Identifying chromatin interactions at high spatial resolution

by

Christopher Campbell Reeder

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2014

Signature redacted

Author .....
Department of Electrical Engineering and Computer Science
May 21, 2014

Signature redacted

Certified by .....
David K. Gifford
Professor
Thesis Supervisor

Signature redacted

Accepted by .....
Leslie A. Kolodziejski
Chair, Department Committee on Graduate Theses

Identifying chromatin interactions at high spatial resolution

by

Christopher Campbell Reeder

Submitted to the Department of Electrical Engineering and Computer Science
on May 21, 2014, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science

Abstract

This thesis presents two computational approaches for identifying chromatin interactions at high spatial resolution from ChIA-PET data. We introduce SPROUT which is a hierarchical probabilistic model that discovers high confidence interactions between binding events that it accurately locates. We apply SPROUT to CTCF ChIA-PET data from mouse embryonic stem cells and demonstrate that SPROUT discovers interactions that are more consistently supported by biological replicates than an alternative method called The ChIA-PET Tool. We also introduce GERM which models genome-wide distributions of protein occupancy without assuming that proteins can be accurately modeled as binding to point locations. We demonstrate that the locations that GERM identifies as interacting with transcription start sites of genes accurately align with ChIP-Seq data that are associated with active enhancers. Finally, we apply GERM to RNA Polymerase II ChIA-PET data from embryonic stem cells and motor neuron progenitors and make several observations about the usage of enhancers during motor neuron development.

Thesis Supervisor: David K. Gifford
Title: Professor

Acknowledgments

This thesis represents the end of a long journey. I surely would not have gotten to this point without the guidance and support of many people. During my undergraduate education I was very fortunate to have apprenticed with Rainer Sachs. It amazes me looking back on the patience with which he guided me through my first forays into applying computational methods to genomic data. He leads by example with an amazing work ethic and it is hard for me to imagine a better mentor. I was also very fortunate to work with Stuart Russell who introduced me to the level of dedication and attention to detail that would be required of me in graduate school.

During graduate school I have interacted with several professors who have all impacted the way that I think about science and research. Tommi Jaakkola taught one of the first classes that I took at MIT which embodied the "drinking from the fire hose" type of learning that is typical of the MIT experience. I later had the pleasure of working with him as a teaching assistant. Ernest Fraenkel and his research group made me feel very welcome at their group meetings. He is a very effective leader and to see the way in which experimental and computational scientists work together productively in his group is very exciting. David Gifford is an excellent communicator of science. He is able to successfully communicate complicated scientific concepts to audiences ranging from the general public to undergraduates to other professors. Without his ability to establish and maintain collaborations and to identify important scientific questions this thesis would not have come to fruition.

Many other individuals have been crucial to my survival of graduate school. Jeanne Darling has been a source of kindness and support during times of immense stress. Patrice Macaluso has also been very helpful. I can't thank Jeanne and Patrice enough. Janet Fischer has helped me navigate the various stages of my graduate program. Leslie Kolodziejski provided a beacon of hope during the very challenging time leading up to the completion of this thesis. I have had many fantastic labmates over the years. Tim Danford and Alex Rolfe had enormous impacts on how I have approached research and much of my work has relied on systems that they established.

Georg Gerber acted as a research mentor to me during my earliest time at MIT and helped me discover interesting problems and methods that became important parts of my graduate work. Shaun Mahony seemed always willing to chat about interesting research ideas, exhibited great leadership, and developed several tools to help the group do research more efficiently. Charlie O'Donnell and I had many productive and unproductive conversation about research and academia. Yuchun Guo and I shared many of the same experiences working in the Gifford Lab. I have valued our conversations about science and life in general and attending RECOMB in Beijing with him was one of the highlights of my graduate experience. Tatsu Hashimoto exemplifies the future of computational biology. His intellectual prowess is daunting and how he is able to be so productive and still be the instigator of trips to the Muddy is beyond me. I have had the great pleasure of having Matt Edwards as an officemate. He is a very talented researcher and all around great guy. Its hard to beat sitting next to a good friend who has provided crucial assistance with my research and with whom I have had crucial diversionary conversations.

I most definitely would not have made it through the last few years without my close friends. Michael Carbin has had a huge impact on how I think about life. Many of my best all time life memories involve Mike. It was incredible to reconnect with my old friend Rob Rykowski. Having a close friend who lives nearby and with whom I share many childhood memories is a very special thing. My first roommate in graduate school was Joseph Kovac. He has been a friend for my entire graduate school experience. I have trusted my life to him literally while on climbing adventures and he has provided essential guidance for surviving MIT. Sam Weiner and Alex Strauss have helped keep me connected to my roots. Friends that I met since moving to Massachusetts such as Jeff Sayabovorn and Drew Bisset have helped me to cultivate new interests. Fellow MIT students such as Giorgos Papachristoudis, Bob Altshuler, and Emily Shen have shared many common experiences.

Family, in various senses of the word, have been my backbone through this whole experience. It has been wonderful having the Dudes nearby to keep me connected to my California family and to share their love of the outdoors with me. I have also

enjoyed getting to know the Browns and the Schuberts. They have welcomed me into their homes and provided a sense of family right here in the Boston area. I am so fortunate to have two strong, independent, intelligent women as grandmothers. My entire extended family has been nothing but supportive throughout this experience and it has been awesome to watch my cousins grow up and achieve amazing things. As strange as it might sound I am very grateful to have had my cat Zolie to help me through the last couple years of graduate school. We spent most of a summer sitting in bed together when I broke my leg and the experience was infinitely better than it would have been without her. The most important people to me are my parents and Stephanie. My parents have done so much for me. They have provided a deep well of emotional support during challenging times and have helped me celebrate in times of joy. They have patiently helped me to slowly transition to adulthood and have always encouraged me to pursue my interests and abilities which is the greatest gift. Stephanie has experienced the full brunt of the highs and lows of my graduate school experience. She is brilliant and beautiful. She has done so many kind things for me. I look forward to a lifetime full of adventures with her.

Contents

List of Figures

List of Tables

Chapter 1

Considerations for discovering chromatin interactions from high-throughput sequencing data

Complex regulatory mechanisms allow for the great diversity of gene expression patterns observed in different cell types. Despite containing DNA sequence for the same set of genes in the genomes of their cells, different cell types within the same organism will express widely different sets of genes [56]. These differences in gene expression are fundamental to allowing different cell types to play different functional roles within the organism. During development, cells express different sets of genes as they move through the stages of differentiation [38]. Fully differentiated cell types in mature organisms turn genes on or off in response to stimuli to allow the organism to maintain homeostasis [2]. Even in the absence of environmental changes, different genes are expressed as a cell progresses through the cell cycle [5]. The variability of gene expression patterns even in cells containing the same genome sequence illustrates one of the core problems of genomics research which is to understand how the expression of genes is regulated.

1.1 Interpreting gene regulation through linear genomics

Until recently, genomics research has taken a mostly linear view of the genome. It is convenient to think of the genome as a one dimensional sequence scattered with regions of importance to gene regulation. The sequence corresponding to genes that contain the information necessary for producing proteins, known as protein-coding genes, makes up only about 3% of the total Human genome sequence [6]. Protein-coding genes have well defined structural components including transcription start sites (TSSs), exons made up of codons that specify amino acids, and introns which are post-transcriptionally spliced out of mRNAs transcribed from protein-coding genes that are specified by splicing signals. The characteristic sequence structure of genes allows for very accurate computational prediction which, along with experimental methods, have led to a thorough inventory of the protein coding genes present in most sequenced genomes [27].

The regions of importance to gene regulation other than the transcribed regions of genes are more difficult to identify from genome sequence alone. The genome sequence that is proximal and just upstream of the TSS of a gene, referred to as the promoter region, plays an important role in regulating the expression of the gene [35]. The binding of a transcription factor within a few kilobases of a TSS will often have an affect on transcription, usually in concert with the binding of other transcription factors. The dominant methods for measuring transcription factor binding involve a chromatin immunoprecipitation (ChIP) step. An antibody that recognizes a protein of interest is used to isolate fragments of genomic DNA that are bound by the protein of interest from a sample of fragmented chromatin extracted from a cell population. Until the advent of high-throughput sequencing technologies, ChIP-enriched fragments were probed for sequences that match a known location in the genome by quantitative PCR or several known locations using microarray technology. Even very dense modern microarray designs are limited in the number of probes that they contain. This limitation prevents high resolution genome-wide profiling of transcription

factor binding and requires that researchers choose the genomic regions that they are most interested in profiling. Given the convenience of looking near annotated TSSs as opposed to the vast majority of the genome which is much less well characterized, much of genomics research until relatively recently focused on characterizing genomic features near annotated TSSs.

The recently developed ability to inexpensively sequence millions of short reads from DNA fragments by high-throughput sequencing has enabled the development of technologies for profiling genomic features genome-wide. By analyzing genome-wide datasets, it was noted that many transcription factors bind very frequently to regions distal to any annotated TSS. This has led to a focus on the identification and characterization of distal regulatory elements such as enhancers and insulators. Experiments have profiled not just the binding of transcription factors by ChIP followed by high-throughput sequencing (ChIP-Seq), but also the association of other proteins with the genome. An important class of proteins that associate with DNA are the histone proteins that make up the protein component of nucleosomes. DNA wraps around histone octamers to form nucleosomes which act as the fundamental unit of chromatin structure [29]. Nucleosome positioning plays a role in gene regulation and can be profiled by ChIP-Seq. Yet, the position of nucleosomes is only part of the role that they play in gene regulation. An interesting property of histone proteins is that they have a "tail" that is not part of the core structure around which DNA is wrapped.

Histone tails are covalently modified in many different ways by nuclear enzymes. It has been observed that certain modifications are correlated with functional activity in the genomic region surrounding the nucleosome containing the modified histone. We will denote histone modifications by the histone type, the residue that is modified, and the type of modification. For example, H3K4me3 refers to trimethylation (me3) of the fourth residue which is a lysine (K4) of histone H3. This particular modification tends to appear near the start sites of actively transcribed genes [66]. Other histone modifications such as H4K4me1 [65] and H3K27ac (acetylation) [11] have been associated with enhancer activity. In some cases, the enzymes that catalyze certain histone

modifications are known. For example, p300 is an acetyltransferase that acetylates H3K27 and p300 binding is frequently used as an indicator of enhancer activity [61].

Other factors that have been profiled by ChIP-Seq to help identify distal regulatory elements include CTCF and components of the Cohesin and Mediator complexes. The role of CTCF in genome function is a field of active inquiry and it may be the case that it has many different functions [51]. The binding of CTCF is thought to have an insulating effect in that the regulatory influence of elements such as enhancers on genes may be blocked by CTCF binding events that exist between the element and the gene. Mediator is known to bind to transcription factors that are bound to enhancers as well as the transcription apparatus which binds to the TSS of genes that are to be transcribed [26]. Cohesin has been shown to bind to DNA at locations that are bound by CTCF or Mediator in a mutually exclusive manner [26]. Cohesin has been shown to act as a stabilizer of chromatin loops, allowing locations that are distal in terms of the genome sequence to be spatially proximal in the nucleus. These observations have led to an understanding that the three dimensional conformations that chromosomes take in the nucleus, including the formation of chromatin loops between distal locations, are a central aspect of genome function.

1.2 Methods for characterizing chromosome conformation

Only very recently have methods been developed for characterizing chromosome conformation in a high throughput fashion. These methods generally incorporate a proximity ligation step inspired by the low throughput method chromosome conformation capture (3C) [12]. Prior to proximity ligation, crosslinked protein and DNA are extracted from cell nuclei and then fragmented by either the application of a restriction enzyme or by sonication. By applying a very low concentration of DNA ligase, the ligation of DNA fragments that are connected by crosslinked proteins is favored over fragments that are not physically connected. This has the effect of favoring the lig-

ation of DNA fragments that were spatially proximal in the nucleus despite the fact that they might not be located proximally in terms of the genome sequence.

In this thesis we develop methods for analyzing data produced by a method known as chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) [18]. This method combines ChIP for isolating fragmented chromatin that contains a protein of interest with proximity ligation. We define two types of ligation events that may occur as part of this process. When one DNA fragment ligates to itself, we call this a self-ligation event. When two DNA fragments ligate to each other, we call this an inter-ligation event. Prior to the proximity ligation step, DNA linkers are ligated to the ends of the DNA fragments that were isolated by ChIP. The proximity ligation of two DNA ends containing the added DNA linkers results in a sequence that is recognized by the restriction enzyme MmeI. This enzyme cuts the DNA 20 bp away from the recognition site formed by the ligated linkers allowing a small DNA fragment containing a portion of the sequences of the genomic DNA ends involved in the proximity ligation to be extracted. These fragments are paired-end sequenced and aligned to a reference genome.

Results obtained from ChIA-PET data provide the opportunity to help fill a large gap in our understanding of genome function. ChIP-Seq data have enabled the identification of genomic locations associated with particular proteins. However, ChIP-Seq data do not directly provide information about the connectivity of distal genomic locations. Recent observations have suggested that enhancers may regulate the expression of genes that are located megabases or more away [1]. It has also been suggested that looping between CTCF binding events demarcate large regulatory domains [51]. ChIA-PET data contain information about chromatin interactions between locations bound by proteins which allows enhancers to be associated with their target genes and the discovery of regulatory domain boundaries. However, like other high-throughput sequencing technologies, ChIA-PET datasets are large and plagued by experimental noise. Analyzing these datasets requires sophisticated computational methods for distilling datasets with tens of millions of datapoints into manageable sets of interpretable results.

At present, both ChIP-Seq and ChIA-PET require the use of chromatin extracted from millions of cells. The results obtained from these experiment types reflect the behavior of a protein averaged over a large population of cells. Because of this, we take the perspective that ChIP-Seq and ChIA-PET results should be interpreted as reflecting the likelihood of a protein behaving in a certain way. For example, ChIP-Seq results reflect the likelihood that a protein occupies a particular location in the genome. Likewise, ChIA-PET results reflect the likelihood that a protein simultaneously occupies two locations in the genome. The joint occupancy of two genomic locations by a protein implies that those locations are involved in a chromatin interaction and that a chromatin loop has formed between them.

The most common approach to analyzing ChIA-PET data is implemented by the ChIA-PET Tool [36]. In this approach, read pairs are classified as having been generated by self-ligation or inter-ligation based on a heuristically determined cutoff on the distance spanned by the read pairs. Self-ligation read pairs are used to determine point locations that are bound by the protein. Inter-ligation read pairs are used to discover chromatin interactions. The locations of binding events computed from the self-ligation read pairs are not used to inform the discovery of interactions. Regions potentially involved in interactions, called anchors, are determined by extending the ends of inter-ligation read pairs by several hundred base pairs and then identifying regions where a significant number of extended read pair ends overlap. These anchors are generally several kilobases in length, much wider than the amount of DNA that would be occupied by any single instance of a protein. The number of inter-ligation read pairs that connect a pair of potential anchors is used to determine the significance of the interaction between the anchors.

In this thesis we present two novel methods for analyzing ChIA-PET data. In both methods we make the assumption that inter-ligation read pairs provide evidence about the simultaneous occupation of two genomic locations by a protein. Based on this assumption, both methods incorporate information from the alignment of self-ligation read pairs about the marginal occupancy of the protein in the interaction discovery process. We will demonstrate that this assumption allows our methods to identify the

pairs of locations that are simultaneously occupied by a protein with a high degree of spatial accuracy. We will also present evidence from examining other forms of high-throughput sequencing data that the high spatial accuracy of the interaction anchors that we identify allow our methods to discover high confidence interactions of functional importance.

1.3 Thesis outline

We have developed two methods for analyzing ChIA-PET data because we recognized that different types of proteins occupy the genome in different ways. Factors such as CTCF tend to occupy consistently narrow genomic regions that are mostly isolated from each other. In Chapter 2 we introduce SPROUT which models interaction anchors as point binding event locations. SPROUT utilizes models of the way that reads align relative to the locations of binding events and incorporates both self-ligation and inter-ligation read pairs when discovering these locations. In Chapter 3 we present GERM which does not make the assumption that the protein being studied binds to isolated point locations. GERM builds high resolution genome-wide models of the occupancy of a protein. This approach is appropriate for factors such as RNA Polymerase II (PolII) which tend to occupy the genome in broad regions of variable width. By relaxing the assumption of point location binding, GERM is able to provide a detailed view of protein occupancy without introducing the types of artifacts that arise from trying to force a point location model to fit PolII data. In Chapter 4 we apply GERM to embryonic stem cell and motor neuron progenitor data and make several observations about the usage of enhancers during motor neuron development.

Chapter 2

Probabilistic modeling of binding events and chromatin interactions between them

ChIA-PET data consist of read pairs that were generated by two different types of ligation events. Self-ligation read pairs are the result of a DNA fragment circularizing to ligate to itself. Inter-ligation read pairs are the result of two distinct DNA fragments ligating to each other. However, the reads that make up self-ligation and inter-ligation read pairs all correspond to the ends of the DNA fragment(s) involved in ligation events. The DNA fragments subjected to the proximity ligation step of the ChIA-PET procedure are enriched for fragments that are bound by the protein of interest by the preceding ChIP step of the procedure. We assume then that the reads that make up both types of read pairs align to positions in the genome that are arranged stochastically around binding events. The fragmentation step that precedes ChIP induces a distribution over genomic locations that describes where reads are likely to align.

Accurately modeling the positions of binding events is important for extracting high quality results from ChIA-PET data. Assuming proteins bind to fixed, punctate locations in the genome, we will demonstrate that we can combine information from both self-ligation and inter-ligation read pairs to estimate the positions of binding

events accurately. By modeling the distribution of read alignment relative to binding events from both types of read pairs, we will demonstrate that we are also able to accurately assign inter-ligation read pairs to pairs of binding events. This reduces the false positive rate of interaction discovery by reducing the degree to which inter-ligation read pairs are assigned to the same interaction when in fact they do not correspond to the same pair of binding events. Likewise, we are able to distinguish between nearby binding events to discover interactions at greater spatial resolution.

2.1 Prior work

The development of methods for analyzing ChIA-PET data has been quite limited. Most published ChIA-PET analyses have been performed using the ChIA-PET Tool software. One paper analyzing H3K4me2 ChIA-PET data [7] applied a method called Density-Based Spatial Clustering of Applications with Noise. The implementation of this method was not made available. On the other hand, many methods for analyzing ChIP-Seq data have been developed. The analysis of ChIP-Seq data shares many concerns with the analysis of ChIA-PET data. A variety of approaches have been taken for identifying binding event locations from ChIP-Seq read alignments including extending the length of aligned reads [53], shifting aligned reads [67], and estimating the distributions of positive and negative strand reads separately [23, 25, 60]. The ChIA-PET tool has a component for identifying binding events from self-ligation read pairs which takes advantage of the fact that the ends of self-ligation read pairs correspond to the ends of an individual DNA fragment that was bound by the protein, therefore removing the need for extension or shifting of reads. All of these approaches identify locations where the overlap of reads or the estimated read density is greatest as locations that are bound by the protein.

The approach taken by SPROUT for identifying binding events which it considers as interaction anchors is most similar to the approach taken by the ChIP-Seq algorithm GPS [21]. SPROUT and GPS both utilize models of the expected distribution of reads relative to a location bound by the protein. Furthermore, they are both generative

models which estimate binding event locations which maximize the likelihood of the observed data. SPROUT extends GPS in several ways. Since ChIA-PET data consist of paired reads, SPROUT utilizes models of the distribution of pairs of reads relative to binding events rather than individual reads. SPROUT also estimates the type of each read pair and uses models of read pair distribution appropriate to the estimated read pair type. A difference in modeling decision between SPROUT and GPS is that GPS initially considers every position in the genome as a potential binding event and then iteratively removes positions from consideration to result in a sparse set of positions that explain the data. SPROUT is initialized with a number of binding events that is much smaller than the size of the genome. These binding events can be removed from consideration in a similar manner to GPS. However, the positions of the binding events in SPROUT are variable so that they can be repositioned. Through a combination of repositioning and removal, SPROUT discovers a sparse set of binding events that explain the data. This approach to locating binding events is more similar to MultiGPS [41] than the original GPS formulation.

2.2 Modeling ChIA-PET read pairs with a hierarchical generative model

In this section we formulate a hierarchical generative model for accurately modeling ChIA-PET read pairs that we call SPROUT. SPROUT is a hierarchical generative model for ChIA-PET data that discovers interaction anchors, and a set of binary interactions between anchors. SPROUT models read-pair data with a mixture over distributions describing the generation of self-ligation pairs and inter-ligation pairs. The components of the model describing these two types of read pairs are themselves mixtures of distributions corresponding to the way pairs of reads are expected to be distributed around anchors. We assume that the paired-end sequence data generated by a ChIA-PET experiment have been processed appropriately resulting in a set $\mathbf{R} = \{r_1, \ldots, r_N\}$ such that each $r_i = \langle r_i^{(1)}, r_i^{(2)} \rangle$ is a pair of genomic coordinates

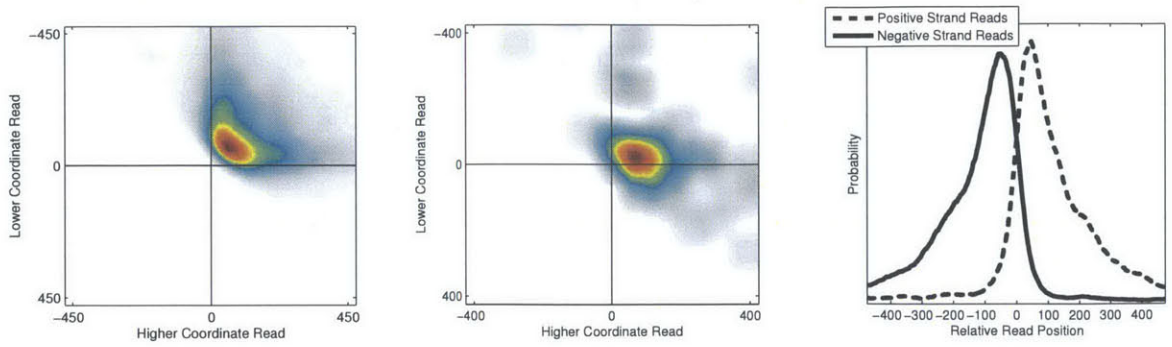
corresponding to the aligned positions of a pair of reads. Such processing includes removing linker tags from the reads, filtering out pairs that are identified as chimeric because of their heterogeneous linker tags, and aligning the reads to the genome. The following is the likelihood of $\mathbf{R}$

$$\Pr(\mathbf{R}, \pi, \psi, \rho, l) = \prod_{i=1}^{N} \left[\rho \left[\sum_{j=1}^{M} \pi_j \Pr(r_i|l_j) \right] + (1-\rho) \left[\sum_{j=1}^{M} \sum_{k=1}^{M} \psi_{j,k} \Pr(r_i|l_j, l_k) \right] \right] \tag{2.1}$$

Where $0 \leq \rho \leq 1$, $\sum_{i=1}^{N} \pi_i = 1$, $\sum_{i=1}^{N} \sum_{j=1}^{N} \psi_{i,j} = 1$

SPROUT identifies a set $l = \{l_1, \ldots, l_M\}$ that specifies the locations of sites that are bound by the protein of interest and are potential anchors for interactions. ρ is the probability that a pair of reads was generated by self-ligation. Self-ligation pairs reflect the ligation of a DNA fragment to itself to form a circular fragment. Such pairs are associated with one anchor and the self-ligation component of the model is a mixture of distributions each taking a single parameter to specify the location of the anchor position. These distributions take the form $\Pr(r_i|l_j)$ as shown in Figure 2-1a. A relative weight π_j is associated with each anchor j. These distributions describe the length and arrangement of fragments around an anchor which are induced by the fragmentation step of the ChIA-PET protocol.

Inter-ligation pairs can be associated with either the same anchor or two different anchors that were in close proximity in the nucleus. The inter-ligation component of the model is a mixture of distributions each taking two parameters that specify the locations of the anchor(s) that the fragments were associated with. A relative weight $\psi_{j,k}$ is associated with each pair of anchors j and k. The distributions $\Pr(r_i|l_j, l_k)$ take different forms because if $j = k$, such as in Figure 2-1b, then there are constraints on the ends of the fragments involved in the ligation. For example, the fragments cannot have been overlapping since they were part of the same chromosome prior to fragmentation. If $j \neq k$, such as in Figure 2-1c, it is assumed that the ends were generated independently by two one-dimensional distributions centered around the two anchors $\Pr(r_i|l_j, l_k) = \Pr(r_i^{(1)}|l_j) \Pr(r_i^{(2)}|l_k)$. We also assume that r_i implicitly



(a) (b) (c)

Figure 2-1: **Examples of read distributions learned from CTCF ChIA-PET data.** SPROUT is initially run with "generic" distributions and then the distributions are re-estimated using the strongest events and SPROUT is re-run with the empirically learned distributions to discover more accurate predictions. (a) The positions of the ends of self-ligation pairs are modeled using a two dimensional distribution. (b) The positions of the ends of inter-ligation pairs where both ends are assigned to the same anchor are also modeled using two dimensional distributions. Each of the four possible strand combinations has its own constraints in terms of where the ends are likely to be positioned relative to each other and to the anchor. This figure demonstrates the distribution associated with inter-ligation pairs where both ends map to the positive strand. (c) The positions of the ends of inter-ligation pairs are modeled separately using one dimensional distributions.

carries information about the strandedness of the reads because in both the case where $j = k$ and $j \neq k$ the distributions depend on strandedness.

ChIA-PET data are noisy, and we observe reads that do not correspond to anchors. To account for these reads, we introduce a noise component with dummy variable l_B ($B \notin \{1, \ldots, M\}$). In this work we consider uniform $\Pr(r_i|l_B)$, however knowledge about the propensity for genomic regions to generate background noise could be incorporated into a more refined noise distribution. We assume that $\Pr(r_i|l_j, l_k)$ where $j = B$ or $k = B$ is defined in the same way as the case in which j and k specify two different anchors: $\Pr(r_i|l_j, l_k) = \Pr(r_i^{(1)}|l_j)\Pr(r_i^{(2)}|l_k)$ and $\Pr(r_i^{(\cdot)}|l_j)$ is uniform when $j = B$.

To avoid overfitting, we wish to find a minimal number of anchors that explain the data well while allowing the noise distribution to account for reads that are not accounted for by anchors. Additionally, we assume that among all possible pairs of

anchors most pairs are not interacting. Thus, we wish to find a minimal number of interacting pairs of anchors that explain the observed data. To achieve both of these types of sparsity we introduce negative Dirichlet priors [16] on π and ψ as specified by Equations 2.2 and 2.3.

$$\Pr(\pi|\alpha) \propto \prod_{j=1}^{M} \pi_j^{-\alpha} \tag{2.2}$$

$$\Pr(\psi|\beta) \propto \prod_{j=1}^{M} \prod_{k=1}^{M} \psi_{j,k}^{-\beta} \tag{2.3}$$

As will become apparent when the inference procedure is described, the α and β parameters have the effect of specifying the minimum number of pairs of reads that must be associated with an anchor or an interaction, respectively, in order to avoid being eliminated from the model.

We also introduce priors on l and ρ. For l we introduce a Bernoulli prior which reflects our prior belief that an anchor exists at a particular genomic coordinate and that at most one anchor exists at any genomic coordinate. Given L possible genomic coordinates,

$$\Pr(l|k) = \prod_{i=1}^{L} k_i^{\mathbf{1}(i \in l)} (1 - k_i)^{\mathbf{1}(i \notin l)} \tag{2.4}$$

$$= \prod_{i=1}^{L} (1 - k_i) \prod_{j=1}^{M} \frac{k_{l_j}}{1 - k_{l_j}} \tag{2.5}$$

$$\propto \prod_{j=1}^{M} \frac{k_{l_j}}{1 - k_{l_j}} \tag{2.6}$$

In this work we consider uniform k, but k could be made non-uniform to reflect any prior belief about where anchors should be located. For ρ we introduce a Beta prior

$$\Pr(\rho|a,b) \propto \rho^{a-1}(1-\rho)^{b-1} \tag{2.7}$$

In this work we let $a = 1$ and $b = 1$ which is a uniform prior on ρ.

Each pair of reads is either a result of a self-ligation event or an inter-ligation event and is associated with one or two anchors. We introduce latent variables $\mathbf{Z} = \{z_1, \ldots, z_N\}$ such that each $z_i = \langle z_i^{(1)}, z_i^{(2)} \rangle$ is a pair of anchor indices $1 \ldots M$ or special index B reflecting the noise distribution. Another special index is used to indicate that a pair of reads was generated by self-ligation i.e. $z_i = \langle j, - \rangle$.

The complete data likelihood is

$$\Pr(\mathbf{R}, \mathbf{Z}|\pi, \psi, \rho, l) = \Pr(\mathbf{R}|\mathbf{Z}, l) \Pr(\mathbf{Z}|\pi, \psi, \rho) \tag{2.8}$$

$$= \prod_{i=1}^{N} \left[\prod_{j=1}^{M} [\rho \pi_j \Pr(r_i|l_j)]^{\mathbf{1}(z_i = \langle j, - \rangle)} \prod_{k=1}^{M} [(1-\rho)\psi_{j,k} \Pr(r_i|l_j, l_k)]^{\mathbf{1}(z_i = \langle j,k \rangle)} \right] \tag{2.9}$$

We are interested in inferring likely values for π, ψ, ρ, and l. To accomplish this we employ a variant of the EM algorithm [13] to maximize the complete data log posterior

$$\begin{aligned} \log \Pr(l, \pi, \psi, \rho|\mathbf{R}, \mathbf{Z}, k, \alpha, \beta, a, b) = \sum_{i=1}^{N} \Bigg[\sum_{j=1}^{M} \Bigg[& \mathbf{1}(z_i = \langle j, - \rangle) (\log \rho + \log \pi_j + \log \Pr(r_i|l_j)) \\ & + \sum_{k=1}^{M} \mathbf{1}(z_i = \langle j, k \rangle) (\log(1-\rho) + \log \psi_{j,k} + \log \Pr(r_i|l_j, l_k)) \Bigg] \Bigg] \\ -\alpha \sum_{j=1}^{M} \log \pi_j - \beta \sum_{j=1}^{M} \sum_{k=1}^{M} \log \psi_{j,k} + \sum_{j=1}^{M} \log \frac{k_{l_j}}{1 - k_{l_j}} & + (a-1) \log \rho + (b-1)(1-\rho) + C \end{aligned} \tag{2.10}$$

E Step:

$$\gamma(z_i) = \frac{\prod_{j=1}^{M}\left[[\rho\pi_j \Pr(r_i|l_j)]^{\mathbf{1}(z_i=\langle j,-\rangle)} \prod_{k=1}^{M}[(1-\rho)\psi_{j,k}\Pr(r_i|l_j,l_k)]^{\mathbf{1}(z_i=\langle j,k\rangle)}\right]}{\sum_{j=1}^{M}\left[[\rho\pi_j \Pr(r_i|l_j)] + \sum_{k=1}^{M}[(1-\rho)\psi_{j,k}\Pr(r_i|l_j,l_k)]\right]} \quad (2.11)$$

M Step:

$$\hat{l}_j = \underset{x}{\text{argmax}}\{\sum_{i=1}^{N}\left[\gamma(z_i=\langle j,-\rangle)\log\Pr(r_i|x) + \sum_{k=1}^{M}[\gamma(z_i=\langle j,k\rangle)\log\Pr(r_i|x,l_k)]\right] + \log\frac{k_x}{1-k_x}\} \quad (2.12)$$

$$\hat{\pi}_j = \frac{\max(N_j-\alpha,0)}{N_\pi} \quad (2.13)$$

$$N_\pi = \sum_{j=1}^{M}\max(N_j-\alpha,0) \quad (2.14)$$

$$N_j = \sum_{i=1}^{N}\gamma(z_i=\langle j,-\rangle) \quad (2.15)$$

$$\hat{\psi}_{j,k} = \frac{\max(N_{j,k}-\beta,0)}{N_\psi} \quad (2.16)$$

$$N_\psi = \sum_{j=1}^{M}\sum_{k=1}^{M}\max(N_{j,k}-\beta,0) \quad (2.17)$$

$$N_{j,k} = \sum_{i=1}^{N}\gamma(z_i=\langle j,k\rangle) \quad (2.18)$$

$$\hat{\rho} = \frac{N_\pi + a}{N+a+b} \quad (2.19)$$

The E and M steps are repeated until the posterior approximately converges. The components of l that correspond to non-zero components of π are the estimated anchor locations. Non-zero components of ψ indicate pairs of anchors that are candidates for significance testing as interactions.

The algorithm is initialized with uniform π and l set at regular intervals throughout the genome. Components of π that do not assign probability to any pairs of reads

are set to 0 and effectively eliminated from the model. Components with $N_j < \alpha$ are eliminated shortly thereafter. In the estimation of $\hat{l}_j$ during each M step the components of l other than the jth component are held fixed making this algorithm an instance of the expectation-conditional maximization algorithm [45]. Thus, the posterior is not necessarily maximized at each iteration but convergence to a local maximum is still guaranteed. The estimation of $\hat{l}_j$ is tractable, despite the lack of a closed form solution, because for the set of pairs of reads such that $\gamma(z_i = \langle j, \cdot \rangle) > 0$, $Pr(r_i|x) > 0$ for any pair of reads in the set for x in only a small neighborhood around the previous value of l_i. Only x in that neighborhood need be considered which reduces the search space for the optimal x considerably.

To test the significance of a component $\psi_{j,k}$, the posterior is recomputed with that component removed. The greater the ratio of the posterior with the component to the posterior without the component, the greater the significance of the corresponding interaction. Making the conservative assumption that all components with $N_{j,k} \leq 2$ are false positives, we set a threshold for the posterior ratio to be the value such that 5% of the components deemed significant have $N_{j,k} \leq 2$.

2.3 Evaluating SPROUT

We appied SPROUT to a CTCF ChIA-PET dataset in mouse embryonic stem (ES) cells published by Handoko et al. [22]. We chose to analyze these data for several reasons.

1. CTCF complies with our assumptions about protein binding in that it binds in a punctate fashion

2. CTCF recognizes a well characterized sequence motif which is strongly predictive of binding

3. CTCF has been suggested to play a significant role in the structural organization of the genome

For these reasons, we expected that SPROUT would perform well on these data and that the results would be useful towards understanding genome structure. We processed the paired sequence data by filtering out chimeric ligation read pairs that contain two different linker sequences, aligning the read pairs using BOWTIE [34], and removing paired positional duplicates to avoid spurious results from PCR artifacts. We applied SPROUT to obtain a set of CTCF binding events and a set of pairs of binding events that are determined to be significantly interacting. SPROUT does not place any constraints on the distance between binding events that may interact. It has generally been observed when measuring chromatin interactions using any of the standard approaches that locations that are closer along the linear sequence of a chromosome are more likely to interact. It is often assumed that this reflects the monotonic relationship between the distance along a polymer between two monomers and the distance in space between the monomers [58]. Assuming chromosomes undergo some amount of random movement, one would expect linearly proximal locations to randomly interact at some rate. We were curious to see whether this effect manifested itself in the results from SPROUT. In Figure 2-2 we first plotted the frequency at which two CTCF binding events exist in the genome at distances ranging from 0 bp to 20 kb. We observed that it is relatively common for CTCF binding events to be located between 2 kb and 4 kb away from another CTCF binding event. At distances greater than 4 kb we observed a constant frequency of CTCF binding events separated by up to 20 kb. We then plotted the frequency at which two interacting CTCF binding events were detected at distances of linear separation up to 20 kb. We observed that a majority of pairs of CTCF binding events with linear separation less than 4 kb are detected as interacting. At distances of linear separation greater than 4 kb, very few pairs of CTCF binding events interact relative to the frequency at which pairs of binding events exist separated by distances greater than 4 kb. This is not to say that pairs of binding events that are linearly separated by more than 4 kb are never detected as interacting. Rather, there seems to be a general tendency for linearly proximal CTCF binding events to interact while more distal pairs of binding events only interact in specific cases. This suggests that when pairs of CTCF binding

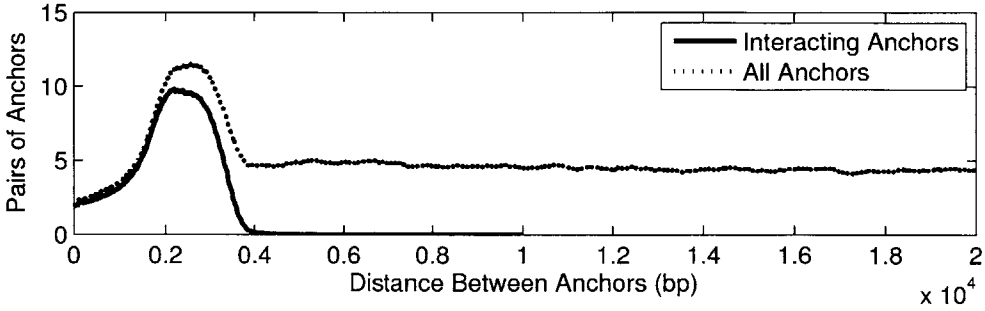
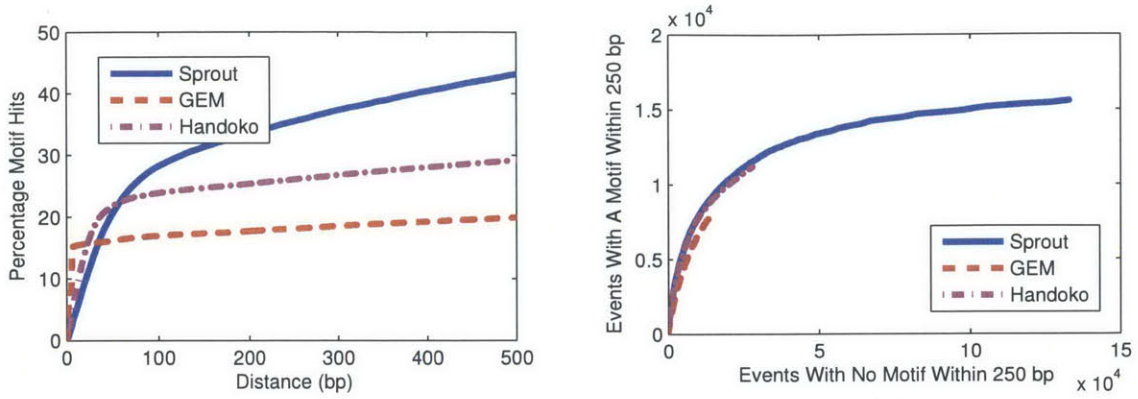


Figure 2-2: **Smoothed plots of the frequency at which binding events are identified by SPROUT as interacting at linear separations up to 20 kb.** Most pairs of binding events that are separated by at most 4 kb are detected as interacting. At linear separations greater than 4 kb relatively few pairs of binding events are detected as interacting.

events are detected as interacting and are separated by a linear distance of more than 4 kb that these interactions were induced by some active mechanism and not random movement of the chromosome.

One of the strengths of SPROUT is the positional accuracy of the binding events that it identifies. Handoko et al. published a set of binding events that they estimated from the read pairs that they determined to have been generated by self-ligation. We also ran a state of the art ChIP-Seq algorithm [20] on an independent ChIP-Seq dataset as a positional "gold standard" for comparison. We scanned the genome for matches to the CTCF motif and computed the percentage of the motif matches that we found that were within distances up to 500 bp from binding events from the three sets as shown in Figure 2-3. The GEM results from ChIP-Seq data identify binding events at about 15% of the motif matches with very high spatial accuracy. If we allow binding events located 100 bp or more to be associated with motif matches, SPROUT is able to identify more binding events near motif matches than the other two methods. We also examined the utility of the measures of significance associated with the binding events in the three sets for identifying binding events associated with motif matches. Assuming that binding events and motif matches may be associated if they are within 250 bp of each other, Figure 2-3b shows that SPROUT identifies more binding events overall that are within 250 bp of a motif match than the other sets of binding events. The measure of significance that SPROUT associates with binding



(a) (b)

Figure 2-3: **Evaluation of the accuracy of CTCF binding events predicted by SPROUT and Handoko et al. from the ChIA-PET data as well as by GEM from an independent ChIP-Seq dataset.** (a) The percentage of CTCF motif matches in the genome that have a binding event identified within distances up to 500 bp. (b) We used the presence of a CTCF motif match within 250 bp of an event as an approximate indicator of true positive anchor calls. As thresholds for significance are varied for each method, the number of true positive and false positive calls are plotted.

events also consistently avoids more binding events that do not have a motif match within 250 bp given a fixed number of binding events with a motif within 250 bp than the other sets.

The chromatin interactions published by Handoko et al. are not pairs of interaction between CTCF binding events, but rather are between anchors whose locations are not directly informed by the locations of CTCF binding events. The Handoko et al. interaction anchors are determined only from the aligned locations of read pairs that they determine to have been generated by inter-ligation. The self-ligation read pairs were not considered and thus do not help refine the locations of interaction anchors as is the case with SPROUT. As shown in Figure 2-4 the interaction anchors published by Handoko et al. are quite broad compared to the binding events considered by SPROUT which are assigned to point locations in the genome. The Handoko et al. anchors are most frequently around 2 kb in width which is about the same as the most frequently observed distance between CTCF binding events as shown in Figure 2-2. This illustrates an advantage of the results from SPROUT in that inter-

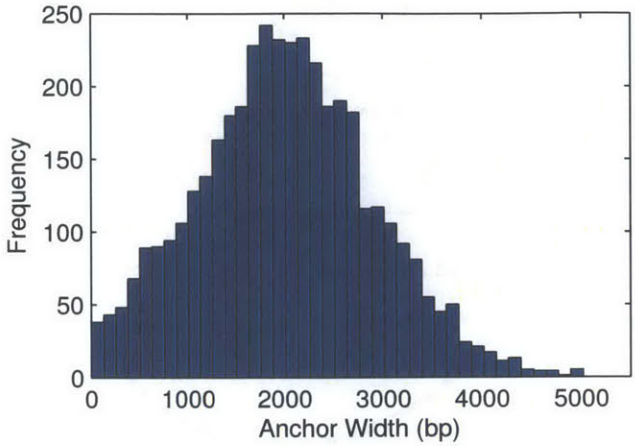


Figure 2-4: **A histogram of the widths of anchors identified by Handoko et al. illustrating the breadth of many of the anchor regions.**

actions are explicitly between pairs of binding events whereas it is more difficult to unambiguously assign the interactions published by Handoko et al. to specific pairs of binding events.

We compared the interactions identified by SPROUT to the interactions published by Handoko et al. and initially found that for many of the Handoko et al. interactions there were no SPROUT identified interactions between pairs of binding events that are within 4 kb of anchors involved in the Handoko et al. interactions (Figure 2-5). We compared the binding event locations identified by Handoko et al. to the anchors of the interactions that they published and discovered that one or both of the anchors for more than half of the interactions for which there are no matching SPROUT identified interactions do not contain binding events. This observation illustrates an assumption that is made implicitly by Handoko et al. when not using binding events to help inform the locations of interaction anchors. This assumption is that chromatin interactions detected from CTCF ChIA-PET data need not be between CTCF binding events. SPROUT does not make this assumption because we assume that both inter-ligation and self-ligation read pairs should reflect the ends of DNA fragments that are bound by CTCF. This large fraction of the interactions published by Handoko et al. are therefore not discoverable by SPROUT. Of the other interactions published by Handoko et al. that do not match SPROUT identified interactions,

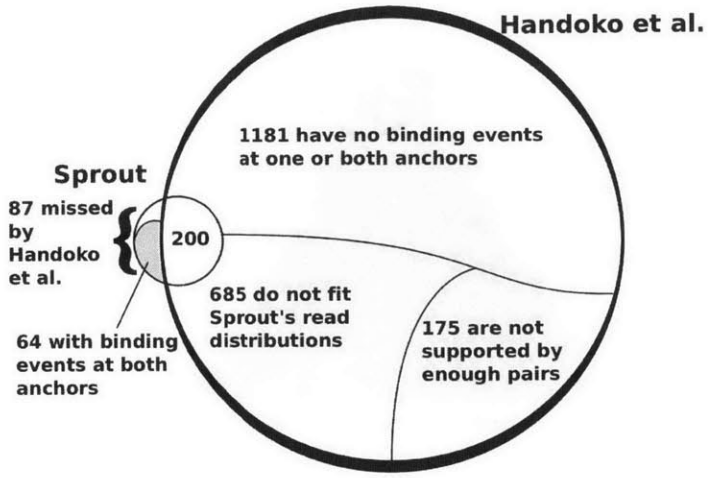
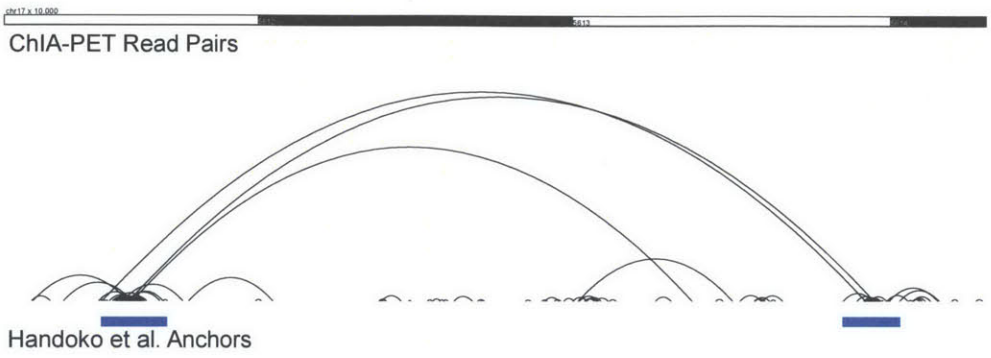


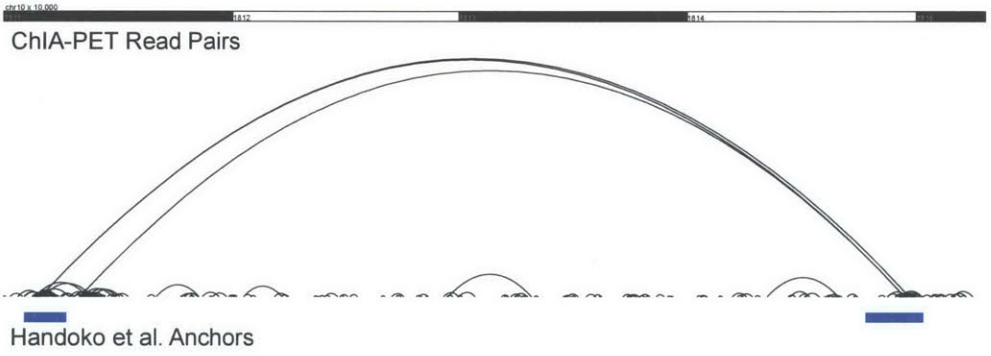
Figure 2-5: **Most of the interactions identified by Handoko et al. are not supported by pairs of reads with ends that fit SPROUT's read distribution.**

more than three quarters of them are based on read pair alignments that do not fit SPROUT's read distributions. An example of such an interaction is shown in Figure 2-6a. The remainder of the interactions published by Handoko et al. that do not match SPROUT identified interactions are supported by fewer than two read pairs. These interactions may reflect weak interactions that are not detected by SPROUT or they may reflect differences in the read alignments used as input for the two methods. The interactions published by Handoko et al. did not contain matches to almost a third of the interactions identified by SPROUT. Most of the unmatched SPROUT identified interactions contained binding events published by Handoko et al. at both anchors and it is unclear why they were not identified by Handoko et al. An example of an interaction that is identified by both methods is shown in Figure 2-6b. The region shown in Figure 2-6b also contains a second distinct SPROUT identified interaction that is not identified by Handoko et al. Figure 2-6b illustrates the ability of SPROUT to provide a detailed view of interactions between distinct binding events.

One of the benefits of the read distributions modeled by SPROUT is that the interactions identified by SPROUT tend to be more uniformly supported by biological replicates. We considered two sets of interactions. One set which we call the good fit set consists of the 200 interactions are are identified by Handoko et al. and SPROUT. The other set which we call the bad fit set consists of the 685 interactions identified



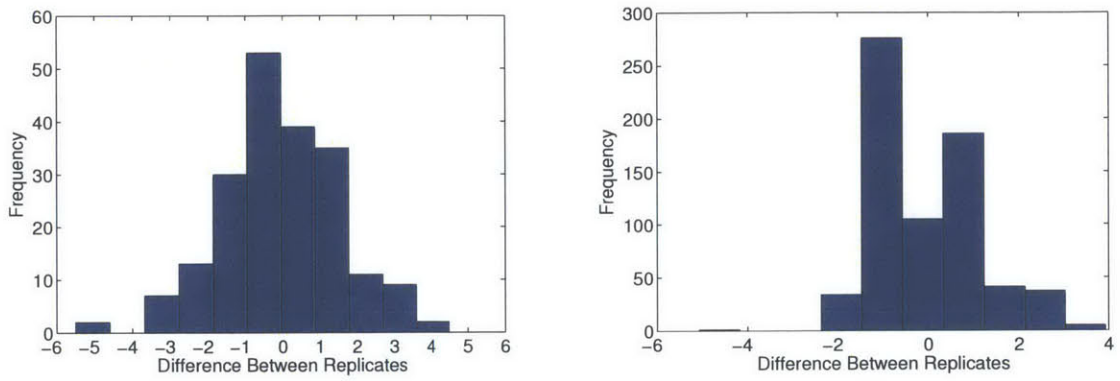
(a)



(b)

Figure 2-6: **Two interactions that are identified by Handoko et al.** The boxes indicate the anchor regions that they identify. (a) This interaction is not called significant by SPROUT because the pairs of reads that connect the anchor regions do not fit SPROUT's model. (b) SPROUT does call a significant interaction between the anchors that fall within the Handoko et al. anchor regions because the pairs of reads that connect the regions were likely to have been generated by the anchors within the regions according to SPROUT's model. Note that there is a second potential anchor on the left side that falls outside of the Handoko et al. identified region. This binding events is identified by both SPROUT and Handoko et al. and is identified by SPROUT but not by Handoko et al. as an independent interaction with the anchor on the right.

by Handoko et al. that are not matched by interactions identified by SPROUT but do contain at least one binding event within both anchors. We chose to define the sets of interactions in this manner to highlight the beneficial effect of explicitly modeling the distribution of read alignments around binding events rather than just counting the number of read pairs whose ends align within two broad regions. We found that the interactions in the good fit set are supported by an average of 4.15 read pairs while the interactions in the bad fit set are supported by an average of only 2.73 read pairs. We then computed the difference for each interaction in both sets between the numbers of read pairs that support the interaction from the two biological replicates that make up the full dataset published by Handoko et al. Figure 2-7a shows that the distribution of differences for the good set is monomodal and centered near zero suggesting that the replicates tend to support the interactions in this set more equally. Figure 2-7b shows that the distribution of differences for the bad set is bimodal and that the support for the interactions in this set from the two replicates tends to be unequal. The differences between these two distributions suggests that the Handoko et al. interactions that are not matched by SPROUT identified interactions are weaker and less replicable implying that they are more likely to be false positives than the interactions identified by both methods.



(a) (b)

Figure 2-7: **Evaluation of biological replicate consistency in interactions discovered by both methods and in interactions identified by Handoko et al. that do not fit** SPROUT**'s read distributions.** (a) A histogram of the difference in the number of pairs of reads from each biological replicate that connect anchors identified by Handoko et al. that subsume interactions called by SPROUT. To account for the overall difference in signal strength, the values were subtracted by the mean per interaction difference. There are interactions that differ in support between the biological replicates. However, the normalized difference in pairs between the biological replicates is most frequently close to 0. (b) A histogram of the difference in the number of pairs of reads from each biological replicate that connect anchors identified by Handoko et al. that are supported by a plausible number of read pairs but do not fit SPROUT's read distributions. As in (a), the differences are subtracted by the mean difference. The biological replicates differ much more frequently than they agree.

Chapter 3

Modeling the joint occupancy of genomic locations by proteins

The assumption that protein binding can be accurately modeled as isolated point binding locations does not hold in some cases. A notable example is the manner in which RNA Polymerase II (PolII) associates with the genome. The enrichment of PolII spreads over much larger domains than are typically observed for transcription factors like CTCF. The distribution of PolII ChIA-PET reads in a region of mouse chromosome 5 is shown in Figure 3-1. When applied to these data, SPROUT attempts to position a number of binding events throughout this region to explain the read alignments. These locations fail to accurately reflect the pattern of enrichment that we observe. SPROUT positions some binding events at locations where the level of enrichment is relatively low in order to help explain the breadth of the domain of enrichment. In general, such domains of enrichment are highly variable in terms of their width and the pattern of enrichment within the domains. SPROUT gains statistical power when analyzing punctate binding data by maintaining a less complex model of protein binding. However, accurately modeling PolII enrichment requires that we alter the assumptions that we made with SPROUT about how proteins associate with DNA.

We relax the assumptions that we made with SPROUT such that we no longer assume that there exists some number of discrete binding events. Rather, we de-

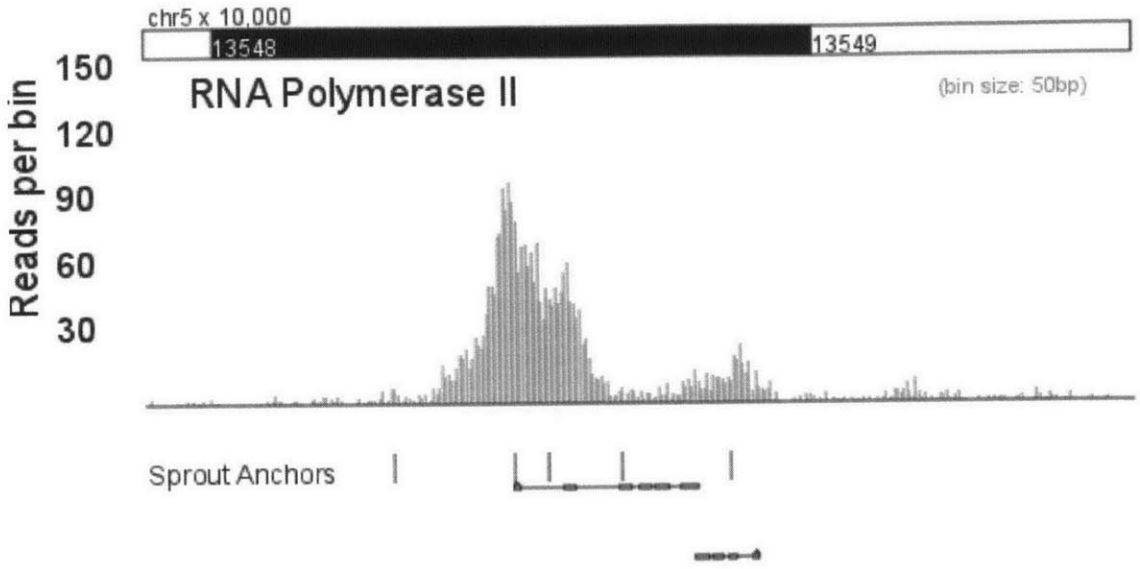


Figure 3-1: **The distribution of PolII ChIA-PET reads is not well modeled by point binding locations**

velop a new algorithm GERM with which we model a full genome-wide distribution of protein occupancy. With SPROUT we assumed that the probability that a protein occupies the vast majority of locations in the genome is zero. With GERM we assume that every location in the genome will have some probability of occupation by the protein. This approach is much more flexible and allows the highly variable domains of PolII enrichment to be modeled more accurately. Aspects of GERM were inspired by methods developed in the image processing literature. To further illustrate the difference in the assumptions made with SPROUT and with GERM we consider two images that one might wish to model in Figure 3-2. Figures 3-2a and 3-2c represent images that we would like to recover from the blurred images in Figures 3-2b and 3-2d respectively. We can safely assume that the image that we would like to recover from Figure 3-2b can be accurately approximated by a number of point locations that is much smaller than the number of pixels in the image. As such, the problem of recovering Figure 3-2a from Figure 3-2b becomes the relatively simple task of estimating the number and locations of the points. Figure 3-2c is much more complicated and contains much more detail than Figure 3-2a. We cannot make the simplifying assumption that Figure 3-2c can be accurately modeled by a number of point locations

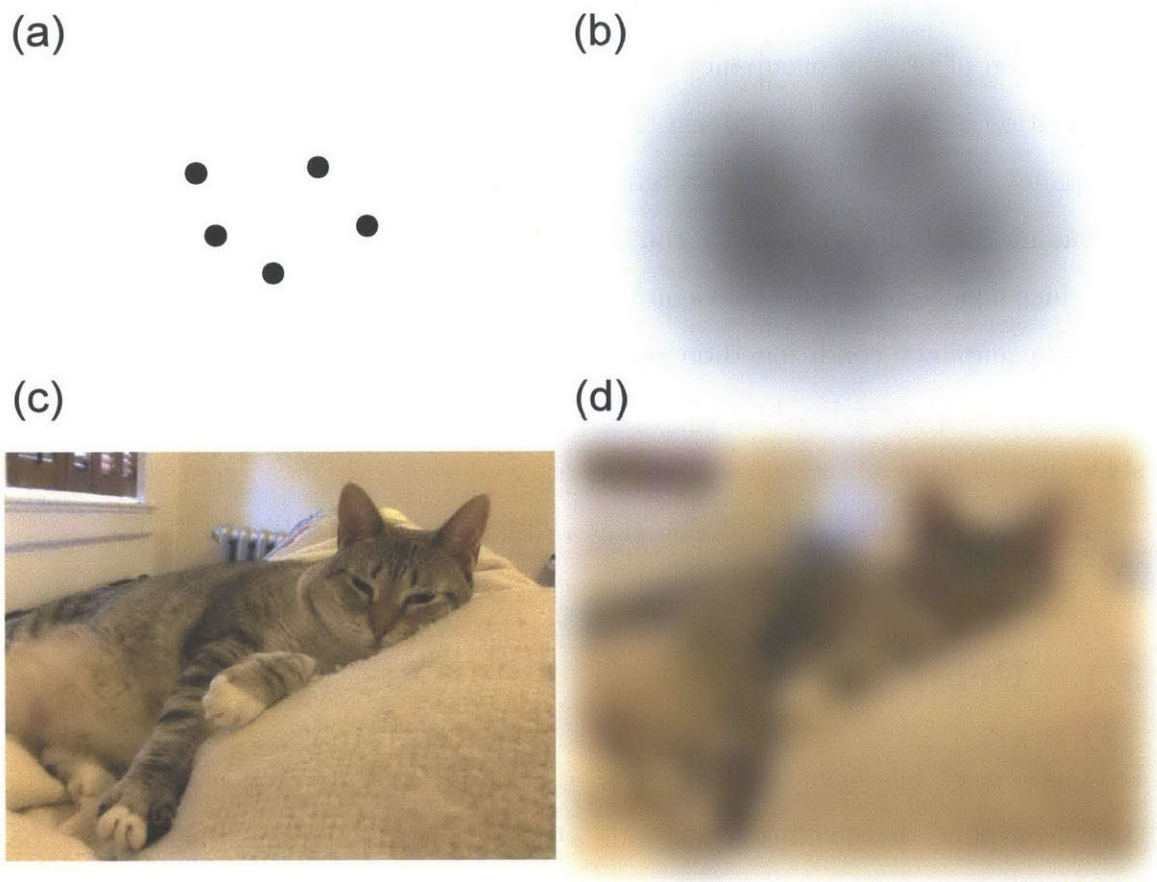


Figure 3-2: **Examples of image reconstruction problems that require different modeling assumptions**

that is much smaller than the number of pixels in the image. We must therefore maintain a more complex model in which we assume that the intensity of every pixel in Figure 3-2c may contribute to the blurred image that we observe in Figure 3-2d.

In the remainder of this chapter we describe GERM which is a novel method for analyzing ChIA-PET data that presents a detailed view of the occupancy of the genome by a protein of interest. An overview of the GERM workflow is shown in Figure 3-3. GERM models the distribution of self-ligation read pairs as a convolution of a model of the chromatin fragmentation process with the marginal distribution of protein occupancy. We apply an adapted blind deconvolution algorithm to simultaneously recover the model of the fragmentation process as well as an estimate of the marginal distribution of protein occupancy for a portion of the genome. The structure

of the estimated fragmentation model allows an estimate of the genome-wide marginal distribution of protein occupancy to be obtained efficiently. The estimated marginal distribution is then used to inform the estimation of the joint distribution of protein occupancy. The joint distribution reflects a detailed view of the likelihood that pairs of genomic locations are simultaneously occupied by a protein of interest. Finally, we introduce a variation on GERM denoted GERM^X in which we compute distributions of protein occupancy conditioned on genomic locations in a set X. A practical example of X when analyzing PolII ChIA-PET data is the set of all annotated transcription start sites (TSSs).

Table 3.1 describes the notation that will be used in this chapter.

3.1 Prior work

Extensive work has been done on deconvolution in the context of image reconstruction. Non-blind deconvolution methods assume that the function that characterizes the blurring effect of the imaging system is known. This function, known as the point spread function (PSF), is often very difficult or impossible to estimate *a priori*. In some systems, this function may even change in unpredictable ways with every image that is captured. To deal with this issue, methods have been developed which do not require knowledge of the PSF *a priori*. These methods are known as blind deconvolution methods. There are several general classes of blind deconvolution algorithms. The distribution of protein occupancy that GERM estimates is nonnegative and GERM does not assume a parametric form for the blurring effect of fragmentation. This places the blind deconvolution component of GERM among the general class of blind deconvolution methods known as nonparametric deterministic image constraints restoration techniques. Examples of methods in this class include iterative blind deconvolution (IBD) [4] which utilizes the fast-Fourier transform and is fast and robust to noise but generally unstable. Another example is the simulated annealing (SA) [44] approach which is more reliable but converges very slowly. The nonnegativity and support constraints recursive inverse filtering (NAS-RIF) [32] ap-

Table 3.1: GERM **notation**

Term	Definition
$r_i = \langle r_i^{(1)}, r_i^{(2)} \rangle$	The aligned locations of the ith read pair
R	The set of all aligned read pair locations
R_{self}, R_{inter}	The sets of aligned self-ligation or inter-ligation read pairs
z_i	The indicator of whether the ith read pair was produced by self-ligation or inter-ligation
$d(r_i)$	The distance between the aligned locations of the ith read pair
N	The total number of aligned read pairs
$N_{++}, N_{+-}, N_{-+}, N_{--}$	The number of aligned read pairs with a particular strand orientation
N_{self}, N_{inter}	The number of aligned self-ligation or inter-ligation read pairs
K_1, K_2	The standard univariate or bivariate Gaussian kernel
$h_{-+}, h_{non-+}, h_{self}$	The bandwidth parameters for kernel density estimates
$ISE(\hat{f})$	The integrated square error of $\hat{f}$ relative to f
q_i	The location occupied by the protein associated with the ith read pair
$RSF(\langle x-u, y-u \rangle)$	The read spread function describing the probability of observing a self-ligation read pair $r = \langle x, y \rangle$ given $q = u$
$\langle -\lambda, \lambda \rangle$	The peak of the estimated RSF
reg	A genomic region
w	The size (in base pairs) of reg
p	The probability of protein occupancy in reg
Z	A random variable representing the number of read pairs associated with reg according to the estimated distribution of occupancy
Y	A random variable representing the number of read pairs associated with reg according to the null model
M	The size of the mappable genome
t_i	$= \sum_u \hat{\Pr}(q = \langle u, v_i \rangle \mid R_{inter})$
m_i	$= \hat{\Pr}(q = v_i)$
τ_i	The estimated mass missing from t_i
f	A significance threshold
i_{max}	The index of the element in X with the greatest estimate mass
c	$(c-1)t_{i_{max}}$ is an estimate of the total amount of mass that should be associated with $v_{i_{max}}$
$eloc$	The location within a region that is jointly occupied with another region that has the greatest probability of being jointly occupied

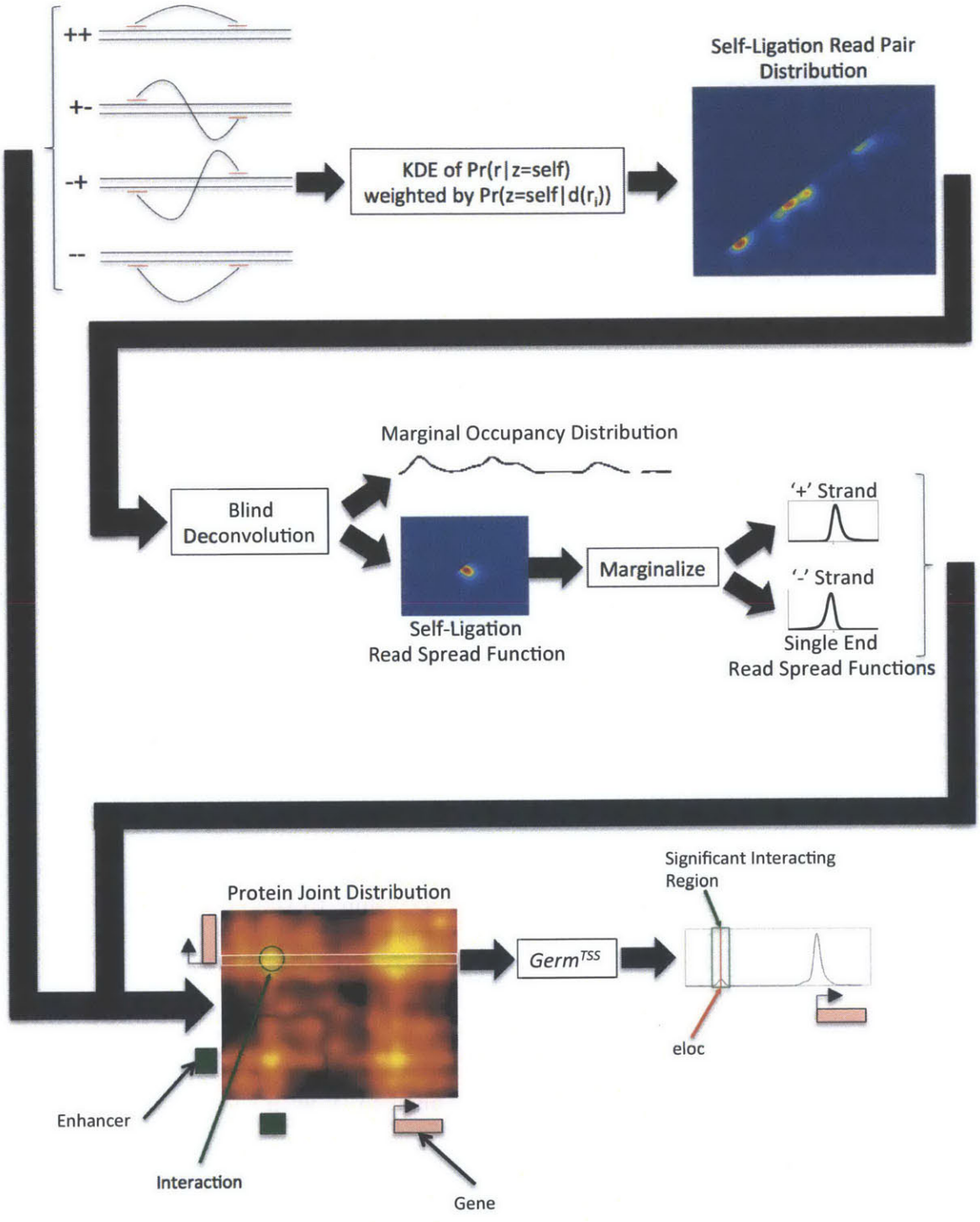


Figure 3-3: **The workflow of** *Germ* **and** $Germ^X$.

proach presents a compromise between the computational complexity of SA and the efficiency of IBD but has been shown to be quite sensitive to noise. The blind deconvolution component of GERM is an adaptation of an approach sometimes referred to as a double iteration algorithm. One of the most popular approaches to non-blind deconvolution is the Richardson-Lucy (RL) [52, 39] algorithm which applies the EM algorithm to image reconstruction. The double iteration approach recognizes that the original image and the PSF are symmetric in the model of convolution that describes the process of generating the blurred image. Based on this symmetry, EM iterations are applied alternately to update estimates of the original image and the PSF. More extensive reviews of blind deconvolution methods are contained in [30, 31, 24].

The only other explicit application of a blind deconvolution method to sequencing data that we are aware of is the CSDECONV [40] algorithm for analyzing ChIP-Seq data. This method has a similar double optimization structure to RL blind deconvolution. However, a significant difference is that the distribution of protein occupancy is modeled as a set of point locations. This modeling assumption is related to the assumptions made by the GPS and SPROUT methods. Also, rather than utilizing the EM algorithm for optimization, CSDECONV utilizes random-restart gradient descent.

3.2 The GERM algorithm

3.2.1 Estimating the 2D Self-Ligation Read Pair Distribution

We assume that ChIA-PET linker tags have been removed from the read pair sequences, that read pairs that are known to have resulted from chimeric ligation events because they contain two different linker tags have been removed, and that the remaining linkerless read pairs have been aligned to the reference genome. Let R be the set of all aligned read pairs such that each read pair $r_i \in R$ is represented by the pair of genomic coordinates to which the ends of the read pair align. We assume that the coordinates for each read pair are ordered so that if $r_i = \langle r_i^{(1)}, r_i^{(2)} \rangle$, then $r_i^{(1)} \leq r_i^{(2)}$. We also assume that each read pair has an associated label according to

the chromosome strands to which the ends align. There are four possible strandedness labels given the imposed ordering on the read pair ends. They are ++, -+, +-, and --. All self-ligation read pairs have strand orientation -+, but not all -+ read pairs were produced by self-ligation.

A distribution estimated from all -+ read pairs would not accurately model the distribution of self-ligation read pairs because self-ligation read pairs are much more likely to align within a short distance than inter-ligation read pairs. This is because the fragment length distribution induced by fragmentation limits the distance between which the ends of self-ligation read pairs may align whereas there is no constraint on the distance between which the ends of inter-ligation read pairs may align. To more accurately estimate the distribution of self-ligation read pairs, we weight the contribution of each -+ read pair by the estimated likelihood that the read pair was produced by self-ligation according to the distance between the aligned locations of the read pair ends.

Let z_i indicate whether -+ read pair r_i was produced by self-ligation or inter-ligation and $d(r_i)$ be the distance between the aligned locations of the ends of -+ read pair r_i. The likelihood that -+ read pair r_i was produced by self-ligation according to $d(r_i)$ can be expressed in terms of quantities that can be estimated from the data

$$\Pr(z_i = self|d(r_i)) = \frac{\Pr(d(r_i)|z_i = self)\Pr(z_i = self)}{\Pr(d(r_i))} \tag{3.1}$$

$\Pr(d(r_i))$ for all -+ read pairs can be estimated by applying an unweighted kernel approach

$$\hat{\Pr}(d(r) = x) = \sum_{i=1}^{N_{-+}} \frac{1}{h_{-+}N_{-+}} K_1\left(\frac{x - d(r_i)}{h_{-+}}\right) \tag{3.2}$$

N_{-+} is the total number of -+ read pairs and K_1 is a standard univariate Gaussian distribution. The bandwidth h_{-+} is a parameter that controls the trade-off between fitting the training data and discovering a smooth estimate. To choose an appropriate h_{-+} we use a least-squares cross-validation approach that minimizes the integrated square error (ISE) of $\hat{\Pr}(x)$.

$$ISE(\hat{f}) = \int (\hat{f} - f)^2 \tag{3.3}$$

The $ISE(\hat{\Pr}(d(r) = x))$ can be approximately minimized by minimizing for all -+ read pairs [54]

$$\sum_i \sum_j \frac{1}{\sqrt{2}h_{-+}} K_1 \left(\frac{d(r_i) - d(r_j)}{\sqrt{2}h_{-+}} \right) - \frac{2}{N_{-+}} \sum_i \left[\frac{\hat{\Pr}(d(r_i)) - \frac{1}{\sqrt{2\pi}}}{N_{-+} - 1} \right] \tag{3.4}$$

We cannot estimate $\Pr(d(r_i)|z_i = self)$ directly for the same reason that we cannot estimate the self-ligation read pair distribution directly. We can estimate $\Pr(d(r_i)|z_i = inter)$ directly because all non -+ read pairs are produced by inter-ligation. We also apply an unweighted kernel approach to estimate this distribution

$$\hat{\Pr}(d(r) = x|z = inter) = \sum_{i=1}^{N_{non-+}} \frac{1}{h_{non-+} N_{non-+}} K_1 \left(\frac{x - d(r_i)}{h_{non-+}} \right) \tag{3.5}$$

We choose an appropriate h_{non-+} by approximately minimizing the $ISE(\hat{\Pr}(d(r) = x|z = inter))$.

Given estimates for $\Pr(d(r_i))$ and $\Pr(d(r_i)|z_i = inter)$, we can estimate $\Pr(d(r_i)|z_i = self)$ by assuming that $\Pr(d(r_i))$ is a mixture of the distributions $\Pr(d(r_i)|z_i = self)$ and $\Pr(d(r_i)|z_i = inter)$

$$\Pr(d(r_i)) = \Pr(z_i = self) \Pr(d(r_i)|z_i = self) + \Pr(z_i = inter) \Pr(d(r_i)|z_i = inter) \tag{3.6}$$

By rearranging the terms in this equation we can obtain

$$\Pr(d(r_i)|z_i = self) = \frac{\Pr(d(r_i)) - \Pr(z_i = inter) \Pr(d(r_i)|z_i = inter)}{\Pr(z_i = self)} \tag{3.7}$$

The final missing component is $\Pr(z_i = self) = 1 - \Pr(z_i = inter)$. We assume that the average number of read pairs with each of the three strand orientations other

than -+ is a good estimator for the number of -+ read pairs that were produced by inter-ligation. We use this information to estimate $\Pr(z_i = inter)$

$$\hat{\Pr}(z_i = inter) = \frac{\text{avg. \# non -+ read pairs}}{\text{\# -+ read pairs}} \tag{3.8}$$

This allows us to estimate the self-ligation read pair distribution using a weighted kernel approach weighted by $\Pr(z = self|d(r_i))$

$$\hat{\Pr}(r = \langle x, y\rangle | z = self) = \sum_{i=1}^{N_{-+}} \frac{\Pr(z = self|d(r_i))}{h_{self}} K_2\left(\frac{\langle x, y\rangle - r_i}{h_{self}}\right) \tag{3.9}$$

where in this case K_2 is a bivariate standard Gaussian distribution with no correlation between the dimensions. To choose an appropriate bandwidth h_{self} we approximately minimize $ISE(\hat{\Pr}(r = \langle x, y\rangle | z = self))$ by minimizing

$$\begin{aligned}\sum_i \sum_j \frac{\Pr(z = self|d(r_i))\Pr(z = self|d(r_j))}{\sqrt{2}h_{self}} K_2\left(\frac{r_i - r_j}{\sqrt{2}h_{self}}\right) \\ -\frac{2}{N}\sum_i \left[\frac{\hat{\Pr}(r_i|z_i = self) - \frac{\Pr(z=self|d(r_i))}{\sqrt{2\pi}}}{\sum_{j\neq i}\Pr(z = self|d(r_j))}\right]\end{aligned} \tag{3.10}$$

3.2.2 Estimating the 1D Marginal Distribution of Protein Occupancy

We assume that the self-ligation read pair distribution is the result of the convolution of the marginal distribution of protein occupancy and a distribution that models DNA fragmentation which we will refer to as the read spread function (RSF). If we let q be the genomic location occupied by the protein,

$$\Pr(r = \langle x, y\rangle | z = self) = \sum_u \Pr(q = u) RSF(\langle x - u, y - u\rangle) \tag{3.11}$$

Simultaneously deconvolving the marginal distribution of protein occupancy and the RSF from the self-ligation read pair distribution is an example of a blind deconvolution problem. This problem commonly arises in the context of image processing.

It is often the case that a camera will systematically blur the images that it captures because of flaws in its lens. This blurring process is modeled as a convolution of the distribution of light that enters the camera lens with a point spread function (PSF) that is induced by the flaws in the lens. The PSF specifically describes the effect that the lens flaws will have on a theoretical point source of light. In our case, the RSF describes the manner in which self-ligation read pairs are likely to be distributed given the theoretical occupancy of the protein at a genomic location.

If we assume at first that the RSF is known, the marginal distribution of protein occupancy can be approximately recovered using a standard approach known as Richardson-Lucy (RL) deconvolution [39, 52]. The RL algorithm iteratively applies the following EM-like update

$$\hat{\Pr}_{i+1}(q = u) = \hat{\Pr}_i(q = u) \left\{ \sum_x \sum_y \left[\frac{\hat{\Pr}(r = \langle x, y \rangle | z = self)}{\sum_v \hat{\Pr}_i(q = v) RSF(\langle x - v, y - v \rangle)} \right] RSF(-\langle x - u, y - u \rangle) \right\} \tag{3.12}$$

RL deconvolution has been shown empirically to converge to a maximum-likelihood estimate for $\Pr(q = u)$ and preserves the non-negativity and sum of the initial guess $\Pr_0(q = u)$. To extend RL deconvolution to the blind case, we take an approach similar to that proposed in [17] and alternate the updates described by Equation 3.12 with the following updates

$$\widehat{RSF}_{i+1}(\langle x, y \rangle) = \widehat{RSF}_i(\langle x, y \rangle) \left\{ \sum_u \left[\frac{\hat{\Pr}(r = \langle x - u, y - u \rangle | z = self)}{\sum_v \widehat{RSF}_i(\langle x - u - v, y - u - v \rangle) \hat{\Pr}(q = v)} \right] \hat{\Pr}(q = -u) \right\} \tag{3.13}$$

The overall procedure then entails going back and forth between updating $\hat{\Pr}(q =$

u) for several iterations while holding $\widehat{RSF}(\langle x-u, y-u\rangle)$ fixed and then updating $\widehat{RSF}(\langle x-u, y-u\rangle)$ for several iterations while holding $\hat{\Pr}(q=u)$ fixed. Despite the unconstrained nature of this approach, the recovered RSF conforms to our expectations. The RSF in Figure 3-4 is typical of what is recovered from RNA PolII ChIA-PET data. Given a location bound by the protein, we would expect the most likely alignment of the ends of self-ligation read pairs to be roughly equidistant to the occupied location with the distance from the occupied location determined by the degree of fragmentation. The typical RSF that we estimate has the greatest value along the line through the origin that is perpendicular to the identity line. Points along this line reflect self-ligation read pairs that align equidistantly to the occupied location which is represented by the origin in the RSF. The distance of the peak in the RSF from the origin reflects the most likely fragment size generated by the sonication step. Thus, the RSF that we recover using our blind deconvolution approach conforms to our expectations and provides useful information about the fragmentation step of the ChIP procedure.

Note that this distribution is similar to the self-ligation read distribution learned by SPROUT from punctate data (Figure 2-1a). Both distributions model the arrangement of self-ligation read pairs relative to a location occupied the protein of interest. The similarity between these distributions is to be expected because these distributions are induced by the DNA fragmentation step that is common to all ChIA-PET experiments. SPROUT and GERM differ in the assumptions that they make about the manner in which proteins associate with the genome. Yet, despite taking different approaches to modeling ChIA-PET data, both methods recover appropriately similar information about the ChIA-PET method itself.

Efficiently estimating the genome-wide protein occupancy distribution

RL blind deconvolution works well for deconvolving the protein occupancy distribution for regions of the genome that are on the order of megabases in size. However, the time that it would take to deconvolve the full genome-wide distribution of protein occupancy is impractical. Based on observations made about typical RSFs estimated

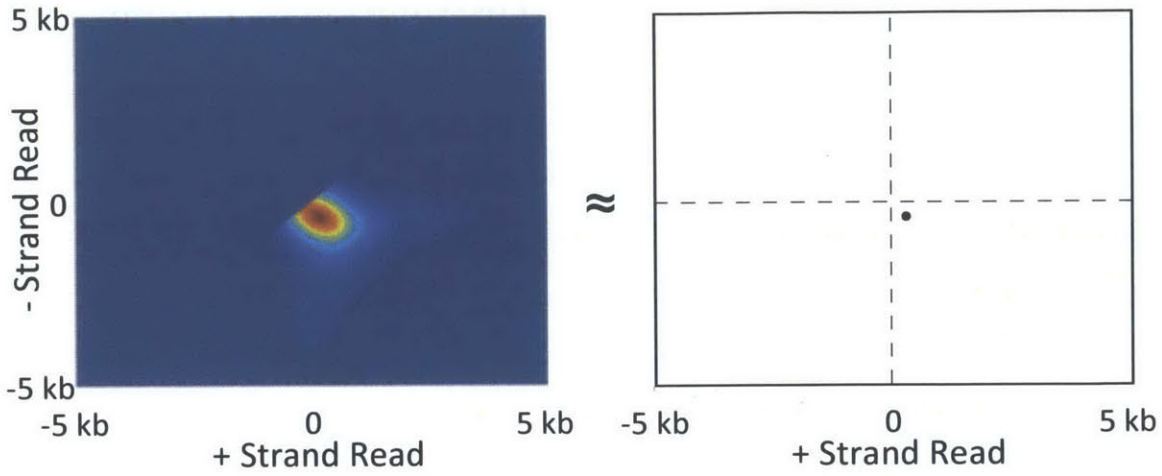


Figure 3-4: **A typical read spread function estimated from RNA PolII ChIA-PET data and an approximation of it that makes deconvolution efficient**

by RL blind deconvolution from portions of real ChIA-PET datasets, we devised a highly efficient procedure that achieves a level of accuracy comparable to full RL blind deconvolution. We observed that typical RSFs estimated by RL blind deconvolution from portions of real datasets are unimodal and sharply peaked. This implies that the RSF can be approximated by a function with all of its mass at the peak of the RSF as in Figure 3-4. This approximation allows for a very efficient deconvolution procedure. If the peak of the estimated RSF is at $\langle -\lambda, \lambda \rangle$, we estimate the protein occupancy distribution as

$$\hat{\Pr}(q = u) \propto \hat{\Pr}(r = \langle u - \lambda, u + \lambda \rangle | z = self) \tag{3.14}$$

In summary, to estimate the marginal distribution of protein occupancy from a full genome-wide ChIA-PET dataset we first estimate the genome-wide self-ligation read pair distribution. We then apply RL blind deconvolution to a 5 megabase region of the genome to obtain a good estimate for the RSF. Finally, we identify the peak of the estimated RSF and estimate the distribution of RNA PolII occupancy as in (3.14).

3.2.3 Estimating the 2D Joint Distribution of Protein Occupancy

Chromatin looping allows proteins to simultaneously occupy two genomic locations [64]. Inter-ligation read pairs can be thought of as samples from a joint distribution of protein occupancy with positional noise introduced by fragmentation. We make several assumptions about this process. We assume that the inter-ligation read pairs are based on independent samples from the joint distribution of protein occupancy. We associate the lower coordinate protein location $q^{(1)}$ with the lower coordinate end of the read pair $r^{(1)}$ and the higher coordinate protein location $q^{(2)}$ with the higher coordinate end of the read pair $r^{(2)}$.

$$\Pr(q = \langle u, v \rangle | R_{inter}) = \frac{1}{N_{inter}} \sum_{r_i \in R_{inter}} \Pr(q = \langle u, v \rangle | \langle r_i^{(1)}, r_i^{(2)} \rangle) \tag{3.15}$$

$$= \frac{1}{N_{inter}} \sum_{r_i \in R_{inter}} \left[\Pr(q^{(1)} = u | \langle r_i^{(1)}, r_i^{(2)} \rangle) \Pr(q^{(2)} = v | q^{(1)} = u, \langle r_i^{(1)}, r_i^{(2)} \rangle) \right] \tag{3.16}$$

$$= \frac{1}{N_{inter}} \sum_{r_i \in R_{inter}} \Pr(q^{(1)} = u | r_i^{(1)}) \Pr(q^{(2)} = v | q^{(1)} = u, r_i^{(2)}) \tag{3.17}$$

The last equality reflects an assumption that we make that the location occupied by the protein is independent of the read pair end that it is not associated with. We will demonstrate that these terms are non-zero in only a relatively small window around their associated read pair end and that the non-associated read pair end has minimal effect on the manner in which we compute these terms. We transform the first term within the sum into quantities that we can compute using Bayes' Theorem

$$\Pr(q^{(1)} = u | r_i^{(1)}) = \frac{\Pr(r_i^{(1)} | q^{(1)} = u) \Pr(q^{(1)} = u)}{\Pr(r_i^{(1)})} \tag{3.18}$$

We assume that we can obtain $\Pr(r_i^{(1)} | q^{(1)} = u)$ by marginalizing the RSF that was estimated during the blind deconvolution step. For read pair ends that align to

the - strand

$$\Pr(r_i^{(\cdot)}|q^{(\cdot)} = u) = \sum_y RSF(\langle r_i^{(\cdot)} - u, y - u\rangle) \tag{3.19}$$

Correspondingly, for read pair ends that align to the + strand

$$\Pr(r_i^{(\cdot)}|q^{(\cdot)} = u) = \sum_x RSF(x - u, r_i^{(\cdot)} - u\rangle) \tag{3.20}$$

$\Pr(q^{(1)} = u)$ is the distribution of protein marginal occupancy that was estimated in the previous step. The prior read distribution $\Pr(r_i^{(1)})$ reflects any factors that might influence the alignment of reads to locations in the genome. Such factors might include the uniqueness of the sequence around that location in the genome and bias in the library preparation or sequencing for the sequence around that location. We assume that $\Pr(r_i^{(1)})$ is uniform in this work. However, future work may be improved by utilizing a more informative prior distribution.

We also transform the second term within the sum in (3.17) using Bayes' Theorem

$$\begin{aligned}
&\Pr(q^{(2)} = v|q^{(1)} = u, r_i^{(2)}) \\
&= \frac{\Pr(r_i^{(2)}|q^{(1)} = u, q^{(2)} = v)\Pr(q^{(2)} = v|q^{(1)} = u)}{\Pr(r_i^{(2)}|q^{(1)} = u)} &(3.21)\\
&\approx \frac{\Pr(r_i^{(2)}|q^{(2)} = v)\Pr(q^{(2)} = v)}{\Pr(r_i^{(2)})} &(3.22)
\end{aligned}$$

The approximation in (3.22) incorporates assumptions to simplify all terms involved. We assume that $r_i^{(2)}$ only depends on the location of protein occupancy that it is associated with, and hence $\Pr(r_i^{(2)}|q^{(1)} = u, q^{(2)} = v) \approx \Pr(r_i^{(2)}|q^{(2)} = v)$ which we obtain by marginalizing the estimated RSF. We next assume that $q^{(1)}$ and $q^{(2)}$ are independent. This is clearly not true, since otherwise we would have no need of estimating their joint distribution. But, since $\Pr(r_i^{(2)}|q^{(2)} = v)$ is only non-zero in a relatively small range around v, the purpose of $\Pr(q^{(2)} = v|q^{(1)} = u)$ is mainly to fine tune the probability that $q^{(2)} = v$ if $r_i^{(2)}$ falls within that range. We expect the loca-

tions of peaks of $\Pr(q^{(2)} = v|q^{(1)} = u)$ to roughly agree with peaks of $\Pr(q^{(2)} = v)$ if they exist, and so we assume that we can swap one for the other in this case. Finally, we assume that $r_i^{(2)}$ is independent of the location of protein occupancy that it is not associated with, allowing us to substitute $\Pr(r_i^{(2)})$ for $\Pr(r_i^{(2)}|q^{(1)} = u)$.

These transformations allow us to write the estimated joint distribution of protein occupancy as

$$\hat{\Pr}(q = \langle u, v\rangle | R_{inter}) \propto \sum_{r_i \in R_{inter}} \Pr(r_i^{(1)}|q^{(1)} = u) \Pr(q^{(1)} = u) \Pr(r_i^{(2)}|q^{(2)} = v) \Pr(q^{(2)} = v) \quad (3.23)$$

$Germ^X$: Estimating the Conditional Distribution of Protein Occupancy with a Set of Locations X

In many situations we are interested in estimating the joint occupancy of a protein with a set of genomic locations X. For example, when analyzing RNA PolII ChIA-PET data, a common query might be to detect regions that are jointly occupied by RNA PolII along with a location from set of annotated transcription start sites (TSSs). If we define TSS to be a set of annotated TSSs, we refer to $Germ^{TSS}$ as the process of estimating $\Pr(q = \langle u, v\rangle | R_{inter})$ only for $v \in TSS$.

3.2.4 Evaluating the Significance of Portions of Estimated Distributions of Marginal and Joint Protein Occupancy

Once we have estimated distributions of marginal and joint protein occupancy from ChIA-PET data we evaluate the significance of the estimated protein occupancy within a given region or the joint occupancy within a given pair of regions. We describe our approach as applied to a marginal distribution of protein occupancy and then extend the approach to joint distributions. Given a genomic region reg of size w base pairs, let $p = \sum_{u \in reg} \hat{\Pr}(q = u)$. If we let $Z \sim \text{Binomial}(N_{self}, p)$ and $Y \sim \text{Binomial}(N_{self}, \frac{w}{M})$ where M is the size of the mappable genome, we then

evaluate the significance of the protein occupancy within reg as $\Pr(Y > Z)$. In other words, we calculate the probability that more self-ligation read pairs would be associated with reg according to a uniform distribution of protein occupancy than would be associated with reg according to the estimated distribution of protein occupancy.

We extend this approach to evaluating the significance of pairs of regions according to a joint distribution of protein occupancy. Given a pair of regions reg_a and reg_b, let $p_{joint} = \sum_{u \in reg_a} \sum_{v \in reg_b} \hat{\Pr}(q = \langle u, v \rangle | R_{inter})$, $p_a = \sum_{u \in reg_a} \hat{\Pr}(q = u)$, and $p_b = \sum_{u \in reg_b} \hat{\Pr}(q = u)$. If we then let $Z \sim \text{Binomial}(N_{inter}, p_{joint})$ and $Y \sim \text{Binomial}(N_{inter}, p_a p_b)$, we then evaluate the significance of the joint protein occupancy of the regions reg_a and reg_b as $\Pr(Y > Z)$.

Significance evaluation for $Germ^X$

The estimate $\hat{\Pr}(q = \langle u, v \rangle | R_{inter})$ for $v \in X$ that is obtained by applying $Germ^X$ is void of mass for much of its domain. This is because not enough inter-ligation read pairs can be sequenced to fully explore this space given current technologies. Without considering the mass that is missing from the estimate of $\hat{\Pr}(q = \langle u, v \rangle | R_{inter})$, the significance of portions of the distribution for which mass is estimated will be overestimated. To remedy this issue, we introduce a method for estimating how much mass is missing from the estimate of $\hat{\Pr}(q = \langle u, v \rangle | R_{inter})$ in order to more accurately evaluate the significance of portions of this distribution. We assume an ordering on the $v_i \in X$ and let $t_i = \sum_u \hat{\Pr}(q = \langle u, v_i \rangle | R_{inter})$ and $m_i = \hat{\Pr}(q = v_i)$. If we assume that there is some amount of mass τ_i that is missing from t_i, then we can find a setting of the τ_i such that $\frac{t_i + \tau_i}{\sum_i t_i + \tau_i} = \frac{m_i}{\sum_i m_i}$. However, there are many valid settings of the τ_i and larger values of the τ_i will cause portions of the estimated distribution to be evaluated as less significant.

To choose an appropriate setting of the τ_i we introduce a procedure that allows us to choose τ_i large enough to avoid overestimating the significance of portions of the estimated distribution. We first choose a set of candidate regions for each $v_i \in X$ which we will evaluate for significance based on $\hat{\Pr}(q = \langle u, v \rangle | R_{inter})$. We

do this by setting a threshold f and adding a region reg to the set for v_i if $\forall u \in reg, \hat{\Pr}(\langle u, v_i \rangle | R_{inter}) > f$. We then identify an i_{max} such that $\forall i, t_{i_{max}} \geq t_i$. We choose some $c > 1$ and set $\tau_{i_{max}} = (c-1)t_{i_{max}}$. We hold $\tau_{i_{max}}$ fixed and apply an iterative procedure to find settings for τ_i ($i \neq i_{max}$) such that $\frac{t_i+\tau_i}{\sum_i t_i+\tau_i} = \frac{m_i}{\sum_i m_i}$. For each iteration, we cycle through $i \neq i_{max}$ and compute

$$\tau_i = \frac{m_i \sum_{j \neq i} (t_j + \tau_j)}{\sum_{j \neq i} m_j} \tag{3.24}$$

Once this converges, we evaluate the significance of the regions defined using the threshold f in the following way. For a region reg in the set for v_i we let $p = \frac{\sum_{u \in reg} \hat{\Pr}(\langle u, v_i \rangle | R_{inter})}{t_i+\tau_i}$ and $p' = \sum_{u \in reg} \hat{\Pr}(u)$. If we then let $Z \sim \text{Binomial}(N_{inter}, p)$ and $Y \sim \text{Binomial}(N_{inter}, p')$, the significance of the estimated joint protein occupancy of v_i and reg is $\Pr(Y > Z)$. We evaluate the significance of the regions in the sets for all $v \in X$ and identify the regions that have an associated $\Pr(Y > Z)$ less than some threshold such as 0.05. We call these regions significant. For each region, we also note the number of read pairs in R_{inter} that contributed to p for that region. If the ratio of the number of significant regions supported by only one read pair to the total number of significant regions is greater than some target threshold, such as 0.1, we increase c and begin the process of finding a new set of τ_i. If there are too few significant regions supported by one read pair with $\Pr(Y > Z) < 0.05$ we reduce c and find new τ_i. In this manner we search for c that achieves a target fraction of weakly supported jointly occupied regions within the set of all regions that evaluate as significant.

3.3 Evaluating GERM

We applied GERM^{TSS} to RNA PolII ChIA-PET data from mouse embryonic stem (mES) cells to evaluate the ability of GERM^{TSS} to identify regions that interact with TSSs and exhibit characteristic features of enhancers. As described in Chapter 1, enrichment for Mediator, p300, Cohesin, and H3K27ac is associated with active enhancers. We also know that Oct4, Sox2, and Nanog frequently bind to active enhancers in mES cells because they have been shown to be important regulators of

the mES cell state. We assume that chromatin loops form to allow active enhancers to become spatially proximal to the TSSs of the genes that they regulate. In this section we examine the regions that we discover to be jointly occupied by PolII with annotated TSSs based on the assumption that joint occupation by PolII should reflect chromatin looping between active enhancers and TSSs. We will demonstrate that the regions that we discover do in fact exhibit the characteristic features of active enhancers that we examined.

We processed the RNA PolII ChIA-PET sequence data by filtering out chimeric ligation read pairs that contain two different linker sequences, aligning the read pairs using BOWTIE, and removing paired positional duplicates to avoid spurious results from PCR artifacts. We obtained ChIP-Seq sequence data for Med1, p300, Smc1a, H3K27ac, Oct4, Sox2, and Nanog as well as whole cell extract (WCE) data in mES cells [11, 26, 62]. These data were also aligned to the reference genome using BOWTIE. We used annotated TSSs from the UCSC knownGene database [27] to discover regions jointly occupied by PolII with TSSs using GERM^{TSS}. For each region reg in this set, we identified the location $\forall v \in TSS, eloc = \max_{u \in reg} \hat{\Pr}(q = \langle u, v \rangle | R_{inter})$. $eloc$ is the location within reg that is most likely to be jointly occupied by RNA PolII with some $v \in TSS$.

In order to avoid detecting interactions between TSSs, we conservatively selected for regions with $eloc$ that is at least 2 kb away from any annotated TSS. This left us with 2924 regions that interact with a TSS and do not contain TSSs themselves. We centered 500 bp windows on the $eloc$ within each region and evaluated the enrichment of each of the ChIP-Seq datasets compared to WCE within those windows as shown in Table 3.2. Notably, over 90% of the GERM^{TSS} identified regions are enriched for the Mediator component Med1. Enrichment for p300 and the Cohesin component Smc1a is high as well (84.7% and 78.8% respectively). Almost 90% of the regions are enriched for H3K27ac. Oct4 is enriched in almost 80% of these regions while Sox2 and Nanog are enriched in nearly 50%. The enrichment of these transcription factors is somewhat less than Mediator, p300, and Cohesin, although still strongly suggestive that these are active enhancers. It may be that these factors are not necessarily

Table 3.2: **Regions identified to interact with TSSs that are enriched for enhancer-associated ChIP-Seq data** 2924 TSS-distal, TSS jointly occupied regions were identified using $Germ^{TSS}$ and 3098 were identified from the results from [68]. For the $Germ^{TSS}$ regions, the most likely jointly occupied location was identified and for the regions taken from the [68] results the midpoint was identified. A 500 bp region centered on the identified location within each region was evaluated for ChIP-Seq data enrichment. For each ChIP-Seq dataset, this table includes the number of regions that are enriched and the percentage of the total number of each type of region that number constitutes.

Factor	# $Germ^{TSS}$	% $Germ^{TSS}$	# Zhang *et al.*	% Zhang *et al.*
Med1	2648	90.6	1908	61.6
p300	2477	84.7	1782	57.5
Smc1a	2303	78.8	1675	54.1
H3K27ac	2629	89.9	2014	65.0
Oct4	2257	77.2	1457	47.0
Sox2	1336	45.7	952	30.7
Nanog	1433	49.0	1015	32.8

present at all active enhancers despite the importance of these factors in maintaining pluripotency in mES cells.

Through a comparison with the results published with the ChIA-PET data that we analyzed, we discovered that a greater percentage of the GERM^{TSS} identified regions are enriched for all of the ChIP-Seq datasets that we considered and that GERM^{TSS} identifies a larger absolute number of locations enriched for each dataset. To make this comparison, we obtained the interaction calls from [68] based on the same data to which we applied GERM^{TSS}. We filtered out the interactions that do not contain a TSS within either anchor region. Since these interactions do not include estimates of the most likely locations within the anchor regions that are jointly occupied by RNA PolII, we chose the midpoint of each anchor region as the approximate *eloc*. We further filtered the interactions to identify the set of interactions that contain a TSS within one anchor region and for which the midpoint of the other anchor region is at least 2 kb away from any TSS. In this manner we identified 3098 regions from the published results. We examined the same ChIP-Seq data to evaluate whether these regions exhibit properties of active enhancers. We centered a 500bp window

around the midpoints of these regions and evaluated the enrichment in that window for each ChIP-Seq dataset. We found that much lower percentages of the regions were significantly enriched for each of the datasets than the GERM^{TSS} identified regions.

To further demonstrate the spatial accuracy of GERM^{TSS}, we visualized the ChIP-Seq data contained within the regions identified by both methods as shown in Figure 3-5. Each row within each box represents ChIP-Seq data from one of the regions identified by each method. The rows are sorted by the measure of significance assigned to the corresponding interaction by the method. We have already demonstrated in Table 3.2 that GERM^{TSS} identifies a greater number of regions that are enriched for all of the ChIP-Seq datasets. Figure 3-5 illustrates the spatial accuracy of the ChIP-Seq enrichment within 6 kb windows centered on the *eloc* or midpoint of the GERM^{TSS} and Zhang et al. regions respectively. Stronger enrichment, as indicated by darker shades of blue, tends to exist in the center of the GERM^{TSS} identified regions. The Zhang et al. regions do not exhibit the same centering of ChIP-Seq enrichment. This comparison illustrates the usefulness of the detailed estimates of joint occupation by GERM^{TSS} for identifying putative enhancers that regulate genes through chromatin interactions.

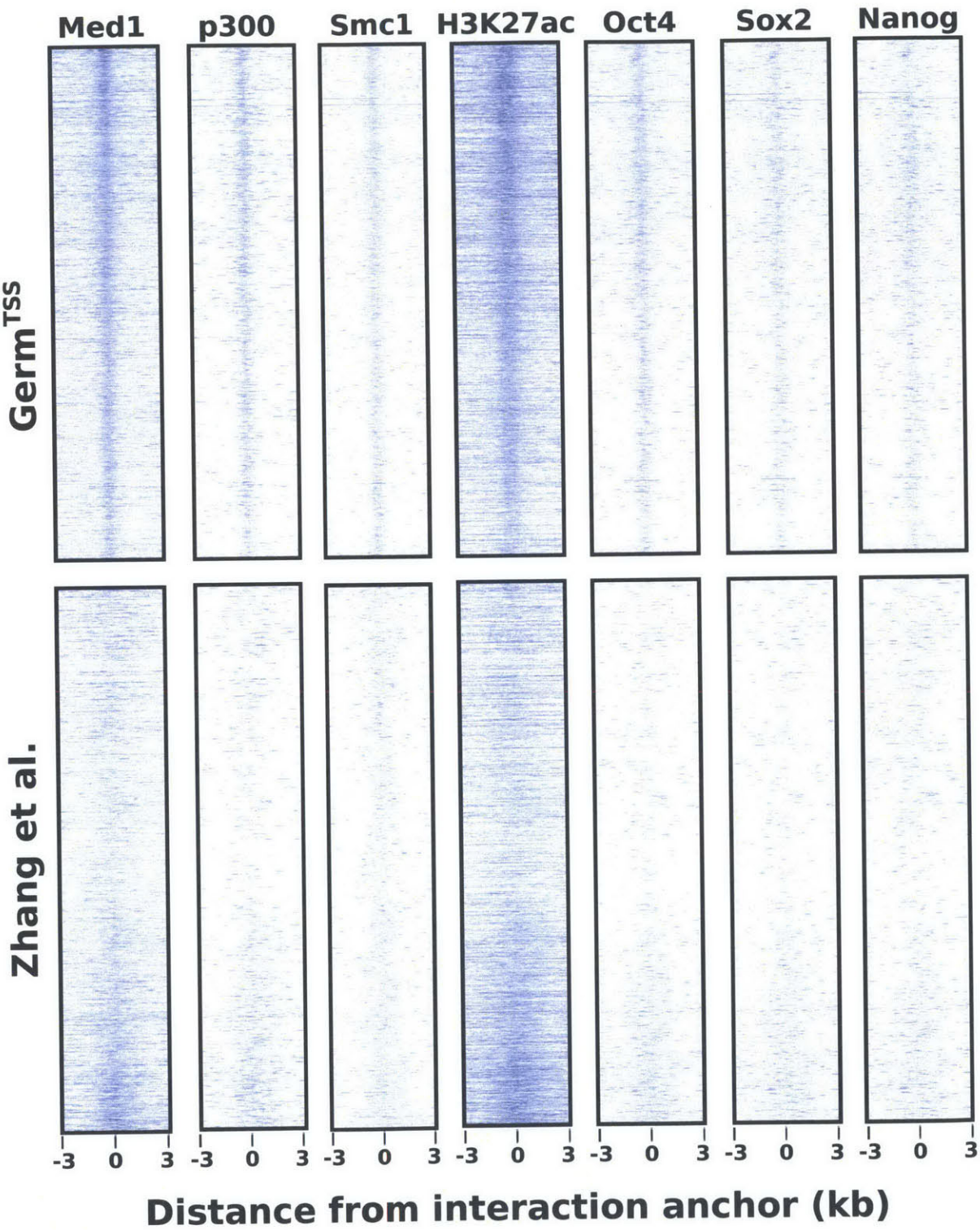


Figure 3-5: **Visualization of ChIP-Seq data in regions detected to interact with TSSs.** The top row of boxes contains TSS-distal, TSS jointly occupied regions identified by $Germ^{TSS}$. The bottom row of boxes contains the corresponding regions from [68]. The 6 kilobase regions are centered on the estimated *eloc* or midpoint and are ordered by the significance associated with the interaction. Each column represents data from a ChIP-Seq dataset that is associated with active enhancers.

Chapter 4

Enhancer utilization during motor neuron development

In Chapter 3 we introduced GERM, a novel algorithm for analyzing ChIA-PET data that presents a detailed view of the occupancy of the genome by a protein of interest. We applied a variant of GERM denoted GERM^{TSS} to RNA Polymerase II (PolII) ChIA-PET data in mES cells in order to discover regions that interact with annotated transcription start sites (TSSs). We demonstrated that the regions that interact with TSSs that are not themselves TSSs exhibit characteristics of active enhancers. Furthermore, we showed that with GERM^{TSS} we are able to identify putative enhancers with high spatial accuracy. Here we apply GERM^{TSS} to an additional PolII ChIA-PET dataset in mouse motor neuron progenitors (pMN). We compare the results obtained by analyzing independently derived biological replicate datasets and make observations about the replicability and sensitivity of the ChIA-PET methodology. Keeping in mind the caveats that we learn from this comparison, we make several observations about differential enhancer usage during motor neuron development.

4.1 Sensitivity and Specificity of GERM^{TSS} results

We generated replicate PolII ChIA-PET libraries from independent derivations of pMN cells that were produced by in vitro differentiation [63]. After filtering out

chimeric ligation read pairs, aligning the read pairs, and removing paired positional duplicates we obtained more than 30 million aligned read pairs. We processed the published ES PolII ChIA-PET data [68] in the same fashion resulting in over 60 million aligned read pairs. To demonstrate that GERM^{TSS} discovers consistent results from independent ChIA-PET experiments in the same cell type, we first applied GERM^{TSS} independently to the replicate pMN datasets. We discovered 56,913 interactions in one replicate and 46,466 interactions in the other replicate. In 23,954 cases the two sets contained interactions between the same TSS and locations that are within 500 bp. Since we are particularly interested in interactions between TSSs and distal enhancers, we identified the interactions from the two sets such that one location involved in the interaction is at least 2 kb from any annotated TSS and examined the overlap between these TSS-nonTSS interactions. Out of the 6,634 and 5,872 TSS-nonTSS interactions in the two sets, 1,822 pairs of interactions involve nonTSS locations within 500 bp.

There are several potential factors that may limit the overlap between the results from the two replicates. One factor that is not well understood and is difficult to characterize given current technology is biological variability. It is not clear how stable chromatin interactions are and to what degree cells in a population take on similar chromatin conformations even if the population is homogeneous in terms of cell type [9, 33, 46]. ChIA-PET datasets also contain very little dynamic range compared to more mature technologies such as ChIP-Seq. The number of possible chromatin interactions is quadratic in the number of possible binding events. As a consequence, the ideal complexity of ChIA-PET libraries would be quadratic relative to the complexity of high quality ChIP-Seq libraries. However, it is not yet feasible to prepare ChIA-PET libraries with this complexity. Furthermore, sequencing such libraries would be very expensive even given relatively inexpensive current sequencing technologies. Because of these limitations, we expect that there is a fairly high false-negative rate inherent to existing ChIA-PET datasets that results in limited overlap between the sets of chromatin interactions detected from independently performed ChIA-PET experiments.

To assess the degree to which GERM^{TSS} results are affected by experimental noise, we constructed a randomly permuted ChIA-PET dataset from one of the pMN replicate datasets. We took each read pair and randomly swapped an end with an end of another read pair from the same chromosome. This resulted in a dataset with the same number of read pairs as well as the same marginal read distribution as a real ChIA-PET dataset, but in which the read pairings have been scrambled such that they should not contain real information about PolII joint occupancy. We applied the GERM^{TSS} algorithm and discovered 4,658 significant interactions of which only 231 overlap with interactions discovered from the original dataset. We also limited our comparison to the 521 TSS-nonTSS interactions called from the randomized dataset of which only 10 overlap with TSS-nonTSS interactions discovered from the original dataset. The much smaller number of interactions identified from the randomized dataset as well as the low overlap between the results from the randomized dataset and the original dataset suggest that the interactions identified by GERM^{TSS} from nonrandomized datasets reflect real signal in the data and that the effects of experimental noise are relatively minimal.

We estimated the total number of discoverable TSS-nonTSS interactions in pMNs and discovered that it is very unlikely that the same overall set of TSS-nonTSS interactions exist in ES cells. We assumed that the same set of TSS-nonTSS interactions were discoverable in the cell populations used for the replicate pMN ChIA-PET experiments. We also assumed that the TSS-nonTSS interactions discovered by GERM^{TSS} are sampled from this set with equal probability. Given these assumptions we estimated that 21,380 (Figure 4-1A) GERM^{TSS} discoverable TSS-nonTSS interactions are present in pMN cells. We computed this by maximizing the likelihood of the hypergeometric distribution given the size of the overlap between the TSS-nonTSS interactions discovered by GERM^{TSS} from the two replicates. We then applied GERM^{TSS} to one of the ES cell replicates to discover 11,974 TSS-nonTSS interactions. Despite the much larger number of TSS-nonTSS interactions discovered by GERM^{TSS} from this dataset compared to either of the pMN replicates, only 733 of these interactions overlap with the TSS-nonTSS interactions discovered from the first pMN replicate. We

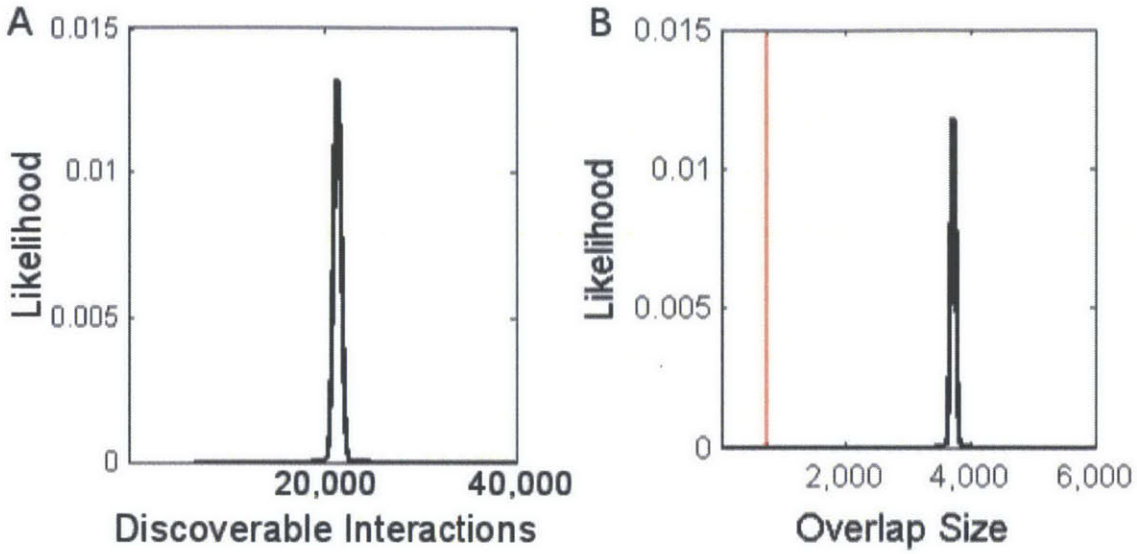


Figure 4-1: **Examining the likelihood that the same set of TSS-nonTSS interactions exist in pMN and ES cells.** (A) The likelihood of the hypergeometric distribution given the sizes of the sets of TSS-nonTSS interactions discovered by GERM^{TSS} from the two replicate pMN datasets while varying the total number of discoverable TSS-nonTSS interactions. The most likely total number of discoverable TSS-nonTSS interactions is 21,380. (B) The likelihood of the hypergeometric distribution assuming the 6,634 TSS-nonTSS interactions discovered from the first pMN replicate and the 11,974 discovered from the first ES replicate were sampled from the same total discoverable set of 21,380 TSS-nonTSS interactions while varying the size of the overlap between the sets of discovered interactions. The actual overlap between the two sets is 733 and is indicated by the red line. The probability of observing an overlap this small or smaller is effectively zero. Given our assumptions this suggests that it is very unlikely that the same set of discoverable TSS-nonTSS interactions are present in both pMN and ES cells.

find that the probability that the GERM^{TSS} discovered ES and pMN TSS-nonTSS interactions were sampled from the same set of interactions is effectively zero (Figure 4-1B). Therefore, despite the high false negative rate inherent to ChIA-PET, it is very likely that the data analyzed for ES and pMN cells reflect different sets of TSS-nonTSS interactions that are present in the two cell types.

Based on the observations made from analyzing the permuted dataset, we made the assumption that experimental noise is not a major source of variation in the results that we obtain from replicate experiments and combined the replicates for each cell type to maximize the number and confidence of interactions that we detect. We

applied GERM^{TSS} to the two combined datasets and discovered 93,582 chromatin interactions involving TSSs in ES cells and 82,177 such interactions in pMNs. As a positive control we ensured that we detect several previously characterized interactions in ES cells (Figure 4-2). We found that a majority of detected TSS based interactions are with other annotated TSSs (Figure 4-3). These interactions have been proposed to indicate the gathering of co-regulated genes into so-called transcription factories [8]. A minority of the detected TSS-based interactions are with genomic locations distal to annotated TSSs that we call nonTSS locations. To ensure that nonTSS locations are distinct from TSSs we conservatively define nonTSS locations to be locations that are at least 2 kb from any annotated TSS.

4.2 Enhancer properties of regions that interact with TSSs

To investigate the hypothesis that TSS-nonTSS interactions represent functional interactions with active enhancers we gathered ChIP-Seq data for the active enhancer related marks H3K27ac, Med1, Med12, p300, and Smc1a in ES cells [11, 26]. Most of the TSS-nonTSS interactions (72.6%) that we discovered in ES cells are enriched in a 500 bp window centered on the nonTSS end for all of the enhancer related marks we considered. These results suggest that enhancers that interact with TSSs are likely to harbor nucleosomes acetylated at H3K27 by p300 and that an ensemble including Mediator and Cohesin stabilize the chromatin interaction allowing PolII to jointly occupy the enhancer along with the TSS.

We reasoned that the ChIP-seq enhancer related marks and the GERM identified enhancers would be highly spatially concordant. We applied the GEM algorithm [20] to accurately identify locations of binding events in the Med1, p300, and Smc1a data. We also computed the peaks of the PolII conditional joint occupancy distribution within each nonTSS location. These peaks reflect the most likely anchor position of the interaction between the nonTSS location and the TSS. Strikingly, the GERM

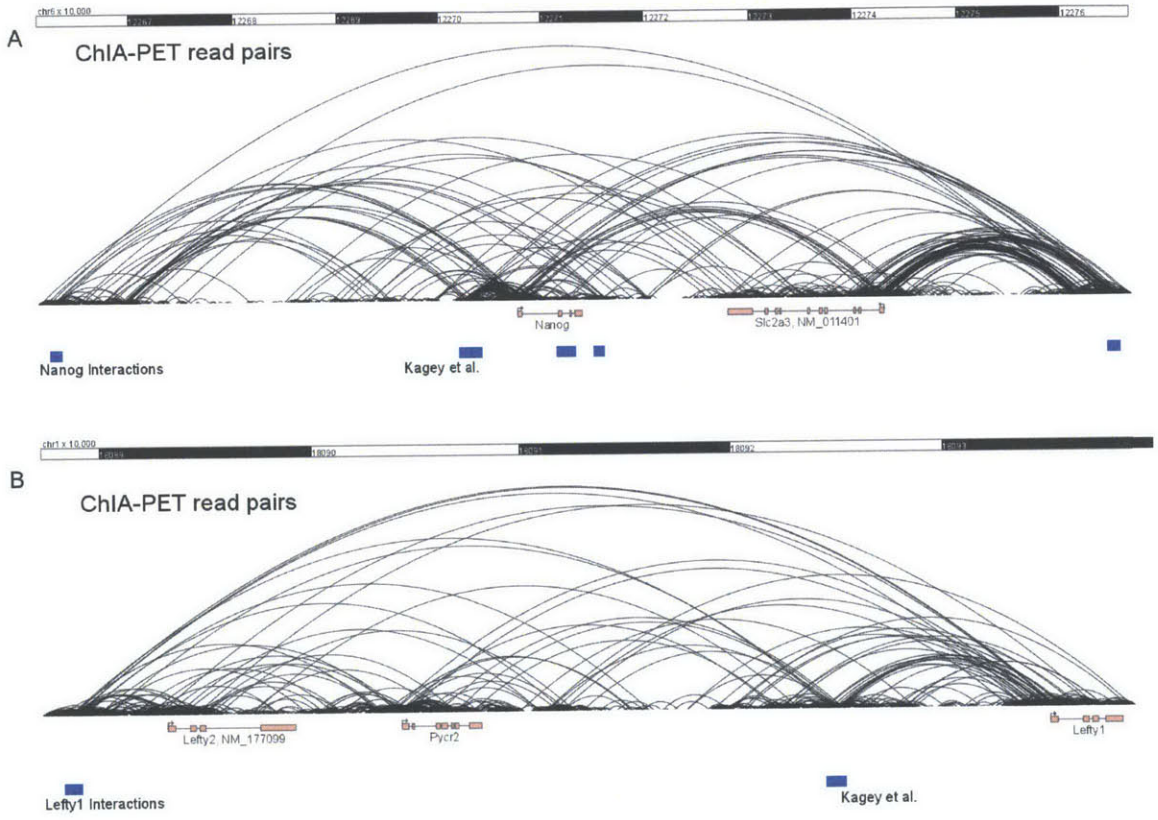


Figure 4-2: **Discovery of interactions that were previously characterized in the literature.** GERM^{TSS} identified regions that interact in ES cells with the TSSs of (A) Nanog and (B) Lefty1 and are at least 2 kb from any TSS are represented by blue boxes. The interacting regions that were verified by 3C [26] are labeled "Kagey et al."

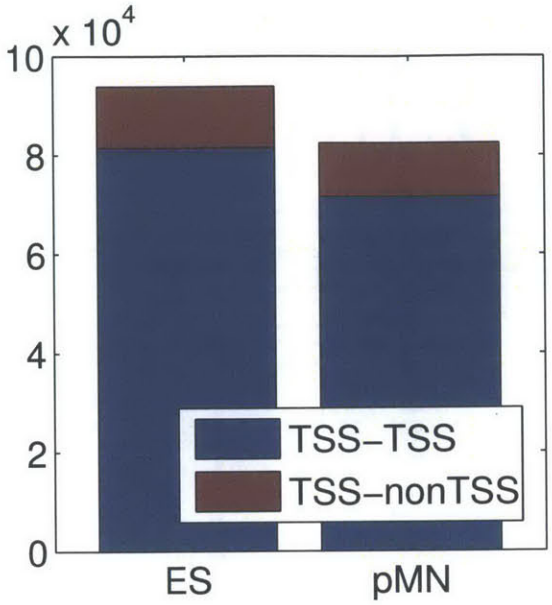


Figure 4-3: **Breakdown of GERM^{TSS} identified interactions.**

interaction peaks line up very closely with the locations of enhancer related binding events (Figure 4-4). This evidence further suggests that the joint occupancy of enhancers and TSSs by PolII is directly related to the occupancy of enhancers by mediator, cohesin, and p300.

We discovered that individual GERM identified enhancers and transcription start sites can exhibit 20 or more distinct interactions. By clustering proximal nonTSS ends of TSS-nonTSS interactions, we discovered 9,127 putative enhancers that are utilized by genes in one or both cell types. 9,574 TSSs interact with putative enhancers in one or both cell types. TSSs in both cell types interact with as many as 20 distinct putative enhancers (Figure 4-5A). We also found that several highly utilized enhancers engage in an even greater degree of connectivity (Figure 4-5B).

We performed RNA-Seq with both ES cells and pMNs and discovered that transcription is correlated with the degree of connectivity of TSSs with nonTSS locations (Figure 4-6). It has been previously demonstrated that greater enrichment of PolII at TSSs is correlated with higher levels of transcription [57]. The observation that the degree of connectivity of a TSS with nonTSS locations is correlated with higher

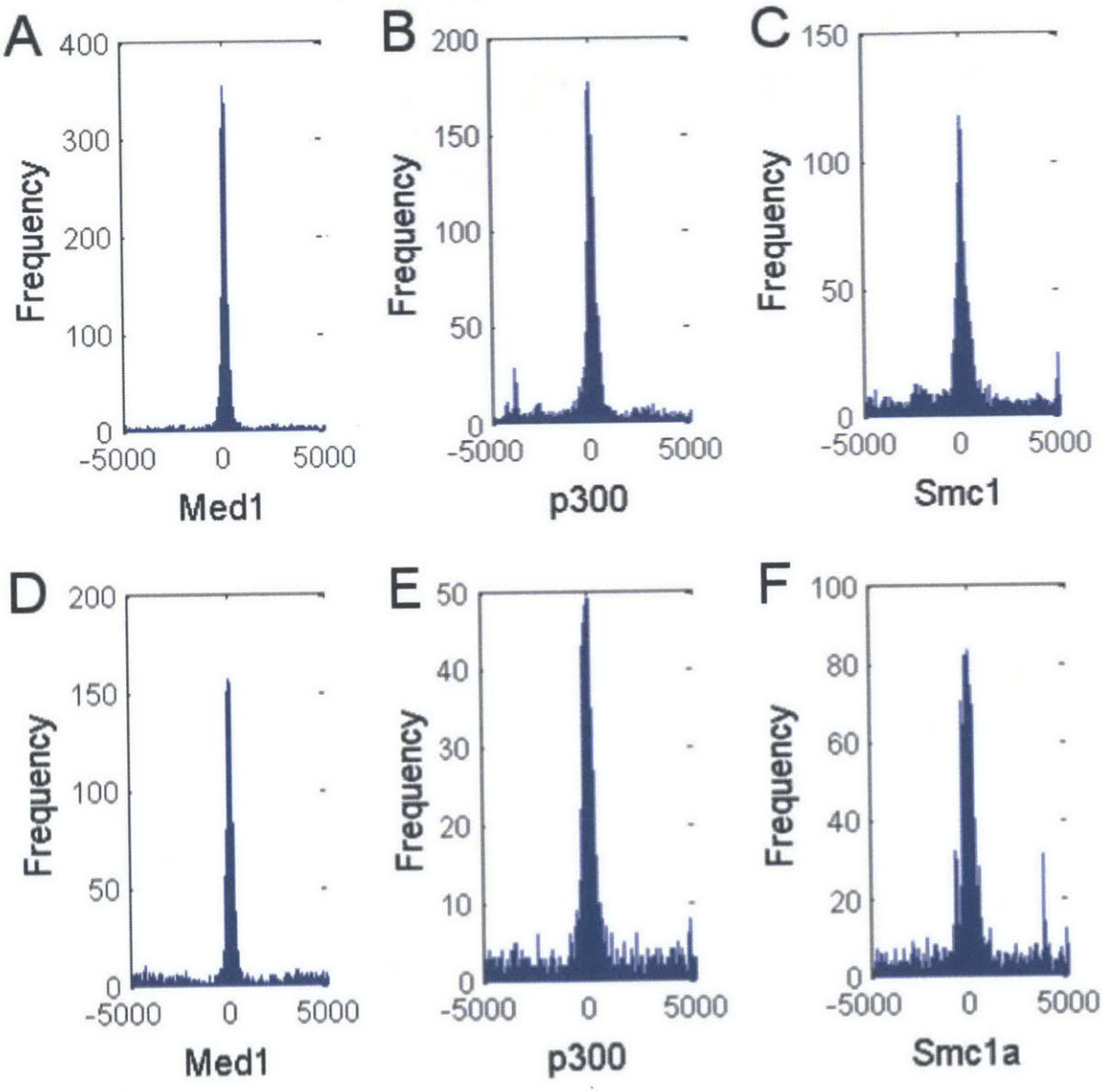


Figure 4-4: **Alignment of ChIP-Seq binding events with GERM^{TSS} identified nonTSS locations that interact with TSSs.** Binding events were identified from ChIP-Seq data using the GEM algorithm. The frequency of binding event locations relative to the nonTSS ends of TSS-nonTSS interactions is shown for (A-C) ES data and (D-F) pMN data.

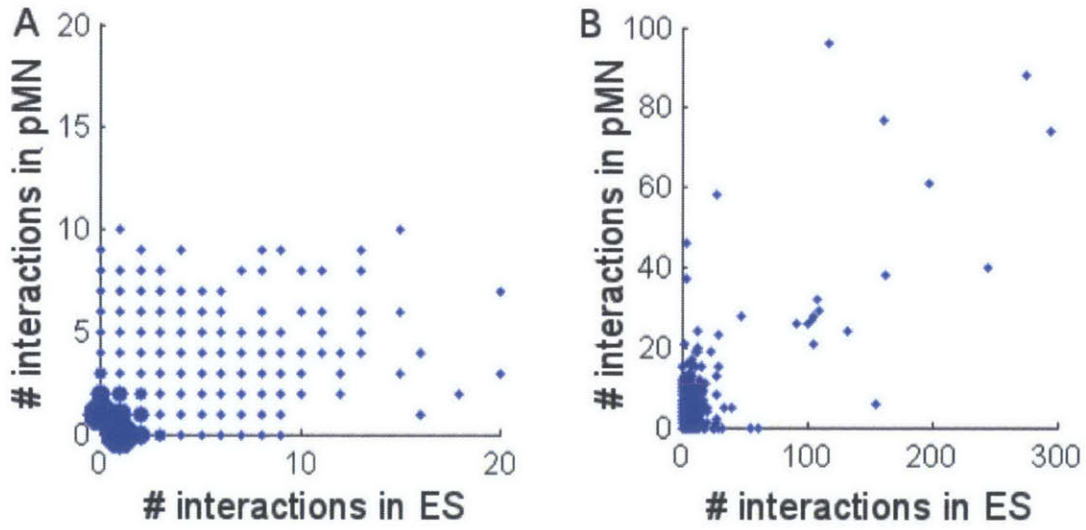


Figure 4-5: **Degrees of connectivity of TSSs and enhancers in ES cells and pMN cells.** (A) TSSs interact with varying numbers of putative enhancers in each cell type. The dot size reflects the frequency of TSSs that interact with the numbers of enhancers in ES cells and in pMN cells denoted by the position of the bubble. (B) Enhancers interact with varying numbers of TSSs in each cell type. The positions of the dots denote combinations of numbers of interactions in ES cells and pMN cells

levels of transcription may imply that the greater degree of enrichment of PolII at the TSSs of highly transcribed genes is related to the greater degree of joint occupation of such TSSs with distal regulatory elements. However, since PolII ChIA-PET reads are more likely to be observed at locations that are more strongly enriched for PolII, it is also possible that interactions that involve a TSS that is less enriched for PolII are more likely to be underrepresented in the data.

By comparing the levels of transcription of genes involved in interactions with putative enhancers to the levels of the 4,321 genes closest to the same set of enhancers, we discovered that chromatin interactions with GERM identified enhancers predict higher levels of transcription than genomic proximity to the same set of enhancers (Figure 4-7A). A greater fraction of the genes most proximal to the GERM identified putative enhancers have transcription levels less than any given threshold than the genes that interact with the putative enhancers (Figure 4-7B). This evidence suggests that GERM identified putative enhancers have a more strongly activating influence on genes that they interact with than the genes that are closest to them.

We also hypothesized that enhancers that engage in greater numbers of interactions with TSSs may exhibit stronger enhancer characteristics. We measured levels of enrichment in 1kb windows centered on the enhancers from both cell types for the H3K27ac, Med1, Med12, p300, and Smc1a ChIP-Seq data (Figure 4-8). All five features are correlated with the degree of connectivity of enhancers in ES cells.

4.3 Enhancer switching and gene switching

Cells at various stages of development are known to utilize dramatically different sets of enhancers as has been demonstrated by observing differential histone modification enrichment at enhancers [59]. Our efforts to detect differential enhancer usage between cell types will be confounded to some degree by the high false negative rate that is inherent to ChIA-PET data. However, since we have shown that it is likely that the sets of TSS-nonTSS interactions present in ES and pMN cells are quite different, we examined the dynamics of enhancer usage between ES and pMN cells keeping in

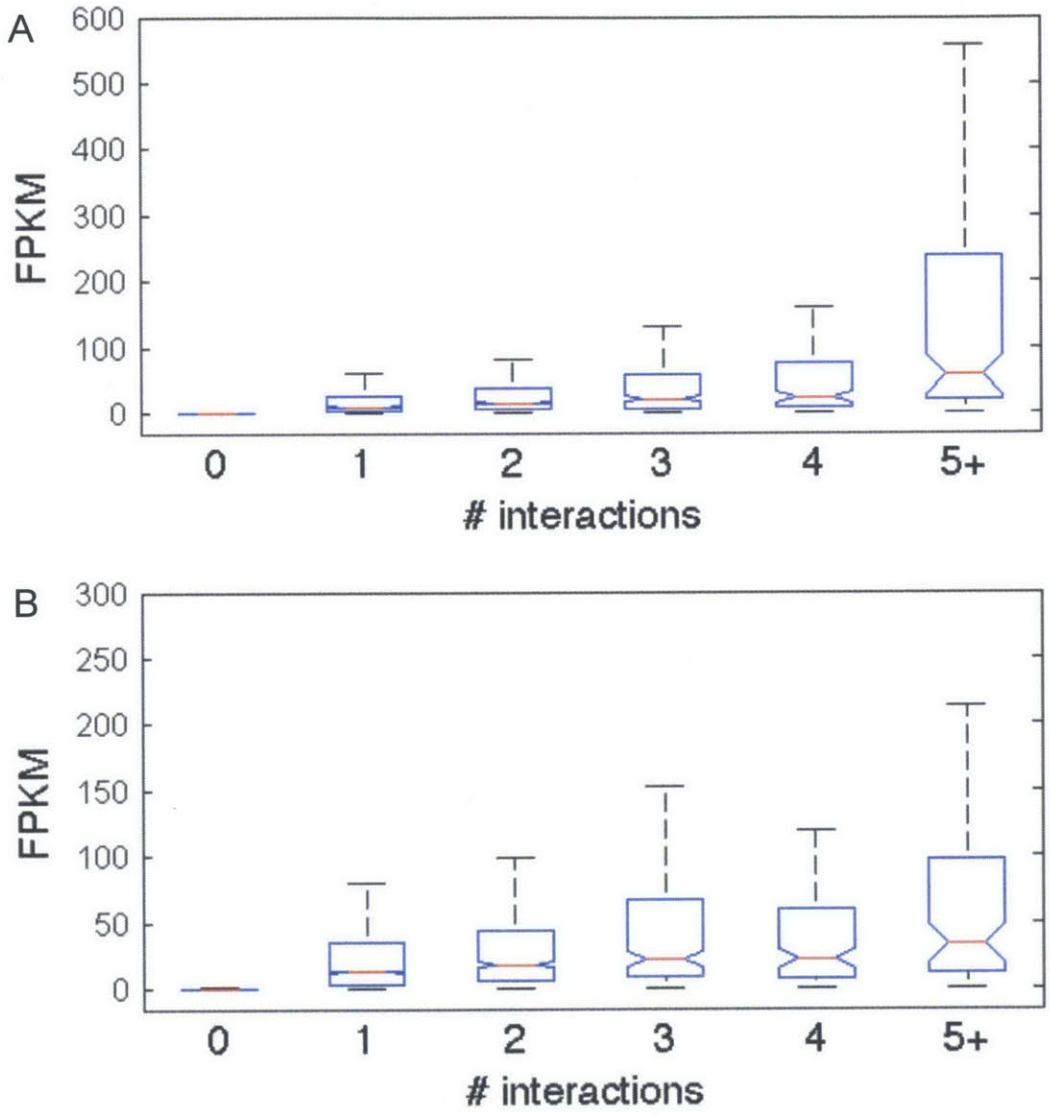


Figure 4-6: **Transcription levels are correlated with the number of nonTSS locations with which a TSS interacts.** Genes are categorized based on the number of nonTSS locations that their TSSs interact with in (A) ES cells and (B) pMNs. The boxplots reflect the distribution of FPKM values computed for the genes in each group from RNA-Seq data.

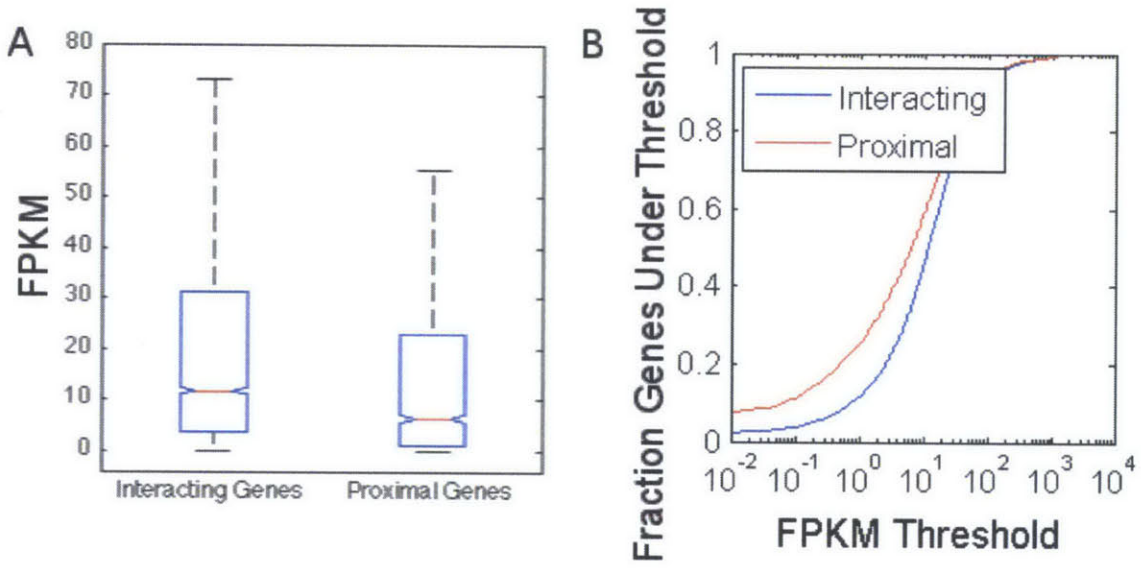


Figure 4-7: **Considering interactions allows more highly transcribed genes to be identified than the set of genes that are closest to the locations that are detected to interact with TSSs.** (A) The set of Interacting Genes is the set of genes for which their TSS is identified by GERM^{TSS} as interacting with at least one nonTSS location. The set of Proximal Genes is the set of genes for which their TSS is the closest TSS to the set of nonTSS locations that are identified by GERM^{TSS} as interacting with at least one TSS. The boxplots reflect the distribution of FPKM values computed for the genes in each group from the ES cell RNA-Seq data. (B) The cumulative distributions of the transcription levels of the two sets of genes in ES cells demonstrate that a greater percentage of the genes proximal to the GERM^{TSS} identified nonTSS locations have transcription levels less than any FPKM threshold than the set of genes that interact with the nonTSS locations.

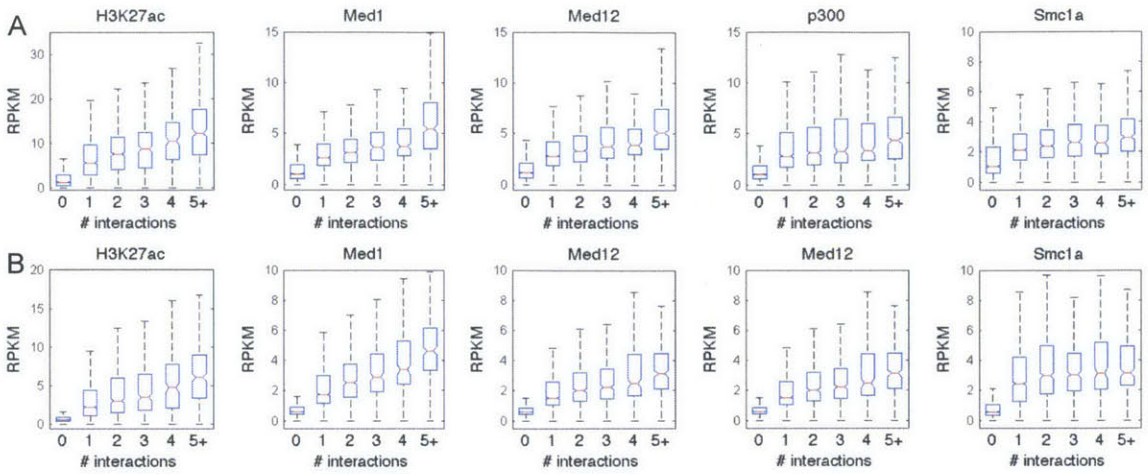


Figure 4-8: **Enrichment for enhancer associated features is correlated with the number of TSSs with which a nonTSS location interacts.** All nonTSS locations that are involved in an interaction with a TSS in at least one of the cell types were considered. The nonTSS locations were categorized based on the number of TSSs that they interact with in (A) ES cells and (B) pMN cells. RPKM values were computed from ChIP-Seq data in 1 kb windows centered on each nonTSS location. The boxplots reflect the distributions of RPKM values for the nonTSS locations in each group for each ChIP-Seq dataset.

mind that many of the interactions that appear to be missing in either cell type may in fact be present. Of the 9,127 putative enhancers that we detect from both cell types, 1,102 (12.1%) of them interact with TSSs in both cell types. Of the 39 highly utilized enhancers that interact with at least 20 TSSs in at least one of the cell types, 31 of them are utilized in both cell types. All 8 of the highly utilized enhancers that are not utilized in both cell types only interact with TSSs in ES cells. This suggests that highly utilized enhancers are established in ES cells and that many of them continue to be utilized during development. Of the 3,554 TSSs that interact with enhancers in both cell types, only 966 (27.2%) TSSs maintain an interaction with the same enhancer in both cell types. Given the caveat about the high false negative rate, differentially active sets of enhancers appear to be not only related to genes that turn on or off during development but also to genes that switch enhancers to maintain or adjust their expression. The differential enhancer usage we observed led us to note that in some cases enhancers also switch the genes with which they interact. Of the

1,102 enhancers that interact with TSSs in both cell types, 661 (60.0%) maintain an interaction with the same TSS. Thus, it appears that many of the enhancers that are active in both cell types do maintain their associations with the same genes but that the genes that an enhancer regulates may not be entirely fixed between cell types.

4.4 Motif analysis of enhancer regions

Of the enhancer related factors that we examined, Med1 was most strongly aligned with the nonTSS ends of TSS-nonTSS interactions. Mediator is an integral component of enhancer-gene interactions that links enhancer-bound transcription factors to the promoter-bound PolII complex. Given the known association between Mediator and enhancer-bound transcription factors we hypothesized that cell type appropriate transcription factor motifs would be present near Med1 binding events that interact with TSSs in one or both cell types. We assigned the 8,867 TSS-nonTSS interactions in ES cells and 4,089 TSS-nonTSS interactions in pMN cells that have a nonTSS anchor that is within 500 bp of a Med1 binding event to the nearest Med1 binding event. This resulted in 4,097 Med1 binding events being associated with at least one TSS-nonTSS interaction. We further grouped together Med1 events from either cell type that are within 500 bp to form 3,481 enhancer units. This resulted in 2,217 Med1$^+$ enhancers that are only interacted with in ES cells, 950 that are only interacted with in pMN cells, and 314 that are interacted with in both cell types.

We discovered the presence of cell-type appropriate motifs within the GERM discovered cell-type specific enhancers (Figure 4-9) from a set of 1,182 PWMs collected from the JASPAR, UniPROBE, and TRANSFAC databases [42, 43, 47]. We looked in 1 kb windows around the most central Med1 binding event within each of the 3,481 enhancer units for motif matches. The stem cell factor Klf4 [37] motif is present in almost half of the ES cell enhancers, and is the most common motif present in these enhancers. Both the Klf4 and Oct4 [49] motifs are present in about twice the percentage of ES specific enhancers as they are in pMN specific and shared enhancers. pMN specific enhancers are enriched for the RXR::RAR [48] motif and many of the Hox

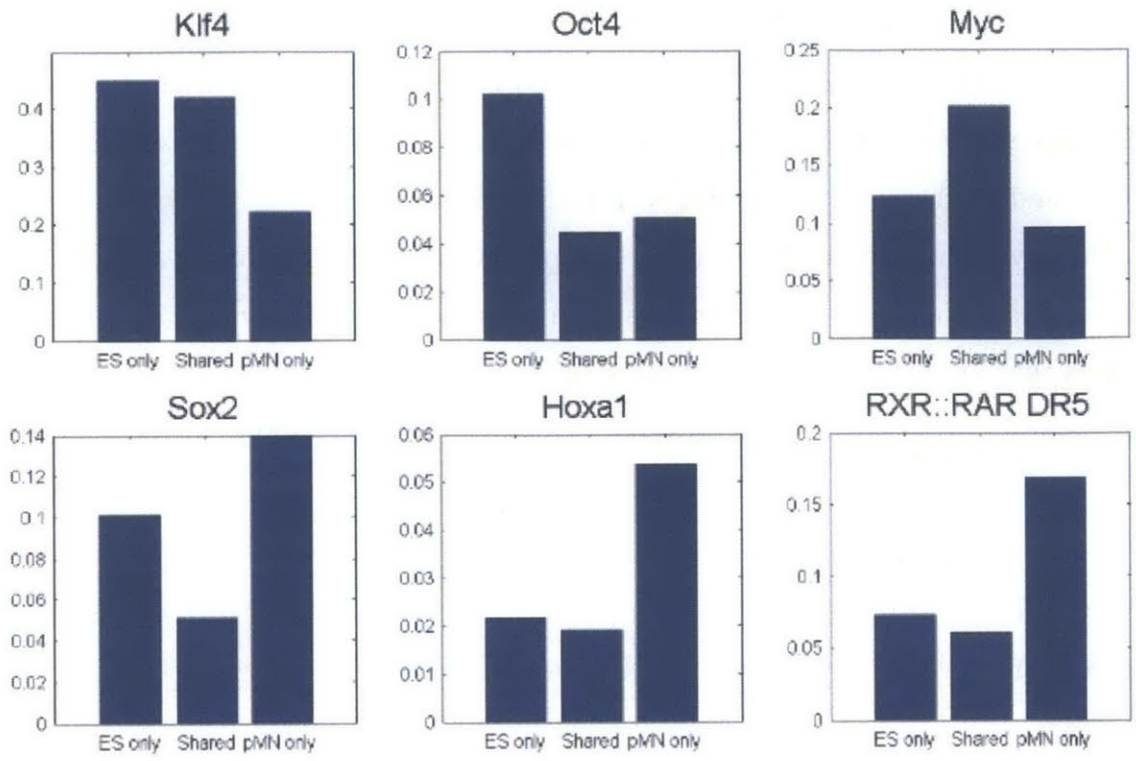


Figure 4-9: **Enhancer usage reflects cell-type appropriate motif enrichment.** 1 kb windows centered on Med1 binding events involved in interactions with TSSs in one or both cell types were scanned for matches to known transcription factor motifs. Med1 binding events were categorized based on whether they interact with TSSs in one or both cell types. The bar graphs reflect the percentages of Med1 binding events in each group that have a motif match within 500 bp for several important transcription factors.

[15] factor motifs compared to ES specific enhancers. Interestingly, the Sox2 [3, 19] motif is at least twice as common in enhancers specific to either cell type as in the shared enhancers. Sox2 is an important transcription factor for both cell types and it may be the case that the two cell types utilize mostly non-overlapping sets of Sox2 binding events to regulate gene expression.

4.5 Discussion

We have demonstrated that applying GERM to ChIA-PET data successfully recovers genomic locations that are enriched for enhancer-related ChIP-Seq data. Their iden-

tity as enhancers is further supported by the observation that the interactions that we identify between these locations and TSSs is correlated with transcription levels. Technologies for profiling chromatin interactions genome-wide such as ChIA-PET, Hi-C, and 5C have yet to reach maturity and present analytical challenges such as inherently high false negative rates. Our observations suggest that gene regulation by long-range chromatin interactions with enhancers is a highly dynamic process. Genes that are expressed in more than one cell type may utilize different enhancers to maintain or adjust their expression. This hypothesis is supported by the observation that differentially utilized enhancers contain varying sets of motifs that are recognized by cell-type appropriate transcription factors. This observation that the relationships between enhancers and genes may be not fixed between cell types has been previously noted [28], although caveats about the high false negative rate inherent to ChIA-PET data have been largely ignored. Theories have been proposed [10, 14, 50, 55] which have begun to characterize the principles underlying regulatory relationships in the genome, yet the logic behind the placement of enhancers relative to the genes that they regulate has yet to be fully elucidated. We hope that the observations about enhancer usage that we have characterized will help guide future studies that address these important questions regarding transcriptional regulation.

Chapter 5

Conclusion

5.1 Summary of results

In this thesis we presented two novel computational methods, SPROUT and GERM, which are both designed to recover information about the manner in which chromatin folds within the nucleus to allow proteins to simultaneously occupy distal genomic locations.

SPROUT is best suited for analyzing ChIA-PET data that profile proteins that bind to the genome in a punctate fashion. SPROUT assumes that protein binding can be accurately modeled by point locations in the genome. The positions of binding events and the existence of interactions between them are estimated by simultaneously considering self-ligation and inter-ligation read pair alignments. Models of the aligned positions of both types of read pairs relative to binding events are utilized to position binding events at high resolution. These models are also used to accurately assign read pairs to binding events in order to avoid overestimating the significance of an interaction from spurious read to binding event assignments. We demonstrate by examining independently derived biological replicate CTCF ChIA-PET datasets that interactions discovered by the ChIA-PET Tool that are not discovered by SPROUT are less uniformly supported by the replicates. This finding suggests that by modeling the distribution of read alignments relative to binding events, SPROUT is able to reduce the false positive interaction calls that are discovered when patterns of read

pair alignment are not modeled in this way.

GERM makes different assumptions from SPROUT about how proteins associate with the genome. While SPROUT assumes that protein binding can be approximated by point locations in the genome, GERM does not make this assumption and estimates genome-wide distributions of protein occupancy. The assumptions that SPROUT makes are appropriate for many proteins, such as most transcription factors, and allow SPROUT to gain statistical power. However, some proteins such as PolII do not appear to bind to distinct point locations in the genome. Rather, PolII is observed to occupy broad regions within which PolII seems to favor specific locations to varying degrees. The detailed distributions of protein occupancy that GERM estimates help preserve a more detailed view of the manner in which a protein associates with the genome. We applied GERM to PolII ChIA-PET data and showed that the sites that GERM detects as interacting with transcription start sites (TSSs) are more enriched for features that are associated with active enhancers than the corresponding sites detected by the ChIA-PET Tool.

We expanded our analysis of PolII ChIA-PET data with GERM to include both data collected from embryonic stem (ES) cells as well as motor neuron progenitors (pMN). This analysis provided the opportunity to examine the dynamics of enhancer usage during neural development. We observed a correlation between the number of interactions that a TSS engages in and the transcription level of the corresponding gene. We also observed a correlation between the number of interactions that a distal location engages in with TSSs and the strength of the enhancer features at the distal location. We presented some caveats about the high false negative rates inherent to current ChIA-PET datasets. Given these caveats, we noted that in many cases, TSSs that are involved in interactions in the two cell types interact with different distal locations. By examining the prevalence of motifs in enhancers that are interacted with in only one of the two cell types, we observed that motifs that are important to a particular cell type are more abundant in the enhancers that are only interacted with in that cell type. We hypothesize that genes use different enhancers in order to maintain their expression in the different transcription factor environments of different

cell types.

5.2 Future work

Technologies that measure chromatin conformation in a high throughput fashion are still very young. There is much work yet to be done to characterize sources of noise or biases that may affect the results that are obtained from these data. This thesis characterized the high false negative rate that is inherent to current ChIA-PET datasets. More work should be done to better understand how confident we can be in the existence of the chromatin interactions that we identify and the nonexistence of the interactions that we do not identify. As more labs are able to successfully perform ChIA-PET experiments, it will be important for experimentalists to collaborate with computationalists in order to maximize the quality of the results that can be obtained. To more fully characterize the set of chromatin interactions that are present in a given cell type, it may be necessary to create more biological replicate datasets or to find ways to increase the complexity of the libraries that are sequenced. A related issue which affects the interpretation of ChIA-PET data is that the stability of chromatin interactions has not been well characterized. ChIA-PET data reflect the conformational state of chromosomes averaged over the millions of cells that were used to perform the experiment. It is unclear which chromatin interactions may occur simultaneously or may be mutually exclusive and how variable the overall conformation of the genome is in a population of cells. Improvements in imaging technology and perhaps single-cell technologies may help provide clarification. At the moment these issues make it difficult to make statements about interactions that are differential between cell types and synergistic effects of interactions within the same cell type.

One of the strengths of the ChIA-PET approach to measuring chromatin interactions is that it utilizes an antibody to detect chromatin interactions that involve locations bound by a particular protein. In theory, this enables different functional types of chromatin interactions to be characterized. However, ChIA-PET datasets

only exist for a handful of proteins in mostly non-overlapping cell types. Performing ChIA-PET experiments for several proteins with different functional roles in the same cell type would provide the opportunity to create more complex models of chromatin conformation. Both SPROUT and GERM could conceivably be extended to analyze combined datasets with read pairs corresponding to different antibodies. If assumptions are made about the consistency of chromatin conformation within separate cell populations of the same cell type, ChIA-PET data from experiments that use different antibodies could be considered together to improve the sensitivity and accuracy of the overall model of chromatin conformation. Labels could then be assigned to the discovered chromatin interactions corresponding to the proteins involved in the interactions. Such a model could improve our understanding of how chromatin conformation contributes to the regulation of gene expression.

Another direction in which ChIA-PET-based studies could be extended is in discovering differential interactions across a greater number of cell types. At present, the high false negative rate inherent to current ChIA-PET datasets would reduce the confidence of any claims made based on differential interaction calls. However, as the ChIA-PET experimental method matures, considering multiple ChIA-PET datasets will provide interesting opportunities for computational modeling. Experiments performed on cells taken from multiple stages along a differentiation pathway or cells that are responding to some sort of environmental stimulus will provide information about the dynamics of chromatin conformation. If the ChIA-PET method is made more efficient, allowing experiments to be performed using fewer cells, experiments performed on samples taken from different tissues within an organism would allow the variability of chromatin conformation within different tissues to be examined. Similarly, experiments performed on different types of cancerous cells would allow the effects of genome rearrangements and other mutations on chromatin conformation to be studied. Datasets with temporal structure or other underlying relationships will provide opportunities for interesting modeling challenges.

Advances in technologies that measure genomic properties in a high throughput fashion have been a boon to functional genomics. Long stretches of mammalian

genomes that were once thought to be of little functional importance are now thought to contain crucially important functional elements. As experimental technologies that utilize next generation sequencing continue to mature, computational methods will continue to be essential for extracting high confidence, interpretable results from large genomics datasets. We hope that the ideas and methods presented in this thesis will be useful towards the ultimate goal of understanding the way the genome functions.

Bibliography

[1] Takanori Amano, Tomoko Sagai, Hideyuki Tanabe, Yoichi Mizushina, Hiromi Nakazawa, and Toshihiko Shiroishi. Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Dev. Cell*, 16(1):47–57, January 2009.

[2] Ido Amit, Manuel Garber, Nicolas Chevrier, Ana Paula Leite, Yoni Donner, Thomas Eisenhaure, Mitchell Guttman, Jennifer K Grenier, Weibo Li, Or Zuk, Lisa A Schubert, Brian Birditt, Tal Shay, Alon Goren, Xiaolan Zhang, Zachary Smith, Raquel Deering, Rebecca C McDonald, Moran Cabili, Bradley E Bernstein, John L Rinn, Alex Meissner, David E Root, Nir Hacohen, and Aviv Regev. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science*, 326(5950):257–63, October 2009.

[3] Ariel A Avilion, Silvia K Nicolis, Larysa H Pevny, Lidia Perez, Nigel Vivian, and Robin Lovell-Badge. Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev.*, 17(1):126–40, January 2003.

[4] G R Ayers and J C Dainty. Iterative blind deconvolution method and its applications. *Opt. Lett.*, 13(7):547–549, 1988.

[5] Ziv Bar-Joseph, Zahava Siegfried, Michael Brandeis, Benedikt Brors, Yong Lu, Roland Eils, Brian D Dynlacht, and Itamar Simon. Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. *Proc. Natl. Acad. Sci. U. S. A.*, 105(3):955–60, January 2008.

[6] Bradley E Bernstein, Ewan Birney, Ian Dunham, Eric D Green, Chris Gunter, and Michael Snyder. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

[7] Iouri Chepelev, Gang Wei, Dara Wangsa, Qingsong Tang, and Keji Zhao. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Res.*, 22(3):490–503, March 2012.

[8] Peter R Cook. The Organization of Replication and Transcription. *Science (80-.).*, 284(5421):1790–1795, June 1999.

[9] Peter R Cook. Predicting three-dimensional genome structure from transcriptional activity. *Nat. Genet.*, 32(3):347–52, November 2002.

[10] Thomas Cremer and Christoph Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.*, 2(4):292–301, April 2001.

[11] Menno P Creyghton, Albert W Cheng, G Grant Welstead, Tristan Kooistra, Bryce W Carey, Eveline J Steine, Jacob Hanna, Michael a Lodato, Garrett M Frampton, Phillip a Sharp, Laurie a Boyer, Richard a Young, and Rudolf Jaenisch. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.*, 107(50):21931–21936, November 2010.

[12] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–11, February 2002.

[13] AP Dempster, NM Laird, and DB Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, 39(1):1–38, 1977.

[14] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, pages 1–5, April 2012.

[15] Monica Ensini, Tammy N Tsuchida, Heinz-Georg Belting, and Thomas M Jessell. The control of rostrocaudal pattern in the developing spinal cord: specification of motor neuron subtype identity is initiated by signals from paraxial mesoderm. *Development*, 125(6):969–82, March 1998.

[16] Mario A T Figueiredo and Anil K Jain. Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(3), May 2002.

[17] D A Fish, A M Brinicombe, E R Pike, and J G Walker. Blind deconvolution by means of the Richardson-Lucy algorithm. *J. Opt. Soc. Am. A*, 12(1):58, January 1995.

[18] Melissa J Fullwood, Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yusoff Bin Mohamed, Yuriy L Orlov, Stoyan Velkov, Andrea Ho, Poh Huay Mei, Elaine G Y Chew, Phillips Yao Hui Huang, Willem-Jan Welboren, Yuyuan Han, Hong Sain Ooi, Pramila N Ariyaratne, Vinsensius B Vega, Yanquan Luo, Peck Yean Tan, Pei Ye Choy, K D Senali Abayratna Wansa, Bing Zhao, Kar Sian Lim, Shi Chi Leow, Jit Sin Yow, Roy Joseph, Haixia Li, Kartiki V Desai, Jane S Thomsen, Yew Kok Lee, R Krishna Murthy Karuturi, Thoreau Herve, Guillaume Bourque, Hendrik G Stunnenberg, Xiaoan Ruan, Valere Cacheux-Rataboul, Wing-Kin Sung, Edison T Liu, Chia-Lin Wei, Edwin Cheung, and Yijun Ruan. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462(7269):58–64, November 2009.

[19] Victoria Graham, Jane Khudyakov, Pamela Ellis, and Larysa Pevny. SOX2 functions to maintain neural progenitor identity. *Neuron*, 39(5):749–65, August 2003.

[20] Yuchun Guo, Shaun Mahony, and David K Gifford. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.*, 8(8):e1002638, January 2012.

[21] Yuchun Guo, Georgios Papachristoudis, Robert C Altshuler, Georg K Gerber, Tommi S Jaakkola, David K Gifford, and Shaun Mahony. Discovering homotypic binding events at high spatial resolution. *Bioinformatics*, 26(24):3028–34, December 2010.

[22] Lusy Handoko, Han Xu, Guoliang Li, Chew Yee Ngan, Elaine G Y Chew, Marie Schnapp, Charlie W H Lee, Chaopeng Ye, Joanne Lim Hui Ping, Fabianus Mulawadi, Eleanor Wong, Jianpeng Sheng, Yubo Zhang, Thompson Poh, Chee Seng Chan, Galih Kunarso, Atif Shahab, Guillaume Bourque, Valere Cacheux-Rataboul, Wing-Kin Sung, Yijun Ruan, and Chia-Lin Wei. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.*, 43(7):630–638, 2011.

[23] Hongkai Ji, Hui Jiang, Wenxiu Ma, David S Johnson, Richard M Myers, and Wing H Wong. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, 26(11):1293–300, November 2008.

[24] Ming Jiang and Ge Wang. Development of blind image deconvolution and its applications. *J. Xray. Sci. Technol.*, 11(1):13–9, January 2003.

[25] Raja Jothi, Suresh Cuddapah, Artem Barski, Kairong Cui, and Keji Zhao. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, 36(16):5221–31, September 2008.

[26] Michael H Kagey, Jamie J Newman, Steve Bilodeau, Ye Zhan, David A Orlando, Nynke L van Berkum, Christopher C Ebmeier, Jesse Goossens, Peter B Rahl, Stuart S Levine, Dylan J Taatjes, Job Dekker, and Richard A Young. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430–5, September 2010.

[27] Donna Karolchik, Galt P Barber, Jonathan Casper, Hiram Clawson, Melissa S Cline, Mark Diekhans, Timothy R Dreszer, Pauline A Fujita, Luvina Guruvadoo, Maximilian Haeussler, Rachel A Harte, Steve Heitner, Angie S Hinrichs, Katrina Learned, Brian T Lee, Chin H Li, Brian J Raney, Brooke Rhead, Kate R Rosenbloom, Cricket A Sloan, Matthew L Speir, Ann S Zweig, David Haussler, Robert M Kuhn, and W James Kent. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.*, 42(1):D764–70, January 2014.

[28] Kyong-Rim Kieffer-Kwon, Zhonghui Tang, Ewy Mathe, Jason Qian, Myong-Hee Sung, Guoliang Li, Wolfgang Resch, Songjoon Baek, Nathanael Pruett, Lars

Grontved, Laura Vian, Steevenson Nelson, Hossein Zare, Ofir Hakim, Deepak Reyon, Arito Yamane, Hirotaka Nakahashi, Alexander L Kovalchuk, Jizhong Zou, J Keith Joung, Vittorio Sartorelli, Chia-Lin Wei, Xiaoan Ruan, Gordon L Hager, Yijun Ruan, and Rafael Casellas. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell*, 155(7):1507–20, December 2013.

[29] Tony Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, February 2007.

[30] Deepa Kundur and Dimitrios Hatzinakos. Blind Image Deconvolution. *IEEE Signal Process. Mag.*, pages 43–64, May 1996.

[31] Deepa Kundur and Dimitrios Hatzinakos. Blind Image Deconvolution Revisited. *IEEE Signal Process. Mag.*, (NOVEMBER):61–63, November 1996.

[32] Deepa Kundur and Dimitrios Hatzinakos. A Novel Blind Deconvolution Scheme for Image Restoration Using Recursive Filtering. *IEEE Trans. Signal Process.*, 46(2):375–390, 1998.

[33] Christian Lanctôt, Thierry Cheutin, Marion Cremer, Giacomo Cavalli, and Thomas Cremer. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat. Rev. Genet.*, 8(2):104–15, February 2007.

[34] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3):R25, January 2009.

[35] Boris Lenhard, Albin Sandelin, and Piero Carninci. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.*, 13(4):233–45, April 2012.

[36] Guoliang Li, Melissa J Fullwood, Han Xu, Fabianus Hendriyan Mulawadi, Stoyan Velkov, Vinsensius Vega, Pramila Nuwantha Ariyaratne, Yusoff Bin Mohamed, Hong-Sain Ooi, Chandana Tennakoon, Chia-Lin Wei, Yijun Ruan, and Wing-Kin Sung. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol.*, 11(2):R22, January 2010.

[37] Yanjun Li, Jeanette McClintick, Li Zhong, Howard J Edenberg, Mervin C Yoder, and Rebecca J Chan. Murine embryonic stem cell differentiation is promoted by SOCS-3 and inhibited by the zinc finger transcription factor Klf4. *Blood*, 105(2):635–7, January 2005.

[38] Rong Lu, Florian Markowetz, Richard D Unwin, Jeffrey T Leek, Edoardo M Airoldi, Ben D MacArthur, Alexander Lachmann, Roye Rozov, Avi Ma'ayan, Laurie a Boyer, Olga G Troyanskaya, Anthony D Whetton, and Ihor R Lemischka. Systems-level dynamic analyses of fate change in murine embryonic stem cells. *Nature*, 462(7271):358–62, 2009.

[39] LB Lucy. An Iterative Technique for the Rectification of Observed Distributions. *Astron. J.*, 79(6):745–754, 1974.

[40] Desmond S Lun, Ashley Sherrid, Brian Weiner, David R Sherman, and James E Galagan. A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from ChIP-seq data. *Genome Biol.*, 10(12):R142, January 2009.

[41] Shaun Mahony, Matthew D Edwards, Esteban O Mazzoni, Richard I Sherwood, Akshay Kakumanu, Carolyn A Morrison, Hynek Wichterle, and David K Gifford. An Integrated Model of Multiple-Condition ChIP-Seq Data Reveals Predeterminants of Cdx2 Binding. *PLoS Comput. Biol.*, 10(3):e1003501, March 2014.

[42] Anthony Mathelier, Xiaobei Zhao, Allen W Zhang, François Parcy, Rebecca Worsley-Hunt, David J Arenillas, Sorana Buchman, Chih-Yu Chen, Alice Chou, Hans Ienasescu, Jonathan Lim, Casper Shyr, Ge Tan, Michelle Zhou, Boris Lenhard, Albin Sandelin, and Wyeth W Wasserman. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 42(1):D142–7, January 2014.

[43] V Matys, O V Kel-Margoulis, E Fricke, I Liebich, S Land, A Barre-Dirrie, I Reuter, D Chekmenev, M Krull, K Hornischer, N Voss, P Stegmaier, B Lewicki-Potapov, H Saxel, a E Kel, and E Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, 34(Database issue):D108–10, January 2006.

[44] BC McCallum. Blind Deconvolution by Simulated Annealing. *Opt. Commun.*, 75(2):101–105, 1990.

[45] X Meng and D B Rubin. Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*, 80:267–278, 1993.

[46] Tom Misteli. Self-organization in the genome. *Proc. Natl. Acad. Sci. U. S. A.*, 106(17):6885–6, April 2009.

[47] Daniel E Newburger and Martha L Bulyk. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, 37(Database issue):D77–82, January 2009.

[48] Karen Niederreither and Pascal Dollé. Retinoic acid in development: towards an integrated view. *Nat. Rev. Genet.*, 9(7):541–53, July 2008.

[49] Hitoshi Niwa, Jun-ichi Miyazaki, and Austin G Smith. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat. Genet.*, 24(4):372–6, April 2000.

[50] Elphège P Nora, Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L van Berkum, Johannes

Meisig, John Sedat, Joost Gribnau, Emmanuel Barillot, Nils Blüthgen, Job Dekker, and Edith Heard. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381–5, May 2012.

[51] Jennifer E Phillips and Victor G Corces. CTCF: master weaver of the genome. *Cell*, 137(7):1194–211, 2009.

[52] William Hadley Richardson. Bayesian-Based Iterative Method of Image Restoration. *J. Opt. Soc. Am.*, 62(1):55, January 1972.

[53] Joel Rozowsky, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, 27(1):66–75, 2009.

[54] M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.*, 9:65–78, 1982.

[55] Tom Sexton, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, 148(3):458–72, February 2012.

[56] Radha Shyamsundar, Young H Kim, John P Higgins, Kelli Montgomery, Michelle Jorden, Anand Sethuraman, Matt van de Rijn, David Botstein, Patrick O Brown, and Jonathan R Pollack. A DNA microarray survey of gene expression in normal human tissues. *Genome Biol.*, 6(3):R22, January 2005.

[57] Hao Sun, Jiejun Wu, Priyankara Wickramasinghe, Sharmistha Pal, Ravi Gupta, Anirban Bhattacharyya, Francisco J Agosto-Perez, Louise C Showe, Tim H-M Huang, and Ramana V Davuluri. Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq. *Nucleic Acids Res.*, 39(1):190–201, January 2011.

[58] Mariliis Tark-Dame, Roel van Driel, and Dieter W Heermann. Chromatin folding–from biology to polymer models and back. *J. Cell Sci.*, 124(Pt 6):839–45, March 2011.

[59] Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, Kavita Garg, Sam John, Richard Sandstrom, Daniel Bates, Lisa Boatman, Theresa K Canfield, Morgan Diegel, Douglas Dunn, Abigail K Ebersol, Tristan Frum, Erika Giste, Audra K Johnson, Ericka M Johnson, Tanya Kutyavin, Bryan Lajoie, Bum-Kyu Lee, Kristen Lee, Darin London, Dimitra Lotakis, Shane Neph, Fidencio Neri, Eric D Nguyen, Hongzhu Qu, Alex P Reynolds, Vaughn Roach, Alexias Safi, Minerva E Sanchez, Amartya Sanyal, Anthony Shafer, Jeremy M Simon, Lingyun Song, Shinny Vong, Molly Weaver, Yongqi Yan, Zhancheng Zhang, Zhuzhu Zhang, Boris Lenhard, Muneesh Tewari,

Michael O Dorschner, R Scott Hansen, Patrick a Navas, George Stamatoyannopoulos, Vishwanath R Iyer, Jason D Lieb, Shamil R Sunyaev, Joshua M Akey, Peter J Sabo, Rajinder Kaul, Terrence S Furey, Job Dekker, Gregory E Crawford, and John a Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, September 2012.

[60] Anton Valouev, David S Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M Myers, and Arend Sidow. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, 5(9):829–834, 2008.

[61] Axel Visel, Matthew J Blow, Zirong Li, Tao Zhang, Jennifer A Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, Feng Chen, Veena Afzal, Bing Ren, Edward M Rubin, and Len A Pennacchio. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–8, February 2009.

[62] Warren A Whyte, David A Orlando, Denes Hnisz, Brian J Abraham, Charles Y Lin, Michael H Kagey, Peter B Rahl, Tong I Lee, and Richard A Young. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, pages 307–319, 2013.

[63] Hynek Wichterle, Ivo Lieberam, Jeffery A Porter, and Thomas M Jessell. Directed differentiation of embryonic stem cells into motor neurons. *Cell*, 110(3):385 97, August 2002.

[64] Christopher L Woodcock and Stefan Dimitrov. Higher-order structure of chromatin and chromosomes. *Curr. Opin. Genet. Dev.*, 11(2):130–5, April 2001.

[65] Joanna Wysocka, Tomek Swigut, Thomas A Milne, Yali Dou, Xin Zhang, Alma L Burlingame, Robert G Roeder, Ali H Brivanlou, and C David Allis. WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. *Cell*, 121(6):859–72, June 2005.

[66] Joanna Wysocka, Tomek Swigut, Hua Xiao, Thomas A Milne, So Yeon Kwon, Joe Landry, Monika Kauer, Alan J Tackett, Brian T Chait, Paul Badenhorst, Carl Wu, and C David Allis. A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature*, 442(7098):86–90, July 2006.

[67] Yong Zhang, Tao Liu, Clifford a Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9(9):R137, January 2008.

[68] Yubo Zhang, Chee-Hong Wong, Ramon Y Birnbaum, Guoliang Li, Rebecca Favaro, Chew Yee Ngan, Joanne Lim, Eunice Tai, Huay Mei Poh, Eleanor Wong, Fabianus Hendriyan Mulawadi, Wing-Kin Sung, Silvia Nicolis, Nadav Ahituv, Yijun Ruan, and Chia-Lin Wei. Chromatin connectivity maps reveal dynamic

promoter-enhancer long-range associations. *Nature*, 504(7479):306–10, December 2013.