THE ROLE OF TOPOLOGICALLY ASSOCIATING DOMAINS IN ENHANCER-PROMOTER INTERACTIONS

JOSHUA M. PRICE

A Thesis
submitted in partial fulfillment of the degree of
Bachelor of Arts
with Department Honors in Molecular and Cell Biology

University of California, Berkeley
2018

Table of Contents

Abstract

Enhancer-promoter interactions (EPIs) are highly regulated molecular events between specific DNA loci that are necessary for transcription to occur in mammalian cells. The physical packing of interphase DNA into topologically-associating domains (TADs) plays a key role in determining enhancer-promoter specificity as 80 to 90% of EPIs take place within the same TAD. However, the remaining 10 to 20% of EPIs cross TAD boundaries, overcoming the insulation mechanisms experienced by most EPIs to achieve long-range and consistent interaction. In this project we used a genome-wide bioinformatic approach to quantify differences between intra-TAD and boundary-crossing inter-TAD EPIs. By mapping genomic regions enriched in transcriptionally relevant histone modifications and transcription factors as determined by ChIP-seq onto a non-random sample of EPIs taken from a ultra-high-resolution contact matrix of mouse ES E14 cells, we discover that enhancers and promoters in our sample have similar enrichment frequencies for both intra-TAD and inter-TAD EPIs. However, the scope of our conclusions about the sample are limited due to outstanding challenges in defining an unbiased global set of enhancers, an issue that we explore in depth. We conclude by considering various approaches to improving global EPI analyses and offer our approach for further research with alternatively defined EPI sets.

Background and Introduction

Transcription Regulation in Three-Dimensional Space

Transcription regulation, as is the case for all biochemical processes, is mediated by Ångstrom-scale interactions and transfers of atoms and electrons between the participating biomolecules. As an inherent result, the spatial organization of transcription-relevant biomolecules within the nucleus at any given time plays a consequential role in transcription regulation (Cremer and Cremer, 2001). Categories of biomolecules identified as essential factors that interact with interphase DNA in transcription regulation are listed and described in Table 1 (Lee and Young, 2000). Mutational or regulatory changes to these biomolecules and their chemical interactions form the basis of many developmental diseases and cancers and a significant portion of molecular biology research is focused on understanding their functions and relationships (Lee and Young, 2013).

Recent research in the field of transcription regulation has shown that the three-dimensional structure and organization of DNA itself also plays a central role in determining transcriptional patterns (Gonzales-Sandoval et al, 2016). Some structural effects are intuitive considering spatial constraints, such as how genes in tightly packed regions of DNA are transcribed at much lower levels than those in openly accessible regions of DNA due to the differential ability of transcription factors to interact with those chromatin regions (Grewal and Moazed, 2003). Another important spatial consideration is the capacity for cis-regulatory elements to physically interact with each other and perform their chemical functions. Simulations of DNA folding using polymer models as well as physical experiments assessing spatial

distribution of chromatin within the nucleus have shown that DNAs in interphase cells fold not randomly but in a highly organized structure that is conserved across evolutionarily domains (Giorgetti et al, 2014; Franke and Gómez-Skarmeta, 2018). This structural conservation suggests that not only genes but also the regulatory mechanisms of transcription are conserved across related species.

Table 1: Categories of biomolecules that interact with DNA to mediate transcription regulation. This list is non-exhaustive and only contains proteins; other biomolecules such as microRNAs also play a significant role in the regulation of gene expression (He and Hannon, 2004).

Biomolecule Category	**Role in Transcription Regulation**	**Example(s)**
RNA Polymerase	Synthesizes mRNA transcripts from NTPs complementary to DNA template.	RNA Pol II
General Transcription Factors	Bind the upstream promoter of all genes with RNA polymerase to form the transcription preinitiation complex required to begin transcription of a given gene.	TFIIA, TFIIB, Mediator
Regulatory Transcription Factors	Selectively bind upstream gene promoters to positively or negatively regulate the transcription of the corresponding gene.	Oct4, Nanog, Sox2
Histones	Strongly bind DNA and organize it into nucleosomes.	H1, H2A, H2B
Chromatin Remodeling Factors	Modify the structure or identity of chromatin by adding (or removing) functional groups to (or from) histones.	Histone acetyltransferases, deacetylates, and methyltransferases
Repressors	Bind DNA to inhibit expression of one or a group of nearby genes.	CTCF

A typical transcriptional event in mammalian cells only takes place when the transcription preinitiation complex, which includes approximately 100 proteins, assembles on the upstream promoter site of a gene (Poss et al, 2013). Cis-regulatory elements called enhancers which can be proximal or distal are required to recruit the preinitiation complex, making enhancer-promoter interaction (EPI) a necessary step before transcription initiation can take place (Tjian and Maniatis, 1994). Individual enhancer modules interact with one or a small number of gene promoters, suggesting EPIs have features leading to specificity and are not promiscuous (Whalen et al, 2016).

Enhancers share common sequence and histone modification characters (Müller et al, 1988). For example, specific histone modifications such as acetylation of histone subunits are strongly correlated with enhancer identity while histone methylation inversely correlates with enhancer activity (Bannister and Kouzarides, 2011). It has been shown that DNA and histone modifications affect the ability of enhancers and other DNA elements such as promoters to bind the transcription factors necessary to form the transcription preinitiation complex on the upstream promoter a given gene (Spitz and Furlong, 2012). As a result, the capacity for an enhancer to perform its active function depends on both the spatial distribution and chemical modifications of DNA and the other biomolecules relevant to transcription.

Current models of spatial transcription regulation stemmed from discoveries made possible by the numerous high-throughput assays developed in the past decade, including next-generation sequencing (NGS). Two techniques particularly relevant to studying the 3D interactions of interphase DNA are chromatin immunoprecipitation and sequencing and Hi-C, both of which are utilized in this work.

The ChIP-Seq Assay

Much of our modern understanding of how chromatin-associating proteins interact with DNA is the result of an assay called chromatin immunoprecipitation and sequencing (ChIP-Seq), the steps of which are outlined in Figure 1. ChIP-seq allows for the isolation of only the DNA regions that interact with a specific protein of interest, such as the transcription-relevant proteins listed in Table 1 (Park, 2009). As ChIP-seq is conducted across a population of cloned cells, the sequence reads can be analyzed to determine the frequency of interaction with the protein of interest as a function of position on the linear genome. This frequency distribution can be used to infer whether the protein binds genetic elements such as enhancers or promoters and which proteins form complexes along DNA (Visel, 2009). By combining ChIP-seq results with data generated in other assays such as RNA-seq, the frequency distribution of protein binding can be used to infer whether that factor upregulates or downregulates the transcription of certain genes (Ouyang et al, 2009). ChIP-seq has also led to the identification of proteins strongly correlated with regulatory DNA elements such as enhancers and promoters, a subset of which is shown in Table 2 (Visel, 2009).

While the distribution of the binding frequency of protein of interest along the linear genome is valuable information in many molecular biology experiments, ChIP-seq data does not provide information about the temporal nature of the protein-DNA interactions of interest. Many interactions in transcription, such as those between non-general transcription factors, are transient but nonetheless necessary molecular events (Woringer and Darzacq, 2018). Naive interpretations of ChIP-seq data can result in transient interactions being underconsidered or

overlooked altogether, so ChIP-seq data analysts must be vigilant in data processing and consider the natures of the molecular events of interest when generating enrichment sets.

Table 2: Cis-regulatory DNA elements, their functions, and proteins determined to commonly bind with them by ChIP-seq.

Cis-Regulatory DNA Element	**Function**	**Binding Proteins Identified by ChIP-seq**	**Histone Modifications Identified by ChIP-seq**
Promoters	Serve as binding site for RNA Pol II upstream of a gene.	RNA Pol II, Transcription Factors	H3K27ac, H3K4me3, H3K4me2, and H3K9ac (activation); H3K27me3 and H3K9me3 (repression)
Enhancers	Recruits transcription factors to corresponding promoters to facilitate positive or negative regulation of the corresponding gene.	RNA Pol II, Transcription Factors	H3K4me1 and H3K27ac
Insulators	Binding site for repressors; prevents interaction between cis-regulatory elements on opposite sides of the site.	CTCF	None Identified

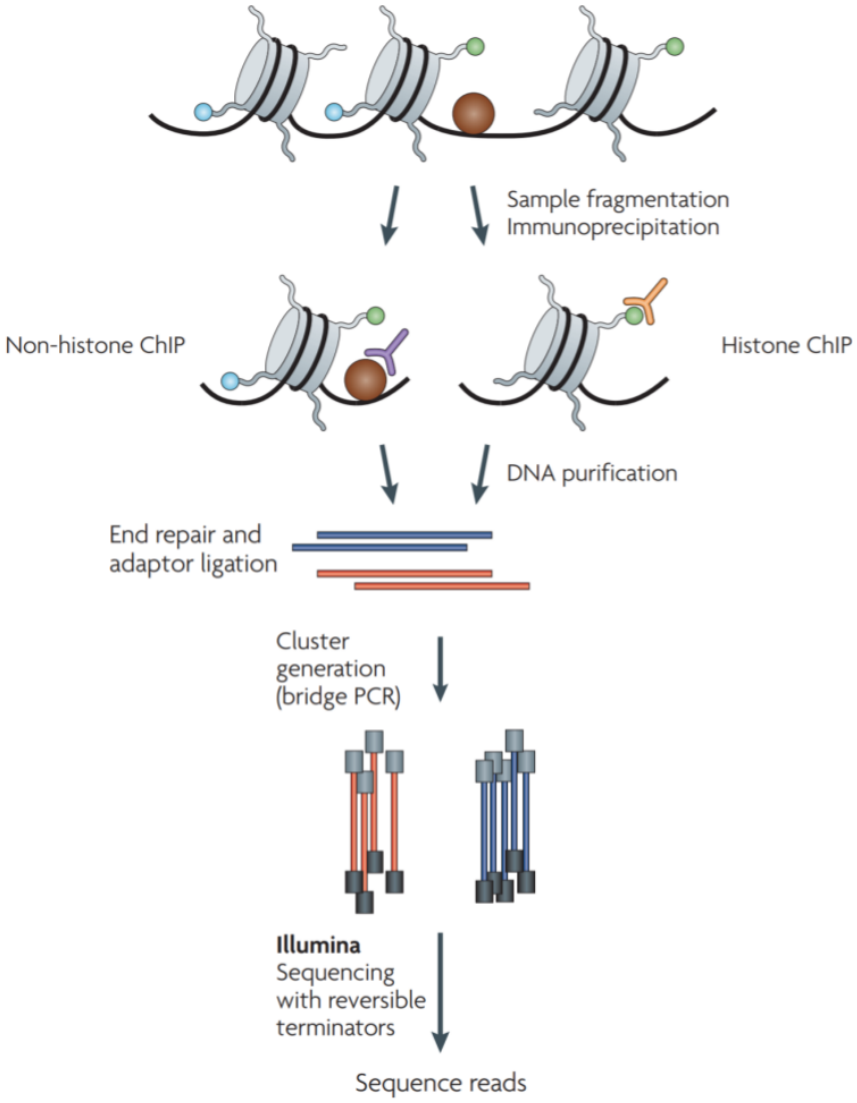

Figure 1: The two major subparts of ChIP-Seq are (1) the isolation of specific regions of DNA by chromatin immunoprecipitation and (2) massively parallel sequencing using Illumina or another NGS method. Both DNA-binding proteins ("non-histone ChIP") and histones with specific modifications ("Histone ChIP") can be targeted by chromatin immunoprecipitation in ChIP. Figure is derived from (Park, 2009).

Despite the widespread adoption of the ChIP-seq assay in modern molecular biology laboratories, methodological issues about how to analyze the data exist. There is no universal approach to peak calling, which is the process used to determine where the proteins bind from raw NGS reads. As a result, different researchers can generate different processed data and conclusions from the same raw dataset, resulting in issues in cross-experimental consistency. This issue of inconsistent peak calling is a general issue in NGS-based approaches.

Though ChIP-based methods allow for the analysis of protein-DNA interactions in transcription-relevant regions such as enhancers and promoters, they provide limited insight into the three-dimensional interactions between cis-regulatory elements in DNA. In light of this limitation, another set of techniques called chromatin conformation capture (3C) methods has emerged to quantify DNA-DNA interactions (Miele et al, 2006).

The Hi-C Assay and Topologically Associating Domains

The most popular of the 3C methods, termed Hi-C, is outlined in Figure 2. Hi-C measures the frequency at which different elements of DNA interact across the genome, which facilitates EPI identification. Several Hi-C derived techniques have been developed such as the micro-C technique used in this work, which uses a different technique to fragment DNA that allows for nucleosome resolution contact maps (Hsieh et al, 2015). The resolution of the micro-C dataset used in this work is 200 base pairs, meaning both long- and short-range interactions can be detected and the specific interacting elements can be annotated for subsequent analysis (Rao et al, 2014).

Hi-C contact matrices showing the frequency of contacts between loci pairs in the genome have revealed several previously unknown properties of 3D DNA interactions. It is immediately obvious when analyzing Hi-C contact matrices that regions of high interaction with clearly defined boundaries exist, as shown in Figure 2 (Rao et al, 2014). These regions have been termed topologically-associating domains (TADs) and have become a widespread paradigm in the field of transcription regulation (Dixon et al, 2012).

Though it may be straightforward to visually identify regions of high DNA-DNA interaction in a Hi-C matrix, defining TADs can be difficult because regions of high contact frequencies exist at every scale. One region of high interaction that appears to be a TAD can be revealed to contain several sub-regions of extra-high interaction within itself. Regions of high interactions within TADs are often called sub-TADs while groups of proximal TADs that show increased interaction are sometimes called super-TADs (Mehra and Kalani, 2018). Many studies have used insulation scores to define TAD boundaries, making TAD boundary definition a function of interaction frequency across the boundary (Narenda et al, 2016). To date, no universally accepted TAD-calling technique has been developed and TAD definitions are inconsistent across studies. As defined in this work, the mouse embryonic stem cell (mESC) genome has about 4500 TADs averaging 622 kbp and ranging from 60 kbp to 2.03 Mbp. A representative set of neighboring TADs is shown in Figure 3.

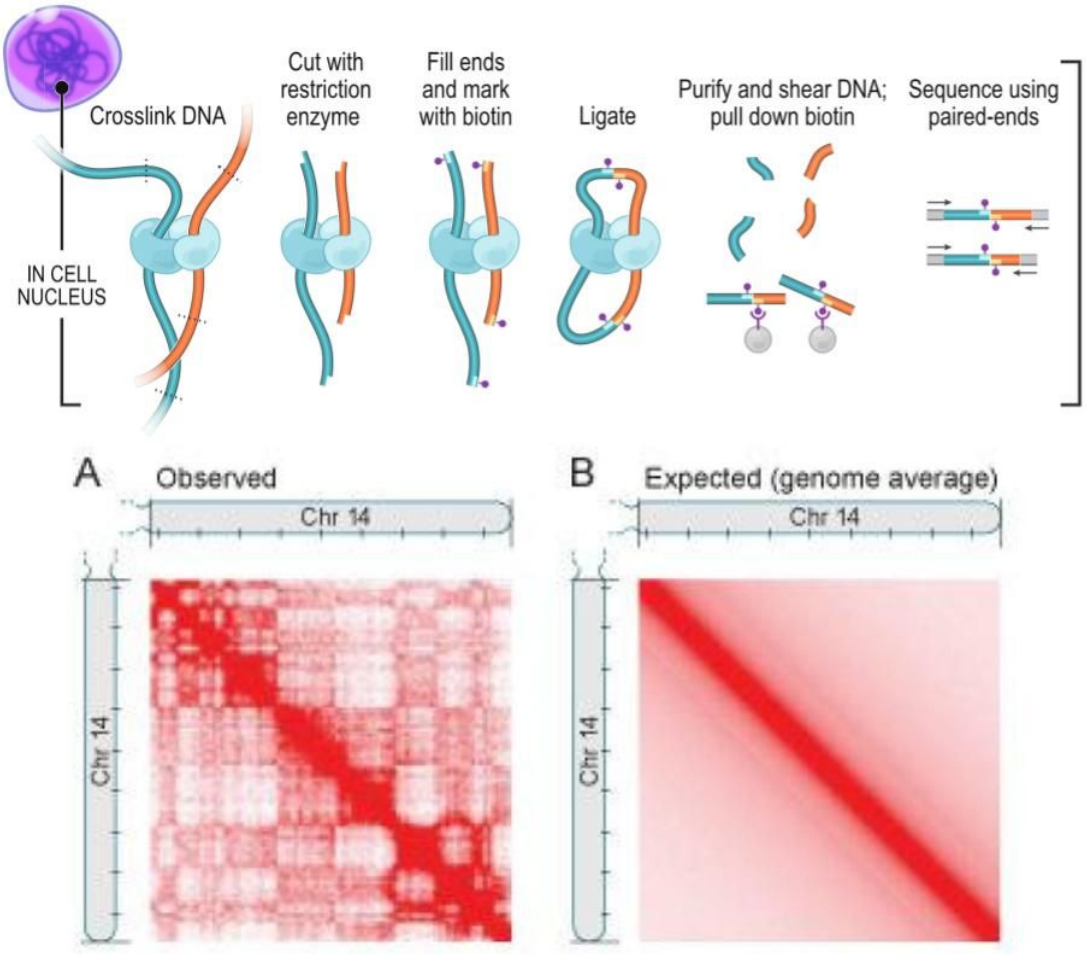

Figure 2: (Top) The Hi-C procedure involves cross-linking and cutting DNA, ligating physically co-located regions, then sequencing ligated loci pairs using next-generation sequencing. (Bottom) The paired-end reads can be analyzed and visualized to show a contact matrix, at left. The theoretical contact matrix at right represents the expected matrix if interactions in the genome were randomly distributed. Figure is derived from (van Berkum et al, 2010).

The borders of TADs, as well as those of sub-TADs and super-TADs, are enriched with repressors and show especially high abundances of CTCF (Ong and Corces, 2014). DNA-DNA

interactions of all types, including EPIs, are much higher within the same TAD (intra-TAD) than in different TADs (inter-TAD) and promoters within the same TAD have been shown to be regulated by the same enhancers across related species (Symmons et al, 2014). This suggests that proteins that give rise to TAD boundaries block inter-TAD enhancer-promoter interactions, which has been confirmed in experiments where TAD boundaries are removed and new EPIs between enhancers and promoters in the merged TAD develop (Lupiáñez et al, 2016).

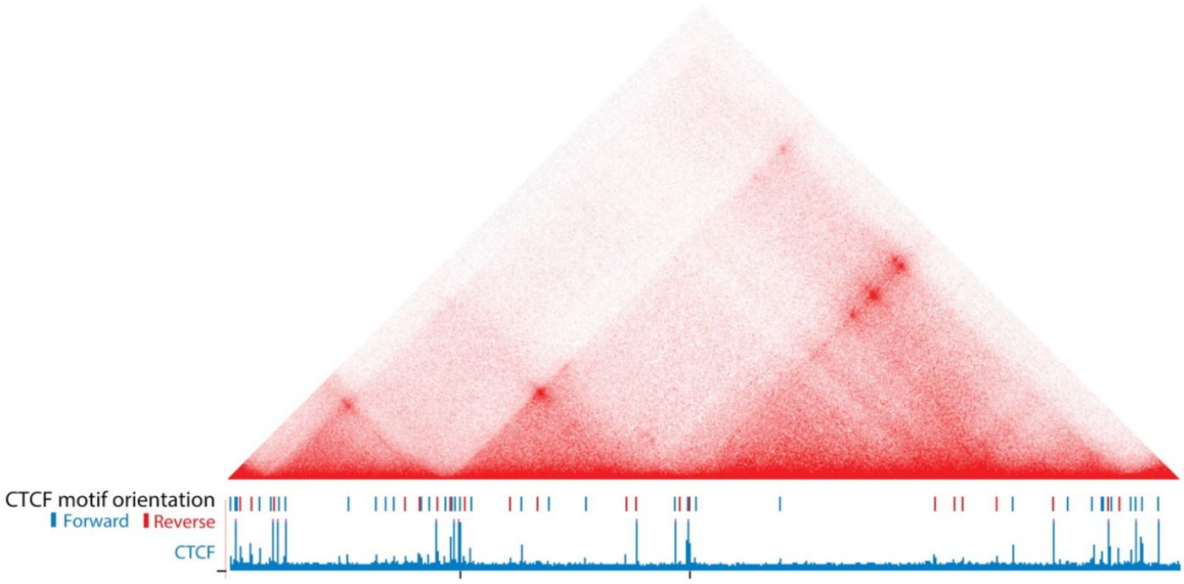

Figure 3: Hi-C contract matrix illustrating TADs, sub-TADs, E-P contacts, and CTCF boundaries. The darker triangles in the matrix are TADs; vertical distance correlates to the distance between the elements in contact. The darker dots represent chromatin loops, which may be due to enhancer-promoter interaction, another mechanistic chromatin looping interaction, or noise. The CTCF enrichment profile at bottom illustrates CTCF enrichment at TAD boundaries. Figure is derived from (Merkenschlager and Nora, 2016).

Intra-TAD and Inter-TAD EPIs

While existing work has shown convincingly that insulation sites at TAD boundaries block EPIs in many cases, 10 to 25% of inter-TAD interactions are EPIs in Metazoan cell lines (Stadhouders et al, 2011; Whalen et al, 2016), suggesting that some inter-TAD EPIs are more influenced by TAD boundaries than others. Numerous papers in the transcription literature have quantified the average determinants of EPIs genome-wide, but this approach results in a bias toward intra-TAD EPI determinants as those comprise the majority of EPIs (Whalen et al, 2016). Important outstanding questions about inter-TAD EPIs remain: are the determinants of inter-TAD and intra-TAD EPIs the same or do they differ? What determines which EPIs remain active and which are blocked in the presence of a TAD boundary?

This work begins to address these outstanding questions by comparing the factors that bind with active inter-TAD EPIs with those that bind "ordinary" or intra-TAD EPIs to understand if they function differently *in vivo*. To accomplish this, we compare the chromatin modification and transcription factor binding frequencies between these categories of EPIs in mES cells with an EPI sample set generated from ultra-high-resolution Hi-C-derived contact matrices and publicly available ChIP datasets. By using bioinformatics techniques to investigate whether different proteins mediate intra-TAD and inter-TAD EPIs, we discover that few differences in histone modification and transcription factor binding frequencies exist between the two categories in our sample. We also consider alternative approaches to defining EPI sets and discuss the tradeoffs of each, including those of our conservative approach. Our findings contribute a new approach to understanding how spatial interactions mediate transcription

regulation, which is an important element in the long march to understanding human biology and the many diseases related to gene expression.

Methods and Materials

TAD and Contact Calling

To probe differences between intra- and inter-TAD EPIs, we used reads from a micro-C dataset generated for a separate study (Hsieh et al, manuscript in progress) to call TADs and generate a list of contacts. The micro-C data were generated for a separate study with manuscript in progress and a list of all valid pairs of interacting elements was generated using HiC-Pro (Servant et al, 2015).

TADs and the DNA-DNA contact list used for analysis were generated using Juicer, a commonly used tool for Hi-C analysis (Durand et al, 2016). We used insulation scores as the metric for determining TAD boundaries, which means we defined TAD boundaries based on the propensity for elements on either side of the boundary to contact. Intrachromosomal contacts with a distance of greater than 2500 bp and all TADs labeled by Juicebox for all chromosomes were considered in the analysis.

ChIP-Seq Factor Selection and Analysis

To choose which transcription factors (TFs) to consider in our analysis, we generated a list of TFs and histone modifications known to play an important role in embryonic development and EPIs. Histone modifications documented in the literature as present at enhancer and

promoter sites were selected for analysis as well (Strahl and Allis, 2000). Only ChIP-seq reads of ES E14 cells in publicly accessible online databases were used in this analysis. ChIP-seq data were obtained from the ENCODE project database (ENCODE, 2007) and the European Nucleotide Archive (Leinonen et al, 2010).

ChIP-seq data from the various sources were formatted in the mm8, mm9, and mm10 genome builds. Reference genomes and chain files for converting between genome builds were obtained from the UCSC Sequence and Annotations database (http://hgdownload.cse.ucsc.edu/downloads.html). Raw ChIP-seq reads were aligned to the respective mouse genome using Bowtie (Langmead and Salzberg, 2012) and peaks were called using MACS14 (Zhang et al, 2008), both standard tools for ChIP-seq analysis. Only peaks of MACS intensity greater than 50 were included in subsequent analysis. Peak loci were then converted to the mm10 genome using CrossMap, a library the lifts over genomic locations from one genome build to another.

Enhancer-Promoter Interaction Sample Definition and Characterization

Several approaches were attempted to define a set of enhancer-promoter interactions. One approach was to take a published list of EPIs in mESC cells, such as the FOCS dataset (Hait et al, 2018), and define this list as the set of EPIs for analysis. A different approach was to label all elements in the micro-C contact set that correspond with known promoter regions as promoters and take all intrachromosomal interactions involving at least one promoter as the EPI set. A third approach combined the two previous approaches to generate a set of observed interactions in the

micro-C data that are also document as known EPIs in the literature. This EPI set was generated by taking the intersection of the the EPI sets created in the first two approaches.

Inter- vs. Intra-TAD Enrichment Analysis

The overall data processing procedure, including the pre-processing steps, is illustrated in Figure 4. All computations on annotated genomic sets were conducted in Python 3 using the pandas and numpy libraries. All visualizations were generated with matplotlib and seaborn.

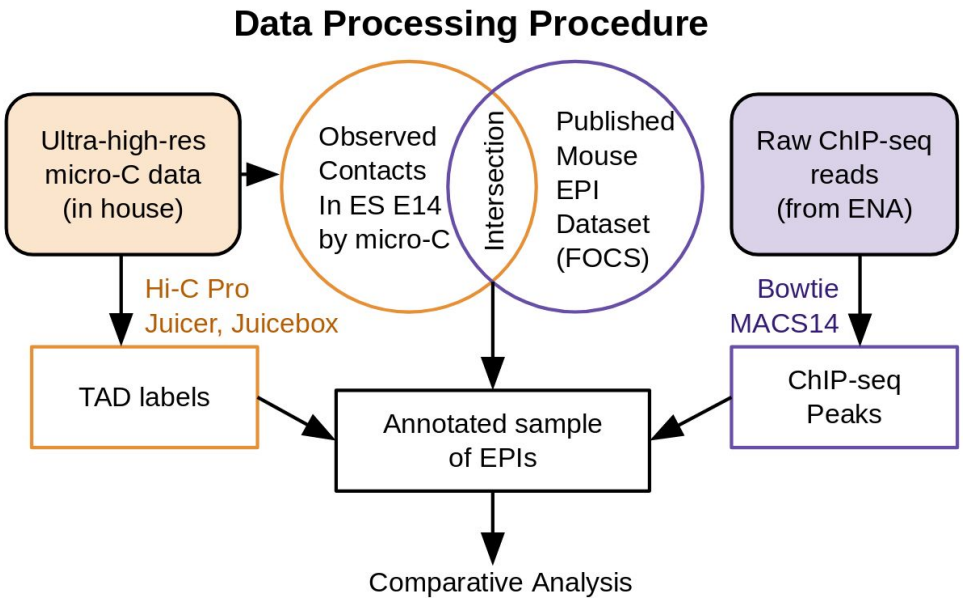

Figure 4: Flowchart overview of the data processing procedure used to generate the EPI sample used for analysis. Arrows indicate the direction of information flow, starting with data collected from online and physical experiments and ending with the annotated sample of EPIs.

Abbreviations: EPI = Enhancer-Promoter Interaction, TAD = Topologically Associating Domain, ENA = European Nucleotide Archive, FOCS = published enhancer set used.

Once the final set of EPIs was generated by intersecting the observed and published FOCS contact sets, each EPI was labeled with a TAD number and TF or histone modification enrichment label for each of the ChIP-seq peak sets. We then assessed the quality and distributions of this EPI sample and compared its characteristics to those published in the literature.

We compared enrichment levels of each histone modification and TF for contacts seen to be within the same TAD (intra-TAD) vs those seen to cross a TAD boundary (inter-TAD). We determined whether enrichment differences between intra-TAD and inter-TAD EPIs was significantly higher or lower than that expected by random chance by bootstrapping the sample with 100,000 simulations of randomly distributed ChIP-seq peaks. Finally, we compared our results with what is document in the literature to both validate known roles of various histone modifications and transcription factors and discover newly uncovered trends from this analysis.

Results

TAD and Contact Calling

Our TAD calling procedure resulted in a set of 4448 TADs averaging 622 kbp and ranging from 60 kbp to 2.03 Mbp, which falls within the range previously published in the literature (Rao et al, 2014). These TADs covered 39% of the genome as a result of the TAD-calling technique. Our contact calling procedure with a binning resolution of 200 bp resulted in a set of 34 million contacts.

ChIP-Seq Data Collection Results

The list of histone modifications and and TFs considered for this analysis are shown in Table 3. Most TFs of interest did not have publicly available ChIP-seq results for the ES E14 cell line so this analysis is limited to those TFs that had available ChIP-seq data.

Table 3: TFs and histone modifications analyzed along with the reason they were selected for analysis, their data source, and their GEO accession number.

TF or Histone Modification	**Reason Selected for Analysis**	**Data Source**	**GEO Accession Number**
H3K9ac	Known transcription activation mark	ENCODE	GSM1000123
H3K27ac	Known transcription activation mark	ENCODE	GSM1000126
H3K4me1	Known transcription activation mark	ENCODE	GSM1003750
H3K4me3	Known transcription activation mark	ENCODE	GSM1000124
H3K36me3	Known transcription activation mark	ENCODE	GSM1000125
CTCF	TAD boundary insulating protein	ENA	GSM699165
YY1	Structural regulator of EPIs	ENA	GSM788496
Suz12	In Polymcomb Repressive Complex 2	ENA	GSM288360
Rad21	Helps maintain stem cell identity	ENA	GSM2099837
Pol2A	Transcriptional polymerase	ENA	GSM699166
P300	Known enhancer-binding TF	ENA	GSM699164
Sox2	Maintains ESC pluripotency	ENA	GSM288347
Klf4	Maintains ESC pluripotency	ENA	GSM288354
Oct4	Maintains ESC pluripotency	ENA	GSM288346
Nanog	Maintains ESC pluripotency	ENA	GSM288345
c-Myc	Maintains ESC pluripotency	ENA	GSM288356
Esrrb	Maintains ESC pluripotency	ENA	GSM288355

Enhancer-Promoter Interaction Sample Characteristics

A total of 914 EPIs were included in the sample used for analysis. The distributions of intra-TAD, inter-TAD, and TAD-less contacts in the sample are shown in Figure 5A. The sample contained 516 intra-TAD contacts, 135 inter-TAD contacts, and 259 TAD-less contacts. EPIs were labeled TAD-less if neither enhancer nor promoter lied within a TAD; TAD-less EPIs were not considered in ChIP-seq analyses.

The linear genomic distance distribution of intra-TAD and inter-TAD contacts is shown in figure 5B. All contacts were of linear distance greater than 2500 bp by definition and the inter-TAD and intra-TAD EPI distributions overlapped significantly. The median linear genomic distance of intra-TAD EPIs was 14.8 kb and while that of inter-TAD EPIs was 44.6 kb.

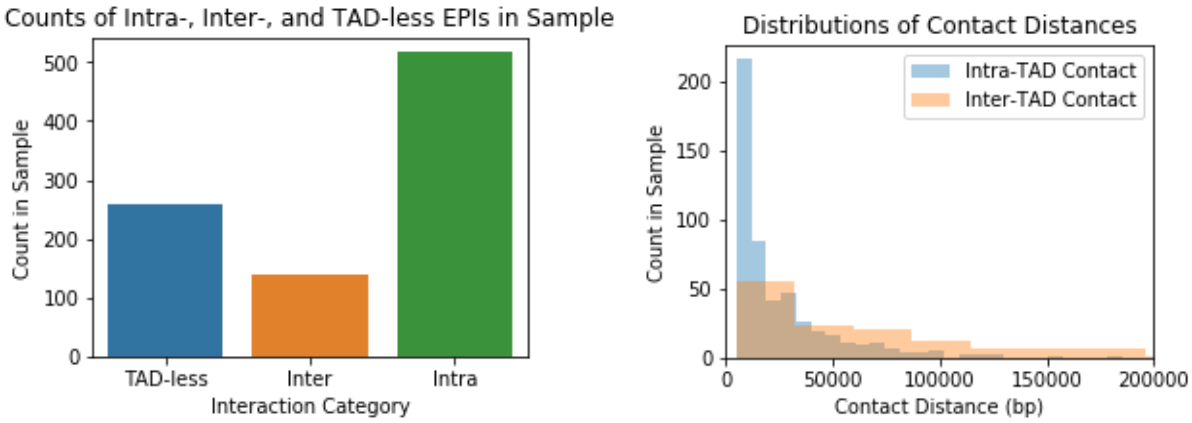

Figure 5: (A) Distribution of intra-TAD, inter-TAD, and TAD-less contacts in the sample. (B) Distribution of linear genomic distances of contacts for the intra-TAD and inter-TAD subsets of the sample.

Intra-TAD vs Inter-TAD Histone Modification Trends

Differences between enrichment of H3K9ac, H3K27ac, H3K4me3, H3K4me1, and H3K36me1 for intra- and inter-TAD EPIs are illustrated in Figure 6A-E. H3K9ac, H3K27ac, and H3K4me3 were enriched for >50% of promoters in the EPI sample while H3K4me1 and H3K36me3 both showed <5% enrichment for promoters. Only enrichment of H3K27ac on promoter sites showed a significant difference ($p<0.05$) between intra-TAD and inter-TAD EPIs with 95% confidence.

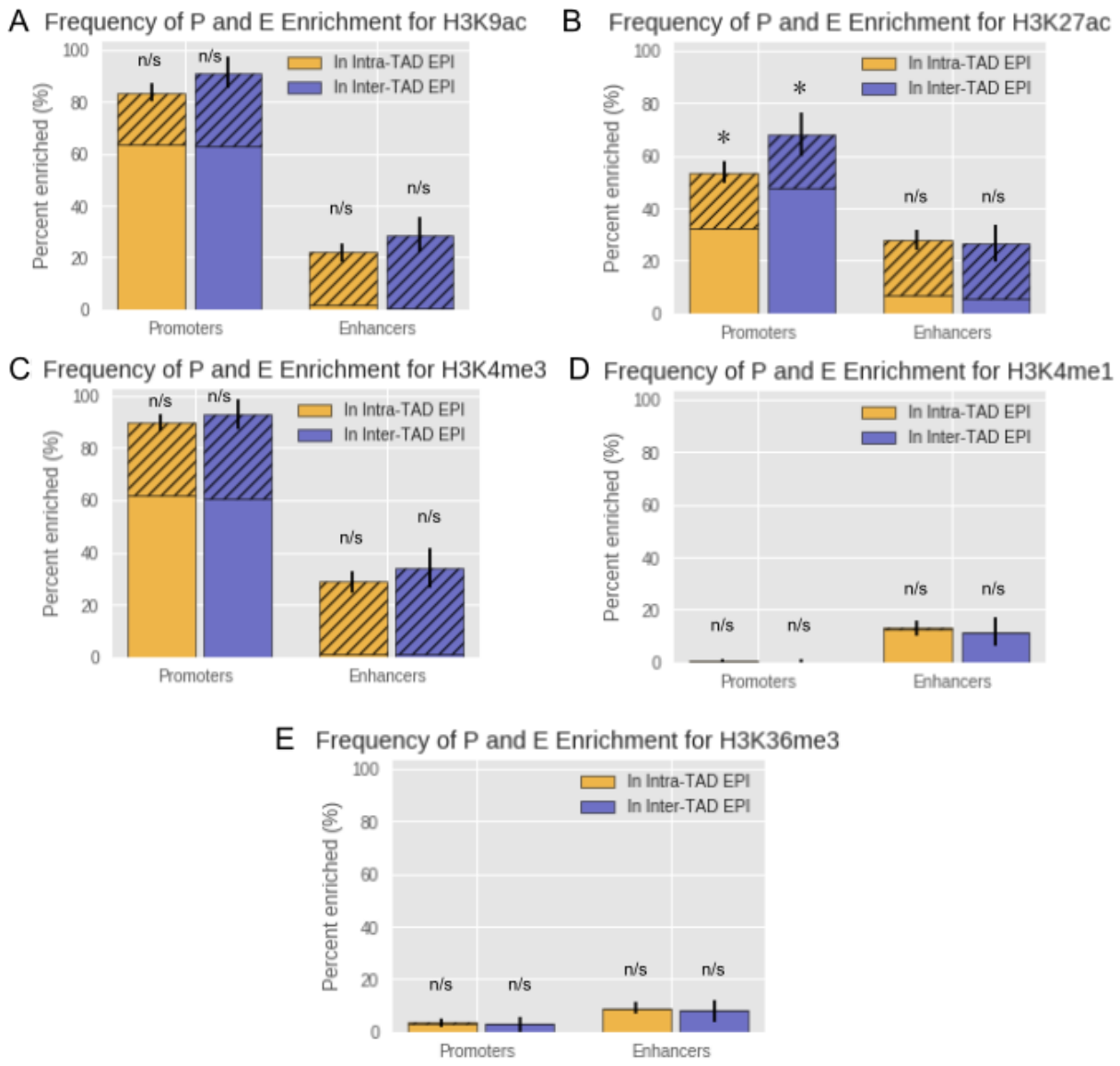

Figure 6: Enrichment levels for (A) H3K9ac, (B) H3K27ac, (C) H3K4me3, (D) H3K4me1, and (E) H3K36me3 on promoter and enhancer sites in intra-TAD and inter-TAD EPIs. The hatched portions indicate the percentage of enriched elements whose interacting element was also enriched for the histone modification of interest. For example, in panel A the first bar indicates that 82% of promoters seen to be involved in inter-TAD EPIs fell within H3K9ac ChIP-seq peaks, and about 20% of those promoters' corresponding enhancer also fell within an H3K9ac ChIP-seq peak.

Intra-TAD vs Inter-TAD Transcription Factor Trends

Differences between enrichment of transcription factors CTCF, YY1, Suz12, Rad21, Polr2A, and p300 for intra- and inter-TAD EPIs are illustrated in Figure 7A-F. Only Polr2A showed promoter enrichment of >50%. Polr2A and Rad21 enrichment were found to be significantly greater ($p<0.05$) on the promoter in inter-TAD EPIs than intra-TAD EPIs. Enhancer enrichment for all TFs and promoter enrichment for CTCF, YY1, Suz12, and p300 did not show significant differences.

Enrichment of pluripotent stem cell identity retaining TFs Oct4, Sox2, Klf4, Nanog, c-Myc, and Esrrb all showed infrequent enrichment on promoters and enhancers. Comparisons between the enrichment of Klf4, c-Myc, and Esrrb for intra- and inter-TAD EPIs are illustrated in Figure 7A-F; none showed significant differences between intra-TAD and inter-TAD EPI element enrichment. Oct4, Sox2, and Nanog did not show an enrichment frequency of >5% on promoters or enhancers, indicating that these TFs did not play a role in regulating the expression of the genes corresponding to the promoters in the EPI set.

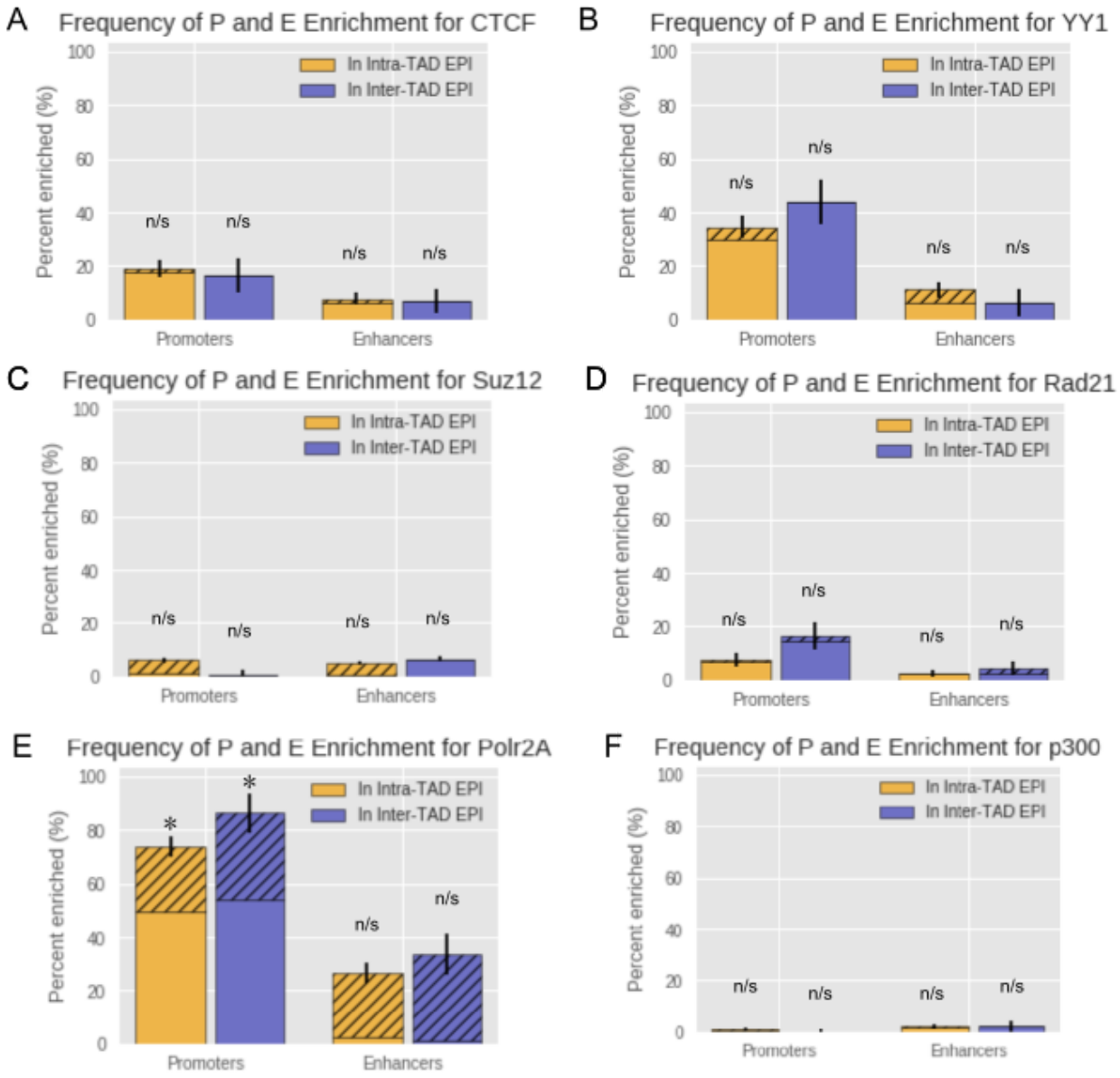

Figure 7: Enrichment levels for (A) CTCF, (B) YY1, (C) Suz12, (D) Rad21, and (E) Polr2a, and (F) p300 on promoter and enhancer sites in intra-TAD and inter-TAD EPIs. The hatched portions indicate the percentage of enriched elements whose interacting element was also enriched for the transcription factor of interest.

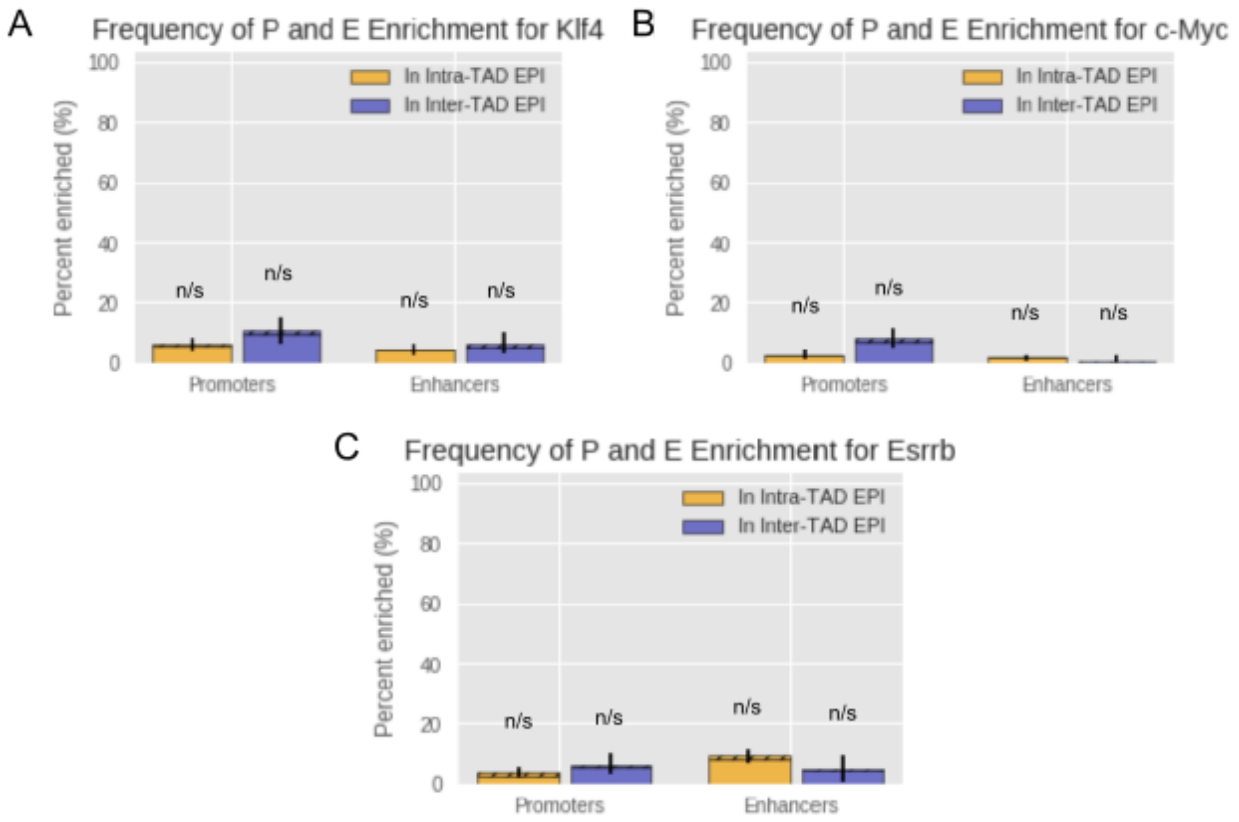

Figure 8: Enrichment levels for (A) Klf4, (B) c-Myc, and (C) Esrrb on promoter and enhancer sites in intra-TAD and inter-TAD EPIs. The hatched portions indicate the percentage of enriched elements whose interacting element was also enriched for the transcription factor of interest.

Discussion

Sample Characteristics Suggest Biases in EPI Set Definition

The procedure used to generate the sample of EPIs resulted in 914 EPIs, which is far fewer than the number of EPIs in mESCs reported in the literature, which is approximately 50,000 (Hait et al, 2018). The procedure identified EPIs that were both predicted based on their features via the FOCS dataset and seen to be interacting in micro-C data, which is a stronger

requirement than most EPI set definitions. As a result, the sample is not a random sample of EPIs from mESC cells, but rather a sample of strong contacts with one element that is a promoter and another that has features that are indicative of enhancer elements. This conservative definition lends confidence that our set is indeed comprised of EPIs, but it results in the sample not being representative of all EPIs. All subsequent discussion must be taken in light of sample bias -- while statistical differences may be true for the sample, we cannot state whether these differences would extend to other EPI definitions.

Another important result of the sampling procedure is that because our set consists of strong EPIs, only the strongest EPI involving a given promoter in the set in most instances. This is noteworthy because every promoter usually has multiple enhancers (Mumbach et al, 2017), which could complicate our measures of enrichment for histone modifications and transcription factors in promoters. For example, it could be the case that only one interacting enhancer need be enriched for transcription to take place, but that enhancer may not be the one in our set. Indeed, if the strongest interacting enhancer is not the one that is enriched, our set would show that the factor must be enriched on the promoter but not the enhancer. Our analysis of promoter enrichment is conducted conducted with this issue of superimposed enhancer interactions in mind.

Histone Modification Enrichment Validates Known Trends With Varying Prevalence

Histone modifications H3K9ac, H3K27ac, and H3K4me3 both showed strong (>50%) promoter enrichment frequency in both intra-TAD and inter-TAD interactions, which is expected because they are all known transcription activation marks (Strahl and Allis, 2000). Interestingly,

the majority of EPI-involved promoter elements enriched in these histone modifications did not have corresponding enriched enhancer elements, but almost all enriched enhancer elements had interacting promoter elements that were enriched. In other words, very few or none of the EPIs showed enrichment in H3K9ac, H3K27ac, or H3K4me3 on solely the enhancer site, while many of the EPIs showed enrichment solely on the promoter site. However, a noteworthy fraction (>20%) of enhancers are enriched for these histone marks. It could be that the enrichment of promoter sites in H3K9ac, H3K27ac, and H3K4me3 is essential for EPIs in this sample but the enrichment of enhancers, though frequent, is as essential for interaction. Alternatively, it could be that the promoter is also interacting with other enhancer sequences which actually are enriched, meaning the promoter is indeed interacting with an enriched enhancer but not this specific enhancer.

Histone modifications H3K36me3 and H3K4me1 showed a much lower enrichment frequency on both promoters and enhancers than that of H3K9ac, H3K27ac, and H3K4me3, even though they are also considered activating factors. This difference in enrichment frequency suggests that H3K9ac, H3K27ac, and H3K4me3 are more universally required to activate transcription than H3K36me3 and H3K4me1.

Few significant differences in histone modification and TF enrichment between inter-TAD and intra-TAD EPIs

With a few exceptions, ChIP-seq enrichment for the TFs and histone modifications considered in this analysis did not significantly differ ($p<0.05$) between intra-TAD and inter-TAD subsets. The exceptions are promoter enrichment for Polr2A (RNA Polymerase II;

Pol II), Rad21, and H3K27ac, which each showed stronger enrichment in promoters involved in inter-TAD EPIs. No significant differences in enhancer enrichment were detected.

This result suggests that, at least for the EPIs sample used, there is no reason to suspect that markedly different mechanisms are involved in mediating intra-TAD versus inter-TAD interactions. In fact, our results indicate that no single histone modification or TFs required for all enhancer-promoter interactions because none of the enrichment frequencies for the conservatively defined EPI sample is 100%, even on promoters. Bar the possibility of experimental error, this means the best we can say is that *most* promoters involved in EPIs are enriched for a given factor or mark, not that *all* are.

Pol II is more frequently enriched in inter-TAD EPIs than intra-TAD EPIs

A noteworthy difference between intra-TAD and inter-TAD EPIs seen in Figure 7 is in Pol II enrichment on promoters: promoters in inter-TAD EPIs in this sample are more frequently enriched with Pol II. While Pol II enrichment of a given promoter does not necessarily mean that its corresponding gene is transcribed, it does indicate that the EPI successfully recruited the transcription initiation complex, which includes Pol II. One possible explanation as to why Pol II is enriched in more inter-TAD EPIs than intra-TAD EPIs could be that more intra-TAD EPIs are interacting due to chance rather than as part of transcription initiation. By our definition of TADs that was based on insulation scores, it is expected that random elements within the same TAD interact more frequently than those outside the TAD. In addition, enhancer and promoter elements are known to have a high propensity for interacting with other DNA elements non-specifically (Mao et al, 2018). Thus it is possible that an enhancer and promoter element in

the same TAD could interact frequently not as a result of a transcription-related mechanism but rather as a consequence of the high interaction propensities of these elements within the same TAD in general.

ESC-Specific TFs showed low enrichment in all sample subsets

Transcription factors Sox2, Klf4, Oct4, Nanog, c-Myc, and Esrrb are all known to play important roles in maintaining the identity of pluripotent stem cells (Schmidt and Plath, 2012). Unexpectedly, none of those TFs are seen to be enriched at noteworthy levels in the EPIs in this sample. While this result would be surprising if it were the trend for for EPIs in mESCs in general, it is not necessarily surprising considering the nature of the sample used in this analysis. As mentioned previously, this sample consists of relatively strong EPIs, such as those known to be present for promoters encoding unregulated genes such as those involved in metabolism (Herzig, 2004). Since Sox2, Klf4, Oct4, Nanog, and Esrrb are are all regulatory factors, they may not be involved in the EPIs that are not regulated, explaining the lack of ChIP-seq enrichment seen in our sample. However, c-Myc is known to bind promoter sites preceding transcription, which is not seen in our results; we suspect that this could be due to an issue with the ChIP-seq sample used for analysis.

Outstanding Challenges in EPI Definition

The limitations brought by our EPI sample generation procedure points to a general challenge in studying EPIs globally: it is difficult to define what an enhancer is, and global results depend on which definition is used. Whereas promoters have specific sequence elements

where RNA Polymerase II binds, enhancers vary in their characteristics. The current state-of-the-art procedure for defining enhancer elements is to use a multivariate hidden markov model trained on select histone modification and TF ChIP-seq data to annotate elements of the genome (Ernst and Kellis, 2012). This is usually accomplished using the popular software ChromHMM (compbio.mit.edu/ChromHMM). However, defining enhancers using hidden markov models inevitably produces a set of loci that resembles known enhancers, and whether those loci will behave as enhancers must be validated by other means.

An alternative, and rather naive, approach would be to define all elements that interact with the promoter of an actively expressed gene as an enhancer. The flaw in this approach is that promoters have a high propensity to interact with all local DNA, not just enhancer elements (Mao et al, 2018). Several other approaches exist to define a set of global enhancers, but each results in a different set with either bias or uncertainty. So long as bias or uncertainty regarding enhancer definitions remain, any approach to global EPI analysis will be subject to those biases and uncertainties.

Potential for Modified Inter-TAD Analyses with New EPI Sets

Despite the limitations brought by the sample used for this study, the bioinformatics procedure developed for the analysis can be applied to any set of EPIs. Whether or not a definition of enhancer is ever universally accepted, this procedure can be used to study subsets of EPIs of interest, such as those known to be especially active in a certain cell line or gene cluster. Numerous outstanding questions about TAD boundaries and EPIs remain, including the underlying question of this study: why can a small proportion of EPIs bypass the insulation

boundaries blocking most EPIs? For the sample considered in this analysis, comparisons of the presence or absence of the histone modifications and transcription factors considered did not answer this question. Regardless, the limitations of this work highlight core challenges in EPI research and can help inform future researchers as to what considerations to make when conducting global genomic analyses involving enhancers.

References

Bannister, A.J., and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Research *21*, 381–395.

van Berkum, N.L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L.A., Dekker, J., and Lander, E.S. (2010). Hi-C: A Method to Study the Three-dimensional Architecture of Genomes. J Vis Exp.

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nature Reviews Genetics *2*, 292–301.

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

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Systems *3*, 95–98.

The ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636–640.

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

Giorgetti, L., Galupa, R., Nora, E.P., Piolot, T., Lam, F., Dekker, J., Tiana, G., and Heard, E. (2014). Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription. Cell *157*, 950–963.

Gonzalez-Sandoval, A., and Gasser, S.M. (2016). On TADs and LADs: Spatial Control Over Gene Expression. Trends in Genetics *32*, 485–495.

Grewal, S.I.S., and Moazed, D. (2003). Heterochromatin and Epigenetic Control of Gene Expression. Science *301*, 798–802.

Hait, T.A., Amar, D., Shamir, R., and Elkon, R. (2018). FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer–promoter map. Genome Biology *19*, 56.

He, L., and Hannon, G.J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. Nature Reviews Genetics *5*, 522–531.

Herzig, S. (2016). Metabolic Control (Springer).

Hsieh, T.-H.S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., and Rando, O.J. (2015). Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell *162*, 108–119.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lee, T.I., and Young, R.A. (2000). Transcription of Eukaryotic Protein-Coding Genes. Annual Review of Genetics *34*, 77–137.

Lee, T.I., and Young, R.A. (2013). Transcriptional Regulation and its Misregulation in Disease. Cell *152*, 1237–1251.

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., et al. (2011). The European Nucleotide Archive. Nucleic Acids Res. *39*, D28-31.

Lupiáñez, D.G., Spielmann, M., and Mundlos, S. (2016). Breaking TADs: How Alterations of Chromatin Domains Result in Disease. Trends in Genetics *32*, 225–237.

Mao, W., Kostka, D., and Chikina, M. (2018). The perils of interaction prediction. BioRxiv 435065.

Mehra, P., and Kalani, A. (2018). What's in the "fold"? Life Sciences *211*, 118–125.

Merkenschlager, M., and Nora, E.P. (2016). CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. Annu. Rev. Genom. Hum. Genet. *17*, 17–43.

Miele, A., Gheldof, N., Tabuchi, T.M., Dostie, J., and Dekker, J. (2006). Mapping Chromatin Interactions by Chromosome Conformation Capture. Current Protocols in Molecular Biology *74*, 21.11.1-21.11.20.

Müller, M.M, Gerster, T., Schaffner, W. (1988). Enhancer sequences and the regulation of gene transcription. European Journal of Biochemistry *176*, 485-495.

Mumbach, M.R., Satpathy, A.T., Boyle, E.A., Dai, C., Gowen, B.G., Cho, S.W., Nguyen, M.L., Rubin, A.J., Granja, J.M., Kazane, K.R., et al. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nature Genetics *49*, 1602–1612.

Narendra, V., Bulajić, M., Dekker, J., Mazzoni, E.O., and Reinberg, D. (2016). CTCF-mediated topological boundaries during development foster appropriate gene regulation. Genes Dev. *30*, 2657–2662.

Ong, C.-T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nature Reviews Genetics *15*, 234–246.

Ouyang, Z., Zhou, Q., and Wong, W.H. (2009). ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proceedings of the National Academy of Sciences *106*, 21521–21526.

Park, P.J. (2009). ChIP–seq: advantages and challenges of a maturing technology. Nature Reviews Genetics *10*, 669–680.

Poss, Z.C., Ebmeier, C.C., and Taatjes, D.J. (2013). The Mediator complex and transcription regulation. Crit. Rev. Biochem. Mol. Biol. *48*, 575–608.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell *159*, 1665–1680.

Schmidt, R., and Plath, K. (2012). The roles of the reprogramming factors Oct4, Sox2 and Klf4 in resetting the somatic cell epigenome during induced pluripotent stem cell generation. Genome Biol *13*, 251.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biology *16*, 259.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nature Reviews Genetics *13*, 613–626.

Stadhouders, R., Heuvel, A. van den, Kolovos, P., Jorna, R., Leslie, K., Grosveld, F., and Soler, E. (2012). Transcription regulation by distal enhancers. Transcription *3*, 181–186.

Strahl, B.D., and Allis, C.D. (2000). The language of covalent histone modifications. Nature

403, 5.

Symmons, O., and Spitz, F. (2013). From remote enhancers to gene regulation: charting the genome's regulatory landscapes. Philos Trans R Soc Lond B Biol Sci *368*.

Tjian, R., and Maniatis, T. (1994). Transcriptional activation: A complex puzzle with few easy pieces. Cell *77*, 5–8.

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854–858.

Whalen, S., Truty, R.M., and Pollard, K.S. (2016). Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat Genet *48*, 488-496.

Woringer, M., and Darzacq, X. (2018). Protein motion in the nucleus: from anomalous diffusion to weak interactions. Biochemical Society Transactions *46(4)*, 945-956.

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

Acknowledgements

I would like to thank my generous advisers Prof. Xavier Darzacq and Maxime Woringer for their extensive support and advice, as well as Tsung-Han (Stanley) Hsuang, PhD for sharing his micro-C data and depth of knowledge. I also extend my gratitude to the others who provided enthusiasm and support towards my project, including Prof. Robert Tjian, Prof. Michael Eisen, Claudia Cattoglio, PhD, Hervé Marie-Nelly, PhD, and Alec Heckert.