Conformation and Dynamics of the Mammalian Chromosome

Lyndon N. Zhang
Bachelor of Science in Biology
Stanford University, 2012

Submitted to the Department of Biology in partial fulfillment of the requirements for the degree of Master of Science at the Massachusetts Institute of Technology.

February 2016
© 2016 Massachusetts Institute of Technology. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

Lyndon N. Zhang
Department of Biology
October 19, 2015
Author Signature:

Ibrahim I. Cissé
Assistant Professor of Physics
Thesis Supervisor
Certification Signature:

Amy E. Keating
Professor of Biology and Biological Engineering
Co-Director, Biology Graduate Program
Acceptance Signature:
Conformation and Dynamics of the Mammalian Chromosome

Lyndon N. Zhang

Abstract

The control of transcription represents a fundamental, initial mechanism by which the regulation of gene expression is implemented. However, while much research has been done on the biochemistry and cellular function of transcription, comparatively little is known on the dynamics of transcriptional mechanisms, their impact on chromatin structure, and concomitant functional consequences. Employing chromatin immunoprecipitation measurements, we report progress towards this goal. We characterize the ensemble chromosome conformation in mouse embryonic stem cells, by measuring interaction, or contact, probabilities between distal genomic loci. We map and describe chromosome loops, consisting of two interacting CTCF sites co-bound by cohesin, that maintain the expression of genes known to promote cell identity, and restrict the expression of genes specifying repressed developmental lineages.

1 Contributions and Prior Publication

The writing of this thesis is entirely the work of the author, with background and methods sourced from the indicated references. The data analysis methods and results, described in sections 3 and 4, respectively, were based on the work of Charles Y. Lin, David A. Orlando, and Brian J. Abraham, and performed in collaboration and under the supervision of Zi Peng Fan and Brian J. Abraham. As the research was performed collaboratively, the author has elected to use “we” in the description of the work.

Experimental methods were performed in collaboration with and courtesy of Jill M. Dowen, Gang Ren, Denes Hnisz, Abraham S. Weintraub, Jurian Schuijers, and Alla Sigova. Direct intellectual oversight originated from Zi Peng Fan, with additional oversight originating from Tong Ihn Lee and Richard A. Young. Work was performed primarily in the lab of Richard A. Young, in collaboration with the lab of Keji Zhao. Phillip A. Sharp and members of the Sharp lab provided valuable discussion and insight.

All research results were published in reference 25, reproduced here: Jill M. Dowen, Zi Peng Fan, Denes Hnisz, Gang Ren, et. al. “Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes.” Cell. 2014; 159:374-87. The paper is published under the Cell open access policy, and listed in Pubmed Central (PMC4197132). All figures in section 4, with the exception of Figure 10 (which was adapted from reference 23), were adapted from figures in this paper.

Research in section 6 was conducted in the lab of Ibrahim I. Cissé. The project was conceived and supervised by Ibrahim I. Cissé. Takuma Inoue implemented the optical microscope used. The cell line was generated by various members of the Cissé lab. Analysis was implemented according to details in the manuscript İzeddin, et. al., “Single-molecule tracking in live cells reveals distinct target-search strategies of transcription factors in the
nucleus.” eLife. 2014; 3:e02230, from the laboratories of Xavier Darzacq and Maxime Dahan, with optimization in the Cissé lab.

Additional analytical steps were written in Mathematica by Ibrahim I. Cissé and optimized by the author. All data presented were acquired and analyzed by the author. Direct assistance was received from James Owen Andrews, Arjun Narayanan, and Takuma Inoue. Other members of the Cissé lab, including Jan-Hendrik Spille, Namrata Jayanth, and Wonki Cho, provided valuable discussion and insight.

2 Introduction

Transcription, the first step of gene expression, is the cellular mechanism by which genes are copied from the DNA template to the complementary mRNA molecule. Mediating this templated polymerization reaction is the RNA polymerase II holoenzyme, which catalyzes the formation of phosphodiester bonds to link nucleoside triphosphates in a hydrolysis reaction [1]. Eukaryotic transcription initiates with the binding of an activator to its cognate distal enhancer sequence to enable the sequential recruitment of the general transcription factors and RNA polymerase II to the cognate gene promoter [2]. Subsequently, proximal response elements upstream of the promoter, such as the TATA sequence, are recognized and bound by cognate proteins, such as the TATA-binding protein (TBP). In this case, the TBP binds, as part of the TFIID complex, to the DNA, 35 base pairs upstream of the transcription start site. Binding of TFIID then promotes the recruitment of the remaining general transcription factors, TFIIB, F, E, H, and the Mediator complex, to form the transcriptional preinitiation complex [3].

However, by themselves, RNA polymerase II and the general transcription factors do not fully reconstitute the in vivo response to transcription regulators [4]. DNA is packaged with histone proteins to form chromatin, and the compaction this achieves enables the 2-meter DNA polymer to package inside eukaryotic nuclei with diameters on the order of microns. An additional host of enzymatic apparatus, including chromatin modifiers — (de)acetylases, (de)methylases, (de)phosphorylases — and ATP-dependent chromatin remodelers — the SWI/SNF and the NuRD remodeling complexes — modulates access to the DNA template, from the “open” 10-nm filament, to the 30-nm fiber, and to more compacted, higher order packings [5]. More recent studies suggest that a fraction of noncoding RNA, transcribed at enhancers, may collaborate with activating transcription factors, and stabilize chromatin loops, to further upregulate gene expression [6].

The regulation of transcription is essential for the specification of cell state [7]. Post-transcriptional gene regulatory programs and external factors are also known to be involved in the control of cell identity. For instance, the miR-290-295 microRNA family maintains embryonic stem cell identity by repressing Pax6 expression, and consequently ectoderm differentiation [8]; extrinsic signaling factors, acting through the cell’s molecular signaling network (e.g., in human embryonic stem cells, the binding of the nodal signal to its cognate receptor, activin), promote the expression of transcription factors to further alter cell identity [9]. However, here we focus on the transcriptional regulatory circuitry required for the specification and maintenance of cell identity.

To establish the transcriptional program required for maintaining cell identity, a set of “master” transcription factors — those whose protein products bind to the enhancers of themselves and of each other — activate their own expression and that of downstream effector genes [7]. This set of transcription factors forms an interconnected autoregulatory loop to maintain cell state. The identity and behavior of this set of factors, that form
the core regulatory circuitry, are best studied in embryonic stem cells (in both human and mouse), in which the transcription factors Oct4, Sox2, and Nanog (OSN) were identified as governing the pluripotent state [7, 10]. In murine embryonic stem cells (ESCs), Oct4 forms a heterodimer with Sox2. While Nanog is not necessary for self-renewal (Nanog−/− ESCs retain wild-type Oct4 expression and the capability to propagate in culture), Nanog-deleted embryonic stem cells are recruited to the germ line, and reach the soma, but do not migrate to the genital ridge beyond 11.5 days after embryogenesis [11]. Additionally, by ChIP-seq, Nanog colocalizes at the majority of Oct4 and Sox2 binding sites throughout the ESC genome, including regions proximal to the Sox2 promoter [12]. Collectively, these three transcription factors — Oct4, Sox2, and Nanog — operationally define the enhancers that transcriptionally activate genes that drive the embryonic cell state [7].

We examine the first step of transcriptional initiation: the binding of activating transcription factors at distal enhancers, and the formation of a physical loop between the enhancer and promoter. To experimentally characterize this, recent molecular technologies, based on modifications of the chromatin immunoprecipitation (ChIP) method, are used to capture a substantial fraction of pairwise interactions between distal genomic loci. In ChIA-PET (chromatin interaction analysis by paired-end tag sequencing), sonication and pulldown of paired DNA linkers colocalizing with the protein of interest are performed as in chromatin immunoprecipitation. Linkers are ligated to the tethered DNA fragments; a subsequent ligation reaction, under dilute reagent concentrations, joins linkers ligated to different fragments. Sequencing and analysis of the paired-end reads identifies local chromosomal structures across the genome [13].

Here, we focus on the chromosome structures associated with the cohesin complex. Cohesin is one of multiple structural maintenance of chromosome proteins, ring-shaped protein complexes that link two DNA molecules. (The SMC family also includes the condensin I and II complexes.) The cohesin complex is comprised of Smc1 and Smc3, which contain coiled-coiled arms, the heat repeat domain-containing STAG, and the Rad21 bridging protein. In eukaryotes, cohesin is responsible for maintaining cohesion of sister chromatids [14]. In yeast, an 80% knockdown of cohesin expression led to transcriptional defects, but sister chromatid cohesion and chromosome structure remained intact, suggesting that cohesin plays a functional role in transcription [15]. Additionally, chromatin immunoprecipitation assays show cohesin colocalized with actively transcribed promoters and enhancers. In mouse embryonic stem cells, colocalization of cohesin is observed, at enhancers, with Med1/Med12 (components of the Mediator complex) and master transcription factors that bind to enhancers and drive the expression of cell identity-specifying genes; at promoters, cohesin colocalizes with Mediator and Pol II [16]. The molecular interactions suggested by these genome-wide ChIP data were biochemically validated in western blot, through the co-purification of Smc3 with Med12 in nuclear cell extract [16]. Finally, reduction of Mediator, cohesin, and Nipbl (a cohesin loading protein) — achieved through shRNA knockdown — resulted in lower Oct4 protein levels, aberrant ESC colony morphology, and reduced OSN expression levels. These show, genetically, the role of cohesin-associated proteins, and Mediator, in maintaining the transcriptional control of cell state.

More recent studies identified a subset of OSN enhancers in stem cells as containing a disproportionately larger proportion of the transcriptional machinery. By aggregating enhancers in linear proximity (within 12.5 Kbp), we define, based on total Mediator signal, a small number of enhancer domains, termed “super-enhancers,” that dominate the transcriptional program specifying cell state. These super-enhancers contain high densities of not only Oct4, Sox2, and Nanog, but also Klf4 and Esrrb, and are enriched for the concomitant
transcription factor binding motifs. Genes associated with super-enhancers (also by linear proximity) show higher expression; they also show increased sensitivity to Oct4 and Med12 levels, upon shRNA knockdown. Super-enhancers, cloned in luciferase assays driven by the Oct4 promoter, activate expression to higher levels than typical enhancers. Examining super-enhancers in multiple cell types suggests that genes associated with super-enhancers are more cell-type specific, and drive the developmental functions of their respective cell types (e.g., MyoD, which has been shown to reprogram fibroblasts to muscle cells, occupies a super-enhancer that activates its own expression in skeletal muscle), than those associated with typical enhancers [17]. Thus, to study the chromosome structures responsible for controlling cell identity, we specialize to the cohesin-associated loops surrounding super-enhancers, and similar genomic regions, that specify cell lineage.

To gain insight into previously identified chromosome structures, we examined previous chromatin interaction studies performed by a related experimental method, Hi-C. Hi-C modifies standard chromosome conformation capture, to comprehensively detect the set of chromatin interactions in the nucleus. Similar to ChIA-PET, chromatin is crosslinked, fragmented, ligated, and purified to create a library of interacting DNA loci. However, without immunoprecipitation, Hi-C measures all interactions, rather those mediated by a specific factor. Data from these studies capture the chromatin conformation ensemble; contact density maps of sufficient quality require samples on the order of 10^6 cells. Recent Hi-C studies showed that contact probabilities between loci P(s) in condensed chromatin (seen in the mitotic phase) initially scale inversely with the square root of the genomic distance apart s (P(s) \propto s^{-0.5}), followed by a rapid fall-off at 10^7 base pairs, at which P(s) \propto s^{-1} [18]. Human chromatin organization within the 10 Mb cutoff exhibits scaling that falls between the fractal (P(s) \propto s^{-0.1}) and equilibrium globule states (P(s) \propto s^{-1} for s \leq 10^6 bp, as in the fractal globule, but plateaus with P(s) \propto s^0 for s \geq 10^6 bp) [18]. This contrasts with human interphase chromatin organization, in which contact probability scales inversely with genomic distance from 500 Kb to 7 Mb [19]. Application of the statistical mechanics of polymers informs the physical mechanisms of chromatin folding, showing that unlike proteins, which fold into a unique native conformation, chromatin fiber adopts unique conformations in individual cells, of which only the ensemble average is shown through these data [20] (barring more recent single-cell Hi-C methods [21]). The length to chain diameter ratio for the chromatin polymer in mammals is on the order of 10^3 to 10^4 higher than for single protein domains (\sim 10^5 – 10^6 for human chromosomes, but \sim 50 – 250 for proteins) [20]. Overall, the statistics of the long-range interactions indicate the chromatin as organizing into a long-lived, nonequilibrium conformation that forms as a consequence of rapid condensation [19, 20].

Over smaller length scales, the Hi-C experiment identities local chromosome structures. Expanding the contact probability matrix T by eigenvector decomposition (T_{ij} = \sum_k \lambda_k E_i^k E_j^k + \langle T \rangle) provides the interaction preference eigenvectors \{E_i^k, E_j^k\} of genomic regions \{i, j\}, weighted by eigenvalue \lambda_k, of order k [24]. The first eigenvector E_1 provides a linear model of chromatin interaction preferences, and is used to separate the chromatin into two compartments, “A” and “B”. Compartments preferentially self-interact and correspond, roughly, to active/euchromatic (“A”) and inactive/heterochromatic (“B”) chromatin, correlating with indicators such as DNA I hypersensitivity, gene density, GC content, and histone marks [22]. Compartments have a characteristic size of \sim 5 Mb and are tissue-specific [19, 24].

At finer resolutions, further Hi-C studies identified so-called topological associating domains (TADs) [18, 22, 23]. Schematics of compartments and topological activating domains, adapted from references 22 and 23, are shown in Figures 1-3. Topological associating do-
mains have a characteristic size of $\sim 400 - 500 \text{ Kb}$, and can correspond to either active or inactive chromatin. In contrast to compartments, TADs are not cell-type specific, nor do adjacent TADs necessarily correspond to opposite transcriptional activity [22]. DNA elements at the boundary of TADs include CTCF binding sites, the H3K4me3 and H3K36me3 histone marks, transcriptional start sites, housekeeping genes, and nascent mRNA signal (as measured by global nuclear run-on sequencing) [22, 23]. Topological domains are conserved between syntenic regions in human and mouse embryonic stem cells. Finally, in contrast to the G1 state, TADs (and compartments) were observed to be lost in HeLa S3 cells in the M phase [18]. Therefore, topological associating domains are hypothesized to function as an evolutionarily-conserved mechanism to partition the genome into regions, containing long-distance interactions between genes and distal regulatory elements [23].

Figure 1. A schematic of chromosome compartments, adapted from [22]. Chromosome compartments, with a characteristic size of order $\sim 5 \text{ Mb}$, are defined by eigenvector expansion (also known as principal component analysis) of the contact probability map. Scores along the first eigenvector can be interpreted as a measurement of “interaction preference,” the probability the region will interact with a gene-rich region of active transcription. Here, orange corresponds to active and blue to repressed chromatin regions.
Figure 2. Schematic of topological activating domains (TADs), adapted from [22]. TADs, with a characteristic size of order $\sim 400 - 500$ Kb, are identifiable by visual inspection, when “zooming in” on the chromatin contact map. TADs fall into both active and repressive chromosome compartments, and adjacent TADs do not necessarily have opposite chromatin activity status.

Figure 3. Identification and proposed function of topological activating domains (TADs), adapted from [23]. TADs were identified by devising a simple “directionality index”: $DI = \left( \frac{B-A}{|B-A|} \right) \left( \frac{(A-E)^2}{E} + \frac{(B-E)^2}{E} \right)$, where $A$ is the number of reads that map upstream, $B$ the number of reads that map downstream, and $E$, as the average of $A$ and $B$, the expected value, based on the chi-squared statistic. Transitions from a negative to positive directionality index, or from an upstream to downstream interaction bias, are observed at the putative boundary between TADs. A Hidden Markov Model was applied to identify these biased states and infer the locations of the topological domains in the genome. This proposes a model in which TADs partition the genome into locally self-interacting regions.

We begin the study by examining, with cohesin ChIA-PET data, for the presence of local chromosome structures, analogous to topological domains, that control the transcription of cell-identity specifying genes. Topological domains are part of a hierarchy of genome structures, so we sought to identify and understand domains of lower characteristic size that act specifically to turn on gene regulatory programs. Cohesin mediates gene looping, so SMC1 was selected for ChIA-PET in embryonic stem cells. ChIP-seq data identified where proteins co-bound with interaction sites. We identify super-enhancer domains (“SDs”) and polycomb domains (“PDs”). Super-enhancer domains, surrounding super-enhancers and their target genes, contain an exceptional amount of transcriptional apparatus, and the domain boundaries insulate transcriptional activation of nearby genes. Polycomb domains surround polycomb-bound genes, marked by H3K27me3, whose expression, corresponding to repressed developmental lineages, is suppressed. Evidence for these domains are seen in more than one cell type. Deletion of the CTCF boundaries (marked by ChIP-seq peaks), leads to disruptions in the regulatory control that the domains impose [25].
3 Methods

3.1 Cell Culture

V6.5 cells mouse ESCs were grown on irradiated mouse embryonic fibroblasts (MEFs). Cells were grown on 0.2% gelatinized tissue culture plates (Sigma, G1890) in media for embryonic stem cells. ESC media contains the following: knockout DMEM (Invitrogen, 10829-018) with 15% fetal bovine serum (Hyclone, characterized SH3007103); 1000 U/mL LIF (ESGRO, ESG1106); 100 μM nonessential amino acids (Invitrogen, 11140-050); 2 mM L-glutamine (Invitrogen, 25030-081), 100 U/ml penicillin, 100 μg/ml streptomycin (Invitrogen, 15140-122), and 8 nl/ml of 2-mercaptoethanol (Sigma, M7522).

3.2 ChIA-PET Library Preparation

The protocol from [25] was used. Embryonic stem cells were grown to a minimum of 80% confluence (up to $1 \times 10^8$ cells) and were treated with 1% formaldehyde at room temperature for 20 minutes; the formaldehyde was neutralized with 0.2M glycine. The crosslinked chromatin was fragmented by sonication to sizes of 300-700 base pairs. The anti-SMC antibody (Bethyl, A300-055A) was used to enrich for Smc1-bound chromatin fragments. Some ChIP DNA was eluted; this aliquot was used to quantify concentration, and to determine the degree of environment in the pulldown procedure. For construction of the ChIA-PET library, ChIP DNA fragments were end-repaired using the T4 DNA polymerase (NEB). The fragments were separated into two aliquots, with one of two linkers (“A” and “B”, see table below) ligated to barcode the sequence. After ligation, the samples were recombined and prepared for proximity-based ligation under dilute reaction conditions in a 20 mL volume; by relying on the higher probability of a self ligation, this minimizes ligations between different DNA-protein complexes. Proximity ligation was performed with T4 (bacteriophage) DNA ligase; incubation was performed, without rocking, at 22 °C for 20 hours. During proximity ligation, DNA fragments with the same linker sequence are ligated to form paired-end tags in a single chromatin complex, producing ligation products with a homodimer linker. However, heterodimer linker products can also occur, if ligations between different chromatin fragments occur to form chimeric products, with heterodimeric linker composition. DNA fragments with heterodimeric linker sequence were removed from the final pool, because they cannot represent the ligation of between pairs of a single chromatin complex. Samples were then treated with proteinase K to digest proteins; DNA was purified from the resulting complex.

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linker A</td>
<td>3'-GAC GAC AGG CTA T(biotin) AG CGC CGG-5'</td>
</tr>
<tr>
<td></td>
<td>5'-CTG CTG TCC GAT ATC GC-3'</td>
</tr>
<tr>
<td>Linker B</td>
<td>3'-GAC GAC AGT ATA T(biotin) AG CGC CGG-5'</td>
</tr>
<tr>
<td></td>
<td>5'-CTG CTG TCA TAT ATC GC-3'</td>
</tr>
</tbody>
</table>

A digestion with EcoP15I (NEB), a restriction enzyme that cuts a set distance (27 base pairs) away from its recognition site, was then performed at 37 °C for 17 hours, to linearize the ligated chromatin fragments. This generates the 27-base-pair paired end reads obtained in this dataset. The biotinylated chromatin fragments were then immobilized on Streptavidin beads (Dynabeads M280). The chromatin fragments were then end-repaired (Epicentre #ER81050); As were added to the ends with the Klenow polymerase fragment, with rotation at 37 °C for 35 minutes. Illumina paired-end sequencing adapters were ligated to the ends; 18 PCR cycles were performed to amplify the library. PCR purification was
performed to extract the paired-end constructs, and the prepared samples were sequenced in a 50 × 50 bp Illumina HiSeq 2000 instrument. Figure 4 and 5 show schematics of the ChIA-PET workflow, adapted from [27].

Figure 4. The initial part of the ChIA-PET workflow, adapted from [27]. Chromatin is crosslinked and fragmented, a factor of interest is enriched through immunoprecipitation, the paired ends are ligated, and the crosslinks are reversed. These steps are common to all chromosome conformation capture methods (3C, 4C, Hi-C, etc.)

Figure 5. The second, method-specific part of the ChIA-PET workflow, adapted from [27]. First, before crosslink reversal, the library is aliquoted, and “A” and “B” adapters are ligated. After crosslinks are reversed, fragments are digested by a site-specific restriction endonuclease that cuts a distance away from the recognition site. Sequencing adapters are then ligated, the library is PCR-amplified, and the sample submitted for sequencing.

3.3 ChIP Data Analysis

ChIP-seq datasets were analyzed by Bowtie using the mm9 build of the mouse genome; the parameters -k 1 -m 1 -n 2 were used to report the best alignment if and only one
alignment was found, with a maximum of 2 mismatches in the seed (first 28 bases of the match, so the entire read for our dataset). MACS was used as a peak-finding algorithm to identify regions of ChIP-seq enrichment over input DNA control. A “p-value” threshold of $10^{-9}$ was used for calling enrichment for all ChIP-seq data sets; paired reads in ChIA-PET datasets can be decoupled to produce ChIP datasets. The “-nolambda” parameter is set to use a single background lambda (enriched regions, or “peaks”, are called against a background Poissonian distribution, and sometimes a local lambda can be used during peak calling); the “-nomodel” parameter is used to bypass shifting reads to their midpoint in the background model building process. Both parameters are additionally used when calling peaks with the H3K27me3 histone modification, which tends to exhibit broad signal over large regions of the genome.

### 3.4 Definition of Enhancers

As described above, co-occupancy of Oct4, Sox2, and Nanog is used to operationally define the location of enhancers; Mediator is typically associated with these sites [16]. Reads from Oct4, Sox2, Nanog, Med1, and Med12 ChIP datasets (the latter two, components of the Mediator complex) were pooled, and MACS was run (ChIP Data Analysis) on the combined dataset to define the locations of putative enhancers. Regions of enriched ChIP-seq signal falling outside promoters (defined as more than 2.5 Kb away from RefSeq transcriptional start sites) were defined as possible enhancers.

### 3.5 Assignment of Interactions to Regulatory Elements

In general, we only required a minimum overlap of 1 base pair between a tag in a paired-end tag (PET) and a DNA element (enhancers, transcription start sites, promoters, ChIP and PET peaks) to assign the tag to the DNA element.

### 3.6 Enhancer-Gene Assignments

Following previous studies [17, 23], enhancer-promoter assignments (for both super-enhancers and typical enhancers) were done by proximity. However, with ChIA-PET data, we additionally confirmed proximity assignments by high-confidence interactions. If a high-confidence interaction was not available, we looked for support in non-chimeric PETs with spans passing the self-ligation cutoff.

### 3.7 Super-enhancer Definition

Super-enhancers are defined as in Whyte, et. al. (2013) [17], using ROSE (Rank Ordering of Super-Enhancers). We describe the methodology of ROSE here: in embryonic stem cells, genomic regions cobound by Oct4, Sox2, and Nanog (by pooling the corresponding ChIP-seq reads and calling MACS peaks on the combined library) were defined as enhancers. Visual inspection of the tracks on the UCSC Genome Browser [28] reveals that certain enhancer regions are closely spaced and have high signal density. To capture these closely spaced, high signal “constituent” enhancers as a single entity, we group constituent enhancers occurring within 12.5 Kb of each other into a single, larger enhancer domain. Then, the set of all enhancers (grouped or not) ranked in a cell type in order of increasing total background-subtracted Mediator signal (Med1 ChIP; background signal obtained from whole cell extract). A normalized, rank-ordered plot of enhancer Mediator signal (Figure 6) identifies a coordinate where the first derivative exceeds unity. We define all enhancers beyond this coordinate as super-enhancers, and the remainder as typical enhancers.
Other factors or marks were considered, such as OSN, DNase I, H3K27ac, and H3K4me1. Med1 signal provided the clearest transition between the two populations, and so was chosen as the standard for annotation. In other cell types, the corresponding master transcription factors were used to define enhancers: PU.1 for pro-B cells, MyoD for myotubes, T-bet for T helper cells, and C/EBPα for macrophages. If the identity of the master transcription factor for the cell type is not available, as in neural progenitor cells, then H3K27ac signal was used instead.

**3.8 ChIA-PET Data Analysis**

ChIA-PET datasets were processed similarly to the methods of [26]. Raw sequences were analyzed for barcode composition (linker “A” or “B”), partitioning the dataset into homodimeric (“AA” or “BB”) versus heterodimeric (“AB”) ligation products, the latter of which unambiguously results from non-specific, non-local ligation events. Paired-end tags (PETs) were trimmed of the linker sequence once a complete match, up until the identifying sequence of the linkers (“CG” for linker A and “AT” for linker B), was found — in other words, a perfect match of the 10 base pairs of linker sequence. This produces 27 base pair reads per paired end, because EcoP15I cuts 27 base pairs away from its recognition sequence. The sequences were then mapped onto the mm9 genome using Bowtie using the parameters -k 1 -m 1 -v 1 (where the -v parameter indicates that at most one mismatch can be found to the entire sequence, not only the seed). Read identifiers were used to pair the reads; further, to remove the possibility of redundancy introduced by PCR artifacts, read pairs with identical coordinates and strand sign were collapsed into a single PET. PETs with ends mapping to different chromosomes (“interchromosomal PETs”) were additionally filtered from the dataset. PET peaks, representing local sequence enrichment, were called by MACS, using “-p 1e-09 -nolambda -nomodel -keep-dup=2” as parameters, where “-keep-dup=2” restricts to a maximum of 2 duplicate tags at the exact same location.

PETs that overlapped at each end, and for which each end fell into a PET peak, were grouped into putative interactions. To set a cutoff for the minimum number of PETs each

---

Figure 6. The normalized, rank-ordered plot of Mediator signal at enhancers, adapted from [17]. Total Med1 ChIP signals of enhancers, after stitching, were plotted in rank order. Super-enhancers were defined as the set of enhancers beyond the coordinate where the first derivative exceeds unity.

ROSE can be found at https://bitbucket.org/young_computation/rose.
putative interaction must have in order for an interaction between two genomic loci to be called with confidence, we examined PETs with both chimeric and non-chimeric linker sequences. For PETs with heterodimer linkers, more than 99.8% of putative interactions are singletons; 0.1% consist of 2 PETs (Figure 7). Therefore, we set a cutoff of at least 3 PETs for an interaction to be called with high confidence. This is in contrast to PETs with homodimer linkers, where at least 10% of interactions had 3 or more PETs.

Figure 7. The scaling of the proportion of intrachromosomal interactions with the number of PETs assigned to each interaction, adapted from [25]. Both PETs with chimeric and non-chimeric linker sequences are plotted. Based on the PET count distribution for chimeric linker-interactions, we require at least 3 non-chimeric linker PETs for an interaction to be called with high confidence.

As part of the ChIA-PET library preparation, DNA can not only ligate with its crosslinked partner or with another chromatin fragment; it can also self-ligate. Self-ligations will contain homodimer linker sequences. However, due to the length of each DNA fragment (< 100 base pairs), we sought to remove PETs from our dataset that fell below a minimum length cutoff. Therefore, we studied the proportion of intrachromosomal interactions scaled with PET genomic span (or the distance between the paired tags), for PETs with both chimeric and non-chimeric linker sequences. We found that while chimeric PET frequency did not scale with genomic span, nonchimeric PET frequency does. As in previous studies [26], we imposed a minimum span cutoff of 4 Kb for non-chimeric interchromosomal PETs to be included in our high-confidence set of interactions.
Figure 8. The scaling of the proportion of intrachromosomal interactions with the genomic span between paired ends, adapted from [25]. Both PETs with chimeric and non-chimeric linker sequences are plotted. We require PETs to span at least 4 Kb for an interaction to be included in the high-confidence set.

We sought to further refine the set of interactions that formed our dataset. We removed PETs where each end did not fall into different PET peaks, as those may result from self-ligation events. To check the minimum PET cutoff set previously, we applied a statistical model based upon the hypergeometric distribution to calculate the expected number of PETs per interaction under the null model. Specifically, we counted the number of PETs originating from each PET peak. We then determined, given the number of PETs originating any two peaks, the probability that we can see the observed number of PETs linking the two PET peaks. This method yielded a minimum of two independent PETs to call high-confidence interactions between pairs of interacting sites.

To understand the limit of our data, and whether we can make statements on the upper limit of interaction probabilities between certain DNA elements (e.g., enhancers and promoters), we sought to determine whether the degree of saturation within our library. To this end, we modeled the number of sampled genomic positions as a function of sequencing depth using a phenomenological Michaelis-Menten, first-order kinetic model. Intrachromosomal PETs with distance spans above the self-ligation cutoff of 4 Kb were subsampled at various depths, and the number of interactions (or unique genomic positions occupied) plotted as a function of the sequencing depth. Fitting by non-linear least squares regression, under the assumptions of this phenomenological model, suggested a sampling of approximately 70% of the available interchromosomal PET space, or 2.2/3.2 million positions (Figure 9) [25].

Figure 9. Saturation analysis of the SMC1 ChIA-PET data set, adapted from [25]. Subsampling of various fractions of PETs within the merged ChIA-PET data set was performed, and the number of unique genomic positions of intrachromosomal PETs beyond the self-ligation distance cutoff of 4 Kb was plotted. The solid line, depicting the non-linear least-squares regression fitting of data to a phenomenological Michaelis-Menten model, suggests a sampling of approximately 70% of the available intrachromosomal PETs beyond the 4 Kb in our library. The dotted line represents a putative 100% saturation.

We also sought to understand the reproducibility of our data. In collaboration with Zhao and colleagues, we produced two replicates of the Smc1 ChIA-PET. Taking only PETs
passing the self-ligation distance cutoff (4 Kb), each chromosome was partitioned in 10 Kb bins, and 21 symmetric two-dimensional matrices were constructed for the two replicates. Element $a_{i,j}$ in the matrices represented the number of PETs with one end in bin i and the other in bin j. Normalization by the number of reads, as well as the bin size $\times 1000$ bp was performed, creating a metric similar to FPKM for the bins in both matrices. The datasets reproduced well, as Pearson correlation yielded $r = 0.912$.

3.9 Definition of Super-enhancer and Polycomb Domains

Super-enhancer domains (SDs) were defined by proximity to the nearest transcriptional start site (TSS), and by confirmed ChIA-PET interactions. A region of 2.5 Kb around the transcriptional start site was taken, to form super-enhancer (SE)-gene units. On occasion, a super-enhancer could belong to more than one SD, if a super-enhancer interacted with more than one gene. CTCF-CTCF PET interactions that enclosed SE-gene units were then searched for; if nested CTCF-CTCF interactions enclosing the SE-gene unit were found, the smallest domain was taken. This leads SDs being defined as a cohesin/CTCF-cohesin/CTCF interaction surrounding, also colocalizing with cohesin, an SE-promoter interaction. All 231 super-enhancers were assigned to a total of 302 SE-gene units. In contrast, only 4128/8563 (48%) of typical enhancers were contained in CTCF-CTCF structures co-bound by cohesin [25].

Genes regulating developmentally repressed lineages are frequently bound by extended regions of the Polycomb complex, originating from the transcription start site and spanning from 2-35 Kb [10]. H3K27me3 is a histone mark present at most inactive promoters, and modulated by the EZH2 methyltransferase, the catalytic subunit of the PRC2 Polycomb complex. Therefore, we sought genes showing H3K27me3 enrichment spanning more than 2 Kb. Similar to the definition of super-enhancers, H3K27me3 domains within 2 Kb were aggregated. 546 genes showed broad H3K27me3 signal around their transcription start sites. By examining for CTCF-CTCF structures, co-bound by cohesin, that surrounded these Polycomb-bound genes, we defined Polycomb domains (PDs). Only the smallest of multiple nested CTCF-CTCF structures were selected for. By this criterion, 349 PDs, surrounding 380 Polycomb-bound genes (if a cohesin/CTCF loop surrounded more than one gene), were found [25].

4 Results

4.1 Functional Chromosome Structures

The partitioning of the genome into a hierarchy of structures (Figure 10), including topological associating domains at the 500 Kb length scale, and enhancer-promoter loops at the 10 Kb length scale (Figure 11), led us to ask whether we can identify chromosome structures, perhaps of intermediate scale, that function in the transcriptional regulation of cell state. The expression of genes within a topological domain is somewhat correlated [23], but the biological role of topological domains remains unclear. Therefore, we sought to understand, in detail, the structure and function of chromosomal loops that regulate the transcription of genes specifying cell identity.
cohesin PET peaks binding sites at TAD boundaries (Figure 14). This is evident through both direct visual inspection (Figure 12) and in meta-plots of interaction frequencies near TAD boundaries (Figure 13). Meta-plots of binding site frequencies (ChIP or PET peaks) around TAD boundaries showed an enrichment of CTCF, Smc1, and cohesin PET peaks binding sites at TAD boundaries (Figure 14).

Two cohesin Smc1 ChIA-PET datasets were generated; in combination, ~ 400 million reads were produced. Analysis of read densities showed high reproducibility between the replicates, with Pearson correlation $r = 0.912$ (Methods). Following filtering and data analysis (Methods) to remove PETs with genomic spans falling within a 4 Kb cutoff which may represent self-ligation events, we identified ~ 19 million unique paired-end tags (PETs) that were used to define PET peaks, by similar methods as those used to identify ChIP-seq binding sites (Methods). The analysis produced ~ 1 million cohesin-associated DNA loops; the majority linked enhancers, promoters, and CTCF binding sites [25]. A high-confidence dataset was produced by grouping all PETs that overlapped with a PET peak (by at least one base pair); a minimum of three PETs between any given two PET peaks was required to call an interaction (Methods). Saturation analysis by fitting to a phenomenological model suggested that our dataset did not fully sample all interacting sites in the genome.

We observe that the majority of the interactions fell within topological associating domains; this is evident through both direct visual inspection (Figure 12) and in meta-plots of interaction frequencies near TAD boundaries (Figure 13). Meta-plots of binding site frequencies (ChIP or PET peaks) around TAD boundaries showed an enrichment of CTCF, Smc1, and cohesin PET peaks binding sites at TAD boundaries (Figure 14).
Figure 12. The majority of high-confidence Smc1 ChIA-PET (blue lines) interactions fall within the boundaries of previous-defined topological associating domains. Hi-C interaction frequencies from [23] are shown at top. Figure adapted from [25].

Figure 13. A meta-plot of mean Hi-C interaction frequencies across a normalized TAD length, extending, on each side, 10% of the TAD length past TAD boundaries. Figure adapted from [25].
4.2 Super-enhancer Domains

We sought to define chromosome structures that control the expression of genes that transcriptionally regulate cell state. Super-enhancers (SEs), first described in Whyte, et. al. (2013) [17], are clusters of enhancers with high Mediator signal densities that dominate cellular transcriptional programs. Followup studies identified the cohesin and condensin II complexes as occupying super-enhancers at high densities [29]. By visual inspection, we identified, with high frequency, cohesin/CTCF-cohesin/CTCF loops that enclose a super-enhancer and its associated gene; we termed these super-enhancer domains (SDs).

We sought to characterize the basic properties of SDs (Figure 15). The SD has a characteristic length scale of $\sim 100$ Kb; they typically contain one or two super-enhancer regulated genes. Cohesin-associated interactions are often found within an SD; these are usually not associated with CTCF binding at either or both ends (Figure 16). Rather, they tend to loop different constituent enhancers within a super-enhancer, or between super-enhancers and their target genes. SDs exhibit high signal densities of Mediator, Pol II, cohesin, histone marks correlated with active transcription, and the master transcription factors of the corresponding cell type (Figure 17). The occupancy of these factors do not typically exceed the CTCF boundaries of these domains (Figure 18).
Figure 15. A schematic of the basic properties of super-enhancer domains. On the top, a super-enhancer domain is defined by a CTCF-CTCF loop that simultaneously associates with the cohesin protein Smc1. It encloses a super-enhancer, defined by Oct4/Sox2/Nanog occupancy, that loops to a target promoter that activates a gene responsible for the specification of a cell state. Note that the enhancer-promoter loop is not associated with CTCF binding on either or both ends. On the bottom, a graphical depiction of a super-enhancer domain. Figure adapted from [25].

Figure 16. Additional examples of super-enhancer domains. In general, we do not find
CTCF binding at either or both ends of Smc1 loops internal to the super-enhancer domain boundaries. Figure adapted from [25].

**Figure 17.** Occupancy metaplots of (ChIP) signal at super-enhancer domains for factors with transcriptional repressing/insulating (CTCF, SUZ12, EZH2) and activating functions (Smc1, Oct4, Sox2, Nanog, Klf4, Esrrb, Mediator, Pol II), and of histone marks associated with gene activation (H3K27ac, H3K4me3) and repression (H3K36me3, H3K27me3). Signal units are in normalized rpm (reads per million base pairs), and metaplots of DNA elements are scaled to normalized lengths and aligned to the center of the element, with extensions on either side in the indicated number of kilobases. Figure adapted from [25].
Figure 18. A heatmap representation of Mediator, Pol II, and H3K27ac signal across a normalized SD length. Signal for these factors is largely contained within SD boundaries. Figure adapted from [25].

In collaboration with Hnisz, Weintraub, and Schuijers, we tested the function of super-enhancer domains. Previous literature has attributed CTCF with insulator activity [30]; therefore, we hypothesized that CTCF at SD boundaries acted to insulate nearby genes from ectopic activation. To test this, we studied the super-enhancer domain that enclosing miR-290-295, a microRNA family that has been shown to promote the embryonic stem cell in mice. We delete one of the CTCF boundaries ($\Delta C1$) and assay, by qRT-PCR, expression in the boundary-deleted mutant. We find that the primary miR-290-295 transcript (prior to downstream RNA processing) is reduced in expression by two-fold, while neighboring gene Nirp12, positioned, with respect to the concomitant enhancer, on the same side as miR-290-295, is increased in expression by eight-fold. The expression of distal genes, AU018091 and Myadm, did not change. This leads to the model that CTCF boundaries are essential for the maintenance of super-enhancer domain integrity. Their deletion lead to a loss of insulating activity and ectopically activates the expression of nearby genes at the expense of the wild-type target promoter.
4.3 Polycomb Domains

Maintenance of cell state not only requires the expression of cell-lineage specifying genes, but additionally the repression of alternative cell lineage-specifying genes. Repressed developmental genes are characterized by the H3K27me3 mark, deposited by the Polycomb protein EZH2 [7]. Therefore, we sought to understand if there existed chromosome structures, similar to super-enhancer domains, that locally confine Polycomb-mediated repression.

Analysis of looping data in the vicinity of H3K27me3-marked sites revealed the presence of similar Smc1/CTCF-Smc1/CTCF loops that surrounded Polycomb-repressed genes, which we term polycomb domains (PDs). Similar to super-enhancer domains, polycomb domains have a characteristic length scale of ~100 Kb, and typically contain one or two H3K27me3 marked sites (Figure 20). High signal densities of concomitant chromatin modifiers SUZ12 and EZH2 — but not of master transcription factors Oct4, Sox2, Nanog; active enhancer histone mark H3K27ac; or Mediator and Pol II — are observed (Figure 21). Furthermore, the boundaries of polycomb domains coincide with the restriction of H3K27me3, SUZ12, and EZH2 signal (Figure 22).
Figure 20. A schematic of the basic properties of polycomb domains. On the top, a polycomb domain is defined by a CTCF-CTCF loop that simultaneously associates with the cohesin protein Smc1. It encloses a polycomb-repressed gene that is marked by the repressive histone mark H3K27me3. On the bottom, a graphical depiction of a polycomb domain. Figure adapted from [25].
Figure 21. Occupancy metaplots of (ChIP) signal at polycomb domains for factors with transcriptional repressing/insulating (CTCF, SUZ12, EZH2) and activating functions (Smc1, Oct4, Sox2, Nanog, Klf4, Esrrb, Mediator, Pol II), and of histone marks associated with gene activation (H3K27ac, H3K4me3) and repression (H3K36me3, H3K27me3). Signal units are in normalized rpm (reads per million base pairs), and metaplots of DNA elements are scaled to normalized lengths and aligned to the center of the element, with extensions on either side in the indicated number of kilobases. Figure adapted from [25].
Figure 22. A heatmap representation of CTCF, H3K27me3, SUZ12, and EZH2 signal in a ±10 Kb window around both Polycomb domain boundaries, for 120 Polycomb domains. Heatmap intensities correspond to a Z-transformed ChIP-seq read density signal. Signal for these factors is largely contained within PD boundaries. Figure adapted from [25].

To test the possibility that the CTCF boundaries of polycomb domains confine Polycomb-mediated gene silencing, in collaboration with Hnisz, Weintraub, and Schuijers, we deleted the CTCF boundaries of the Tcfap2e polycomb domain. Deletion of the upstream boundary (ΔC1) lead to a 1.7-fold increase in expression of Tcfap2e, while deletion of the downstream boundary (ΔC2) led a 4-fold expression increase; no significant effect was seen on the expression of neighboring genes Psmb2 and Ncdn (Figure 23). Therefore, we similarly posit that CTCF boundaries are essential for the maintenance of polycomb domain integrity.
Figure 23. The CRISPR-mediated deletion of both CTCF boundaries for the Tcfap2e polycomb domain. Smc1 ChIA-PET interactions are depicted at top, followed by tracks displaying CTCF, Smc1, and OSN ChIP signal. Deletion of the upstream (ΔC1) and downstream (ΔC2) CTCF boundaries led to the ectopic activation of target gene Tcfap2e to expression levels 1.7 and 4 fold above wild-type, respectively. No significant effect was seen on the expression levels of neighboring genes Psmb2 and Ncdn. Transcript levels are assayed by qRT-PCR and normalized to that of housekeeping gene GAPDH. Figure adapted from [25].

4.4 Cell-type Specificity of Super-enhancer and Polycomb Domains

In light of the cell-type dependence of chromosome compartments but the cell-type independence of topological domains, we sought to understand whether the cell-type specificity of super-enhancer and polycomb domains. Previous work, by Phillips-Cremins, et.
al. [30], had suggested, by 5C (a chromosome conformation capture method) and chromatin IP, that cohesin/CTCF loops are more cell-type independent than cohesin/Mediator loops, which typically represent cell-type specific enhancer-promoter interactions. This proposes a model that between syntenic regions in different cell types, we expect to see conserved cohesin/CTCF-cohesin/CTCF interaction, regardless of super-enhancer presence in the vicinity of a cell-identity specifying gene. (Figure 24).

![Figure 24](image)

Figure 24. A model depicting, between two cell types, conserved cohesin-mediated interactions between two CTCF sites, regardless of the presence of a super-enhancer in the vicinity of a cell-identity specifying gene. Figure adapted from [25].

To test this model, we used our ESC Smc1 ChIA-PET data in combination with mouse neural precursor cell (NPC) 5C data from Phillips-Cremins, et. al. [30]. Super-enhancers in NPCs were identified by the co-occupancy of NPC master transcription factors Sox2 and Brn2. At loci for which 5C and Smc1 ChIA-PET data were available for the respective cell types (Nanog and Olig1/Olig2), we saw evidence of ChIA-PET and 5C interactions at at both regions, but the cell-type specific presence of a super-enhancer driving the expression of the corresponding cell-identity specifying gene (Figure 25).

![Figure 25](image)

Figure 25. At the Nanog and Olig1/Olig2 loci, Smc1 ChIA-PET and 5C interactions are displayed for the mouse embryonic stem cell and neural precursor cell, respectively. The constitutive presence of a cohesin/CTCF loop, independent of the cell-type specificity of the corresponding superenhancer (OSN for ESCs, Sox2/Brn2 for NPCs), was observed. Figure adapted from [25].

To determine whether the constitutive presence of the cohesin/CTCF loop related to the cell-type independence of CTCF binding, we profiled the cell-type specificity of ESC CTCF...
binding sites. Using ChIP-seq data from ENCODE across 18 mouse cell types, we characterized the proportion of ESC CTCF peaks that were found in more than one cell type. We found, in general, that CTCF sites that co-bound with super-enhancer domain and polycomb domain boundaries, and more generally with cohesin-cohesin gene loops, were more cell-type independent than the total population of ESC CTCF peaks (Figure 26).

![Figure 26](image.png)

Figure 26. Cell-type specificity of embryonic stem cell CTCF binding sites, across 18 mouse cell types. CTCF peaks colocalizing with SD/PD boundaries, and more generally, with all Smc1 PET peaks, are more constitutive than the global set of CTCF peaks. Figure adapted from [25].

5 Discussion

Here we identify and describe, using Smc1 cohesin ChIA-PET, chromosome structures, of intermediate length scale (~100 Kb), that control the transcriptional of cell identity-specifying genes. Evidence provided earlier of the role of CTCF and cohesin in enhancer-promoter looping, as well as their presence at larger chromosomal structures, such as chromosome compartments and topological associating domains, suggested that they may also play a role in restricting the regulatory activity of activating super-enhancers and repressive Polycomb-group proteins [4, 7, 10, 14, 16-24, 30].

The two chromosome structures that we have identified, super-enhancer (SDs) and polycomb domains (PDs), surround super-enhancer/target gene units and H3K27me3 (Polycomb-repressed) developmentally silenced genes. They are surrounded by a cohesin/CTCF loop and, by the insulator activity of CTCF, restrict either the activating (SD) or repressive (PD) activity of super-enhancers or of EZH2, respectively. Deletion of either or both CTCF boundaries disrupt the insulating function of both domain types.

We further find that both SDs and PDs are cell-type constitutive, most likely due to the constitutive co-localization of cohesin/CTCF gene loops. Further studies should extend this work by addressing the biochemistry of the factors that regulate chromosomal loop formation, to understand the molecular mechanisms that govern the relationship between the transcriptional regulation of cell identity and local chromatin structure.

6 Appendix

The author discusses briefly the background and some attempts, using the methods of single molecule imaging, to measure the dynamics of protein clustering in the living cell.

A recent study [31], using in vivo single-molecule analysis, sought to clarify the temporal dynamics of RNA Pol II foci that form transiently in live cells. Thought to possibly represent transcriptional factories, when observed in fixed cells [32], measurement of the
lifetimes of Pol II clusters in live cells yielded an average life-time of $\tau_{on} = 5.1 \pm 0.4$ seconds. This therefore suggested a transient interaction of Pol II at active gene loci.

The author attempted to extend these studies by applying the previously developed single particle tracking, a complementary technique [33]. Key steps include detection, localization, and linking detections between frames. Technical difficulties include particle merging/splitting, high particle density, and heterogeneity in particle motion [34].

Using the following experimental parameters: 5 ms exposure time; nucleus-restricted region of interest (ROI), leading to a 14.85-16.22 ms interval between frames; 50% AOTF setting for the 561 nm laser (corresponding to a 10.6 mW output), and maximum camera gain, the author captured single particle trajectories for the Rpb1 subunit of RNA polymerase II (Figure 27: screenshot from raw movie, Figure 28: preliminary data).

An interesting question that one might consider asking is how various treatments, such as serum induction or flavopiridol (transcriptional inhibitor) addition, might alter the proportion of Rpb1 molecules that fall in bound/slow-/fast-diffusing populations.

Figure 27. Still frame from a raw movie (frame 381/1000) taken in the reproduction of single particle tracking. Scale bar spanning 100 pixels represents 1.6 mm; interval between consecutive time steps is 15 ms.
Figure 28. Preliminary single particle tracking data. On the left are example single molecule traces of a Rpb1-Dendra2 fluorophore fusion protein; the 52 longest trajectories in a single experiment are shown. On the right is a display of translocation distances for a single Rpb1 trajectory. Plotted are, for time steps \( t = [1, 12] \), the distance the particle translocated from time step \( t-1 \) to \( t \). Horizontal axis represents time step \( t \) (interval between consecutive time steps is 15 ms) and vertical axis, distance translocated in microns.

7 Funding

Work described in this thesis is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant Number 1122374. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Further funding comes from National Institutes of Health grant HG002668 to Richard A. Young. Additional funding comes from the NIH Director’s New Innovator Award and National Cancer Institute grant 1DP2CA195769 to Ibrahim I. Cissé. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

8 Acknowledgements

Graduate school has certainly been a very interesting process; I came in thinking that I would work on gene regulation, but have, over the course of a few years, realized that I am much more interested in working on something quite different. I thank the professors, classmates, co-workers, friends, and acquaintances who have helped me, in various capacities and at different times, in this long and sometimes quite challenging, but ultimately rewarding process. I suspect that there were a few others who tried to help me out at various points, but either your involvement was too subtle for me to detect, or I simply don’t know your name. I hope that the lessons that I have learned here, at MIT, will be of substantial benefit to me in future research.

An alphabetical, non-exhaustive list of people to thank, is as follows, so please don’t take offense if I accidentally forgot to include you in the list.


And of course my parents, W. Zhang and J. Inagaki.
9 Bibliography


