for histone modifications and RNA Pol II binding data in human CD4$^+$ T cells resulted in the prediction of 2,542 putative chromatin boundary elements genome-wide. Predicted boundary elements display two distinct features: first, position-specific open chromatin and histone acetylation that is coincident with the recruitment of sequence-specific DNA binding factors such as CTCF, EVI1 and YYI, and second, a directional and gradual increase in histone lysine methylation across predicted boundaries coincident with a gain of expression of non-coding RNAs, including examples of boundaries encoded by tRNA and other non-coding RNA genes. Accordingly, a number of the predicted human boundaries may function via the synergistic action of sequence-specific recruitment of transcription factors leading to non-coding RNA transcriptional interference and the blocking of facultative

heterochromatin propagation by transcription-associated chromatin re-modeling complexes.

Introduction

Eukaryotic chromosomes are functionally organized into alternating active and repressive chromatin domains, referred to as euchromatin and heterochromatin respectively (64,104). Active chromatin domains are characterized by histone modifications that facilitate gene expression via the opening of chromatin, which provides transcription factors access to genomic DNA, whereas repressive domains are enriched with histone modifications that yield more tightly compact and less accessible chromatin leading to the repression of gene expression (1,2,105-109). Accordingly, the establishment and maintenance of distinct chromatin domains has important implications for gene regulation specific to cellular development and function (110,111).

The organization of eukaryotic chromatin into functionally distinct domains implies the existence of chromatin partitioning elements that can be used both to delineate active euchromatic and repressive heterochromatic domains, while preserving their structural integrity, and to prevent regulatory cross-talk between different domains (43,44,57,75). Such chromatin partitioning elements do in fact exist and they are known as ‘boundary elements’ (46,56,78). Boundary element functionality is characterized by two fundamental properties: 1) the ability to protect from chromosomal position effects by acting as barriers against the self-propagation of repressive chromatin (46,80,112) and 2) the ability to insulate or block regulatory interactions between distal enhancers and proximal gene promoters (57,113,114). Some boundary elements are able to act both as chromatin barriers and enhancer blocking insulators (56,115). Boundary elements that are cell-type specific help to establish alternating facultative, as opposed to constitutive, euchromatic and heterochromatic domains.

Known boundary elements are diverse, and several different mechanisms of boundary element activity have been uncovered. First, fixed boundary elements consist of specific DNA sequences and their associated proteins, which establish boundaries with well defined positions. Such precisely located boundaries are thought to form discrete physical barriers that partition distinct chromatin and/or regulatory domains. For example, the HS4 boundary element found upstream of the chicken β-*globin* locus is bound by the CCCTC-binding factor (CTCF), a well known vertebrate insulator associated protein with demonstrated enhancer blocking activity (49,116). The scs/scs' elements in Drosophila provide fixed boundaries at the heat-shock domain locus (112,114,117), and the chromatin barrier activity of the scs/scs' boundaries is dependent upon the binding of two protein factors Zw5 and BEAF (118).

Second, there are variable boundary elements that do not occupy specific DNA sequences or genomic locations. These variable boundaries are thought to be established and maintained through a dynamic balance of collisions between opposing chromatin modifying enzyme complexes responsible for the formation of euchromatin on one side of the boundary and heterochromatin on the other (119,120). For example, the phenomenon of position effect variegation (PEV) in Drosophila can be attributed to variable boundary elements (44,121). PEV refers to the variegated expression of genes located between adjacent euchromatic and heterochromatic domains. PEV occurs due to the changing locations of variable boundaries between cells, which result in genes being located in alternating euchromatic or heterochromatic environments in different cells.

Third, boundary element activity can depend upon transcriptional interference from small non-protein-coding transcriptional units, such as tRNA genes in yeast (52,53,74,80,81) or tRNA-derived SINE retrotransposons in mouse (56,79). Boundary elements that function via transcriptional interference contain specific sequence features needed to recruit transcription factors (*e.g.* the Pol II and Pol III machineries), and they may also provide a physical barrier to the propagation of heterochromatin via nucleosomal gaps close to transcription start sites. These nucleosomal gaps may also serve as entry sites for chromatin remodeling complexes that help to establish the boundaries (43,53).

Thus, many of the currently known boundary elements have been defined functionally, based on experimental confirmation of their activity, rather than categorically based on the presence of well defined features. Indeed, as detailed above, there are diverse mechanisms that underlie boundary element activity and no common sequence or protein features that unite all known boundaries. This lack of common boundary element features makes comprehensive prediction of boundaries difficult. To date, boundary element prediction methods have relied on specific features to identify mechanistically coherent subsets of boundaries. For example, genome-wide distributions of CTCF binding sites considered together with chromatin domain borders have been used to infer the locations of putative fixed boundaries (3,122). This feature-based approach to boundary element prediction may overlook boundaries that function via diverse and possibly as yet unknown mechanisms.

Recently, a number of genome-wide maps of histone modifications have been computationally analyzed in order to describe chromatin architecture in terms of the distribution of distinct domains within and between cell types. For instance, studies in *Drosophila melanogaster* (25,123), *Caenorhabditis elegans* (15) and human (6,7) have characterized the genomic distributions of euchromatic and heterochromatic domains at high levels of resolution. The ability to characterize chromatin domain distributions in this way suggests that it should also be possible to more precisely define the locations of putative chromatin boundaries between domains along with their local properties. To address this issue here, we employed a computational analysis of histone modification maps in human CD4$^+$ T cells. To date, CD4$^+$ T cells represent the single best characterized system for studying chromatin architecture as there exist genome-wide maps for 38 histone modifications and one histone variant (3,18). The existence of multiple (five) repressive modifications, in particular, is a unique aspect of this data set that provides increased resolution for delineating active versus repressive domains. Furthermore, experimentally characterized genome-wide maps of chromatin accessibility (DNase I hypersensitive sites), binding sites for RNA Pol II and Pol III as well as several other protein factors exist for CD4$^+$ T cells along with RNA-seq data for genome expression.

The goal of this study was to take advantage of the detailed genome-wide chromatin maps that exist for CD4$^+$ T cells in order to predict and analyze a collection of putative human boundary elements that is unbiased with respect to the mechanisms of boundary activity. Such a set of predicted boundary elements could help to prioritize experimental interrogation of boundaries and further define the scope of possible boundary element mechanisms. To this end, we developed a boundary element prediction algorithm that does not rely on any previously characterized features of boundary element sequences, such as the binding of specific protein factors (*e.g.* CTCF), the presence of tRNA or tRNA-derived sequences or the expression of non-coding RNAs. Rather, our approach defines the genomic positions of putative boundaries in cell-type specific manner based solely on the locations of transition points between facultatively active (euchromatic) and repressive (heterochromatic) domains, along with the distributions of Pol II binding sites. We chose this objective approach to avoid biasing our boundary element predictions with respect to a limited set of previously known features, and more importantly, to allow for the opportunity to discover boundary elements that may operate via novel, previously unreported mechanisms of action. Boundary element prediction proceeded in two steps. First, we defined euchromatic and heterochromatic domains based on the distributions of active versus repressive histone modifications, and the regions between adjacent domains were taken as possible locations for boundary elements. Second, the regions between chromatin domains were further analyzed with respect to the distributions of Pol II binding sites to more precisely locate putative boundaries.

Application of this two-stage chromatin boundary element prediction algorithm to human CD4$^+$ T cell chromatin data resulted in the prediction of 2,542 cell-type specific boundary elements genome-wide. The functional relevance of the predicted boundaries, with respect to facultative chromatin and cell-type specific expression, was supported by the finding that pairs of genes immediately flanking the boundaries are more divergently expressed in CD4$^+$ T cells than in other human cells. Feature analysis of the predicted human boundaries suggests the possibility of several novel and distinct modes of action: 1) predicted boundaries show a distinct local chromatin environment including peaks of open chromatin marked by enrichment for numerous histone acetylations. These results suggest that the establishment of boundaries involves the local action of specific chromatin remodeling proteins, 2) while many of the predicted boundaries are shown to be bound by the well known insulator protein CTCF, there are a number of boundaries that may function in a CTCF-independent manner via the binding of protein factors that are known to function in chromatin remodeling but were not previously implicated in boundary activity, *e.g.* EVI1 and YY1, 3) a number of predicted boundaries show evidence for the action of transcriptional interference including examples of putative tRNA derived boundaries. tRNA genes were previously shown to function as boundaries in yeast (52,53,80,81) but these are the first examples of putative tRNA derived boundaries in human.

<u>Materials and methods</u>

Datasets of histone modifications and Pol II binding in CD4+ T cells

We used publicly available genome-wide ChIP-seq data for 38 histone modifications and one histone variant (H2A.Z) defined in human CD4$^+$ T cells (3,18). These 39 histone modifications are classified into active histone modifications and repressive histone modifications, based on previous results (18), for use in chromatin domain prediction. Active modifications are positively correlated with gene expression levels and are known to mark euchromatic genomic regions, whereas repressive modifications are negatively correlated with expression levels and mark heterochromatic domains. The 34 active modifications used here are: H2BK5ac, H2BK12ac, H2BK20ac, H2BK120ac, H2AK5ac, H2AK9ac, H2AZ, K3K4ac, H3K9ac, H3K14ac, H3K18ac, H3K23ac, H3K27ac, H3K36ac, H4K8ac, H3K12ac, H4K5ac, H4K16ac, H4K91ac, H2BK5me1, H3K4me1, H3K4me2, H3K4me3, H3K9me1, H3K27me1, H3K36me1, H3K36me3, H3K79me1, H3K79me2, H3K79me3, H3R2me1, H3R2me2, H4K20me1 and H4R3me2. The 5 repressive modifications are: H3K9me2, H3K9me3, H3K27me2, H3K27me3 and H4K20me3. Genome-wide ChIP-seq data for Pol II binding in CD4$^+$ T cells was also obtained from Barski et al. 2007.

General scheme of chromatin boundary element prediction algorithm

In order to predict chromatin boundary elements in CD4$^+$ T cells, we designed a two-stage algorithm (Figure 5.1.A). First, we employed active versus repressive histone modification distribution information to define the locations of large-scale euchromatic and heterochromatic domains respectively (Figure 5.1.B). Regions in transitions (RIT) between adjacent euchromatic and heterochromatic domains are taken as possible locations containing chromatin boundary elements. Second, we predicted the specific locations of boundary elements using Pol II binding inside RITs. Boundary elements were taken as 8kb windows flanking the precise transition points between high versus low Pol II binding regions. Only RITs with one such Poll II transition point were considered to contain unambiguous boundary elements. Details for each stage of the algorithm are provided below.

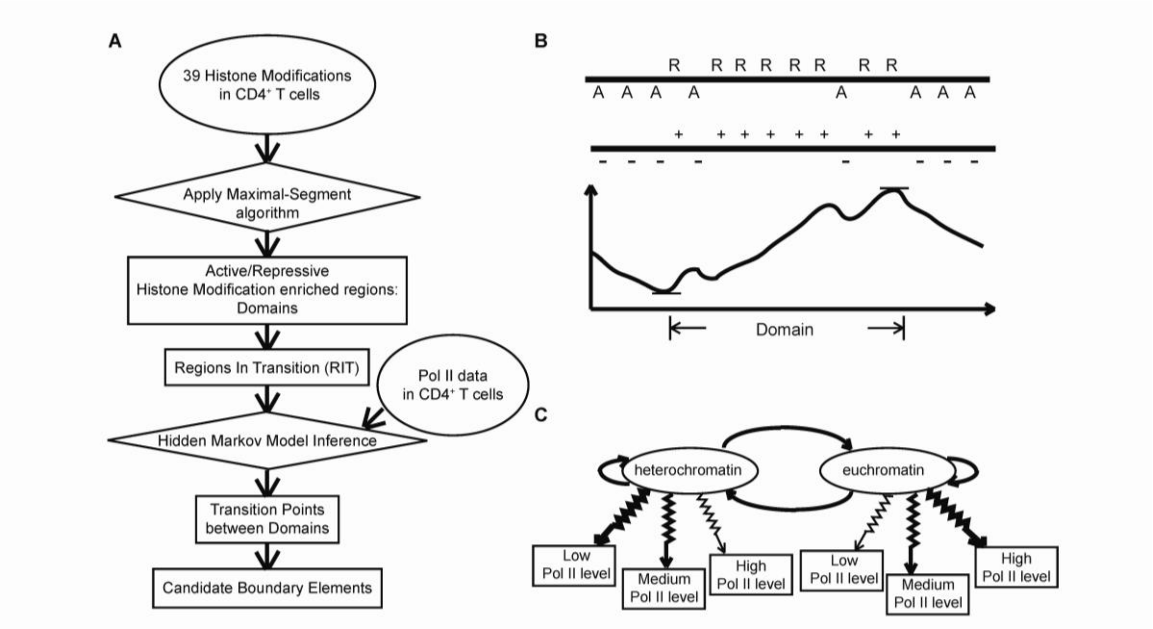

A
39 Histone Modifications in CD4+ T cells
Apply Maximal-Segment algorithm
Active/Repressive Histone Modification enriched regions: Domains
Regions In Transition (RIT)
Pol II data in CD4+ T cells
Hidden Markov Model Inference
Transition Points between Domains
Candidate Boundary Elements
B
R R R R R R R R R
A A A A A A A A
+ + + + + + + + +
- - - - - - - -
Domain
C
heterochromatin
euchromatin
Low Pol II level
Medium Pol II level
High Pol II level
Low Pol II level
Medium Pol II level
High Pol II level

Figure 5.1: Boundary element prediction algorithm scheme. A: Pipeline of the boundary element prediction algorithm. B: Scheme of domain prediction: repressive modifications (R) and active modifications (A) at each genomic site are transformed to positive or negative scores. A maximal-segment algorithm is then applied on the score strings to locate contiguous regions with local maximal cumulative scores; such regions correspond to euchromatic or heterochromatic domains. C: Scheme of the hidden Markov model for boundary element prediction. The two hidden states are heterochromatin and euchromatin. Each state is characterized by distinct emission probabilities of low, medium and high Pol II binding levels.

Domain localization with a maximal-segment algorithm

Histone modifications were characterized as active versus repressive based on their correlation with gene expression levels as previously described (18). All active modifications were then considered together as a single set for subsequent analysis as were all repressive modifications. In order to infer heterochromatic domains, we set a positive score for each genomic location which has repressive histone modification ChIP-seq tags and a negative score for each location with active modification tags. The tag counts of repressive and active modifications were further classified as small (<= 8 tags), medium (>8 tags and <=15 tags) and large (>15 tags). Based on Karlin's theorems (72), the scores for individual genomic sites are set as $s_{ij} = \ln(\frac{p_{ij}}{q_{ij}})$, where i ={repressive, active} and j ={small, medium, large}. p_{ij} represents the estimated frequency of the specific kind of sites in real heterochromatin domains, and q_{ij} represents the genomic background frequency of the specific kind of sites. Intuitively, in heterochromatic domains, the frequency of repressively modified sites is higher than the genomic frequency of repressively modified sites and the corresponding scores are positive and larger for sites with more tags. Likewise, the scores for actively modified sites are negative. We use the peri-centromeric regions to estimate p_{ij}, since peri-centromeric regions are believed to be heterochromatic regions. Peri-centromeric regions are defined as the regions on both sides of centromeres extending to the most proximal gene as previously described (124). After the scoring step, we applied the maximal-segment algorithm (73) to detect contiguous genomic regions with local maximal cumulative scores. Such contiguous regions represent domains that are enriched with repressive histone modifications, *i.e.* heterochromatic domains (Figure 5.1.B). As previously suggested (64), we removed the candidate heterochromatic domains that are <10kb. This cut-off was chosen to reflect that fact that domains, by definition, are thought to be broad and widely spread, and relatively short genomic regions <10kb are more likely to represent discrete regulatory elements than *bona fide* domains. The remaining inferred heterochromatic domains were used in subsequent steps.

In order to infer euchromatic domains, we set positive scores for actively modified sites and negative scores for repressively modified sites, and the other steps were the same as described for inference of heterochromatic domains. As with heterochromatic domains, predicted euchromatic domains <10kb were eliminated from further consideration. In order to estimate the frequency of actively modified sites in real euchromatic domains, we used the histone modification data for the top 5% of genes that are most highly expressed in CD4$^+$ T cells (99) assuming those genes must be inside euchromatic regions.

After obtaining heterochromatic domains and euchromatic domains in this way, we define a list of RITs between adjacent heterochromatic and euchromatic domains. All possible boundary elements should reside within RITs, but it is not necessary that every RIT contains a boundary element. The next step in the algorithm narrows down these RITs to more precisely define the location of putative boundary elements.

Boundary element localization with a hidden Markov model

In order to more accurately predict specific chromatin boundary element locations within RITs, we took advantage of the fact that euchromatic regions have higher Pol II binding signal levels than heterochromatic regions. We built a two-state hidden Markov model (HMM) on Pol II binding data, and employed the Viterbi algorithm to find the most possible hidden state chain (Figure 5.1.C). The two states in this chain are heterochromatin and euchromatin respectively. The emission probabilities of the Pol II signal in euchromatic regions are estimated based on Pol II data in genes which are the top 5% most highly expressed in CD4$^+$ T cells, and the emission probabilities of Pol II signal in heterochromatic regions are estimated based on Pol II data in genes which are not expressed (the lowest 5%). The total size of heterochromatic domains is denoted as s_1 and the total size of euchromatic domains as s_2. The total size of RITs that go from heterochromatin to euchromatin is denoted as t_{12}, and the total size of RITs that go from euchromatin to heterochromatin as t_{21}. Then the transition probability from heterochromatin to euchromatin is estimated as $\frac{t_{12}}{(t_{12}+s_1)}$, and the transition probability from heterochromatin to heterochromatin is estimated as $\frac{s_1}{(t_{12}+s_1)}$. The transition probability from euchromatin to heterochromatin is estimated as $\frac{t_{21}}{(t_{21}+s_2)}$, and the transition probability from euchromatin to euchromatin is estimated as $\frac{s_2}{(t_{21}+s_2)}$.

After running the Viterbi algorithm over all RITs, we recorded the most probable hidden state chains for each RIT. Transition points from one state to the other were taken as possible boundary element locations. To avoid bivalently modified regions and to eliminate small scale variations in Pol II binding, boundary elements were only predicted for RITs that show a single transition point in the hidden state chain. Since boundary elements may be expected to contain a combination of multiple regulatory elements around the precise transition points, putative boundary elements were taken as 8kb regions around the exact transition points.

DNase I hypersensitivity analysis

Genome-wide DNase I hypersensitivity data in human CD4$^+$ T cells were taken from (4). The genomic locations of DNase I hypersensitive sites are transformed to NCBI36/hg18 using the UCSC Genome Browser program Liftover (70,125). To check whether the predicted boundary elements are more DNase I hypersensitive than flanking regions on average, we extended the predicted boundary elements by 8kb upstream and downstream and divided the extended regions into 1kb non-overlapping bins. For each bin, we calculated the average DNase I hypersensitive scores and normalized them by the genomic average DNase I hypersensitive scores.

Histone modification signature analysis

Tag counts for each individual histone modification were computed for predicted boundary elements extended by 8kb upstream and downstream. Extended regions were divided into 1kb non-overlapping bins, and for each bin, the average tag counts are normalized by genomic averages.

Analysis of CTCF binding

Genome-wide ChIP-seq data for CTCF binding in human CD4$^+$ T cells were taken from (3). We only considered locations with more than 5 tags as reliable CTCF binding sites. To check whether predicted boundary elements have higher affinity to CTCF binding than flanking regions on average, we extended the predicted boundary elements by 8kb upstream and downstream and divided the extended regions into 1kb non-overlapping bins. For each bin, we calculated the average CTCF tag counts and normalized them by the genomic average CTCF tag count for 1kb regions.

TFBS analysis

In order to look for putative protein factors associated with predicted chromatin boundary elements, we used the "TFBS Conserved" track from the UCSC Genome Browser. We gathered those computationally predicted conserved TFBS (with Zscore above 1.96) inside predicted boundary elements. For each transcription factor, we counted the number of its appearance within boundary elements and statistically tested whether the specific transcription factor is significantly associated with boundary elements using the hypergeometric test.

Boundary element transcription analysis

RNA-seq data of transcription in human CD4$^+$ T cells were taken from (95). We extended the putative chromatin boundary elements by 8kb upstream and downstream and divided them into 1kb non-overlapping bins. We calculated the average non-protein-coding RNA-seq tag counts for each bin and normalized them by the genomic average tag counts. The data was then $\log_2$ transformed. Predicted boundary elements were classified into two groups: boundaries containing RNA genes and boundaries without RNA genes, and the above calculations were done on the two groups of boundaries separately. The annotations of RNA gene locations are from the "RNA gene" track (126,127) on UCSC Genome Browser.

Gene expression analysis

Gene expression profiles were taken from (99). For genes located within predicted euchromatic domains and heterochromatic domains, we calculated their average expression levels in human CD4$^+$ T cells. For each predicted boundary element, we took the two genes most proximal to it on the two opposite sides (the euchromatic side and the heterochromatic side) and calculated the expression differences between these pairs for CD4$^+$ T cells and for all other tissues together.

Gene function annotations

Gene Ontology analysis and KEGG pathway analysis were performed using MSigDB (128,129) for predicted euchromatic domains with high gene densities (> 1 gene/20 kb).

Results

Datasets and chromatin boundary element prediction algorithm

In recent years, a substantial body of data detailing the chromatin structure of eukaryotic genomes has been accumulated. For the human genome in particular, there are now genomic maps with experimentally characterized locations of numerous histone modifications as well as binding sites for a variety of proteins. Such data provide opportunities for the discovery of novel chromatin related regulatory elements across the genome.

Human CD4$^+$ T cells represent one of the best characterized systems for the genome-scale analysis of chromatin. Keji Zhao and colleagues have used chromatin immunoprecipitation followed by high-throughput sequencing experiments (ChIP-seq) to generate genome-wide maps for 38 histone modifications and one histone variant (H2A.Z), CTCF binding, Pol II binding and Pol III binding (3,18,95). Chromatin accessibility in CD4$^+$ T cells has been evaluated genome-wide using DNase I hypersensitivity assays coupled to high-throughput sequencing (4), and genome-wide CD4$^+$ T cell expression levels have been determined using microarray and RNA-seq technologies (95,99).

We took advantage of the existence of these genome-scale chromatin datasets to facilitate the discovery of boundary elements in the human genome. The goal of this work was to provide a comprehensive list of likely boundary element candidates, and then to evaluate the features of these putative boundaries with respect to possible mechanisms of action. We designed a two-stage algorithm to predict the locations of putative boundary elements (Figure 5.1.A). In the first stage, we defined the locations of large-scale active (euchromatic) and repressive (heterochromatic) chromatin domains based on the genomic distributions of active and repressive histone modifications. The histone modifications analyzed here were characterized active or repressive as previously described (see Materials and Methods) (18). For each genomic position, a specific score (negative or positive) was assigned according to the relative abundance of active or repressive modifications. A maximal-segment algorithm was then applied to the resulting string of scores to locate contiguous genomic regions with maximal local cumulative scores (Figure 5.1.B). The maximal-segment algorithm was chosen because it can detect such contiguous regions over variant lengths, and it is robust to small scale stochastic noise in the ChIP-seq data. The maximal-segment algorithm also worked well here because the parameters that define the relative negative or positive scores can be directly estimated from the ChIP-seq data. Further details on our maximal-segment algorithm for domain detection can be found in the Materials and Methods section (see Domain localization with a maximal-segment algorithm).

We searched for chromatin boundary elements that reside within regions between adjacent euchromatic and heterochromatic domains – hereafter referred to as regions in transition (RITs). However, it should be noted that not all RITs will necessarily contain discretely located boundary elements. For instance, some RITs may contain regions with fuzzy patterns of active and repressive modification distributions that would not allow for precise delineation of boundary element locations. Such fuzzy patterns may represent boundaries that act via PEV related mechanisms, owing to different boundary locations among heterogeneous cell populations, and these imprecisely located boundaries will not be detected by our method. Furthermore, because the sizes of RITs can be relatively large (>50kb) in some cases, a method is needed to narrow down the genomic regions where predicted boundary elements can be located. In light of both of these issues, we developed a second stage of the algorithm that uses a hidden Markov model (HMM) of Pol II binding distributions along RITs in order to more precisely locate boundary elements (Figure 5.1.C). This approach is based on the rationale that euchromatin is transcriptionally active, whereas heterochromatin is largely transcriptionally silent. Accordingly, euchromatin is expected to have higher levels of Pol II binding, and heterochromatin is expected to have lower levels of Pol II binding. Furthermore, Pol II protein complexes are known to associate with proteins that have acetyltransferase and/or chromatin re-modeling functions (130). Thus, boundary elements are expected to be located in genomic regions with particularly sharp transitions between low and high Pol II binding; HMMs are ideal for delineating such abrupt transitions.

HMMs were used to model RITs by predicting the facultative chromatin state – euchromatin or heterochromatin – for each genomic site that best explains the Pol II binding distribution along each RIT. To do this, the Viterbi algorithm was used to infer the most probable chromatin state chain along the RITs based on Pol II binding emission probabilities and chromatin state transition probabilities (Figure 5.1.C). Details on the HMM we used for boundary element localization can be found in the Materials and Methods section (see Boundary element localization with a hidden Markov model).

After obtaining the most probable hidden state chains of euchromatin and heterochromatin, we removed RITs that contain more than one transition point between the two chromatin states, since these represent ambiguously located boundaries. Sequence features of the remaining RITs are summarized in Table D.1. For RITs with single chromatin state transition points, we take 8kb regions centered on those transition points as putative boundary element regions. The 8kb window size was chosen to strike a balance between the utility of precisely locating predicted boundary elements and the biological reality that boundary element activity may be spread over multiple adjacently located regulatory elements.

Chromatin domain localization

In the first stage of the algorithm (Figure 5.1.B), we predicted the locations of large-scale active and repressive chromatin domains, *i.e.* facultative euchromatic and heterochromatic regions. An example of several adjacent euchromatic and heterochromatic domains on chromosome 2 can be seen in Figure 5.2. The predicted euchromatic domains are enriched with the active histone modification H3K79me1, and the predicted heterochromatic domains are enriched with the repressive modification H3K27me2. The same pattern can be seen when all 34 active and all 5 repressive modifications are considered together (Figure D.1). In this example, we also observe higher Pol II binding and RNA-seq expression levels in the predicted euchromatic domains than seen for the predicted heterochromatic domains (Figure 5.2), consistent with the expectation that euchromatin is more actively transcribed than heterochromatin. Furthermore, predicted euchromatic domains genome-wide have significantly higher average CD4^{+} T cell expression levels than the predicted heterochromatic domains

(Figure 5.3; Mann-Whitney *U* test *P*<1E-10). The observations on expression levels serve to validate the maximum segment algorithm we use to delineate active (euchromatic) and repressive (heterochromatic) domains based on the analysis of histone modification data alone.

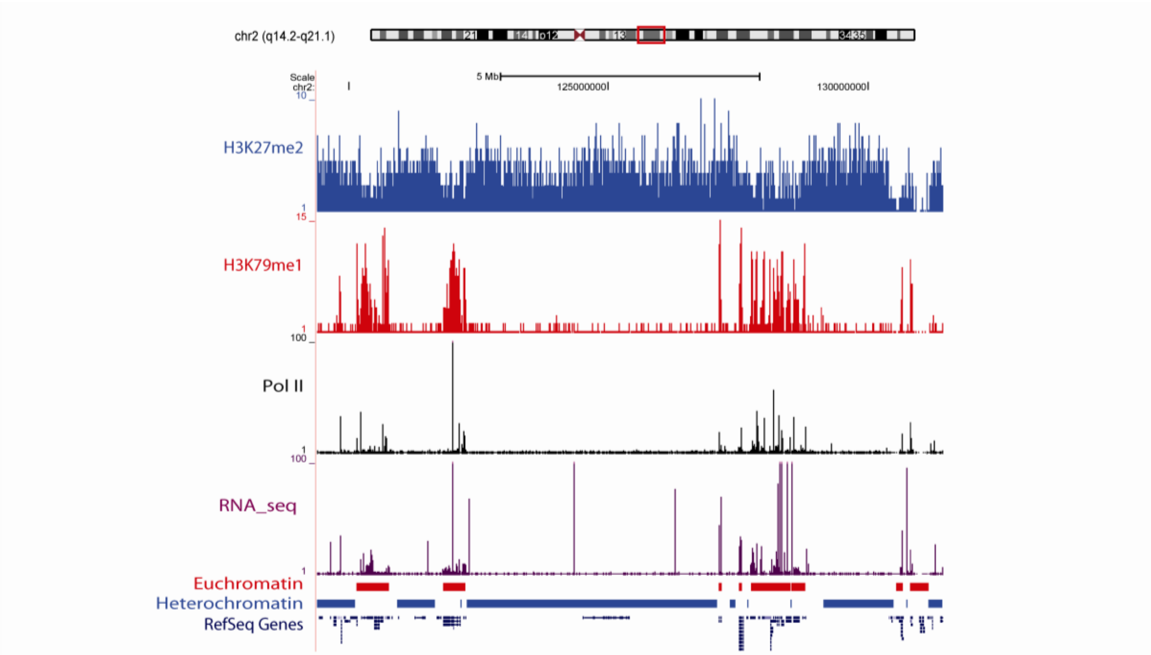


Figure 5.2: Example of predicted chromatin domains. An ideogram of chromosome 2 shows the cytogenetic banding pattern along with the location of this specific example. The distributions of ChIP-seq tag mapping peaks for the active histone modification H3K79me1 (red bars), the repressive histone modification H3K27me2 (blue bars), Pol II binding (black bars) and RNA-seq tags (purple bars) are shown in separate tracks. The predicted euchromatic domains (red bands) and heterochromatic domains (blue bands) are shown in the tracks denoted as 'Euchromatin' and 'Heterochromatin'. The locations of RefSeq Genes are shown below the chromatin domains.

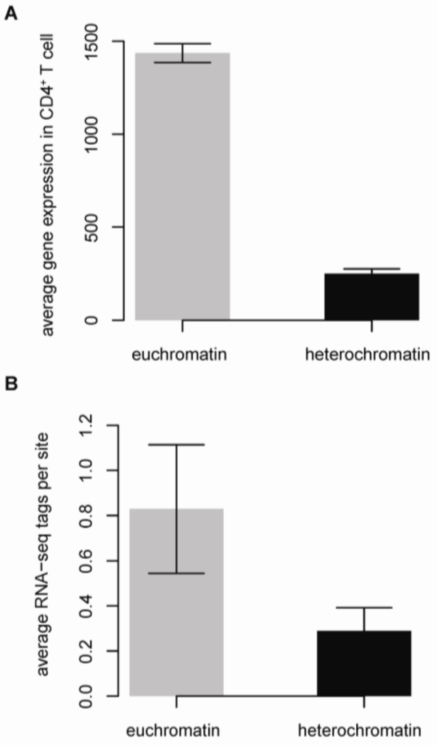


Figure 5.3: Validation and analysis of predicted chromatin domains. A: Average human CD4+ T cell expression levels for genes located in predicted euchromatic domains (grey bar) and heterochromatic domains (black bar). B: Average RNA-seq tags per site in CD4+ T cell for predicted euchromatic domains (grey bar) and heterochromatic domains (black bar).

We also used Gene Ontology (GO) and KEGG pathway analyses to interrogate the functional relevance of the euchromatic and heterochromatic domains predicted with our algorithm. Genes found in predicted euchromatic domains are enriched with functional terms and pathways related to CD4$^+$ T cell functions, such as defense response (GO), systemic lupus erythematosus (KEGG) and antigen processing and presentation (KEGG) (Table D.2).

Boundary element prediction

Application of the two-stage maximal segment algorithm and HMM approach (Figure 5.1) to the CD4$^+$ T cell ChIP-seq data resulted in the identification of 2,542 putative chromatin boundary elements. Sequence features of these boundary elements are summarized in Table D.1. It should be noted that our prediction method is not mechanistically biased in the sense that it does not rely on any previously known features of boundary element sequences, *e.g.* CTCF protein binding (122,131), the presence of tRNA genes (53) or the expression of non-coding RNAs originating from SINE repeats (56,79). By predicting boundaries in this way, without regard to previously known features, we can evaluate the associations of putative boundaries with such features *a posteriori* and, more importantly, look for novel boundary element related features, which may be indicative of as yet unknown boundary element mechanisms.

Examples of three predicted chromatin boundaries are shown in Figure 5.4; the locations of the boundaries are compared to the locations of the chromatin domains defined by active and repressive histone modification distributions along with the locations of CTCF binding, Pol II binding and RNA-seq expression levels. All of these boundaries are located close to the edges of borders between adjacent chromatin domains and at sharp transition points of Pol II binding and RNA-seq levels. The two boundaries shown in Figure 5.4.A are co-located with CTCF binding sites. The boundary shown in Figure 5.4.B shows a similar chromatin profile to those in Figure 5.4.A but is not related to CTCF binding. More detailed illustrations of these boundaries showing all of the individual histone modifications can be found in Figures D.2 & D.3.

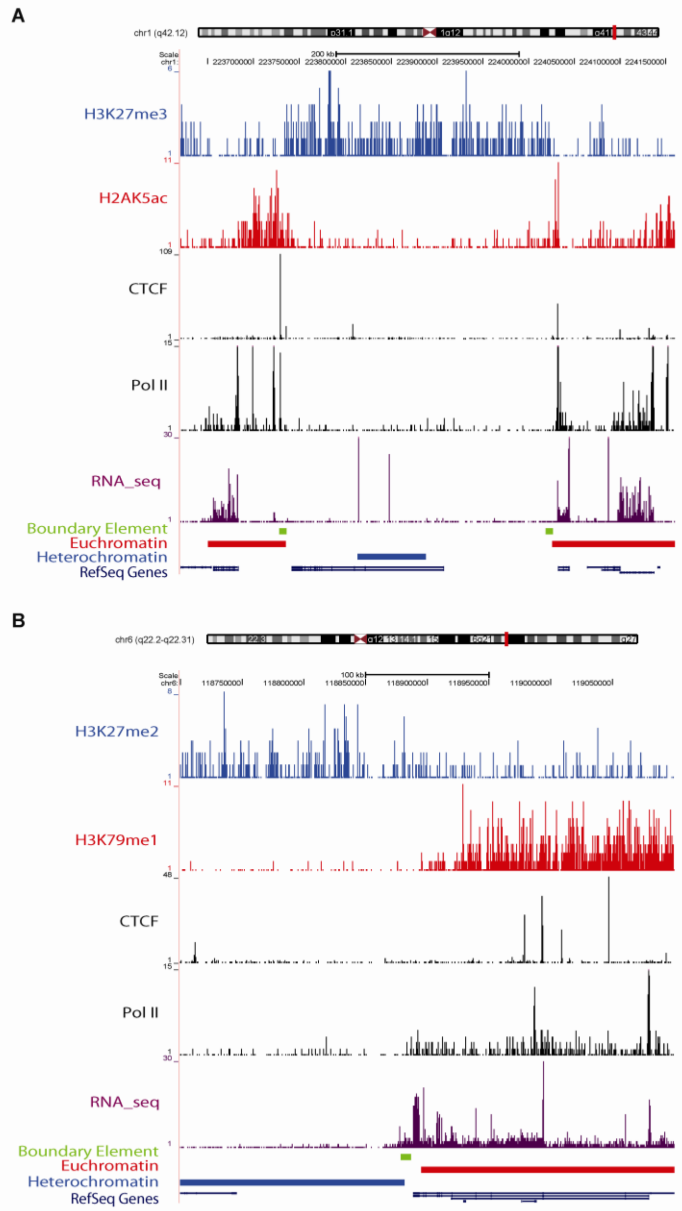


Figure 5.4: Examples of predicted chromatin boundary elements. A: Examples of predicted boundary elements with CTCF binding. B: Example of a predicted boundary element without CTCF binding. The predicted boundary elements are shown as green bands. ChIP-seq peaks for active and repressive histone modifications, CTCF binding, Pol II binding and RNA-seq tags along with the locations of euchromatic domains, heterochromatic domains and RefSeq genes are illustrated as separate tracks (as in Figure 5.2).

In order to test the relevance of the predicted chromatin boundaries to facultative chromatin and cell-type specific gene regulation, we compared the expression level differences for pairs of genes located on immediately opposing sides of the boundaries for CD4$^+$ T cells to their expression level differences among a set of 78 different human tissues and cell types (99). If the predicted boundary elements do in fact represent CD4$^+$ T cell specific regulatory elements that help to establish facultative chromatin domains, then the expression level differences of gene pairs that flank the boundaries should be greater for CD4$^+$ T cells than for other tissue-types. Consistent with this expectation, gene pairs that flank the predicted boundaries have significantly greater expression level differences in CD4$^+$ T cells than in other tissues and cell-types (Figure 5.5; Mann-Whitney *U* test *P*<1E-10).

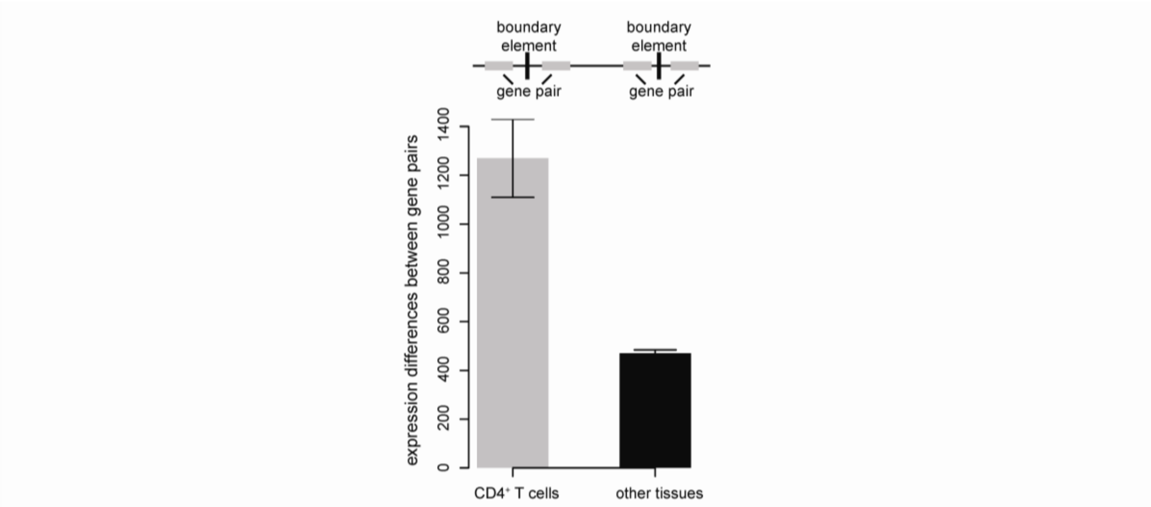


Figure 5.5: Expression differences between gene pairs that flank boundary elements. Expression differences of gene pairs located on immediately opposing sides of predicted boundary elements are shown for CD4+ T cells (grey bar) and 78 other human tissues together (black bar).

In an attempt to further evaluate the potential functional significance of the boundaries predicted here, we searched for overlaps between the predictions and previously experimentally characterized boundaries. Among the few known boundaries that have been functionally verified, only one boundary element, the BEAD-1 element, was identified in human T cells. BEAD-1 is a ~2kb region located between the divergently transcribed Vδ3 and TEA gene segments within the T cell receptor α/δ locus, and it has been shown to have enhancer-blocking activity (132). BEAD-1 is located within a RIT defined by our algorithm and overlaps one of the predicted boundary elements (Figure 5.6 and Figure D.4). Previously, the BEAD-1 sequence was shown to have a CTCF binding site and its enhancer blocking activity was found to be CTCF dependent in an erythroleukemia cell line (51). However, there is no evidence for CTCF binding of BEAD-1 from the genome-wide ChIP-seq analysis of CD4$^+$ T cells (3) suggesting that boundary element activity at this locus may be CTCF independent in some conditions.

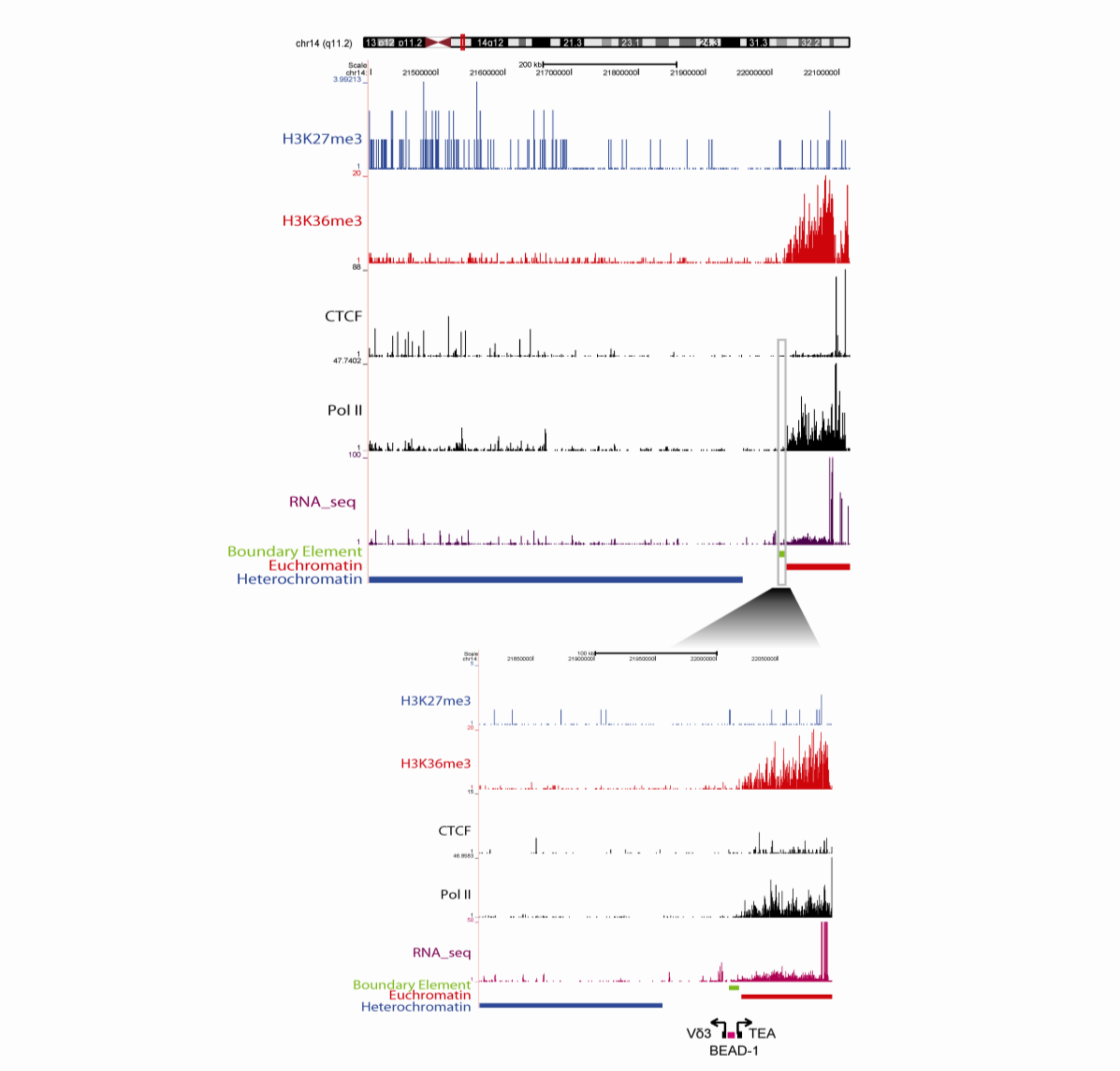


Figure 5.6: Co-location of a predicted boundary element with BEAD-1. A boundary element predicted by our method (green band) is shown to overlap with the experimentally characterized BEAD-1 boundary element (purple band). The BEAD-1 element is located between the Vδ3 and TEA gene segments (black boxes) of the T cell receptor α/δ locus. ChIP-seq peaks for active and repressive histone modifications, CTCF binding, Pol II binding and RNA-seq tags along with the locations of euchromatic domains, heterochromatic domains are illustrated as separate tracks (as in Figure 5.4). The inset shows greater detail at the BEAD-1 locus.

Chromatin features of predicted boundaries

The boundary element predictions reported here are based solely on chromatin states inferred from histone modifications and Pol II binding and do not rely on any previously characterized features of boundary element sequences. Since boundary elements are known to have diverse mechanisms of action (43,44,48,133), we analyzed our predicted boundaries for enrichment with a number of previously characterized boundary features and also with respect to as yet unknown features that may suggest novel mechanisms of boundary element activity.

We evaluated the chromatin environment of predicted boundaries using enrichment analysis of a number of genome-scale chromatin data sets. To do this, the 2,542 predicted boundary element regions were co-oriented and center aligned in such a way as to observe 8kb boundary element regions flanked by 8kb heterochromatic and euchromatic regions respectively. Predicted boundary elements show marked enrichment for DNase I hypersensitivity consistent with an open chromatin environment (Figure 5.7.A). Twelve histone acetylation marks all show similar peaked patterns of enrichment over predicted boundaries compared to flanking heterochromatic and euchromatic regions, suggesting that the predicted boundary elements are specifically acetylated to facilitate opening of the chromatin and recruitment of sequence-specific DNA binding factors (Figure 5.7.B).

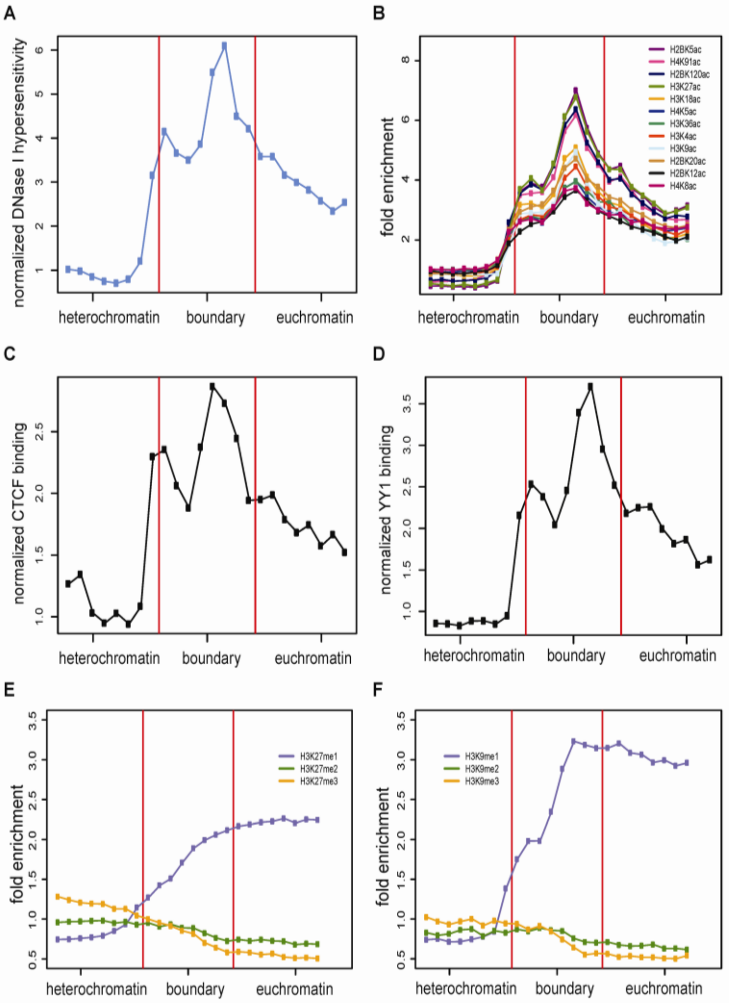


Figure 5.7: Chromatin signatures of predicted boundary elements. A-F: 8kb boundary regions are shown together with 8kb flanking heterochromatic and euchromatic regions. Normalized levels of DNase I hypersensitivity (A), fold enrichment profiles of 12 histone acetylations (B), normalized levels of CTCF binding (C), normalized levels of YY1 binding (D), fold enrichment profiles of H3K27 mono-, di- and tri-methylations (E) and fold enrichment profiles of H3K9 mono-, di- and tri-methylations (F) are compared for flanking regions and boundaries.

Levels of binding for the CTCF insulator protein are also elevated in predicted boundary element regions compared to adjacent heterochromatic and euchromatic regions (Figure 5.7.C). Thus, the apparent acetylation activity at predicted boundary elements may be recruited by specific protein factors such as CTCF. The importance of

CTCF in establishing chromatin regulatory domains recently was underscored by results indicating that numerous functional CTCF binding sites are constitutively occupied among different cell types, and more remarkably, conserved among syntenic regions in the human, mouse and chicken genomes (94). However, it should be noted that only a minority of predicted boundary elements (777 or 30.6%) contain CTCF binding sites, suggesting that at some of the predicted boundaries acetylation events occur in a CTCF independent manner or may be indicative of the recruitment of different DNA-binding factors.

Table 5.1: Protein factors enriched in predicted boundary elements.

Protein	**No.**[1]	***P*-value**[2]	**Annotations**[3]
EVI1	382	0.022	Interacts with histone deacetylase, histone methyltransferases and CBP and P/CAF
CEBP	249	2.27E-17	Interacts with CBP and p300 and promotes histone acetylation
YY1	157	1.44E-17	Directs histone deacetylases and histone acetyltransferases to promoter
CREBP1	150	5.87E-24	Essential in H2B and H4 acetylation, can interact with CBP HAT domain
USF	140	2.50E-28	Recruits histone modifications at vertebrate boundary elements

[1] The number of boundary elements containing the corresponding protein factor binding sites.

[2] The statistical significance of the enrichment of the protein factor in predicted boundary elements assessed by hypergeometric test.

[3] Functional annotations for the proteins based on the relevant literature (cited in the text).

We used the conserved TFBS data from the UCSC Genome Browser (70,125) to search for protein binding sites that are significantly enriched among the set of predicted chromatin boundaries. There are a number of significantly enriched TFBS that interact with proteins directly or indirectly involved in chromatin remodeling events (Table 5.1). For example, EVI1, CEBP, CREBP1, USF and YY1 are all involved in chromatin re-modeling via their interactions with chromatin modifying enzymes such as HAT, HDAC and HMT (134-140). In addition, the transcription factor USF has previously been implicated as mediating chromatin boundary element activity (47,141). The presence of distinct TFBS often overlap at individual boundaries indicating that a number of predicted boundaries have common binding sites (Figure D.5).

Inferences on protein binding based on the presence of TFBS are prone to false positives (although the use of conserved sites greatly mitigates this possibility) and also do not yield information on cell-type specific binding. For these reasons, we searched for ChIP-seq data sets from CD4$^+$ T cells to validate the TFBS observed to be enriched at our predicted boundaries with experimentally characterized cell-type specific binding events. There are CD4$^+$ T cell ChIP-seq data for YY1 (142), and analysis of these data reveal that the predicted boundaries are significantly overrepresented for YY1 binding (n=918; $P \leq 10^{-16}$ hypergeometric test), and YY1 binding peaks at boundaries relative to adjacent chromatin (Figure 5.7.D). Interestingly, there are far more boundaries bound by YY1 (n=918) than boundaries with conserved YY1 TFBS (n=157). This may be due to the presence of lineage-specific or non-canonical YY1 binding site motifs among the predicted boundaries. Consistent with observations that YY1 is a cofactor of CTCF for X-chromosome inactivation (143), there is a highly significant overlap between boundaries bound by CTCF and YY1 (n=534; $P \leq 10^{-113}$ hypergeometric test) suggesting the possibility of synergistic action between these two factors. Nevertheless, there remain 384 boundaries with YY1 binding only suggesting CTCF-independent mechanisms of action. For example, evidence showing that YY1 can interact with both HDAC and HAT (144-150) led to a potential model proposing that YY1 can activate or repress transcription via changing the local chromatin environment (148). YY1 was also shown to be able to interact with components of nuclear matrix (151,152), which may also facilitate partitioning of active and repressive chromatin domains.

The specific methylation status, mono- di- or tri-methylation, of the H3K27 and H3K9 histone marks show divergent trends across predicted boundary elements containing regions and adjacent heterochromatic and euchromatic regions (Figure 5.7.E and Figure 5.7.F). H3K27 and H3K9 mono-methylation (H3K27me1 and H3K9me1) levels increase steadily from facultative heterochromatic domains across boundary element containing regions and into euchromatic domains. On the other hand, di- and tri-methylation of the same residues (H3K27me2, H3K27me3, H3K9me2 and H3K9me3) gradually decrease from heterochromatin through the boundary element regions to euchromatin.

A number of other histone methylation marks, along with non-protein-coding RNA-seq accumulation, also show steadily increasing levels across boundary element regions from facultative heterochromatin to euchromatin (Figure 5.8.A and Figure 5.8.B), consistent with a gradual opening of the chromatin. However, all of the modifications of histone H3K4 analyzed here (H3K4me1, H3K4me2, H3K4me3 and H3K4ac) show distinct peaks over the predicted boundaries relative to flanking heterochromatic and euchromatic regions (Figure 5.8.C). These particular histone modifications have been associated with promoter and/or enhancer activity, suggesting that boundary element mechanisms may be related to initiation of transcription (43), in the case of promoters, and/or perturbation of the local chromatin environment, as has been suggested for enhancers (44). The enrichment profiles of all histone modifications could be found in Figures D.6, D.7 and D.8.

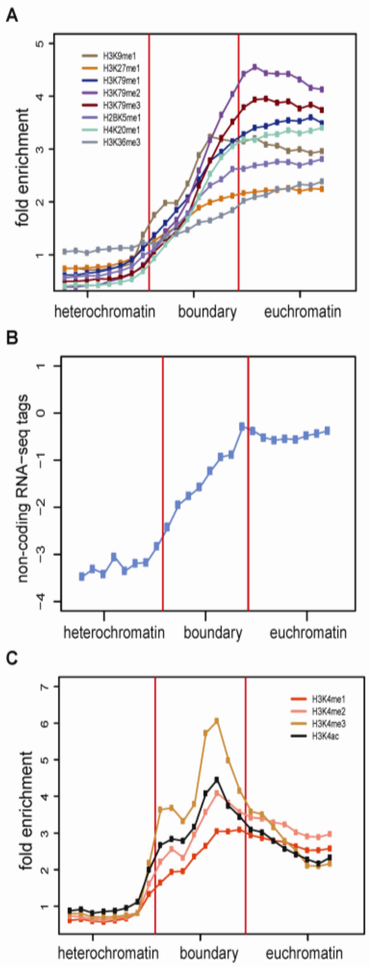


Figure 5.8: Chromatin and transcriptional transitions across predicted boundary elements. A-C: 8kb boundary regions are shown together with 8kb flanking heterochromatic and euchromatic regions. Fold enrichment profiles of 8 histone methylations (A), log2 transformed normalized non-protein-coding RNA-seq tags (B) and fold enrichment profiles of H3K4 histone modifications (C) are compared for flanking regions and boundaries.

Transcriptional interference at predicted boundaries

Transcription of non-coding RNA has been shown to be important for boundary element function from yeast to higher eukaryotes (52,53,56,79). Therefore, we analyzed RNA-seq data from CD4$^+$ T cells in order to evaluate whether our predicted boundaries are transcriptionally active (95). Across the predicted boundary elements, RNA-seq levels increase steadily with the transition from heterochromatin (low levels) to euchromatin (high levels) (Figure 5.8.B). Interestingly, a subset of 77 predicted boundary elements contain annotated non-coding RNA genes (126,127) and show distinct peaks of RNA accumulation relative to the adjacent chromatin domains (Figure 5.9.A), which coincide with Pol III binding (Figure 5.9.B). The RNA-seq peaks indicate that these particular boundary locations are transcribed at markedly higher levels than genomic background consistent with a role for transcriptional interference.

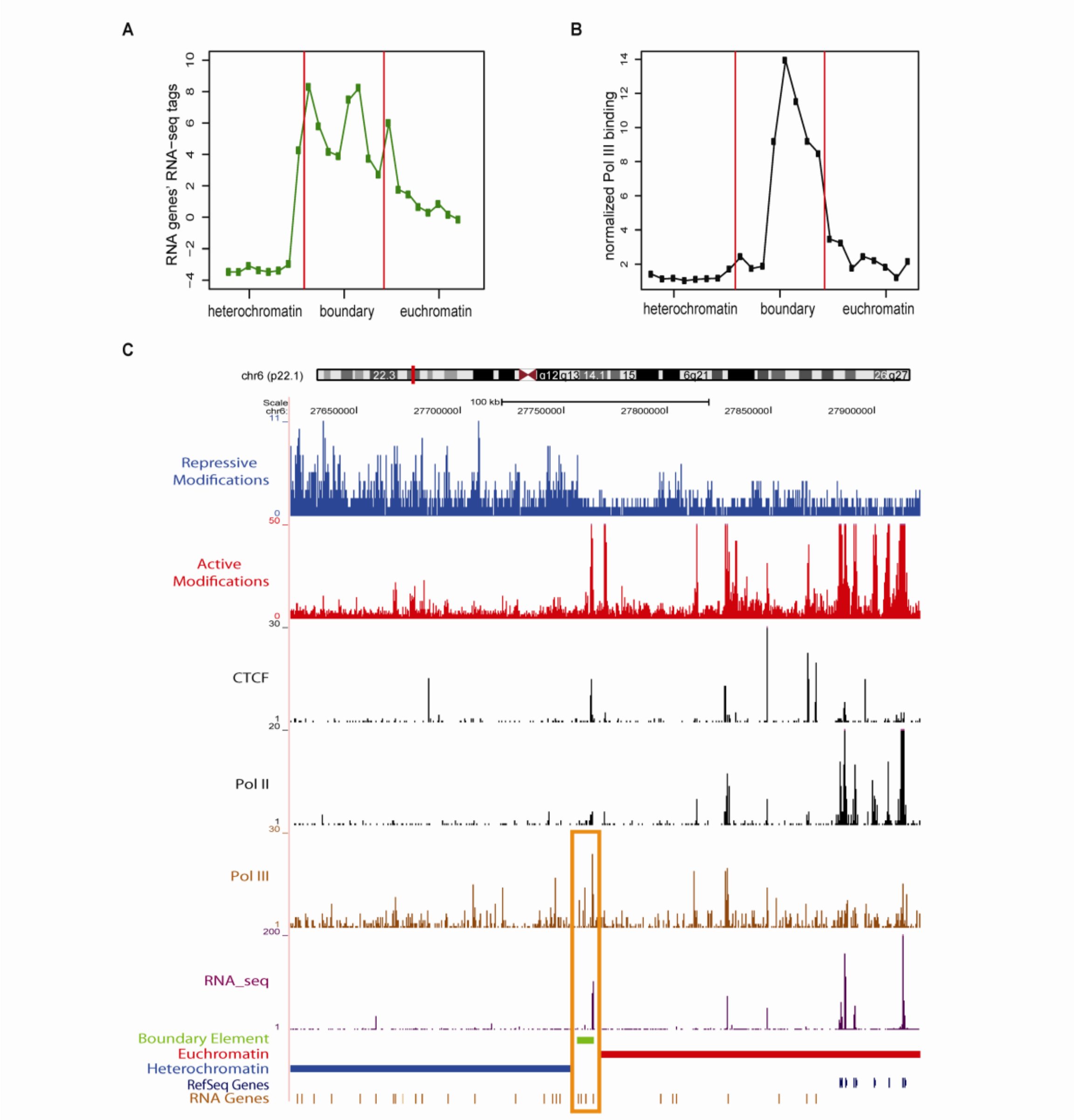


Figure 5.9: Features of boundary elements containing RNA genes. A-B: 8kb boundary regions are shown together with 8kb flanking heterochromatic and euchromatic regions. Log2 transformed normalized RNA-seq tags (A) and normalized Pol III binding levels (B) of boundary elements containing RNA genes are compared for flanking regions and boundary regions. C: Example of boundary element containing tRNA genes. The predicted boundary element is shown as the green band. ChIP-seq peaks for active and repressive histone modifications, CTCF binding, Pol II binding and RNA-seq tags along with the locations of euchromatic domains, heterochromatic domains and RefSeq genes are illustrated as separate tracks (as in Figure 5.2). Pol III binding (yellow bars) and RNA genes are also shown separately.

Figure 5.9.C shows an example of a predicted boundary element that contains a cluster of 4 tRNA genes along with peaks of RNA-seq expression and Pol III and CTCF binding, suggesting a possible relationship between CTCF binding and tRNA gene transcription. The example shown in Figure 5.9.C suggests that, similar to yeast, tRNA genes in the human genome may operate as genomic boundaries, although definitive assessment of their functional significance awaits further experimental analysis. Consistent with this prediction, clusters of mouse tRNA genes have been shown to encode chromatin barrier activity (153).

Discussion

A chromatin based approach to unbiased boundary element prediction

Boundary elements are known to organize chromatin into functionally distinct domains and to prevent regulatory cross-talk between domains. Distinct boundary elements may act through a variety of mechanisms, and accordingly boundaries have been characterized phenotypically based on their activity rather than the presence of characteristic features. Thus, boundary element prediction algorithms that use pattern detection methods to search for known boundary element characteristic features will result in biased sets of predictions that only reflect one or another of the known mechanisms of action. This fundamental challenge to the computational prediction of boundary elements motivated our development and application of an unbiased algorithm that predicts the locations of putative boundary elements genome-wide based on their functional consequences, with respect to both chromatin and transcription states, as opposed to any intrinsic characteristics that they may possess.

Our approach to boundary element prediction relies on the delineation of adjacent active (euchromatic) and repressive (heterochromatic) domains based on the genomic distributions of active versus repressive histone modifications. Regions in transition (RITs) between adjacent chromatin domains are further interrogated for the presence and location of putative boundaries using distributions of Pol II binding sites that serve as marks of active cell-type specific transcription. Application of this two-stage chromatin boundary element prediction algorithm (Figure 5.1) to CD4$^+$ T cell data resulted in the prediction of 2,542 boundary elements across the human genome. The role of these predicted boundary elements in cell-type specific chromosomal domain organization was confirmed by the finding that genes immediately flanking boundaries are more highly differentially expressed in CD4$^+$ T cells than seen for other human cells/tissues (Figure 5.5). Having predicted boundary elements in this way, we then analyzed the putative boundaries for the presence of a variety of features that may yield specific clues as to their potential mechanisms of action.

Models for human boundary element activity

Previous studies on boundary elements have suggested competing models that explain the mechanisms underlying boundary element activity. The fixed model for boundary element activity implicates specific DNA sequences and their associated proteins, whereas the transcriptional interference model emphasizes the role of transcription from non-protein-coding transcriptional units. We have previously noted that these two models are not necessarily mutually exclusive (43). Under the fixed model, boundaries are precisely located and contain specific sequences that form discrete physical barriers between domains. Specific sequence features are also needed to recruit

Pol II and Pol III machineries for the transcriptional interference model, and transcriptional units that act as boundaries may also form physical barriers that block the propagation of repressive chromatin. The features uncovered for our predicted boundary elements can similarly be taken to suggest that the mechanisms of human boundary activity include aspects of both the fixed and transcriptional interference models.

Analysis of the predicted boundary elements and surrounding RITs revealed two main features: 1) position-specific acetylation and open chromatin coincident with the recruitment of transcription factors such as EVI1, YY1 and USF (Figure 5.7.A, 5.7.B & 5.7.D; Table 5.1), and 2) a gradual transition across RITs, from heterochromatin to euchromatin, of increasing histone lysine methylation and non-protein-coding RNA levels (Figure 5.7.E & 5.7.F; Figure 5.8.A & 5.8.B). Considered together, these two observations lead us to propose a possible model for human boundary element activity (Figure 5.10). Under this model, the specific positions of boundaries are established via the local recruitment of histone acetyltransferase (HAT) activity and transcription factors leading to the expression of non-protein-coding RNAs (Figure 5.10.A). Boundary element function is maintained more broadly across RITs by the superposition of distinct and opposing chromatin modifying activities leading to the observed gradual transitions between heterochromatic and euchromatic histone lysine methylation and mediated by transcriptional interference (Figure 5.10.B).

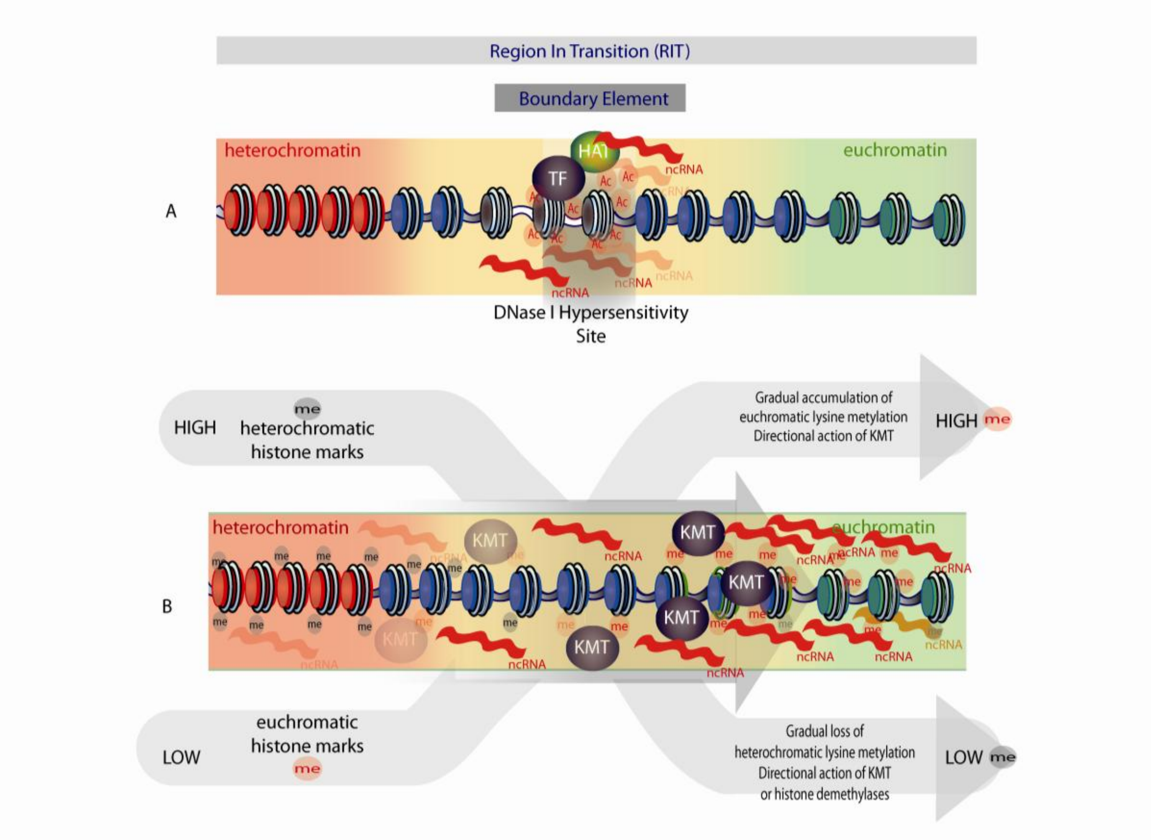

Region In Transition (RIT)
Boundary Element
heterochromatin
euchromatin
A
HAT
TF
Ac
ncRNA
DNase I Hypersensitivity Site
HIGH
me
heterochromatic histone marks
Gradual accumulation of euchromatic lysine metylation
Directional action of KMT
HIGH me
heterochromatin
euchromatin
B
KMT
me
ncRNA
euchromatic histone marks
LOW
Gradual loss of heterochromatic lysine metylation
Directional action of KMT or histone demethylases
LOW me

Figure 5.10: Model for human chromatin boundary element activity. Broader regions in transition (RITs) from heterochromatin to euchromatin are shown along with more precisely located boundary elements. A: Boundary element locations are characterized by position-specific open chromatin environment (DNase I Hypersensitivity Site) and hyperacetylation (Ac) that are coincident with the recruitment of transcription factors (TF) and non-protein-coding RNA transcription. B: RITs are characterized by gradual changes in the levels of histone methylation (me) from heterochromatin to euchromatin.

Predicted boundary elements reside in regions of distinctly open chromatin and also show position-specific accumulations of 12 different histone acetylation marks (Figure 5.7.A and 5.7.B). Previous studies have suggested boundary element activity is dependent upon the local recruitment of histone acetyltransferase activities to counteract the spread of repressive chromatin (141,154,155). The patterns of histone lysine acetylation enrichment observed at position-specific location within predicted boundaries are in agreement with already reported prominent role for histone acetylation at boundary elements and further corroborate the boundary prediction method used here.

Along with the position-specific chromatin features and recruitment seen at predicted boundaries, we also observe distinct chromatin dynamics spread across the RITs that lie between adjacent facultative heterochromatic and euchromatic domains. For instance, H3K27 and H3K9 mono-methylation levels increase steadily from heterochromatic domains across boundary element containing regions and into euchromatic domains, whereas H3K27 and H3K9 di- and tri-methylation levels gradually decrease across the same intervals (Figure 5.7.E and 5.7.F). This pattern can be taken to indicate a unidirectional activity of histone demethylation across RITs from heterochromatin to euchromatin. At the same time, a number of other mono- di- and tri-methylation histone marks show steady accumulations across RITs from heterochromatin to euchromatin (Figure 5.8.A) and are indicative of increased transcriptional activity (Figure 5.8.B) and/or the action of chromatin modifying enzymatic complexes associated with transcriptional elongation.

H3K79 mono- di- and tri-methylation all show progressively increasing levels across RITs from facultative heterochromatin to euchromatin (Figure 5.8.A). While the exact function of H3K79 methylation is currently unknown, accumulation of these marks, catalyzed by the lysine methyltransferase (KMT) DOT1 (156), is correlated with actively transcribed protein-coding genes (157). Accordingly, it is possible that H3K79 methylation also marks active transcription of non-protein-coding RNAs across RITs as observed here (Figure 5.8.A & 5.8.B). In fact, H3K79 methylation has previously been implicated in the stable maintenance of distinct chromatin states in yeast and mammalian cells (158), and our data also suggests a possible, and previously unexplored, role for DOT1 in the establishment and maintenance of chromatin boundaries.

Transcriptional regulators at predicted boundary elements

The observations that predicted boundary elements contain binding site motifs for a number of proteins implicated in both the regulation of transcription and chromatin remodeling (Table 5.1), along with experimentally characterized YY1 binding (Figure 5.7.D), are consistent with a role for transcriptional interference in human boundary element activity. Involvement of transcription factors capable of maintaining a local active chromatin environment at boundaries has previously been reported by the Felsenfeld group in the context of the USF1 factor (47). USF transcription factors can regulate Pol II transcription via direct interaction with components of the basic transcription machinery, such as TFIID and TBP associated factors (159), or through the recruitment of co-factors such as the histone acetyltransferase PCAF or the H3K4 histone methyltransferase SET7/92 (141). Here, we observe a significant enrichment of the USF binding site motif (E-box element) among predicted boundaries. Thus, we speculate that USF participates in the establishment and/or maintenance of human boundary element activity by triggering transcriptional interference, which may be mediated, at least in part, by the action of the aforementioned co-factors.

EVI1 is another sequence-specific transcription regulator with binding sites that are over-represented among the boundary elements predicted here (Table 5.1). EVI1 has been shown to interact with the histone acetyltransferase PCAF, the histone deacetylase HDAC1 and the histone methyltransferases SUV39H1 and G9A (134,135). Thus, we speculate that EVI1 may function in boundary element activity by serving as a switch between distinct chromatin remodeling activities thereby mediating the transition from heterochromatin to euchromatin in a cell-type dependent manner.

Conclusions and prospects

Chromatin boundary elements are major players in genome organization and regulation, but at this time there are relatively few examples of known boundary elements. Here, we report a large collection of putative boundary elements for CD4$^+$ T cells that span the entire human genome. The boundaries reported here are computational predictions and thus must be treated with all due caution; nevertheless, analysis of the features of these boundaries yields results that are consistent with their roles as chromatin related regulatory elements. We hope that the boundaries predicted here can serve as a prioritized list of targets for further experimental validation. If validated experimentally, the predictions reported here could help to substantially enlarge the catalog of known chromatin boundary elements. Our feature analysis of the predicted boundaries also raises the possibility of a mechanism of chromatin boundary activity in the human genome related to transcriptional interference. This possibility awaits further detailed investigations.

Acknowledgments

We would like to acknowledge members of the Jordan and Lunyak labs along with Lluis Montoliu for useful discussions. This work was supported by an Alfred P Sloan Research Fellowship in Computational and Evolutionary Molecular Biology [BR-4839 to J.W. and I.K.J.], a pilot program grant from the Georgia Tech Integrative BioSystems Institute [to J.W. and I.K.J.], and the Buck Institute Trust Fund [to V.V.L.].

CHAPTER 6

CHROMATIN SIGNATURE DISCOVERY VIA HISTONE MODIFICATION PROFILE ALIGNMENTS

Abstract

We report on the development of an unsupervised algorithm for the genome-wide discovery and analysis of chromatin signatures. Our Chromatin-profile Alignment followed by Tree-clustering algorithm (ChAT) employs dynamic programming of combinatorial histone modification profiles to identify locally similar chromatin sub-regions and provides complementary utility with respect to existing methods. We applied ChAT to genomic maps of 39 histone modifications in human CD4$^+$ T cells to identify both known and novel chromatin signatures. ChAT was able to detect chromatin signatures previously associated with transcription start sites and enhancers as well as novel signatures associated with a variety of regulatory elements. Promoter associated signatures discovered with ChAT indicate that complex chromatin signatures, made up of numerous co-located histone modifications, facilitate cell-type specific gene expression. The discovery of novel L1 retrotransposon associated bivalent chromatin signatures suggests that these elements influence the mono-allelic expression of human genes by shaping the chromatin environment of imprinted genomic regions. Analysis of long gene associated chromatin signatures point to a role for the H4K20me1 and H3K79me3 histone modifications in transcriptional pause release. The novel chromatin signatures

and functional associations uncovered by ChAT underscore the ability of the algorithm to yield novel insight on chromatin based regulatory mechanisms.

Introduction

Histone proteins are subject to a variety of covalent modifications, including methylation, acetylation, phosphorylation and ubiquitylation. The identities and locations of these histone modifications have profound effects on the structure and regulatory properties of eukaryotic chromatin (21). Indeed, over the last several years specific genomic regulatory elements, such as promoters, enhancers and boundary elements have been associated with distinct combinatorial patterns of histone modifications (3,5,8-11,18,23,28,33,95). The discovery and characterization of such combinatorial histone modification patterns, or chromatin signatures as they are often referred to, can provide valuable information with respect to the location and activity of cell-type and developmentally-specific genomic regulatory features (7,13,14,17,22,25,31,34,160).

Next-generation sequencing based technologies, chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq) in particular, provide an opportunity for the systematic analysis of combinatorial histone modification patterns genome-wide (19,161). Computationally, the inference of combinatorial histone modification signatures is a pattern recognition problem in high-dimensional space. There are currently two classes of computational approaches designed for this purpose: supervised and unsupervised methods. Supervised methods identify histone modification signatures characteristic of a pre-defined set of known genomic features, *e.g.* promoters or enhancers (9,27,28,34). Regulatory element characteristic combinatorial modification patterns identified in this way can then be used to query the genome to identify the locations of additional regulatory elements of the same kind. The use of supervised methods in this way was critically important for the discovery that specific genomic regulatory elements bear distinct chromatin signatures. However, supervised methods are unsuited for the discovery of novel histone modification patterns that may be associated with as yet unknown regulatory activities. Unsupervised methods do not rely on training data sets derived from previously annotated features, and as such they have the potential to discover the kinds of unknown chromatin signatures that characterize novel regulatory elements. Here, we are more interested in the unsupervised approach to the analysis of chromatin given the potential this approach holds for novel discoveries.

There are a number of available unsupervised algorithms for the analysis of histone modification patterns. The program ChromaSig utilizes probabilistic profiles that are characteristic of specific histone modification patterns (29,30). The CoSBI algorithm applies a biclustering method to search for regions with common histone modification patterns (32). Hidden Markov Model (HMM) based methods are widely used to segment eukaryotic genomes into various combinatorial chromatin states with distinct histone modification profiles (6,7,62). The most recently developed method of this kind, Segway, employs Dynamic Bayesian Networks (DBN) to achieve greater precision for the detection of known regulatory elements along with superior accommodation of missing data (162).

We have developed an unsupervised algorithm for analysis of combinatorial histone modification patterns that extends the capabilities of existing methods in a number of ways. First, our method does not apply any restriction to the size of co-located histone modification patterns. Second, our method does not utilize any motif seed to initialize the subsequent inference of histone modification patterns. Third, our method is capable of detecting histone modification patterns with multiple modes, *e.g.* co-located signatures made up of constituent individual modifications that are spatially shifted with respect to one another. Fourth, our method is capable of detecting co-located signatures composed of alternating segments with conserved and variant combinatorial patterns. Fifth, our method discriminates between chromatin signatures composed of the same histone modifications but with different shapes. Sixth, our method provides an inherent statistical criterion that allows related chromatin signatures to be classified into distinct groups, and thereby delineates the total number of patterns observed in any data set. The first four features described above distinguish our method from the ChromaSig and CoSBI programs. The fifth feature provides added utility beyond what is available for the HMM and Segway methods, and the sixth statistical feature is uniquely implemented in our approach.

We call our method ChAT, for Chromatin-profile Alignment followed by Tree-clustering, and we applied this approach to the genome-wide analysis of 39 histone modifications characterized by ChIP-seq analysis of human CD4$^+$ T cells (3,18). Application of ChAT on this data set resulted in the discovery of chromatin signatures previously shown to be characteristic of specific genomic regulatory elements along with a number of novel chromatin signatures and features that point to as yet unexplored chromatin related regulatory mechanisms. We report these discoveries in light of the design and implementation of the ChAT algorithm, with an emphasis on comparison to existing methods. The ability of the ChAT algorithm to discern combinatorial histone modification patterns previously observed to be associated with known regulatory elements serves as proof of its utility for the discovery of functionally relevant chromatin signatures. The characterization of previously undiscovered chromatin signatures and functional associations with ChAT supports the potential utility of the algorithm to yield novel biological insight.

Materials and methods

General scheme of the ChAT algorithm

The ChAT algorithm analyzes genome-wide histone modification data sets produced via ChIP-seq in order to characterize distinct chromatin signatures. ChAT is an unsupervised algorithm; its use does not require any training set based on pre-defined genomic annotations such as the locations of promoters, enhancers or transcription factor binding sites. There are three major steps in the ChAT algorithm: 1) ChIP-seq data transformation, 2) dynamic programming on histone modification profiles, and 3) hierarchical clustering of genomic regions that correspond to related chromatin signatures (Figure 6.1).

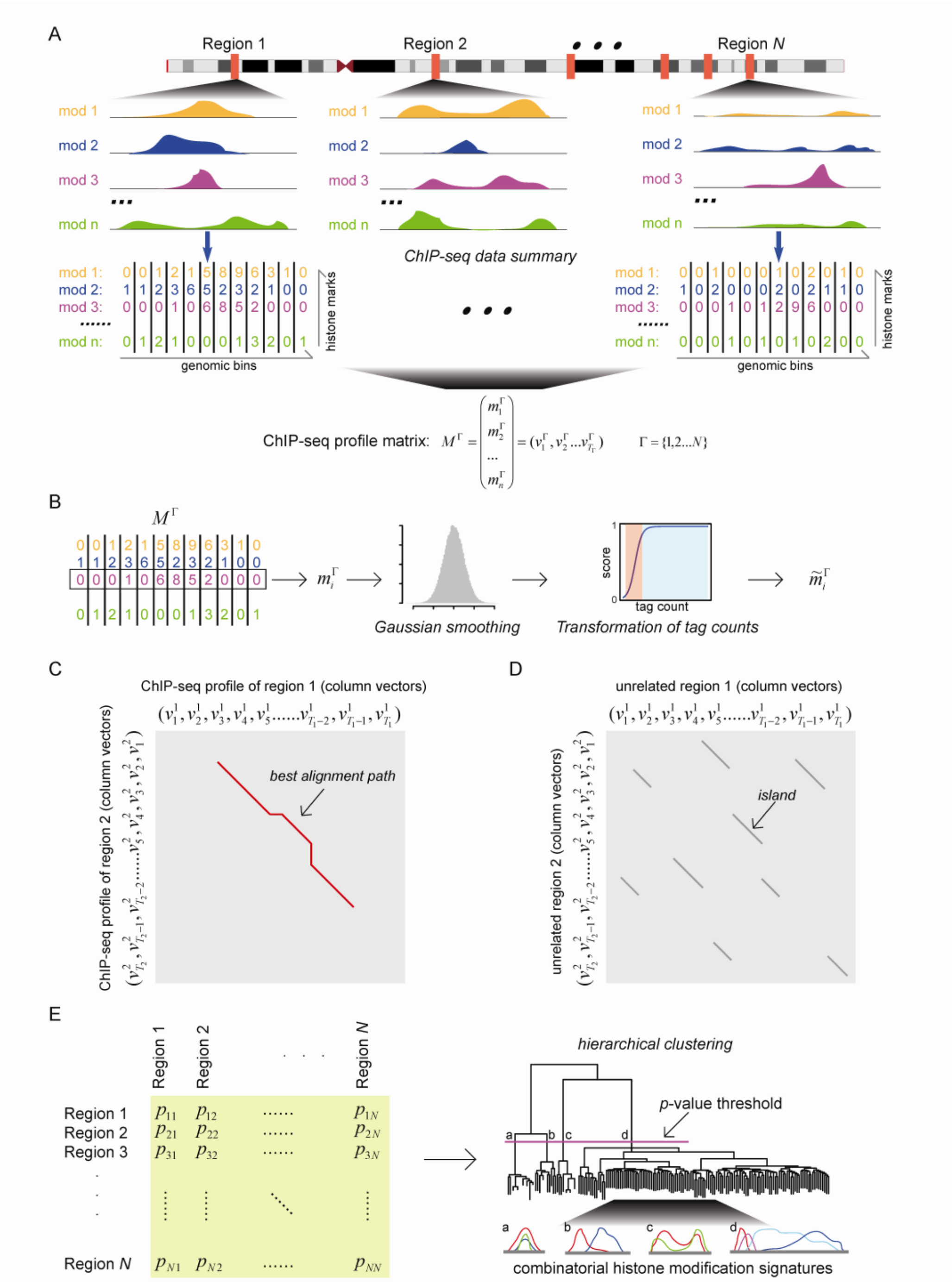

A
Region 1
Region 2
Region N
mod 1
mod 2
mod 3
mod n
mod 1: 0 0 1 2 1 5 8 9 6 3 1 0
mod 2: 1 1 2 3 6 5 2 3 2 1 0 0
mod 3: 0 0 0 1 0 6 8 5 2 0 0 0
mod n: 0 1 2 1 0 0 0 1 3 2 0 1
genomic bins
histone marks
ChIP-seq data summary
mod 1: 0 0 1 0 0 0 1 0 2 0 1 0
mod 2: 1 0 2 0 0 0 2 0 2 1 1 0
mod 3: 0 0 0 1 0 1 2 9 6 0 0 0
mod n: 0 0 0 1 0 1 0 1 0 2 0 0
genomic bins
histone marks
ChIP-seq profile matrix: $M^\Gamma = \begin{pmatrix} m_1^\Gamma \\ m_2^\Gamma \\ \dots \\ m_n^\Gamma \end{pmatrix} = (v_1^\Gamma, v_2^\Gamma \dots v_{T_\Gamma}^\Gamma)$ $\Gamma = \{1,2 \dots N\}$
B
M^Γ
m_i^Γ
Gaussian smoothing
score
1
0
tag count
Transformation of tag counts
$\widetilde{m}_i^\Gamma$
C
ChIP-seq profile of region 1 (column vectors)
$(v_1^1, v_2^1, v_3^1, v_4^1, v_5^1 \dots\dots v_{T_1-2}^1, v_{T_1-1}^1, v_{T_1}^1)$
ChIP-seq profile of region 2 (column vectors)
$(v_{T_2}^2, v_{T_2-1}^2, v_{T_2-2}^2 \dots\dots v_5^2, v_4^2, v_3^2, v_2^2, v_1^2)$
best alignment path
D
unrelated region 1 (column vectors)
$(v_1^1, v_2^1, v_3^1, v_4^1, v_5^1 \dots\dots v_{T_1-2}^1, v_{T_1-1}^1, v_{T_1}^1)$
unrelated region 2 (column vectors)
$(v_{T_2}^2, v_{T_2-1}^2, v_{T_2-2}^2 \dots\dots v_5^2, v_4^2, v_3^2, v_2^2, v_1^2)$
island
E
Region 1
Region 2
Region N
Region 1 p_{11} p_{12} ...... p_{1N}
Region 2 p_{21} p_{22} ...... p_{2N}
Region 3 p_{31} p_{32} ...... p_{3N}
Region N p_{N1} p_{N2} ...... p_{NN}
hierarchical clustering
p-value threshold
a
b
c
d
combinatorial histone modification signatures

Figure 6.1: Scheme of the ChAT algorithm. (A) For a series of genomic regions, combinatorial histone modification distributions are represented by ChIP-seq profile matrices. (B) Histone modification ChIP-seq tag counts are smoothed and transformed to produce normalized scores. (C) Dynamic programming is used to identify sub-regions with similar chromatin signatures. (D) Pairwise p-values are computed based on a null distribution of high-scoring chromatin segment pairs (islands) found between unrelated genomic regions. (E) Pairwise p-values are organized into a distance matrix that is used for hierarchical clustering of similar chromatin sub-regions.

ChIP-seq data transformation

The genome is divided into 200bp non-overlapping bins, and for each bin arrays of ChIP-seq signals (*i.e.* tag counts) for all histone modifications in the data set are computed. In this way, combinatorial histone modification profiles are represented as a matrix, where the column vectors correspond to combinatorial histone modification tag counts within individual genomic bins and the row vectors correspond to the contiguous genomic landscape of individual histone modifications (Figure 6.1.A). Then for each individual histone modification (*i.e.* each row vector), Gaussian smoothing is applied to remove noise resulting from spurious tag counts in the ChIP-seq experiments (Figure 6.1.B). The resulting smoothed ChIP-seq tag counts for each histone modification are transformed to a score between 0 and 1 for all subsequent analysis (Figure 6.1.B).

The transformation is: $sc = 1 \Big/ 1 + \tau \cdot e^{-\eta \frac{t - T_i}{T_i}}$, where sc is the transformed score and t is the smoothed tag count. T_i is the genomic median of tag counts of histone modification i. The transformation is performed for two reasons. First, the vast majority of bin tag counts for each histone modification are very small (*e.g.* 1 or 2 tags), and the transformation allows such regions to be effectively excluded from subsequent analysis. Second, large differences between high bin tag count values (*e.g.* 100 versus 150 tags) can bias subsequent alignment steps, and the transformation allows the magnitude of such differences to be dampened.

Having quantified and transformed ChIP-seq histone modification tag count signals in this way, the algorithm then divides the genome into discrete genomic regions

(Figure 6.1.A) by delineating contiguous regions that contain high ChIP-seq signals for at least one histone modification from intervening regions that do not contain any such signal. The intervening genomic regions that do not contain any high ChIP-seq signal are excluded from subsequent analysis, and the contiguous genomic regions with high ChIP-seq signal are taken as discrete units for subsequent alignment and chromatin signature analysis. To do this, consecutive genomic bins with high ChIP-seq signals ($sc > 0.5$) are first merged into a single region, and regions which are close to each other (<4kb) are further merged together. Importantly, at this step no size threshold or limit for contiguous regions is used. This allows the algorithm to characterize chromatin signatures across a wide range of genomic sizes. In addition, consecutive bins do not need to be enriched with the same histone modification in order to be merged. This allows the algorithm to characterize chromatin signatures with spatially shifted patterns of individual histone modifications.

To make the algorithm more computationally efficient, individual genomic regions with similar histone modification profiles are grouped together prior to profile alignment with dynamic programming. This grouping is achieved via a simple two-step clustering procedure. First, genomic regions are checked for presence or absence of a set of user-defined histone modifications (*e.g.* H3K4me3, H3K27ac, H3K27me3 and H3K36me3), and regions are grouped together if they contain the same sets of these modifications. This step reflects the fact that regions which differ with respect to the presence/absence of critical user-defined histone modifications are unlikely to have similar chromatin signatures. Second, genomic regions are further grouped into three size categories: small (≤5kb), medium (>5kb and <10kb) and large (≥10kb). This initial grouping greatly reduces the number of pairwise profile alignments needed to be performed. It also allows for intelligent user input with respect to the coherence of functionally related (*e.g.* active versus repressive) histone modifications.

Dynamic programming on histone modification profiles

For every pair of genomic regions within the same group, local pairwise alignment of transformed histone modification profile matrices is performed using dynamic programming. The dynamic programming approach entails a number of advantages: it does not require any prior chromatin signature motif seed, it guarantees optimal local alignments that can include gaps, it allows for the discovery of chromatin signatures of vastly different sizes, and it allows for the calculation of *p*-values that quantitatively measure chromatin signature similarities between genomic regions.

To perform dynamic programming, the transformed histone modification profile matrix of each discrete genomic region is considered as a string of column vectors and a modified cosine similarity is used as the score to measure the similarity between each pair of column vectors (Figure 6.1.C). For example, the column vector for bin i of the first region (region 1) of a pair under comparison is denoted as v_i^1. Each entry of this column vector corresponds to the transformed score for the level of a specific histone modification, *e.g.* v_{ik}^1 is the value for the kth histone modification in bin i. Similarly, the vector for bin j of the second region (region 2) of a pair under comparison is denoted as v_j^2 and v_{jk}^2 is the value for the kth histone modification in bin j. The raw score for the similarity between v_i^1 and v_j^2 is calculated as: $\tilde{s}_{ij} = \cos(f \cdot \arccos(\frac{v_i^1 \bullet v_j^2}{|v_i^1||v_j^2|}))$.

The factor f is an amplification factor ($1 < f < 2$) that enlarges the angle between v_i^1 and v_j^2. The value of $\tilde{s}_{ij}$ is more likely to be negative with higher values of f and accordingly the two bins will have lower probability of being aligned. Thus, increasing the value of f will cause the alignment to be more stringent. Here, f is set to 2 for small sized region comparisons in order to focus on highly similar sub-regions and is set to 1.5 for medium and large size comparisons.

The raw score is further multiplied by a weight factor to calculate the final score for v_i^1 and v_j^2. The final score is $s = w \cdot \tilde{s}_{ij}$ and the weight factor is related to $m_{ij} = \min\{|v_i^1|, |v_j^2|\}$. The relation between w and m_{ij} is $w = 1 - e^{-m_{ij}/\sigma}$. Thus vectors with small norms are given small weight; the rationale being that vectors with small norms have low levels of ChIP-seq signals and therefore should contribute less to the final signatures even if they are very similar with each other. σ is used to control the stringency of the weight factor. Larger values of σ result in smaller weights, and accordingly only genomic regions with abundant ChIP-seq signals will be aligned. Here, σ is set as 0.3.

The gap penalty is designed to be proportional to the vector norm. For example, the gap penalty of aligning v_i^1 to a gap is $g_i^1 = k \cdot |v_i^1|$. The gap penalty scheme is designed such that it highly penalizes the alignment of vectors with large norms (*i.e.* high levels of ChIP-seq signals) to gaps. The parameter k is used to control the stringency of the alignment, and it is designed to be larger for small size region comparisons and smaller for medium and large size comparisons. The introduction of gaps using this scheme enables the discovery of multi-modal chromatin signatures, particularly for large-sized signatures that often contain combinations of conserved and variant segments.

Having parameterized the dynamic programming algorithm in this way, it is then used to search for the most similar sub-regions between pairs of transformed histone modification matrices representing discrete genomic regions. Each entry of the alignment matrix for dynamic programming is: $c_{i+1,j+1} = \max\{c_{i,j} + s_{ij}, c_{i+1,j} - g^{2}_{j+1}, c_{i,j+1} - g^{1}_{i+1}, 0\}$, and $c_{i,0} = 0, c_{0,j} = 0$. Each pair of regions is compared twice: in the same and in the opposite orientations. In this way, sub-regions with the highest combinatorial histone modification profile similarities will be found.

P-values are calculated to quantify the similarities between genomic sub-regions aligned in this way (Figure 6.1.D). To do this, the algorithm employs the island method, based on the extreme value distribution of high-scoring segment pairs, originally developed for DNA sequence comparisons (72). This method creates a null distribution of random similarity scores, against which the observed similarity scores can be compared in order to compute *p*-values for aligned pairs of sub-regions. To create the null distribution of random similarity scores, pairs of unrelated genomic regions are randomly sampled from the entire set of regions under consideration. Then for each pair of unrelated regions, dynamic programming with the same parameter settings is applied and all high-scoring islands of similarity, with scores above a threshold t, are retained. Using those high-scoring islands, the parameters K_t and λ_t for the extreme value distribution are estimated as suggested by Altschul et al (163), and finally the *p*-value is calculated as: $p \approx 1 - e^{-K_t mne^{-\lambda_t x}}$.

Hierarchical clustering of related chromatin signatures

All *p*-values for pairwise profile alignments are organized into a pairwise distance matrix, and hierarchical clustering is applied on this matrix (Figure 6.1.E). In this way, sub-regions with the same combinatorial histone modification signatures will be grouped together and the branch lengths among them in the hierarchical tree will be shorter. Furthermore, since *p*-values are used as pairwise distances, the branch lengths can be viewed as approximate *p*-values among sub-groups or clusters. Then, for a given *p*-value threshold (*e.g.* 0.05), the hierarchical tree divided by this threshold will yield clusters of related sub-regions at user-defined levels of statistical confidence (Figure 6.1.E). Cluster-characteristic combinatorial histone modification signatures can then be derived.

Chromatin signature feature enrichment analysis

Chromatin signatures discovered via the application of ChAT to genome-wide histone modification data sets are evaluated for the enrichment over annotated genomic features (*e.g.* promoters and enhancers) using a fold enrichment (FE) criterion: $FE = p / q$, where p is the fraction of the patterns overlapping with specific genomic features, and q is the fraction of the specific genomic feature in the genome. Here, an FE threshold of 3 was taken to indicate that a given chromatin signature is enriched over a particular genomic feature. The features analyzed include TSS (8kb sequences centered on the transcription start sites of Refseq gene models), TTS (8kb sequences centered on the transcription termination sites of Refseq gene models), enhancers (CD4$^+$ T cell specific p300 binding sites) (164) and CD4$^+$ T cell DNase I hypersensitive sites (4).

Results

The ChAT algorithm for chromatin signature discovery

As its name implies, the ChAT algorithm analyzes genome-wide maps of histone modifications characterized by ChIP-seq studies via a process of Chromatin-profile Alignment followed by Tree-clustering. To do this, chromatin profiles are represented as numeric matrices with transformed scores for each histone modification along the genomic sequence (Figure 6.1.A and 6.1.B). Alignment of these profiles is performed using an implementation of the local dynamic programming algorithm, which allows for the detection of genomic sub-regions with shared chromatin profiles (Figure 6.1.C). Dynamic programming also allows for the introduction of gaps in the chromatin profile alignments. Gaps are critical since they allow the algorithm to extend beyond regions with variant (or diffuse) chromatin enrichment signatures, and in so doing facilitate the discovery of chromatin signatures that span long genomic regions as well as those with complex multi-modal patterns of histone modification enrichment. For each resulting pairwise chromatin profile alignment, an approximate *p*-value is calculated (Figure 6.1.D), and hierarchical clustering is then applied on these pairwise values to organize genomic regions into related groups of chromatin signatures (Figure 6.1.E). The use of *p*-values for clustering allows for an inherent statistical criterion by which the hierarchical tree can be divided into groups of coherent chromatin signatures.

Application of ChAT to CD4+ T cell chromatin

We applied the ChAT algorithm to the analysis of genome-wide maps of 39 histone modifications characterized using ChIP-seq on human CD4$^+$ T cells (3,18) in an attempt to discover all discernible histone modification patterns. ChAT was run using the parameter values described in the Materials and Methods section, and a *p*-value threshold of 0.05 was used to partition the resulting hierarchical trees of patterns in order to explicitly delineate individual chromatin signatures. As stated previously, application of ChAT to ChIP-seq histone modification data sets does not require any restriction on the size of potential chromatin signatures or the use of motif seeds to initialize the search.

ChAT identified a total of 206 distinct combinatorial histone modification patterns genome-wide, which were subsequently grouped into small- (144), medium- (35) and large-sized (27) categories as explained in the Materials and Methods. Overall, the features of these observed chromatin signatures are consistent with the intended design of the algorithm and point to the additional utility provided by its use. For instance, we detected a number of large-sized patterns, ranging from 10kb – 100kb, which demonstrate the utility of allowing alternating conserved and variant segments in the detection scheme. We also find a number of signatures with multiple modes of histone modifications as well as spatially shifted patterns for individual constituent modifications. Combinatorial patterns that bear the same individual histone modifications with different relative profile shapes are recognized as distinct chromatin signatures.

Inspection of the small-sized patterns revealed that a substantial fraction of these signatures are associated with known regulatory features, such as TSS, TTS and p300 binding sites (Table E.1). 41.7% of the small-sized patterns are enriched with DNase I hypersensitive sites, using a fold enrichment threshold of 3 (FE>3), implying that they are located in open chromatin and possibly co-located with individual regulatory elements. In the following sections, we describe a number of the chromatin signatures discovered by ChAT, with an emphasis on the characterization of known regulatory features, which serve as a kind of positive control for the approach, along with descriptions of previously uncharacterized patterns that underscore the ability of the algorithm to facilitate novel discoveries.

TSS associated chromatin signatures

Since chromatin signatures around active TSS have been previously well-characterized (3,9), we searched for ChAT identified chromatin signatures that are co-located with annotated TSS in an attempt to evaluate the performance of the algorithm. There are 36 small-sized signatures that were found to be enriched at TSS (Table E.1; FE>3), and the common characteristic histone modifications of these patterns include the canonical TSS associated marks H3K4me3, H2AZ, H3K4me1 and H3K9me1 as well as a number of other combinations of histone acetylations, which are known active marks. Examples of several TSS associated signatures detected by ChAT are shown in Figure 6.2.

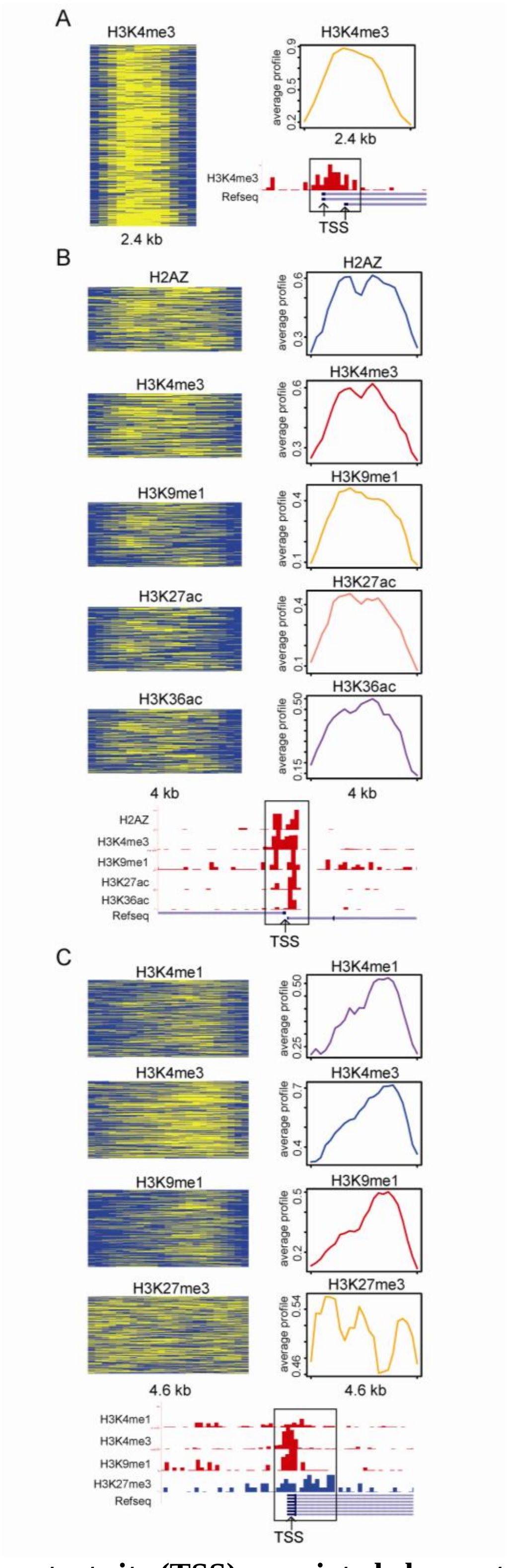


Figure 6.2: Transcription start site (TSS) associated chromatin signatures. (A) A TSS associated signature based on enrichment of H3K4me3. (B) A TSS associated signature composed of 5 active histone modifications. (C) A bivalent TSS associated signature with 3 active modifications and 1 repressive modification.

Figure 6.2.A shows the histone modification enrichment profile of the simplest TSS signature, which is characterized by H3K4me3 alone. In Figure 6.2.B, the TSS associated signature is shown to be enriched with 5 co-located active histone modifications. Interestingly, a number of bivalent TSS associated signatures were also found by ChAT. For example, the bivalent signature shown in Figure 6.2.C is characterized by 3 co-located active marks and a spatially shifted and multi-modal enrichment of the repressive mark H3K27me3. From the perspective of the ChAT algorithm design, the enrichment profiles of the bivalent signature example (Figure 6.2.C) illustrate the ability of the program to find patterns with multiple modes caused by shifted enrichments of different histone modifications.

Analysis of expression levels (99) in CD4$^+$ T cells for sets of genes with TSS marked by distinct signatures show that bivalent signatures are associated with lower gene expressions than seen for active signatures (p=4.1x10^{-4}, Mann-Whitney test) (Figure 6.3.A). Furthermore, the lower gene expression levels associated with bivalent signatures, and higher gene expression levels associated with active signatures, are specific to T cells and B cells compared with expression levels in other cell types (Figure 6.3.B). This observation indicates cell-type specific regulatory functions of distinct TSS associated combinatorial histone modification signatures discovered by ChAT for CD4$^+$ T cells.

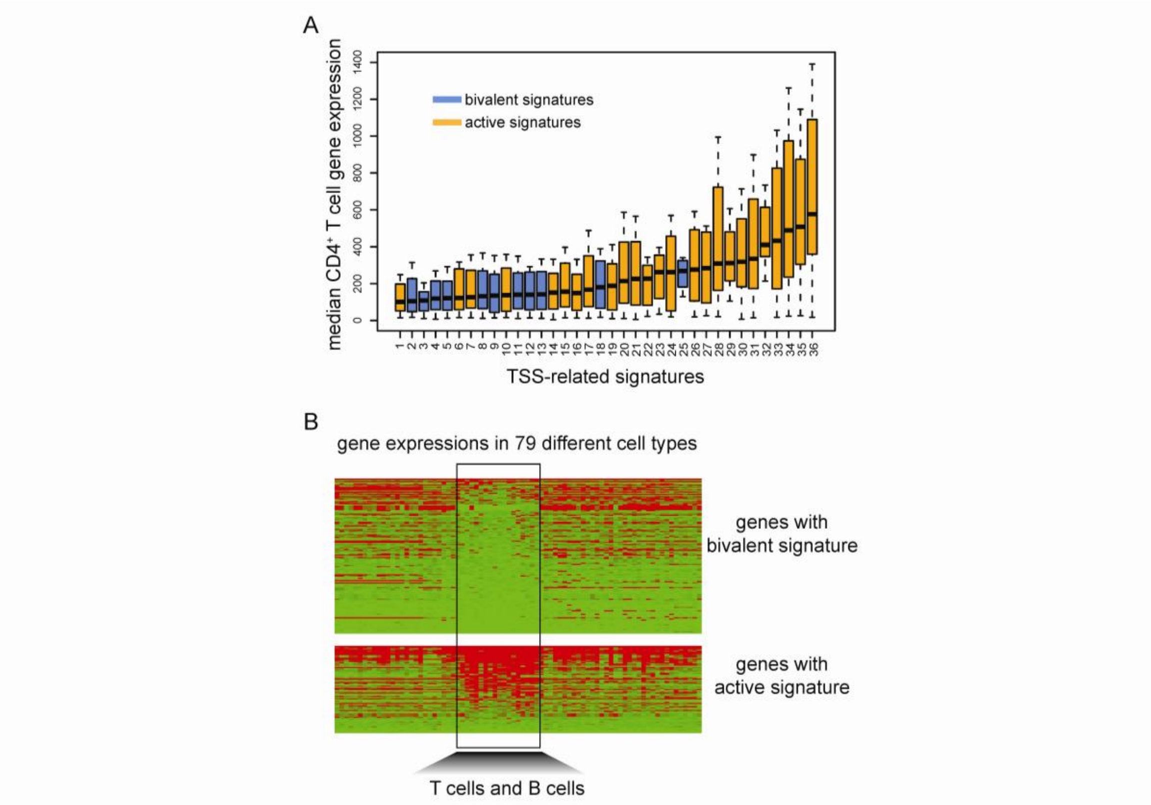


Figure 6.3: Differential gene expression associated with specific TSS chromatin signatures. (A) Median CD4+ T cell expression levels (+/- 1 quartile) of genes with TSS marked by 36 distinct chromatin signatures. Bivalent TSS signatures (blue bars) correspond to lower overall expression levels than active signatures (orange bars). (B) Cell-type specific gene expression patterns associated with different TSS chromatin signatures. Gene expression levels across 79 cell types (red=high and green=low) are shown for genes with TSS marked by a bivalent signature versus genes with TSS marked by an active signature. Expression differences are most pronounced for the indicated T cells and B cells.

We also observed that sets of genes with similar T or B cell expression levels can show very different TSS associated chromatin signatures. For instance, Figure 6.4.A shows two sets of genes with indistinguishable T or B cell expression levels (*p*=0.7, Mann-Whitney test), but different levels of expression (*p*=4.9x10^{-3}, Mann-Whitney test) across a panel of numerous other cell-types and tissues (99). In other words, the first set (s1) has a narrower cell-type specific expression pattern, whereas the second set (s2) shows broad expression over numerous cell-types and tissues (Figure 6.4.A). The chromatin signature for the set of cell-type specific genes (s1, Figure 6.4.B) is far more complex, being comprised of six different histone modifications, than the signature made up of two histone modifications seen for the set of broadly expressed genes (s2, Figure 6.4.C). This suggests the possibility that cell-type specific expression is regulated via a more complex chromatin promoter landscape. In fact, when all 36 of the TSS related chromatin signatures are evaluated, more complex signatures are found to be associated with gene sets that have higher T or B cell-type specific expression levels (Figure 6.4.D). The acetylation marks H3K36ac and H3K27ac in particular are associated with high levels of T or B cell-type specific expression.

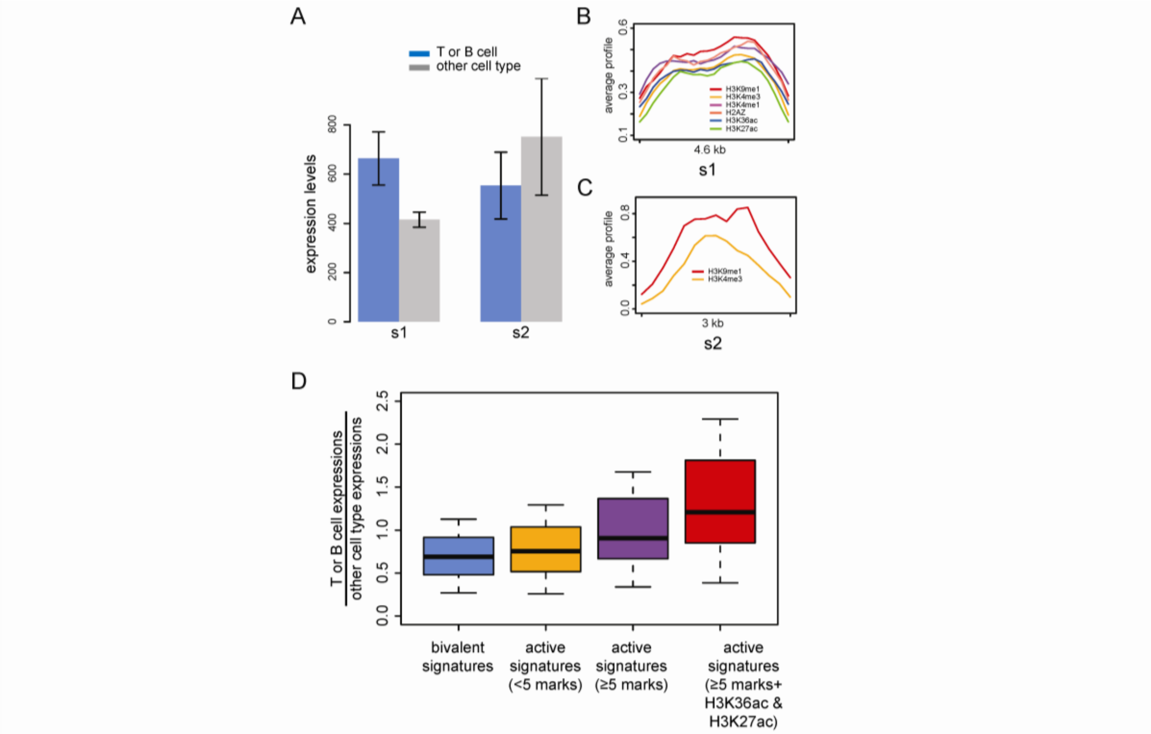


Figure 6.4: Cell-type specific expression associated with complex chromatin signatures. (A) Average (±sd) expression levels (blue-T or B cell expression, grey-other cell-type expressions) of genes with TSS marked by two different chromatin signatures (s1 and s2). (B) Enrichment profiles showing the average histone modification scores across signature s1. (C) Enrichment profiles showing the average histone modification scores across signature s2. (D) Box-plots showing T or B cell specific expression level distributions for different sets of chromatin signatures.

TTS associated chromatin signatures

The nature of chromatin signatures around TTS have not been previously characterized as well as those associated with TSS (6,15,162), and this may be due to a lack of coherence in the histone modification patterns found at gene termini. Nevertheless, ChAT was able to discern 9 small-sized patterns associated with TTS in CD4$^+$ T cells (Table E.1; FE>3). The common characteristic marks for these TTS signatures are quite distinct from those seen around TSS and include H2BK5me1, H4K20me1 and H3K27me1. Two examples of TTS associated signatures are shown in

Figure 6.5.A and 6.5.B. A single genomic region showing adjacent locations of each of these two signatures close to an annotated TTS is shown in Figure 6.5.C. Both of these TTS patterns are bi-modal with two enriched peaks linked by a relatively depleted central region. The relatively low levels of histone modifications seen in the central regions of these patterns may be related to specific protein binding events as has been suggested for the bi-modal patterns of enhancers (7). Consistent with this possibility, these same sets of regions show peaks of RNA polymerase II (Pol II) binding that corresponds to the locations of the depleted regions in the bi-modal patterns (Figure 6.5.D and 6.5.E). With respect to the ChAT algorithm design, the bi-modal patterns seen at TTS point to the utility of gaps in the chromatin profile alignments, which allow chromatin patterns to extend beyond variant regions and include multiple peaks of individual histone modifications.

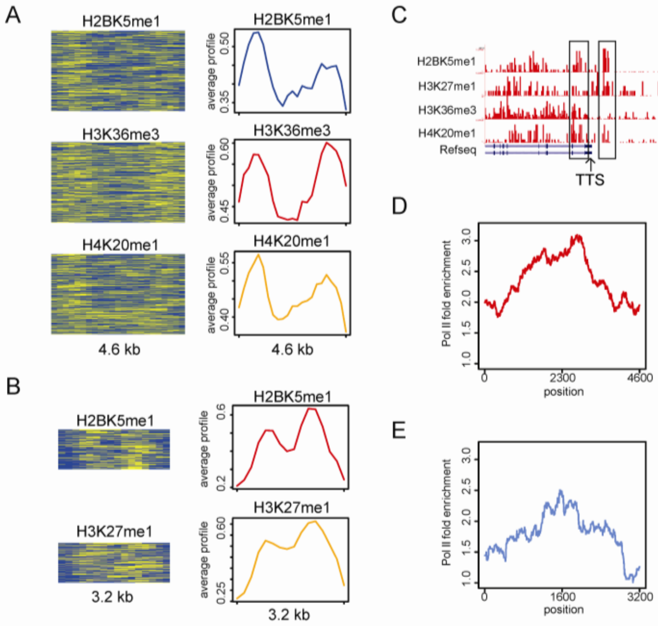


Figure 6.5: Transcription termination site (TTS) associated chromatin signatures. TTS signatures associated with three (A) and two (B) histone modification combinations are shown (histone modification representations described as for Figure 6.2). (C) A specific TTS proximal locus showing adjacent locations of each of these two patterns. (D) Pol II enrichment profile within genomic regions marked by the signature shown in panel A. (E) Pol II enrichment profile within genomic regions marked by the signature shown in panel B.

Enhancer associated chromatin signatures

Chromatin signatures characteristic of enhancers have been characterized in a number of studies (7,9,11,27-30), many of which rely on the positions of p300 binding sites to identify enhancer locations. We also took the locations of p300 binding sites (164) to indicate putative enhancers and found that ChAT characterized 18 small-sized signatures that are co-located with these sites (Table E.1; FE>3). The common characteristic marks of these patterns include the canonical enhancer associated marks H3K4me1 and H3K4me3 along several other histone acetylations (Figure 6.6.A). Examples of enhancer associated signatures detected by ChAT are shown in Figure 6.6.B and 6C; these two distinct signatures are characterized by similar sets of histone modifications with markedly different profile shapes, *i.e.* mono-modal (Figure 6.6.B) versus bi-modal (Figure 6.6.C). The different shapes of this kind discovered by ChAT may point to distinct dynamics of histone modifying enzymes and/or DNA binding proteins between the two sets of enhancers, indicative of the utility of the algorithm for discovering specific chromatin based regulatory mechanisms.

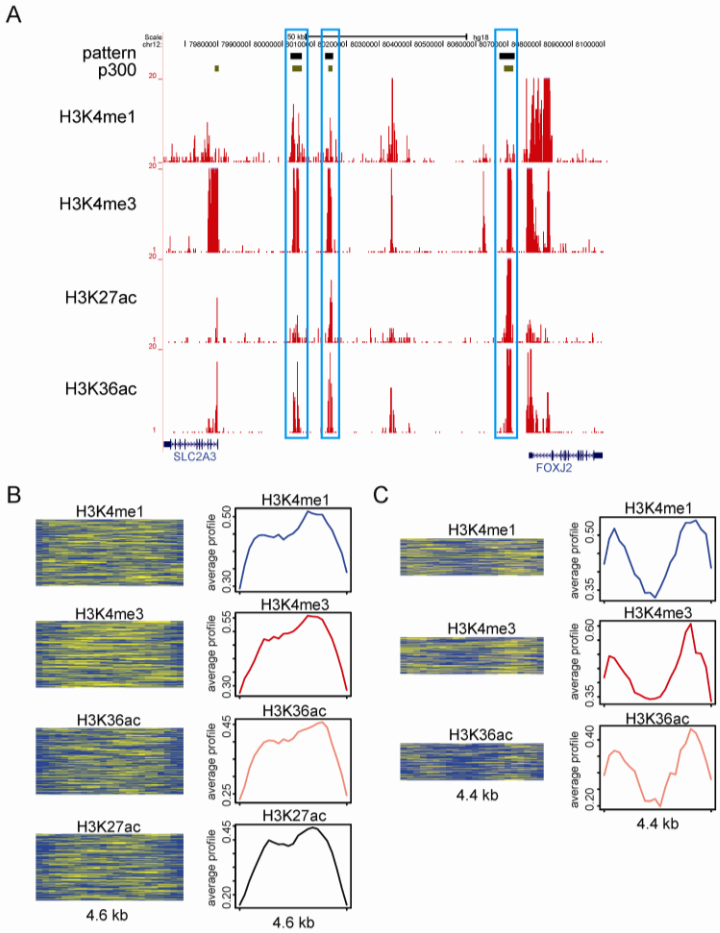


Figure 6.6: Enhancer associated chromatin signatures. A ~100kb genomic region with three locations marked by a specific signature composed of co-located peaks. (B) Histone modification enrichment profiles of an enhancer associated mono-modal signature. (C) Enrichment profiles of an enhancer associated bi-modal signature.

Conserved non-coding element associated chromatin signatures

Conserved non-coding elements (CNEs) are non protein-coding sequences that have been found to be anomalously conserved between species; CNEs are of interest because they are thought to correspond to regulatory regions that have been conserved by purifying selection based on their functional utility (165). We evaluated CNEs characterized via the comparison of genome sequences from 28 vertebrate species for the presence of chromatin signatures discovered with the ChAT algorithm and found that all 144 signatures show substantial overlap (FE>3) with the CNEs (Figure 6.7.A and Table E.1). This result is consistent with the presumed regulatory activity of CNEs. Not surprisingly, most of the CNE associated signatures are made up of active histone marks and tend to be associated with TSS or enhancers; such CNEs are likely to be active regulatory elements in CD4$^+$ T cells. However, a number of CNEs were also found to be associated with repressive chromatin signatures. For example, a simple chromatin signature made up of the repressive mark H3K27me3 (Figure 6.7.B) is highly enriched over CNEs (FE=18.4). We surmised that these CNEs may represent regulatory elements that are active in other cell-types but repressed in a specific manner in T or B cells. To evaluate this possibility, we checked the expression levels of the genes most proximal to these CNEs for their expression across 79 human tissues and cell-types (99). These genes do appear to be repressed in T or B cells in a cell-type specific manner, since they are expressed at higher levels across other cell types compared to T or B cells (Figure 6.7.C and 6.7.D).

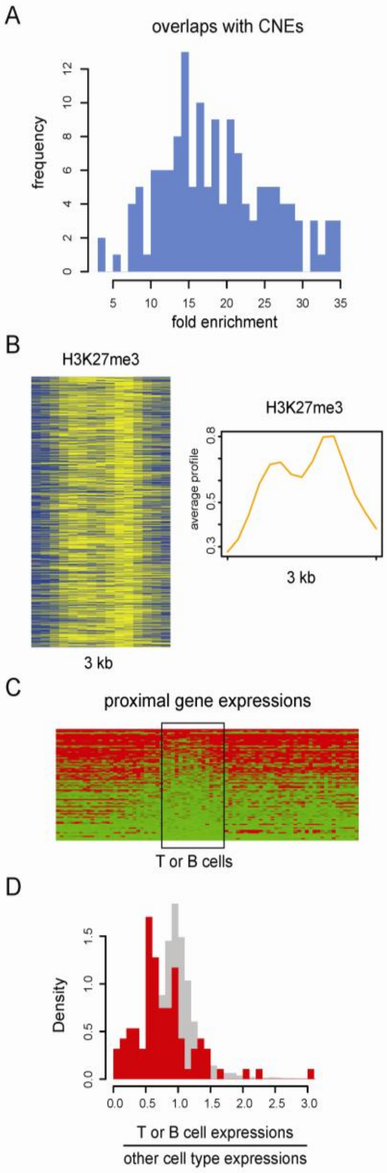


Figure 6.7: Conserved non-coding element (CNE) associated chromatin signatures. (A) Distribution of fold enrichments of CNEs for all small-sized signatures. (B) Histone modification enrichment profiles (as described for Figure 6.2) for a repressive signature highly enriched within CNEs. (C) Cell-type specific expression levels for genes proximal to CNEs bearing the repressive signature shown in panel B. (D) Distribution of the ratios of T or B cell average expressions and other cell type average expressions for genes shown in panel C (observed=red expected=grey). Observed ratios are significantly smaller than expected ratios calculated from gene expression levels randomly simulated across cell-types and tissues (p=1.3x10-10, Mann-Whitney test).

Bivalent chromatin signatures associated with L1 retrotransposons

Bivalent chromatin signatures, composed of co-located active and repressive histone modifications (12,16), have previously been associated with TSS sequences, and the ChAT algorithm was also able to detect such bivalent signatures at TSS in CD4$^+$ T cells (Figure 6.2.C and 6.3). Application of ChAT here revealed two bivalent signatures that were not found to be associated with TSS: H3K9me3 and H3K36me3 (Figure E.1) along with H3K4me3 and H3K9me3 (Figure 6.8.A). Interestingly, both of these bivalent signatures were found to be highly enriched within L1 retrotransposon sequences; 68.4% of the genomic regions marked by the H3K9me3-H3K36me3 signature overlap with L1 as do 77.0% of genomic regions marked by H3K4me3 and H3K9me3. A broad genomic region with several L1 encoded segments that overlap the H3K4me3-H3K9me3 signatures can be seen in Figure 6.8.B.

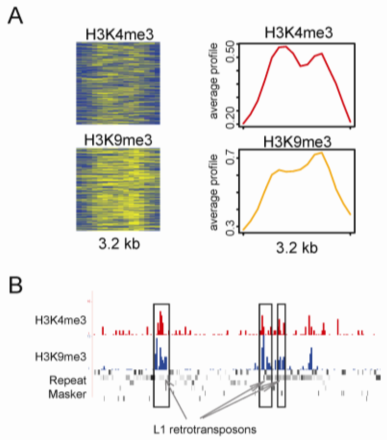


Figure 6.8: A bivalent chromatin signature associated with L1 retrotransposons. (A) Histone modification enrichment profiles (as described for Figure 6.2) for the bivalent signature. (B) A single genomic region with three locations marked by the L1 characteristic bivalent signature.

This particular bivalent pattern has previously been associated with imprinted genomic loci wherein genes tend to be expressed in a mono-allelic fashion based on the parent of origin for the allele (16). Interestingly, a number of studies have also shown that L1 retrotransposons are enriched in-and-around imprinted genomic loci (166-169). Thus, the enrichment of these bivalent signatures on L1 retrotransposons may point to a chromatin based mechanism by which L1 sequences contribute to the mono-allelic expression of human genes. On the other hand, such bivalent patterns may actually result from ChIP-seq analyses performed heterogeneous cell populations with the locations in some cells marked by active modifications and others with repressive modifications. In this case, the patterns revealed by the algorithm would represent an artifact of the ChIP-seq experimental design.

Large-sized chromatin signatures

The ChAT algorithm places no restriction on the size of chromatin signatures that it can identify, and we found 27 large-sized signatures in CD4+ T cells ranging from 10kb – 100kb in length. These large-sized chromatin signatures can be classified into two groups. The first group contains long contiguous co-located blocks of repressive marks, presumably representing heterochromatic or repressive chromatin domains. The second group shows more complex and potentially interesting patterns resembling the known H3K4me3-H3K36me3 domains, which are associated with gene bodies and long non-coding RNAs (3,8,170). For example, the signatures shown in Figure 6.9.A and 6.9.B (see also Figures E.2 and E.3) are characterized by the presence of similar active marks albeit over different size ranges. In both cases, the long chromatin signatures show punctate enrichments of several active marks at one end of the pattern together with broader enrichments of different active marks throughout the rest of the signature. These two large-sized signatures show substantial overlaps with gene bodies (Figure 6.9.C), suggesting the utility of ChAT for annotating genes.

However, while more than 90% of these two large-sized signatures do overlap with known gene bodies (Figure 6.9.D), there is still a small fraction which does not overlap with gene bodies. For example, Figure 6.9.E shows two specific genomic regions where the signatures do not overlap with annotated gene models. Inspection of RNA-seq and spliced EST data from these regions suggests the possibility that the regions marked by these chromatin signatures represent as yet uncharacterized alternative promoters of nearby genes.

The biggest difference in the enrichment levels for any individual mark between these two patterns is seen for H3K36me3, a mark of transcriptional elongation (3,7). Consistent with this observation, genes marked by these two chromatin signatures show different expression levels in CD4$^+$ T cells (p=0.016; Figure 6.9.F). These data underscore the functional relevance of slight differences in chromatin signatures that are able to be distinguished by the ChAT algorithm.

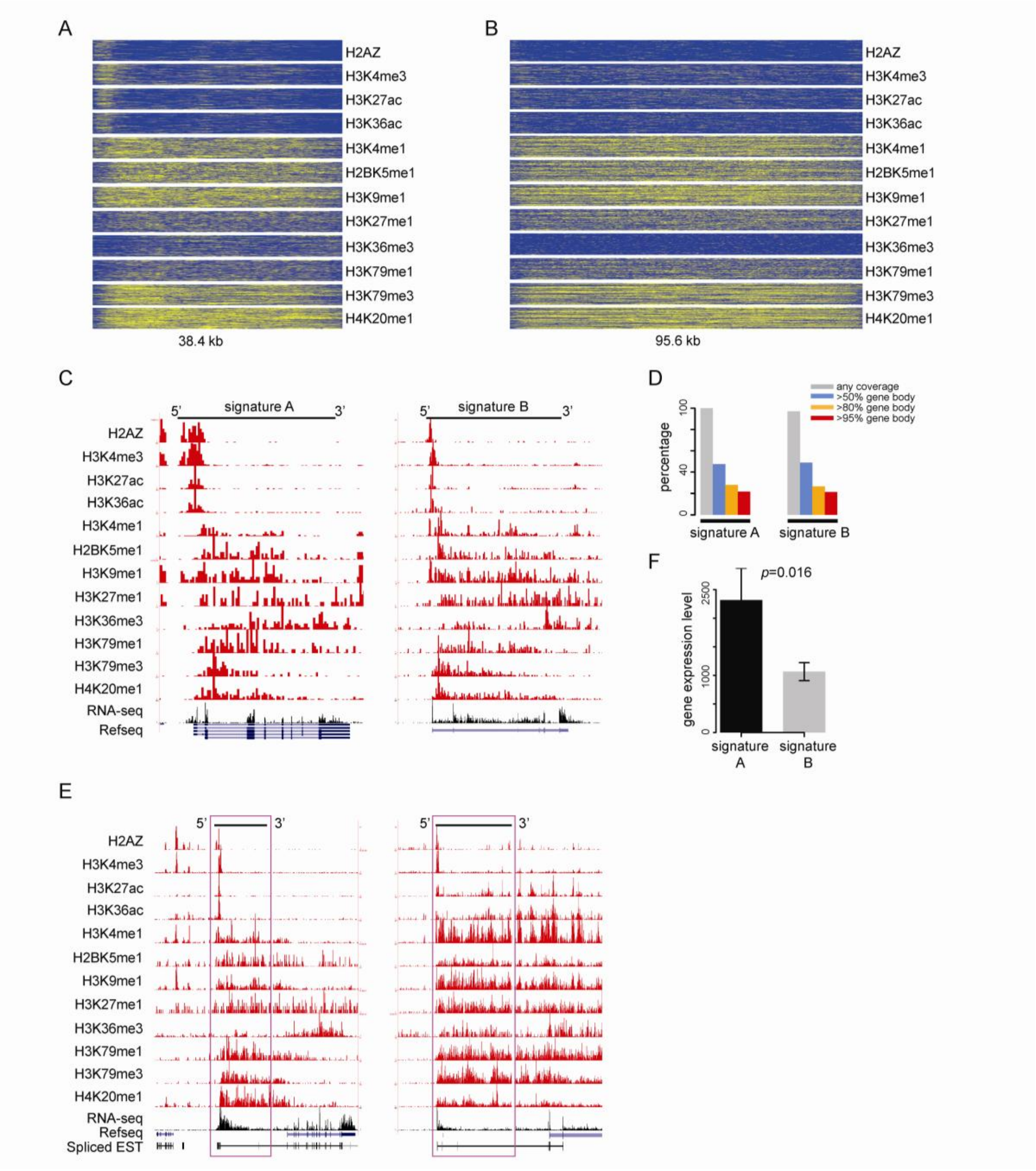


Figure 6.9: Large-sized chromatin signatures associated with gene bodies. (A & B) Histone modification enrichment profiles are shown for two chromatin signatures composed of the same constituent modifications and spatial patterns with distinct sizes. (C) Specific instances of each signature co-located with human gene bodies. (D) Percentage of these two large-sized signatures that overlapping with gene bodies (grey=any coverage, blue>50% coverage, orange>80% coverage, red >95% coverage of the gene body). (E) Two examples where signature B is co-located with individual genomic regions that are annotated as intergenic but show evidence of being genic. (F) Average CD4+ T cell expression levels for genes marked by signatures A and B.

Both of these long chromatin signatures show enrichment of H4K20me1 and H3K79me3 that tend to be located within gene bodies and start just downstream of TSS (Figure 6.9.A-C). This suggests the possibility that these marks are associated with transcriptional pause release, a phenomenon whereby Pol II complexes paused at promoter regions are allowed to proceed into gene bodies to facilitate active transcription of the genes (171,172). Previously, the relative levels of bound Pol II seen in promoter proximal versus downstream regions have been used to evaluate the extent of transcriptional pause release (173,174). Here, we show that the ratio of gene body-to-TSS Pol II density is positively correlated with the gene body levels of H4K20me1 (Figure 6.10.A) and H3K79me3 (Figure 6.10.B) consistent with a role for these marks in transcriptional pause release.

The discoveries of those complex large-sized signatures highlight the performance of ChAT with respect to several aspects of the algorithm design. First of all, the large-size of these signatures underscores the advantage of predicting chromatin signatures without size restrictions. Second, the prediction of large-sized signatures was facilitated by the ability of the algorithm to extend histone modification profile alignments through the use of gaps in the dynamic programming implementation. Third, the complex histone modification enrichment profiles apparent in these signatures, *i.e.* the specific enrichments of several histone modifications over a narrow range of the pattern and the broad enrichments of other marks in the rest of the pattern, demonstrates the ability of the algorithm to detect patterns with spatially shifted multi-modal enrichments of multiple modifications.

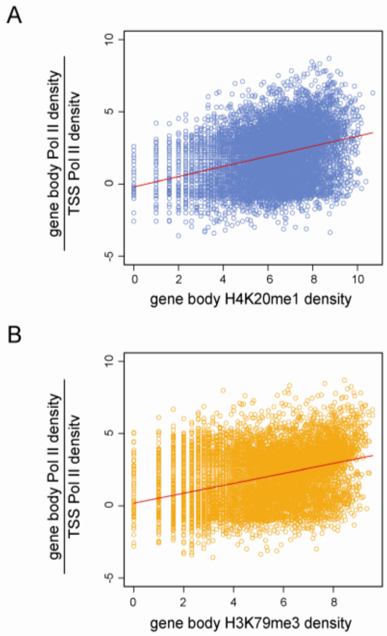


Figure 6.10: Transcriptional pause release associated with H4K20me1 and H3K79me3. The ratio of Pol II density downstream of TSS (+1kb~+5kb) over its density around TSS (-1kb~+1kb) is positively correlated with the density of downstream H4K20me1 (A, Spearman's ρ=0.54) and H3K79me3 (B, Spearman's ρ=0.51).

Conclusions

We developed ChAT (Chromatin-profile Alignment followed by Tree-clustering) an unsupervised algorithm for the discovery and characterization of recurrent combinatorial histone modification patterns, *i.e.* chromatin signatures. ChAT utilizes a novel dynamic programming and hierarchical clustering approach to relate and group similar chromatin signatures dispersed across the genome. The algorithm was explicitly designed to provide complementary utility with respect to existing methods. For example, ChAT can identify chromatin signatures across a vast range of different sizes, it finds multi-modal chromatin signatures composed of individual histone modifications that are spatially shifted as well as complex signatures composed of conserved and variant segments, and ChAT can also distinguish between chromatin signatures that are made up of the same constituent histone modifications with different shapes. The algorithm also employs an explicit statistical criterion that provides confidence levels for the grouping of similar chromatin signatures.

We applied ChAT to the analysis of genome-wide histone modification maps in human CD4$^+$ T cells. The algorithm was able to discern combinatorial histone modification patterns previously observed to be associated with genomic regulatory features such as TSS and enhancers, serving as a proof of its utility for the discovery of functionally relevant chromatin signatures. Perhaps more interestingly, we were also able to discover a number of previously unknown chromatin signatures with ChAT. For example, we discovered novel chromatin signatures associated with TTS, enhancers and CNEs. We were also able to uncover functional associations, based on enrichment of chromatin signatures at specific genomic regulatory features, which point to novel chromatin based mechanisms of gene regulation. For example, we found evidence for the role of complex chromatin signatures, made up of numerous co-located histone modifications, in the cell-type specific regulation of human genes. We also found evidence suggesting that L1 retrotransposons can influence the mono-allelic expression of human genes by creating a local genomic environment enriched for specific bivalent chromatin signatures. Finally, novel long chromatin signatures found to be associated with human genes suggest a role for the H4K20me1 and H3K79me3 histone modifications in transcriptional pause release. The discovery of these novel chromatin signatures and functional associations underscores the potential utility of the algorithm to provide novel biological insight and to help focus future experimental efforts for the characterization of chromatin based regulatory mechanisms.

Acknowledgements

The authors would like to acknowledge members of the Jordan and Lunyak labs for helpful discussions. Jianrong Wang and I. King Jordan were supported by Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839) and Georgia Tech Integrative BioSystems Institute pilot program grant. Victoria V. Lunyak was supported by the Buck Institute Trust Fund.

CHAPTER 7

CONCLUSIONS

In summary, this dissertation is composed of five computational algorithms developed for ChIP-seq datasets of epigenomics research. The first two algorithms belong to the basic data processing field, and they serve essentially to reduce noise and retrieve real genomic locations of histone modifications and/or transcription factors. In CHAPTER 2, a read mapping algorithm is developed to deal with ambiguous ChIP-seq tags, and in CHAPTER 3 a peak calling method is designed to identify broad peaks for diffuse ChIP-seq signals. The next three algorithms are question-driven methods that apply pattern recognition techniques for basic biological discoveries. While CHAPTER 4 focuses on a hypothesis-driven pipeline for insulator predictions, CHAPTER 5 introduces an unbiased hypothesis-free approach for predicting chromatin boundary elements. In CHAPTER 6, an unsupervised algorithm is developed to explore novel combinatorial chromatin signatures that are associated with various genomic features.

While next-generation sequencing technologies have produced large amounts of sequence tags that make many large-scale biological analyses applicable now, the very short sequences cause new computational problems when they are mapped back to the reference genomes. One problem is related to the ambiguity of multi-mapping tags. CHAPTER 2 presents a Gibbs sampling strategy to solve this problem. Theoretical derivations are discussed, and it guarantees the optimality of the performance from the Bayesian statistics point of view. Applying the method on simulated datasets, it has substantial improvements on the fractions of correctly mapped ambiguous reads compared with previous methods. Furthermore, the accuracy of recovering the real sites is also improved by using this algorithm. More detailed analysis of the recovered sites in different repetitive genomic regions supports the utility of this algorithm for finding more signals within those previously under-investigated regions.

As another critical step for basic data processing, several peak calling methods have been developed. Most of those methods restricted on calling sharp peaks that are characteristic for transcription factors and some histone modifications. There are certain histone modifications well known for their diffuse distributions and methods for identifying broad peaks are lacking. CHAPTER 3 presents a maximal scoring segment algorithm based method for broad peak calling. A parameter estimation module is constructed using Gibbs sampling procedures on non-homogeneous Poisson processes. The global observations of the shifted enrichments of H3K36me3 and H3K79me2 broad peaks along gene bodies, along with the enrichments of CTCF bindings around the edges of the resulted H3K27me3 broad peaks, indicate that the performance of this algorithm fits well with *a priori* biological knowledge. Evaluations on simulated datasets further prove the superior performance compared with existing methods for large broad peak calling.

One of the important epigenomics question relates to the identity and locations of insulator elements in the human genome. Inspired by experimental observations summarized from a subset of insulators, a hypothesis-driven pipeline is designed in CHAPTER 4 to predict locations of a subset of insulators: MIR-insulators. This pipeline integrates both genomics and epigenomics features and finally generated a set of 1,178 MIR-insulators in CD4$^+$ T cells in the human genome. Several selected MIR-insulators are experimentally tested using EBAs in both human kidney cell lines and zebrafish embryos. Specific local chromatin signatures are found for the putative MIR-insulators. Their distance distributions to TSS imply the evolutionary dynamics of insulators. Functional annotations of genes proximal to those MIR-insulators show an interesting enrichment of TCR pathway. This observation, along with specific examples, raises the importance of MIR-insulators on cell type specific regulations. A global analysis of those MIR-insulators found a large fraction of them to be functional in a cell type specific manner.

A related, but different, type of regulatory element is chromatin boundary elements. These boundary elements play an important role in epigenomics because they can organize large-scale chromatin domain configurations and presumably related with three-dimensional structures. In order to address the lack of unified features of the currently known barriers, CHAPTER 5 developed an unbiased hypothesis-free algorithm to search for boundary elements in an attempt to discovering novel features. In order to do that, chromatin boundaries are modeled as transition points between chromatin states and a HMM based method is designed. As an indication of the good performance, the canonical boundary element, BEAD1 element, is successfully found. The resulted boundaries can be classified into CTCF dependent and independent groups. To search for novel features, sequence analysis shows a set of transcription factor binding motifs enriched within the predicted boundaries. It includes EVI1, CREBP1, USF and YY1. All of these proteins have interesting interactions with chromatin modifying enzymes, and USF has even been shown to be related to one canonical boundary element experimentally before. The most interesting feature analysis comes from the finding of a subset of boundaries containing non-coding RNA genes. It is the first report of potential non-coding RNA gene, specifically tRNA, derived boundaries in the human genome, although similar observations were found in yeast and mouse before. This computational predication was later experimentally confirmed.

Given the large number of different histone modifications existing in the genome, the complex relationships between combinations of histone modifications and various genomic features have been an important topic in epigenomics research. CHAPTER 6 built an unsupervised algorithm for the discovery of combinatorial chromatin signatures in the genome. This algorithm is based on a high-dimensional profile alignment strategy and bears a set of inherent advantages compared to previous methods, such as free of size restrictions, the capability of finding multi-mode patterns, the discrimination between patterns with different profile shapes and the statistical criteria for pattern identifications. Applications of this method on human CD4$^+$ T cell epigenome datasets produced a set of interesting combinatorial chromatin signatures. These signatures are further analyzed by comparing with various genomic features. These associations support the performance of this method for discovering novel combinatorial chromatin signatures and the utility of this algorithm for biological research.

APPENDIX A

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

Table A.1. Parameters of sequence tag libraries.

Libraries	tag length	signal to noise ratio	sequencing error rate
Library 1	35	99	2/5L
Library 2	35	99	4/5L
Library 3	35	9	2/5L
Library 4	35	9	4/5L
Library 5	20	99	2/5L
Library 6	20	99	4/5L
Library 7	20	9	2/5L
Library 8	20	9	4/5L
Big Library	20	9	4/5L

Table A.2. Comparison of the number of sites identified using unique tags only versus using unique and ambiguous tags with the Gibbs sampling method.

		Lib 1	Lib 2	Lib 3	Lib 4	Lib 5	Lib 6	Lib 7	Lib 8	Lib Big
th	Total site	51278	51278	51278	51278	51278	51278	51278	51278	173877
4	Unique	42886	42454	42729	42302	38631	36760	38515	36729	144751
	Unique TP	42869	42446	42712	42290	38625	36759	38515	36729	144638
	Gibbs	46127	45857	45957	45762	43848	43225	43693	43036	162991
	Gibbs TP	45470	45264	45315	45124	43123	42551	43002	42347	162501
	Improvement	2601	2818	2603	2834	4498	5792	4498	5624	17863
	Fraction	5.07%	5.50%	5.08%	5.53%	8.77%	11.30%	8.77%	10.97%	10.27%
6	Unique	41264	40600	41120	40454	36364	32253	36284	32100	126452
	Unique TP	41262	40599	41118	40452	36363	32252	36282	32099	126402
	Gibbs	45142	44500	44957	44369	42383	41347	42254	41229	158139
	Gibbs TP	44622	44060	44452	43902	41826	40879	41724	40734	157929
	improvement	3360	3461	3334	3450	5463	8627	5442	8635	31527
	fraction	6.55%	6.75%	6.50%	6.73%	10.65%	16.8%	10.61%	16.84%	18.13%
8	Unique	39351	37967	39214	37856	32816	24852	32667	24918	98044
	Unique TP	39350	37966	39214	37855	32815	24851	32666	24918	98018
	Gibbs	43520	42023	43317	41879	40631	38340	40501	38287	149200
	Gibbs TP	43115	41705	42921	41555	40221	38031	40098	37950	149076
	improvement	3765	3739	3707	3700	7406	13180	7432	13032	51058
	fraction	7.34%	7.29%	7.23%	7.22%	14.44%	25.70%	14.49%	25.41%	29.36%

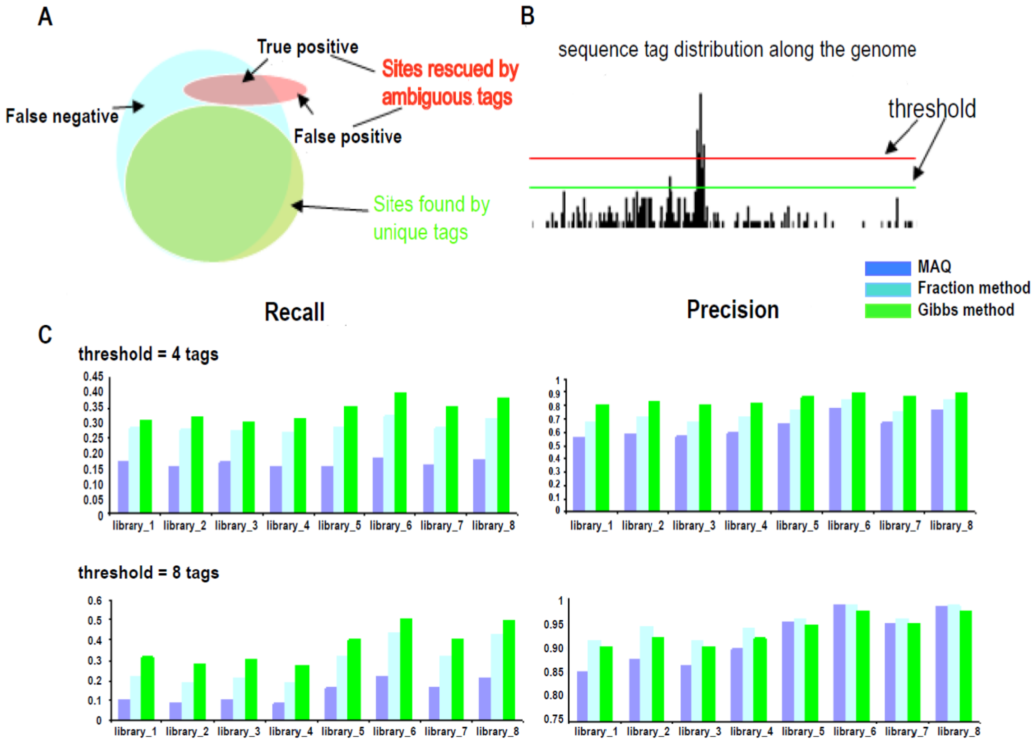


Figure A.1. Recall and precision of 3 algorithms under various tag count thresholds (4 tags and 8 tags). A. Illustration of data used to test algorithm performances. B. Variant tag thresholds could cause differences in the performance test. The lines (red and green) are two tag thresholds. C. Barplots of recall and precision for the three methods (MAQ-dark blue, fraction method-light blue, Gibbs method-green) on 8 libraries under tag thresholds = 4 and 8.

Table A.3. Algorithm performance on 8 sequence tag libraries.

Th	libraries	sites to recover	MAQ	MAQ (TP)	Fraction method	Fraction method (TP)	Gibbs	Gibbs (TP)
4	library 1	8409	2581	1435	3558	2387	3241	2601
	library 2	8832	2378	1395	3434	2452	3403	2818
	library 3	8566	2573	1444	3490	2352	3228	2603
	library 4	8988	2378	1418	3397	2432	3460	2834
	library 5	12653	3009	1997	4737	3588	5217	4498
	library 6	14519	3409	2643	5583	4669	6465	5792
	library 7	12774	3077	2049	4806	3628	5178	4498
	library 8	14555	3384	2594	5462	4544	6307	5624
6	library 1	10016	1912	1366	3164	2582	3878	3360
	library 2	10679	1749	1325	2994	2580	3900	3461
	library 3	10160	1849	1371	3110	2579	3837	3334
	library 4	10826	1699	1303	2968	2555	3915	3450
	library 5	14915	2718	2287	4777	4230	6019	5463
	library 6	19026	4155	3883	7427	7060	9094	8627
	library 7	14996	2701	2282	4699	4140	5970	5442
	library 8	19179	4264	3979	7469	7094	9129	8635
8	library 1	11928	1476	1258	2875	2632	4169	3765
	library 2	13312	1352	1186	2685	2533	4056	3739
	library 3	12064	1446	1252	2753	2521	4103	3707
	library 4	13423	1309	1176	2643	2491	4023	3700
	library 5	18463	3186	3032	6174	5930	7815	7406
	library 6	26427	5868	5798	11547	11425	13488	13180
	library 7	18612	3256	3089	6190	5944	7834	7432
	library 8	26360	5633	5553	11428	11282	13369	13032

Table A.4. Algorithm performance for recall and precision together.

thresholds		**4**			**6**			**8**		
criteria		***recall***	***precision***	***F***	***recall***	***precision***	***F***	***recall***	***precision***	***F***
Library_1	**MAQ**	**0.17**	**0.56**	**0.26**	**0.14**	**0.71**	**0.23**	**0.11**	**0.85**	**0.19**
	Fraction	**0.28**	**0.67**	**0.39**	**0.26**	**0.82**	**0.39**	**0.22**	**0.92**	**0.36**
	Gibbs	**0.31**	**0.80**	**0.45**	**0.34**	**0.87**	**0.49**	**0.32**	**0.90**	**0.47**
Library_2	**MAQ**	**0.16**	**0.59**	**0.25**	**0.12**	**0.76**	**0.21**	**0.09**	**0.88**	**0.16**
	Fraction	**0.28**	**0.71**	**0.40**	**0.24**	**0.86**	**0.38**	**0.19**	**0.94**	**0.32**
	Gibbs	**0.32**	**0.83**	**0.46**	**0.32**	**0.89**	**0.47**	**0.28**	**0.92**	**0.43**
Library_3	**MAQ**	**0.17**	**0.56**	**0.26**	**0.13**	**0.74**	**0.22**	**0.10**	**0.87**	**0.18**
	Fraction	**0.27**	**0.67**	**0.38**	**0.25**	**0.83**	**0.38**	**0.21**	**0.92**	**0.34**
	Gibbs	**0.30**	**0.81**	**0.44**	**0.33**	**0.87**	**0.48**	**0.31**	**0.90**	**0.46**
Library_4	**MAQ**	**0.16**	**0.60**	**0.25**	**0.12**	**0.77**	**0.21**	**0.09**	**0.90**	**0.16**
	Fraction	**0.27**	**0.72**	**0.39**	**0.24**	**0.86**	**0.38**	**0.19**	**0.94**	**0.32**
	Gibbs	**0.32**	**0.82**	**0.46**	**0.32**	**0.88**	**0.47**	**0.28**	**0.92**	**0.43**
Library_5	**MAQ**	**0.16**	**0.66**	**0.26**	**0.15**	**0.84**	**0.25**	**0.16**	**0.95**	**0.27**
	Fraction	**0.28**	**0.76**	**0.41**	**0.28**	**0.89**	**0.43**	**0.32**	**0.96**	**0.48**
	Gibbs	**0.36**	**0.86**	**0.51**	**0.37**	**0.91**	**0.53**	**0.40**	**0.95**	**0.56**
Library_6	**MAQ**	**0.18**	**0.78**	**0.29**	**0.20**	**0.93**	**0.33**	**0.22**	**0.99**	**0.36**
	Fraction	**0.32**	**0.84**	**0.46**	**0.37**	**0.95**	**0.53**	**0.43**	**0.99**	**0.60**
	Gibbs	**0.40**	**0.90**	**0.55**	**0.45**	**0.95**	**0.61**	**0.50**	**0.98**	**0.66**
Library_7	**MAQ**	**0.16**	**0.67**	**0.26**	**0.15**	**0.84**	**0.25**	**0.17**	**0.95**	**0.29**
	Fraction	**0.28**	**0.75**	**0.41**	**0.28**	**0.88**	**0.42**	**0.32**	**0.96**	**0.48**
	Gibbs	**0.35**	**0.87**	**0.50**	**0.36**	**0.91**	**0.52**	**0.40**	**0.95**	**0.56**
Library_8	**MAQ**	**0.18**	**0.77**	**0.29**	**0.21**	**0.93**	**0.34**	**0.21**	**0.99**	**0.35**
	Fraction	**0.31**	**0.83**	**0.45**	**0.37**	**0.95**	**0.53**	**0.43**	**0.99**	**0.60**
	Gibbs	**0.39**	**0.89**	**0.54**	**0.45**	**0.95**	**0.61**	**0.49**	**0.97**	**0.65**
Library_Big	**MAQ**	**0.28**	**0.92**	**0.43**	**0.30**	**0.99**	**0.46**	**0.29**	**1.00**	**0.45**
	Fraction	**0.46**	**0.95**	**0.62**	**0.53**	**0.99**	**0.69**	**0.56**	**1.00**	**0.72**
	Gibbs	**0.61**	**0.98**	**0.75**	**0.66**	**0.99**	**0.79**	**0.67**	**1.00**	**0.80**

Table A.5. Algorithm performance on the bigger sequence tag library.

Th	sites to recover	MAQ	MAQ (TP)	Fraction method	Fraction method (TP)	Gibbs	Gibbs (TP)
4	29239	8963	8272	14330	13594	18240	17863
6	47475	14475	14314	25181	25034	31687	31527
8	75859	21700	21655	42789	42746	51156	51058

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

Table B.1: The summary of recall, precision and F score for BroadPeak, SICER and RSEG on simulated libraries with different overlapping criterion.

Feature	Software	overlap criteria = 20%			overlap criteria = 50%			overlap criteria = 80%		
		dataset 1	dataset 2	dataset 3	datase t 1	dataset 2	dataset 3	dataset 1	dataset 2	dataset 3
Recall	BroadPeak	0.44	0.53	0.55	0.39	0.46	0.51	0.38	0.43	0.47
	SICER	0.07	0.07	0.07	0.04	0.04	0.04	0.03	0.03	0.03
	RSEG	0.60	0.38	0.19	0.31	0.21	0.12	0.24	0.18	0.11
Precision	BroadPeak	0.69	0.74	0.73	0.60	0.66	0.64	0.55	0.60	0.58
	SICER	0.78	0.84	0.89	0.66	0.74	0.82	0.52	0.63	0.74
	RSEG	0.29	0.65	0.89	0.24	0.55	0.82	0.18	0.44	0.71
F score	BroadPeak	0.54	0.62	0.63	0.47	0.54	0.57	0.45	0.50	0.52
	SICER	0.12	0.12	0.13	0.08	0.07	0.08	0.06	0.06	0.06
	RSEG	0.39	0.48	0.31	0.27	0.30	0.21	0.21	0.26	0.19

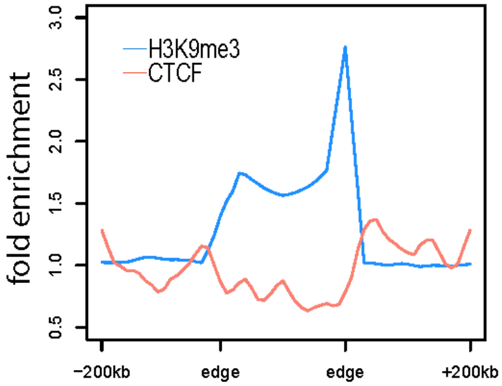


Figure B.1: Enrichment of CTCF binding around broad peak edges of H3K9me3 identified by BroadPeak. The blue curve shows the enrichment profile of H3K9me3 within and around the identified broad peaks and the pink curve shows the enrichment profile of CTCF binding.

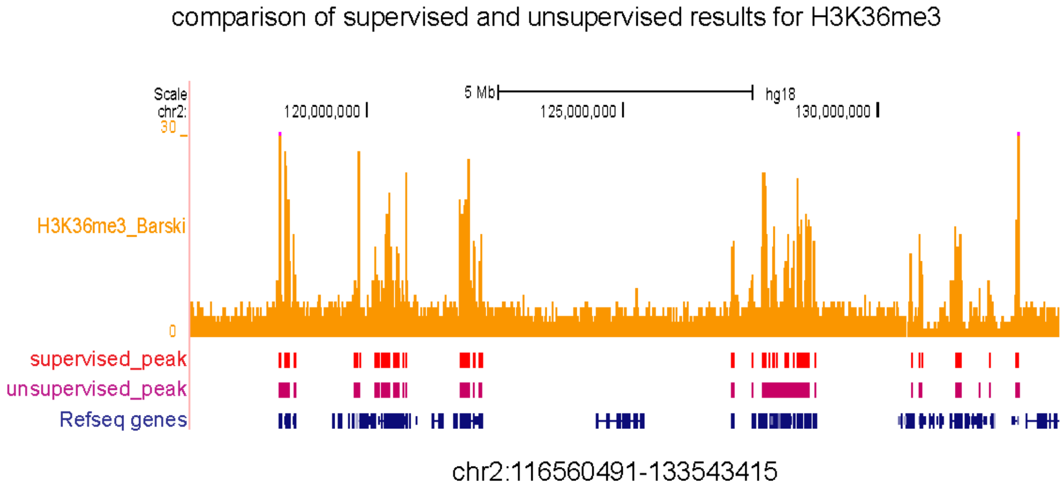


Figure B.2: Examples of H3K36me3 broad peaks identified by supervised and unsupervised parameter estimations. Examples of the supervised peaks (red) and unsupervised peaks (purple) are compared with the gene bodies (blue) and the tag profiles of H3K36me3 (orange).

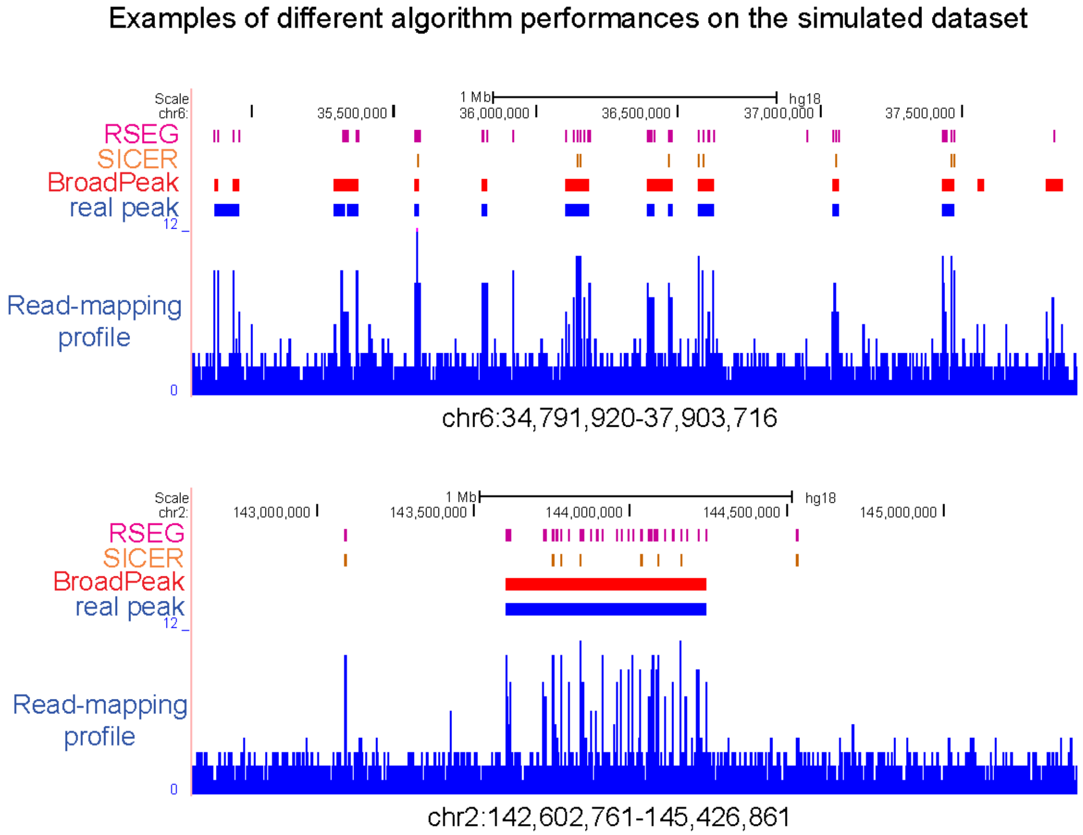


Figure B.3: Examples of broad peak calling by BroadPeak, SICER and RSEG on one simulated library. The identified broad peaks by RSEG (purple), SICER (orange) and BroadPeak (red) are compared with the real peaks and the tag profiles (blue).

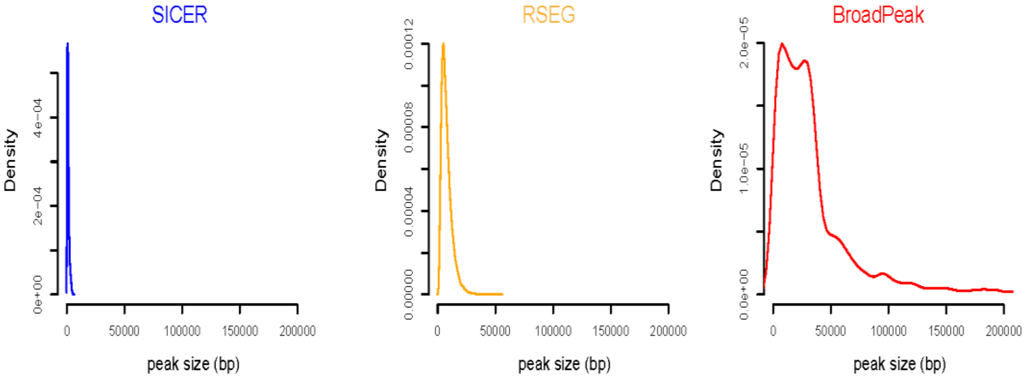


Figure B.4: The size distributions of broad peaks identified by BroadPeak, SICER and RSEG.

APPENDIX C

SUPPLEMENTARY INFORMATION FOR CHAPTER 4

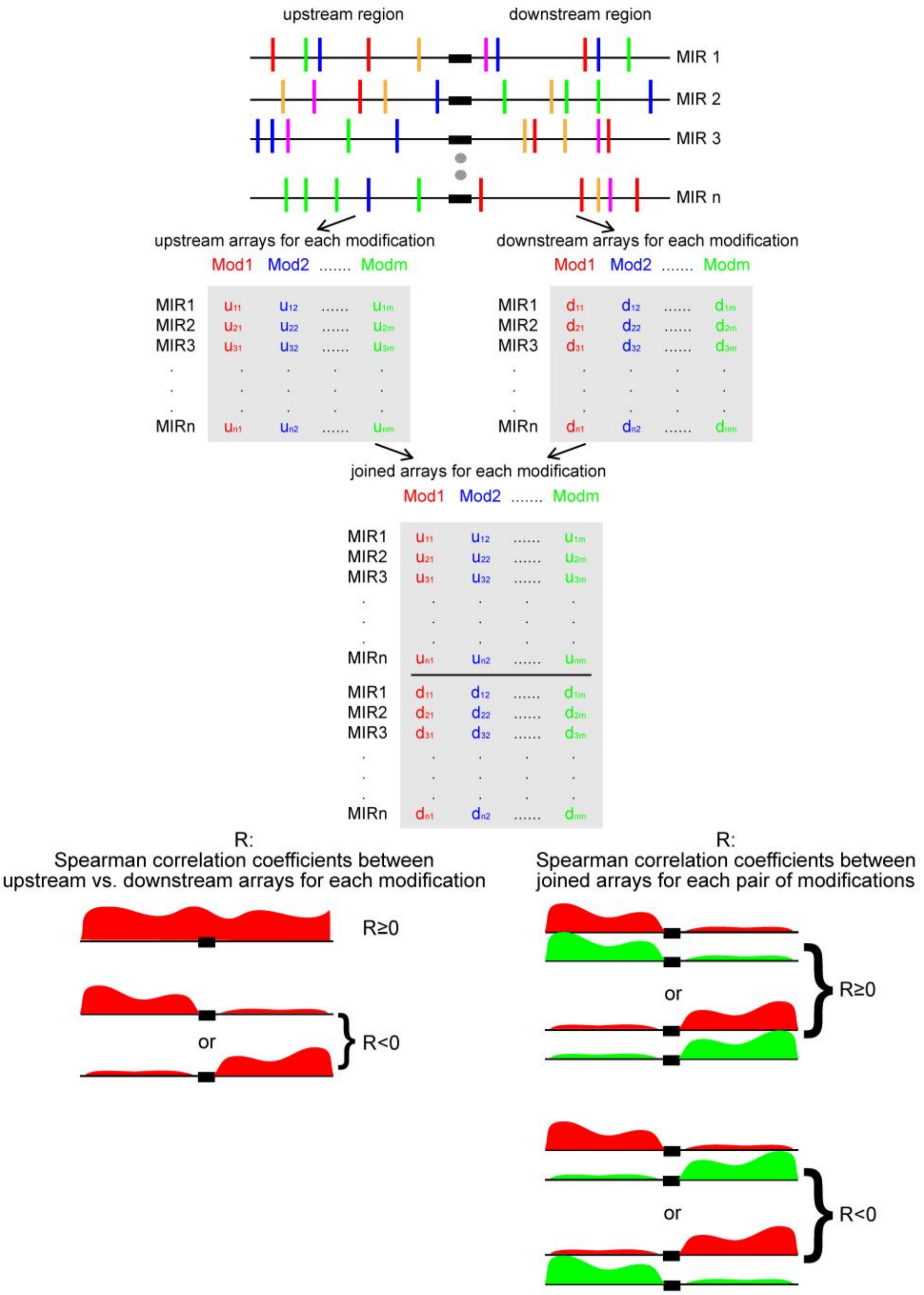

upstream region
downstream region
MIR 1
MIR 2
MIR 3
MIR n
upstream arrays for each modification
downstream arrays for each modification
Mod1 Mod2 .... Modm
MIR1 u_{11} u_{12} ...... u_{1m}
MIR2 u_{21} u_{22} ...... u_{2m}
MIR3 u_{31} u_{32} ...... u_{3m}
MIRn u_{n1} u_{n2} ...... u_{nm}
MIR1 d_{11} d_{12} ...... d_{1m}
MIR2 d_{21} d_{22} ...... d_{2m}
MIR3 d_{31} d_{32} ...... d_{3m}
MIRn d_{n1} d_{n2} ...... d_{nm}
joined arrays for each modification
R:
Spearman correlation coefficients between upstream vs. downstream arrays for each modification
R:
Spearman correlation coefficients between joined arrays for each pair of modifications
R≥0
or
R<0
R≥0
or
or
R<0

Figure C.1: Scheme illustrating the MIR-insulator computational validation procedure.

Table C.1: Spearman correlations between upstream and downstream histone modification levels across putative MIR-insulators.

Histone modifications	Spearman correlations	*P* value
H2AK5ac	-0.23	2.3E-8
H2AK9ac	-0.79	3.1E-81
H2BK5ac	-0.39	1.2E-41
H2BK12ac	-0.47	2.7E-55
H2BK20ac	-0.24	3.7E-16
H2BK120ac	-0.36	5.9E-36
H3K4ac	-0.41	1.4E-44
H3K9ac	-0.55	1.3E-82
H3K14ac	-0.85	1.6E-21
H3K18ac	-0.20	3.1E-12
H3K23ac	-0.76	3.6E-61
H3K27ac	-0.37	2.3E-37
H3K36ac	-0.43	4.9E-49
H4K5ac	-0.32	1.7E-27
H4K8ac	-0.28	9.6E-21
H4K12ac	-0.70	8.5E-84
H4K16ac	-0.44	5.7E-45
H4K91ac	-0.33	6.8E-29
H2AZ	-0.18	10.0E-10
H2BK5me1	-0.36	3.7E-34
H3K4me1	-0.06	1.5E-2
H3K4me2	-0.38	4.1E-41
H3K4me3	-0.32	1.0E-29
H3K9me1	-0.41	5.0E-49
H3K9me2	-0.78	1.6E-37
H3K9me3	-0.63	6.5E-48
H3K27me1	-0.42	3.4E-47
H3K27me2	-0.72	7.2E-66
H3K27me3	-0.47	1.0E-31
H3K36me1	-0.70	1.0E-56
H3K36me3	-0.43	3.7E-52
H3K79me1	-0.54	3.2E-80
H3K79me2	-0.72	3.5E-146
H3K79me3	-0.69	3.4E-139
H3R2me1	-0.33	6.6E-20
H3R2me2	-0.72	1.2E-12
H4K20me1	-0.45	1.3E-55
H4K20me3	-0.64	1.2E-17
H4R3me2	-0.82	1.6E-29

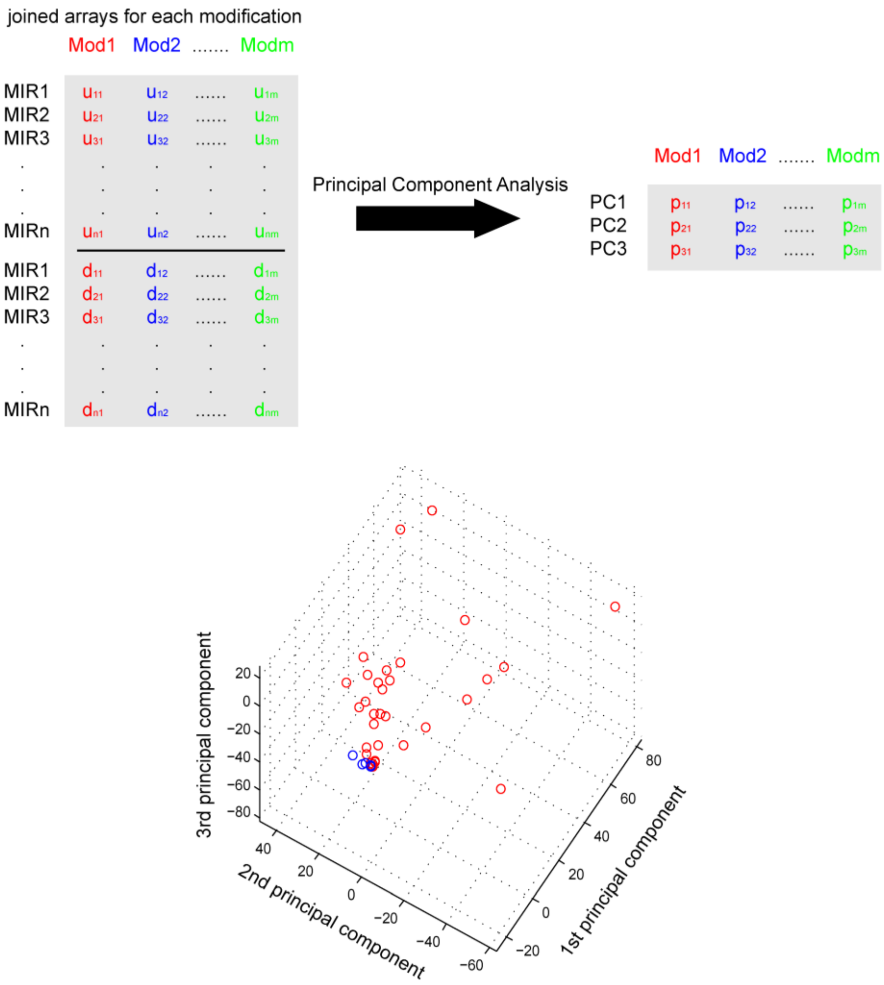


Figure C.2: Scheme and results of the principal components analysis around predicted MIR-insulators. Above, the joined histone modification arrays are shown along the first three principal component arrays that result from the PCA analysis. Below, a three-dimensional plot showing the locations of individual active (red) and repressive (blue) histone modifications in the principal component space.

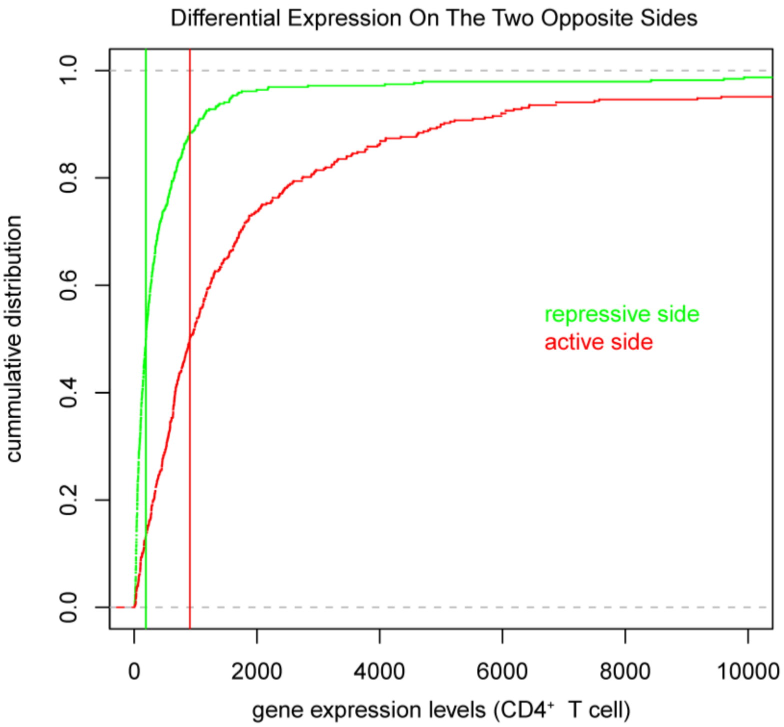


Figure C.3: Cumulative distributions of the CD4+ T cell gene expression levels for MIR-insulator proximal genes.

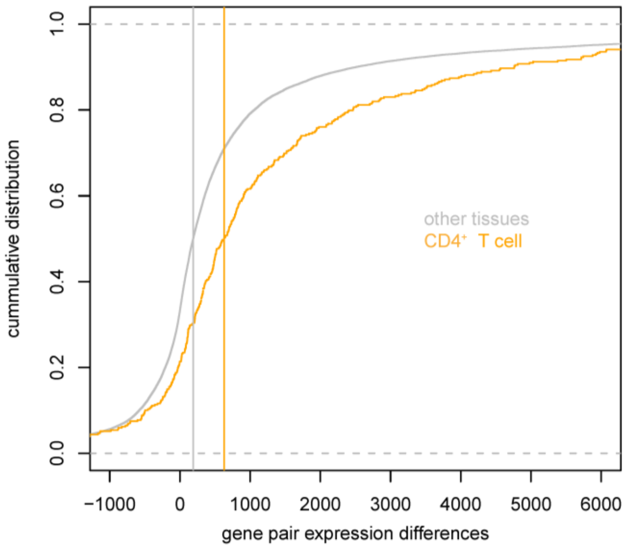


Figure C.4: Cumulative distributions of the differences in the gene expression levels for genes proximal to MIR-insulators. Difference distributions are shown for CD4+ T cell expression levels (orange) and for expression levels across 78 different tissues (grey).

Table C.2: Genomic coordinates of three MIR-insulators and their corresponding primers for EBA validations.

MIR element locations	Type	Coordinates of tested sequences	Size (bp)	Primer ID	Primer Sequences
chr1:23555914-23556047	MIR	chr1:23555859-23556088	230	1	ATACACTCGAGATGCATGATATGGCCCAGTGATGGTC
				2	ATACACTCGAGATGCATAGTCATGCCCATACCACCTC
		chr1:23555438-23556521	1084	3	ATACACTCGAGATGCATTGATTGGGATAAACCCAGGA
				4	ATACACTCGAGATGCATTCCCATTGCATGATCTGTTT
chr2:97999554-97999807	MIR	chr2:97999495-97999868	374	5	ATACACTCGAGCTGCAGTGAACATAGGAGGGGAGGTG
				6	ATACACTCGAGCTGCAGAAGATGATCCACCCTGCAAT
		chr2:97999109-98000252	1144	7	ATACACTCGAGCTGCAGAGGAGCCAGTCACAGAAGGA
				8	ATACACTCGAGCTGCAGTGCTTTGAAACCCTTTACGC
chr11:82289556-82289817	MIRb	chr11:82289550-82289843	294	9	ATACACTCGAGATGCATAACGGCAATAACAGCTACCA
				10	ATACACTCGAGATGCATTAGGGAGTGGTTAGGCTCCA
		chr11:82289143-82290278	1136	11	ATACACTCGAGATGCATCAGAAGCGCACAGGCTAAG
				12	ATACACTCGAGATGCATAGTCTTTCTCCCCGACAGGT

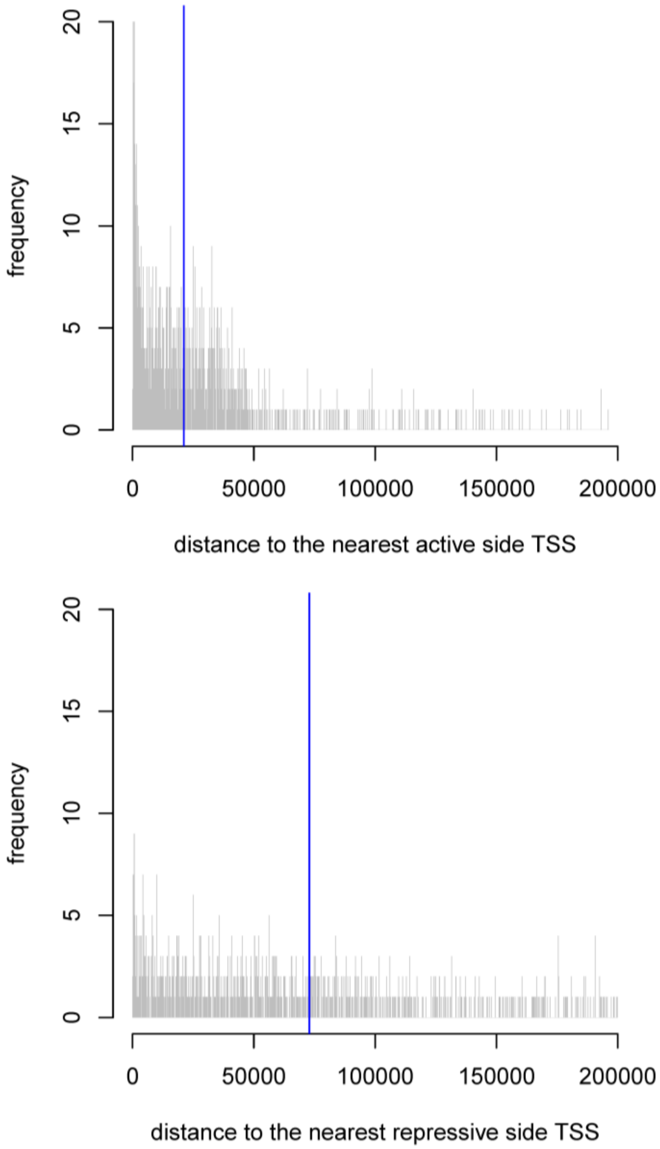


Figure C.5: Distance distributions between MIR-insulators and the nearest gene promoters. Median values of the distributions are shown in blue on each plot.

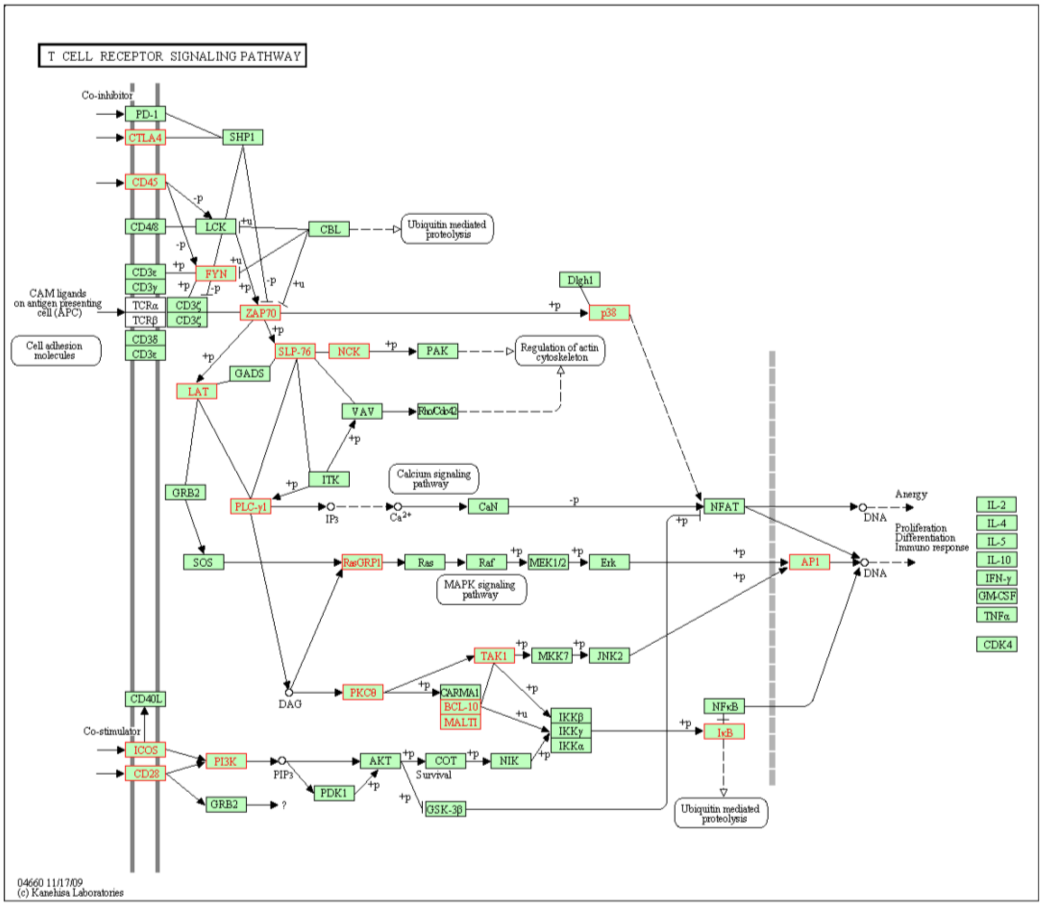


Figure C.6: T cell receptor pathway illustration from KEGG database. Genes located proximal to MIR-insulators, on the active domain side, are highlighted in red.

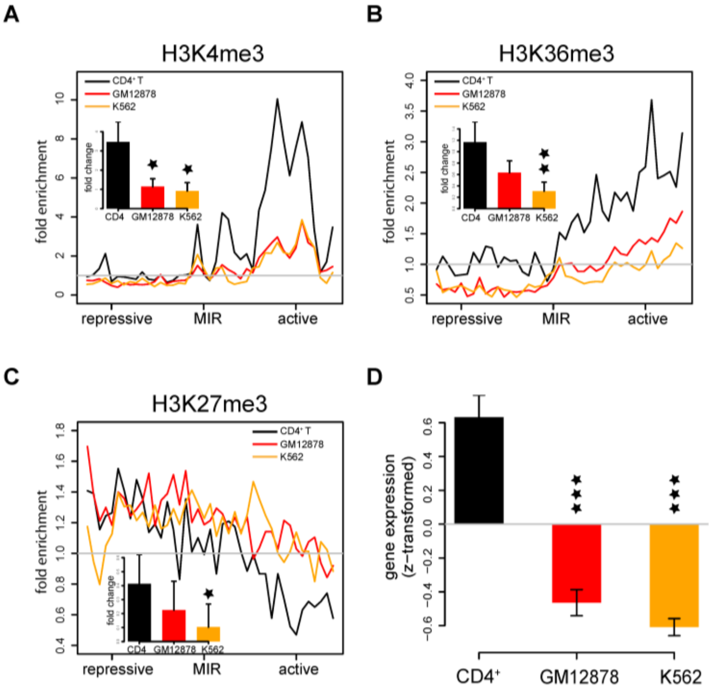


Figure C.7: Cell type-specific chromatin barrier activity and gene regulation by MIR-insulators from the T cell receptor pathway. ChIP-seq fold enrichment levels around MIR-insulators proximal to the 21 T cell receptor genes are shown for H3K4me3, H3K36me3 and H3K27me3 in CD4+ T cells (black), GM12878 cells (red) and K562 (orange) cells. Insets show the average differences (± standard error) between the active versus repressive domains surrounding MIR-insulators for the marks and cells. Significance of the differences between CD4+ T cells and other cells are indicated as * P<0.05 ** P<0.01 *** P<0.001. Average gene expression levels (± standard error) are shown for genes located in the active domain side proximal to MIR-insulators at the 21 T cell receptor genes.

APPENDIX D

SUPPLEMENTARY INFORMATION FOR CHAPTER 5

Table D.1: Sequence features of RITs and predicted boundary elements.

	RIT	Boundary element
Median size	68.6kb	8kb
GC content	0.421	0.423
CpG O/E	0.229	0.316
Genic fractions	43.0%	40.9%

[1] The ratio of observed CpG frequency to expected CpG frequency.
[2] The length fractions of regions within gene bodies.

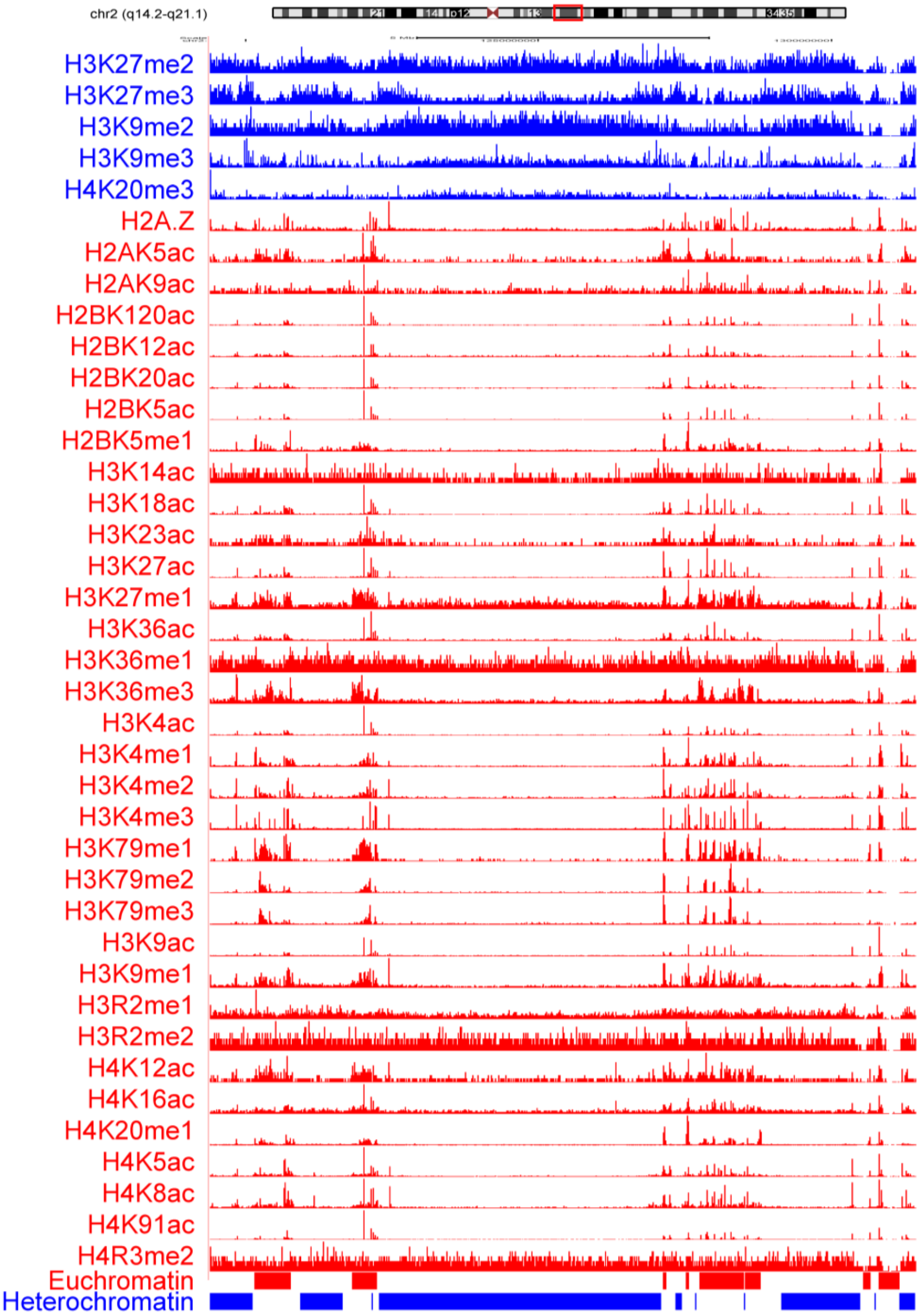


Figure D.1: Example of predicted chromatin domains. An ideogram of chromosome 2 shows the cytogenetic banding pattern along with the location of this specific example. The distributions of ChIP-seq tag mapping peaks for the active histone modification (red bars), the repressive histone modification (blue bars) are shown in separate tracks. The predicted euchromatic domains (red bands) and heterochromatic domains (blue bands) are shown in the tracks denoted as ‘Euchromatin’ and ‘Heterochromatin’.

Table D.2: Enriched gene ontology and KEGG terms of genes in predicted euchromatin domains with high gene densities.

	Term	*P*-value
Gene Ontology	ATP dependent helicase activity	0.039
	Defense Response	0.045
	Glycerophospholipid Biosynthetic Process	0.056
	Regulation of Response to External Stimulus	0.069
	Inflammatory Response	0.070
KEGG Pathway	Systemic Lupus Erythematosus	0
	Antigen Processing and Presentation	0.017

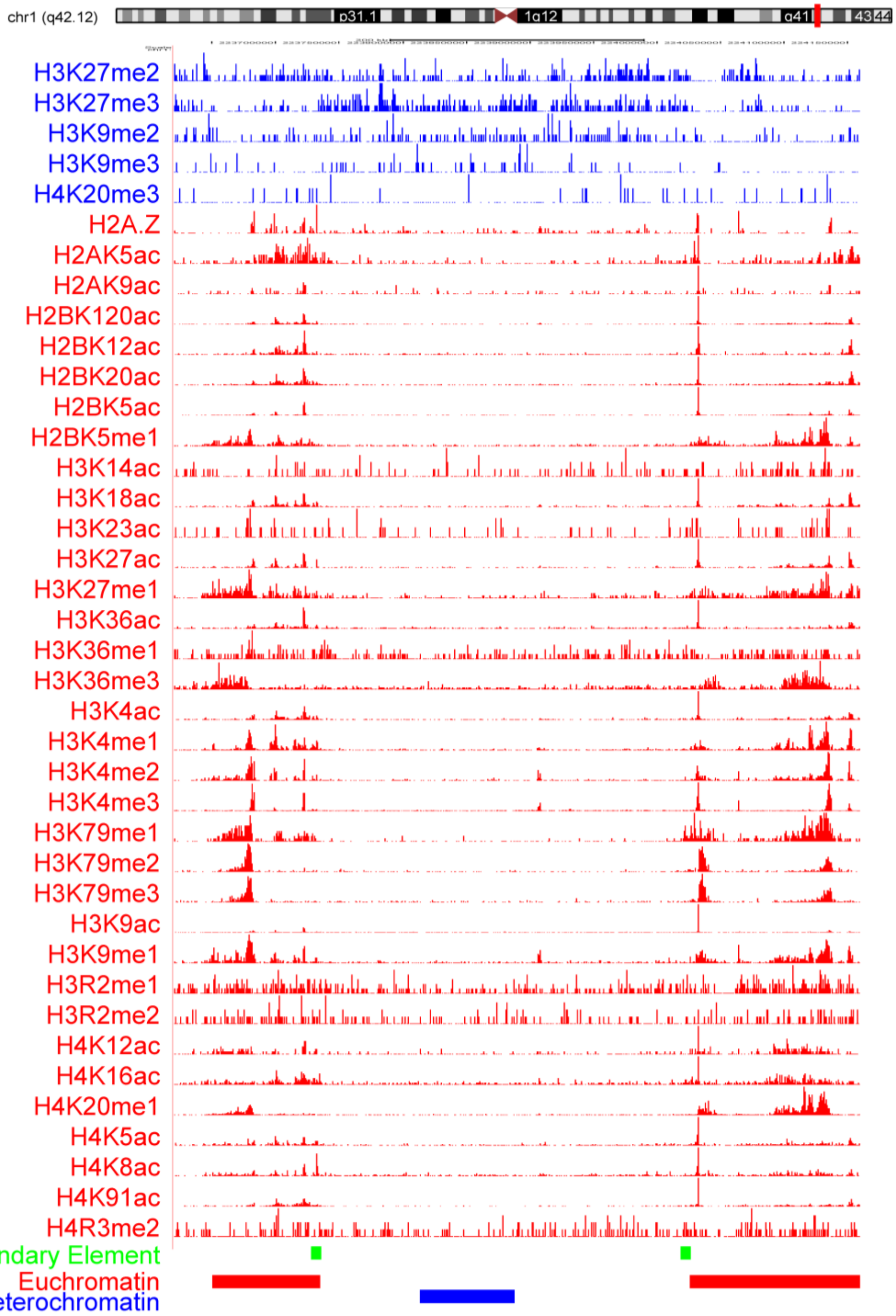


Figure D.2: Examples of predicted boundary elements with CTCF binding. The predicted boundary elements are shown as green bands. ChIP-seq peaks for active and repressive histone modifications, along with the locations of euchromatic domains and heterochromatic domains are illustrated as separate tracks (as in Figure D.1).

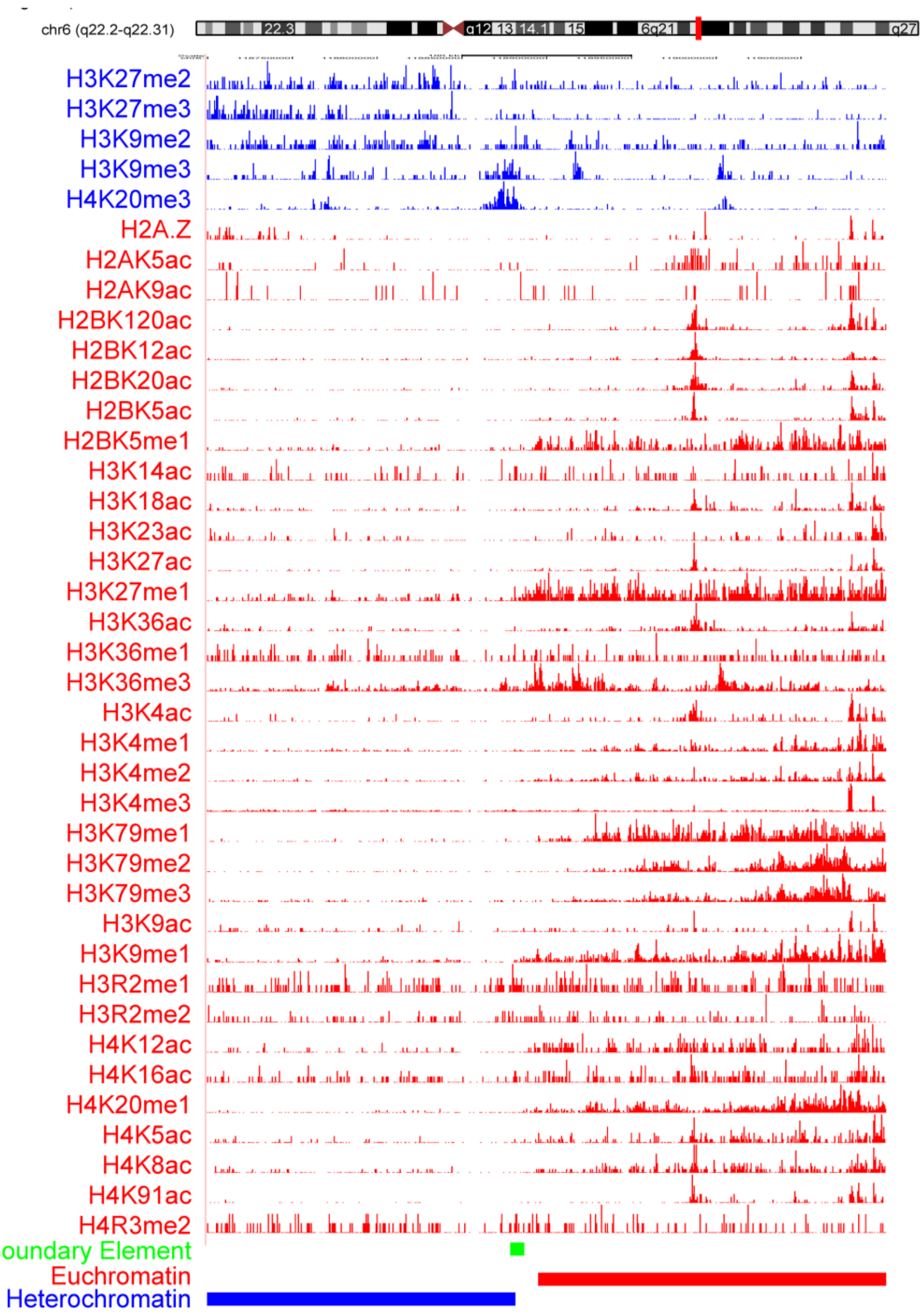


Figure D.3: Example of a predicted boundary element without CTCF binding. The predicted boundary element is shown as green bands. ChIP-seq peaks for active and repressive histone modifications, along with the locations of euchromatic domains and heterochromatic domains are illustrated as separate tracks (as in Figure D.1).

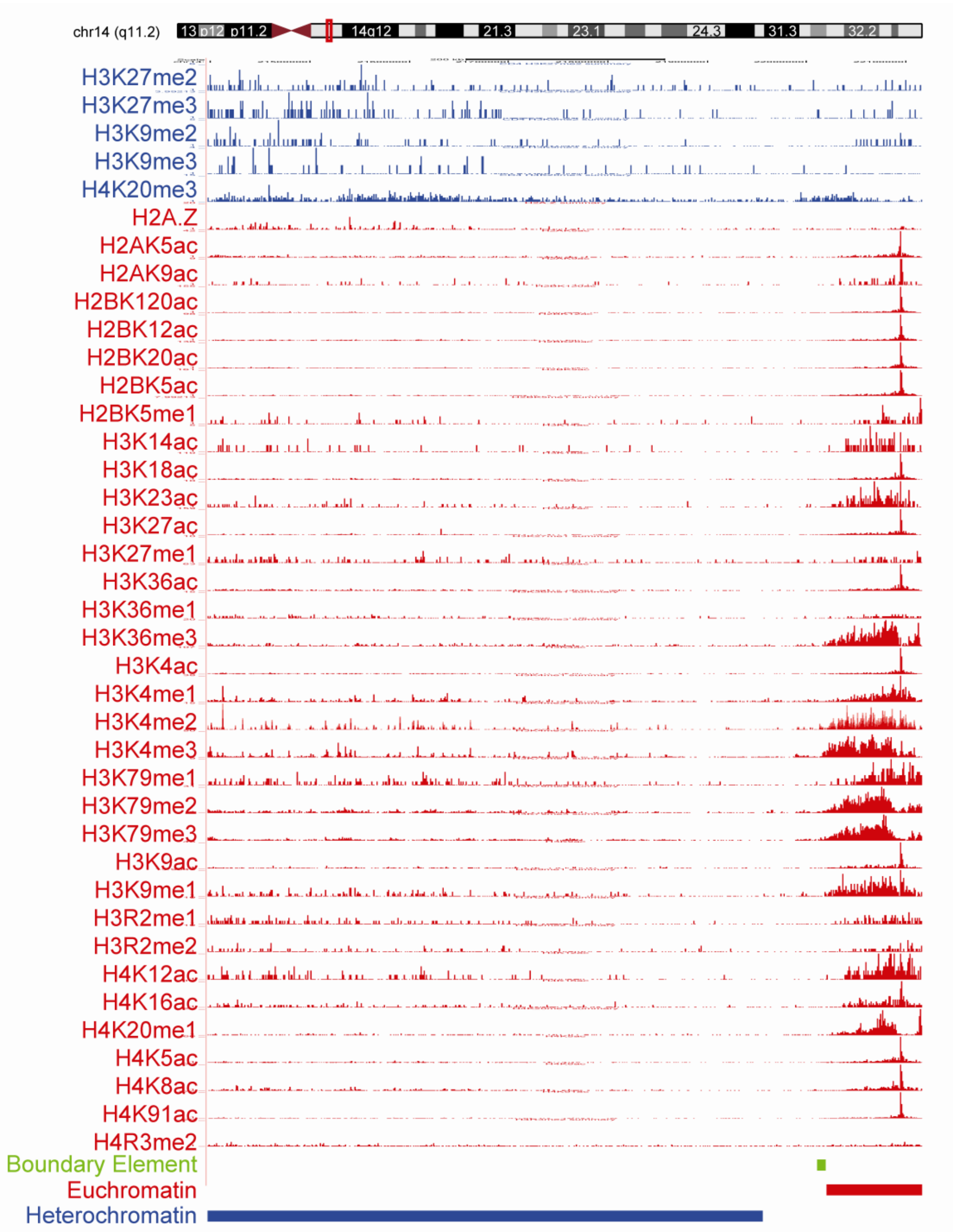
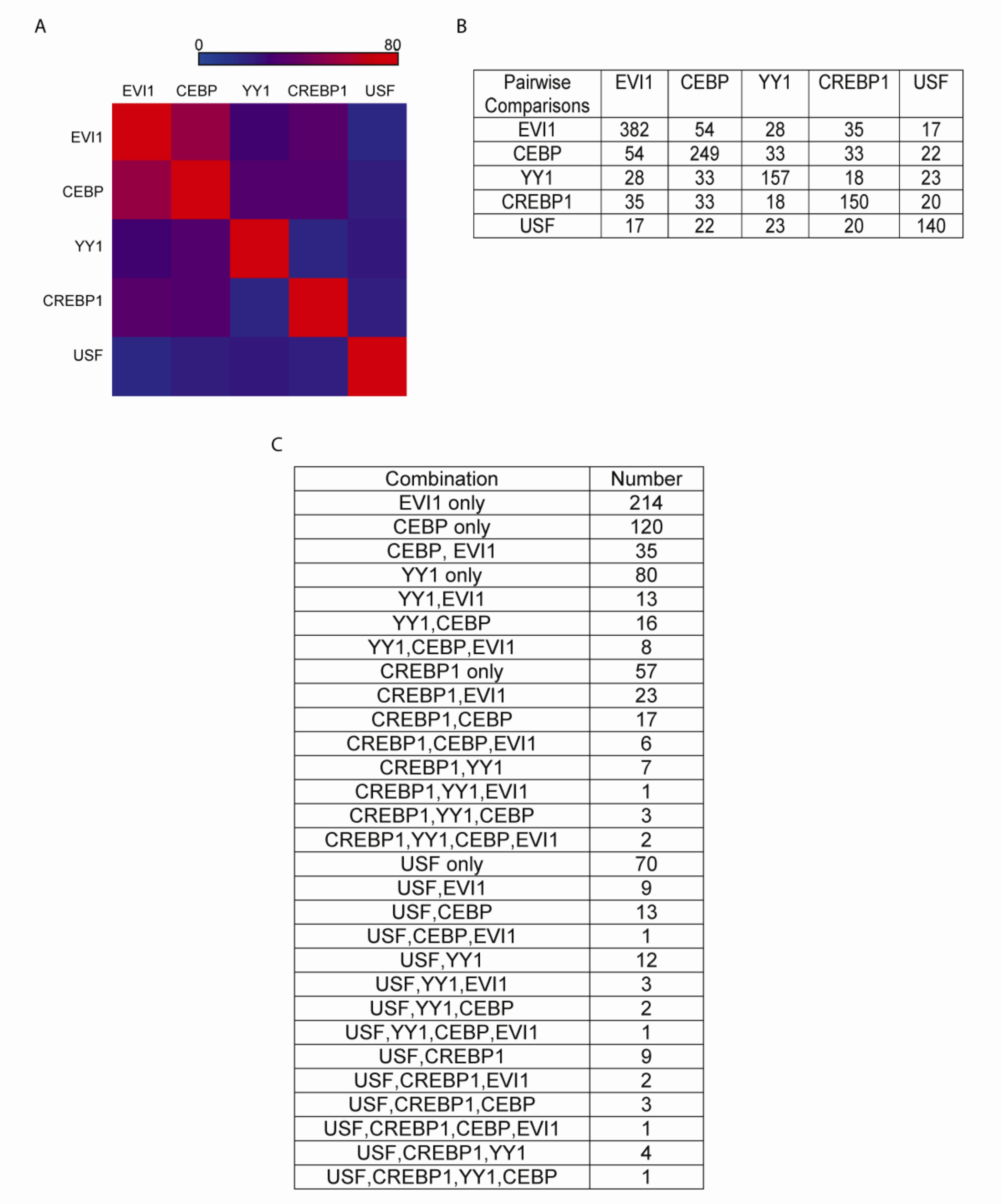


Figure D.4: The predicted boundary element overlapping with BEAD-1. The predicted boundary element is shown as the green band. ChIP-seq peaks for active and repressive histone modifications, along with the locations of euchromatic domains and heterochromatic domains are illustrated as separate tracks (as in Figure D.1).



Pairwise Comparisons	EVI1	CEBP	YY1	CREBP1	USF
EVI1	382	54	28	35	17
CEBP	54	249	33	33	22
YY1	28	33	157	18	23
CREBP1	35	33	18	150	20
USF	17	22	23	20	140

Combination	Number
EVI1 only	214
CEBP only	120
CEBP, EVI1	35
YY1 only	80
YY1,EVI1	13
YY1,CEBP	16
YY1,CEBP,EVI1	8
CREBP1 only	57
CREBP1,EVI1	23
CREBP1,CEBP	17
CREBP1,CEBP,EVI1	6
CREBP1,YY1	7
CREBP1,YY1,EVI1	1
CREBP1,YY1,CEBP	3
CREBP1,YY1,CEBP,EVI1	2
USF only	70
USF,EVI1	9
USF,CEBP	13
USF,CEBP,EVI1	1
USF,YY1	12
USF,YY1,EVI1	3
USF,YY1,CEBP	2
USF,YY1,CEBP,EVI1	1
USF,CREBP1	9
USF,CREBP1,EVI1	2
USF,CREBP1,CEBP	3
USF,CREBP1,CEBP,EVI1	1
USF,CREBP1,YY1	4
USF,CREBP1,YY1,CEBP	1

Figure D.5: Overlaps between conserved TFBSs. A. Heatmap showing the degrees of pairwise overlaps between TFBSs; B. Matrix showing the numbers of pairwise overlaps between TFBSs; C. List of numbers of all observed combinations of TFBSs.

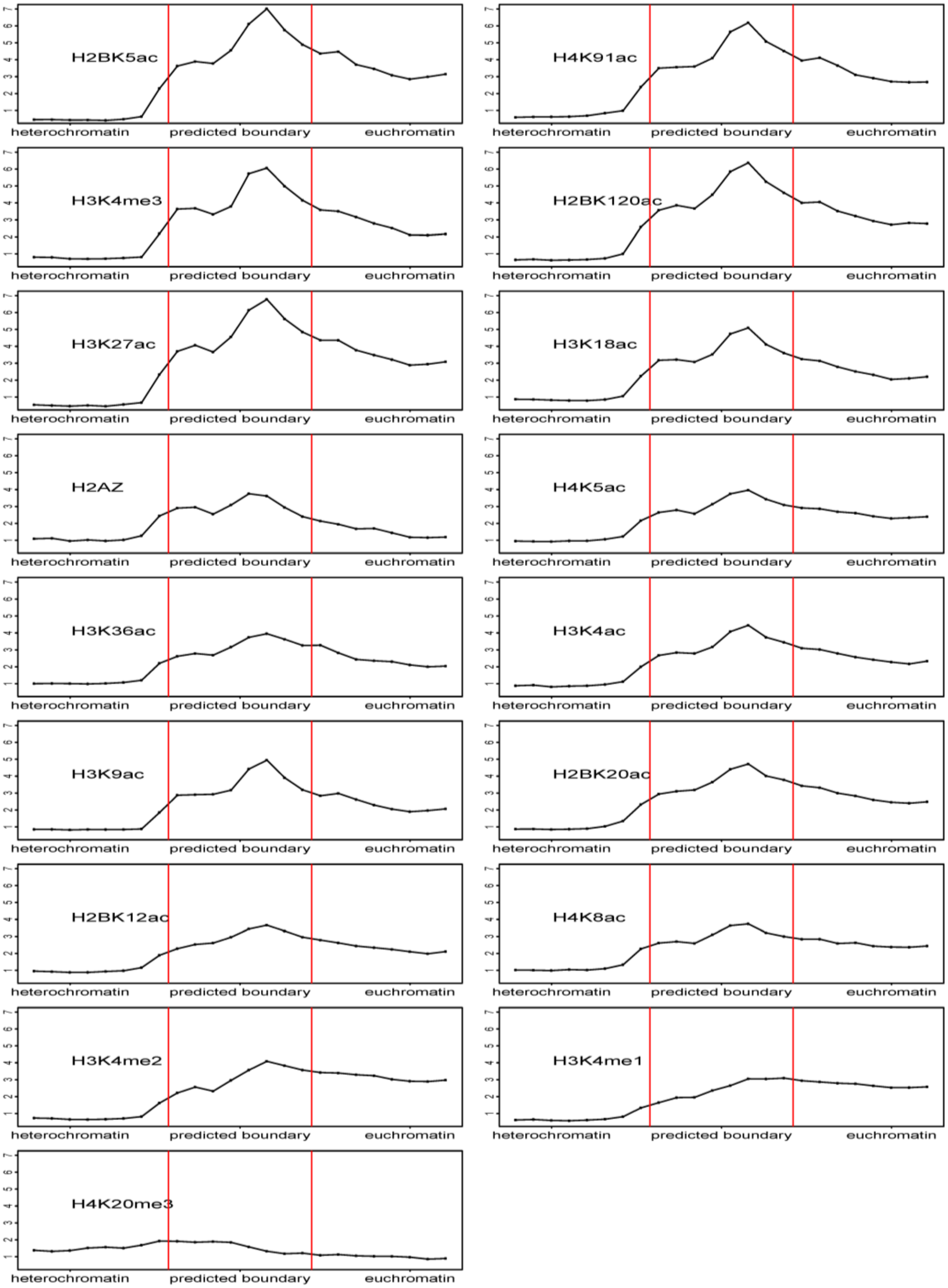


Figure D.6: Enrichment profiles around boundary elements of histone modifications which show distinct peaks. The average fold enrichments (y-axis) of individual histone modifications are ploted for the predicted boundary elements (8kb), the heterochromatin sides (8kb) and the euchromatin sides (8kb).

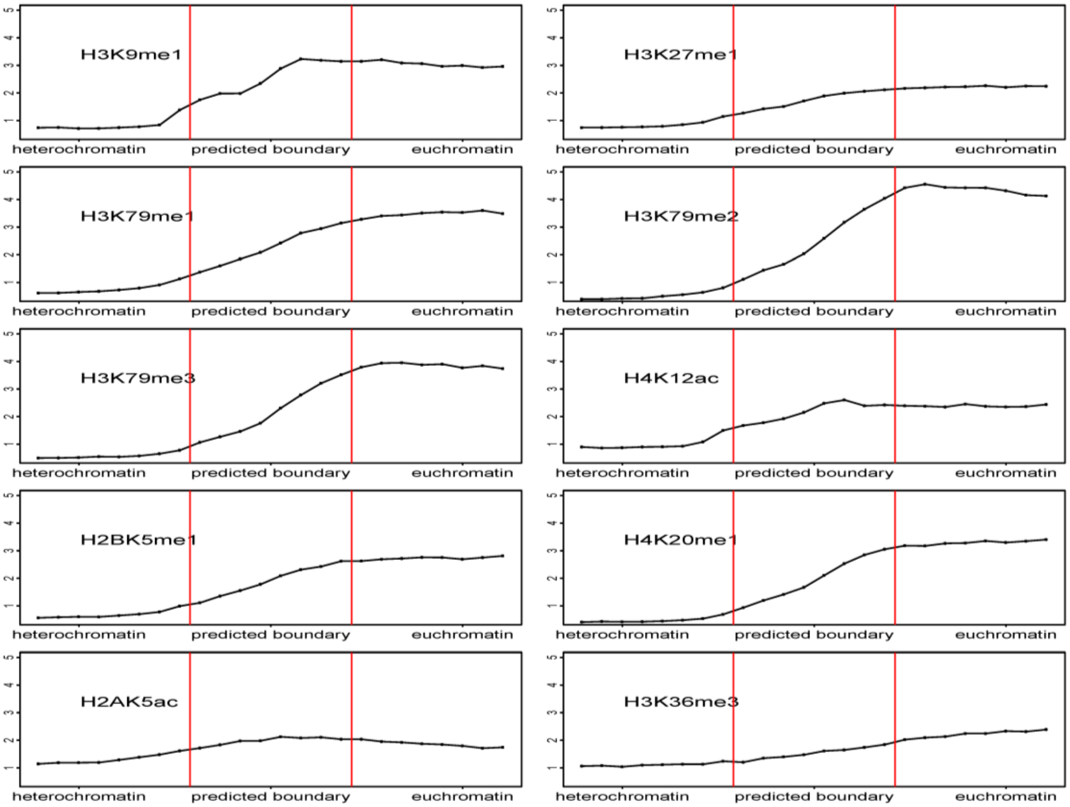


Figure D.7: Enrichment profiles around boundary elements of histone modifications which increase from heterochromatin to euchromatin. The average fold enrichments (y-axis) of individual histone modifications are ploted for the predicted boundary elements (8kb), the heterochromatin sides (8kb) and the euchromatin sides (8kb).

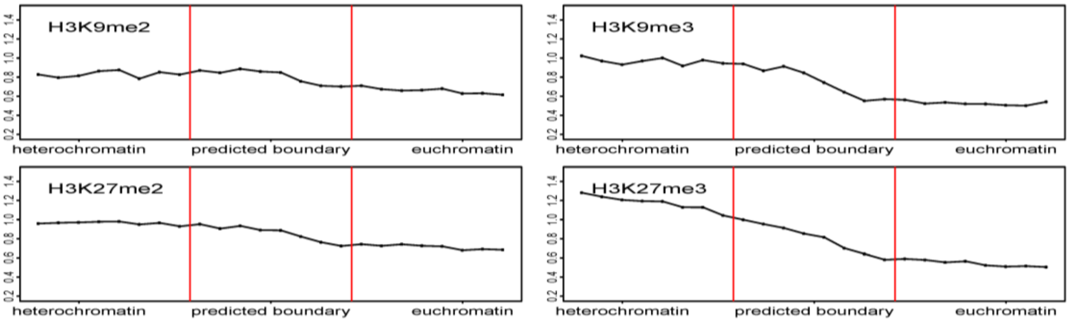


Figure D.8: Enrichment profiles around boundary elements of histone modifications which decrease from heterochromatin to euchromatin. The average fold enrichments (y-axis) of individual histone modifications are ploted for the predicted boundary elements (8kb), the heterochromatin sides (8kb) and the euchromatin sides (8kb).

APPENDIX E

SUPPLEMENTARY INFORMATION FOR CHAPTER 6

Table E.1: Enrichments of small-size combinatorial histone modification patterns with functional genomic features.

Genomic Features	No. patterns enriched with FE[a]>3	No. patterns enriched with FE>5	No. patterns enriched with FE>8
TSS[b]	36 (25.0%)	21 (14.6%)	8 (5.6%)
TTS[c]	9 (6.3%)	0	0
p300[d]	18 (12.5%)	16 (11.1%)	12 (8.3%)
DNase I[e]	60 (41.7%)	51 (35.4%)	40 (27.8%)
CNE[f]	144 (100.0%)	142 (98.6%)	137 (95.1%)

[a] FE: ratios of the fractions of patterns overlapping with the specific features over the genomic fractions of the corresponding features.
[b] TSS: transcription start site.
[c] TTS: transcription termination site.
[d] p300: binding sites of p300.
[e] DNase I: DNase I hypersensitive sites.
[f] CNE: Conserved non-coding elements predicted based on sequence alignments of 28 vertebrate species (data downloaded from UCSC genome browser).

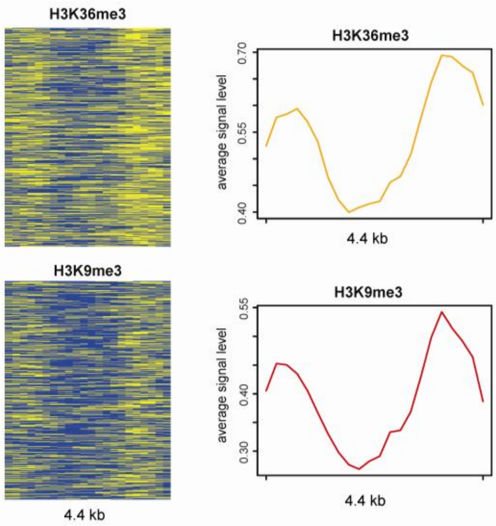


Figure E.1: Histone modification profiles for H3K36me3-H3K9me3 bivalent pattern. Genomic locations with this specific bivalent pattern are aligned and levels of H3K36me3 and H3K9me3 are shown as heatmaps on the left (yellow - higher levels, blue - lower levels). The average profiles of histone modifications of this pattern are shown on the right.

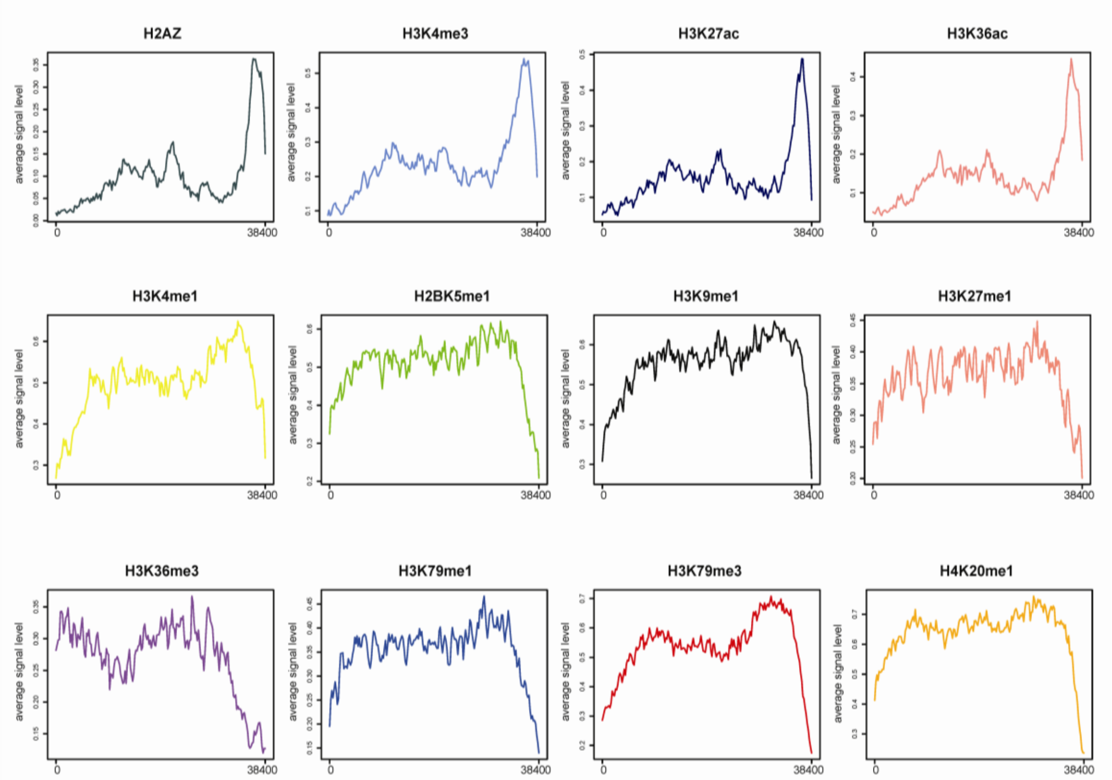


Figure E.2: Average histone modification profiles for the large pattern example A. Each curve shows the average profile of a specific histone modification of genomic locations with the same pattern.

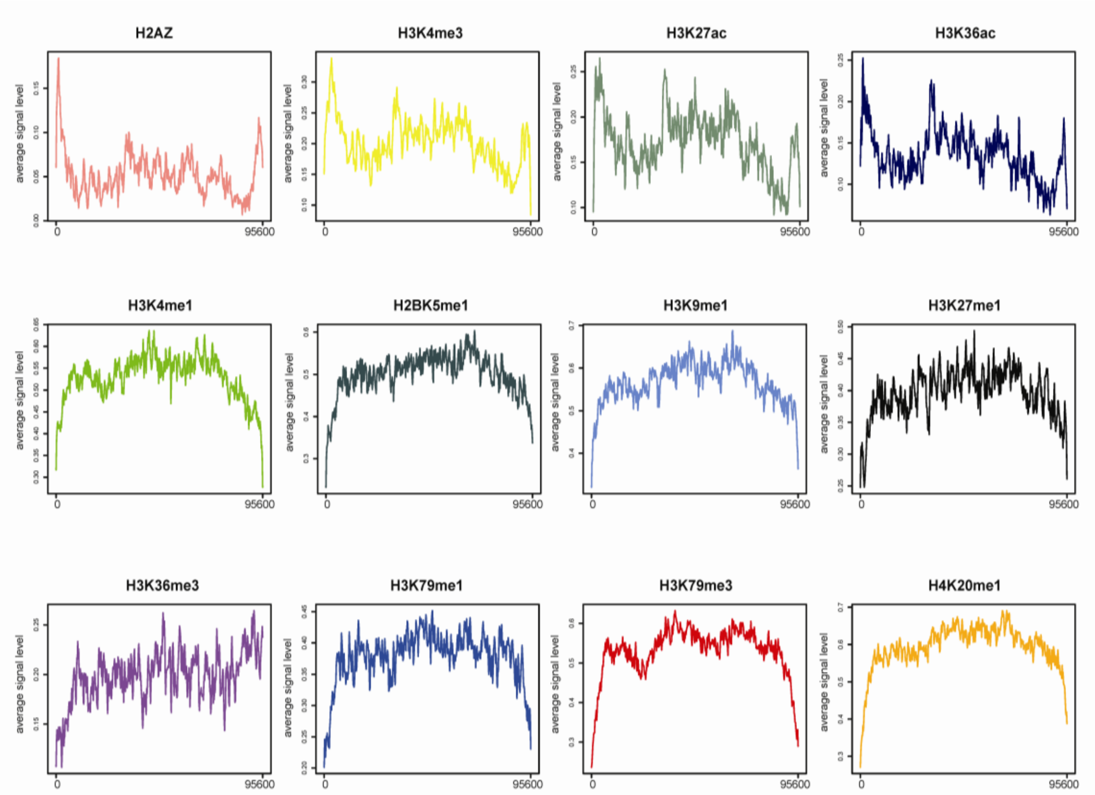


Figure E.3: Average histone modification profiles for the large pattern example B. Each curve shows the average profile of a specific histone modification of genomic locations with the same pattern.

PUBLICATIONS

1. Jianrong Wang, Victoria V. Lunyak and I. King Jordan. 2012. BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP-seq datasets. (submitted to Bioinformatics)

2. Jianrong Wang, Victoria V. Lunyak and I. King Jordan. 2012. Chromatin signature discovery via histone modification profile alignments. (in revision for Nucleic Acids Research)

3. Jianrong Wang, Cristina Vicente-García, Eduardo Moltó, Ana Fernandez-Miñán, Ana Neto, José Luis Gómez-Skarmeta, Lluís Montoliu, Elbert Lee, Victoria V. Lunyak and I. King Jordan. 2012. MIR retrotransposons provide insulators to the human genome. (in revision for PNAS)

4. Jianrong Wang, Victoria V. Lunyak and I. King Jordan. 2012. Genome-wide prediction and analysis of human chromatin boundary elements. Nucleic Acids Research 40:511-529.

5. Jianrong Wang*, Glenn J. Geesman*, Sirkka Liisa Hostikka, Michelle Atallah, Benjamin Blackwell, Elbert Lee, Peter J. Cook, Bogdan Pasaniuc, Goli Shariat, Eran Halperin, Marek Dobke, Michael G. Rosenfeld, I. King Jordan[#] and Victoria V. Lunyak[#]. 2011. Inhibition of activated pericentromeric SINE/Alu repeat transcription in senescent human adult stem cells reinstates self-renewal. Cell Cycle 10:3016-3030.

6. Jianrong Wang, Ahsan Huda, Victoria V. Lunyak and I. King Jordan. 2010. A Gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. Bioinformatics 26:2501-2508.

7. Jianrong Wang, Nathan J. Bowen, Leonardo Marino-Ramirez and I. King Jordan. 2009. A c-Myc regulatory subnetwork from human transposable element sequences. Molecular Biosystems 5:1831-1839.

REFERENCES

1. Bernstein, B.E., Meissner, A. and Lander, E.S. (2007) The mammalian epigenome. *Cell*, **128**, 669-681.

2. Kouzarides, T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693-705.

3. Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823-837.

4. Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S. and Crawford, G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311-322.

5. Eaton, M.L., Prinz, J.A., MacAlpine, H.K., Tretyakov, G., Kharchenko, P.V. and MacAlpine, D.M. (2011) Chromatin signatures of the Drosophila replication program. *Genome Res*, **21**, 164-174.

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

7. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43-49.

8. Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223-227.

9. Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*, **39**, 311-318.

10. Magklara, A., Yen, A., Colquitt, B.M., Clowney, E.J., Allen, W., Markenscoff-Papadimitriou, E., Evans, Z.A., Kheradpour, P., Mountoufaris, G., Carey, C. *et al.* (2011) An epigenetic signature for monoallelic olfactory receptor expression. *Cell*, **145**, 555-570.

11. Zentner, G.E., Tesar, P.J. and Scacheri, P.C. (2011) Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res*, **21**, 1273-1283.

12. Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., 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**, 315-326.

13. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R. *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol*, **28**, 1045-1048.

14. Gerstein, M.B., Lu, Z.J., Van Nostrand, E.L., Cheng, C., Arshinoff, B.I., Liu, T., Yip, K.Y., Robilotto, R., Rechtsteiner, A., Ikegami, K. *et al.* (2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, **330**, 1775-1787.

15. Liu, T., Rechtsteiner, A., Egelhofer, T.A., Vielle, A., Latorre, I., Cheung, M.S., Ercan, S., Ikegami, K., Jensen, M., Kolasinska-Zwierz, P. *et al.* (2011) Broad chromosomal domains of histone modification patterns in C. elegans. *Genome Res*, **21**, 227-236.

16. Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553-560.

17. Negre, N., Brown, C.D., Ma, L., Bristow, C.A., Miller, S.W., Wagner, U., Kheradpour, P., Eaton, M.L., Loriaux, P., Sealfon, R. *et al.* (2011) A cis-regulatory map of the Drosophila genome. *Nature*, **471**, 527-531.

18. Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q. *et al.* (2008) Combinatorial patterns of

histone acetylations and methylations in the human genome. *Nat Genet*, **40**, 897-903.

19. Park, P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, **10**, 669-680.

20. Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev*, **16**, 6-21.

21. Strahl, B.D. and Allis, C.D. (2000) The language of covalent histone modifications. *Nature*, **403**, 41-45.

22. Cheng, C. and Gerstein, M. (2012) Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res*, **40**, 553-568.

23. Pekowska, A., Benoukraf, T., Ferrier, P. and Spicuglia, S. (2010) A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome Res*, **20**, 1493-1502.

24. Young, M.D., Willson, T.A., Wakefield, M.J., Trounson, E., Hilton, D.J., Blewitt, M.E., Oshlack, A. and Majewski, I.J. (2011) ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Res*, **39**, 7415-7427.

25. Kharchenko, P.V., Alekseyenko, A.A., Schwartz, Y.B., Minoda, A., Riddle, N.C., Ernst, J., Sabo, P.J., Larschan, E., Gorchakov, A.A., Gu, T. *et al.* (2011) Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature*, **471**, 480-485.

26. Onder, T.T., Kara, N., Cherry, A., Sinha, A.U., Zhu, N., Bernt, K.M., Cahan, P., Marcarci, B.O., Unternaehrer, J., Gupta, P.B. *et al.* (2012) Chromatin-modifying enzymes as modulators of reprogramming. *Nature*, **483**, 598-602.

27. Firpi, H.A., Ucar, D. and Tan, K. (2010) Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics*, **26**, 1579-1586.

28. Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108-112.

29. Hon, G., Ren, B. and Wang, W. (2008) ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. *PLoS Comput Biol*, **4**, e1000201.

30. Hon, G., Wang, W. and Ren, B. (2009) Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput Biol*, **5**, e1000566.

31. Hon, G.C., Hawkins, R.D. and Ren, B. (2009) Predictive chromatin signatures in the mammalian genome. *Hum Mol Genet*, **18**, R195-201.

32. Ucar, D., Hu, Q. and Tan, K. (2011) Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering. *Nucleic Acids Res*, **39**, 4063-4075.

33. Wang, J., Lunyak, V.V. and Jordan, I.K. (2012) Genome-wide prediction and analysis of human chromatin boundary elements. *Nucleic Acids Res*, **40**, 511-529.

34. Won, K.J., Chepelev, I., Ren, B. and Wang, W. (2008) Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics*, **9**, 547.

35. Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, **18**, 1851-1858.

36. Faulkner, G.J., Forrest, A.R., Chalk, A.M., Schroder, K., Hayashizaki, Y., Carninci, P., Hume, D.A. and Grimmond, S.M. (2008) A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics*, **91**, 281-288.

37. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**, R25.

38. Qin, Z.S., Yu, J., Shen, J., Maher, C.A., Hu, M., Kalyana-Sundaram, S. and Chinnaiyan, A.M. (2010) HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics*, **11**, 369.

39. Rozowsky, J., Euskirchen, G., Auerbach, R.K., Zhang, Z.D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M. and Gerstein, M.B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol*, **27**, 66-75.

40. Song, Q. and Smith, A.D. (2011) Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics*, **27**, 870-871.

41. Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K. and Peng, W. (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, **25**, 1952-1958.

42. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, **9**, R137.

43. Lunyak, V.V. (2008) Boundaries. Boundaries...Boundaries??? *Curr Opin Cell Biol*, **20**, 281-287.

44. Gaszner, M. and Felsenfeld, G. (2006) Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet*, **7**, 703-713.

45. Wallace, J.A. and Felsenfeld, G. (2007) We gather together: insulators and genome organization. *Curr Opin Genet Dev*, **17**, 400-407.

46. Gdula, D.A., Gerasimova, T.I. and Corces, V.G. (1996) Genetic and molecular analysis of the gypsy chromatin insulator of Drosophila. *Proc Natl Acad Sci U S A*, **93**, 9378-9383.

47. Huang, S., Li, X., Yusufzai, T.M., Qiu, Y. and Felsenfeld, G. (2007) USF1 recruits histone modification complexes and is critical for maintenance of a chromatin barrier. *Mol Cell Biol*, **27**, 7991-8002.

48. West, A.G., Gaszner, M. and Felsenfeld, G. (2002) Insulators: many functions, many mechanisms. *Genes Dev*, **16**, 271-288.

49. Chung, J.H., Whiteley, M. and Felsenfeld, G. (1993) A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. *Cell*, **74**, 505-514.

50. Yusufzai, T.M., Tagami, H., Nakatani, Y. and Felsenfeld, G. (2004) CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Mol Cell*, **13**, 291-298.

51. Bell, A.C., West, A.G. and Felsenfeld, G. (1999) The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, **98**, 387-396.

52. Noma, K., Cam, H.P., Maraia, R.J. and Grewal, S.I. (2006) A role for TFIIIC transcription factor complex in genome organization. *Cell*, **125**, 859-872.

53. Donze, D. and Kamakaka, R.T. (2001) RNA polymerase III and RNA polymerase II promoter complexes are heterochromatin barriers in Saccharomyces cerevisiae. *EMBO J*, **20**, 520-531.

54. Simms, T.A., Dugas, S.L., Gremillion, J.C., Ibos, M.E., Dandurand, M.N., Toliver, T.T., Edwards, D.J. and Donze, D. (2008) TFIIIC binding sites function as both heterochromatin barriers and chromatin insulators in Saccharomyces cerevisiae. *Eukaryot Cell*, **7**, 2078-2086.

55. Ebersole, T., Kim, J.H., Samoshkin, A., Kouprina, N., Pavlicek, A., White, R.J. and Larionov, V. (2011) tRNA genes protect a reporter gene from epigenetic silencing in mouse cells. *Cell Cycle*, **10**, 2779-2791.

56. Lunyak, V.V., Prefontaine, G.G., Nunez, E., Cramer, T., Ju, B.G., Ohgi, K.A., Hutt, K., Roy, R., Garcia-Diaz, A., Zhu, X. *et al.* (2007) Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science*, **317**, 248-251.

57. Raab, J.R. and Kamakaka, R.T. (2010) Insulators and promoters: closer than we think. *Nat Rev Genet*, **11**, 439-446.

58. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376-380.

59. Botta, M., Haider, S., Leung, I.X., Lio, P. and Mozziconacci, J. (2010) Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. *Mol Syst Biol*, **6**, 426.

60. Splinter, E., Heath, H., Kooren, J., Palstra, R.J., Klous, P., Grosveld, F., Galjart, N. and de Laat, W. (2006) CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev*, **20**, 2349-2354.

61. Burke, L.J., Zhang, R., Bartkuhn, M., Tiwari, V.K., Tavoosidana, G., Kurukuti, S., Weth, C., Leers, J., Galjart, N., Ohlsson, R. *et al.* (2005) CTCF binding and higher order chromatin structure of the H19 locus are maintained in mitotic chromatin. *EMBO J*, **24**, 3291-3300.

62. Ernst, J. and Kellis, M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*, **9**, 215-216.

63. Bock, C. and Lengauer, T. (2008) Computational epigenetics. *Bioinformatics*, **24**, 1-10.

64. Thurman, R.E., Day, N., Noble, W.S. and Stamatoyannopoulos, J.A. (2007) Identification of higher-order functional domains in the human ENCODE regions. *Genome Res*, **17**, 917-927.

65. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, **9**, R137.

66. Feschotte, C. (2008) Transposable elements and the evolution of regulatory networks. *Nat Rev Genet*, **9**, 397-405.

67. Hashimoto, T., de Hoon, M.J., Grimmond, S.M., Daub, C.O., Hayashizaki, Y. and Faulkner, G.J. (2009) Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using MuMRescueLite. *Bioinformatics*, **25**, 2613-2614.

68. Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F. and Wootton, J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208-214.

69. Neuwald, A.F., Liu, J.S. and Lawrence, C.E. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci*, **4**, 1618-1632.

70. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D. and Kent, W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*, **32**, D493-496.

71. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res*, **12**, 996-1006.

72. Karlin, S. and Altschul, S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A*, **87**, 2264-2268.

73. Ruzzo, W.L. and Tompa, M. (1999) A linear time algorithm for finding all maximal scoring subsequences. *Proc Int Conf Intell Syst Mol Biol*, 234-241.

74. Valenzuela, L. and Kamakaka, R.T. (2006) Chromatin insulators. *Annu Rev Genet*, **40**, 107-138.

75. Capelson, M. and Corces, V.G. (2004) Boundary elements and nuclear organization. *Biol Cell*, **96**, 617-629.

76. Gerasimova, T.I., Gdula, D.A., Gerasimov, D.V., Simonova, O. and Corces, V.G. (1995) A Drosophila protein that imparts directionality on a chromatin insulator is an enhancer of position-effect variegation. *Cell*, **82**, 587-597.

77. Geyer, P.K. and Corces, V.G. (1992) DNA position-specific repression of transcription by a Drosophila zinc finger protein. *Genes Dev*, **6**, 1865-1873.

78. Labrador, M. and Corces, V.G. (2002) Setting the boundaries of chromatin domains and nuclear organization. *Cell*, **111**, 151-154.

79. Roman, A.C., Gonzalez-Rico, F.J., Molto, E., Hernando, H., Neto, A., Vicente-Garcia, C., Ballestar, E., Gomez-Skarmeta, J.L., Vavrova-Anderson, J., White, R.J. *et al.* (2011) Dioxin receptor and SLUG transcription factors regulate the insulator activity of B1 SINE retrotransposons via an RNA polymerase switch. *Genome Res*, **21**, 422-432.

80. Oki, M. and Kamakaka, R.T. (2005) Barrier function at HMR. *Mol Cell*, **19**, 707-716.

81. Scott, K.C., Merrett, S.L. and Willard, H.F. (2006) A heterochromatin barrier partitions the fission yeast centromere into discrete chromatin domains. *Curr Biol*, **16**, 119-129.

82. Raab, J.R., Chiu, J., Zhu, J., Katzman, S., Kurukuti, S., Wade, P.A., Haussler, D. and Kamakaka, R.T. (2011) Human tRNA genes function as chromatin insulators. *EMBO J*, **31**, 330-350.

83. Schmidt, D., Schwalie, P.C., Wilson, M.D., Ballester, B., Goncalves, A., Kutter, C., Brown, G.D., Marshall, A., Flicek, P. and Odom, D.T. (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, **148**, 335-348.

84. de Koning, A.P., Gu, W., Castoe, T.A., Batzer, M.A. and Pollock, D.D. (2011) Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet*, **7**, e1002384.

85. Jurka, J., Zietkiewicz, E. and Labuda, D. (1995) Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. *Nucleic Acids Res*, **23**, 170-175.

86. Silva, J.C., Shabalina, S.A., Harris, D.G., Spouge, J.L. and Kondrashovi, A.S. (2003) Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. *Genet Res*, **82**, 1-18.

87. Marino-Ramirez, L. and Jordan, I.K. (2006) Transposable element derived DNaseI-hypersensitive sites in the human genome. *Biol Direct*, **1**, 20.

88. Piriyapongsa, J., Marino-Ramirez, L. and Jordan, I.K. (2007) Origin and evolution of human microRNAs from transposable elements. *Genetics*, **176**, 1323-1337.

89. Huda, A., Bowen, N.J., Conley, A.B. and Jordan, I.K. (2011) Epigenetic regulation of transposable element derived human gene promoters. *Gene*, **475**, 39-48.

90. Huda, A., Tyagi, E., Marino-Ramirez, L., Bowen, N.J., Jjingo, D. and Jordan, I.K. (2011) Prediction of transposable element derived enhancers using chromatin modification profiles. *PLoS One*, **6**, e27513.

91. Jjingo, D., Huda, A., Gundapuneni, M., Marino-Ramirez, L. and Jordan, I.K. (2011) Effect of the transposable element environment of human genes on gene length and expression. *Genome Biol Evol*, **3**, 259-271.

92. Smit, A.F. and Riggs, A.D. (1995) MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res*, **23**, 98-102.

93. Tiana, M., Villar, D., Perez-Guijarro, E., Gomez-Maldonado, L., Molto, E., Fernandez-Minan, A., Gomez-Skarmeta, J.L., Montoliu, L. and Del Peso, L. (2011) A role for insulator elements in the regulation of gene expression response to hypoxia. *Nucleic Acids Res*.

94. Martin, D., Pantoja, C., Fernandez Minan, A., Valdes-Quezada, C., Molto, E., Matesanz, F., Bogdanovic, O., de la Calle-Mustienes, E., Dominguez, O., Taher, L. *et al.* (2011) Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes. *Nat Struct Mol Biol*, **18**, 708-714.

95. Barski, A., Chepelev, I., Liko, D., Cuddapah, S., Fleming, A.B., Birch, J., Cui, K., White, R.J. and Zhao, K. (2010) Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. *Nat Struct Mol Biol*, **17**, 629-634.

96. McClintock, B. (1951) Chromosome organization and genic expression. *Cold Spring Harb Symp Quant Biol*, **16**, 13-47.

97. Britten, R.J. and Davidson, E.H. (1971) Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q Rev Biol*, **46**, 111-138.

98. Thomas, D.J., Rosenbloom, K.R., Clawson, H., Hinrichs, A.S., Trumbower, H., Raney, B.J., Karolchik, D., Barber, G.P., Harte, R.A., Hillman-Jackson, J. *et al.* (2007) The ENCODE Project at UC Santa Cruz. *Nucleic Acids Res*, **35**, D663-667.

99. Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, **101**, 6062-6067.

100. Bemmo, A., Benovoy, D., Kwan, T., Gaffney, D.J., Jensen, R.V. and Majewski, J. (2008) Gene expression and isoform variation analysis using Affymetrix Exon Arrays. *BMC Genomics*, **9**, 529.

101. Gardina, P.J., Clark, T.A., Shimada, B., Staples, M.K., Yang, Q., Veitch, J., Schweitzer, A., Awad, T., Sugnet, C., Dee, S. *et al.* (2006) Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics*, **7**, 325.

102. Recillas-Targa, F., Bell, A.C. and Felsenfeld, G. (1999) Positional enhancer-blocking activity of the chicken beta-globin insulator in transiently transfected cells. *Proc Natl Acad Sci U S A*, **96**, 14354-14359.

103. Bessa, J., Tena, J.J., de la Calle-Mustienes, E., Fernandez-Minan, A., Naranjo, S., Fernandez, A., Montoliu, L., Akalin, A., Lenhard, B., Casares, F. *et al.* (2009) Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. *Dev Dyn*, **238**, 2409-2417.

104. Pauler, F.M., Sloane, M.A., Huang, R., Regha, K., Koerner, M.V., Tamir, I., Sommer, A., Aszodi, A., Jenuwein, T. and Barlow, D.P. (2009) H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome Res*, **19**, 221-233.

105. Bernstein, B.E., Humphrey, E.L., Erlich, R.L., Schneider, R., Bouman, P., Liu, J.S., Kouzarides, T. and Schreiber, S.L. (2002) Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc Natl Acad Sci U S A*, **99**, 8695-8700.

106. Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., 3rd, Gingeras, T.R. *et al.* (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, **120**, 169-181.

107. Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K. *et al.* (2006) Polycomb

complexes repress developmental regulators in murine embryonic stem cells. *Nature*, **441**, 349-353.

108. Roh, T.Y., Cuddapah, S., Cui, K. and Zhao, K. (2006) The genomic landscape of histone modifications in human T cells. *Proc Natl Acad Sci U S A*, **103**, 15782-15787.

109. Roh, T.Y., Cuddapah, S. and Zhao, K. (2005) Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev*, **19**, 542-552.

110. Dillon, N. and Sabbattini, P. (2000) Functional gene expression domains: defining the functional unit of eukaryotic gene regulation. *Bioessays*, **22**, 657-665.

111. Kamakaka, R.T. and Thomas, J.O. (1990) Chromatin structure of transcriptionally competent and repressed genes. *EMBO J*, **9**, 3997-4006.

112. Kellum, R. and Schedl, P. (1991) A position-effect assay for boundaries of higher order chromosomal domains. *Cell*, **64**, 941-950.

113. Recillas-Targa, F., Pikaart, M.J., Burgess-Beusse, B., Bell, A.C., Litt, M.D., West, A.G., Gaszner, M. and Felsenfeld, G. (2002) Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities. *Proc Natl Acad Sci U S A*, **99**, 6883-6888.

114. Udvardy, A., Maine, E. and Schedl, P. (1985) The 87A7 chromomere. Identification of novel chromatin structures flanking the heat shock locus that may define the boundaries of higher order domains. *J Mol Biol*, **185**, 341-358.

115. Noma, K., Allis, C.D. and Grewal, S.I. (2001) Transitions in distinct histone H3 methylation patterns at the heterochromatin domain boundaries. *Science*, **293**, 1150-1155.

116. Pikaart, M.J., Recillas-Targa, F. and Felsenfeld, G. (1998) Loss of transcriptional activity of a transgene is accompanied by DNA methylation and histone deacetylation and is prevented by insulators. *Genes Dev*, **12**, 2852-2862.

117. Kellum, R. and Schedl, P. (1992) A group of scs elements function as domain boundaries in an enhancer-blocking assay. *Mol Cell Biol*, **12**, 2424-2431.

118. Zhao, K., Hart, C.M. and Laemmli, U.K. (1995) Visualization of chromosomal domains with boundary element-associated factor BEAF-32. *Cell*, **81**, 879-889.

119. Fourel, G., Magdinier, F. and Gilson, E. (2004) Insulator dynamics and the setting of chromatin domains. *Bioessays*, **26**, 523-532.

120. Kimura, A. and Horikoshi, M. (2004) Partition of distinct chromosomal regions: negotiable border and fixed border. *Genes Cells*, **9**, 499-508.

121. Henikoff, S. (1990) Position-effect variegation after 60 years. *Trends Genet*, **6**, 422-426.

122. Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.Y., Cui, K. and Zhao, K. (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res*, **19**, 24-32.

123. Riddle, N.C., Minoda, A., Kharchenko, P.V., Alekseyenko, A.A., Schwartz, Y.B., Tolstorukov, M.Y., Gorchakov, A.A., Jaffe, J.D., Kennedy, C., Linder-Basso, D. *et al.* (2011) Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. *Genome Res*, **21**, 147-163.

124. Rosenfeld, J.A., Wang, Z., Schones, D.E., Zhao, K., DeSalle, R. and Zhang, M.Q. (2009) Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics*, **10**, 143.

125. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*, **31**, 51-54.

126. Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M.I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., Muller, P. *et al.* (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, **408**, 86-89.

127. Mourelatos, Z., Dostie, J., Paushkin, S., Sharma, A., Charroux, B., Abel, L., Rappsilber, J., Mann, M. and Dreyfuss, G. (2002) miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev*, **16**, 720-728.

128. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**, 15545-15550.

129. Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, **34**, 267-273.

130. Cho, H., Orphanides, G., Sun, X., Yang, X.J., Ogryzko, V., Lees, E., Nakatani, Y. and Reinberg, D. (1998) A human RNA polymerase II complex containing factors that modify chromatin structure. *Mol Cell Biol*, **18**, 5355-5363.

131. Phillips, J.E. and Corces, V.G. (2009) CTCF: master weaver of the genome. *Cell*, **137**, 1194-1211.

132. Zhong, X.P. and Krangel, M.S. (1997) An enhancer-blocking element between alpha and delta gene segments within the human T cell receptor alpha/delta locus. *Proc Natl Acad Sci U S A*, **94**, 5219-5224.

133. Bushey, A.M., Dorman, E.R. and Corces, V.G. (2008) Chromatin insulators: regulatory mechanisms and epigenetic inheritance. *Mol Cell*, **32**, 1-9.

134. Spensberger, D. and Delwel, R. (2008) A novel interaction between the proto-oncogene Evi1 and histone methyltransferases, SUV39H1 and G9a. *FEBS Lett*, **582**, 2761-2767.

135. Chakraborty, S., Senyuk, V., Sitailo, S., Chi, Y. and Nucifora, G. (2001) Interaction of EVI1 with cAMP-responsive element-binding protein-binding protein (CBP) and p300/CBP-associated factor (P/CAF) results in reversible acetylation of EVI1 and in co-localization in nuclear speckles. *J Biol Chem*, **276**, 44936-44943.

136. Bruhat, A., Cherasse, Y., Maurin, A.C., Breitwieser, W., Parry, L., Deval, C., Jones, N., Jousse, C. and Fafournoux, P. (2007) ATF2 is required for amino acid-regulated transcription by orchestrating specific histone acetylation. *Nucleic Acids Res*, **35**, 1312-1321.

137. Karanam, B., Wang, L., Wang, D., Liu, X., Marmorstein, R., Cotter, R. and Cole, P.A. (2007) Multiple roles for acetylation in the interaction of p300 HAT with ATF-2. *Biochemistry*, **46**, 8207-8216.

138. Sano, Y., Tokitou, F., Dai, P., Maekawa, T., Yamamoto, T. and Ishii, S. (1998) CBP alleviates the intramolecular inhibition of ATF-2 function. *J Biol Chem*, **273**, 29098-29105.

139. Kovacs, K.A., Steinmann, M., Magistretti, P.J., Halfon, O. and Cardinaux, J.R. (2003) CCAAT/enhancer-binding protein family members recruit the coactivator CREB-binding protein and trigger its phosphorylation. *J Biol Chem*, **278**, 36959-36965.

140. Yao, Y.L., Yang, W.M. and Seto, E. (2001) Regulation of transcription factor YY1 by acetylation and deacetylation. *Mol Cell Biol*, **21**, 5979-5991.

141. West, A.G., Huang, S., Gaszner, M., Litt, M.D. and Felsenfeld, G. (2004) Recruitment of histone modifications by USF proteins at a vertebrate barrier element. *Mol Cell*, **16**, 453-463.

142. Cuddapah, S., Schones, D.E., Cui, K., Roh, T.Y., Barski, A., Wei, G., Rochman, M., Bustin, M. and Zhao, K. (2011) Genomic profiling of HMGN1 reveals an association with chromatin at regulatory regions. *Mol Cell Biol*, **31**, 700-709.

143. Donohoe, M.E., Zhang, L.F., Xu, N., Shi, Y. and Lee, J.T. (2007) Identification of a Ctcf cofactor, Yy1, for the X chromosome binary switch. *Mol Cell*, **25**, 43-56.

144. Austen, M., Luscher, B. and Luscher-Firzlaff, J.M. (1997) Characterization of the transcriptional regulator YY1. The bipartite transactivation domain is independent of interaction with the TATA box-binding protein, transcription factor IIB, TAFII55, or cAMP-responsive element-binding protein (CPB)-binding protein. *J Biol Chem*, **272**, 1709-1717.

145. Galvin, K.M. and Shi, Y. (1997) Multiple mechanisms of transcriptional repression by YY1. *Mol Cell Biol*, **17**, 3723-3732.

146. Lee, J.S., Galvin, K.M., See, R.H., Eckner, R., Livingston, D., Moran, E. and Shi, Y. (1995) Relief of YY1 transcriptional repression by adenovirus E1A is mediated by E1A-associated protein p300. *Genes Dev*, **9**, 1188-1198.

147. Shi, Y., Lee, J.S. and Galvin, K.M. (1997) Everything you have ever wanted to know about Yin Yang 1. *Biochim Biophys Acta*, **1332**, F49-66.

148. Thomas, M.J. and Seto, E. (1999) Unlocking the mechanisms of transcription factor YY1: are chromatin modifying enzymes the key? *Gene*, **236**, 197-208.

149. Yang, W.M., Inouye, C., Zeng, Y., Bearss, D. and Seto, E. (1996) Transcriptional repression by YY1 is mediated by interaction with a mammalian homolog of the yeast global regulator RPD3. *Proc Natl Acad Sci U S A*, **93**, 12845-12850.

150. Yang, W.M., Yao, Y.L., Sun, J.M., Davie, J.R. and Seto, E. (1997) Isolation and characterization of cDNAs corresponding to an additional member of the human histone deacetylase gene family. *J Biol Chem*, **272**, 28001-28007.

151. Bushmeyer, S.M. and Atchison, M.L. (1998) Identification of YY1 sequences necessary for association with the nuclear matrix and for transcriptional repression functions. *J Cell Biochem*, **68**, 484-499.

152. Guo, B., Odgren, P.R., van Wijnen, A.J., Last, T.J., Nickerson, J., Penman, S., Lian, J.B., Stein, J.L. and Stein, G.S. (1995) The nuclear matrix protein NMP-1 is the transcription factor YY1. *Proc Natl Acad Sci U S A*, **92**, 10526-10530.

153. Ebersole, T., Kim, J.H., Samoshkin, A., Kouprina, N., Pavlicek, A., White, R.J. and Larionov, V. (2011) tRNA genes protect a reporter gene from epigenetic silencing in mouse cells. *Cell Cycle*, **10**.

154. Chiu, Y.H., Yu, Q., Sandmeier, J.J. and Bi, X. (2003) A targeted histone acetyltransferase can create a sizable region of hyperacetylated chromatin and counteract the propagation of transcriptionally silent chromatin. *Genetics*, **165**, 115-125.

155. Donze, D. and Kamakaka, R.T. (2002) Braking the silence: how heterochromatic gene repression is stopped in its tracks. *Bioessays*, **24**, 344-349.

156. Feng, Q., Wang, H., Ng, H.H., Erdjument-Bromage, H., Tempst, P., Struhl, K. and Zhang, Y. (2002) Methylation of H3-lysine 79 is mediated by a new family of HMTases without a SET domain. *Curr Biol*, **12**, 1052-1058.

157. Steger, D.J., Lefterova, M.I., Ying, L., Stonestrom, A.J., Schupp, M., Zhuo, D., Vakoc, A.L., Kim, J.E., Chen, J., Lazar, M.A. *et al.* (2008) DOT1L/KMT4 recruitment and H3K79 methylation are ubiquitously coupled with gene transcription in mammalian cells. *Mol Cell Biol*, **28**, 2825-2839.

158. Ng, H.H., Ciccone, D.N., Morshead, K.B., Oettinger, M.A. and Struhl, K. (2003) Lysine-79 of histone H3 is hypomethylated at silenced loci in yeast and mammalian cells: a potential mechanism for position-effect variegation. *Proc Natl Acad Sci U S A*, **100**, 1820-1825.

159. Meisterernst, M. and Roeder, R.G. (1991) Family of proteins that interact with TFIID and regulate promoter activity. *Cell*, **67**, 557-567.

160. Consortium, T.m., Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L. *et al.* (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, **330**, 1787-1797.

161. Kidder, B.L., Hu, G. and Zhao, K. (2011) ChIP-Seq: technical considerations for obtaining high-quality data. *Nature immunology*, **12**, 918-922.

162. Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A. and Noble, W.S. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods*.

163. Altschul, S.F., Bundschuh, R., Olsen, R. and Hwa, T. (2001) The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res*, **29**, 351-361.

164. Wang, Z., Zang, C., Cui, K., Schones, D.E., Barski, A., Peng, W. and Zhao, K. (2009) Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell*, **138**, 1019-1031.

165. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, **15**, 1034-1050.

166. Allen, E., Horvath, S., Tong, F., Kraft, P., Spiteri, E., Riggs, A.D. and Marahrens, Y. (2003) High concentrations of long interspersed nuclear element sequence

distinguish monoallelically expressed genes. *Proc Natl Acad Sci U S A*, **100**, 9940-9945.

167. Luedi, P.P., Dietrich, F.S., Weidman, J.R., Bosko, J.M., Jirtle, R.L. and Hartemink, A.J. (2007) Computational and experimental identification of novel human imprinted genes. *Genome Res*, **17**, 1723-1730.

168. Luedi, P.P., Hartemink, A.J. and Jirtle, R.L. (2005) Genome-wide prediction of imprinted murine genes. *Genome Res*, **15**, 875-884.

169. Walter, J., Hutter, B., Khare, T. and Paulsen, M. (2006) Repetitive elements in imprinted genes. *Cytogenet Genome Res*, **113**, 109-115.

170. Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, 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.

171. Holstege, F.C., Fiedler, U. and Timmers, H.T. (1997) Three transitions in the RNA polymerase II transcription complex during initiation. *EMBO J*, **16**, 7468-7480.

172. Rahl, P.B., Lin, C.Y., Seila, A.C., Flynn, R.A., McCuine, S., Burge, C.B., Sharp, P.A. and Young, R.A. (2010) c-Myc regulates transcriptional pause release. *Cell*, **141**, 432-445.

173. Reppas, N.B., Wade, J.T., Church, G.M. and Struhl, K. (2006) The transition between transcriptional initiation and elongation in E. coli is highly variable and often rate limiting. *Mol Cell*, **24**, 747-757.

174. Zeitlinger, J., Stark, A., Kellis, M., Hong, J.W., Nechaev, S., Adelman, K., Levine, M. and Young, R.A. (2007) RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat Genet*, **39**, 1512-1516.