Abstract

Elucidating Transcriptional Regulation Using High-Throughput Sequencing, Data Integration, and Computational Methods

Raymond K. Auerbach

2012

Transcription of DNA into mRNA is a first step towards translating the genetic code of an individual into functional proteins necessary for life. By regulating transcription levels, a cell can quickly react to changes that may otherwise negatively affect its long-term viability. Understanding this process is a central theme to understanding the very fabric of molecular biology itself. High-throughput techniques combined with computational, *in silico* methods have allowed the field to advance passed the Central Dogma of molecular biology and to expand beyond the limitations of studying the mechanisms of transcription regulation at a small number of loci. Instead, transcription can now be queried on a genome-wide basis by examining transcription factor binding, mRNA abundance, chromatin remodeling, and other regulatory mechanisms. In short, the past five years have seen the rise of several new assays to probe transcription, but each assay has various biases and other nuances that must be understood to accurately analyze and interpret experiments as well as enable proper methods and tools to be developed. My graduate work has focused upon understanding these various assays and determining the best way to integrate the resulting data to gain a more holistic view of transcription regulation. This dissertation is focused upon bioinformatics approaches to study transcription regulation at multiple scales through data integration, computational

analysis, and the design of new software tools. Chapter 2 focuses on understanding the different reference DNA types that were set forth for scoring ChIP-Seq experiments and their downstream effects on calling significant regions, or peaks. Chapter 3 takes the lessons that we learned from Chapter 2 when studying ChIP-Seq and applies them to a complex with very different properties than ChIP-Seq had been applied to previously: four subunits of the human SWI/SNF chromatin remodeling complex. In Chapter 4, we analyze data generated from an RNAPII ChIA-PET experiment to hypothesize different models of transcription and how DNA folding affects the formation of protein complexes by integrating these data with other sources. Chapter 5 discusses Coupled Analysis of Polymerase Binding and Expression (CAPE), a Java program designed to identify transcripts with an unexpected relationship between transcription factor binding at the promoter and mRNA abundance. Finally, Chapter 6 concludes this dissertation with a summary and discussion of possible future directions.

Elucidating Transcriptional Regulation Using High-Throughput Sequencing, Data Integration, and Computational Methods

A Dissertation
Presented to the Faculty of the Graduate School
Of
Yale University
In Candidacy for the Degree of
Doctor of Philosophy

by
Raymond Kyle Auerbach

Dissertation Director: Mark B. Gerstein

December 2012

Acknowledgements

Pursuing a PhD is a long process that would not be possible without the abetment of many individuals. I would like to thank my dissertation advisors, Drs. Mark Gerstein and Michael Snyder, for their support and guidance over the past five years as well as the opportunity to work on several interesting projects in their laboratories. I would also like to thank Drs. Hongyu Zhao and Sherman Weissman for helpful conversations and for agreeing to serve on my dissertation committee. I would also like to acknowledge Drs. Yijun Ruan and Guoliang Li from the Genome Institute of Singapore for our collaboration on the ChIA-PET project.

I would also like to thank current and former members of the Gerstein and Snyder labs for their assistance and friendship over the years. In particular, thanks to Joel Rozowsky for never turning me away when I had a question or a concern about an analysis. I would also like to thank Lukas Habegger, Arif Harmanci, Andrea Sboner, Ashish Agarwal, Declan Clarke and Tara Gianoulis. I greatly miss Tara's enthusiasm and positive outlook on all things science, but know that her spirit lives on. She will always be remembered.

Every good bioinformaticist relies upon a star biologist with which to discuss analysis strategies, to learn how to view things from a wet bench perspective as a complement to the computational viewpoint, and last but not least, to generate data. From the Snyder lab, Ghia Euskirchen has been instrumental in many of the works presented in this dissertation and much of it would not have been possible without her tireless efforts. Philippe LeFrançois has also been an invaluable friend and resource, comparing notes and sharing in many insightful discussions in the biological and bioinformatics realms

after the rest of the Snyder lab moved to Stanford University. Both Ghia and Philippe serve as reminders as to why I enjoy working closely with biologists.

My network of friends and family during this time has been instrumental and also quite large. I would like to thank my fellow CBB class members Lukas Habegger (again), Pedro Alves, Rebecca Robilotto, and Taiwo Togun for listening and offering encouragement over the past five years. I would also like to thank my past advisors from Northern Arizona University, Drs. Paul Keim, David Wagner, Stephen Beckstrom-Sternberg, and Bernard Carey, for their continued backing and for helping me keep things in perspective during this journey. In addition, my former labmates and group members from NAU in both Biology and Computer Science have always been rooting for me. I am proud to call you all my friends as well as my colleagues.

From Yale, Lisa Sobel and Anne Nicotra have been very helpful since I first arrived. Dr. Perry Miller has always been willing to listen and offer advice. The CBB students will be losing a valuable asset when Perry moves on from his co-director role this summer. Nick Carriero and Rob Bjornson have provided indispensible high-performance computing assistance without which this work would not have been possible and have never hesitated to answer my queries quickly, even when they came at odd hours. I would also like to thank Mark Lufkin, who had nothing to do with this dissertation directly but was always willing to listen when I needed to chat with someone at Yale outside the scientific realm as well as put up with my ever-morphing schedule each week to play squash at Payne Whitney Gym.

Last but certainly not least, I would like to thank my parents for their

unconditional love and support. The PhD has been a long (and sometimes, arduous) road filled with many late nights, long meetings, and ever-changing expectations. Thanks for always being there when I needed to talk.

Table of Contents

Table of Main Figures

Table of Main Tables

List of Abbreviations

bp - base pair

ChIA-PET - Chromatin Interaction Analysis by Paired End diTag Sequencing

ChIP - Chromatin immunoprecipitation

DNA - Deoxyribonucleic acid

FISH - Fluorescence *in situ* hybridization

kb - kilobase pairs (1,000 bp)

Mb - Megabase pairs (1,000,000 bp)

mRNA - messenger RNA

PET - Paired End Tag

qPCR - quantitative real-time polymerase chain reaction

RNA - Ribonucleic Acid

RNAPII, Pol2, PolII, or Pol 2 - RNA Polymerase II

SWI/SNF - Switch/Sucrose NonFermentable

UML - Unified Modeling Language

Chapter 1: Background and Dissertation Outline

1.1 Introduction

The sequencing revolution has not been as simple as replacing older genome-wide experimental methods with newer methods. The transition to sequencing-based methods and the corresponding increase in the quality and amount of data available requires new computational analysis methods as well as renewed focus on experimental and analytical design. In my opinion, these new and exciting technologies require the roles of wet-lab biologists and bioinformaticists to shift. While the traditional perception or notion of a computational biologist has been someone who designs new algorithms and analytical methods in the background to be applied by the community, when working with new assays a computational biologist must involve himself or herself in a much earlier stage of the process and should have a working knowledge of biological underpinnings. In short, high-throughput experiments and next-generation sequencing have opened a world of possibilities, but each experimental assay has its own quirks and analysis must be customized to fit the biological question being asked. This is, in a sense, the mission of the computational biologist during the era of next-generation sequencing. Although a challenging task, the advantage is that very complex systems can begin to be teased apart. This dissertation focuses on one such complex process as its overarching theme: transcription of mRNA from DNA in the eukaryotic cell.

1.2 Understanding Transcription in the High-Throughput Age

Transcription of DNA into mRNA is a first step towards translating the genetic code of an individual into functional proteins necessary for life. Understanding this process is a central theme to understanding the very fabric of molecular biology itself. High-throughput techniques combined with computational, *in silico* methods have allowed the field to advance beyond the Central Dogma of molecular biology first described by Francis Crick (1) and expand beyond the effects of transcription at a small number of loci. Instead, now transcription can be queried on a genome-wide basis by examining transcription factor binding and mRNA abundance. Chromatin immunoprecipitation-based techniques such as ChIP-chip enable the identification of transcription factor binding sites while expression microarrays provided researchers with a quick way to compare gene expression profiles between different cells, tissues, and even organisms. Next-generation sequencing technologies such as those from Illumina, Life Technologies, and 454 Life Sciences have ushered in a new era in genomic analysis still, allowing the genome to be studied in higher resolution. Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) vastly improves upon the localization of transcription factor binding sites while RNA-sew still allows for mRNA quantification but boasts a greater dynamic range and decreased sources of error over microarrays (2).

1.3 The ChIP-Seq assay: Past, Present and Future

As most of this dissertation centers on the analysis of data from ChIP-Seq and related technologies as well as its integration with other data types to answer interesting biological questions, one must first discuss the evolution of these technologies. The

analysis of ChIP-Seq data provides several unique computational challenges even though from an experimental standpoint, this assay was a logical extension of ChIP-chip. In the early days of the ChIP-Seq technology, many different methods were attempted to obtain a "signal map" of TF binding. These methods included which reads to map to a genome (e.g. does one map all reads obtained from a next-generation sequencing study at the risk of overmapping, map only unique reads at the risk of omitting reads in repeated regions, or apply a sampling approach representing a compromise), how to call peaks (e.g. whether to represent peaks with a model such as a Poisson or negative binomial distribution or to use a non-model-based approach such as a kernel density function (3-5)), and whether to score the data against a reference sample. The resolution of ChIP-Seq mapping has also improved. When this technology was first being developed, computational algorithms assumed a standard library fragment size (usually 200 or 250 bp) for all reads. As paired-end sequencing became more practical on the new platforms, however, bioinformaticists were able to use this extra information to determine the fragment size for each read-pair empirically, which in turn improves the peak calling process. Several peak calling algorithms now support the use of paired end data. The next improvement to this assay will likely bring single-nucleotide resolution to the identification of transcription factor binding sites (6).

1.4 Extensions to ChIP-Seq

Although ChIP-Seq can identify transcription factor binding sites on a genome-wide scale, the technology cannot identify which regions of the genome are close in three-dimensional space. This information is particularly useful to the bioinformaticist as it can be used to identify potential enhancer sites, particularly when coupled with a ChIP

experiment targeted towards RNAPII, as well as to model DNA folding and to identify all regions of a genome in close 3-D proximity. The ChIA-PET assay combines a standard ChIP-Seq experiment with a proximity ligation step to retain spatial information of a DNA looping event (7). This is particularly important for studying transcription. In 3-D space, necessary transcription factors are thought to occupy certain regions of the nucleus in very high concentrations. These regions are termed "transcription factories" and are analogous to the nucleosome essentially being a large concentration of RNA Polymerase I complexes (8, 9). In this manner, DNA can actually be brought into contact with these transcription factories via DNA looping, overcoming the physical complications with the previously accepted models of transcription (e.g. the random association and disassociation of RNAPII on DNA and the slide-and-bind model (10)). ChIA-PET currently has a resolution on the order of kilobases and this will improve as the sequencing depth achievable by next-generation sequencing technologies improves.

Computationally, mapping ChIA-PET data also offers a significant challenge. The distribution of ChIA-PET paired-end tags is currently compared against a control consisting of random ligations taken from the same solution and a "true" interaction determined by both the distance between "diTags" and the depth of coverage for a particular interaction compared to the control. As with all genome-wide, long range interaction technologies, determining statistically significant interactions becomes challenging as the number of interactions increases. In addition, because the span between diTags is one of the criteria used to determine the significance of interactions it is currently not possible to effectively examine ChIA-PET interactions between chromosomes. Hi-C, another method to study long-range interactions based on the

chromatin conformation capture family of assays, is primarily used to study large-scale DNA structure because the resolution of these experiments are typically on the order of megabases (11). Hi-C has been adapted to study inter-chromosomal interactions by adding an additional biotin pull-down step to the protocol to increase the signal-to-noise ratio (12). This method should also work with ChIA-PET, although no one has yet reported any efforts along these lines. Further advancing these technologies will require collaboration between bioinformaticists and wet-lab biologists to develop improved methodologies. However, even with long-range interaction assays still in their infancy and experiments being conducted in pilot phases, the information that we glean from these assays allows us to better understand how DNA folding plays a role in transcription and to study how particular transcription factors may be recruited to target sites.

1.5 Understanding Transcription Through Data Integration

As described above, there are many new and exciting technologies that can be used to study transcription. However, with any new technology there must always be room for skepticism. One way to confirm that a technology may actually be identifying the features that it advertises is to integrate it with other, known data types. For example, an assay that is designed to identify promoters of active genes should overlap with other known signatures for promoters of active genes. In a similar manner, data integration can also be used to identify characteristics and trends of a biological system that would be missed by only examining a single experiment (e.g. the formation of TF complexes or the relationship between TF binding and mRNA abundance). As a bioinformaticist, recognizing and respecting the underlying biology behind various assays is vital to

designing an effective analytical strategy. In that vein, several different data types are described in this dissertation.

The identification of transcription factor binding sites, histone modifications, and chromatin structure are important prerequisites to discuss, particularly as they work together to initiate transcription. Proteins such as those comprising the RNA polymerase II (RNAPII) complex and other factors that can stabilize the RNAPII complex at the promoter region to enhance transcriptional efficacy, but first chromatin must be in an open conformation in order for transcription factors to bind. Histone modifications at the promoter such as H2A.Z, H3K4me3, and H3K27ac are hallmarks of genes that are poised to be actively transcribed or are being actively transcribed (13-15). These histone modifications mark regions of open chromatin where transcription factors and other proteins can actively bind. Conversely, transcription factors cannot bind as easily in regions where chromatin is wound very tightly, also known as heterochromatin. These regions are typically marked by H3K27me3 modifications, which are thought to be established through the actions of the Polycomb complex methylating lysine-27 of the H3 histone subunits. These actions are in direct competition of the Trithorax group proteins, which are thought to be responsible for establishing euchromatin-associated histone marks by demethylating lysine 27 of H3 and by methylating lysine 4 of H3. In addition to identifying transcription factor binding sites and histone modifications, there are also protein complexes that help to maintain or change the structure of chromatin. Chromatin remodelers can act to regulate access of transcription factors and other proteins to chromatin by converting heterochromatin regions to euchromatin and visa-versa (16). Due to their unique binding patterns compared to traditional transcription factors and

their status as large, multi-subunit complexes, analyzing these types of factors offers a particular challenge. Other features associated with regions of open chromatin include DNase hypersensitive sites and the results from the Formaldehyde-Assisted Isolation of Regulatory Elements assay (FAIRE) (17).

1.6 Dissertation Outline

With the prerequisites dispatched, our focus now turns to the organization of this dissertation. As stated, the dissertation is focused upon bioinformatics approaches to study transcription at multiple scales. There are six chapters, including this introduction.

1.6.1 Chapter 2

Chapter 2 focuses on understanding the different reference DNA types that were set forth for scoring ChIP-Seq experiments and their downstream effects on calling significant regions, or peaks. Rozowsky et al (3) was the first paper to propose the use of a reference DNA type for scoring ChIP-Seq data, but it did not consider the spectrum of reference samples that could be used in ChIP-Seq including Input DNA, a mock IgG control, and MNase-digested DNA.

In addition to the rationale given in the chapter's introduction, the reader should note that the experimental and computational work for this project was begun in the early days of short-read sequencing using the Solexa/Illumina technology. Analyzing ChIP-Seq data was a new challenge and the idea of using a reference DNA type to aid scoring had just been introduced. In the beginning, labs used a completely computational approach by generating a randomized background against which ChIP DNA could be scored. Like most purely computational approaches developed based on incomplete assumptions,

however, randomized background turned out to be a very poor choice because it does not take into account biases introduced by chromatin structure. Desiring a true experimental control, laboratories began to experiment with Input DNA or fractionated DNA that is prepared alongside ChIP DNA and sonicated but that never undergoes a chromatin immunoprecipitation step. It was thought that this would essentially produce a randomized background, as sonication was assumed to be a random process. Upon examining the Input DNA control mapped against a reference genome, however and to much surprise, stark peaks were observed in the Input DNA control. Through data integration and by designing a strategy combining experimental and computational approaches, my coauthor and I characterized this phenomenon. This work was the first to show that input DNA, in what had become the *de facto* standard reference for ChIP-Seq analysis conducted by the ENCODE Consortium, was actually biased towards a subset of deprotected chromatin near regions with proteins bound nearby. This bias manifests itself in high signals near 5' ends of genes and other regions, which can have a direct effect on the peaks identified by ChIP-Seq peak calling algorithms. It also shows that input DNA as a reference type is particularly problematic for ChIP-Seq experiments targeting factors that bind in these regions such as RNA Polymerase II and chromatin remodeling complexes such as SWI/SNF (discussed in Chapter 3). In short, one must understand that the biases in the experimental control may not necessarily correspond to the biases inherent to the ChIP DNA depending upon the characteristics of the protein being queried. For this reason, effective ChIP-Seq analysis will likely never be fully reduced to the domain of a push-button, computational tool.

The findings presented herein comprise a major reason why our lab switched from using input DNA as our preferred reference DNA type to a mock IgG control (which our paper also examined and discussed for the first time). For human data, the use of a mock IgG control instead of input DNA has become fairly standard practice. For ChIP-Seq in general, scoring against a reference sample to correct for the phenomena we observed has also become common and depending upon the antibody target, our observations have a profound effect on the biological interpretation of ChIP-Seq data.

This work was published in the *Proceedings of the National Academy of Sciences* in 2009 by Auerbach and Euskirchen, et al., and has been well-cited (18).

1.6.2 Chapter 3

Chapter 3 takes the lessons that we learned from Chapter 2 and applies them when studying ChIP-Seq to a complex with very different properties than ChIP-Seq had been applied to previously: four subunits of the human SWI/SNF chromatin remodeling complex. In addition to the rationale given in the chapter's introduction, the reader should note that the experimental and computational work for this project was begun in the early days of short-read sequencing using the Solexa/Illumina technology. The process of ChIP-Seq peak calling was still new and the best practices for determining a high confidence list were still under investigation. In addition, at the time most of the ChIP-Seq assays being run by members of the ENCODE Consortium were targeted toward RNAPII. The RNAPII antibody used (8WG16) is considered to be one of the strongest ChIP-grade antibodies available and is not indicative of the data quality one should expect from a ChIP-Seq experiment using a different antibody. In addition, as discussed in Chapter 2 the control DNA types being used for scoring were biased towards promoter

regions. SWI/SNF is a general chromatin remodeler and represented one of the early efforts to apply ChIP-Seq to a non-promoter associated factor within the ENCODE Consortium. To attack this problem, we designed a set of criteria combining several parameters including but not limited to the q-value output by peak callers (which at the time was the sole method that many labs used to determine peak quality), the ratio of ChIP tags to control tags, and the number of tags present in each experiment. Data integration played a major role on several fronts, as well. Since ChIP-Seq was performed against four subunits of the SWI/SNF complex independently and each subunit may have roles outside of SWI/SNF, my coauthor (who conducted the experiments) and I devised one of the early strategies to examine ChIP-Seq data in the context of a protein complex. We introduce the concept of a ChIP-Seq "domain" in scoring, i.e. a region where a complex is likely to be bound in some form, and use this information to infer how chromatin remodeling works on a genome-wide scale in HeLa cells. Finally, as SWI/SNF is a chromatin remodeling complex with a bromodomain that binds to acetylated histones at promoter regions (16), our findings provide a genome-wide bridge between genomic and epigenomic factors that affect transcription.

This work was published in *PLoS Genetics* in 2011 by Euskirchen and Auerbach, et al (19).

1.6.3 Chapter 4

One very interesting aspect of the work described in Chapter 3 was how subunits of the SWI/SNF complex can appear in disparate locations in 2-D space but can be brought into close proximities in 3-D space via DNA folding, possibly completing the complex. Applying this thought process to study the transcriptional machinery was a logical

extension. In Chapter 4, we analyze data generated from an RNAPII ChIA-PET experiment to hypothesize different models of transcription and how DNA folding affects the formation of protein complexes by integrating these data with other sources. As mentioned above, the transcription factory theory had been gaining traction in the field supported by other biological assays such as DNA fluorescent *in situ* hybridization (DNA-FISH). Several questions remained to be explored on a genome-wide scale, however, including possible models for transcription and how subunits are recruited for key protein complexes involved in transcription. In the case of the former question, we specifically examined whether genes are typically transcribed on a single-gene basis as opposed to being transcribed as part of a multi-gene transcription factory, how these different models affect the mRNA abundance, and whether transcription factories can act as essentially a eukaryotic operon. For the latter question, we examined whether subunits of known protein complexes for which ChIP-Seq data was publicly available are typically found near promoters or if subunits are actually being recruited to the promoter via DNA folding. This distinction is important for two reasons. First, DNA looping is how distal enhancers are thought to work, as the binding of distal enhancer proteins to promoter proximal proteins stabilizes the RNAPII complex, resulting in increased transcription efficacy. Second, given the limitations of the ChIP-Seq assay and its lack of spatial information, ChIP-Seq signal would appear at both proximal and distal sites for the same instance. By deciphering which proximal and distal regions are linked, it becomes possible to hypothesize whether a transcription factor is more likely to be found proximally or distally to promoters.

In addition to the rationale given in the chapter's introduction, the reader should note that the experimental and computational methods to study long-range interactions on a genome-wide scale are still in their infancy. Some whole-genome chromatin conformation capture methods such as Hi-C have proven useful to examine DNA folding on a macro level, but lack the resolution at this time to effectively study folding between individual factors. Conversely, although ChIA-PET can be used to examine all interactions involving a particular transcription factor or protein, the method falls victim to all of the limitations of a typical ChIP experiment. Computationally, the algorithms and statistics behind Hi-C and ChIA-PET are still under active development.

This project was particularly satisfying as a way to use bioinformatics to bridge data generated by the ENCODE consortium with systems biology and DNA folding, as well as learn more about how transcription works in 3-D space. We propose several models of transcription based on our findings and also look at recruitment of individual factors to form a protein complex at the promoter. Surprisingly, several subunits known to form protein complexes were actually completed by subunits recruited from distal sites. We observed this in SWI/SNF based on chromatin capture data, as described in Chapter 3, but this allowed us to quickly look at all transcription factors available as part of ENCODE. These findings will undoubtedly have a profound effect on modeling TF binding and the formation of protein complexes, as the spatial component would have been missing with a standard ChIP-Seq experiment. Whether used to predict new, putative enhancer sites computationally or to confirm predictions experimentally, we expect that this work lays the foundation for an exciting new type of analysis and will further future work in both the biological and computational arenas.

This work was published in *Cell* in 2012 by Li, Ruan, Auerbach, and Sandhu, et al (20).

1.6.4 Chapter 5

The penultimate chapter of this dissertation discusses Coupled Analysis of Polymerase Binding and Polymerase (CAPE), a Java program designed to identify transcripts with an unusual relationship between transcription factor binding at the promoter and mRNA abundance. Typically, the amount of RNA Polymerase II binding at the promoter is correlated to the normalized depth-of-coverage obtained from RNA-seq experiments. Some transcripts, however, deviate from this correlation for biologically relevant reasons. For example, a transcript that exhibits high levels of RNAPII binding and low levels of mRNA transcript abundance is a hallmark of a stalled or poised promoter. Transcripts with stalled promoters are often transcribed as a result of some external stimulus or change to the cell that requires a very quick response. In one manner, one can think of RNAPII stalling as similar to the use of different sigma factors in the bacterial RNA Polymerase complex. In bacteria, different sigma factors target different promoter recognition sites (21). By swapping sigma factors, a cell's transcriptional program can quickly respond to conditions such as heat shock by controlling how many RNA polymerase complexes are actively transcribing heat shock genes (21, 22). Eukaryotic transcription is a bit more complicated, but by RNAPII becoming poised at promoters the transcriptional machinery is in position to quickly increase transcription levels of the affected transcripts. Transcripts present high abundance but with low RNAPII binding levels at the promoter are also interesting, as they can indicate an issue

with the annotation (e.g. the misannotation of the promoter region), an experimental or mapping anomaly, or a transcript that is transcribed by a complex other than RNAPII.

As high-throughput sequencing continues to become more accessible and as large consortia such as ENCODE/modENCODE and the Epigenetics Roadmap continue to release data sets to the public, certain matched experiments have become common. One such experiment pair is ChIP-Seq targeting RNAPII coupled with RNA-seq. RNAPII represents the ideal ChIP experiment due to the efficacy of the antibody and to the role of RNAPII during transcription. RNA-seq is a simple yet powerful assay to obtain transcript abundance. Together, this experiment pair can be used to study differences in transcriptional programs between organisms, between developmental stages within the same organism, or between cells exhibiting normal and diseased phenotypes (23). Despite the power inherent in this particular pairing of experiments, the community currently lacks a tool to quickly and simply summarize the results of these two experiments together. CAPE fills this niche.

CAPE is written in Java and is designed to run from the command-line of a computer for which version 1.6 or higher of the Java Virtual Machine is available. A conscious decision was made to avoid the trend of web services in this case, as the signal files produced by ChIP-Seq and RNA-seq experiments can be fairly large and uploading these files to a web service can become prohibitive. Given an annotation set, a list of mRNA abundance measurements such as reads per kilobase per million mapped reads (RPKM), and a ChIP-Seq signal track, CAPE calculates the mRNA abundance with respect to ChIP-Seq binding for each transcript, identifies transcripts with unusual relationships of binding relative to mRNA abundance, and produces a summary report. A

set of default metrics for determining unusual relationships between binding and transcription is provided, but CAPE also allows a user to override these default metrics to refine the analysis for his or her particular question or organism of interest. In addition to cataloguing this relationship for a single experiment, CAPE can also analyze multiple experiment pairs and identify transcripts in which the relationships between ChIP-Seq binding and mRNA abundance change. This is particularly applicable to analyses designed to examine changes during a developmental time course or to identify differences between organisms. In fact, CAPE will also automatically limit its comparison to transcripts from orthologous features if a list of orthologs is provided to the comparison tool. We expect that CAPE will be well-received by the community.

This work was submitted to *Bioinformatics* for review at the time of dissertation composition.

1.6.5 Chapter 6

Chapter 6 concludes this dissertation with a summary and discussion of possible future directions.

1.7 References

1. Crick F (1970) Central dogma of molecular biology. *Nature* 227:561–563.
2. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10:57–63.
3. Rozowsky JS et al. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* 27:66–75.
4. Zhang Y et al. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9:R137.
5. Valouev A et al. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* 5:829–834.
6. Rhee HS, Pugh BF (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* 147:1408–1419.
7. Fullwood MJ et al. (2009) An oestrogen-receptor-alpha-bound human chromatin

interactome. *Nature* 462:58–64.
8. Cook PR (2010) A Model for all Genomes: The Role of Transcription Factories. *Journal of Molecular Biology* 395:1–10.
9. Xu M, Cook PR (2008) Similar active genes cluster in specialized transcription factories. *J Cell Biol* 181:615–623.
10. Cook PR (1999) The organization of replication and transcription. *Science* 284:1790–1795.
11. Lieberman-Aiden E et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326:289–293.
12. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol* 30:90–98.
13. Zhou VW, Goren A, Bernstein BE (2011) Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* 12:7–18.
14. Barski A et al. (2010) Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. *Nat Struct Mol Biol* 17:629–634.
15. Barski A et al. (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129:823–837.
16. Clapier CR, Cairns BR (2009) The Biology of Chromatin Remodeling Complexes. *Annu Rev Biochem* 78:273–304.
17. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* 17:877–885.
18. Auerbach RK et al. (2009) Mapping accessible chromatin regions using Sono-Seq. *Proc Natl Acad Sci USA* 106:14926–14931.
19. Euskirchen GM et al. (2011) Diverse roles and interactions of the SWI/SNF chromatin remodeling complex revealed using global approaches. *PLoS Genet* 7:e1002008.
20. Li G et al. (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 148:84–98.
21. Gruber TM, Gross CA (2003) Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu Rev Microbiol* 57:441–466.
22. Sharma UK, Chatterji D (2010) Transcriptional switching in Escherichia coli during stress and starvation by modulation of sigma activity. *FEMS Microbiol Rev* 34:646–657.
23. Gerstein MB et al. (2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* 330:1775–1787.

Chapter 2: Understanding the Choice of Reference DNA Type in a ChIP-Seq Experiments and its Effects on Downstream Computational Analyses

2.1 Statement of Prior Publication and Bioinformatics Contributions

This work is reprinted from a 2009 paper in the *Proceedings of the National Academy of Sciences* by Auerbach and Euskirchen, et al., and has been well cited [25]. No additional permissions were required. As the first comprehensive analysis of ChIP-Seq reference DNA types used for peak scoring, this work exposed the effects of local chromatin structure on downstream scoring results. Peak scoring for ChIP-Seq was initially performed *in silico* using a randomized background. Desiring a biological control for scoring, the field began to use input DNA for this purpose expecting that it would produce a random distribution of reads; however, this paper shows that this is often not the case. As a consequence, peak-calling algorithms will produce biased results that depend upon the characteristics of the factors being ChIPped and the reference DNA type used for scoring. Understanding these biases is central to the proper interpretation of a ChIP-Seq experiment. Using a very early version of the PeakSeq algorithm, I designed and implemented the analysis strategy used in this paper, integrated our results against several other features from both published and unpublished sources, and interpreted the results. This included all processing and scoring of the data, identifying and pre-processing external data sets for comparison, designing a strategy to score reference

DNA types against a "reference of references" (for example, identifying parameters that needed to be altered from those used for scoring traditional ChIP-Seq experiments), comparing our results against similar assays such as FAIRE, and interpreting the characteristics of each reference DNA type in a system-wide context.

2.2 Abstract

Disruptions in local chromatin structure often indicate features of biological interest such as regulatory regions. We find that sonication of crosslinked chromatin when combined with a size selection step and massively parallel sequencing can be used as a method (Sono-Seq) to map locations of high chromatin accessibility in promoter regions. Sono-Seq sites frequently correspond to actively transcribed promoter regions as evidenced by their co-association with RNA Polymerase II ChIP regions, transcription start sites, histone H3 lysine 4 trimethylation (H3K4me3) marks, and CpG islands. The pattern of breakage by Sono-Seq overlaps with, but is distinct from, that observed for FAIRE and DNase hypersensitive sites. Our results demonstrate that Sono-Seq can be a useful and simple method for mapping many local alterations in chromatin structure. Furthermore, our results provide insights into the mapping of binding sites using ChIP-Seq experiments and the value of reference samples that should be used in such experiments.

2.3 Introduction

The accessibility of regulatory elements in chromatin is critical for many aspects of gene regulation. Nucleosomes positioned over regulatory elements inhibit access of transcription factors to DNA; deprotection of the DNA arises from local changes in

chromatin conformation. Previous methods for mapping chromatin accessibility include mapping DNase I hypersensitivity sites or FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) regions and analyzing the DNA using microarrays or DNA sequencing [1-3]. These methods have mapped many open chromatin sites to promoters of actively transcribed genes as well as to enhancers.

The in vivo mapping of regulatory elements is often performed by chromatin immunoprecipitation (ChIP) of a factor of interest followed by analysis of associated DNA [4-6]. Chromatin complexes are preserved through cell fixation with formaldehyde, fragmentation of the chromatin, and isolation of protein-bound DNA regions using antibodies to a specific DNA-associated protein. DNA fragments are purified and used to probe DNA microarrays (ChIP-chip) or, more recently, identified by high throughput DNA sequencing (ChIP-Seq) thereby locating transcription factor binding sites (TFBSs) on a genome-wide scale [4-7]. In ChIP experiments, significant targets representing binding regions are found by analyzing signal levels produced by an experimental sample relative to a reference sample. Although several automated scoring algorithms exist for ChIP-Seq data [6-11], an appreciation of the characteristics and biases inherent to different reference DNA samples and preparation methods is important for understanding the significance of the results obtained.

In the work presented here, we examine the signal distributions of commonly used reference samples including sonicated chromatin and investigate the aggregate signals relative to annotated regions (Table 2.1). We show that even without immunoprecipitation, crosslinked chromatin fragments can be size-selected for novel chromatin regions and many of these regions are proximal to promoters. We investigate

the causes of these signals and develop this observation as a method for mapping these chromatin domains.

Table 2.1 Data sources

Library	Size selection	Cell conditions	Sonication conditions	Uniquely Mapped Reads	Biological Replicates
RNA Pol II (HeLa S3)	100-350 bp	formaldehyde-crosslinked	7 x 30 sec	29,060,928	3
Sono-Seq (HeLa S3)	100-350 bp	formaldehyde-crosslinked	7 x 30 sec	29,840,987	3
Sono-Seq (HeLa S3)	350-800 bp	formaldehyde-crosslinked	7 x 30 sec	19,729,371	3
Naked DNA	100-350 bp	not crosslinked	1 x 30 sec	34,550,812	3
Normal IgG (mouse, HeLa S3)	100-350 bp	formaldehyde-crosslinked	7 x 30 sec	28,960,961	2
MNase (HeLa S3)	100-200 bp	not crosslinked	not sonicated	20,924,734	2

2.4 Results

2.4.1 Sonicated chromatin fragments reveal peaks over promoter regions

While examining the signal tracks of reference DNA samples for ChIP-Seq, we observed the presence of "peak" regions that appeared to have greater signal relative to the genome as a whole (Figures 2.1, A.1 and A.2). Sonicated chromatin was prepared from nuclear lysates of formaldehyde-crosslinked HeLa S3 cells and either subjected to chromatin immunoprecipitation with a specific antibody to RNA polymerase II (ChIP DNA) or DNA was purified without immunoprecipitation ("Input" or "Sono-Seq" DNA) (Figure 2.2). Both preparations of DNA were size selected for 100-350 bp fragments and converted to libraries for sequencing on the Illumina Genome Analyzer II platform. 29.0 M uniquely mapped reads were obtained for RNA polymerase II and 29.8 M reads for Sono-Seq.

The peaks in sonicated chromatin are often similar, albeit of lower magnitude, than those obtained from the Pol II ChIP-Seq experiment. For HeLa S3 cells 106,958 Sono-

Seq DNA peaks are observed as compared with 49,377 peaks for Pol II ChIP-Seq with a total coverage for the peaks of 27.7 and 36.7 Mb, respectively. Filtering for strong targets (see Materials and Methods) reduced the data set to 27,773 peaks in Pol II and 21,762 peaks in Sono-Seq DNA. Using these strong targets, 65.0% of Sono-Seq regions are within 1 kb of a Pol II region and 49.4% of Sono-Seq regions are within 2.5 kb of a 5 end of an Ensembl gene. To further investigate Sono-Seq characteristics we examined its aggregated signal over the proximal promoter regions of expressed and non-expressed Ensembl genes [12]. In general, Sono-Seq DNA displays elevated signal at the 5′ ends of Ensembl genes compared to background (see Materials and Methods; Figures 2.3A-C). Sono-Seq DNA enriched regions heavily overlap with those of Pol II (Figure A.3); however not all Sono-Seq regions co-occur with Pol II. Locations of Sono-Seq DNA peaks were intersected against Pol II peaks and we found 6,892 peaks where no corresponding Pol II peaks were identified within 1 kb. We found that some of these unique Sono-Seq peaks correspond to HeLa-derived small (< 200 nucleotides) RNAs [13] as shown in Figure A.4.

2.4.2 Sono-Seq requires crosslinked chromatin

The signals from Sono-Seq DNA could either be due to regions of preferential breakage intrinsic to the DNA sequence or breaks that occur in regions made accessible by biological activity. To further investigate the source of Sono-Seq DNA signal, we prepared HeLa S3 genomic DNA from non-crosslinked, deproteinized cells and sonicated the DNA into fragments of 100-350 bp on average to produce “naked DNA” (Figure 2.2). Naked DNA did not show visible peaks either in signal tracks or over promoter regions and examination of its aggregated signal near transcription start sites did not reveal any

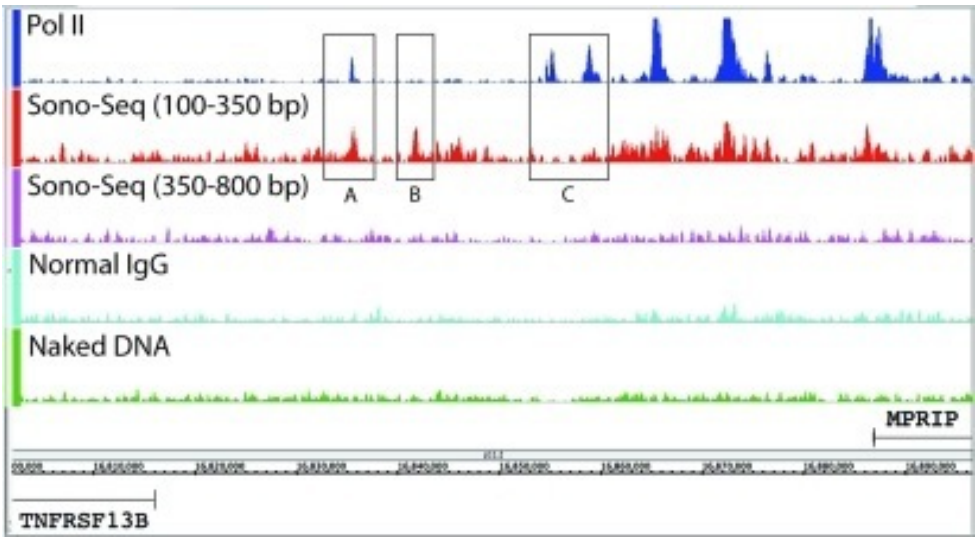


Figure 2.1 Signal map. Signal map showing Pol II ChIP DNA, Sono-Seq DNA (small and large fragment sizes), normal IgG, and naked DNA. All signals are in HeLa S3 cells. Signal levels between positions 16,802,000-16,896,000 of chromosome 17 are shown. Tracks are scaled based upon the number of uniquely-mapped reads obtained for each sample type. Both TNFRSF13B and MPRIP are not expressed in HeLa S3 based on RNA-Seq data (14). Several regions of disagreement between Sono-Seq and Pol II signal are shown, such as A) a large Sono-Seq peak with a less pronounced Pol II peak, B) the absence of a Pol II peak and the presence of a Sono-Seq peak, and C) Pol II peaks without corresponding Sono-Seq peaks.

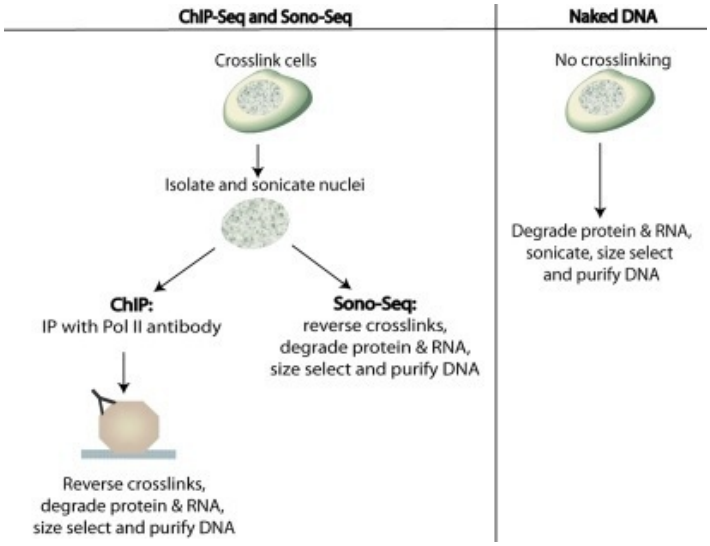


Figure 2.2 Steps to prepare ChIP DNA, Sono-Seq DNA, and naked DNA.

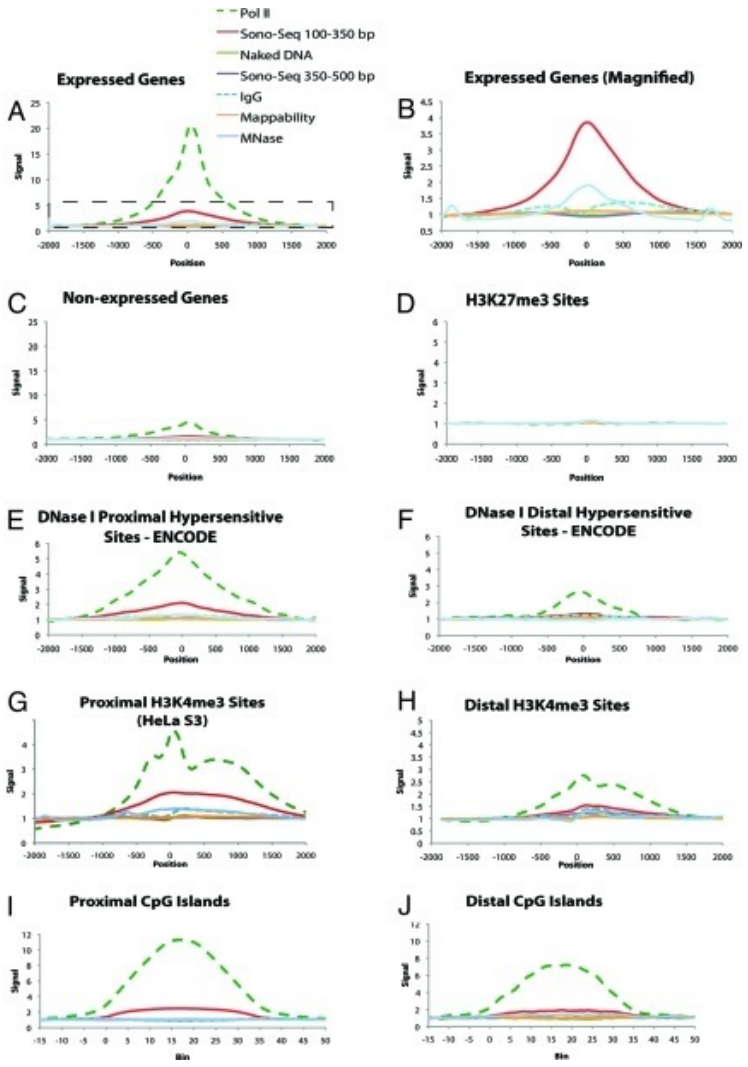


Figure 2.3 Aggregation plots depicting average ChIP signal across a variety of genomic features. For plots E-J, proximal is defined as lying within ±2.5 Kb of an Ensembl gene. Values are given in "fold enrichment" compared to a background signal (see Materials and Methods). A value of 1.0 indicates enrichment equal to background. Signal was calculated in Pol II, two different size selections of Sono-Seq DNA, naked DNA, normal IgG, and MNase-digested DNA. Mappability is a measure of how well reads mapped to the features being compared (see Materials and Methods). A mappability of 1.0 indicates an equal mappability level as background. Panel B is a magnified view of the regions enclosed by the dotted box in panel A in which Pol II is removed and scales are altered to allow for better comparison between reference sample types. Vertical axis units are consistent between all plots. Horizontal axis units are given in nucleotides from the feature start site in plots A-H and in bins each representing 1/35th of the feature size in plots I-J. In all figures, position/bin 0 corresponds to the start of the target feature.

enrichment at these regions (Figure 2.1). These results indicate that Sono-Seq peaks require crosslinked chromatin, presumably because crosslinking preserves the in vivo state of DNA.

As an additional control we prepared DNA according to the exact protocol as ChIP

DNA, but substituted affinity-purified IgG from a non-immunized animal for the antibody that recognizes Pol II. We call this data set "normal IgG". Interestingly examination of the aggregated peak signals indicates that normal IgG signal is near baseline (0.9-fold) over TSSs (Figures 2.3A-C), lower than that of Sono-Seq DNA. We also prepared DNA in which chromatin from non-crosslinked cells was treated with micrococcal nuclease (MNase). MNase-treated DNA exhibits elevated signals over promoters analogous to Pol II and Sono-Seq DNA signals (Figures 2.3A-B).

2.4.3 Sono-Seq DNA peaks reside over expressed promoters

Using HeLa S3 expression data determined from an RNA-Seq experiment [14], we examined Sono-Seq and Pol II ChIP aggregate signals over genes that are expressed in HeLa S3 cells as well as those that are not expressed. We define an expressed gene as having an average coverage of at least one-fold across each nucleotide in a gene. The remaining genes were classified as non-expressed. 10,993 genes are expressed and 19,273 genes are non-expressed using these criteria. Aggregated signals from Sono-Seq DNA are enriched 4-fold over expressed genes (Figures 2.3A-B). MNase-treated DNA also gave a signal over 5′ ends of expressed genes as expected, indicating that open chromatin is present in these regions. We found that 31.8% of all Ensembl genes and 67.9% of all expressed Ensembl genes in HeLa S3 possess a significant peak in Sono-Seq DNA proximal to the 5′ ends.

To ascertain relationships between peak significance and gene expression, we created rank-order lists for Pol II and Sono-Seq DNA peaks by sorting peaks first by tag count followed by fold-enrichment over the corresponding signal in naked DNA. We then calculated the percentage of peaks occurring in promoters of expressed and non-

expressed genes as well as those occurring distal to promoter regions. The top of the list (i.e. the most significant peaks) contains a large percentage (90%) of peaks found in 5′ ends of expressed genes (Figure 2.4). Toward the bottom of the rank-order list, the percentage of enriched regions found proximal to 5′ ends decreases while the percentage of enriched regions found distal to 5′ ends of genes increases. The percentage of enriched reads proximal to 5′ ends of non-expressed genes remains consistent throughout the data set.

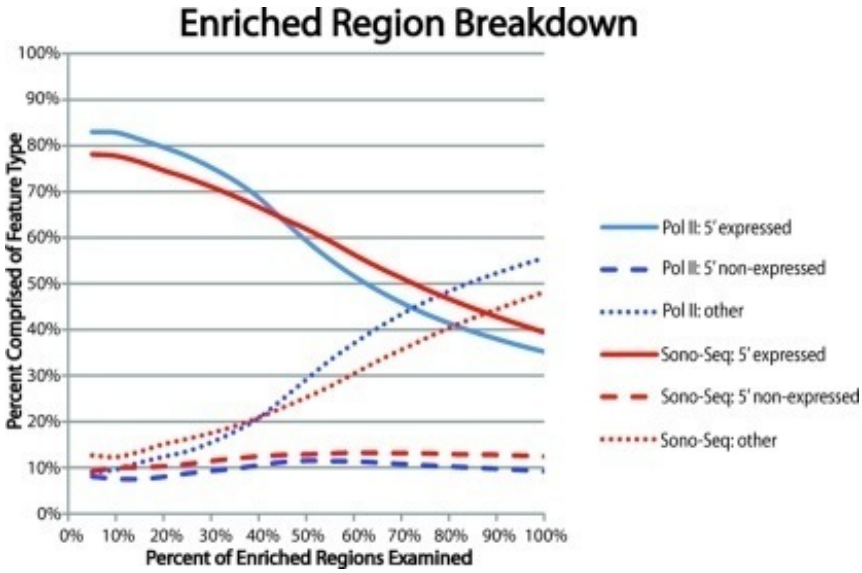


Figure 2.4 Rank order plot. A rank-order plot depicting the percentage of Sono-Seq and Pol II enriched regions located proximal and distal to genes (see Materials and Methods). Regions most highly enriched by Sono-Seq typically lie proximal to the TSSs of expressed genes. Enrichment over promoter regions of non-expressed genes remains constant whereas enriched regions lying distal to known promoter regions are ranked lower (i.e. have lower scores).

2.4.4 Sono-Seq DNA signals are enriched over other markers associated with gene expression

To further explore Sono-Seq signals we compared the Sono-Seq peaks to several other published chromosomal features, including DNase I hypersensitive sites, H3K4me3

sites and CpG islands (Figures 2.3E-J). Using the ENCODE region data of Crawford et al. [2] we selected 2,060 DNase I hypersensitive sites from HeLa cells with a q-value < .05. Of these, 958 were proximal (within 2.5 kb) and 1,103 were distal (greater than 2.5 kb) to the TSSs of Ensembl genes. Aggregation of signals over proximal DNase I hypersensitive sites reveals Pol II signals increase more than 5-fold over these regions. Signal elevation over proximal DNase I hypersensitive sites is also evident for small-fragment Sono-Seq DNA (2-fold). Signal is not enriched in naked DNA.

Mapping Sono-Seq DNA relative to distal DNase I hypersensitive sites reveals a different pattern. Pol II signal is modestly elevated over these distal regions (2.5-fold). Small-fragment Sono-Seq DNA, normal IgG, and naked DNA all show minimal signal elevation over distal DNase I hypersensitive sites. Thus, Sono-Seq DNA signals are preferentially located over proximal DNase I hypersensitive sites as compared to distal ones.

To further investigate the Sono-Seq signal at promoters with proximal DNase I hypersensitive sites, we examined the association of Sono-Seq peaks with H3K4me3 sites, which are also correlated to gene expression level and promoter localization [15]. For these analyses we aggregated Sono-Seq signals over two different genome-wide H3K4me3 ChIP-Seq data sets: one containing a total of 54,467 H3K4me3 sites from HeLa cells [16] and another containing a random sample of 100,000 H3K4me3 sites from CD4+ cells [17]. Aggregation of signals over either source of the H3K4me3 sites revealed that the Pol II and Sono-Seq DNA signals are significantly elevated at H3K4me3 sites (Figures 2.3G-H; Supp. Figure 2.5). For the H3K4me3 sites identified in HeLa S3 cells, separate aggregations were performed for sites located distal and proximal

to Ensembl genes. For proximal H3K4me3 sites, Pol II signal is elevated 4.5-fold and Sono-Seq DNA signal is elevated 2-fold. Normal IgG produced a lower enrichment signal (1.3-fold) whereas signal was not elevated in naked DNA. Enrichment over distal H3K4me3 sites identified in HeLa S3 cells drops to 3-fold and 1.5-fold for Pol II and Sono-Seq DNA, respectively. Other reference samples exhibit signals comparable to those observed over proximal H3K4me3 sites.

Finally, we also analyzed Sono-Seq signals relative to CpG islands, which are associated with promoter regions [18]. For this analysis we used a coordinate list of unique CpG islands represented on the Illumina Infinium HumanMethylation27 BeadChip Assay (Illumina, San Diego, CA). The CpG islands on this array have a mean size of 1,388 bp and query 11,471 unique CpG islands. Of these regions, 7,101 sites lie within 2.5 kb of TSSs of expressed Ensembl genes whereas 4,029 lie within 2.5 kb of TSSs of non-expressed Ensembl genes. We observe signal enrichment over these CpG islands in Pol II, Sono-Seq (150-350 bp), and MNase-digested DNA. Other reference DNA types remain unenriched over these CpG islands (Figures 2.3I-J).

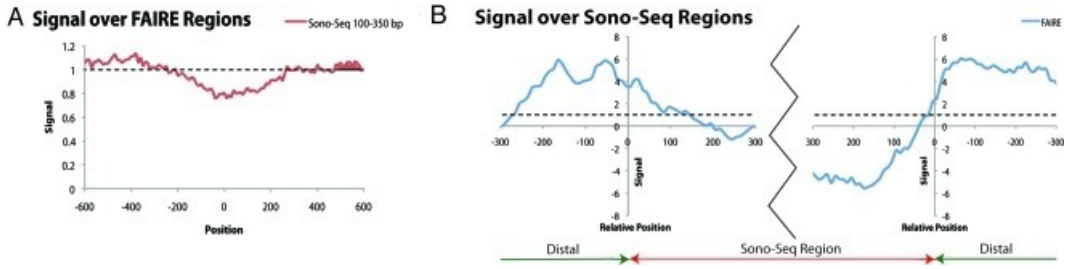


Figure 2.5 Sono-Seq and FAIRE aggregation plots. A) Aggregation of signal from yeast Sono-Seq DNA (selected at 100-350 bp) over regions enriched in yeast FAIRE. Sono-Seq signal is depressed over regions enriched by FAIRE. **B)** Aggregation of FAIRE DNA signals over regions enriched in Sono-Seq. FAIRE signal appears to be enriched in regions flanking Sono-Seq sites. All data shown in this figure originate from *S. cerevisiae* chromosome 2. In all figures, position 0 corresponds to the start of the target feature and signal is given in fold-enrichment compared to background. Yeast FAIRE data from Hogan et al. were used for this analysis [19].

2.4.5 Sono-Seq DNA signals show little increase over H3K27me3 sites

H3K27me3 histone modification sites represent a signature of closed-conformation, facultative heterochromatin and are established by Polycomb group proteins [17,20]. We compared Sono-Seq DNA signals to the signals from 100,000 H3K27me3 sites identified in CD4+ cells [17]. The ChIP-Seq signals for all sample DNA types, including Pol II, remain flat over H3K27me3 sites (Figure 2.3D). We also examined the sequences of these regions to determine if the observed lack of ChIP-Seq signal was real or an artifact arising from an inability to map reads to these locations. As shown in Figure 2.3D, the sequences in H3K27 trimethylation regions and other genome regions can be mapped equally well (i.e. the mappability line in the plots remains close to 1.0 at all times, representing complete mappability). Thus, sonicated chromatin peaks preferentially lie near sites of active chromatin and are absent in closed chromatin regions.

2.4.6 Sono-Seq signal is depressed over FAIRE regions

We next determined whether Sono-Seq regions coincide with FAIRE regions because both protocols rely upon sonication of crosslinked chromatin. We performed this comparison in *S. cerevisiae*, in which FAIRE was first described [1]. Using data from Hogan et al. that was generated from *S. cerevisiae* chromosome 3 [19], we aggregated Sono-Seq signal over FAIRE sites and found that Sono-Seq signal is depressed (Figure 2.5A). When aggregating FAIRE signal over Sono-Seq sites, we observe highly enriched FAIRE signal levels bordering Sono-Seq regions but depressed signal levels over the Sono-Seq regions themselves (Figures 2.5B and A.6). These findings indicate that Sono-Seq is different from FAIRE and that Sono-Seq enriches regions that are protein-bound and exhibit local denaturation.

Hogan et al. [19] also found that FAIRE sites are anti-correlated with MNase-digested DNA signal. Our aggregation plots in HeLa S3 cells show that Sono-Seq and MNase-digested DNA signals exhibit trends similar to each other, further supporting that Sono-Seq and FAIRE experiments produce markedly different results.

2.4.7 Sono-Seq DNA peaks are affected by fragment size

To further investigate the origin of the Sono-Seq DNA signals, we analyzed different fragment sizes. Instead of using only the small 100-350 bp size sample normally recommended for Illumina sequencing, we also analyzed a larger size fraction (350-800 bp) that was prepared from the same sonicated extracts as the 100-350 bp fragments. As shown in Figure 2.3, the size of the fragments determines the presence and magnitude of the sonicated chromatin signals. The smallest fragments (100-350 bp) exhibit the largest signals whereas the largest fragments (350-800 bp) give smaller signals. Greater signals were also observed when qPCR was performed using electrophoretically separated small (100-500 bp) rather than large (1,000-6,000 bp) DNA fragments as template (Appendix C; Figure A.7). Thus size selection is a critical step in the preparation of Sono-Seq DNA and the characterization of the signal obtained.

2.5 Discussion

We demonstrate that sonication of chromatin causes breaks in localized and specific regions of the genome. By comparing Sono-Seq signals to TSSs, Pol II-bound regions, histone H3K4me3 sites, CpG islands and promoter-proximal DNase I hypersensitive sites we show that many of these peaks, as evidenced by aggregated signal, are located within promoter regions. Further analysis reveals that higher Sono-Seq signals are observed over the promoters of expressed genes as compared to those with little or no

expression. These results suggest that the breaks preferentially occur near regions of open chromatin of expressed genes; presumably these promoters have undergone chromatin remodeling, permitting access to transcriptional machinery but also allowing breakage by sonication of deprotected DNA (Figure A.10). In addition, we show that Sono-Seq peaks are also found in intergenic regions where these signals are modestly enriched over features commonly associated with promoter activity of expressed genes such as CpG islands and H3K4me3 sites. Many distal regions enriched for Sono-Seq are enriched for Pol II (21.7%; Figure A.3). As we show that regions enriched for both Sono-Seq and Pol II are generally associated with 5′ ends of expressed genes, we speculate that some of these distal peaks may lie proximal to genes that have not yet been annotated. Although many Sono-Seq peaks overlap with Pol II-bound loci, a number of Sono-Seq peaks do not and may represent binding of various other factors.

The advent of ChIP-Seq permits mapping of DNA regulatory regions at high resolution and low cost and is rapidly replacing ChIP-chip for the mapping of transcription factor binding sites. However, there are important differences. For ChIP-chip, immunoprecipitated fragments are labeled along their entire length and hybridized to a microarray in a mixture containing a differentially labeled reference DNA such that ratios of ChIP-DNA to reference DNA are typically recorded. In ChIP-Seq, breaks are generated in chromatin-bound DNA, short fragments isolated and the ends sequenced. The combined effect of examining one signal per sample and high resolution mapping of short fragment ends in ChIP-Seq reveals features of both the chromatin and the ChIP samples that have not been previously observed when ratios of ChIP to reference samples are analyzed. The implications of this are several-fold. First, reference DNA samples may

exhibit increased signals over 5′ ends of genes (Figures 2.3A-C). Second, we found that Sono-Seq DNA and normal IgG DNA are not equivalent. We expect that normal IgG DNA may be more useful as a reference sample because its treatment closely parallels that of a ChIP DNA sample and signal levels will not be dampened as much over transcribed regions, as would be the case when Sono-Seq DNA is used for scoring.

Fractionation of chromatin is non-random and may have an underlying biological basis depending upon the method by which it is prepared. Studies similar to ours have also demonstrated that chromatin fragments may be associated with annotated regions. One such method, FAIRE, has been shown to isolate regions correlated with nucleosome depletion, increased DNase I hypersensitivity, transcriptional start sites, and active promoters. Sono-Seq sites are different than FAIRE sites even though the Sono-Seq and FAIRE protocols share several common steps. Both protocols necessitate formaldehyde-crosslinking of proteins to DNA and then sonication of the crosslinked DNA. The key difference between FAIRE and Sono-Seq is that in FAIRE phenol-chloroform extraction occurs before reverse-crosslinking, such that protein-protected DNA is trapped at the interface and the open regions of DNA are released into the aqueous phase. However in Sono-Seq the protein-DNA crosslinks are reversed prior to phenol-chloroform extraction such that any protein-crosslinked DNA would be retained during purification. Although both Sono-Seq and FAIRE are associated with active promoters, aggregate Sono-Seq signal is depressed over FAIRE regions (Figure 2.5A). Furthermore, Sono-Seq signal parallels MNase-digested DNA signal over promoter regions [Figure 3; Figure 2C of [19]], FAIRE signal is depressed over Sono-Seq regions, and Sono-Seq regions are bounded by high FAIRE signal (Figure 2.5B), all further differentiating Sono-Seq from

FAIRE.

We speculate that Sono-Seq enriches regions that are protein-bound with local denaturation and detects breaks from neighboring open chromatin sites. For Sono-Seq regions to be recovered they must have sensitivity to sonication, which may arise from regions that are undergoing chromatin remodeling or local denaturation (e.g. by Pol II), most likely in preparation for or during transcription. Interestingly, enrichment over distal DNase I hypersensitive sites was not observed; these regions may be smaller and not readily broken during sonication, or they may reside in areas where chromatin organization is relatively static. Regardless of the mechanism, our analyses illustrate the utility of Sono-Seq as an effective approach for detecting accessible chromatin regions and facilitating the annotation of the human genome particularly in the promoter regions.

2.6 Materials and Methods

2.6.1 Preparation of DNA for ChIP-Seq and Sono-Seq

Cell growth protocols are available in Supplementary Text. For RNA Pol II ChIP-Seq, normal IgG ChIP-Seq and Sono-Seq, fixed HeLa S3 cells were washed in cold Dulbecco's PBS (Invitrogen, Carlsbad, CA) and swelled on ice in 10 mL hypotonic lysis buffer (20 mM Hepes [pH 7.9], 10 mM KCl, 1 mM EDTA [pH 8.0], 10% glycerol, 1 mM DTT, 0.5 mM PMSF, and protease inhibitors). Cell lysates were homogenized with 30 strokes in a Dounce homogenizer. Nuclear pellets were collected and lysed in 10 mL of RIPA buffer per 3 x108 cells (RIPA buffer: 10 mM Tris-Cl [pH 8.0], 140 mM NaCl, 1% Triton X-100, 0.1% SDS, 1% deoxycholic acid, 0.5 mM PMSF, 1 mM DTT, and protease inhibitors). Chromatin was sheared with an analog Branson 250 Sonifier (power setting 2, 100% duty cycle for 7x 30 s intervals) to an average size of less than 500 bp as verified

on a 2% agarose gel. Lysates were then clarified by centrifugation at 20,000xg for 15 min at 4°C. For Sono-Seq, aliquots of clarified lysate were reserved for reversal of crosslinking, followed by RNase and proteinase K treatments. Sono-Seq DNA was further purified by phenol-chloroform extraction and ethanol precipitation. For RNA Pol II and normal mouse IgG ChIP samples, 12 g of either the mouse monoclonal 8WG16 antibody (Covance MMS-126R) or normal mouse IgG (Santa-Cruz sc-2025) were added to 1x108 cells. Chromatin immunoprecipitations were conducted as previously described [21-22]. Libraries were constructed in a manner consistent with those from Rozowsky et al [8]. See Supplementary Text.

2.6.2 Preparation of naked DNA for sequencing

HeLa S3 cells were collected by centrifugation, resuspended in digestion buffer (100 mM NaCl, 10 mM Tris-HCl [pH 8.0], 25 mM EDTA [pH 8.0] and 0.5% SDS) and digested overnight at 50C with 0.1 mg/mL proteinase K (Ambion, Austin, TX). The digest was extracted twice with phenol-chloroform, once with chloroform and ethanol precipitated. The DNA was recovered, treated with RNase (Qiagen) for 3 h at 37C, extracted once with phenol-chloroform, once with chloroform, ethanol precipitated and resuspended at 2.5 x 108 cell equivalents in 5 mL of 1x TE pH 7.5 (10 mM Tris, 1 mM EDTA). A 2.5 mL aliquot was sonicated once for 30 s with a Branson 250 Sonifier (power setting 2, 100% duty cycle) to an average size of less than 500 bp as verified on a 2% agarose gel.

2.6.3 Preparation of MNase-treated DNA for sequencing

HeLa S3 cells were resuspended in MNase buffer (10 mM Tris-HCl [pH 7.5], 10 mM NaCl, 3 mM MgCl2, 1 mM CaCl2, 4% NP-40, and 1 mM DTT) and treated with 50

units of micrococcal nuclease (USB) at 37C for 1 h. The samples were treated with proteinase K for 2 h at 37C, extracted twice with phenol-chloroform, ethanol precipitated, treated with RNase A (Qiagen) and centrifuged through G50 sephadex spin columns. Each sample was treated with 30 units of calf intestinal alkaline phosphatase (NEB) for 2 h at 37C. After a second ethanol precipitation, the samples were treated with 30 units T4 polynucleotide kinase (NEB).

2.6.4 Preparation of yeast Sono-Seq DNA

Saccharomyces cerevisiae strains CMY288-1B (BY background) and YJM339 (clinical isolate) were grown in 500 mL of YPAD to mid-log phase (OD600=1.0). Cells were fixed with 1% formaldehyde for 15 min after which glycine was added to a final concentration of 125 mM. Cells were lysed with five 1-min bursts at 6.0m/s on a FastPrep-24 (MP Biomedicals). Chromatin was sonicated with a Branson 250 sonifier (Amplitude 50% for 5x 30 s intervals) to an average size of 450-500 bp. For each biological replicate, 250 µl of clarified lysate were processed to reverse crosslinks overnight, followed by a proteinase K treatment. The DNA was extracted three times in phenol:chloroform:isoamyl alcohol (25:24:1), and once in chloroform. After ethanol precipitation, DNA was resuspended in 1X TE [pH 8.0], RNase-treated and purified using a Qiagen MinElute PCR purification column. Finally, 100-350 bp Sono-Seq DNA was size-selected using a 2% agarose gel before Illumina library preparation. Sequencing libraries were generated as described above. Buffers are described in Aparicio et al [23].

2.6.5 Computational analysis of Illumina GA II data

Sequencing reads were analyzed using Illumina's Genome Analysis Pipeline version 0.3. Reads were aligned to human genome build 18 using the Eland aligner and

unique reads were used for ChIP-Seq scoring with PeakSeq [8]. Signal maps and aggregation plots were generated as described in the Supplementary Text.

Data are available in NCBI's Gene Expression Omnibus [23] through accession numbers GSE12781 (Pol II and Sono-Seq) and GSE14022 (Naked DNA, DNA treated with MNase, large-fragment Sono-Seq, and normal IgG). Signal files and other data can be accessed at http://archive.gersteinlab.org/proj/Sono-Seq.

2.6.6 Creation of ChIP-Seq mappability aggregations

A mappability profile for 30 nt reads was created and aggregations performed using the same strategies presented in Rozowsky et al. [8]. A mappability fraction of 1.0 for a given position means that a 30 nt read beginning at that position is fully mappable. Low ChIP signals from regions with high mappability indicate a true lack of reads from these regions.

2.6.7 Creation of FAIRE signal files and enriched regions

Block normalized log2 normalized FAIRE data from yeast chromosome 3 tiling arrays from Hogan et al. were downloaded from the GEO (accession number GSE4721) [19,23]. Values for each probe were averaged across the four microarray experiments to produce a composite data set and the probe IDs converted to genomic positions. These positions along with the corresponding average score were used to create a FAIRE signal file by averaging the values from overlapping tiles at each nucleotide position. Regions with a composite average score of at least 0.6 were deemed enriched and used to create a list of discrete FAIRE sites.

2.6.8 Scoring and aggregating Sono-Seq DNA in yeast

A Sono-Seq DNA data set for yeast was created by pooling data from two replicates then scoring against a randomized background using PeakSeq [8]. An aggregation of FAIRE signal over Sono-Seq sites was then created using the method described above. A bin size of 10 bp was used for all yeast aggregations. To create the FAIRE over Sono-Seq plot, aggregation was performed over regions consisting of 400 bp from the endpoints of each Sono-Seq region. The average of the furthest ten bins from each endpoint (corresponding to a region 300-400 bp distal of each Sono-Seq region endpoint) was used to normalize the remaining points.

2.6.9 Scoring Pol II and reference DNA samples against naked DNA and intersecting against promoters of Ensembl genes

PeakSeq was used to score Pol II and each reference DNA type against naked DNA [8]. Regions deemed to be enriched by PeakSeq were then intersected against promoter regions of Ensembl genes from Ensembl Release 50/NCBI36 using a C program leveraging the Bios library and coverage statistics were generated using the Active Region Comparer [24]. For this analysis, we define promoter regions of Ensembl genes to be 2.5 kb of the transcription start site and intersection as two sequences sharing at least one base position.

2.6.10 Calculating percent feature composition and creating a rank-order plot for Sono-Seq DNA and Pol II DNA

Enriched regions for Pol II and Sono-Seq DNA were ranked in descending order according to sequence tag count and fold enrichment versus naked DNA. These peaks were then classified by their proximity to known promoter regions and a rank-order plot

produced (see Appendix A).

2.7 References

1. Nagy, P. L., Cleary, M. L., Brown, P. O. & Lieb, J. D. Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin. *Proc Natl Acad Sci USA* **100**, 6364–6369 (2003).
2. Crawford, G. E. *et al.* DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat Methods* **3**, 503–509 (2006).
3. Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. & Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* **17**, 877–885 (2007).
4. Horak, C. E. & Snyder, M. ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Meth Enzymol* **350**, 469–483 (2002).
5. Iyer, V. R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
6. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
7. Lee, T. I., Johnstone, S. E. & Young, R. A. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nature protocols* **1**, 729–748 (2006).
8. Rozowsky, J. S. *et al.* PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* **27**, 66–75 (2009).
9. Jothi, R., Cuddapah, S., Barski, A., Cui, K. & Zhao, K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* **36**, 5221–5231 (2008).
10. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* **5**, 829–834 (2008).
11. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
12. Flicek, P. *et al.* Ensembl 2008. *Nucleic Acids Res* **36**, D707–14 (2008).
13. Affymetrix ENCODE Transcriptome Project & Cold Spring Harbor Laboratory ENCODE Transcriptome Project Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* **457**, 1028–1032 (2009).
14. Morin, R. *et al.* Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques* **45**, 81–94 (2008).
15. Niwa, H. Open conformation chromatin and pluripotency. *Genes Dev* **21**, 2671–2676 (2007).
16. Robertson, A. G. *et al.* Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Res* **18**, 1906–1917 (2008).
17. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
18. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters.

Proc Natl Acad Sci USA **103**, 1412–1417 (2006).
19. Hogan, G. J., Lee, C.-K. & Lieb, J. D. Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters. *PLoS Genet* **2**, e158 (2006).
20. Ross, P. J. *et al.* Polycomb gene expression and histone H3 lysine 27 trimethylation changes during bovine preimplantation development. *Reproduction* **136**, 777–785 (2008).
21. Euskirchen, G. M. *et al.* Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. *Genome Res* **17**, 898–909 (2007).
22. Aparicio, O., Geisberg, J. V. & Struhl, K. Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Curr Protoc Cell Biol* **Chapter 17**, Unit 17.7 (2004).
23. Barrett, T. *et al.* NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res* **35**, D760–5 (2007).
24. Rozowsky, J. S. *et al.* The DART classification of unannotated transcription within the ENCODE regions: associating transcription with known and novel loci. *Genome Res* **17**, 732–745 (2007).
25. Auerbach, R. K. *et al.* Mapping accessible chromatin regions using Sono-Seq. *Proc Natl Acad Sci USA* **106**, 14926–14931 (2009).

Chapter 3: Analysis of the Human SWI/SNF Chromatin Remodeling Complex through Data Integration

3.1 Statement of Prior Publication and Bioinformatics Contribution

This work is reprinted from a 2011 paper in *PLoS Genetics* by Euskirchen and Auerbach, et al [91]. No additional permissions were required. This work represents one of the early attempts to ChIP members of a chromatin remodeling complex. Informatically this posed several new challenges. ChIP-Seq scoring algorithms had focused on factors with a more defined binding pattern (e.g. factors that bind proximal to promoter regions) rather than factors that are ubiquitous across the genome such as chromatin remodelers. From personal experience with several peak calling algorithms available at the time, many would call too many peaks or produce questionable peak calls based on the reference DNA biases discussed in Chapter 2. Additionally, there are approximately 288 possible subunit combinations and some members of the SWI/SNF chromatin remodeling complex are thought to have other roles outside the complex. These characteristics required a more specialized analysis than what was being done at the time. In consultation with my coauthor in regards to biological underpinnings, we designed a strategy to select high quality peaks based on total read count, the difference between the number of factor and input reads, and other characteristics in addition to the q-value returned by peak calling programs. At the time, few ChIP-Seq datasets were publicly available, those that were in the public domain were typically produced using

ideal antibodies (e.g. RNA Polymerase II), and bioinforamticists relied almost solely upon the q-value produced by peak-calling programs. This approach was insufficient for this project, however, as most peak callers at the time were designed around datasets with defined binding patterns near promoters and that produced a relatively small number of peak calls. I identified additional criteria to filter the ChIP-Seq peaks. In addition, I (with consultation from my coauthor, Ghia Euskirchen, regarding the underlying biology) employed an evidence-based approach to determine whether a SWI/SNF complex was present based on surrounding annotated features, the ChIP-Seq peaks called for each subunit, and what was known about SWI/SNF biology. This paper represents one of the early efforts in our lab to go beyond "off-the-shelf" peak calls and to design an approach to examine non-standard ChIP factors. The results of the above evidence-based approach were consistent with many targets/pathways one would expect given SWI/SNF biology (e.g. signaling pathways, cancer pathways). In addition to formulating the methodology and identifying these targets, I also conceived and implemented the analysis depicted in Figure 3.5 to examine the transcriptional efficacies of different SWI/SNF subunit combinations. Through this evidence-based approach that I designed and implemented, we were able to identify high-confidence targets of a multi-subunit protein complex that also made sense in a biological context.

3.2 Abstract

A systems understanding of nuclear organization and events is critical for determining how cells divide, differentiate and respond to stimuli and for identifying the causes of diseases. Chromatin remodeling complexes such as SWI/SNF have been implicated in a wide variety of cellular processes including gene expression, nuclear

organization, centromere function and chromosomal stability, and mutations in SWI/SNF components have been linked to several types of cancer. To better understand the biological processes in which chromatin remodeling proteins participate we globally mapped binding regions for several components of the SWI/SNF complex throughout the human genome using ChIP-Seq. SWI/SNF components were found to lie near regulatory elements integral to transcription (e.g. 5' ends, RNA Polymerases II and III and enhancers) as well as regions critical for chromosome organization (e.g. CTCF, lamins and DNA replication origins). Interestingly we also find that certain configurations of SWI/SNF subunits are associated with transcripts that have higher levels of expression whereas other configurations of SWI/SNF factors are associated with transcripts that have lower levels of expression. To further elucidate the association of SWI/SNF subunits with each other as well as with other nuclear proteins we also analyzed SWI/SNF immunoprecipitated complexes by mass spectrometry. Individual SWI/SNF factors are associated with their own family members as well as with cellular constituents such as nuclear matrix proteins, key transcription factors and centromere components implying a ubiquitous role in gene regulation and nuclear function. We find an overrepresentation of both SWI/SNF-associated regions and proteins in cell cycle and chromosome organization. Taken together the results from our ChIP and immunoprecipitation experiments suggest that SWI/SNF facilitates gene regulation and genome function more broadly and through a greater diversity of interactions than previously appreciated.

3.3 Author Summary

Genetic information and programming are not entirely contained in DNA sequence

but are also governed by chromatin structure. Gaining a greater understanding of chromatin remodeling complexes can bridge gaps between processes in the genome and the epigenome and offer insights into diseases such as cancer. We identified targets of the chromatin remodeling complex, SWI/SNF, on a genome-wide scale using ChIP-Seq. We also identify proteins that co-purify with its various components via immunoprecipitation combined with mass spectrometry. By integrating these newly-identified regions with a combination of novel and published data sources, we identify pathways and cellular compartments in which SWI/SNF plays a major role as well as discern general characteristics of SWI/SNF target sites. Our parallel evaluations of multiple SWI/SNF factors indicate that these subunits are found in highly dynamic and combinatorial assemblies. Our study presents the first genome-wide and unified view of multiple SWI/SNF components, and also provides a valuable resource to the scientific community as an important data source to be integrated with future genomic and epigenomic studies.

3.4 Introduction

Chromosomes undergo a wide variety of dynamic processes including transcription, replication, repair and packaging. Each of these activities requires the recruitment and congregation of a particular set of factors and chromosomal elements. For example visualization of nascent mRNA in HeLa cells has led to a model of transcription units being clustered into “factories” thereby facilitating optimal engagement of RNA Polymerase II (Pol II) and coordination with other crucial holoenzyme complexes [1-3]. In addition to RNA Pol II and transcription factors, transcriptional assemblages include proteins critical to regulating chromatin. The accessibility of nuclear proteins to DNA is often controlled by ATP-dependent chromatin remodeling complexes, which are thought

to play a role in a number of different cellular transactions by reshaping the epigenetic landscape.

The SWI/SNF (switch/sucrose nonfermentable) chromatin remodeling proteins were first discovered in *Saccharomyces cerevisiae* as components of a 2 MDa complex that repositions nucleosomes for vital tasks such as transcriptional control, DNA repair, recombination, and chromosome segregation [4-5]. Mammalian SWI/SNF is comprised of approximately ten subunits and the combinations of these subunits, some of which have multiple isoforms, enable multiple varieties of SWI/SNF complexes to exist both within a given cell and across cell types [6]. Among these subunits either of the two ATPases, Brg1 or Brm, is sufficient to remodel nucleosome arrays *in vitro*, however maximal nucleosome remodeling activity is achieved when the SWI/SNF subunits BAF155, BAF170 and Ini1 are present in a 2:1 stoichiometry relative to Brg1 [7]. Whereas the ATPases have an obvious catalytic function, the roles of the other SWI/SNF subunits are largely obscure. Several reports indicate that BAF155 and BAF170 provide scaffolding functions for other SWI/SNF subunits as well as regulating their protein levels [8-9]. SWI/SNF also contains -actin and the actin-related protein BAF53, suggesting a possible bridge to nuclear organization or signal transduction, e.g. through phosphatidylinositol signaling [10-11]. Phosphatidylinositol 4,5-bisphosphate has been shown to bind to Brg1 and promote binding to actin filaments [12]. Mutations resulting in loss of Ini1 function are associated with rare but aggressive pediatric cancers [13-14]. The SWI/SNF subunits Brg1 [15] and ARID1A [16-18] are likewise thought to have tumor suppressor roles based on mutations recovered from other tumor types. Curiously, Ini1 alone has a unique and largely undefined role in HIV-1 infection that includes

binding of Ini1 to HIV-1 integrase and the cytoplasmic export of Ini1 and its incorporation into HIV-1 particles [19-21].

The role of SWI/SNF components in cancer and tumor suppression is poorly understood despite extensive study. Detailed investigations of individual loci have implicated SWI/SNF in various transcriptional pathways including the cell cycle and p53 signaling [22], insulin signaling [23], and TGF signaling [24], as well as signaling through several different nuclear hormone receptors [25]. Although *in vitro* experiments and single-gene studies have been informative and have laid the foundation for understanding chromatin remodeling, a global analysis of targets of SWI/SNF is expected to yield a more extensive view into the biological roles of SWI/SNF components and their involvement in human disease.

In this study we present two complementary global analyses of SWI/SNF subunits to provide a more systematic view of SWI/SNF functions. First we performed ChIP-Seq with the ubiquitous SWI/SNF components Ini1, BAF155, BAF170 as well as the Brg1 ATPase. Second, in a parallel set of studies we performed mass spectrometry identification of proteins that co-immunoprecipitate with SWI/SNF components. Using our ChIP-Seq results the resulting chromosomal locations were integrated with published annotations to yield a more complete understanding of SWI/SNF on a genome-wide scale. We find SWI/SNF components frequently occupy transcription start sites (TSSs), enhancers, CTCF regions and many regions occupied by Pol II. Further analyses of the SWI/SNF regions we identified by ChIP-Seq reveals that SWI/SNF factors target genes and signaling pathways involved in cell proliferation and cancer. Our investigation of SWI/SNF protein interactions detected not only the expected co-occurrences of

individual SWI/SNF factors with each other but also with cellular components such as nuclear matrix proteins, key transcription factors and centromere proteins implying a ubiquitous role in gene regulation and nuclear function. We find an overrepresentation of both SWI/SNF-associated chromosomal regions and proteins in cell cycle and chromosome organization. Collectively our results suggest that SWI/SNF is at the nexus of multiple signal transduction pathways, essential chromosomal functions and nuclear organization.

3.5 Results

3.5.1 Genome-wide mapping of SWI/SNF subunits reveals many different co-associations

We identified the targets of four SWI/SNF components, Ini1 (SMARCB1), Brg1 (SMARCA4), BAF155 (SMARCC1) and BAF170 (SMARCC2), using ChIP-Seq. Chromatin complexes were isolated from HeLa S3 nuclei following independent immunoprecipitations with antibodies for each factor. Each of these antibodies was characterized by both immunoblot and mass spectrometry analyses (see Materials and Methods). Reads that mapped uniquely to the genome were retained (29-33 million reads per data set; Table 1) and significant binding regions were identified using the PeakSeq program with q-value < 0.05 [26]. The peaks were compared to a similarly-sized data set of uniquely mapped ChIP DNA reads obtained from control immunoprecipitation experiments using normal IgG (i.e. a control serum that is not directed to any known antigens). Using this approach we identified many Ini1-, Brg1-, BAF155- and BAF170-associated regions (Table 1).

The majority of SWI/SNF binding occurs near (2.5 kb) protein-coding genes, a

distribution that is significant relative to a random target list ($p < 1x10^{-16}$; Genome Structure Correction (GSC) test [27]; see Materials and Methods). Several examples of SWI/SNF positioning relative to genic regions are shown in Figures 3.1 and B.1. In order to further examine SWI/SNF locations with respect to gene-rich and gene-poor regions we obtained a set of histone H3K27me3 domains that were identified in HeLa cells (Table B.1; [28]) because this chromatin mark often occurs in gene-poor and repressed (i.e. heterochromatin) regions. Although most SWI/SNF-binding occurs outside H3K27me3 domains, we observed that SWI/SNF is occasionally found in heterochromatin regions, as shown in Figure B.2. In this example a 7.5 Mb heterochromatin region on Chr16 contains a single gene, the neuronal cadherin *CDH8*, that is repressed and lacks RNA Pol II, however several SWI/SNF binding regions are found nearby.

We have performed considerable analyses of the targets for the individual SWI/SNF factors, particularly with respect to elements representing several major classes of genomic features including promoters (Ensembl protein-coding genes), RNA Pol II sites [26], CTCF sites [28], and predicted enhancers [29]. All of these features were identified in HeLa cells (Tables 3.1, B.1 and B.2; see Materials and Methods). In comparisons between the individual target lists for Ini1, Brg1, BAF155 and BAF170 with promoters, RNA Pol II sites, CTCF sites and enhancers we found that each SWI/SNF factor is significantly overrepresented for each of these major classes of genomic elements ($p < 1x10^{-16}$, GSC test, see Materials and Methods). To arrive at a single unified and more conservative list of SWI/SNF locations, we first took the union of all regions for Ini1, BAF155, BAF170 and Brg1, resulting in 69,658 SWI/SNF regions. We then

trimmed this list to a high-confidence set of 49,555 sites by eliminating those regions where either only a single SWI/SNF subunit was present or that those regions that did not co-occur with either promoters, RNA Pol II sites, CTCF sites or predicted enhancers. We used this list of 49,555 SWI/SNF regions for all subsequent analyses unless otherwise noted (Table B.3). The four major classes of genomic features mentioned above were overrepresented in both the 69,658 SWI/SNF regions as well as the more conservative list 49,555 SWI/SNF regions ($p < 1x10^{-16}$, GSC test).

We next examined the configurations of our 49,555 SWI/SNF regions (Figure 3.3A and Table 3.2). Ini1, BAF155 and BAF170 have been described as forming a 'core' based on their ability to stimulate remodeling activity of the Brg1 ATPase in reconstitution experiments [7]. Among our data 30,310 regions (61%) have two or more SWI/SNF components and 9,760 regions (20%) contain the core of Ini1, BAF155 and BAF170; for the purposes of this study we call this the 'core set'. Among putative complexes comprised of two or more SWI/SNF subunits, we observed BAF155 was the subunit most common to each binding region. Only 770 SWI/SNF subunit co-occurrences were recovered that lacked BAF155 as compared to 6,467 for BAF170 and 14,824 for Ini1. This finding is consistent with several previous studies showing that BAF155 is important for SWI/SNF complex stability [8-9]. BAF155 may increase the stability of the complex during assembly, or BAF155 may be easier to detect by ChIP.

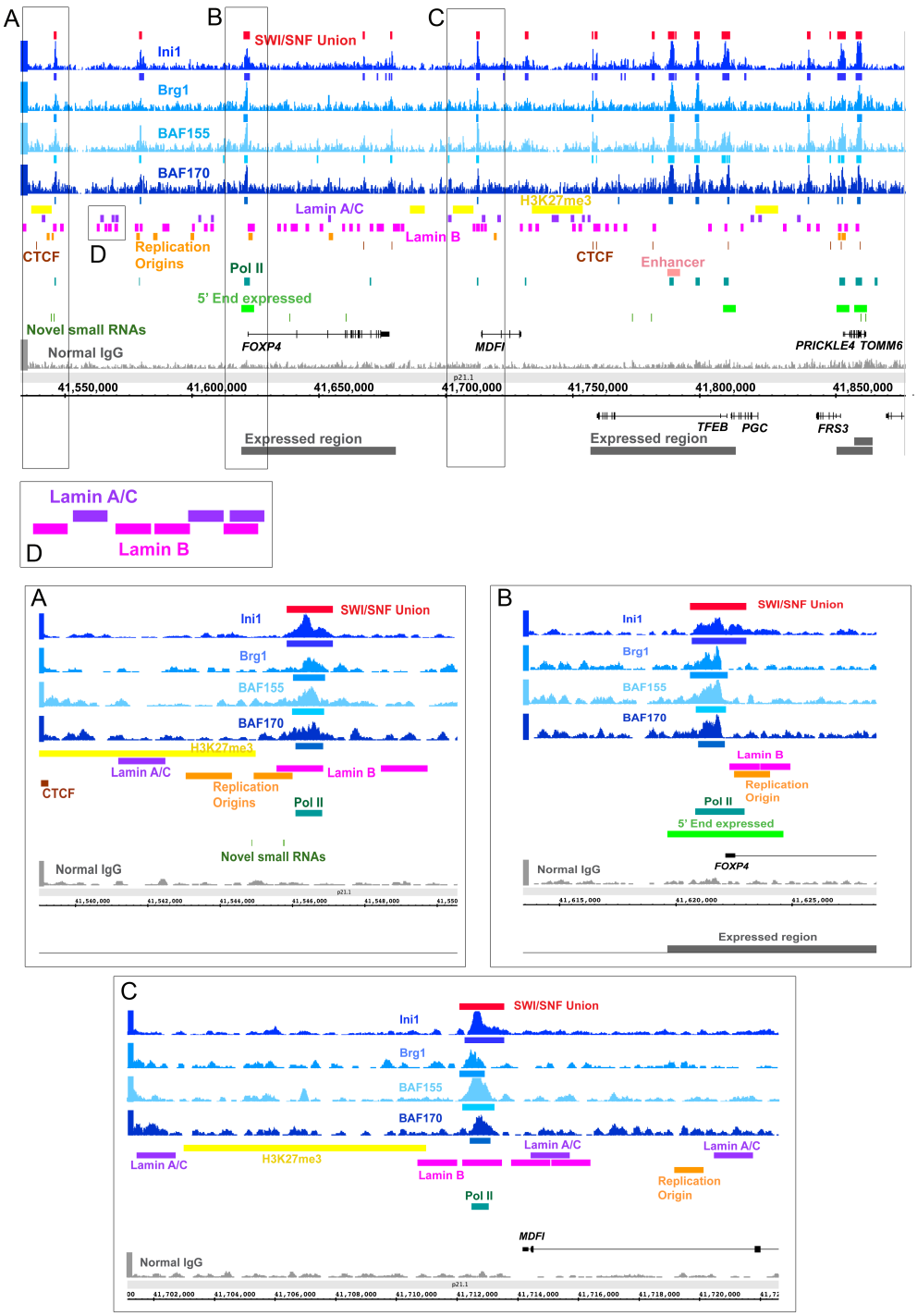


Figure 3.1 SWI/SNF regions co-occur with many diverse genomic elements. The ChIP-Seq regions and signal tracks displayed encompass a ~340 kb region on chromosome 6. The coordinates shown are in hg18 and all regions were identified in HeLa cells as detailed in Table S1 and Materials and Methods. Insets A-D are shown both in the context of the 340 kb region and in magnified view. Inset A displays a ~10 kb region at the edge of an H3K27me3 domain. Inset B displays a ~10 kb region around the 5 end of *FOXP4*. Inset C displays a ~20 kb region around the 5 end of *MDFI*. Inset D shows an example of where lamin A/C and lamin B can both flank and overlap with each other. Annotations above the coordinate axis are for forward-strand genes, and annotations below are for reverse-strand genes. Signal tracks are scaled consistently

based on number of reads. The vertical axis for each signal track is the count of the number of overlapping DNA fragments at each nucleotide position and is scaled from 0 to 40 for each track.

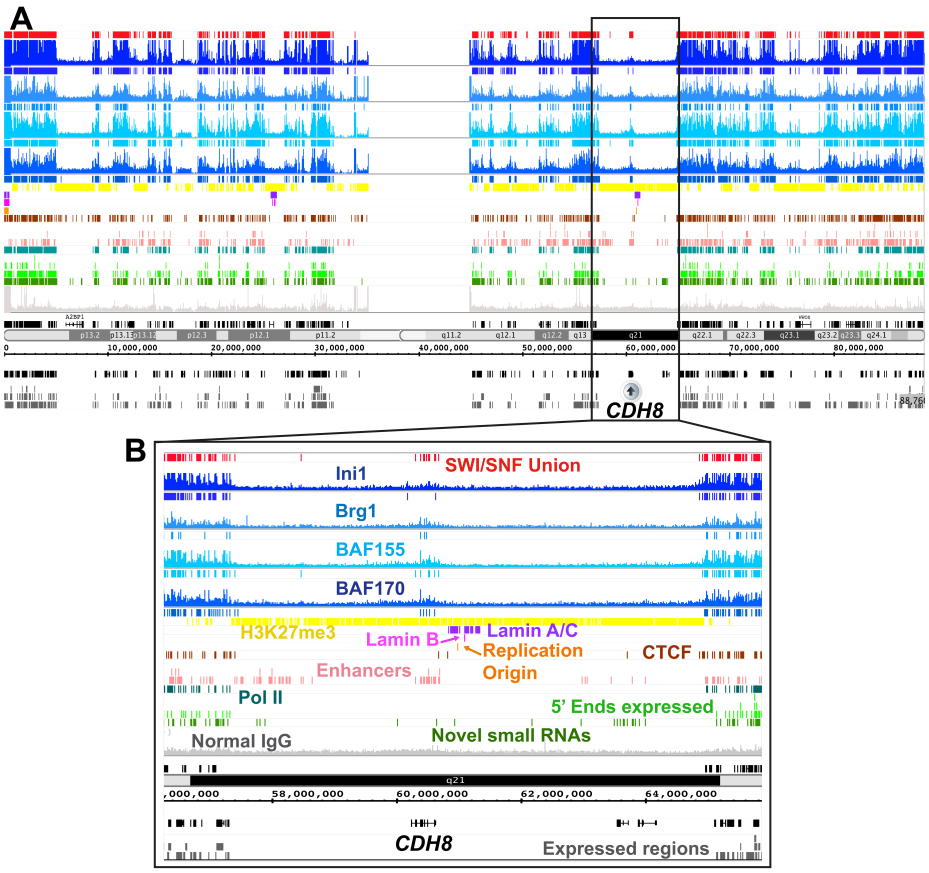


Figure 3.2 SWI/SNF signals and target regions in the context of H3K27me3 domains. As shown in panel A SWI/SNF signals (blue) are sparse in H3K27me3 regions (yellow) along the entire length of chromosome 16. An exception is shown in Panel B where SWI/SNF occurs around the *CDH8* gene that is embedded in an H3K27me3 domain. CDH8 encodes a brain-expressed cadherin that is not expressed in HeLa cells using the data of Morin et al.

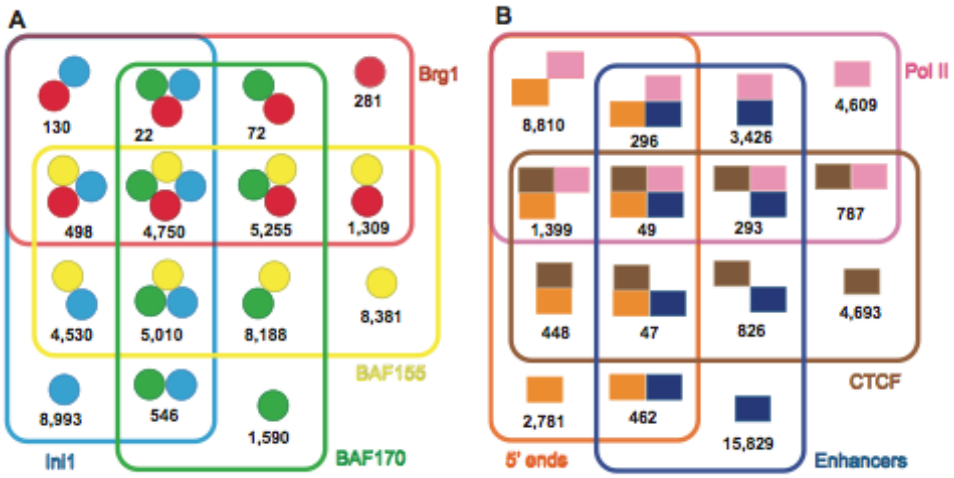


Figure 3.3 Venn diagrams showing overlaps for the SWI/SNF union target regions. Panel A displays the various combinations of all Ini1, BAF155, BAF170 and Brg1 targets in the 49,555 high-confidence union regions (see also Table 2). Panel B displays the various combinations of Pol II regions, 5 ends of Ensembl protein-coding genes, CTCF sites and putative enhancers occurring in the 49,555 SWI/SNF high-confidence union target regions. Of the 49,555 high-confidence union regions, 4,800 (10%) do not contain any of these elements and are defined as 'unclassified' (Table 3.3).

Table 3.1 Read counts and target regions identified by ChIP-Seq.

Data set	Number of uniquely mapped reads	Total number of targets (PeakSeq)[1]	Number of genic targets[2]	Number of targets after filtering (high-confidence union)[3]
Ini1	33,360,976	49,458	32,725 (66%)	24,478 (49%)
Brg1	30,037,219	12,725	7,823 (61%)	12,317 (25%)
BAF155	28,800,740	46,412	28,221 (61%)	37,921 (77%)
BAF170	29,090,374	30,136	18,847 (63%)	25,433 (51%)
Pol II	29,060,928	23,320	18,305 (78%)	19,669 (40%)
IgG control	28,960,961	N/A	N/A	N/A

[1]Uniquely mapped targets were identified by PeakSeq and further filtered using criteria more stringent than the default parameters. See Materials and Methods.
[2]Genic regions were identified for the total number of targets determined by PeakSeq. Genic regions are defined as a window encompassing 2.5 kb up- and downstream of the 5' and 3' ends, respectively, using protein-coding genes from Ensembl build 52 based on hg18.
[3]The high-confidence union list contains 49,555 targets and was formed by creating a union list of all SWI/SNF regions from those identified by PeakSeq and trimming this list to those SWI/SNF regions that co-occur with each other, Pol II regions, 5' ends of protein-coding genes, CTCF sites or putative enhancers. For further details, see Materials and Methods.

Table 3.2 Combinations of SWI/SNF factors found in the high-confidence union regions.

SWI/SNF Union Set	49,555
Two or more subunits	30,310
Three or more subunits	15,535
Core Set: Ini1, BAF155 and BAF170 (may include Brg1)	9,760
Ini1, BAF155, BAF170 and Brg1	4,750

3.5.2 Genome-wide locations of SWI/SNF components suggest diverse roles in gene regulation

One of the primary functions of chromatin remodeling complexes is to assist in gene regulation. Among the SWI/SNF regions in our high-confidence union set of 49,555 sites, 29% correspond to the 5' ends of protein-coding Ensembl genes, 40% correspond to Pol II sites, 17% correspond to CTCF sites and 43% correspond to predicted enhancer regions (Figure 3.3B; Table 3.3). The various combinations of these four elements account for a total of 90% of the SWI/SNF union regions; 4,800 (10%) of the SWI/SNF regions are unclassified using the above elements. Similar trends were observed for the 9,760 SWI/SNF "core" regions where Ini1, BAF155 and BAF170 all co-occur (Table 3.3). None of these four particular SWI/SNF subunits or any combinations thereof exhibited a differential preference for one type of element (Table 3.4).

There are some differences between the SWI/SNF core and union regions. The SWI/SNF core regions are overrepresented for RNA Pol II ($p < 9.9 \times 10^{-16}$; hypergeometric test) and 5' ends ($p < 6.5 \times 10^{-211}$; hypergeometric test) relative to all of the SWI/SNF high-confidence union regions; however the SWI/SNF high-confidence

union regions are overrepresented for enhancer regions relative to the Ini1-BAF155-BAF170 core ($p < 2.4 \times 10^{-67}$; hypergeometric test). Neither the SWI/SNF core nor the high-confidence union regions were over- or underrepresented for CTCF sites relative to each other ($p > 0.05$; hypergeometric test).

Enhancers are often characterized by long-range interactions. We examined the locations of SWI/SNF binding regions in the 150 kb *CIITA* region where numerous chromosomal looping interactions have been mapped at high resolution in HeLa cells using 3C (Chromosome Conformation Capture). Brg1 has been previously mapped at several sites in this locus in these cells [30]. Superimposition of these 3C data on our SWI/SNF ChIP-Seq data (Figure 3.4) reveals that all six of the 3C interacting regions in the *CIITA* locus (-50 kb, -16 kb, -8 kb, pIV, +40 kb and +59 kb) are bound by SWI/SNF components. Moreover certain individual SWI/SNF component binding regions that appeared initially as orphans may now be seen as part of a complete complex when joined with a distal element. For example Ini1 at pIV when joined with BAF155 and BAF170 regions at the -16 kb element forms a SWI/SNF core. Thus in the *CTIIA* locus SWI/SNF regions are often associated with 3C regions and many of the regions bound by individual factors may in fact be part of entire SWI/SNF complexes inside the nucleus.

Overall our ChIP-Seq results are summarized in Tables 3.1 to 3.3 and Figure 3.3 and indicate that SWI/SNF likely contributes to gene regulation through many different avenues in light of its binding to promoters, enhancers and CTCF sites. Furthermore SWI/SNF may facilitate looping interactions among these various elements as it has been shown *in vitro* that SWI/SNF can interact simultaneously with multiple DNA sites and generate loops between them [31]. Interestingly we found a slightly higher presence of

the SWI/SNF core at TSSs and with Pol II than the SWI/SNF union regions with these elements (Table 3.3). Thus a complete core of Ini1, BAF155 and BAF170 may be required for effective promoter function whereas only a subset of these factors may be required for enhancer function. Alternatively a full SWI/SNF core may be more difficult to recover from a single enhancer element as compared to a more compact promoter region due to the enhancer's presumed interaction with many different distal elements.

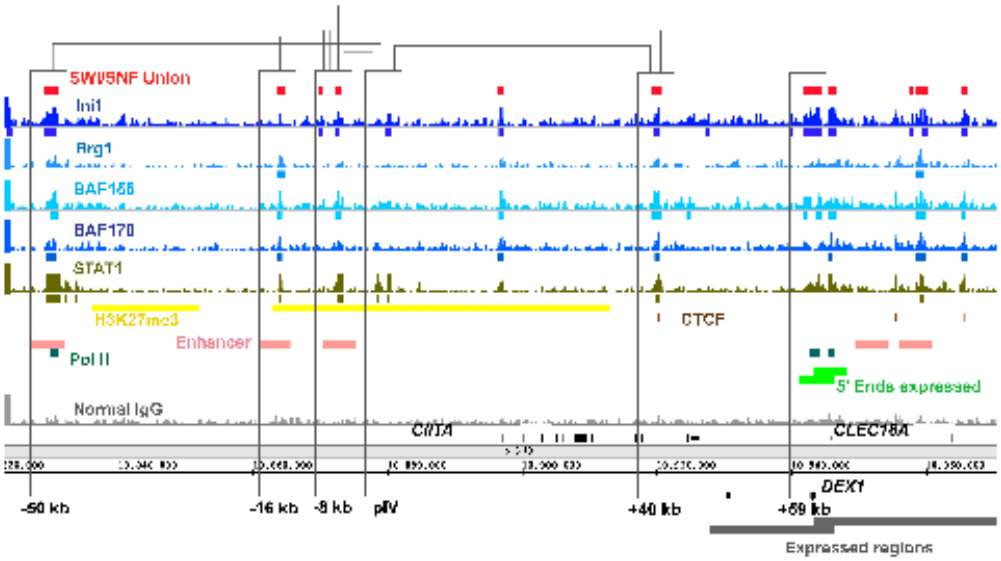


Figure 3.4 SWI/SNF signals relative to 3C sites in the *CIITA* locus. A ~150 kb region surrounding the *CIITA* locus is shown with SWI/SNF signals. Chromosomal loops detected in Ni et al. [30] are displayed as brackets connecting regions that were shown to contact each other using 3C. In the absence of -interferon eight constitutive contacts have been observed by 3C in HeLa cells between the sites at: (–50:–8), (–50:+59), (–8:+59), (pIV:+40), (–50:pIV), (–16:pIV), (–8:+40) and (–8:pIV). *CIITA* contains STAT1 binding regions; for comparison, STAT1 data are also shown from ChIP-Seq signals and target regions obtained from -interferon-stimulated HeLa cells as we previously reported [26]. The Ini1 site at pIV, when joined with BAF155 and BAF170 regions at the -16 kb element, forms complete a SWI/SNF core. The vertical axis for each signal track is the count of the number of overlapping DNA fragments at each nucleotide position and is scaled from 0 to 40 for each track.

Table 3.3 Genomic elements found in SWI/SNF target regions.

Genomic Elements	SWI/SNF union set (49,555 regions total)[1]	SWI/SNF core set (9,760 regions total)[2]
CTCF, Pol II, Enhancers, 5' ends (any combination)	44,755 (90%)	8,968 (92%)
Unclassified	4,800 (10%)	792 (8%)
RNA Pol II sites	19,669 (40%)	6,562 (67%)
Putative Enhancers	21,228 (43%)	3,431 (35%)
CTCF sites	8,542 (17%)	1,692 (17%)
5' ends (within 2.5 kb) of Ensembl protein-coding genes	14,291 (29%)	4,089 (42%)

[1]This high-confidence union list is the same described in Table 1. For further details see Materials and Methods.
[2]The core set is defined as those regions having a co-occurrence of Ini1, BAF155 and BAF170.

3.5.3 RNA polymerases are extensively colocalized with SWI/SNF

As detailed above SWI/SNF regions are enriched for Pol II. To explore the prevalence of SWI/SNF with transcriptional machinery we asked whether the converse would also be true, namely if regions bound by RNA polymerases are enriched for SWI/SNF. Indeed Pol II regions are enriched for SWI/SNF binding regions ($p < 1x10^{-16}$, GSC test). Although Pol II overlaps extensively with SWI/SNF it differs from SWI/SNF in its concordance with CTCF and enhancer regions (Table B.5). Pol II regions lacking SWI/SNF show a five-fold decrease in CTCF sites and a two-fold decrease in enhancer regions as compared to those Pol II regions containing SWI/SNF.

We further compared our SWI/SNF regions with binding intervals identified for RNA polymerase III (Pol III), which in addition to transcribing tRNA and other non-protein coding RNAs has an emerging role in the formation of boundary elements [32-33]. Pol III localization data were obtained from published ChIP-Seq studies using HeLa cells ([34-35]; Tables B.1 and B.6) and constitute 478 known and novel Pol III-associated

regions. Pol II is often associated with Pol III (Table 3.4; reviewed in [32]. Therefore we examined whether SWI/SNF was associated with Pol III binding regions independently of Pol II. Of the 478 Pol III regions, 253 Pol III intervals lack Pol II and among these 39% (98/253) contain one or more SWI/SNF components. These results suggest that SWI/SNF association with Pol III can occur independently of Pol II.

Overall 65% (309/478) of Pol III regions and 84% (19,541/23,320) of Pol II regions have at least one SWI/SNF factor associated with them. The Ini1-BAF155-BAF170 core is found at 41% (195/478) of Pol III regions and 52% (12,079/23,320) of Pol II regions. From the colocalizations of SWI/SNF, Pol II and Pol III we see that there is substantial overlap among these factors yet each of these factors also has distinct characteristics.

3.5.4 SWI/SNF components bind near many expressed regions

SWI/SNF is known to act as both an activator and repressor of transcription [36]. We examined the locations of four SWI/SNF components relative to transcribed regions in HeLa S3 cells using the RNA-Seq data of Morin et al. [37], Ini1, Brg1, BAF155 and/or BAF170 are present at or near the 5' ends (± 3.5 kb) of 71 to 92% of active protein-coding genes. As noted above, SWI/SNF occupancy in promoters is similar to that of Pol II and each of the factors is individually enriched in promoter regions ($p < 1x10^{-16}$, GSC test). Although the majority of Ini1, Brg1, BAF155 and BAF170 target genes are expressed, an appreciable fraction of gene targets have little or no detectable mRNA in HeLa cells. A closer examination of the union regions where a SWI/SNF component is located in the promoter of an inactive gene reveals that 58% (2,063/3,565) of these promoters are co-associated with Pol II suggesting transcriptional stalling (reviewed in [38-39].

Table 3.4 Co-occurrence of RNA Pol II and Pol III with SWI/SNF high-confidence union regions.

	All RNA Pol III[1]	**All RNA Pol II[2]**	**SWI/SNF, Pol II and Pol III[3]**	**SWI/SNF and Pol III present; Pol II absent[4]**	**Pol III present and SWI/SNF absent[5]**	**Pol II present and SWI/SNF absent[6]**
RNA Pol III	478	182 (0.8%)	211	98	169	6 (~0.1%)
RNA Pol II	225 (47%)	23,320	211	0	14 (8%)	3,779
Ini1	274 (57%)	18,674 (80%)	201 (95%)	73 (74%)	0	0
BAF155	235 (49%)	16,175 (69%)	176 (83%)	59 (60%)	0	0
BAF170	248 (52%)	12,790 (55%)	179 (85%)	69 (70%)	0	0
Brg1	75 (16%)	7,303 (31%)	68 (32%)	7 (7%)	0	0
Core	195 (41%)	12,079 (52%)	160 (76%)	35 (36%)	0	0
CTCF	49 (10%)	2,448 (10%)	14 (7%)	25 (26%)	10 (6%)	68 (<2%)
Enhancer	22 (5%)	4,527 (19%)	4 (2%)	14 (14%)	4 (2%)	377 (10%)
Protein-coding[7]	198 (41%)	18,305 (78%)	111 (53%)	34 (35%)	53 (31%)	3,052 (81%)

[1]There are a total of 478 Pol III regions genome-wide.
[2]There are a total of 23,320 Pol II regions genome-wide.
[3]SWI/SNF, Pol II and Pol III co-occur in 211 regions. Percentages shown are relative to these 211 regions.
[4]Pol III co-occurs with 98 SWI/SNF regions in the absence of Pol II.
[5]Of the total 478 Pol III regions 169 lack SWI/SNF.
[6]Of the total 23,320 Pol II regions 3,779 lack SWI/SNF.
[7]Regions within 2.5 kb of an Ensembl protein-coding gene.

Considering that SWI/SNF components bind near many expressed regions and that SWI/SNF factors occur in a multitude of configurations (Figure 3.3 and Table B.3), we examined transcript expression levels for all possible combinations of Ini1, Brg1, BAF155 and BAF170 occurrences. Using the RNA-Seq data of Morin et al. [37], we examined transcript expression levels corresponding to each of these configurations (Figure 3.5). We see that the highest levels of transcription are associated with the following four configurations: 1) the complete core of Ini1, BAF155 and BAF170; 2) the

complete core plus Brg1; 3) Ini1 and BAF155 only and 4) Brg1, BAF155 and BAF170. Although BAF155 is the subunit that is common to all of the configurations associated with the highest levels of transcription, it does not appear to be the sole driver of transcriptional activity. Compared against each other, all three components of the core complex taken individually have nearly indistinguishable profiles. Despite the involvement of Brg1 in two of the four configurations with the highest expression levels, most other configurations involving Brg1 are restricted to profiles associated with the lowest expression levels. One inference from these data is that certain combinations of SWI/SNF subunits are likely synergistic in promoting transcription whereas other combinations may be inhibitory or unstable.

We also examined SWI/SNF occurrences relative to 48,403 non-canonical small RNAs from HeLa cells (≤156 bp; Table B.1) where most (83%; $p < 1x10^{-16}$, GSC test) of these small RNAs are near protein-coding genes [40]. Approximately one third (30%) of this entire small RNA set is within 1 kb of a target from our high-confidence union list of 49,555 SWI/SNF regions. The incidence of small RNA-SWI/SNF co-associated regions was nearly equivalent in protein-coding genes and intergenic regions. From this we surmise that SWI/SNF may contribute to gene regulation of a variety of transcripts, many of which are newly annotated and of unknown function.

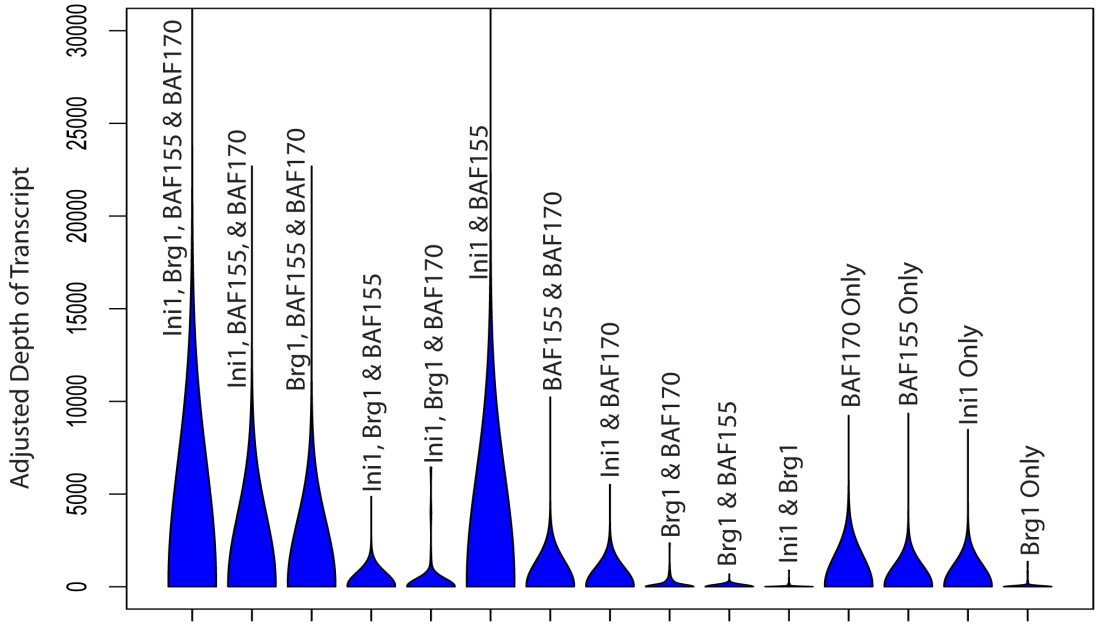


Figure 3.5 Violin plots of expression values across all possible SWI/SNF subunit occurrences. Violin plots display the probability density function plotted against the adjusted depth (i.e. expression) values from Morin et al [37]. Transcript counts for each category are given in Table S7. We find that some combinations of subunits are associated with transcripts with higher expression levels while other combinations are associated with transcripts with lower expression levels.

3.5.5 SWI/SNF targets genes involved in nuclear function and cancer pathways

Prior research has shown that a variety of signaling cascades are linked to SWI/SNF [25]. To gain further insights into potential actions of SWI/SNF components we examined the underlying Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) designations of their gene targets to determine significantly overrepresented annotations and pathways (Tables 3.5 and B.8). SWI/SNF gene targets were associated with 'Pathways in cancer' and several specific cancers types, e.g. chronic myeloid leukemia and pancreatic cancer. A number of signaling pathways and cellular processes that are "hallmarks of cancer" [41] were also overrepresented among the gene targets of Ini1, Brg1, BAF155 and BAF170. These include the Wnt, ErbB, p53, MAPK,

and insulin signaling pathways, and processes endemic to oncogenesis and cancer progression such as DNA repair, the cell cycle and apoptosis. From these analyses we surmise the recruitment of SWI/SNF components is likely to influence the molecular basis of cancer through several potential mechanisms.

The SWI/SNF-enriched pathways are highly interconnected. Using the 49,555 SWI/SNF targets we identified a total of 24 KEGG signaling or biochemical pathways (Figure 3.6, yellow nodes). Interestingly, these pathways partition into three groups (Figure 3.6, panels A-C). Two of the groups (Figure 3.6A and 3.6B) comprise sets of pathways exhibiting at most one degree of separation, e.g. ‘inositol phosphate metabolism’ and ‘amino sugar and nucleotide sugar metabolism’. The third group (Figure 3.6C) consists of three pathways that are unrelated to any other pathways in the KEGG database. As displayed in Figure B.2 directly related pathways such as ‘p53 signaling’ and the ‘cell cycle’ have shared components and many of the genes encoding these components are occupied by SWI/SNF factors. Thus, our results demonstrate that SWI/SNF is involved in many closely related signaling pathways and cellular processes and may help serve to coordinate expression of genes involved in these processes.

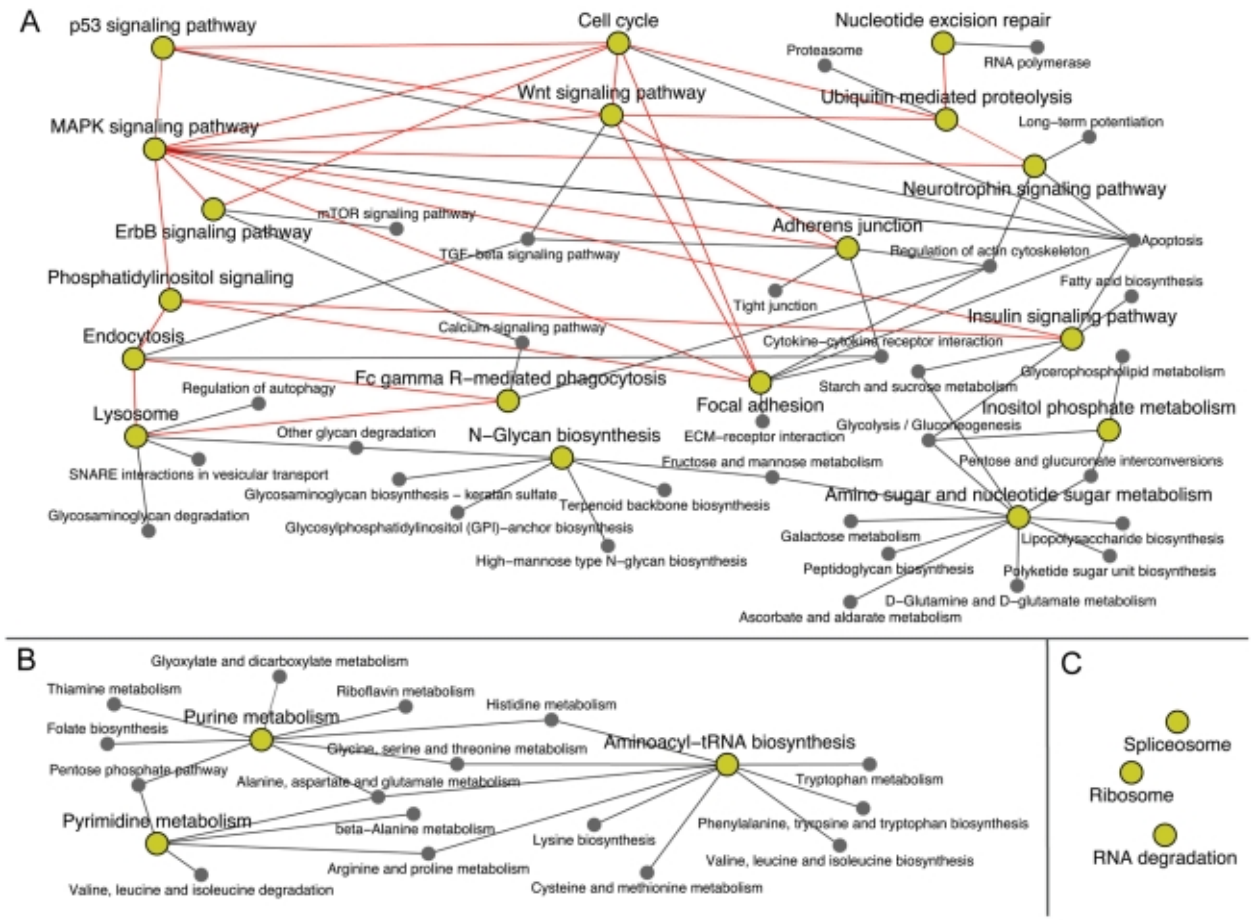


Figure 3.6 Network of overrepresented and other related KEGG pathways identified using SWI/SNF ChIP-Seq union regions. The KEGG pathways identified as overrepresented using the 49,555 SWI/SNF ChIP-Seq union regions are shown as yellow nodes and each of their related, KEGG-designated pathways are also displayed. Pathways related to the overrepresented pathways but that are not overrepresented themselves are shown as gray nodes. Red edges connect yellow nodes. Three of the nodes, spliceosome, ribosome and RNA degradation, are distinct and unrelated to other pathways according to the KEGG

Table 3.5 Significant pathways and biological processes associated with SWI/SNF union ChIP-Seq regions.

Overrepresented categories for the SWI/SNF union regions[1]	Benjamini-corrected p-values
KEGG hsa05200:Pathways in cancer	4.7×10^{-8}
KEGG hsa05212:Pancreatic cancer	4.9×10^{-3}
KEGG hsa05222:Small cell lung cancer	1.7×10^{-3}
KEGG hsa05211:Renal cell carcinoma	2.5×10^{-3}
KEGG hsa05220:Chronic myeloid leukemia	3.4×10^{-4}
KEGG hsa05215:Prostate cancer	8.9×10^{-3}
KEGG hsa05210:Colorectal cancer	1.7×10^{-3}
KEGG hsa05016:Huntington's disease	2.1×10^{-4}
KEGG hsa05010:Alzheimer's disease	1.4×10^{-3}
KEGG hsa04010:MAPK signaling pathway	1.3×10^{-5}
KEGG hsa04012:ErbB signaling pathway	2.6×10^{-3}
KEGG hsa04115:p53 signaling pathway	3.4×10^{-4}
KEGG hsa04310:Wnt signaling pathway	7.8×10^{-3}
KEGG hsa04910:Insulin signaling pathway	8.9×10^{-3}
KEGG hsa04070:Phosphatidylinositol signaling system	7.9×10^{-3}
KEGG hsa04120:Ubiquitin mediated proteolysis	8.4×10^{-7}
KEGG hsa03040:Spliceosome	8.6×10^{-10}
GO:0051056 regulation of small GTPase mediated signal transduction	1.8×10^{-3}
GO:0007049 cell cycle	3.4×10^{-34}
GO:0006260 DNA replication	4.5×10^{-10}
GO:0051301 cell division	4.1×10^{-18}
GO:0006281 DNA repair	5.6×10^{-20}
GO:0006915 apoptosis	8.8×10^{-9}
GO:0051276 chromosome organization	1.5×10^{-11}
GO:0016568 chromatin modification	2.1×10^{-17}
GO:0006357 regulation of transcription from RNA polymerase II promoter	6.1×10^{-7}
GO:0034470 ncRNA processing	2.4×10^{-15}
GO:0001701 in utero embryonic development	1.5×10^{-5}

[1]Overrepresented terms were determined using DAVID tools for Ensembl genes corresponding to the 49,555 SWI/SNF union regions for Benjamini corrected p-values <0.01. A complete list is available in Table B.7.

3.5.6 SWI/SNF components associate with proteins involved in multiple aspects of gene regulation and are nodes in a highly integrated network

The genomic binding data demonstrates that SWI/SNF localization is coupled with a broad range of functional elements, suggesting that SWI/SNF may also be found with a broad range of associated proteins. To further examine the scope of SWI/SNF's roles in the nucleus we analyzed proteins associated with SWI/SNF subunits using co-immunoprecipitation followed by mass spectrometry. The SWI/SNF components Ini1,

BAF155, BAF170, Brg1, Brm and ARID1A were immunoprecipitated from HeLa S3 nuclei, the resulting proteins were gel-separated and peptides were generated for analysis by mass spectrometry (See Materials and Methods; Table B.9). In addition to the factor-specific antibodies, parallel immunoprecipitations were performed using non-specific IgG antibodies. Proteins identified in these "control IgG" immunoprecipitations were excluded as potential SWI/SNF co-purifying factors.

We identified a total of 101 proteins that were specifically associated with at least one of the SWI/SNF components assayed (Figure 7, turquoise edges; Table B.10). Of the non-SWI/SNF subunits detected, 5 of these interactions were found previously in HeLa cells (e.g. estrogen receptor alpha [42], and 96 were new to this study. Interestingly one of the novel interactions we observed in HeLa cells, BAF155 with NUF2, has been previously observed in yeast between the yeast homolog of BAF155 (SWI3) and NUF2 [15]. Using the 101 nodes that we identified as proteins co-purifying with SWI/SNF in our undirected approach we ascertained overrepresented GO categories (Table 3.6). Several of these designations such as 'cell cycle' and 'chromosome organization' coincide with the categories obtained from GO and pathway analyses of SWI/SNF ChIP-Seq targets, suggesting the possibility of highly interactive network structures.

Many of the proteins that were novel to this study reinforce and expand upon other published reports of SWI/SNF characterizations. For example SWI/SNF components have been localized by immunofluorescence to mitotic kinetochores and spindle poles [43], and Brg1-deficient mice show dissolution of pericentromeric heterochromatin domains [44]. From our immunoprecipitations BAF155 and BAF170 were associated with a number of kinetochore and centrosomal proteins (e.g. BUB1B, CENPE and NUF2,

Figure 8, green circles). The role of SWI/SNF in the maintenance of kinetochore and spindle function is unknown. We detected a variety of transcription factor activators and repressors (e.g. NFB1, NFB2, RelA, PML and NFX1) as well as DNA repair (ERCC5 and RAD50) and cell cycle (e.g. CCNB3 and CDCA2) proteins (Figure S3). Some of the SWI/SNF interacting proteins themselves interact with one another. For example we detected several different proteins integral to estrogen and insulin signaling (Figure 7; Table B.10). We also identified proteins associated with only one SWI/SNF factor; these may either be interactions with a specific SWI/SNF component or an inability to detect the protein in the immunoprecipitations.

We developed an expanded network of SWI/SNF associations by including proteins that were found by others to co-purify with SWI/SNF subunits (Figure 7, black edges). Only those factors that showed a one-degree separation with a SWI/SNF component in HeLa cells are displayed and all interactions are annotated in Table B.10. SWI/SNF interacting proteins are associated with numerous UniProt keywords (Figure 8; [45]). Overall these results suggest a role for SWI/SNF components in a wide array of nuclear processes and diseases. Some of these processes may take place in nuclear substructures. Higher order chromatin structure is facilitated by the nuclear lamina and tethering of genes to the nuclear periphery is one epigenetic mechanism of gene regulation [46-47]. Intriguingly we and others have detected SWI/SNF components with various nuclear envelope-associated proteins (Figure 7 and Table B.10) including lamin A, EMD (emerin) and BAF/BANF1 (Barrier to Autointegration Factor, which although similar in name is not a SWI/SNF subunit). Two of the nuclear membrane proteins, SYNE1 and C14orf49, that we isolated in association with BAF155 are part of LINC complexes that

link the nucleoskeleton and cytoskeleton [48-49].

Table 3.6 Over-represented annotations from proteins identified as co-purifying with SWI/SNF in this study.

Overrepresented categories for SWI/SNF co-associated proteins[1]	Benjamini-corrected p-values
GO:0007049 cell cycle	3.8×10^{-4}
GO:0000279 M phase	2.8×10^{-4}
GO:0015630 microtubule cytoskeleton	8.4×10^{-4}
GO:0006323 DNA packaging	7.9×10^{-3}
GO:0000793 condensed chromosome	8.1×10^{-3}
GO:0051276 chromosome organization	9.9×10^{-3}
GO:0006333 chromatin assembly or disassembly	1.0×10^{-2}
GO:0005813 centrosome	1.2×10^{-2}
GO:0034728 nucleosome organization	1.5×10^{-2}
GO:0005815 microtubule organizing center	2.2×10^{-2}
GO:0016584 nucleosome positioning	2.8×10^{-2}
GO:0000777 condensed chromosome kinetochore	2.9×10^{-2}

Overrepresented terms were determined using DAVID for Ensembl genes corresponding to the 101 proteins we identified as co-associated with a SWI/SNF factor as determined by IP-mass spectrometry. We considered Benjamini-corrected p-values <0.05. A complete list is available in Table B.8.

3.5.7 A fraction of SWI/SNF regions are associated with the nuclear lamina

Numerous studies point to a high degree of functional organization in cell nuclei [46]. Emerging nuclear organization models would benefit greatly from a catalogue of processes and chromatin characteristics mapped to particular genomic elements. For example, the nuclear lamins are thought to influence chromatin organization, DNA replication and transcription [47,50]. Our immunoprecipitation results demonstrating that SWI/SNF components are associated with lamin A/C (Figure 3.7 and Table B.10) along with immunoprecipitation, immunolocalization and cell fractionation experiments from others demonstrating an association between SWI/SNF and nuclear lamina (e.g. emerin Figure 3.7; [51]) prompted us to investigate whether SWI/SNF and the lamins can be located to the same genomic sequences.

We isolated lamin A/C and lamin B ChIP DNA from HeLa S3 nuclei and

performed ChIP-chip on tiling arrays covering the ENCODE pilot regions (see Materials and Methods and Table S1). Most of the 1,770 lamin A/C regions mapped to H3K27me3 domains (76%; 1,337/1,770) whereas the 1,270 lamin B regions were less commonly associated with H3K27me3 (29%; 372/1,270). Comparing regions where signal was detectable for the SWI/SNF, lamin A/C and lamin B experiments revealed that SWI/SNF has a much higher overlap with lamin B than lamin A/C. We found that 38% (297/784) of SWI/SNF sites are within 100 bp of a lamin B region whereas only 5% (41/784) of SWI/SNF sites are within 100 bp of a lamin A/C region (Table 3.7). For both lamin types the colocalization with SWI/SNF regions is significant relative to random target lists (lamin B, $p < 1x10^{-16}$; lamin A/C, $p < 1 \times 10^{-15}$; GSC tests). SWI/SNF-lamin B intersecting regions contained approximately the same proportion of CTCF sites in the ENCODE regions as did all SWI/SNF sites in the ENCODE regions ($p > 0.05$ hypergeometric test; Table 3.7). Enhancers are underrepresented in the SWI/SNF-lamin B regions relative to all SWI/SNF locations in the ENCODE regions ($p < 1.9 \times 10^{-36}$; hypergeometric test). The SWI/SNF-lamin B regions are overrepresented for Pol II ($p < 2.9 \times 10^{-39}$; hypergeometric test) and 5' ends ($p < 7.3 \times 10^{-37}$; hypergeometric test) relative to all SWI/SNF locations in the ENCODE regions.

In crosslinked chromatin SWI/SNF is detected primarily with lamin B, but as noted from the above mass spectrometry experiments, in solubilized, non-cross-linked cells SWI/SNF is detected with lamin A/C and not lamin B (Figures 3.1 and 3.7). We interpret these results to indicate that SWI/SNF, lamin A/C and lamin B co-associate in different nuclear contexts but are all part of a broader interacting network with specific sub-associations.

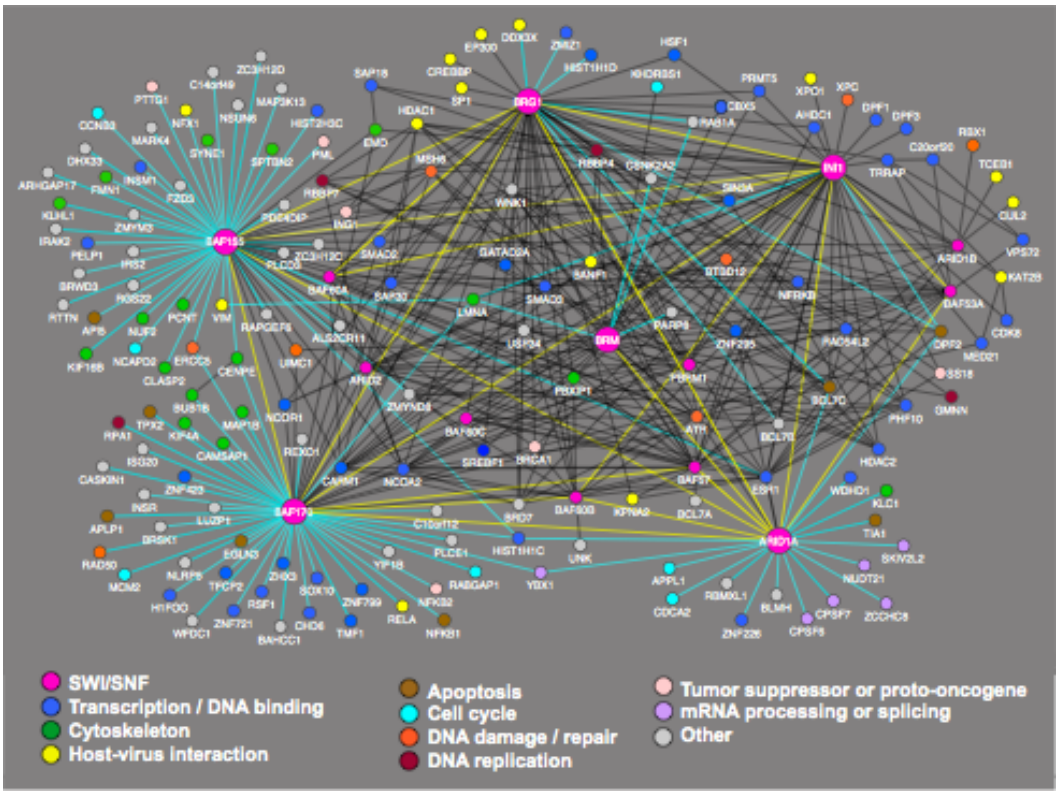


Figure 3.7 Network of proteins that have been shown to co-purify with SWI/SNF factors. 158 proteins have been shown to co-purify with SWI/SNF factors in HeLa cells, either by our IP-mass spectrometry experiments or in other studies. Interactions are annotated in Table B.10. As indicated in the figure key, pink circles denote SWI/SNF components; the larger pink circles are SWI/SNF factors used as bait in this study. Blue edges denote interactions detected in this study. Yellow edges indicate interactions between SWI/SNF factors themselves that were detected in this study. Black edges indicate interactions from other published sources. As noted in Table B.10, the studies used a variety of biochemical methods and SWI/SNF factors were either bait or prey. Non-SWI/SNF factors are color-coded according to UniProt keywords [45].

Table 3.7 Co-occurrence of SWI/SNF factors and lamins in the ENCODE regions.

Factor or Feature	Lamin A/C[1]	Lamin B[2]	SWI/SNF[3]
Ini1	22 (54%)	186 (63%)	598 (76%)
Brg1	6 (15%)	75 (25%)	204 (26%)
BAF155	31 (76%)	216 (73%)	588 (75%)
BAF170	23 (56%)	155 (52%)	394 (50%)
CTCF	7 (17%)	60 (20%)	146 (19%)
Enhancers	8 (20%)	32 (11%)	289 (37%)
RNA Pol II	23 (56%)	214 (72%)	335 (43%)
5' Ends	15 (37%)	186 (63%)	274 (35%)
5' Ends, expressed	12 (29%)	128 (43%)	93 (12%)
5' Ends, not expressed	4 (10%)	66 (22%)	191 (24%)

[1]A total of 41 SWI/SNF regions intersect a lamin A/C region.
[2]A total of 297 SWI/SNF regions intersect a lamin B region.
[3]There are 784 SWI/SNF high-confidence union regions that were used for comparison with lamin ChIP-chip array data.

3.5.8 Association of SWI/SNF with DNA replication origins

SWI/SNF and the lamins have each been implicated in DNA replication (see above; [52-53]). One of the proteins we detected as associated with SWI/SNF is the replication protein RepA and another regulator of DNA replication, geminin, has been found to co-purify with SWI/SNF in HeLa cells (Figure 3.7, red circles; [54]). We investigated whether there might be a relationship among SWI/SNF, lamins and DNA replication origins. We obtained a set of 282 DNA replication origins identified in HeLa cells for the ENCODE regions ([55]; Table B.1). Of these 282 replication origins, 90 (32%) occur within 100 bp of a SWI/SNF region ($p < 1x10^{-16}$, GSC test), 86 (31%) occur at the 5' ends of protein-coding genes and 151 (54%) occur within 100 bp of a lamin B region. In contrast to lamin B, only 17% (48/282) of the replication origins were near a lamin A/C region. These results are consistent with nuclear staining patterns observed in mouse 3T3 cells showing colocalization between lamin B and sites of DNA replication whereas the same colocalization patterns were not observed for replication foci and lamin A [52].

Of the 86 replication origins in promoter regions, 88% (76/86) intersected a lamin

B region and most (78% or 67/86) were within a 100 bp of a SWI/SNF region. These data indicate that SWI/SNF components are located near many DNA replication origins, particularly those located in promoter regions. The coincidence of chromatin remodeling factors, promoters, lamins and replication origins at the same subset of genomic regions suggests that these loci may be particularly favorable for the formation of both DNA and RNA polymerase assembly and chromatin tethering. As shown in Figure 3.1 the interplay among these elements as well as with Pol II, CTCF and heterochromatin regions is complex and interwoven, such that each may share many different supporting and counteracting roles.

3.6 Discussion

SWI/SNF performs a crucial function in gene regulation and chromosome organization by directly altering contacts between nucleosomes and DNA. In the work presented here we undertook a two-pronged approach (ChIP-Seq and IP-mass spectrometry) to move towards a more thorough understanding of these functions. Our ChIP-Seq analyses demonstrate that SWI/SNF components overlap extensively with important regions that require tight control of the dynamics of nucleosome occupancy such as promoters, enhancers and CTCF sites. Not only does the SWI/SNF complex change the accessibility of DNA but it also acts in concert with an extensive host of cooperating factors, thereby facilitating combinatorial control among various genomic elements. In addition to our ChIP-Seq results, the diversity and number of proteins that co-purify with SWI/SNF as identified in our mass spectrometry experiments further supports SWI/SNF's involvement with a variety of functionally distinct complexes.

RNA polymerases II and III are extensively colocalized with SWI/SNF components.

Studies of transcription in HeLa cells have estimated that the number of active RNA II polymerases exceeds the number of transcriptionally active sites by at least one order of magnitude, leading to the proposal of "transcription factories" [1-3]. The number of RNA Pol II transcription factories in HeLa cells has been estimated between 5,000 and 8,000 where each factory can be typified by several looped loci, their resulting transcripts and distal elements such as enhancers. We infer that SWI/SNF regions are prevalent in transcriptional assemblages and their associated regulatory loops, given that >90% of our high-confidence union targets are associated with genic or regulatory regions and that 65% of Pol III and 84% of Pol II regions colocalize with at least one SWI/SNF factor (Tables 3.4, B.5 and B.6).

Interestingly we observed that SWI/SNF components often occur independently of each other and in various configurations across the genome, and similarly our mass spectrometry data point to heterogeneity of SWI/SNF complexes. We speculate that several mechanisms may underlie these various configurations and their associated genomic features, including 1) synergism or antagonism of the individual SWI/SNF factors in influencing expression (e.g. Figure 3.5); 2) failure to detect individual subunits due to epitope masking as a consequence of variation with local environments; 3) the capture of incomplete complexes that may in fact be completed upon superposition of genome-wide 3C data once such data become available (e.g. Figure 3.4); 4) the existence of SWI/SNF sub-complexes that deviate from the conventional composition of SWI/SNF assemblies (e.g. [56]) or 5) the capture of intermediates in a multistep assembly or remodeling process. This last view is consistent with a model of stochastic assembly that may occur through intermediate interactions and that has been described for several other

large, multifactor complexes such as RNA polymerases and associated transcription factors [57], spliceosomes [58], and DNA repair complexes [59].

As shown in Figure 3.6 SWI/SNF occurs throughout many interconnected pathways. The assembly of functional SWI/SNF complexes at many locations in the genome may require the activation of one or more of these related pathways. Consequently some of the SWI/SNF associated regions we observed may reflect constitutive binding of partially assembled complexes that may be poised to receive additional signal inputs for subsequent regulatory activity. Indeed it has been shown that SWI/SNF components are present at regulatory regions even in the absence of stimulatory conditions or tissue-specific cofactors. For example Brg1 is present constitutively at the interferon-inducible genes *IFITM3* [60] and *CIITA* [30] in unstimulated HeLa cells, which is consistent with our own finding of Brg1 and Ini1 at *IFITM3* and various combinations of BAF155, BAF170, Ini1 and Brg1 at different elements in *CIITA*. In solution SWI/SNF factors are associated constitutively with RelB (HEK293 cells, [61]), RelA, NFkB1 and NFkB2 (HeLa cells, this study), the glucocorticoid receptor (T4D7 cells, [62]) and estrogen receptor alpha (HeLa cells, this study and [42]; SW13 cell extracts, [63]). The prevalence of SWI/SNF and the high degree of connectivity of its overrepresented pathways implies that SWI/SNF may assist in many related processes and may even facilitate crosstalk across many constituents of the transcriptional machinery. Notably SWI/SNF binds in the genes of its own subunits (Table B.19) suggesting that SWI/SNF may contribute to auto- and cross-regulation of its subunit levels. Loss-of-function of a particular subunit, as may occur in certain cancers, could initiate oscillations and alter the relative abundance of the levels of the other

SWI/SNF subunits through a variety of feedback and feed-forward loops. Aberrant SWI/SNF expression has been proposed to result in new combinatorial assemblies of SWI/SNF, some of which may deleterious [64].

The gene attributes revealed by our ChIP-Seq data substantiate that SWI/SNF is proximal to targets that comprise sets of fundamental biological processes. Many of the functional categories we found to be significantly overrepresented have disease implications, especially as related to cancer (Figure B.2). For example failures in DNA repair and unchecked cell cycle activity are common characteristics of pre-cancerous cells, and our SWI/SNF analyses identified the p53 and MAPK signaling pathways, which are well known for maintaining checkpoint functions. Growth dysregulation particularly in the context of hormone signaling is another common cancer phenotype. Extracellular growth signals are transduced from the cell membrane to the nucleus by the ErbB, insulin and phosphtidylinositol signaling pathways, all of which we recovered as overrepresented (Table 3.5). The existence of phosphoinositide signaling in the nucleus and the ability of Brg1 to act as an effector for phosphatidylinositol 4,5-bisphosphate (PIP_2) raises the prospect of several levels of control of this signaling pathway with respect to SWI/SNF [65], a hypothesis that can be examined in future studies.

Several of the overrepresented pathways we identified through our ChIP-Seq analyses share proteins detected in SWI/SNF co-purification experiments, thereby providing a resource to explore potential, highly-interactive network structures. For example we found that genes with products critical for ‘nucleotide excision repair’ were enriched using our SWI/SNF union list (Figure 3.6). Within this pathway the excision repair protein ERCC5 co-purified with both BAF155 and BAF170 in our IP

(immunoprecipitation)-mass spectrometry experiments. The excision repair protein, XPC, associates with SWI/SNF in response to UV irradiation in HeLa cells, and BRCA1 and ATR also cooperate with SWI/SNF in DNA repair (Figure 3.7; Table B.10; [66]). Thus we speculate SWI/SNF may participate in DNA repair through both transcriptional regulation as well as recruitment to regions undergoing repair.

Our study uses two strategies to attempt to comprehensively collect a SWI/SNF interaction network. We limited our network to a single model system, HeLa cells, because many attributes of SWI/SNF have been documented in these cells and it has been noted that SWI/SNF associations vary by cell type [67]. We extensively collated SWI/SNF protein interactions described in the literature. This undertaking was necessary because many of the proteins described in the literature as co-associated with SWI/SNF factors are not represented in interaction databases such as BioGRID, Molecular Interactions Database (MINT), IntAct, Human Protein Reference Database (HPRD), Nuclear Protein Database (NPD) and Interologous Interaction Database (I2D). Therefore we attempted to comprehensively collect such information to overcome these limitations. In total 158 SWI/SNF interacting proteins have been described in HeLa cells (Figure 3.8 and Table B.10), which is similar to the number of SWI/SNF interacting proteins that have been described in other cell types [67]. Published molecular associations that were not discerned here might be due to interactions that are: 1) transient or of low affinity, 2) dependent on a specific set of biochemical conditions or 3) undetectable due to masking by the presence of more abundant protein(s) of similar size. In working with protein interaction data, similar degrees of overlap have been noted when comparisons are made across data sets [68-69] and even in a well-studied model such as yeast, mass

spectrometry analyses have found a plasticity of complexes and many previously undetected interactions [70-72]. From the ChIP-Seq and ChIP-chip results we expected that CTCF and lamin B may be among the proteins that co-associate with SWI/SNF, however neither of these factors was recovered in any of the non-directed experiments (Table B.10), including a CTCF immunoprecipitation-mass spectrometry experiment performed in HeLa cells. In addition to the above considerations one possibility is that CTCF or lamin B may associate more strongly with one of the SWI/SNF factors not studied, e.g. BAF53A or one of the BAF60 subunits.

SWI/SNF is most often described in a chromatin remodeling context however data derived from a variety of sources suggests that SWI/SNF has other facets. It is possible that not all of SWI/SNF's functions involve DNA localization and therefore other types of global experiments, such as the IP-mass spectrometry, are valuable as first steps towards recognizing previously unknown roles. Unlike cytoplasmic compartments, nuclear compartments are not separated by a physical barrier but rather are functional assemblies that are typically organized around sets of molecules engaged in common functions. Data from both ChIP-Seq and IP-mass spectrometry illuminate the sectors in which SWI/SNF operates and the integration of these two methods is better than each alone for furnishing a broad comprehension of SWI/SNF action. For example ChIP-Seq enables the global identification of SWI/SNF chromosomal elements except for those regions with highly repetitive sequence such as human centromeres (Figure 3.2A). In this respect IP-mass spectrometry is complementary to ChIP-Seq because it strongly suggests that SWI/SNF occurs at kinetochores as evidenced by its co-purification with CENPE, NUF2, BUB1B and CLASP2 (Figures 3.7 and 3.9). In addition to kinetochore proteins

the SWI/SNF co-purification experiments also uncovered proteins from other substructures including centrosomes, microtubules, the nuclear periphery and PML nuclear bodies, the latter of which is characterized by cryptic foci of PML (promyelocytic leukemia protein) and has been implicated in a variety of diseases [73]. The ChIP-Seq and IP-mass spectrometry data are synergistic as well. Notably both methods found an overrepresentation of regions or proteins enriched for 'cell cycle' and 'chromosome organization'. One possible inference from these studies is that SWI/SNF is well positioned to integrate signals across multiple signaling pathways both by its presence in a variety of cellular structures and its role in gene regulation through chromatin remodeling.

A fraction of SWI/SNF complexes co-associate with elements of the nuclear periphery where they are well situated to contribute to the nuclear organization and position-dependent gene expression (Figure 3.7; [51]). We found that in crosslinked cells SWI/SNF localizes more widely with lamin B than lamin A whereas in non-crosslinked cells SWI/SNF co-purifies with lamin A. As mentioned above lamin B may have escaped detection in SWI/SNF protein interaction studies. A related possibility is that SWI/SNF may exist in different nuclear pools that have varying solubilities and associations, such that recovery of particular SWI/SNF complexes depends upon the proteins with which SWI/SNF is associated. For example lamins A and B are known to have different nucleoplasmic mobilities and localization patterns [50,52]. Immunolocalization experiments in HeLa nuclei have revealed that the A/C- and B-type lamins form distinct meshworks with occasional points of intersection [50], which is consistent with the interspersed patterns of lamin A/C and B that we detected (Figure 3.1). Hence it is

reasonable to expect that SWI/SNF associated with lamin A would behave differently than when associated with lamin B. We surmise that in a chromatin context the dominant association of SWI/SNF with the nuclear lamins occurs in regions where lamin B is present. The purification of SWI/SNF with lamin A may indicate other biological roles, such as cell cycle progression or nuclear assembly [74-75].

Gaining a more detailed understanding of SWI/SNF's activities in or near various heterochromatin environments will be central to comprehending nuclear events over the cell cycle as well as during development. Among the numerous molecular and epigenetic factors that have been found to affect heterochromatin formation or maintenance, the heterochromatin protein 1 alpha (HP1, also known as CBX5; Figure 3.7) and Polycomb complexes (PcG) are of particular relevance to SWI/SNF [76-78]. Polycomb complexes promote gene silencing by catalyzing the trimethylation of H3K27 in its target regions, and SWI/SNF antagonizes this epigenetic silencing [79]. It is tempting to speculate that SWI/SNF found near the edges of H3K27me3 domains (Figure 3.1A and 3.1C) may be contributing to the establishment or maintenance of boundary elements. SWI/SNF may also engage in heterochromatin dynamics through its interaction with HP1, which is often located in the centromeric regions (reviewed in [80]). Curiously HP1 interacts with the lamin B receptor [81] thus providing a potential bridge between heterochromatin and the inner nuclear membrane. Both H3K27me3 and lamin B are associated with spatially regulated genes whose conversion between active and inactive states depends on access to their regulatory regions, as may be conferred by SWI/SNF.

The work presented here provides new insights into the scope of SWI/SNF's influence in gene regulation and nuclear organization. The integration of numerous

studies is beginning to reveal the complexities contributing to the regulation of any given locus. Contemporary models of transcriptional control propose that a series of factors transiently associate with a regulatory region before a decisive event tilts these intermediate reactions towards a productive outcome [57,82]. SWI/SNF may contribute to such intermediate reactions or trigger switches between inactive and active states. The capacity for SWI/SNF to preserve many aspects of homeostasis also makes it vulnerable to being ensnared for aggressive cell proliferation. Our work demonstrates that SWI/SNF in particular and perhaps chromatin remodeling proteins in general will contribute unique insights to our understanding of gene regulation and disease mechanisms through the integration of target regions, spatial positioning and functional annotations. For example the co-occurrence of SWI/SNF with centrosomes, microtubules, kinetochores and the nuclear periphery may suggest that a pool of SWI/SNF is sequestered by these structures during mitosis to assist in the post-mitotic reformation of chromosomal territories. Our collective findings help inform a comprehensive view of SWI/SNF function as well as form a valuable compendium for future studies of nuclear functions as related to chromatin remodeling.

3.7 Materials and Methods

3.7.1 Chromatin immunoprecipitations

Suspension HeLa S3 cells were cultured by the National Cell Culture Center (Biovest International Inc., Minneapolis, MN) in modified minimal essential medium (MEM), supplemented with 10% FBS at 37°C in 5% CO2, to a density of 6 x105 cells/mL. Cells were fixed with 1% formaldehyde at room temperature for 10 min. Fixation was terminated with 125 mM glycine (2 M stock made in 1x PBS).

Formaldehyde-fixed cells were washed in cold Dulbecco's PBS (Invitrogen) and swelled on ice in a 10-mL hypotonic lysis buffer [20 mM Hepes (pH 7.9), 10 mM KCl, 1 mM EDTA (pH 8.0), 10% glycerol, 1 mM DTT, 0.5 mM PMSF, and Roche Complete protease inhibitors, Cat#1697498]. To isolate nuclei, whole cell lysates were

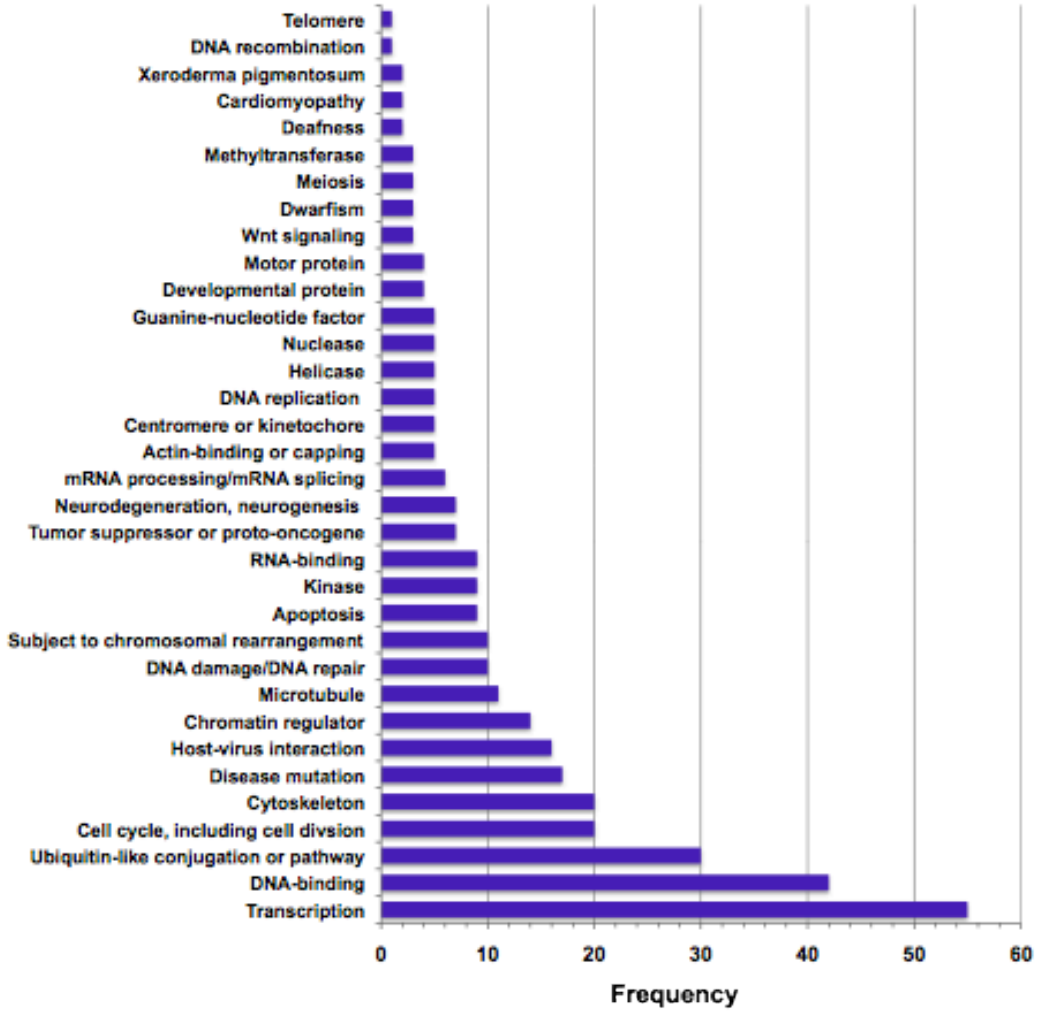


Figure 3.8 Histogram showing the frequencies of UniProt keywords for proteins that co-purify with SWI/SNF factors. Keywords shown were retrieved from the UniProt database [45] for proteins that co-purify with a SWI/SNF factor, as annotated in Table B.10.

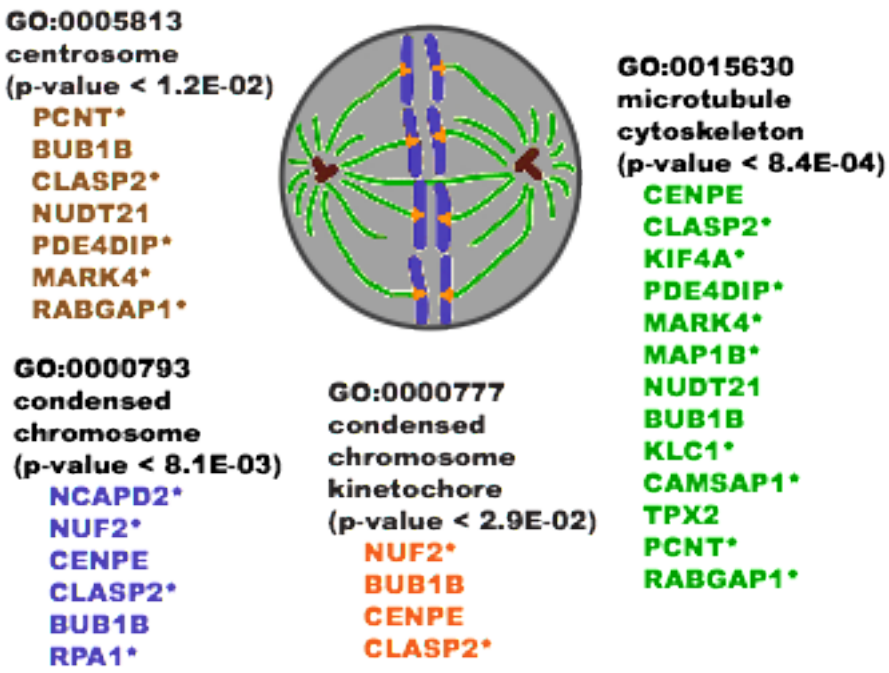


Figure 3.9 Illustration showing overrepresented GO 'cellular component' categories for SWI/SNF co-purifying proteins. Overrepresented GO 'cellular component' categories are displayed for proteins we detected by IP-mass spectrometry. Centrosomal proteins are shaded brown, chromosomal proteins are blue, kinetochore proteins are orange and cytoskeletal proteins are green. Genes encoding starred proteins are targets of SWI/SNF as identified by ChIP-Seq. Based on these annotations SWI/SNF is associated with multiple cellular components.

homogenized with 30 strokes in a 7 mL Dounce homogenizer (Kontes, pestle B). Nuclear pellets were collected by centrifugation and lysed in 10 mL of RIPA buffer per 3 x 108 cells [RIPA buffer: 10 mM Tris-Cl (pH 8.0), 140 mM NaCl, 1% Triton X-100, 0.1% SDS, 1% deoxycholic acid, 0.5 mM PMSF, 1 mM DTT, and protease inhibitors]. Chromatin was sheared with an analog Branson 250 Sonifier (power setting 2, 100% duty cycle for 7 x 30-s intervals) to an average size of less than 500 bp, as verified on a 2% agarose gel. Lysates were clarified by centrifugation at 20,000 x g for 15 min at 4°C.

Clarified nuclear lysates from 1 x 108 cells were agitated overnight at 4°C with 20 µg of one of the following antibodies: 1) anti-Ini1 (C-20), Santa Cruz Biotechnology, sc-16189; 2) anti-BAF155 (H-76), Santa Cruz Biotechnology, sc-10756; 3) anti-BAF170 (H-116), Santa Cruz Biotechnology, sc-10757; 4) anti-Brg1 (G-7), Santa Cruz

Biotechnology, sc-17796; 5) anti-lamin A/C (H-110), Santa Cruz Biotechnology, sc-20681; 6) anti-lamin B antibody, EMD Biosciences, NA12; or 7) normal IgG, Santa Cruz Biotechnology, sc-2025. Antibody incubations were followed by addition of either protein A (Millipore #16-156) or protein G agarose beads (Millipore #16-266). Beads were permitted to bind to protein complexes for 60 min at 4°C. Immunoprecipitates were washed three times in 1x RIPA, once in 1x PBS, and then eluted in 1xTE/1%SDS. Crosslinks were reversed overnight at 65°C. ChIP DNA was purified by incubation with 200 μg/mL RNase A (Qiagen #19101) for 1 h at 37°C, with 200 μg/mL proteinase K (Ambion AM2548) for 2 h at 45°C, phenol:chloroform:isoamyl alcohol extraction, and precipitation with 0.1 volumes of 3M sodium acetate, 2 volumes of 100% ethanol and 1.5 μL of pellet paint (Novagen #69049-3). ChIP DNA prepared from 1 x 108 cells was resuspended in 50 μL of Qiagen Elution Buffer (EB). Three biological replicates were prepared per antibody.

3.7.2 Construction and sequencing of Illumina libraries

ChIP-Seq libraries were prepared and sequenced as previously described [26,83]. Biological replicates for each factor were converted into separate and distinct libraries. To summarize, ChIP DNA samples were loaded onto Qiagen MinElute PCR columns, eluted with 15 μL of Qiagen buffer EB, size-selected in the 100-350 bp range on 2% agarose E-gels (Invitrogen) and gel-purified using a Qiagen gel extraction kit. DNA was end-repaired and phosphorylated with the End-It kit from Epicentre (Cat# ER0720). The blunt, phosphorylated ends were treated with Klenow fragment (3 to 5 exo minus; NEB, Cat# M0212s) and dATP to yield a protruding 3-'A' base for ligation of Illumina adapters (100 RXN Genomic DNA Sample Prep Oligo Only Kit, Part# FC-102-1003),

which have a single 'T' base overhang at the 3 end. After adapter ligation (LigaFast, Promega Cat# M8221) DNA was PCR-amplified with Illumina genomic DNA primers 1.1 and 3.1 for 15 cycles by using a program of (i) 30 s at 98 °C, (ii) 15 cycles of 10 s at 98°C, 30 s at 65°C, 30 s at 72°C, and (iii) a 5 min extension at 72°C. The final libraries were band-isolated from an agarose gel to remove residual primers and adapters. Library concentrations and A_{260}/A_{280} ratios were determined by UV-Vis spectrometry on a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific). Purified and denatured library DNA was capture on an Illumina flowcell for cluster generation and sequenced on an Illumina Genome Analyzer II following the manufacturer's protocols [84].

3.7.3 Identification of proteins by mass spectrometry

Immunoprecipations were performed using the same conditions as for ChIP experiments except the HeLa S3 cells were not crosslinked. In addition to the ChIP antibodies described above we also used anti-Brm, Abcam Cat# ab15597 and anti-BAF250a (PSG3), Santa Cruz Biotechnology, sc-32761. Complexes were resolved on BioRad 4-20% precast Tris-HCl gels (Cat# 161-1159) such that a single gel was used for each specific antibody and normal IgG immunoprecipitation pair. Gels were silver stained using Pierce SilverSNAP stain for mass spectrometry (Cat# 24600) and each lane was excised into 10-12 molecular weight regions. Gel slices were destained, dried in a Savant speed-vac and digested overnight at 42C with Sigma's Trypsin Profile IGD kit for in-gel digests (Cat# PP0100). Following the overnight incubation the liquid was removed from each gel piece and volume reduced by drying to approximately 10 μL. The individual gel slices were analyzed separately.

3.7.4 Mass spectrometry

The samples were subjected to nanoflow chromatography using nanoAcquity UPLC system (Waters Inc.) prior to introduction into the mass spectrometer for further analysis. Mass spectrometry was performed on a hybrid ion trap LTQ Orbitrap mass spectrometer (Thermo Fisher Scientific) in positive electrospray ionization (ESI) mode. The spectra were acquired in a data dependent fashion consisting of full mass spectrum scan (300-2000 m/z) followed by MS/MS scan of the 3 most abundant parent ions. For the full scan in the Orbitrap the automatic gain control (AGC) was set to 1x106 and the resolving power for 400 m/z of 30,000. The MS/MS scans were done using the ion trap part of the mass spectrometer at a normalized collision energy of 24 V. Dynamic exclusion time was set to 100 s to avoid loss of MS/MS spectral information due to repeated sampling of the most abundant peaks.

Sequence data from MS/MS spectra was processed using the SEQUEST database search algorithm (Thermo Fisher Scientific). The resulting protein identifications were brought into the Scaffold visualization software (Proteome Software) where the information was further refined resulting in improved protein id conformation. Scaffold search criteria were set at 98% probability and required at least 2 unique peptides per ID.

3.7.5 Determination of enriched regions in SWI/SNF ChIP-Seq data

All ChIP-Seq data sets (Ini1, Brg1, BAF155, BAF170, and Pol II) were scored against a normal IgG control using PeakSeq [26] with default parameters (q-value < 0.05) to determine an initial set of enriched regions. These lists were then filtered by removing those regions that did not meet all of the following requirements: 1) the q-value from PeakSeq was further restricted to a q-value of < 0.01; 2) a minimum of 20 sequencing

reads per peak from the specific antibody ChIP; 3) an enrichment of 1.5-fold of the specific antibody relative to the normal IgG control; and 4) an excess of at least 10 of the specific antibody reads relative to the normal IgG control reads. Enriched regions satisfying these criteria comprised our initial list of enrichment sites for each factor (Tables 1 and S11-S16). Among these data sources, Pol II and the normal IgG control have been published as part of prior studies and are available in GEO (accession numbers GSE14022 and GSE12781, respectively) [26,83]. Data for Ini1, Brg1, BAF155 and BAF170 can be accessed through GEO series GSE24397.

3.7.6 Generation of a SWI/SNF union list from ChIP-Seq results

After obtaining our initial list of enriched regions for each factor subjected to chromatin immunoprecipitation, we generated a union list of SWI/SNF component targets. Using the method described in Euskirchen et al. [85], we formed the union of Ini1, BAF155, BAF170, and Brg1 enriched regions as identified by ChIP-Seq and merged any unioned regions that were separated by ≤ 100 bp. Each union region was then classified by whether it intersected with one or more of BAF155, BAF170, Ini1, and Brg1. The resulting list consists of 69,658 SWI/SNF union regions (Table S2).

3.7.7 Determination of the 'high-confidence' and 'core' SWI/SNF regions from the ChIP-Seq union regions

We compared our ChIP-Seq target lists for the 69,658 SWI/SNF union regions against genomic features at which chromatin remodeling is expected to play a prominent role: RNA polymerase II sites [26], 5' ends of Ensembl protein-coding genes, CTCF sites [28], and regions predicted to be enhancers in HeLa cells [29]. We also compared individual SWI/SNF component lists against each other. Only those SWI/SNF regions

which intersect another SWI/SNF component or which intersect at least one of the above genomic features were retained for the 'high-confidence' union list. For gene promoter regions, we define overlap as a target region with at least 1 shared bp within ± 2.5 kb of the annotated transcription start site (TSS). SWI/SNF region intersections were calculated both for all genes in the Ensembl 52 database build using annotations from NCBI36 (human genome build hg18) as well as for a subset of genes that Ensembl identifies as protein-coding. The resulting target list consists of 49,555 'high-confidence' SWI/SNF union regions (Table S3). Union regions containing all three of the BAF155, BAF170, and Ini1 subunits are designated as the 9,760 'core' SWI/SNF regions (Table 3).

3.7.8 Generating co-occurrence tables

To determine the co-occurrences of features of interest we used a similar intersection strategy as was used for determining the high-confidence SWI/SNF regions. For all pairwise comparisons, one of the two data sets was extended by 100 bp on each side of the region and then intersected against the other, non-extended dataset. We required an overlap of at least 1 bp to deem two regions as associated. Using a Perl script, the intersection results for all comparisons were combined to form the co-occurrence table. The same procedure was followed to generate SWI/SNF-centric (Tables S2 and S3), Pol II-centric (Table S5) and Pol III-centric (Table S6) co-occurrence tables.

3.7.9 Determination of expressed regions

Using the HeLa RNA-Seq data of Morin et al. [37], we subdivided each list by the expression status of the corresponding gene targets. Expressed genes were defined as any Ensembl gene with an associated Ensembl transcript having an adjusted depth of ≥1, representing an average coverage of 1x across all bases in the transcript. A total of 9,711

expressed protein-coding genes satisfied these criteria.

3.7.10 Comparison of expression levels associated with different SWI/SNF sub-complexes

We created a series of lists based upon the combinations of SWI/SNF components that could co-occur using the 49,555 high-confidence SWI/SNF regions derived from Table S3. Using the RNA-Seq data of Morin et al. [37], we intersected each list against the 5 ends of transcripts queried by that study and recorded the corresponding adjusted depth for any transcript with a 5 end within ±2.5 kb of a SWI/SNF region. Morin et al. treats adjusted depth as a measurement of transcription level for the corresponding transcript. For each list, these measurements were used to build a series of violin plots showing the probability distribution of transcription levels associated with different compositions of SWI/SNF subunits. Note that each SWI/SNF region from table S2 can only be assigned to one list (e.g. a region containing BAF155, BAF170, and Ini1 is not also assigned to the list of regions containing BAF155 and BAF170).

3.7.11 Pathway analyses of SWI/SNF factors

Overrepresented GO categories [86] and KEGG pathways [87] were determined using DAVID tools [16]. Figures S2 and S3 were drawn using KGML-ED [88].

3.7.12 ChIP-chip experimental procedures and array scoring

The ENCODE tiling arrays (NimbleGen Systems Inc., Madison, WI) interrogate the regions from the pilot phase of the ENCODE project [89] and tile the non-repetitive forward strand DNA sequence with 50-mer oligonucleotides spaced every 38 bp (overlapping by 12 bp) for a total of approximately 390,000 features. For array

hybridizations ChIP DNA samples from 1 x 108 cells were labeled according to the manufacturer's protocol by Klenow random priming with Cy5 nonamers (lamin A/C or lamin B ChIP DNA) or Cy3 nonamers (normal IgG ChIP DNA). Biological replicates, defined as ChIP DNA isolations prepared from distinct cell cultures, were each hybridized to separate microarrays. Each lamin data set consists of three biological replicates. ChIP DNA labeling and array hybridizations were conducted by the NimbleGen service facility (Reykjavik, Iceland). Briefly, arrays were hybridized in Maui hybridization stations for 16-18 h at 42°C, and then washed in 42°C 0.2% SDS/0.2x SSC, room temperature 0.2x SSC, and 0.05x SSC. Arrays were scanned on an Axon 4000B scanner.

For each pair of arrays the files (in .GFF file format) corresponding to the two channels for ChIP DNA (635 nm) and reference DNA (532 nm), were uploaded to the Tilescope pipeline for normalization and scoring [90]. Data were scored with the following TileScope program parameters: quantile normalization of replicates, iterative peak identification, window size= 500, oligo length=50, pseudomedian threshold=1.0, p-value threshold=4.0, peak interval=1000, and feature length=1000. Regions called by Tilescope were then filtered and corrected for multiple hypothesis testing by false discovery rate (FDR). To generate our set of background regions for FDR analysis, we randomly shuffle the probe values within each replicate, ensuring that the same probes are swapped for each replicate. This shuffled data set is then used as input to Tilescope and the scores compared against the lamin A/C and the lamin B data sets. The final lists of enriched regions for lamin A/C and lamin B have a final FDR of 0.1. Target coordinates were converted to hg18 using the UCSC 'liftOver' utility

(http://genome.ucsc.edu/cgi-bin/hgLiftOver). Lamin A/C and lamin B data are available through GEO series GSE24382 and Tables S17 and S18.

3.7.13 Comparison of features across the ENCODE regions

To facilitate comparisons between sequencing and array data we retained only those regions that could be queried by both platforms. To this end, we first identified sequences represented on the ENCODE tiling array that possess less than 25% mappability in ChIP-Seq experiments using 30 bp reads. Any enriched regions in the lamin A/C and the lamin B data sets that were entirely contained within these regions of low mappability were removed from our lists, as corresponding signal levels are unlikely to be detected accurately via ChIP-Seq. Mappability was determined using a 30 bp read length and reported in 100 bp windows according to [26]. The end result is a list of lamin A/C and lamin B enriched regions identified by ChIP-chip in areas of the genome that can be queried by ChIP-Seq. Accordingly, regions that are not represented on the ENCODE tiling arrays were also removed from our SWI/SNF ChIP-Seq experiments for this comparison. Because our ChIP-Seq data covers the entire genome, we began by restricting our enriched SWI/SNF regions only to those that occur in the ENCODE pilot regions. We further refined our ChIP-Seq data set by discarding any SWI/SNF regions that occur in a region of the tiling array for which a signal level of 0 was observed via ChIP-chip. Once our SWI/SNF, lamin A/C, and lamin B lists were limited to those regions that could be queried by both platforms, we intersected the remaining lamin regions and the SWI/SNF regions using the same method that generated the all features table for enhancers, Pol II, and other elements, as described above. Similar procedures were followed for intersections with DNA replication origins identified in the ENCODE

regions using tiling arrays [55].

3.7.14 Evaluating enrichment of SWI/SNF components with respect to other genomic features

To determine whether SWI/SNF components, core regions, and union regions are enriched for factors such as enhancers, small RNAs, lamin A/C and B, CTCF sites, Pol II regions, Pol III sites, 5 ends and DNA replication origins, we used the genome structure correction test (GSC). This test determines the significance of observations where there "exists a complex dependency structure between observations" and was specifically designed for large-scale genomic studies [27]. Given two lists of genomic regions to compare and a list of coordinates defining the overall sample space (i.e. the length of each chromosome), a p-value for the significance of the overlap of the two lists is calculated and we report this value where noted.

3.7.15 Data deposition

All data produced for this study can be accessed through GEO and accession numbers for individual series are provided in the relevant sections. Alternatively, data from the lamin ChIP-chip experiments and the Ini1, Brg1, BAF155, and BAF170 ChIP-Seq experiments can be accessed through GEO using the SuperSeries accession number GSE24398.

3.8 References

1. Jackson, D. A., Iborra, F. J., Manders, E. M. & Cook, P. R. Numbers and organization of RNA polymerases, nascent transcripts, and transcription units in HeLa nuclei. *Mol. Biol. Cell* **9**, 1523–1536 (1998).
2. Cook, P. R. The organization of replication and transcription. *Science* **284**, 1790–1795 (1999).
3. Cook, P. R. A Model for all Genomes: The Role of Transcription Factories.

Journal of Molecular Biology **395**, 1–10 (2010).
4. Clapier, C. R. & Cairns, B. R. The Biology of Chromatin Remodeling Complexes. *Annu. Rev. Biochem.* **78**, 273–304 (2009).
5. la Serna, de, I. L., Ohkawa, Y. & Imbalzano, A. N. Chromatin remodelling in mammalian differentiation: lessons from ATP-dependent remodellers. *Nat Rev Genet* **7**, 461–473 (2006).
6. Wu, J. I., Lessard, J. & Crabtree, G. R. Understanding the words of chromatin regulation. *Cell* **136**, 200–206 (2009).
7. Phelan, M. L., Sif, S., Narlikar, G. J. & Kingston, R. E. Reconstitution of a core chromatin remodeling complex from SWI/SNF subunits. *Mol. Cell* **3**, 247–253 (1999).
8. Chen, J. & Archer, T. K. Regulating SWI/SNF subunit levels via protein-protein interactions and proteasomal degradation: BAF155 and BAF170 limit expression of BAF57. *Mol. Cell. Biol.* **25**, 9016–9027 (2005).
9. Sohn, D. H. *et al.* SRG3 interacts directly with the major components of the SWI/SNF chromatin remodeling complex and protects them from proteasomal degradation. *J. Biol. Chem.* **282**, 10614–10624 (2007).
10. Percipalle, P. & Visa, N. Molecular functions of nuclear actin in transcription. *J. Cell Biol.* **172**, 967–971 (2006).
11. Castano, E. *et al.* Actin complexes in the cell nucleus: new stones in an old field. *Histochem. Cell Biol.* **133**, 607–626 (2010).
12. Rando, O. J., Zhao, K., Janmey, P. & Crabtree, G. R. Phosphatidylinositol-dependent actin filament binding by the SWI/SNF-like BAF chromatin remodeling complex. *Proc Natl Acad Sci USA* **99**, 2824–2829 (2002).
13. Versteege, I. *et al.* Truncating mutations of hSNF5/INI1 in aggressive paediatric cancer. *Nature* **394**, 203–206 (1998).
14. Sévenet, N. *et al.* Spectrum of hSNF5/INI1 somatic mutations in human cancer and genotype-phenotype correlations. *Human Molecular Genetics* **8**, 2359–2368 (1999).
15. Wong, J. *et al.* A protein interaction map of the mitotic spindle. *Mol. Biol. Cell* **18**, 3800–3809 (2007).
16. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**, 44–57 (2009).
17. Wiegand, K. C. *et al.* ARID1A mutations in endometriosis-associated ovarian carcinomas. *N. Engl. J. Med.* **363**, 1532–1543 (2010).
18. Jones, S. *et al.* Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science* **330**, 228–231 (2010).
19. Van Maele, B., Busschots, K., Vandekerckhove, L., Christ, F. & Debyser, Z. Cellular co-factors of HIV-1 integration. *Trends Biochem Sci* **31**, 98–105 (2006).
20. Turelli, P. *et al.* Cytoplasmic recruitment of INI1 and PML on incoming HIV preintegration complexes: interference with early steps of viral replication. *Mol. Cell* **7**, 1245–1254 (2001).
21. Das, S., Cano, J. & Kalpana, G. V. Multimerization and DNA binding properties of INI1/hSNF5 and its functional significance. *J. Biol. Chem.* **284**, 19903–19914 (2009).

22. Isakoff, M. S. *et al.* Inactivation of the Snf5 tumor suppressor stimulates cell cycle progression and cooperates with p53 loss in oncogenic transformation. *Proc Natl Acad Sci USA* **102**, 17745–17750 (2005).
23. Lee, Y. S. *et al.* Chromatin remodeling complex interacts with ADD1/SREBP1c to mediate insulin-dependent regulation of gene expression. *Mol. Cell. Biol.* **27**, 438–452 (2007).
24. Xi, Q., He, W., Zhang, X. H.-F., Le, H.-V. & Massagué, J. Genome-wide impact of the BRG1 SWI/SNF chromatin remodeler on the transforming growth factor beta transcriptional program. *J. Biol. Chem.* **283**, 1146–1155 (2008).
25. Simone, C. SWI/SNF: the crossroads where extracellular signaling pathways meet chromatin. *J. Cell. Physiol.* **207**, 309–314 (2006).
26. Rozowsky, J. S. *et al.* PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* **27**, 66–75 (2009).
27. Bickel, P., Boley, N., Brosn, J., Huang, H. & Zhang, N. Subsampling methods for genomic inference. *Annals of Applied Statistics* **4**, 1660–1697 (2010).
28. Cuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* **19**, 24–32 (2009).
29. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
30. Ni, Z., Abou El Hassan, M., Xu, Z., Yu, T. & Bremner, R. The chromatin-remodeling enzyme BRG1 coordinates CIITA induction through many interdependent distal enhancers. *Nat. Immunol.* **9**, 785–793 (2008).
31. Bazett-Jones, D. P., Côté, J., Landel, C. C., Peterson, C. L. & Workman, J. L. The SWI/SNF complex creates loop domains in DNA and polynucleosome arrays and can disrupt DNA-histone contacts within these domains. *Mol. Cell. Biol.* **19**, 1470–1478 (1999).
32. Noma, K.-I., Cam, H. P., Maraia, R. J. & Grewal, S. I. S. A role for TFIIIC transcription factor complex in genome organization. *Cell* **125**, 859–872 (2006).
33. Raab, J. R. & Kamakaka, R. T. Insulators and promoters: closer than we think. *Nat Rev Genet* **11**, 439–446 (2010).
34. Oler, A. J. *et al.* Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. *Nat Struct Mol Biol* **17**, 620–628 (2010).
35. Barski, A. *et al.* Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. *Nat Struct Mol Biol* **17**, 629–634 (2010).
36. Urnov, F. D. & Wolffe, A. P. Chromatin remodeling and transcriptional activation: the cast (in order of appearance). *Oncogene* **20**, 2991–3006 (2001).
37. Morin, R. *et al.* Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques* **45**, 81–94 (2008).
38. Saunders, A., Core, L. J. & Lis, J. T. Breaking barriers to transcription elongation. *Nat Rev Mol Cell Biol* **7**, 557–567 (2006).
39. Fuda, N. J., Ardehali, M. B. & Lis, J. T. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* **461**, 186–192 (2009).
40. Affymetrix ENCODE Transcriptome Project & Cold Spring Harbor Laboratory

ENCODE Transcriptome Project Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* **457**, 1028–1032 (2009).
41. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
42. Malovannaya, A. *et al.* Streamlined analysis schema for high-throughput identification of endogenous protein complexes. *Proc Natl Acad Sci USA* **107**, 2431–2436 (2010).
43. Xue, Y. *et al.* The human SWI/SNF-B chromatin-remodeling complex is related to yeast rsc and localizes at kinetochores of mitotic chromosomes. *Proc Natl Acad Sci USA* **97**, 13015–13020 (2000).
44. Bourgo, R. J. *et al.* SWI/SNF deficiency results in aberrant chromatin organization, mitotic failure, and diminished proliferative capacity. *Mol. Biol. Cell* **20**, 3192–3199 (2009).
45. UniProt Consortium The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* **38**, D142–8 (2010).
46. Misteli, T. Beyond the Sequence: Cellular Organization of Genome Function. *Cell* **128**, 787–800 (2007).
47. Dechat, T. *et al.* Nuclear lamins: major factors in the structural organization and function of the nucleus and chromatin. *Genes Dev* **22**, 832–853 (2008).
48. Crisp, M. *et al.* Coupling of the nucleus and cytoplasm: role of the LINC complex. *J. Cell Biol.* **172**, 41–53 (2006).
49. Haque, F. *et al.* SUN1 interacts with nuclear lamin A and cytoplasmic nesprins to provide a physical connection between the nuclear lamina and the cytoskeleton. *Mol. Cell. Biol.* **26**, 3738–3751 (2006).
50. Shimi, T. *et al.* The A- and B-type nuclear lamin networks: microdomains involved in chromatin organization and transcription. *Genes Dev* **22**, 3409–3421 (2008).
51. Reyes, J. C., Muchardt, C. & Yaniv, M. Components of the human SWI/SNF complex are enriched in active chromatin and are associated with the nuclear matrix. *J. Cell Biol.* **137**, 263–274 (1997).
52. Moir, R. D., Montag-Lowy, M. & Goldman, R. D. Dynamic properties of nuclear lamins: lamin B is associated with sites of DNA replication. *J. Cell Biol.* **125**, 1201–1212 (1994).
53. Cohen, S. M. *et al.* BRG1 co-localizes with DNA replication factors and is required for efficient replication fork progression. *Nucleic Acids Res* **38**, 6906–6919 (2010).
54. Seo, S. *et al.* Geminin regulates neuronal differentiation by antagonizing Brg1 activity. *Genes Dev* **19**, 1723–1734 (2005).
55. Cadoret, J.-C. *et al.* Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc Natl Acad Sci USA* **105**, 15837–15842 (2008).
56. Ryme, J., Asp, P., Böhm, S., Cavellán, E. & Farrants, A.-K. O. Variations in the composition of mammalian SWI/SNF chromatin remodelling complexes. *J. Cell. Biochem.* **108**, 565–576 (2009).
57. Dinant, C. *et al.* Assembly of multiprotein complexes that control genome function. *J. Cell Biol.* **185**, 21–26 (2009).
58. Rino, J. *et al.* A stochastic view of spliceosome assembly and recycling in the

nucleus. *PLoS Comput Biol* **3**, 2019–2031 (2007).
59. Luijsterburg, M. S. *et al.* Stochastic and reversible assembly of a multiprotein DNA repair complex ensures accurate target site recognition and efficient repair. *J. Cell Biol.* **189**, 445–463 (2010).
60. Liu, H., Kang, H., Liu, R., Chen, X. & Zhao, K. Maximal induction of a subset of interferon target genes requires the chromatin-remodeling activity of the BAF complex. *Mol. Cell. Biol.* **22**, 6471–6479 (2002).
61. Bouwmeester, T. *et al.* A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. *Nat Cell Biol* **6**, 97–105 (2004).
62. Hsiao, P.-W., Fryer, C. J., Trotter, K. W., Wang, W. & Archer, T. K. BAF60a mediates critical interactions between nuclear receptors and the BRG1 chromatin-remodeling complex for transactivation. *Mol. Cell. Biol.* **23**, 6210–6220 (2003).
63. Belandia, B., Orford, R. L., Hurst, H. C. & Parker, M. G. Targeting of SWI/SNF chromatin remodelling complexes to estrogen-responsive genes. *EMBO J* **21**, 4094–4103 (2002).
64. Weissman, B. & Knudsen, K. E. Hijacking the chromatin remodeling machinery: impact of SWI/SNF perturbations in cancer. *Cancer Res* **69**, 8223–8230 (2009).
65. Barlow, C. A., Laishram, R. S. & Anderson, R. A. Nuclear phosphoinositides: a signaling enigma wrapped in a compartmental conundrum. *Trends Cell Biol* **20**, 25–35 (2010).
66. Ray, A. *et al.* Human SNF5/INI1, a component of the human SWI/SNF chromatin remodeling complex, promotes nucleotide excision repair by influencing ATM recruitment and downstream H2AX phosphorylation. *Mol. Cell. Biol.* **29**, 6206–6219 (2009).
67. Ho, L. *et al.* An embryonic stem cell chromatin remodeling complex, esBAF, is essential for embryonic stem cell self-renewal and pluripotency. *Proc Natl Acad Sci USA* **106**, 5181–5186 (2009).
68. Ewing, R. M. *et al.* Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol* **3**, 89 (2007).
69. Lambert, J.-P., Mitchell, L., Rudner, A., Baetz, K. & Figeys, D. A novel proteomics approach for the discovery of chromatin-associated protein networks. *Mol Cell Proteomics* **8**, 870–882 (2009).
70. Gavin, A.-C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
71. Gavin, A.-C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
72. Breitkreutz, A. *et al.* A global protein kinase and phosphatase interaction network in yeast. *Science* **328**, 1043–1046 (2010).
73. Lallemand-Breitenbach, V. & de Thé, H. PML nuclear bodies. *Cold Spring Harb Perspect Biol* **2**, a000661 (2010).
74. Muchardt, C., Reyes, J. C., Bourachot, B., Leguoy, E. & Yaniv, M. The hbrm and BRG-1 proteins, components of the human SNF/SWI complex, are phosphorylated and excluded from the condensed chromosomes during mitosis. *EMBO J* **15**, 3394–3402 (1996).
75. Güttinger, S., Laurell, E. & Kutay, U. Orchestrating nuclear envelope disassembly and reassembly during mitosis. *Nat Rev Mol Cell Biol* **10**, 178–191 (2009).

76. Nielsen, A. L. *et al.* Selective interaction between the chromatin-remodeling factor BRG1 and the heterochromatin-associated protein HP1alpha. *EMBO J* **21**, 5797–5806 (2002).
77. Lavigne, M. *et al.* Interaction of HP1 and Brg1/Brm with the globular domain of histone H3 is required for HP1-mediated repression. *PLoS Genet* **5**, e1000769 (2009).
78. Wilson, B. G. *et al.* Epigenetic antagonism between polycomb and SWI/SNF complexes during oncogenic transformation. *Cancer Cell* **18**, 316–328 (2010).
79. Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B. & Cavalli, G. Genome regulation by polycomb and trithorax proteins. *Cell* **128**, 735–745 (2007).
80. Kwon, S. H. & Workman, J. L. The heterochromatin protein 1 (HP1) family: put away a bias toward HP1. *Mol. Cells* **26**, 217–227 (2008).
81. Ye, Q., Callebaut, I., Pezhman, A., Courvalin, J. C. & Worman, H. J. Domain-specific interactions of human HP1-type chromodomain proteins and inner nuclear membrane protein LBR. *J. Biol. Chem.* **272**, 14983–14989 (1997).
82. Métivier, R. *et al.* Cyclical DNA methylation of a transcriptionally active promoter. *Nature* **452**, 45–50 (2008).
83. Auerbach, R. K. *et al.* Mapping accessible chromatin regions using Sono-Seq. *Proc Natl Acad Sci USA* **106**, 14926–14931 (2009).
84. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
85. Euskirchen, G. M. *et al.* Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. *Genome Res* **17**, 898–909 (2007).
86. Gene Ontology Consortium The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res* **38**, D331–5 (2010).
87. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* **38**, D355–60 (2010).
88. Klukas, C. & Schreiber, F. Dynamic exploration and editing of KEGG pathway diagrams. *Bioinformatics* **23**, 344–350 (2007).
89. ENCODE Project Consortium The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).
90. Zhang, Z. D. *et al.* Tilescope: online analysis pipeline for high-density tiling microarray data. *Genome Biol* **8**, R81 (2007).
91. Euskirchen, G. M. *et al.* Diverse roles and interactions of the SWI/SNF chromatin remodeling complex revealed using global approaches. *PLoS Genet* **7**, e1002008 (2011).

Chapter 4: Exploring the Topological Basis for Transcription Regulation and the Formation of Protein Complexes By DNA Folding Using Chromatin Interaction Analysis

4.1 Statement of Prior Publication and Bioinformatics Contribution

This work was published in *Cell* in 2012 by Li, Ruan, Auerbach, and Sandhu, et al [33] and is reprinted with permission. This paper focuses on two analytical goals: the formulation of models to describe how transcription works in three-dimensional cellular space and how the subunits of several known protein complexes are recruited to RNAPII sites. Using the nascent method ChIA-PET against RNAPII, we were able to identify RNAPII binding sites while retaining spatial information to match up regions of DNA that are in close proximities in the cell. We find that for several factors signal is actually higher in sites distal to the promoter, indicating that some subunits are likely being brought to promoter sites via DNA folding to complete known protein complexes. This finding would be missed using a standard ChIP-Seq assay and challenges negative conclusions drawn solely from the use of simple proximity tests. We also propose three models for transcription based upon the number of interacting regions and show that the larger the number of interactions, the larger the transcriptional abundance we observe. In addition to contributing to the refinement of the ChIA-PET interaction calling algorithm and the formulation of the transcription models, I conceived and implemented all of the analysis related to the binding patterns of different protein complex subunits using ChIP-Seq and other data, identified and integrated relevant ChIP-Seq data sets from the public

domain, and processed/analyzed the RNA-seq data used throughout the paper. Beyond performing the subunit analyses and directing which data sets and methods would be most appropriate for our project goals, I also led most analyses on the K562 and HeLa cell lines as well as analyses using ENCODE. My colleagues at the Genome Institute of Singapore focused their analysis efforts on the MCF7 cell line while analysis of the HCT116 and NB4 cell lines was a joint effort.

4.2 Summary

Higher-order chromosomal organization for transcription regulation is poorly understood in eukaryotes. Using genome-wide Chromatin Interaction Analysis with Paired-End-Tag sequencing (ChIA-PET), we mapped long-range chromatin interactions associated with RNA polymerase II in human cells and uncovered widespread promoter-centered intragenic, extragenic, and intergenic interactions. These interactions further aggregated into higher-order clusters, wherein proximal and distal genes were engaged through promoter-promoter interactions. Most genes with promoter-promoter interactions were active and transcribed cooperatively, and some interacting promoters could influence each other thereby implying combinatorial complexity of transcriptional controls. Comparative analyses of different cell lines showed that cell-specific chromatin interactions could provide structural frameworks for cell-specific transcription, and suggested significant enrichment of enhancer-promoter interactions for cell-specific functions. Furthermore, genetically-identified disease-associated noncoding elements were found to be spatially engaged with corresponding genes through long-range interactions. Overall, our study provides insights into transcription regulation by three-dimensional chromatin interactions for both housekeeping and cell-specific genes in

human cells.

4.3 Introduction

A fundamental question in biology is how genes and regulatory regions are organized and coordinated for transcription regulation. While operons, in which one promoter transcribes multiple genes in a single unit, are common in bacteria [1], and bicistronic transcript structures have been described in worms and flies [2][3], eukaryotic genes are thought to be individually transcribed from their own promoters. However, evidence from in situ fluorescence studies in the last decade suggests that transcription is not evenly distributed and is instead concentrated within large discrete foci in mammalian nuclei, raising the possibility that genes are organized into "transcription factories" [4] containing RNA polymerase II (RNAPII) and other components for transcription. However, this theory lacks evidence with molecular and structural details. Thus, the question of how the regulation of genes is coordinated for transcription in mammalian cells remains largely open.

Mammalian genomes are known to be organized intensively into higher-order conformation inside the micron-sized nuclear space. Consequently, three-dimensional (3D) organization must have a role in the mechanisms for transcription regulation and coordination [5]. Chromosome Conformation Capture (3C) and similar techniques [6] along with traditional in situ techniques have demonstrated that chromatin interactions can regulate transcriptional and epigenetic states [7]. However, such analyses are either limited to certain specific domains or of low resolution and lack functional details. Therefore, a global and high-resolution map of functional chromatin interactions is likely to uncover underlying principles of the higher-order genomic architectures regulating

transcription.

Recently, we developed Chromatin Interaction Analysis by Paired-End-Tag sequencing (ChIA-PET) for genome-wide investigation of chromatin interactions bound by specific protein factors [8]. By immunoprecipitation of a factor of interest along with associated DNA fragments and followed by diluted proximity ligation of distant DNA fragments tethered together within individual chromatin complexes, we elucidated the association of regulatory information through nonlinear arrangements. We demonstrated that long-range chromatin interactions occur between the transcription factor Estrogen Receptor α (ERα) bound regions and their target promoters. To globally investigate how all active promoters dynamically interact with their corresponding regulatory regions in vivo, we used ChIA-PET to analyze genome-wide chromatin interactions associated with RNAPII. Our results provide insights into the 3D interplay of active promoters as well as regulatory regions and suggest an architectural model in which related genes in mega-base range are organized for efficient and potentially cooperative transcription.

4.4 Results

4.4.1 Organizational Complexity of RNAPII-Associated Chromatin Interactions

We analyzed five different human cell lines (MCF7, K562, HeLa, HCT116, and NB4) using ChIA-PET with a RNAPII antibody (8WG16) that recognizes the initiation form of the protein. The cell lines originated from a wide range of lineages, and provided a broad representation of human cells. In our pilot analysis, about 20 million uniquely mapped paired-end reads were generated for each of the ChIA-PET experiments (Table S1A available online), which resulted in two genome-wide datasets: the ChIP-enriched RNAPII binding sites and the RNAPII-bound long-range chromatin interactions. Both

intrachromosomal and interchromosomal interaction data were obtained, and the vast majority of chromatin interactions identified by ChIA-PET were intrachromosomal (Table S1B). Twenty-five intrachromosomal and seven interchromosomal interactions were validated either by 3C, DNA-FISH, or both (Figure C.1 and inset of Figure 4.1C).

To present an inclusive view of the RNAPII-associated human chromatin interactome, we combined the ChIA-PET sequence reads from the six pilot experiments into one dataset for analysis (Table S1). Using embedded nucleotide barcode controls and statistical analyses, we assessed the data quality, filtered out the technical noise, and identified high-confidence binding sites and interacting PET clusters (Experimental Procedures). From the combined pilot dataset, we identified 14,604 high-confidence (FDR < 0.05) RNAPII binding sites as well as 19,856 high-confidence intrachromosomal interaction PET clusters (Table S3). The majority (83%) of RNAPII binding sites in the combined dataset were proximal to 5′ Transcription Start Sites (TSS) of genes (Figure 4.1A). There were also distinct but relatively weaker enrichments of peaks at the 3′ Transcription End Sites (TES) of genes. Similar patterns were seen in all the individual experiments. Of the total RNAPII binding sites, 9,487 (65%) were involved in chromatin interactions and these sites showed higher RNAPII occupancy than those not involved in interactions (Figure 4.1B), indicating that most highly-enriched RNAPII binding sites are involved in looped chromatin conformations.

Three basic types of interactions were identified around gene promoters in the combined pilot dataset: intragenic (promoter to gene internal regions, 938, 5%), extragenic (promoter to distal regulatory elements such as enhancer, 6,530, 33%), and intergenic (promoter-promoter of different genes, 8,282, 42%). There was also a

subcategory composed of intermediate enhancer-enhancer interactions (4,106, 20%). Some interactions (2,341, 12%) were standalone duplex interactions between two interacting anchor regions, whereas most (17,515, 88%) were further aggregated into 1,544 interaction complexes.

We speculated that the isolated RNAPII binding at promoter sites, which are not involved in interactions, may reflect the basal promoter function for gene transcription, and thus were termed "basal promoters." By contrast, RNAPII-associated interactions might constitute a structural basis for complex regulatory mechanisms. These basic interactions further aggregated into complex architectures that we classified as "single-gene" or "multigene" complexes depending on the number of genes involved (Figure 4.1C). The single-gene models consisted of single or multiple enhancer interactions with only one gene promoter, whereas the multigene models included intergenic promoter-promoter interactions and could also include intragenic and extragenic enhancer-promoter interactions. Moreover, several such complexes, distantly separated on a chromosome or on different chromosomes, further converged to form higher-order multigene interaction complexes (Figures C.1B, C.1D, C.1F, and C.1G). Many chromatin complexes had genomic spans of 150 kb–200 kb, and a few complexes spanned several megabases. Although there were only 1,328 multigene complexes in this combined pilot dataset, 11,723 genes were engaged in these complexes for an average of 8.8 genes per interaction complex (Figure 4.1D), indicating that promoter-promoter interactions were widespread and may play a significant role in transcription regulation.

To understand how these looping structures influence transcription, we characterized these RNAPII-associated chromatin models (basal promoters, single-gene

and multigene complexes) for structural features (genomic property), functional output (transcription activity), and epigenomic marks (chromatin state).

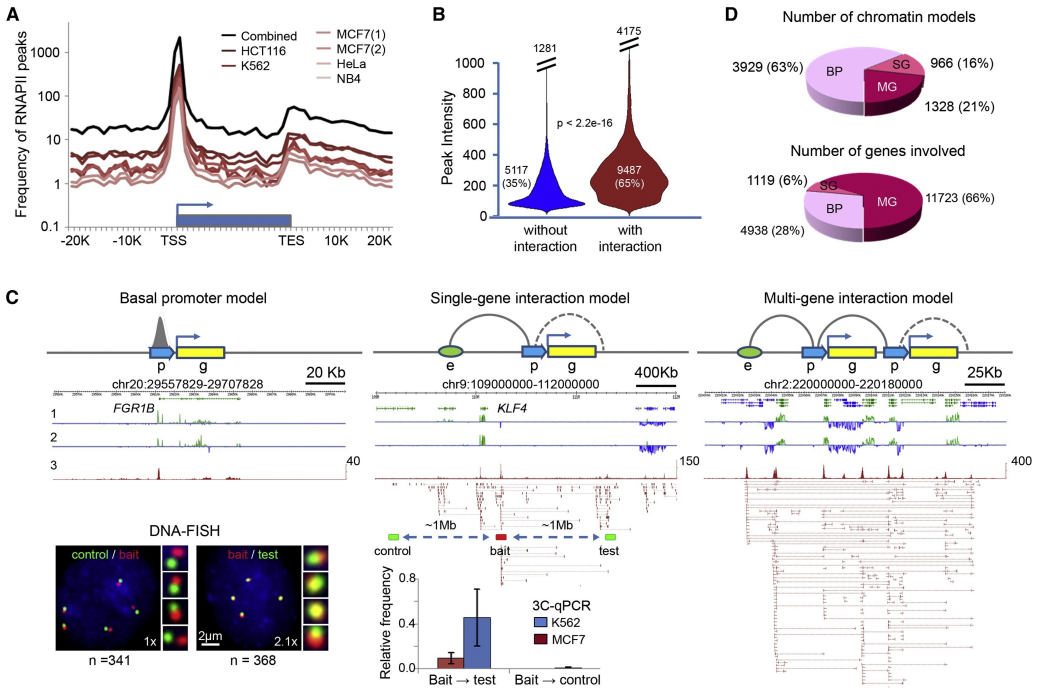


Figure 4.1 Characterization of RNAPII Binding Peaks and Chromatin Interactions. (A) RNAPII binding profile around gene body. **(B)** Violin plots for intensities of RNAPII peaks involved (red, mean intensity = 281) and not involved in interactions (blue, mean intensity = 141). **(C)** RNAPII-associated chromatin models: basal promoter (BP) with RNAPII binding but no chromatin interaction, single-gene (SG) complex with intra- and/or extragenic interactions and multigene (MG) complex with multiple genes in the interaction clusters. p, promoter; g, gene; and e, enhancer. The dotted curve for possible intragenic loop, and the solid curve for potential loop of enhancer-promoter and promoter-promoter interactions. Data tracks are: 1 and 2, strand specific RNA-Seq data of MCF7 and K562; 3, RNAPII binding peaks and ChIA-PET data. Inset (bottom): DNA-FISH and 3C-qPCR validations of the extragenic interaction at the KLF4 locus, where the KLF4promoter and enhancer are ~1 Mb apart. Genomic locations used for 3C bait, test and control sites are indicated. The same locations were also used for DNA-FISH. The numbers (n) of nuclei counted and the fold change (x) in the number of instances showing close proximity (≤1 μm) are indicated. 3C-qPCR mean values and standard error of means (SEM) from three independent experiments are shown. **(D)** Distribution of chromatin models (BP, SG, MG) and the numbers of genes engaged in the models. Also see Figure C.1, Table S1, Table S2, and Table S3.

4.4.2 Distinct Genomic Properties of Single- and Multigene Interaction Models

To determine the genomic characteristics of RNAPII-associated chromatin structures, we mapped several genomic descriptors that were known to associate with the expressivity of the human genome [9], including GC content, gene density, SINE/LINE density, gene length, and the intron/exon ratio. In our analyses (Figure 4.2, Figure C.2A),

the multigene complexes were significantly enriched with higher GC content, higher gene and SINE density, and lower LINE density as compared to the single-gene interaction complexes and the regions of basal promoters, suggesting that multigene complexes were located in open chromatin and highly transcribed regions. In addition, genes in the multigene complex regions were relatively shorter than other gene categories, which is yet another property of highly expressed genes [10]. Conversely, genomic loci associated with the single-gene complexes lay in the regions with lower gene and SINE density. Moreover, the genes engaged in the single-gene complexes were significantly longer and had higher intron/exon ratios than the genes of other chromatin models (Figure 4.2B). These observations suggest that genes with enhancer-promoter interactions in single-gene complexes were more likely to be tissue-specific or developmentally regulated, in line with the previous findings that genes in gene-poor regions associated with several distant regulatory elements, tended to be longer and had a higher noncoding to coding ratio than housekeeping genes [10][11].

4.4.3 Interacting Genes Show Correlated Expression

To investigate the functional output of genes involved in the different chromatin models, as defined by transcriptional activity, we focused our analyses on MCF7 cells, as it is a well-characterized human cancer cell model with complementary datasets including RNA-Seq (Experimental Procedures), time-course microarray gene expression [8], and GRO-Seq datasets [12].

Consistent with the combined pilot dataset, 90% binding sites in MCF7 cells were found proximal to known gene promoters and 97% genes with RNAPII present at their

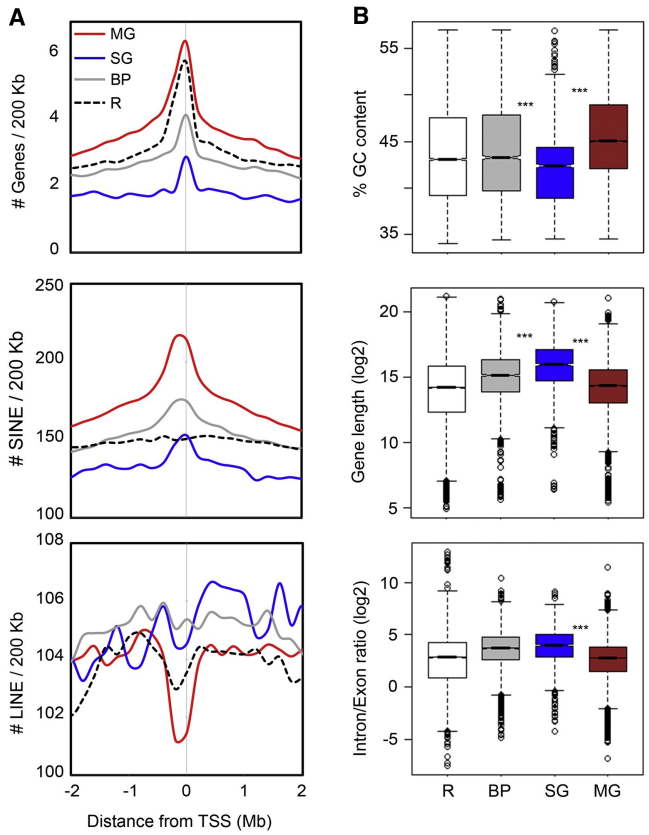


Figure 4.2 Genomic Properties of Promoter-Centered Chromatin Models. (A) Aggregation plots showing enrichment of genes, SINE and LINE elements around the TSS of genes in different chromatin models. Unique RefSeq TSSs were used for analyses. Red curve stands for multigene (MG) model, blue for single-gene (SG) model, gray for basal promoter (BP) model, and black dotted line for the rest of the genes (R).

(B) Box-plots showing distribution of percentage GC content of GC isochore around different models, gene length, and intron/exon ratio of RefSeq genes involved in the models. Triple asterisks (***) signifies p-value < 2.2E-16. Red box stands for MG, blue for SG, and gray for BP. Open box is for R (rest of genic regions) as background. Also see Figure C.2.

promoters had detectable transcriptional activity by RNA-Seq (Figure 4.3A). The interactive RNAPII binding sites that were distal to gene promoters included intra- and extragenic regulatory elements such as enhancers. Approximately 45% of the extragenic distal regulatory sites had detectable RNA signals that could represent possible noncoding RNA (ncRNA) transcripts.

For genes associated with the three chromatin models, we analyzed the transcription levels measured by RNA-Seq reads. As shown in Figure 4.3B, in general, RNAPII binding at promoter sites correlated well with the expression level of the

corresponding genes. Interestingly, the genes involved in the single-gene and the multigene models showed higher correlation between RNAPII binding and RNA-Seq signal (Pearson's correlation coefficient: PCC: 0.46 and 0.45 respectively) as compared to basal promoter genes (PCC: 0.24). Moreover, we observed that genes linked by complex chromatin interactions, especially those in multigene complexes, had significantly higher expression levels than basal promoter genes (Figure 4.3C). This high expression appeared to be limited to genes interacting at the RNAPII anchor sites, as compared to genes located in the intervening chromatin loops. These data indicated that promoter-promoter interactions in multigene complexes were associated with higher transcriptional activity, which is consistent with our observations of their associated genomic features.

Next, we characterized the expression patterns of genes present in the interacting regions using microarray data derived from 84 human tissues [13]. We found distinct representation of tissue-specific and housekeeping genes in the three chromatin models (Figure 4.3D, Figures C.3A and C.3B). Most genes in single-gene complexes with enhancer-promoter connectivity were tissue-specific, consistent with growing evidence that the expression levels of developmental and tissue-specific genes are largely modulated through cis-remote regulatory elements and trans-protein factors [14][15], and consistent with their genomic features (less gene density, longer gene body and higher intron/exon ratio) as previously described. Conversely, genes involved in multigene complexes as well as the basal promoter genes were characterized as both tissue-specific and housekeeping categories. These observations were also supported by normalized CpG content and GC-skew at their promoter regions (Figures C.3C and C.3D).

As promoter-promoter interactions cluster multiple genes, they could provide an

ideal topological framework for potential transcriptional coordination of both tissue-specific and housekeeping genes. This observation agrees with the evidence that "ridges," which are domains of highly transcribed genes, contain both housekeeping and tissue-specific genes [9]. Since large numbers of genes are found in multigene complexes, we propose that promoter-promoter interactions could serve as a dominant mechanism for transcription regulation of both housekeeping and tissue-specific genes in mammalian genomes.

Next, we sought to determine whether genes with promoter-promoter interactions were more likely to be transcriptionally coordinated. RNA-Seq data showed that most of the paired genes with promoter-promoter interactions were expressed together at high levels (Figure 4.3E; Figure C.3E). To further assess the coordinated transcription of paired genes across different conditions, we performed Pearson's correlation analysis using estrogen-induced time course of GRO-Seq data [12] that measured transcription initiation rates of estrogen responsive genes, and observed significant transcriptional correlation (Figure 4.3F; p-value < 2.2E-16). Interestingly, the correlation was even greater for ERα-mediated gene pairs derived from our earlier data [8], suggesting stronger correlation of transcription for genes involved in multigene complexes mediated by specific transcription factors. Similar correlation was also observed from other gene expression datasets (Figures C.3F–C.3I). As expected, housekeeping genes and genes belonging to the same GO classes showed even higher correlation than the rest (Figures C.3J and C.3K). Altogether, our analyses indicated that a significant proportion of gene pairs involved in promoter-promoter interactions tended to be transcribed cooperatively.

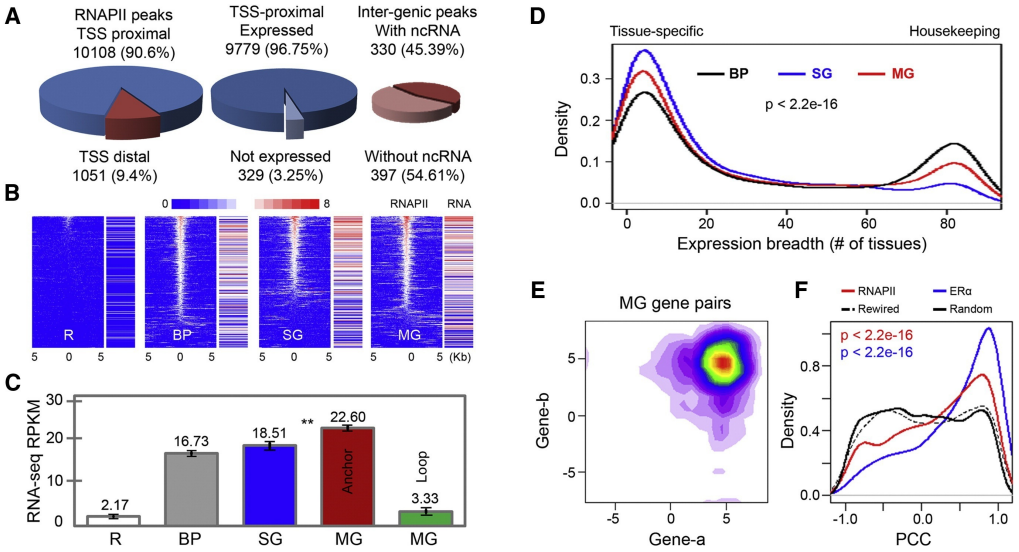


Figure 4.3 Transcriptional Activities in RNAPII-Associated Chromatin Models in MCF7 Cells.(A) Pie charts of RNAPII binding peaks proximal (blue) or distal (red) to TSS of genes (left), RNA-Seq data for genes with RNAPII peaks near TSS (middle), and RNA-Seq enrichment around intergenic RNAPII peaks (right). **(B)** Correlation of RNAPII binding in basal promoter (BP), single-gene (SG) and multigene (MG) models with gene transcription levels measured by RNA-Seq. The RNAPII enrichment heatmap shows binding intensity centered on TSS (±5 kb) along with corresponding gene transcription intensity. **(C)** Bar plots of expression levels of genes in the three models (BP, SG, and MG). RNA-Seq mean values (RPKM) and standard error of means (SEM) from genes in the corresponding models are shown. MG complexes also contain "anchor genes" (TSS proximal to interacting anchors) and "loop genes" (distant from anchors, residing in loop regions). The remaining genes (R) not bound by RNAPII were included as a control. Double asterisks (**) indicate significant differences between the mean expressions of genes from SG and MG models (p-value < 4.02E-08). **(D)** Expression breadth (number of tissues a gene is expressed in) of genes present in three different chromatin models. p-value is calculated using the nonparametric test of Kruskal-Wallis. **(E)** Contour plot of log-transformed RNA-Seq RPKM values for cotranscription of interacting genes involved in MG models in MCF7 cells. **(F)** Distribution of PCC values for RNAPII- and ERα-bound interacting gene pairs, randomly rewired gene pairs, and randomly picked gene pairs from control regions with the same genomic span and gene density distribution as the multigene complex regions. Also see Figure C.3.

4.4.4 Multigene Complexes Provide Structural Framework for Cotranscription

Correlated expression of interacting genes suggests that the multigene interaction complex might provide a molecular basis for the postulated "transcription factory" [4]. To elucidate the link between the multigene complexes revealed by ChIA-PET and transcription factories, we performed 3D DNA-FISH experiments using probes representing distinct multigene complexes in combination with RNAPII-IF staining in MCF7 nuclei (Experimental Procedures). All experiments on four genomic loci randomly chosen from multigene complexes revealed a significant association of the multigene

complex loci with RNAPII foci (Figure 4.4A-B), adding further evidence to support our view that multigene complexes could provide a structural framework for cotranscription.

Furthermore, gene families were significantly over-represented (p-value < 0.006) in the multigene complexes (Figure C.3L), such as HIST, ZNF, KRT, HOXC, etc. (Table S4). Taking the HIST1H family as an example, the 58 genes of this family located on chromosome 6 formed three multigene complexes, and these three complexes converged into a higher-order super-complex, suggesting that all HIST1H genes were organized in a single chromatin architecture for coordinated transcription (Figure 4.4C). All HIST1H genes were actively transcribed in both MCF7 and K562 cells, and were highly coregulated across different tissues and cellular conditions (Figure 4.4D). Interestingly, HFE, a gene was not a part of the HIST1H family but was located in the middle of the first HIST1H multigene complex, was not anchored at the interaction sites and was not expressed. Similarly, the genes located in the intervening loop regions between the three HIST1H interacting complexes were relatively less active and much less coordinated for coregulation across different tissues and cellular conditions. This case exemplifies the model where multigene complexes organize genes with similar functions across genomic space for coordinated expression.

4.4.5 Multigene Complexes Support Synergistic Transcription Regulation

To further investigate the likelihood that the multigene complex structure might provide a topological framework for transcriptional coregulation of interacting genes involved in such topology, we designed a set of perturbation experiments to test this. After comparing the RNAPII and ERα ChIA-PET data from MCF7 cells, we found that the RNAPII-bound multigene complex at the GREB1 locus partially overlaps with the

ERα-bound chromatin loops, suggesting that this interaction complex, in part, is associated with ERα. Therefore, we performed siRNA experiments to knockdown the

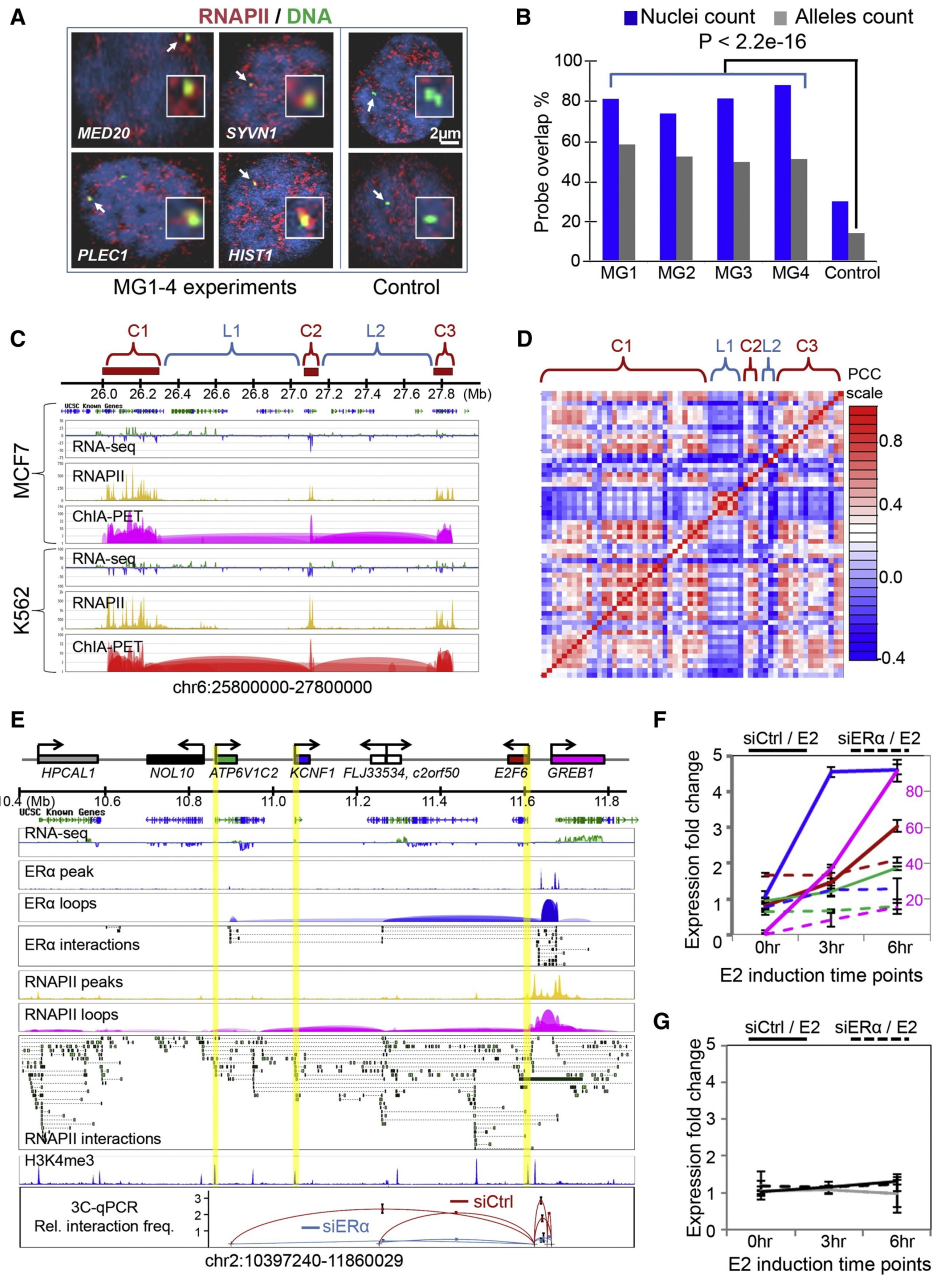


Figure 4.4 Transcriptional Coordination in Multigene Chromatin Complexes.(A) Colocalization of multigene loci with RNAPII foci. Shown are the nuclear images of RNAPII IF-staining with four randomly-selected multigene loci (MG1-4) and 2 control loci. Representative gene loci are MED20, SYVN1, HIST1, and PLEC1. **(B)** Quantitative analysis of nuclei (n = 476) and alleles showing overlap of MG loci and RNAPII foci. Percentage overlaps from MG loci and those from control loci are significantly different. **(C)** Super multigene complex of the histone gene family. Three distant clusters (C1, C2, C3) of HIST1H genes converge together in a super-MG complex. Shown are RNA-Seq, RNAPII and ChIA-PET tracks in MCF7 and K562 cells. **(D)** Cotranscription of HIST1H genes in the super-MG complex in (C). Correlation matrix derived from publicly available microarray data of 4,787 samples (Supplemental Information). The rows and columns correspond to genes in each complex and the intervening regions. **(E)**

RNAPII-bound multigene complex at the GREB1locus. Shown are the ERα- and RNAPII-bound chromatin interactions. Highlighted promoters are anchored by RNAPII, but not by ERα. The bottom panel shows relative interaction frequency by 3C-qPCR data for the perturbation experiments using siERα knockdown and estrogen induction. **(F and G)** Time course RT-qPCR following estrogen (E2) induction after siControl (solid) and siERα (dashed) transfections of MCF7 cells. Colors of the curves correspond to genes shown in (E). A secondary axis (red, right side) is used for GREB1 expression to accommodate its high expression level. Expression data of genes involved in the GREB1 multigene complex are in (F), and the data for genes outside of the complex are in (G). RT-qPCR mean values and standard deviations (SD) from two independent experiments are shown. Also see Figure C.4 and Table S2.

protein level of ERα in MCF7 cells, and monitored the alteration of chromatin interactions and gene transcription in the GREB1 multigene complex. Several chromatin interaction loops at this locus were disrupted by siERα transfection as tested by 3C experiments (Figure 4.4E). In addition to GREB1, which had a strong response to estrogen induction and reduction by siERα knockdown (Figures C.4A–C.4D), we observed that the other genes in this complex such as E2F6, KCNF1 and ATP6VC12 also had various levels of response to induction by estrogen and reduction by siERα knockdown (Figure 4.4F). Interestingly, these genes did not directly interact with ERα at their promoter regions, but indirectly associated with ERα through RNAPII-bound chromatin loops. As a control, this effect was not seen in the nearby genes such as NOL10 and HPCAL1 that were in other RNAPII interaction complexes and also did not interact with ERα (Figure 4.4G). Similar results were observed at another interaction locus centered on the GPR68 and CCDC88C genes (Figure C.4E). Thus, these results indicate that a specific stimulus (estrogen) could lead to coactivation of genes organized primarily through RNAPII-bound multigene complexes, and perturbation at one gene locus (loss of ERα binding in this case) in a multigene complex could alter the transcriptional states of other interacting genes within the same complex. Although genes in close genomic distances with each other had been reported to be correlated in expression levels [16], our data suggests that the conjoint expression can be mediated

through chromatin interactions. The functional significance of such coregulation needs further investigation.

4.4.6 Epigenomic Marks Associated with Chromatin Interaction Sites

To study the association of transcription factors (TFs) with the RNAPII interactions, we examined the enrichment of 20 different TFs in K562 cells at the RNAPII interaction sites from the three chromatin models in our K562 ChIA-PET dataset (Figures 4.5A and 4.5B, Figures C.5A–C.5D). General TFs such as E2F4 and E2F6 (Figure 4.5A, Figure C.5A) directly bound at TSS sites (Figure 4.5B for a specific example). By contrast, specific TFs such as JunD and Max preferentially bound to distal regulatory sites and marked potential enhancers (Figure C.5B). Several chromatin remodeling factors and chromatin organization proteins such as INI1, BRG1, CTCF, and RAD21 associated primarily with non-TSS sites, suggesting that they may mediate long-range interactions with enhancer regions (Figure 4.5A, Figure C.5C). This hypothesis is consistent with other observations that INI1 and BRG1, two subunits of the SWI/SNF complex, were involved in transcriptional looping [17]. A common observation among all the factors was that interaction sites in the multigene complexes consistently showed elevated levels of factor enrichment, suggesting that the cooperative binding of factors in gene-rich domains leads to higher transcriptional activity, or these transcriptionally active open chromatin domains might converge to distinct specialized transcription factories, each enriched with general and specific TFs.

We further explored the histone modification data available from the ENCODE Consortium. Collectively, we found high enrichment of active histone modification marks coupled with a lack of repressive marks in RNAPII interaction sites, confirming

that the RNAPII interaction sites mapped by our ChIA-PET data were located in promoter and distal regulatory regions engaged and/or poised for high transcription levels (Figure C.5D). Interestingly, the enrichment of active marks was highest in the multigene complexes, indicating that these might constitute transcriptional hubs. Our observations matched previous findings that the enrichment of active histone modifications positively correlated with RNAPII occupancy [18].

We observed similar histone modification profiles in MCF7 cells (Figure 4.5C) using data that we generated previously [19]. In particular, we applied the log ratio of H3K4me3/H3K4me1 signal as a quantitative measurement of the likelihood that a genomic locus can act as a promoter or enhancer. Most non-interacting RNAPII sites proximal to TSS in basal promoter model showed high log ratios (Figure 4.5D, plot 1; median = 2.4; > 90% of the binding regions have log ratios > 0), whereas most of the RNAPII interaction sites distal to TSS in the single-gene complex model and the multigene complex model (conventional enhancer sites) showed low H3K4me3/me1 log ratios (Figure 4.5D, plot 4 and 6; median < −0.72), confirming that this log ratio could reflect relative capacities of promoters and enhancers. Surprisingly, examination of RNAPII interaction sites proximal to known TSSs in the multigene complexes (Figure 4.5D plot 5) revealed two peaks in the histogram of the log ratios, suggesting a mixture of enhancer and promoter elements in the promoter regions. Detailed profiles of H3K4me3 and H3K4me1 marks around the center (±5 kb) of those RNAPII interaction sites showed distinct characteristics of promoter-like, enhancer-like sub-groups (Figure 4.5D, heatmap). Moreover, enhancer-like RNAPII interaction sites, on average, showed lower transcriptional activity than the promoter-like RNAPII sites (Figure C.5J).

Thus, a large portion of interacting promoters may also have potential enhancer functions. We observed the same inverse correlation of H3K4me3/me1 log ratio at the TSS proximal and TSS distal RNAPII sites for K562 (Figure 4.5A), indicating that this observation is a general phenomenon applicable to all cell types.

4.4.7 Interacting Promoters Possess Combinatorial Regulatory Functions

To examine potential enhancer activity of promoters, we performed luciferase reporter gene assays, a commonly used method for promoter and enhancer characterization [20]. In these assays, approximately 500 bp fragments of the expected promoter regions were cloned upstream of a luciferase reporter gene construct either in a proximal position as the driving promoter or in a distal position as a presumed enhancer, and the constructs were transfected into MCF7 cells (Experimental Procedures, Figures C.5E–C.5I). As shown in Figure 4.5E, the two interacting loci INTS1 and MAFK were 26 kb apart, and our RNA-Seq data suggested that both genes were active in MCF7 cells. However, the normalized log ratio of H3K4me3/me1 was 0.36 for the INTS1 promoter and 1.13 for the MAFK promoter, suggesting that the INTS1promoter may have enhancer properties. To test this, we cloned the INTS1 promoter fragment in both orientations upstream of the MAFK promoter flanking the luciferase gene. The luciferase reporter gene assay showed at least 7-fold enhancement of luciferase expression from the MAFK promoter activity by the INTS1 promoter fragment, indicating that a bona fide promoter can act as an enhancer to augment the activities of other promoters.

In another example (Figure 4.5F), the promoter of CALM1 interacts with an enhancer element 15 kb upstream and connects to the promoter of C14orf102 further upstream in 65 kb. Both RNA-Seq data and the H3K4me3/me1 log ratio indicated that

the CALM1 promoter was strong, whereas the C14orf102 promoter was weak and enhancer-like. The luciferase reporter gene assay showed marginal enhancement to the CALM1 promoter reporter gene activity by the native CALM1 enhancer and the C14orf102 promoter individually. However, the combined CALM1 enhancer and the C14orf102 promoter together led to a significant ~3-fold enhancement of reporter expression from the CALM1 promoter. This result further validates the enhancer function by interacting promoters and elucidates a possibility of combinatorial effect among interacting elements in multigene interaction complexes for transcription regulation.

Next, we asked whether promoters with enhancer activity act specifically on their target genes. We swapped the promoter elements in the two examples of INTS1-to-MAFK and C14orf102-to-CALM1 for additional reporter genes assays (Figure 4.5G). Intriguingly, when placed upstream to the CALM1 promoter, the INTS1promoter showed remarkable enhancement of CALM1 promoter activity. Similarly, the combined construct of C14orf102 promoter and CALM1 enhancer also increased MAFK promoter activity significantly. Meanwhile, a TATA box deleted promoter and other control promoters (either active or inactive), taken from the nearby genes that are not involved in a promoter-promoter relationship, did not show cooperative enhancement to MAFK and CALM1 promoter activities (Figures C.5H and C.5I). Thus, these results suggest a common property for promoters with enhancer capacity that could influence other promoters.

In addition, we also tested the combination of inserting the enhancer-like promoter fragment in the position proximal to luciferase gene and the strong promoter in the distal position in the reporter gene construct. Of the 20 such luciferase experiments, we

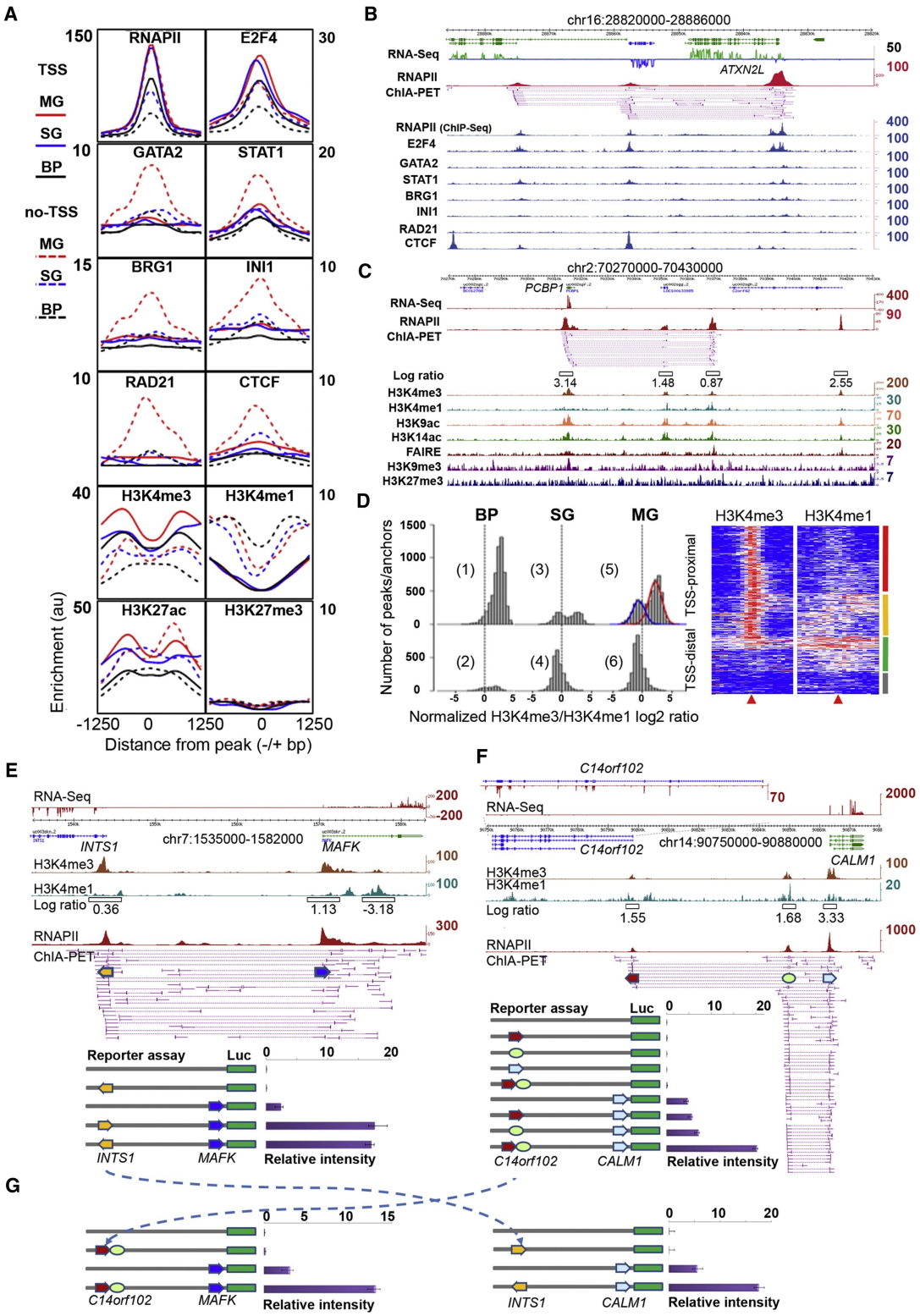


Figure 4.5 Epigenomic Profiles of Chromatin Interactions and Combinatorial Regulation of Interacting Promoters. (A) Enrichment profiles of TFs and histone modifications centered on RNAPII peaks (±1250 bp) of interacting loci of the three models in K562 cells. Solid lines represent "TSS" proximal regions and dotted lines depict "non-TSS" regions. y axis: sliding median for ChIP-Seq enrichment in the region. **(B)** Examples of TF enrichment at RNAPII interacting loci in K562 cells. **(C)**

Histone modification marks and open chromatin mark (FAIRE) associated with chromatin interaction sites in MCF7 cells. The width of the open boxes in the log ratio track reflects the region where the H3K4me3 and H3K4me1 data were used for the log ratio calculation. **(D)** Histograms of normalized H3K4me3/me1 log ratio at RNAPII sites proximal to TSS (TSS) and distal to TSS (non-TSS) of genes in the three chromatin models in MCF7 cells. Two peaks are seen in plot #5 (blue curve for enhancer-like, and the red for promoter-like). The heatmap shows detailed H3K4me3 and H3K4me1 enrichments around RNAPII interaction sites (±5 kb) proximal to TSS. Four distinct clusters of RNAPII sites are promoter-like (red), enhancer-like (green), heterogeneous (yellow) and weak signals (gray). **(E–G)** Reporter gene assay of interacting promoters in MCF7 cells. RNA-Seq, H3K4me3, H3K4me1, H3K4me3/me1 ratio, and RNAPII ChIA-PET data tracks are shown. Numbers on the right side for each track indicate the highest peak intensity. The mean values and standard deviations (SD) of the luciferase activities from at least three independent experiments are shown. (E) Promoter-promoter interaction at the INTS1-MAFK locus. The arrow boxes indicate the aligned promoter regions, which were cloned in reporter gene constructs for luciferase assay. (F) Promoter-enhancer-promoter interactions at the C14orf102-CALM1 locus. RNA-Seq data showed that CALM1 was highly expressed, whereasC14orf102 only marginally transcribed (enlarged RNA-Seq track of the C14orf102 locus). (G) Swap assay of DNA fragments from different multigene complexes. The dotted arrow lines show the swap of elements cloned in the distal positions in the reporter gene constructs for luciferase assay. Also see Figure C.5 and Table S2.

observed that the weaker promoters conveyed significant enhancer function to their stronger interacting partners in luciferase activity rather than the reverse (Figure C.5K). In the case of interacting pair INTS1 (enhancer-like promoter) and MAFK (strong promoter), the strong promoter MAFK did not demonstrate significant enhancer activity (Figure C.5L). Thus, at promoter sites, there is an inverse relationship between enhancer and promoter functions.

4.4.8 Cell-Line Specificity of Long-Range Chromatin Interactions

To elucidate the cell-line specificity of chromatin interactions, we saturated the coverage of chromatin interactions through deep sequencing of more MCF7 and K562 ChIA-PET replicates (Experimental Procedures). The saturated libraries are highly reproducible for interactions, and thus highly reliable for inter-cell line comparative analysis. These libraries exhibit the same pattern of genomic descriptors as the pilot libraries (Figures C.2B and C.2C). With comprehensive ChIA-PET and RNA-Seq datasets, we performed comparative analysis between the two cell lines and identified cell-line specific genes and chromatin interactions (Figure 4.6A). Most of the genes

specifically expressed in their respective cells also showed cell-specific interactions (Figure 4.6B), implying that cell-specific chromatin interactions provide the structural basis for cell-specific transcription. Gene Ontology (GO) analysis revealed significant enrichment of erythroid related GO terms such as response to stimulus and blood circulation for genes with specific expression and chromatin interactions in K562 cells, whereas GO terms such as ectoderm development and related biological process were enriched in MCF7 cells (Figure 4.6C, Figure C.6A). As expected, the genes common in both cell lines showed enrichment of housekeeping functions like metabolism, cell-cycle and signal transduction (Figure C.6B).

Among the chromatin interactions specific to K562 cells, we captured many previously characterized interactions including the α- and β-globin loci ([21][14]. Figure 4.6D shows extensive interactions identified by ChIA-PET data between the α-globin gene locus and the DNase hyper-sensitive (DHS) sites present in the gene body of the C16orf35 gene. Additionally, we found that the α-globin locus in K562 extended its interactions to the neighboring domains, which were constitutively active in both K562 and MCF7 cells, whereas the interactions to α-globin genes are K562-specific, suggesting a complex chromatin architecture for spatiotemporal regulation of both constitutive and cell-specific transcription. Similarly, the β-globin gene locus also displayed previously known K562-specific interactions with the nearby locus control region (Figure C.6C).

GREB1 is a well characterized MCF7-specific gene. As expected, we found abundant chromatin interactions associated with RNAPII at this locus in MCF7, but not in K562 cells (Figure 4.6E). In addition to recapitulating the previously identified ERα-associated interactions [8], RNAPII interaction data showed an additional interaction site

on the far most upstream (left in Figure 4.6E) side of this complex. A strong H3K4me1 mark on this site suggested that this is potentially an enhancer site for a transcription factor other than ERα. Intriguingly, a significant RNA-Seq peak was also identified at this site, indicating a possible enhancer RNA transcript, a new class of noncoding RNA species [22].

4.4.9 Long-Range Enhancer-Promoter Interactions and Disease-Associated Noncoding Elements

Our data showed that the enhancer-promoter interactions were significantly enriched over other types of interactions for cell-specific genes (Figure 4.7A) when compared to genes commonly expressed in both cell lines. This finding supported the general view that distant-acting enhancers tend to be specifically involved in tissue-specific genes, and was consistent with our analysis in Figure 4.3D. Although potential enhancer sites can be identified using high throughput approaches [23], it is still challenging to connect enhancers to their target genes that are hundreds of kilobases away. Moreover, many remote enhancers could be embedded in intronic regions of other distantly located genes [24], making it notoriously difficult to relate enhancers to their specific target genes. In this study, we identified tens of thousands enhancer-promoter interactions (Table S1C) including approximately 1000 ultra-long-distance (500 kb to megabases) events. We observed that ≥ 40% of enhancers do not interact with their nearest promoters and instead jump over to their target promoters, bypassing several intervening genes (Figure 4.7B,Figure C.7).

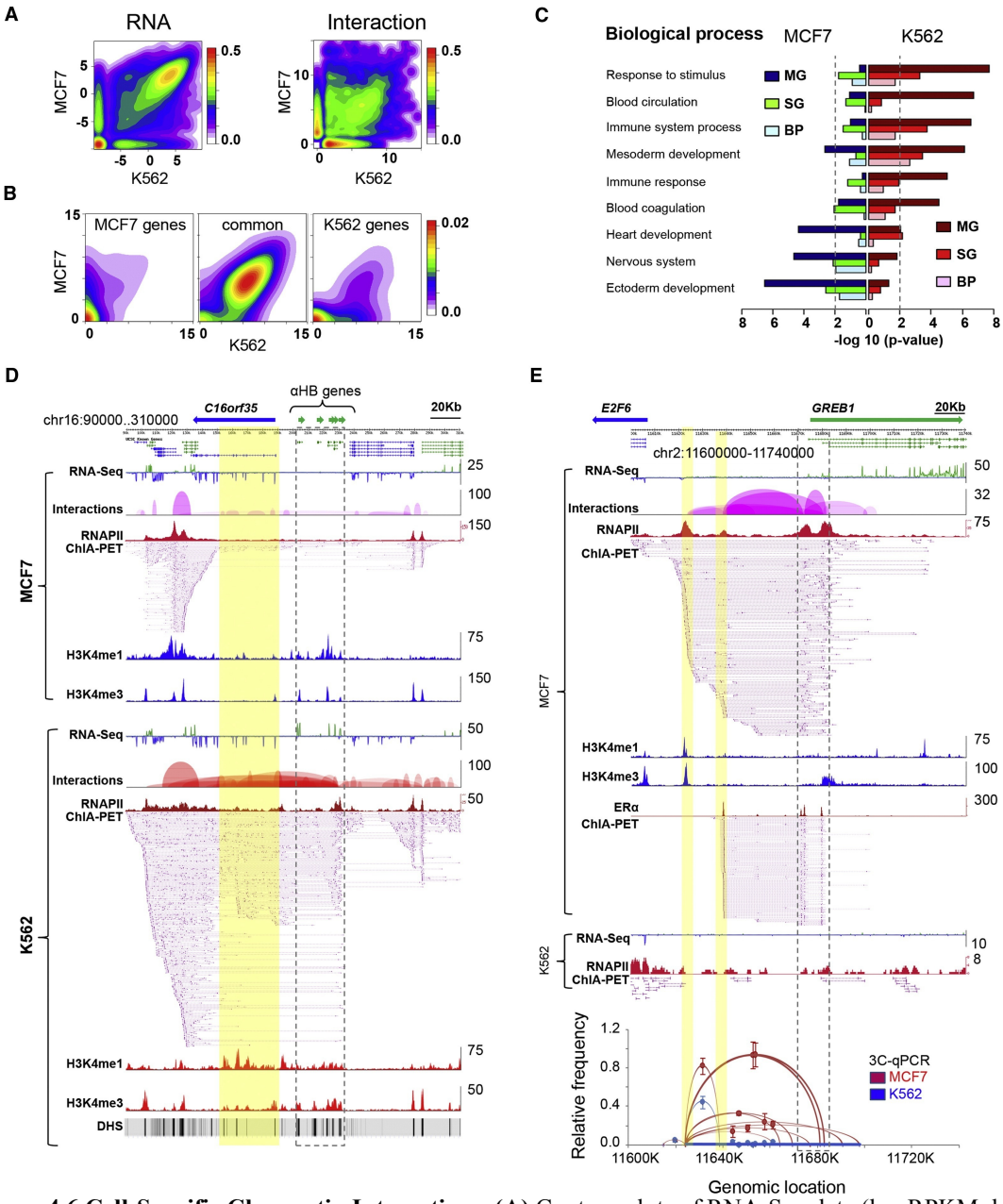


Figure 4.6 Cell-Specific Chromatin Interactions. (A) Contour plots of RNA-Seq data (log RPKM, left) and chromatin interactions (log PET counts, right) in MCF7 and K562 cells, showing common and cell-specific gene expression and chromatin interactions. **(B)** Contour plots of interaction data (log PET counts) for genes specifically and commonly expressed in MCF7 and K562 cells. **(C)** Enrichment of cell-specific GO terms in genes and chromatin interactions specific in MCF7 and K562 cells. The p-value of 0.01 is marked as dotted line. **(D)** An example of K562-specific chromatin interactions. α-globin genes (in dotted line box) interact with distantly located (~20 kb) DHS sites (highlighted in yellow), which are known to interact with α-globin genes. In sharp contrast, the α-globin genes in MCF7 cells are not expressed and have no interactions with the DHS sites. **(E)** An example of MCF7-specific chromatin interactions around the GREB1 locus. The far left highlighted yellow is a RNAPII interaction site that is not overlapped by ERα-bound interactions in this region. It is also the bait site for independent 3C validation of interactions in this region. Tracks included in (D) and (E) are RNA-Seq data, interaction loop view, RNAPII ChIA-PET peaks and interaction PETs, ChIP-Seq density profile of H3K4me1 and H3K4me3, and the ERα-ChIA-PET in (E). The numbers on the right of each track are the highest density value. 3C-qPCR mean values and

standard error of means (SEM) from three independent experiments are shown. Also see Figure C.6 and Figure C.7.

An interesting example is the SHH gene that was expressed in MCF7 but not in K562 cells (Figure 4.7C). SHH is important in development and related to certain cancers [25]. Transcription of SHH is controlled by its enhancer which is located 1 Mb away and embedded in the intronic region of LMBR1; point mutation in this enhancer site is known to cause preaxial polydactyly, a common congenital limb malformation in mammals [25]. We found abundant interaction data between the SHH promoter and the previously characterized SHH enhancer site in the LMBR1 intronic region in MCF7 cells, but no interaction data in K562 cells (Figure 4.7C), which correlated well with their SHH transcription status. This is consistent with earlier observations [26].

In another interesting example, we identified two major interaction sites located ~600 kb and ~1 Mb downstream from the IRS1 gene promoter. IRS1 is known to participate in type-2 diabetes (T2D) mellitus, and is found specifically expressed in MCF7 cells (Figure 4.7D). A recent GWAS study uncovered a cluster of SNPs that is genetically associated with high risk to insulin resistance, T2D, and coronary artery heart disease [27]. This high risk locus is found located in one of the IRS1 enhancer sites (Figure 4.7D). Thus, our data provides experimental evidence to suggest that this disease-risk locus could be physically connected with the IRS1 promoter, potentially serving as a critical long-range enhancer to regulate the expression of IRS1, in a similar manner as the SHH locus. Other examples of long-range and cell-specific enhancer-promoter interactions in MCF7 and K562 are shown in Figure C.7. Taken together, these results suggest that ChIA-PET interaction data may better inform the association of a SNP with a gene involved in a disease process by providing evidence for direct physical interactions.

4.5 Discussion

Through genome-wide mapping, we comprehensively analyzed RNAPII-associated long-range chromatin interactions. Our most interesting finding was the extensive promoter-promoter interactions among proximal and distant genes from 5 human cell-lines, which indicated that this mechanism is common in cells. Our work with reporter gene and siRNA knockdown assays provided experimental evidence that many promoters in the multigene complexes can cooperatively regulate the activity of other promoters with which they interact. Our observations thus blurred the conventional definition of promoter and regulatory elements for transcription. With such promoter-promoter interactions, we speculate that genetic error at one particular promoter might also propagate to other promoters and hence could lead to pleiotropic consequences depending on the interaction network within a cell type. Intriguingly, the multigene complexes illustrated in this study are, in principle, akin to the bacterial operon as a mechanism for coordinated transcriptional regulation of related genes, suggesting the possibility of a chromatin-based operon mechanism (chro-operon or chroperon) for spatiotemporal regulation of gene transcription in eukaryotic nuclei. However, the "chroperon" expression is not dependent on the linear arrangement of the genes, but is highly dynamic and can adopt a multitude of cassette configurations because of the combinatorics permitted by the looping interactions. Alternatively, these interactions could reflect stochastic movement of proximal and distant active genes to localized transcription factories.

An important question is how these multigene complexes are organized. A likely model is that a suite of protein factors for modulating gene expression in a functional

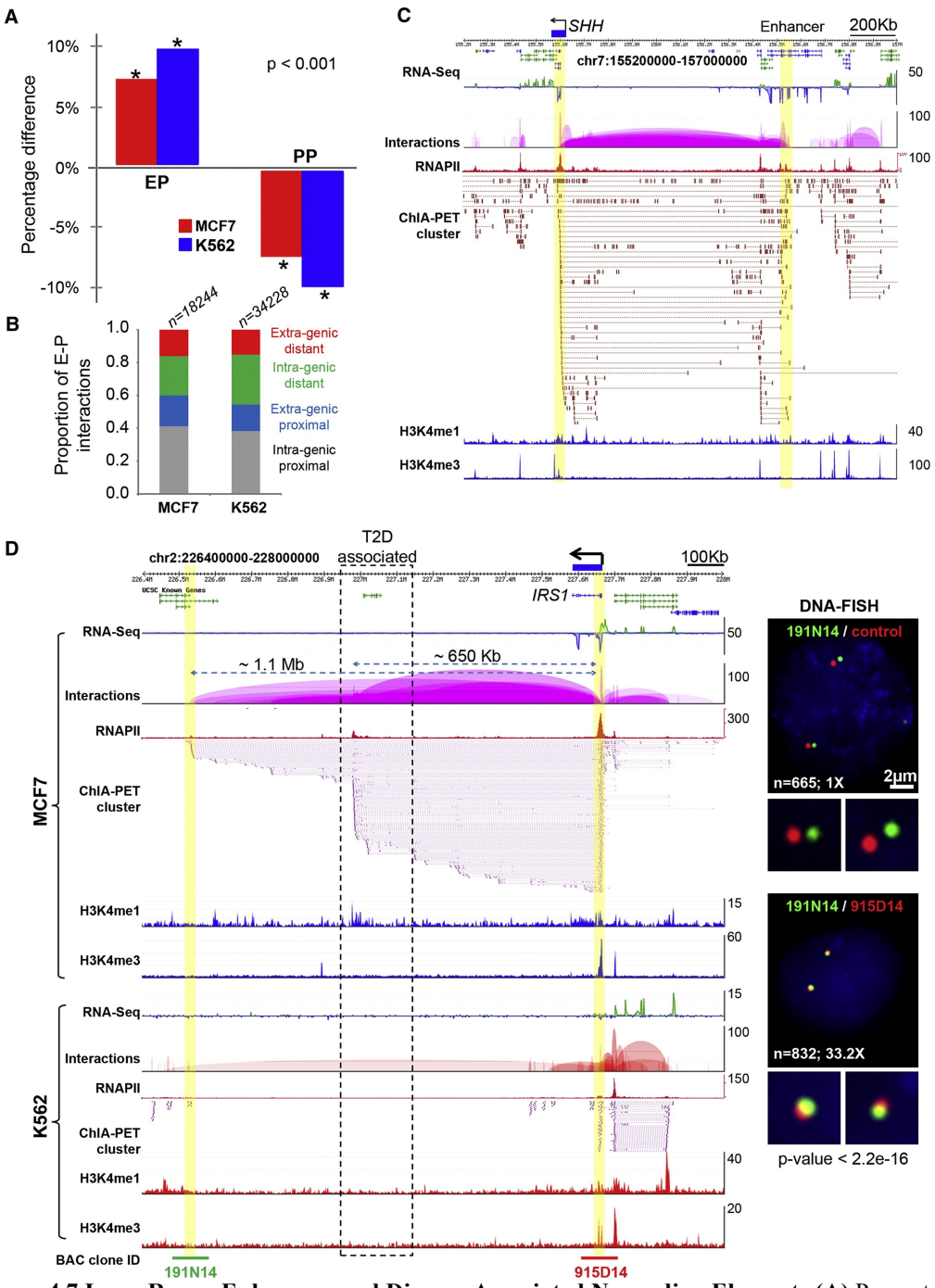


Figure 4.7 Long-Range Enhancers and Disease-Associated Noncoding Elements.(A) Percentage difference of enhancer-promoter (EP) and promoter-promoter (PP) interactions in cell-specific versus common genes from MCF7 and K562 cells. The representation of EP interactions is significantly increased in cell-specific interactions, while the representation of PP interactions is decreased, when compared to interactions that are common to both cell lines. **(B)** Proportional distribution of 4 classes of enhancers observed in two cell lines based on locations in relation to gene coding regions. "Intragenic proximal" enhancers locate inside of gene-body (mostly introns) and interact with the nearby promoters. "Extragenic proximal" enhancers locate outside of gene body and interact with the nearby promoters. "Intragenic distal" enhancers locate inside of gene body (mostly introns), bypass nearby genes and interact with faraway gene promoters in long-distance. "Extragenic distal" enhancers locate outside of gene body, bypass nearby genes and interact with faraway gene promoters in long-distance. **(C)** Long-range interactions between SHH

(highlighted in yellow, left) and its enhancer located about 1 Mb away in an intron of LMBR1 (highlighted yellow, right). The SHH expression is specifically seen in MCF7 cells. **(D)** Long-range interactions between IRS1 promoter and two enhancers as well as strong IRS1expression are seen in MCF7, but not in K562 cells. The dotted line box indicates the enhancer region that contains SNPs associated with insulin resistance, type-2 diabetes (T2D) and coronary artery heart disease identified by a GWAS study. The interactions of enhancer located 1.1 Mb away to IRS1 promoter (highlighted in yellow) is validated by DNA-FISH (right). The BAC clones and genomic segments used for DNA-FISH are indicated at the bottom. Tracks included in (C) and (D) are RNA-Seq density profile, interaction loop view, RNAPII peaks, ChIA-PET interaction PETs, ChIP-Seq density profile of H3K4me1 and H3K4me3 marks. Also see Figure C.7 and Table S5.

regulatory cassette may result in optimal stoichiometry when aggregated in 3D space. This clustering also draws the regulated genes into a common spatial domain, similar to how the nucleolus is organized. The interacting regions can be established and/or maintained by potential chromatin bridging proteins such as cohesins [28] and CTCF [29], and this process might be facilitated by chromatin remodeling proteins [17], all of which are enriched at the interacting sites defined by RNAPII ChIA-PET data.

Long-range chromatin interactions including enhancer-promoter interactions are increasingly being recognized as an important mechanism to regulate many important genes. However, methods to identify such long-range relationships have been technically challenging. High-throughput approaches such as ChIP-Seq and DNase-Seq are efficient in identifying potential regulatory sites, but lack the ability to interrogate the connectivity between the prospective enhancers and their target gene promoters. In this study using RNAPII as the protein target for ChIA-PET analysis, we identified a comprehensive repertoire of distant regulatory elements directly interacting with gene promoters. Many of them act through ultra-long-range chromatin interactions. Such distal enhancer-promoter relationships are particularly difficult to be identified by other approaches. As demonstrated in the cases of SHH and IRS1, long range interactions derived from ChIA-PET data could provide the connectivity of GWAS-identified high-risk loci to their target genes, and thus offer possible mechanistic explanations to the function of disease-

associated noncoding elements. Further investigation of spatial architectures revealed in this study will enhance our understanding of transcription regulation in normal and diseased conditions of human cells.

4.6 Experimental Procedures

4.6.1 Cell Culture

Five cell lines, namely MCF7 (ATCC# HTB-22), K562 (ATCC# CCL-243), HCT116 (ATCC# CCL-247), HeLa (ATCC# CCL-2.2), and NB4, were grown under standard culture conditions and harvested at log phase.

4.6.2 ChIA-PET

Harvested cells were cross-linked using 1% formaldehyde followed by neutralization with 0.2M glycine. Chromatin was isolated and subjected to the ChIA-PET procedure [8]. The ChIA-PET sequence reads were analyzed using ChIA-PET Tool [30]. The data are available from NCBI/GEO (ID: GSE33664). Control and reproducibility analyses are described in Figure C.8.

4.6.3 RNA-Seq Data

MCF7 mRNA was isolated following the protocol described in Ruan et al. [31] for strand-specific RNA-Seq analysis by SOLiD sequencing platform. The rest of the RNA-Seq datasets for other cell-lines were retrieved from the ENCODE data repository site (http://genome.ucsc.edu/ENCODE/).

4.6.4 ChIP-Seq Data

The ChIP-Seq data were retrieved from [19], [32] and the ENCODE data repository

site (http://genome.ucsc.edu/ENCODE/).

4.6.5 RNAPII IF Stain and DNA-FISH

MCF7 cells were fixed using 4% formaldehyde followed by permeabilization with 0.04% Triton-X. After blocking with donkey serum, cells were incubated with primary antibody (8WG16) overnight followed by Cy3 conjugated secondary antibody for 1 hr. IF-stained cells were post-fixed and subjected to dehydration by 70, 80, 100% ethanol series, rehydration with 2× SSC and denaturation in 2× SSC/50% formamide at 80°C for 40 min. Biotin-16-dUTP and digoxigenin-11-dUTP labeled DNA probes were hybridized to cells at 37°C overnight in a humid chamber. Slides were washed, stained with DAPI, mounted and visualized by a Carl Zeiss LSM confocal microscope.

4.6.6 Quantitative Chromosome Conformation Capture Analysis

Targeted 3C products were analyzed by qPCR. The 3C-qPCR protocol was adapted and modified from the previous publication [8].

4.6.7 Luciferase Reporter Gene Assay

Dual luciferase assays were performed as described [20]. Testing fragments were cloned into pGL4.10-basic vector. Constructs were transfected into MCF7 cells, and luciferase activities were measured following standard protocols.

4.6.8 Statistical Analysis

All the statistical tests were conducted with the R statistical package (http://www.r-project.org/). More details are available in Extended Experimental Procedures (online).

4.7 References

1. Jacob, F., Perrin, D., Sanchez, C. & Monod, J. [Operon: a group of genes with the

expression coordinated by an operator]. *C. R. Hebd. Seances Acad. Sci.* **250**, 1727–1729 (1960).
2. Pauli, D., Tonka, C. H. & Ayme-Southgate, A. An unusual split Drosophila heat shock gene expressed during embryogenesis, pupation and in testis. *J Mol Biol* **200**, 47–53 (1988).
3. Zorio, D. A., Cheng, N. N., Blumenthal, T. & Spieth, J. Operons as a common form of chromosomal organization in C. elegans. *Nature* **372**, 270–272 (1994).
4. Cook, P. R. The organization of replication and transcription. *Science* **284**, 1790–1795 (1999).
5. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**, 292–301 (2001).
6. van Steensel, B. & Dekker, J. Genomics tools for unraveling chromosome architecture. *Nat Biotechnol* **28**, 1089–1095 (2010).
7. Cope, N. F., Fraser, P. & Eskiw, C. H. The yin and yang of chromatin spatial organization. *Genome Biol* **11**, 204 (2010).
8. Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
9. Versteeg, R. *et al.* The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res* **13**, 1998–2004 (2003).
10. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes are compact. *Trends Genet* **19**, 362–365 (2003).
11. Taylor, J. Clues to function in gene deserts. *Trends Biotechnol.* **23**, 269–271 (2005).
12. Hah, N. *et al.* A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* **145**, 622–634 (2011).
13. Su, A. I. *et al.* Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci USA* **99**, 4465–4470 (2002).
14. Hou, C., Dale, R. & Dean, A. Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proc Natl Acad Sci USA* **107**, 3651–3656 (2010).
15. Schoenfelder, S. *et al.* Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet* **42**, 53–61 (2010).
16. Singer, G. A. C., Lloyd, A. T., Huminiecki, L. B. & Wolfe, K. H. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. *Mol Biol Evol* **22**, 767–775 (2005).
17. Euskirchen, G. M. *et al.* Diverse roles and interactions of the SWI/SNF chromatin remodeling complex revealed using global approaches. *PLoS Genet* **7**, e1002008 (2011).
18. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
19. Joseph, R. *et al.* Integrative model of genomic factors for determining binding site selection by estrogen receptor-α. *Mol Syst Biol* **6**, 456 (2010).
20. Pan, Y. F. *et al.* Regulation of estrogen receptor-mediated long range transcription via evolutionarily conserved distal response elements. *J. Biol. Chem.* **283**, 32977–32988 (2008).
21. Baù, D. *et al.* The three-dimensional folding of the α-globin gene domain reveals

formation of chromatin globules. *Nat Struct Mol Biol* **18**, 107–114 (2011).
22. Kim, J. *et al.* A Myc Network Accounts for Similarities between Embryonic Stem and Cancer Cell Transcription Programs. *Cell* **143**, 313–324 (2010).
23. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
24. Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461**, 199–205 (2009).
25. Lettice, L. A. *et al.* Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc Natl Acad Sci USA* **99**, 7548–7553 (2002).
26. Amano, T. *et al.* Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Dev. Cell* **16**, 47–57 (2009).
27. Kilpeläinen, T. O. *et al.* Genetic variation near IRS1 associates with reduced adiposity and an impaired metabolic profile. *Nat Genet* **43**, 753–760 (2011).
28. Merkenschlager, M. Cohesin: a global player in chromosome biology with local ties to gene regulation. *Curr Opin Genet Dev* **20**, 555–561 (2010).
29. Handoko, L. *et al.* CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* **43**, 630–638 (2011).
30. Li, G. *et al.* ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol* **11**, R22 (2010).
31. Ruan, Y. *et al.* Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res* **17**, 828–838 (2007).
32. Raha, D., Hong, M. & Snyder, M. ChIP-Seq: a method for global identification of regulatory elements in the genome. *Curr Protoc Mol Biol* **Chapter 21**, Unit 21.19.1–14 (2010).
33. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).

Chapter 5: CAPE - Coupled Analysis of Polymerase Binding and Expression By Comparing ChIP-Seq and RNA-seq

5.1 Abstract

Motivation: Next-generation sequencing assays such as RNA-seq RNA polymerase II (RNAPII) ChIP-Seq enable researchers to study transcription on a genome-wide scale across different time points, samples, and organisms. As next-generation sequencing has become more affordable, researchers are generating many paired (i.e. from the same cell line, tissue, or conditions) RNA-seq and RNAPII ChIP-Seq data sets as a "natural experiment" to elucidate transcription by relating DNA binding to mRNA abundance; however a specialized tool to quickly analyze such paired experiments does not exist. We present CAPE (Coupled Analysis of Polymerase and Expression), a multiplatform tool to analyze these paired experiments. CAPE categorizes transcripts based on mRNA abundance and RNAPII binding, compares orthologous features between different replicates, samples, or organisms, and summaries the results.

Availability and Implementation: CAPE is implemented in Java 1.6. Binaries, source code, and other supplemental documentation are available from our website (http://cape.gersteinlab.org).

5.2 Introduction

Next-generation sequencing (NGS) and short-read technologies have enabled new assays to study transcription. By using an antibody designed to target a particular protein

of interest, Chromatin Immunoprecipitation (ChIP)-based assays can identify transcription factor binding sites on a genome-wide scale (Johnson *et al.*, 2007). The typical output from a ChIP-Seq experiment includes a "signal map" that corresponds to the localization of a transcription factor as well as a list of high-confidence binding sites derived from a peak calling program (Pepke *et al.*, 2009). The RNA-seq assay uses NGS to measure transcript abundance by sequencing RNA-derived cDNA to produce a "signal map" of transcript abundance as well as a quantitative, normalized measure of transcription for each transcript such as Reads Per Kilobase Per Million Mapped Reads (RPKM) (Pepke *et al.*, 2009; Mortazavi *et al.*, 2008; Wang *et al.*, 2009). Combining the results of these assays allows researchers to elucidate the relationship between transcription factor binding and mRNA abundance on a genome-wide scale.

Relating the level of RNAPII binding near a transcript's promoter region to mRNA abundance is a very useful technique to quickly identify whether a transcript is being actively transcribed, is inactive, or if RNAPII is poised for transcription, allowing researchers to understand more about a cell's transcriptional program on a genome-wide level. As the popularity of RNA-seq and ChIP-Seq has grown and been supported by several large consortia, many matched (i.e. produced from the same organism, sample, cell line, or conditions) ChIP-Seq and RNA-Seq datasets have been deposited into public databases such as the UCSC Genome Browser, modMine, and GEO (Barrett *et al.*, 2011; Contrino *et al.*, 2012; Rosenbloom *et al.*, 2012). In fact, the Roadmap Epigenomics, ENCODE, and modENCODE consortia have made 162 RNAPII ChIP-Seq tracks available, 116 of which have matching expression data. The deposition rate of this particular experiment pair should only continue to increase. RNAPII is often used to

calibrate and evaluate ChIP-Seq protocols, as the RNAPII antibody is very robust and produces very strong, clean results. Additionally, RNAPII ChIP-Seq is almost always paired with RNA-seq experiments because RNA-seq is straightforward and unlocks the relationship between RNAPII binding and mRNA abundance with minimal additional work. This relationship can be compared across different organisms or samples to explore variation in transcriptional programs. Many ChIP-Seq-related tools currently focus on punctate signals rather than the broader signals of RNAPII ChIP-Seq and do not integrate RNA-seq. Because the number of matched experiments in public databases should only continue to grow, a specific tool for polymerase analysis is both necessary and timely.

In this manuscript, we present the Coupled Analysis of Polymerase binding and Expression (CAPE), a multiplatform program that integrates the information provided by RNAPII ChIP-Seq and RNA-seq to identify transcripts that are likely active, inactive, or where RNAPII is stalled and poised for future transcription. CAPE also allows for the comparison of transcript states between different samples or organisms given a list of orthologs, allowing researchers to fully unlock the information contained in their RNAPII ChIP-Seq and RNA-seq experiments and facilitate comparative analyses.

5.3 Description

CAPE has been specifically designed with the output from RNAPII ChIP-Seq, and RNA-seq in mind. CAPE is comprised of three components: an analysis module (CAPE-analyze), a comparison module (CAPE-compare) and a Java library of data structures and functions (AnnotationLibrary). A sample workflow for transcript analysis is shown in Figure 1.

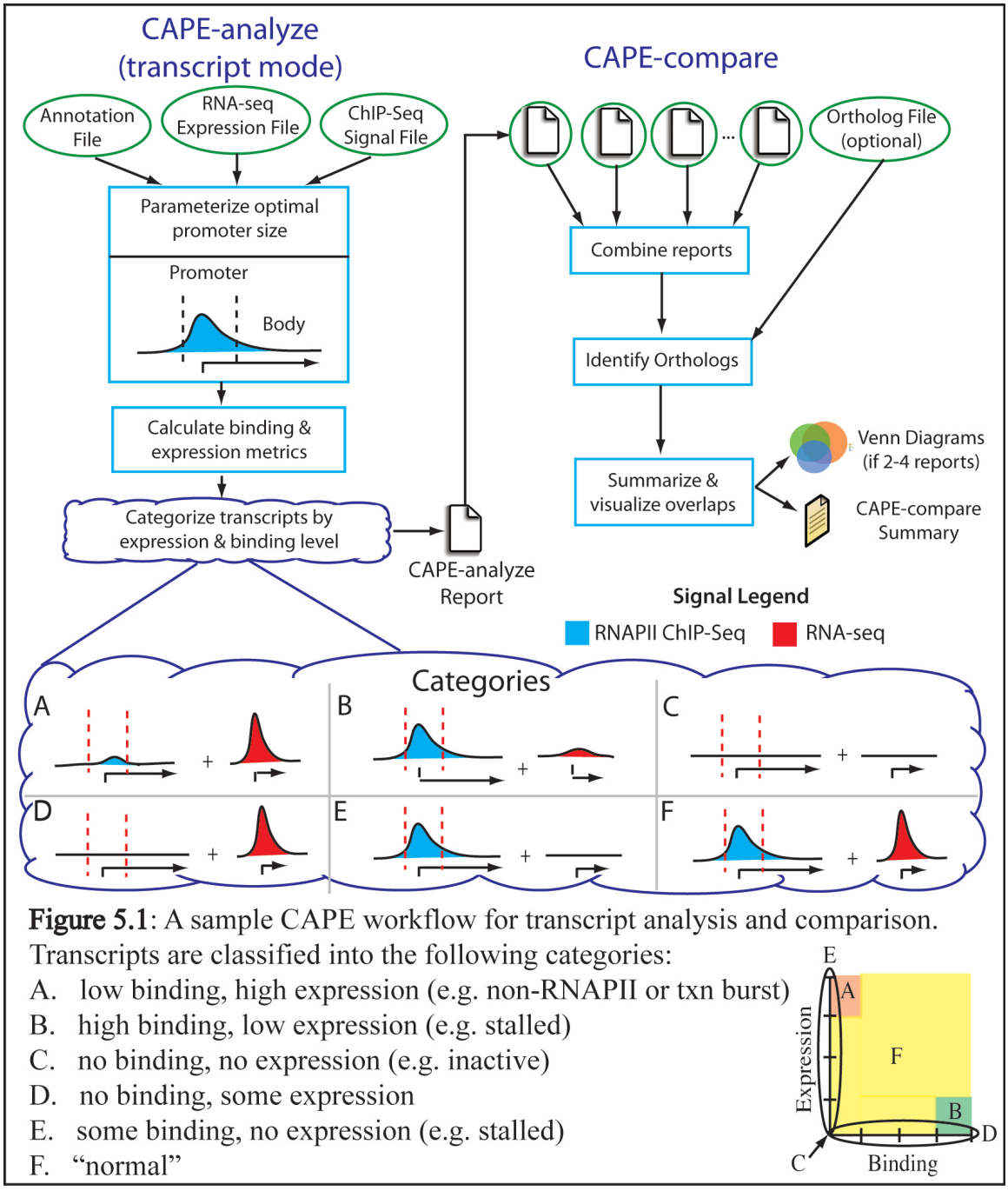


Figure 5.1: A sample CAPE workflow for transcript analysis and comparison. Transcripts are classified into the following categories:

A. low binding, high expression (e.g. non-RNAPII or txn burst)
B. high binding, low expression (e.g. stalled)
C. no binding, no expression (e.g. inactive)
D. no binding, some expression
E. some binding, no expression (e.g. stalled)
F. "normal"

CAPE-analyze produces a straightforward text report listing summary information for each transcript as well as its likely state while CAPE-compare allows for the comparison of orthologous transcripts between different samples or organisms. Given a ChIP-Seq signal file (Kent *et al.*, 2010), a file containing transcript annotations, and an RNA-seq quantification file from Cufflinks (Trapnell *et al.*, 2010), RSeqTools (Habegger

et al., 2011), or provided as part of the annotation file, CAPE-analyze will parameterize the average promoter size by analyzing signal aggregation around TSSs (see Users Guide), thereby enabling CAPE to analyze both compact and sparse genomes. CAPE determines the amount of RNAPII ChIP-Seq signal present in the promoter and the transcript body, then calculates two different metrics. The first, stalling index, is the ratio of promoter signal to body signal and is most appropriate for RNAPII ChIP-Seq or GRO-Seq experiments that target both initiation and elongation phases (Core *et al.*, 2008). The second metric compares the relative level of ChIP-Seq signal in the promoter to the mRNA abundance for the transcript determined by RNA-seq using either percentile-based or absolute cutoff values. CAPE-analyze produces a summary table containing the metrics and classifications for each transcript.

5.3.1 CAPE-compare

Researchers often want to quickly compare sets of matched experiments, for example when comparing different samples of a developmental timecourse, orthologous transcripts between different organisms, or samples with diseased and normal phenotypes. CAPE-compare allows quick comparison of multiple CAPE-analyze reports, producing a summary table and, in the case of two or three samples, an R script to produce Venn diagrams and an additional summary data in HTML. CAPE-compare will compare all common transcripts if comparing samples from the same organism or a tab-delimited ortholog list can be provided to allow comparison between organisms or to limit the transcripts considered. Several example use cases showing the applicability of CAPE-analyze and CAPE-compare in addition to sample input and output can be found at the CAPE website (http://cape.gersteinlab.org).

5.3.2 AnnotationLibrary

AnnotationLibrary is a collection of data structures and classes used by CAPE-analyze that we are making publicly available to the developer community. Further development and extension of AnnotationLibrary is openly encouraged and full Javadoc documentation is available at the CAPE website.

5.4 Discussion and Conclusion

In summary, CAPE allows researchers to quickly and easily analyze paired RNAPII ChIP-Seq and RNA-seq experiments to explore transcription genome-wide between different samples or organisms. As shown in our use cases, CAPE-analyze can be used to identify transcripts with unusual relationships between RNAPII binding and gene expression, identifying features for further analysis. A transcript with a stalled promoter, for example, may be lying in wait for an external stimulus before transcription commences. Transcripts with these unusual relationships are often the most interesting. One can easily imagine such transcripts playing a role in disease or during organism development. In fact, one of our use cases identifies the differences between human, worm, and fly embryo transcripts using publicly available data from the ENCODE and modENCODE consortia. We expect that CAPE will prove increasingly useful to the genomics community.

5.5 Acknowledgements

CAPE uses the following external libraries: Google Guava, Apache Commons Math, Apache Commons CLI, VennDiagram, and the Broad Institute's BigFile. Please see the Users Guide (Appendix D).

5.6 Funding

Funding is provided by grants to MG from the National Institutes of Health and National Human Genomic Research Institute.

5.7 References

Barrett,T. *et al.* (2011) NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic Acids Res*, **39**, D1005–10.

Contrino,S. *et al.* (2012) modMine: flexible access to modENCODE data. *Nucleic Acids Res*, **40**, D1082–8.

Core,L.J. *et al.* (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845–1848.

Habegger,L. *et al.* (2011) RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics*, **27**, 281–283.

Johnson,D.S. *et al.* (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

Kent,W.J. *et al*. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204–2207.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5**, 621–628.

Pepke,S. *et al.* (2009) Computation for ChIP-seq and RNA-seq studies. *Nat Methods*, **6**, S22–32.

Rosenbloom,K.R. *et al.* (2012) ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res*, **40**, D912–7.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, **28**, 511–515.

Wang,Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, **10**, 57–63.

Chapter 6: Summary and Future Directions

This dissertation presents several strategies and methods to analyze data from ChIP-Seq and related technologies, how to effectively integrate such data with those from other technologies to study transcription at multiple scales, and describes a tool to aid researchers in quickly discerning the relationship between RNA Polymerase II binding and mRNA abundance between samples. First, I describe in Chapter 1 the characteristics of different reference DNA types used for scoring ChIP-Seq data, how these characteristics can affect the downstream results from ChIP-Seq peak calling, and through the integration of different data sources hypothesize why peaks are present in cross-linked, sonicated DNA. This work also disproved early assumptions about ChIP-Seq reference DNA types, particularly that a simulated randomized background track was an adequate approximation of an experimental control.

After discussing the ChIP-Seq assay as well as considerations that may affect peak calling when choosing a reference sample in Chapter 2, Chapter 3 uses ChIP-Seq to explore the binding sites of four subunits of the human SWI/SNF chromatin remodeling complex and examines their interrelationships with each as well to other genomic features. Despite many RNAPII and histone mark datasets being deposited in public repositories, the bridge between genomic and epigenomic control of transcription was understudied. Chromatin remodelers can bridge this gap, as they can control the access of transcription factors to DNA and in the case of SWI/SNF, bind to acetylated histones typically present in promoter regions. Moreover tight control of nucleosome positioning is important for many different aspects of nuclear function. To underscore this point, I

integrated SWI/SNF regions with a diversity of data types including lamin-associated regions, replication origins and regions associated with RNA Pol III-based transcription. Bioinformatically, the analysis for this project was particularly challenging, as ChIP-Seq analysis had focused primarily upon single factors at the time that this work was conducted. Some subunits are also thought to have functions outside of the SWI/SNF complex, further complicating analysis. To solve this problem, my coauthor and I introduce the concept of a multi-factor domain in ChIP-Seq scoring when analyzing multiple members of a protein complex as well as provide a roadmap for scoring and analyzing non-standard factors (e.g. factors that do not only bind predominantly near promoter regions of expressed genes).

Chapter 4 extends upon the theme of ChIP-Seq to study transcription by describing long-range interaction analysis of RNAPII via ChIA-PET. By studying how DNA folds to bring components of the transcriptional machinery into close contact with each other, we were able to propose several different models of transcription in the nucleus and describe characteristics for each model. These models include basal transcription (i.e. a RNAPII peak at the promoter with no other interactions), the single-gene model where a promoter interacts with a distal region that may enhance transcription, and the multi-gene model where promoters of multiple genes and multiple distal regions share common interactions. Through the integration of our data with several other ChIP-Seq data sets from the same cell lines, I also showed that some complexes have subunits that are typically found in regions distal to promoters and are brought into close proximity with other subunits at the promoter via DNA folding.

Finally, Chapter 5 presents Coupled Analysis of Polymerase Binding and

Expression (CAPE). CAPE is a Java program designed to analyze the large number of paired RNAPII ChIP-Seq and RNA-seq experiments by cataloguing transcripts based on the observed relationship between RNAPII binding and mRNA abundance. The program is constructed to handle comparisons from multiple samples, whether these samples come from the same organism or from different, related organisms. I expect that CAPE will prove very useful for identifying and comparing variations in transcriptional programs between samples. This may take the form of comparing disparate organisms given a list of orthologs, exploring transcription at different developmental time points within the same organism, or identifying differences between diseased and normal states. CAPE is provided as publicly available, open-source software.

In the future, I envision that transcription factor binding assays such as ChIP-Seq will continue to play a major role in understanding how biological systems function. New assays are already being developed that can identify transcription factor binding sites at single-nucleotide resolution. The various DNA reference samples discussed in Chapter 1 will have to be re-evaluated in light of these new assays to see if their characteristics and biases remain consistent with their ChIP-Seq counterparts. New strategies for data integration will also become necessary as laboratories continue to ChIP new, non-standard factors and as they try to discern the interactions between subunits of a protein complex. As more experiments of this type become available in the public repositories, tools can be developed to simulate the binding of protein complex subunits and probabilistic methods designed to predict whether ChIP-Seq peaks from two related subunits are likely to form a complex. As described in Chapter 4, long-range interaction assays are still in their infancies. Current efforts are being made on the experimental side

to increase the resolution of these experiments while saturating the set of possible interactions obtained. The latter will pose a particular problem for statistical analysis, as current approaches to calling ChIA-PET interactions breakdown when a large number of possible interactions are obtained. Additionally, improved statistical methods and interaction calling will allow the field to confidently analyze interactions between different chromosomes. I expect that this type of analysis will prove particularly interesting when viewing transcriptional regulation on a system-wide level. Finally, CAPE represents a useful tool for comparing RNAPII binding levels to mRNA abundance for a set of transcripts, but as with any tool there are always possible improvements. At present, CAPE uses a single, accepted measurement to represent mRNA abundance, but it can be extended to instead use the RNA-Seq signal track data in much the same way that CAPE uses ChIP-Seq signal data. This will allow for comparisons between components of a transcript such as exons and introns. More specific still, this would allow CAPE to compare first exons against other exons from the same transcript, first exons against other first exons in the same sample, or even first exons between orthologous transcripts from different organisms. In short, I expect the data deluge in the biological sciences to continue unabated and as such, all tools and methods will need to continue to evolve to meet analytical needs as larger, more complex questions are explored.

Appendix A: Supporting Documentation for Chapter 2

A.1 Supporting Materials and Methods

A.1.1 Growth of cells for ChIP-Seq, Sono-Seq, MNase digestion, naked DNA and qPCR

For RNA Polymerase II ChIP-Seq, normal IgG ChIP-Seq and Sono-Seq, HeLa S3 cells were grown in suspension in Joklik's modified minimal essential medium (MEM), supplemented with 10% FBS to a density of 6 x105 cells/mL. Cells were fixed with 1% formaldehyde at room temperature for 10 min and the fixation was terminated by the addition of glycine to a final concentration of 125 mM. The cells were washed in Dulbecco's PBS (Invitrogen), snap frozen as cell pellets in liquid nitrogen, and provided to us by the National Cell Culture Center (Biovest International Inc., Minneapolis, MN). For the preparation of the MNase-treated cells and naked DNA samples, HeLa S3 cells were grown in SMEM (Invitrogen), supplemented with glutamine, 10% FBS (Invitrogen), and antibiotics (penicillin-streptomycin) and harvested without crosslinking at a density of 5 x105 cells/mL. For qPCR, HeLa cells were grown in MEM supplemented with 10% fetal bovine serum (Atlanta Biologicals), 100 U/mL penicillin and 100 g/mL streptomycin. 107 cells were trypsinized, fixed in 10 mL MEM supplemented with 1% formaldehyde for 10 min at room temperature, and quenched by the addition of glycine to a final concentration of 125 mM.

A.1.2 Construction and sequencing of Illumina libraries

DNA samples were run though Qiagen MinElute PCR columns, eluted with 15 l of

Qiagen buffer EB and size-selected on 2% agarose E-gels (Invitrogen). Band-isolated fragments were gel-purified using a Qiagen gel extraction kit. Libraries were prepared according to DNA Sample Kit instructions (Illumina Part# 0801-0303) but substituting kit enzymes with those available from other suppliers. Briefly, DNA was end-repaired and phosphorylated with the End-It kit from Epicentre (Cat# ER0720). The blunt, phosphorylated ends were treated with Klenow fragment (3 to 5 exo minus; NEB, Cat# M0212s) and dATP to yield a protruding 3-'A' base for ligation of Illumina's adapters, which have a single 'T' base overhang at the 3 end. After adapter ligation (LigaFast, Promega Cat#M8221) DNA was PCR-amplified with Illumina genomic DNA primers 1.1 and 2.1 for 15 cycles using a program of 1) 30 s at 98C 2) 15 cycles of [10 s at 98C, 30 s at 65C, 30 s at 72C] and 3) a 5 min extension at 72C. For the Sono-Seq DNA with large inserts of 350-800 bp, the extension time for the 15 cycles was increased to 1 min. The final libraries were band isolated from an agarose gel to remove residual primers and adapters. Library concentrations and A260/A280 ratios were determined by UV-Vis spectrometry on a NanoDrop ND-1000 spectrophotometer (NanoDrop, Wilmington, DE). Purified library DNA was captured on an Illumina flowcell for cluster generation and sequenced on an Illumina Genome Analyzer II following the manufacturer's protocols.

A.1.3 Preparation of DNA for qPCR

Formaldehyde-fixed HeLa cell pellets were resuspended in cell lysis buffer (25 mM HEPES [pH 7.9], 1.5 mM MgCl2, 10 mM KCl, 1 mM DTT, 0.1% NP-40) supplemented with EDTA-free protease inhibitor cocktail (Roche, Indianapolis) and 0.5 mM PMSF at a concentration of 107 cell equivalents/mL and incubated on ice for 10 min. After centrifugation, the crude nuclear pellet was resuspended in nuclear lysis buffer (50 mM

HEPES [pH 7.9], 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS) supplemented with EDTA-free protease inhibitor cocktail and 0.5 mM PMSF at a concentration of 107 cell equivalents/mL. Chromatin was sheared at 4C, 10 x 30 s at 30+ s intervals on a Branson Microtip Sonifier 450 set at constant duty and an output level of 4. After centrifugation for 10 min at 16,000xg, chromatin was sonicated for an additional 0 to 5 min at 10 s intervals in 0.5 mL aliquots using a cup horn on a Misonix sonicator 4000 set at level 6. The sonicated chromatin was treated with RNase A (Invitrogen, Carlsbad, CA) for 10 min at room temperature and decrosslinked by boiling for 10 min. After an additional centrifugation for 10 min at 16,000xg, DNA was extracted with phenol:chloroform:isoamyl alcohol (25:24:1) (Ambion, Austin, TX) and purified through Qiagen PCR purification columns (Qiagen Inc. Valencia, CA). DNA was resolved by agarose gel electrophoresis and 100-500 bp and 1000-6000 bp sized DNA bands were excised and purified again through Qiagen PCR columns. The concentration and purity of the DNA was measured by A260 and A280 UV-Vis spectrometry on a NanoDrop ND-1000 spectrophotometer.

A.1.4 Quantitation by real-time PCR

For chromatin size selection experiments samples were assayed by quantitative PCR to assess the enrichment of genomic regions in either the 100-500 bp or 1000-6000 bp chromatin samples. PCR reactions contained 2 l DNA template, 3 l of 3.3 mM primer pairs, and 5 l of 2X EvaGreen reaction mix (FluoProbes Interchim, France). Quantitative PCR was performed on an Applied Biosystems 7500 Fast unit using a 10 min soak at 95oC, followed by 40 cycles of 5 s at 95C, 5 s at 55C and 20 s at 72C. Ct values were determined at threshold of 0.01. For each amplification product, the relative enrichment

in the 100-500 bp sample versus the 1000-6000 bp sample was determined using the formula relative enrichment = a*1.9Ct(1000-6000bp)-Ct(I100-500bp), where a was the constant associated with the ratio of the DNA concentration of the 100-500 bp and 1000-6000 bp samples and Ct is the threshold cycle.

A.1.5 Creation of ChIP-Seq and reference DNA sample aggregation plots

Uniquely-mapped reads were extracted from the corresponding standard Eland output files for each factor/reference DNA type, signal maps created, and aggregation plots created. A Python script was then used to create a signal map file in sgr format using a sliding window approach. The size of the sliding window used for each data set is shown in Table A.1. For each list of features, coordinates were obtained and converted to build hg18 of the human genome, when necessary, using UCSC's Liftover tool. A Perl script was then used to perform the aggregation. This script divides the region immediately upstream and downstream of a feature's start site into several bins. For each bin, all reads present for each nucleotide within the bin are summed and the average signal for the bin calculated. The bin scores corresponding to the same relative position for each feature are then averaged to produce a mean signal for each bin upstream and downstream of a feature's start position. These signals are finally normalized to the sum of the averages for the first four and last four bins for each feature type to produce the final ChIP-Seq aggregated signal values. For all factors/reference types in this analysis, we chose to use a total of 46 non-overlapping bins (23 on each side of a feature's start position) of length 90 base pairs for all features other than CpG islands. Due to the varied lengths of CpG islands, we could not apply a standard bin size across all islands. Instead, we partitioned each CpG island into 35 equal-sized bins and aggregated over each

fraction. We then extended this method to include additional bins of the same size regions flanking each CpG island.

A.1.6 Calculating Percent Feature Composition for Sono-Seq DNA and Pol II DNA

From the ranked lists described in the manuscript, all enriched regions comprised of fewer than 20 tags or possessing a fold-enrichment less than 5 were discarded. A subset of the peaks was then analyzed in a stepwise fashion by intersecting enriched regions against promoter regions of expressed Ensembl genes and against promoter regions of non-expressed genes. Remaining enriched regions were deemed to lie outside of promoter regions and classified as "Other." For this analysis, promoter regions are again defined as within 2.5 kb of a TSS. Data was added to the subset in 5% increments (i.e. iteration 1 would intersect enriched regions above the fifth percentile, iteration 2 above the tenth percentile, etc) until 100% of enriched regions were analyzed.

A.1.7 Creation of the Pol II and Sono-Seq rank-order plot

To create the rank-order plot, enriched regions are ranked by tag count (minimum: 20) and enrichment factor (minimum: 5), and q-value (maximum: 0.05). This subset of enriched regions is deemed to be "high-quality enriched regions." For Pol II and Sono-Seq, data is added in a stepwise-fashion at 5% increments in decreasing order of enrichment to create an analysis set. During each iteration, enriched regions contained in the analysis set are intersected against 5 ends of Ensembl genes. This process is repeated until all enriched regions are included in the analysis set. Regions intersecting promoter regions of Ensembl genes are further subdivided based upon whether they overlap promoters of expressed or non-expressed Ensembl genes. Enriched regions lying distal (2.5 kb) to the TSS of an Ensembl gene are classified as "other."

A.2 Supporting Results

A.2.1 Sono-Seq DNA signals show little increase over CTCF regions distal to promoters

Since Sono-Seq DNA regions are often associated with promoter regions, particularly those of expressed genes, we examined whether Sono-Seq signal is depleted over regions of closed chromatin or insulators. CCCTC-binding factor (CTCF) plays many roles in the human genome including behaving as a chromatin barrier, binding insulator elements to restrict transcriptional enhancers from activating unrelated promoters, and acting as an anchor for positioning neighboring nucleosomes [1]. We analyzed the association of CTCF sites distal to promoters by removing sites found within 2.5 kb of 5′ ends of known genes from a list of 127,172 CTCF sites obtained from Barski et al. [2]. ChIP signals were then aggregated using a random sample of 100,000 sites from the remaining 119,940 distal sites. We find that Pol II signal is elevated over both proximal and distal CTCF sites, as well as Sono-Seq and MNase-digested DNA signals to a lesser degree (Figure A.9).

A.2.2 Highly-transcribed regions are sonication-sensitive whereas centromeric repeats are sonication-resistant

As a complementary approach for examining how different genomic regions are affected by the size of DNA fragments in sonicated samples, we performed quantitative PCR analysis. DNA from sonicated chromatin was electrophoretically separated into small (100-500 bp) and large (1,000-6,000 bp) DNA fragments, and the amounts of DNA for various genomic regions were determined by quantitative PCR analysis (Figure A.8). The results are presented as the small:large ratio, where a value of 1.0 indicates

equimolar representation in the two samples. Several Pol II promoter regions are strongly overrepresented among the small DNA fragments (ratios ranging between 5-20). In contrast, the corresponding coding regions as well as two non-annotated, transcriptionally inactive regions of chromosome 21 are comparably represented in both the small and large DNA samples. Two of four Pol III genes as well as the 18S and 28S regions of the ribosomal DNA genes are also highly overrepresented among the small DNA fragments (ratios between 20-30). Interestingly, whereas a telomeric region is equally represented in the two samples, a centromeric region is extremely under-represented among the small DNA fragments (ratio of 0.03). Thus, sonication of crosslinked chromatin samples occurs in a highly non-random fashion (variation among genomic regions occurs over a 200-fold range), with preferential fragmentation occurring at promoters and in highly-transcribed regions, and strong resistance to fragmentation occurring near centromeric repeats.

A.3 Supporting Figures

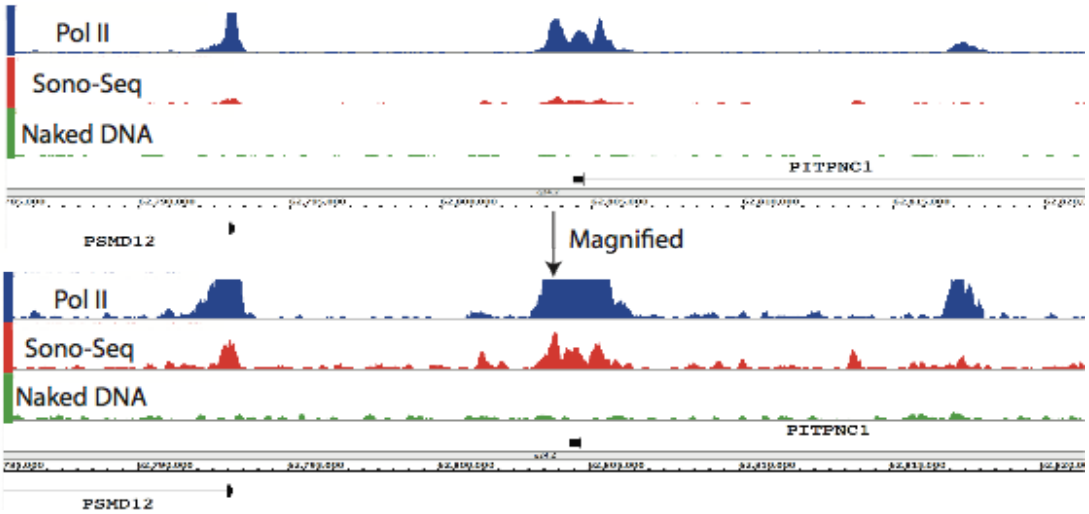


Figure A.1 Signal map showing Pol II ChIP DNA, Sono-Seq DNA from HeLa S3 cells and naked DNA. Signal maps are created with the IGB Browser (Affymetrix, Santa Clara, CA) and tracks are scaled based upon the number of uniquely mapped reads obtained for each sample type. The magnified view shows the same region but uniformly alters the scale for all tracks to show additional peak detail. This

figure shows signal levels between positions 62,755,000-62,821,500 of chromosome 17. Both PSMD12 and PITPNC1 are expressed in HeLa S3 based on RNA-Seq data [3].

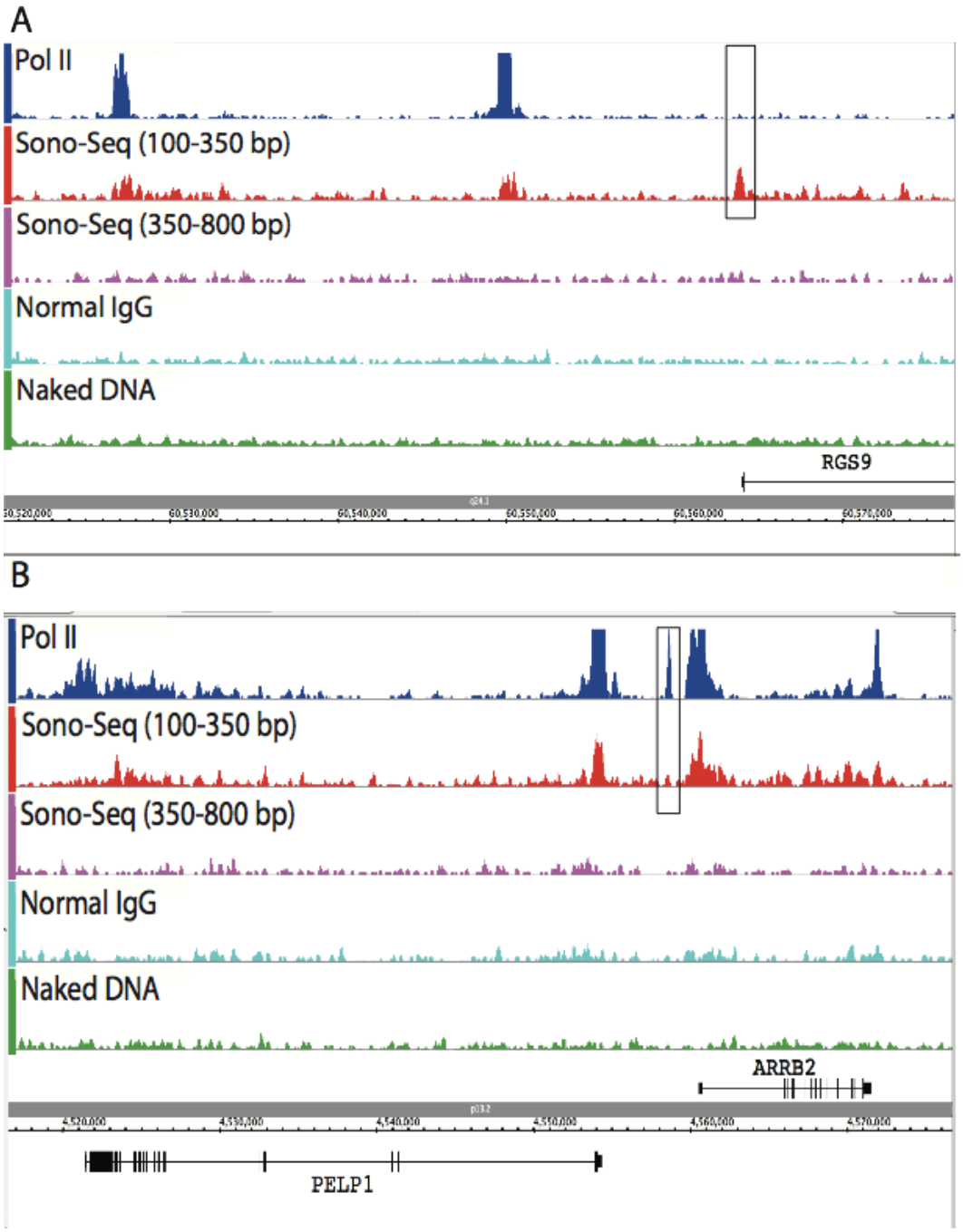


Figure A.2 Signal maps. A) Signal map showing Pol II ChIP DNA, Sono-Seq DNA (small and large fragment sizes), normal IgG, and naked DNA. All signals are in HeLa S3 cells. This figure shows signal levels between positions 4,517,000-4,576,000 of chromosome 17. Signal maps are created with the IGB Browser (Affymetrix, Santa Clara, CA) and tracks are scaled based upon the number of uniquely-mapped reads obtained for each sample type. Both PELP1 and ARRB2 are expressed in HeLa S3 based on RNA-Seq data [3]. The boxed region illustrates a region where an auxiliary Pol II peak is observed without an accompanying Sono-Seq peak near the ARRB2 promoter region, although other large Sono-Seq and Pol II

peaks are located at this loci. **B)** Signal map showing Pol II ChIP DNA, Sono-Seq DNA (small and large fragment sizes), normal IgG, and naked DNA. All signals are in HeLa S3 cells. This figure shows signal levels between positions 60,520,000-60,576,000 of chromosome 17. Signal maps are created with the IGB Browser (Affymetrix, Santa Clara, CA) and tracks are scaled based upon the number of uniquely-mapped reads obtained for each sample type. RGS9 is not expressed in HeLa S3 based on RNA-Seq data [3]. The boxed region shows a large Sono-Seq peak in the absence of a corresponding peak in Pol II.

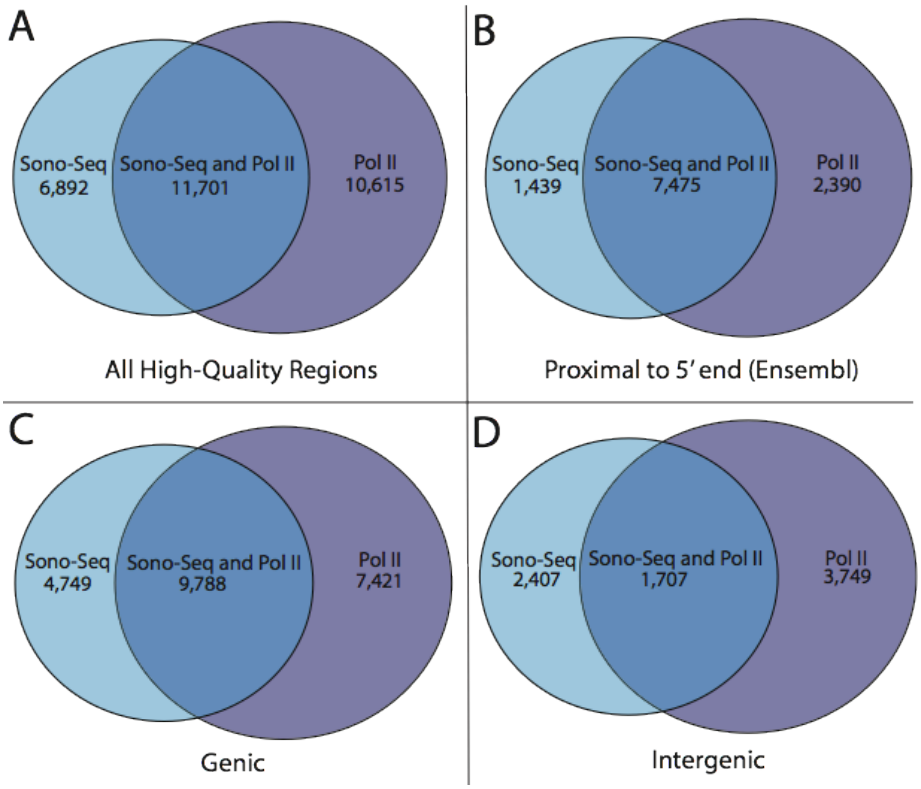


Figure A.3 Venn diagrams. Venn diagrams showing the number of Sono-Seq DNA and Pol II ChIP DNA regions, where both data sets were collected from HeLa S3 cells. Diagrams show intersections of all highly-enriched regions of Pol II and Sono-Seq DNA **A)** in the entire genome, **B)** proximal (within ±2.5 kb) to an Ensembl gene TSS, **C)** within or proximal to Ensembl genes, and **D)** distal to Ensembl genes. Intersections were performed using the Active Region Comparison Tool, which merged all peaks occurring within 500 bp before performing the intersections [4]. Results are different than one-way intersections used in the paper, as Sono-Seq and Pol II hits do not necessarily exhibit a one-to-one relationship.

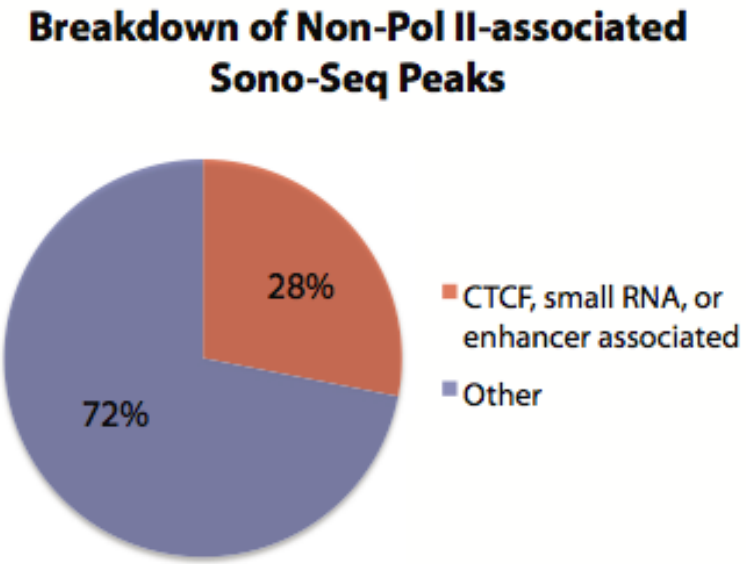


Figure A.4 Breakdown of non-Pol II-associated Sono-Seq peaks with respect to small RNAs, CTCF sites, and enhancers. All three of these intersecting sets were HeLa-derived [5-7]. Sono-Seq regions located > 1 kb from a Pol II peak were intersected against CTCF and enhancer sites (within 200 bp) and small RNAs (within 2 kb). 28% of the non-Pol II-associated peaks fall within one of these categories.

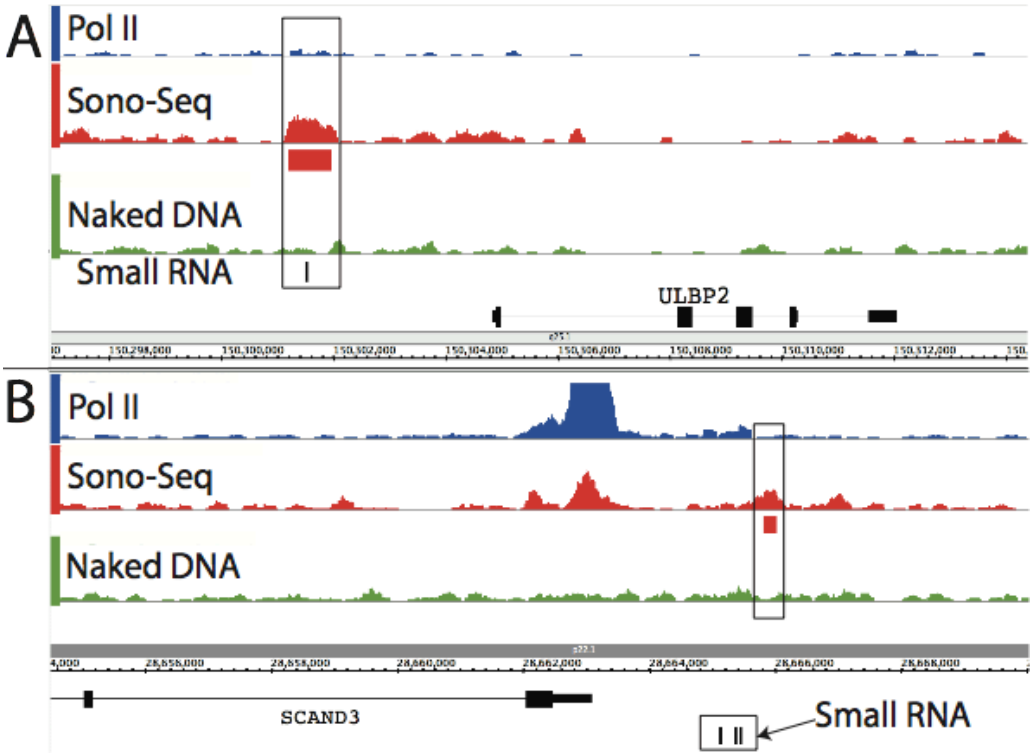


Figure A.5 Signal map. Signal map showing Pol II ChIP DNA, Sono-Seq DNA (small fragment sizes) and naked DNA relative to several small (< 200 nucleotides) RNAs from HeLa cells [5]. Signal tracks are scaled based upon the number of uniquely-mapped reads obtained for each sample type. Boxed regions in A) and B) show Sono-Seq peaks in the absence of corresponding RNA Pol II peaks and where several

small RNAs (small black rectangles near gene annotations) are within 1 kb of the Sono-Seq peaks. Neither SCAND3 nor ULBP2 are expressed in HeLa S3 cells based on RNA-Seq data [3]. Regions shown are from 28,654,000-28,670,000 on chromosome 6 for SCAND3 and from 150,297, 200-150,314,000 on chromosome 6 for ULBP2. Signal maps are created with the IGB Browser (Affymetrix, Santa Clara, CA).

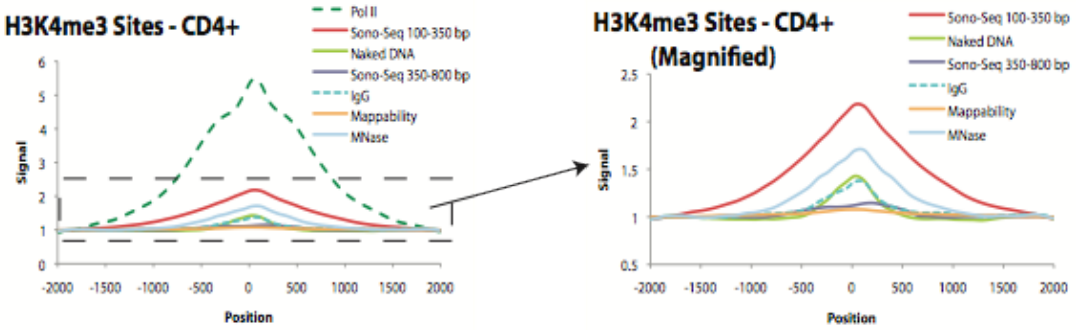


Figure A.6: Aggregation plot depicting average ChIP signal across a random sample of 100,000 H3K4me3 sites identified in CD4+ cells. The right panel is a magnified view of the region enclosed by the dotted box in the left panel. In the right panel, Pol II is removed and the scale is altered to allow for better comparison between reference sample types. Vertical axis units are consistent between all plots. Horizontal axis units are given in terms of nucleotides from the feature start site.

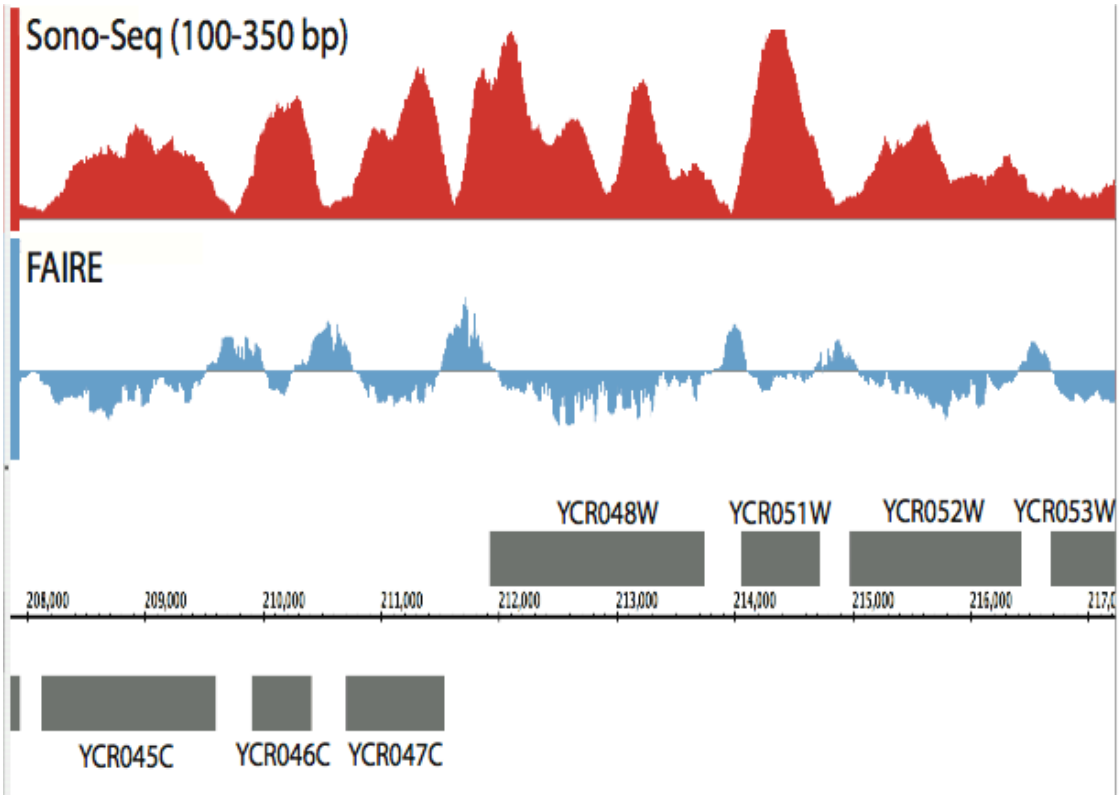


Figure A.7 Signal map showing Sono-Seq DNA and FAIRE signals in *Saccharomyces cerevisiae*. For FAIRE, signal levels above the axis are enriched whereas levels beneath the axis are depressed. Regions enriched in Sono-Seq appear anti-correlated with FAIRE signal. This figure shows signal levels between

positions 207,900-217,200 of *S. cerevisiae* chromosome 3. Signal maps are created with the IGB Browser (Affymetrix, Santa Clara, CA).

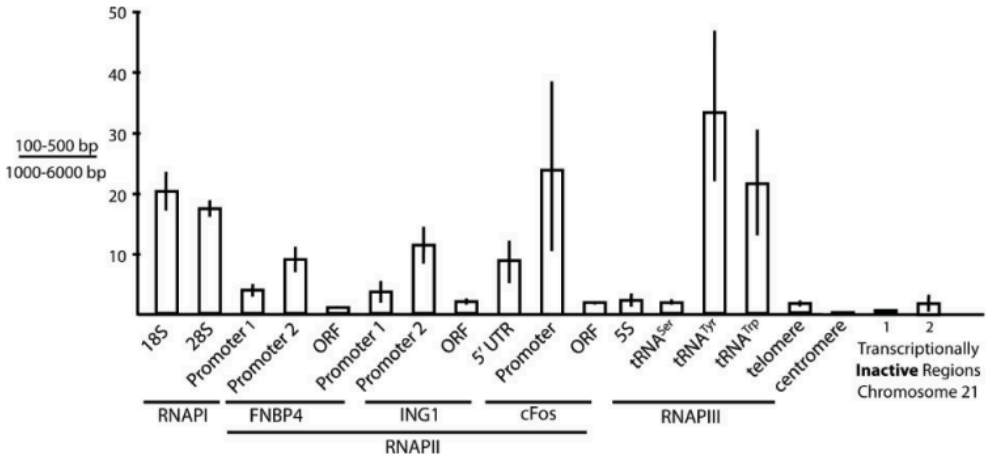


Figure A.8 Sonication efficiency varies greatly among genomic regions. DNA samples derived from 100-500 bp and 1,000-6,000 bp chromatin fragments were analyzed by qPCR and normalized to the concentration of the input DNA. The amount of DNA in the 100-500 bp sample was 2.8 ± 0.5 higher than in the 1,000-6,000 bp sample.

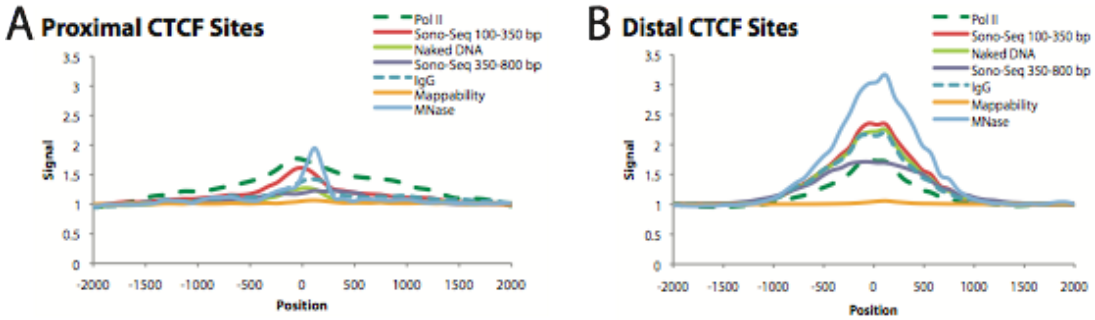


Figure A.9 Aggregation plots depicting average ChIP signal across 7,232 proximal and 100,000 distal CTCF sites. The same units and conventions as those from Figure 3.3 in the main text are used for the axes.

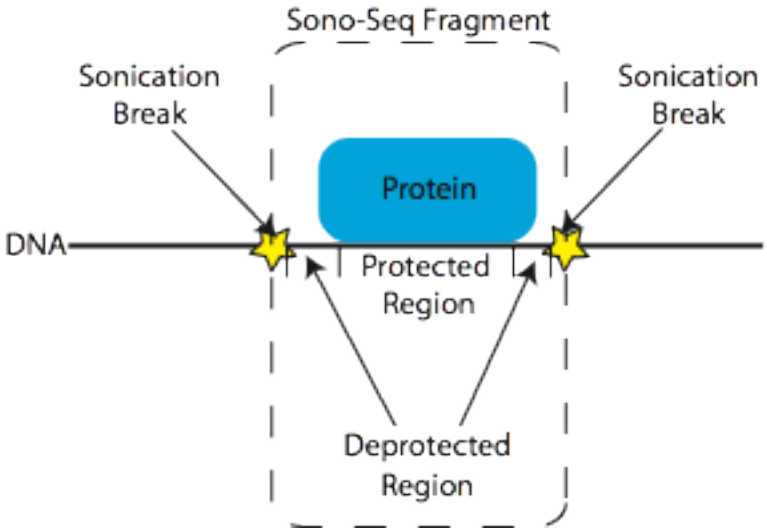


Figure A.10 Sono-Seq enriches for regions flanked by sonication-sensitive sites and are comprised of both protected and deprotected chromatin **regions.** Signal peaks are only seen at smaller Sono-Seq fragment sizes (100-350 bp).

A.4 Supporting Tables

Table A.1 Sliding windows sizes used to generate signal maps.

Factor/Reference Sample Type	Signal Map (SGR) Window Size
Pol II	200
Sono-Seq (100-350 bp, HeLa S3)	200
Sono-Seq (350-800 bp, HeLa S3	575
Naked DNA	200
IgG	200
MNase	200

A.5 Supporting References

1. Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet* 4, e1000138 (2008).
2. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* 129, 823–837 (2007).
3. Morin, R. *et al.* Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques* 45, 81–94 (2008).
4. Rozowsky, J. S. *et al.* The DART classification of unannotated transcription within the ENCODE regions: associating transcription with known and novel loci. *Genome Res* 17, 732–745 (2007).
5. Affymetrix ENCODE Transcriptome Project & Cold Spring Harbor Laboratory ENCODE Transcriptome Project Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* 457, 1028–1032 (2009).

6. Cuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* 19, 24–32 (2009).
7. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459, 108–112 (2009).

Appendix B: Supporting Documentation for Chapter 3

B.1 Supporting Figures

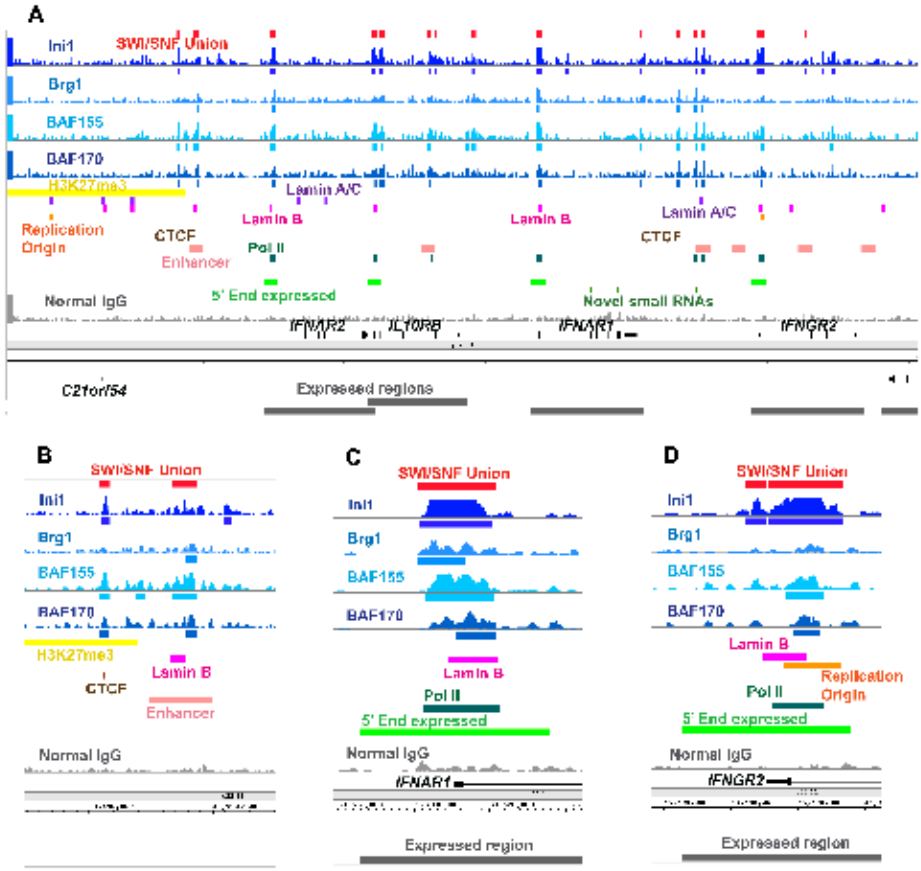


Figure B.1 SWI/SNF signals and target regions in the context of interferon receptor genes on chromosome 21. The coordinates shown are in hg18 and all regions were identified in HeLa cells as detailed in Table S1 and Materials and Methods. The vertical axis for each signal track is the count of the number of overlapping DNA fragments at each nucleotide position and is scaled from 0 to 40 for each track. Panel A displays a ~370 kb region on chromosome 21 containing genes encoding cytokine receptors. Panel B displays a ~20 kb region at the edge of an H3K27me3 domain. Panels C and D each display ~6 kb regions around the 5 ends of expressed genes.

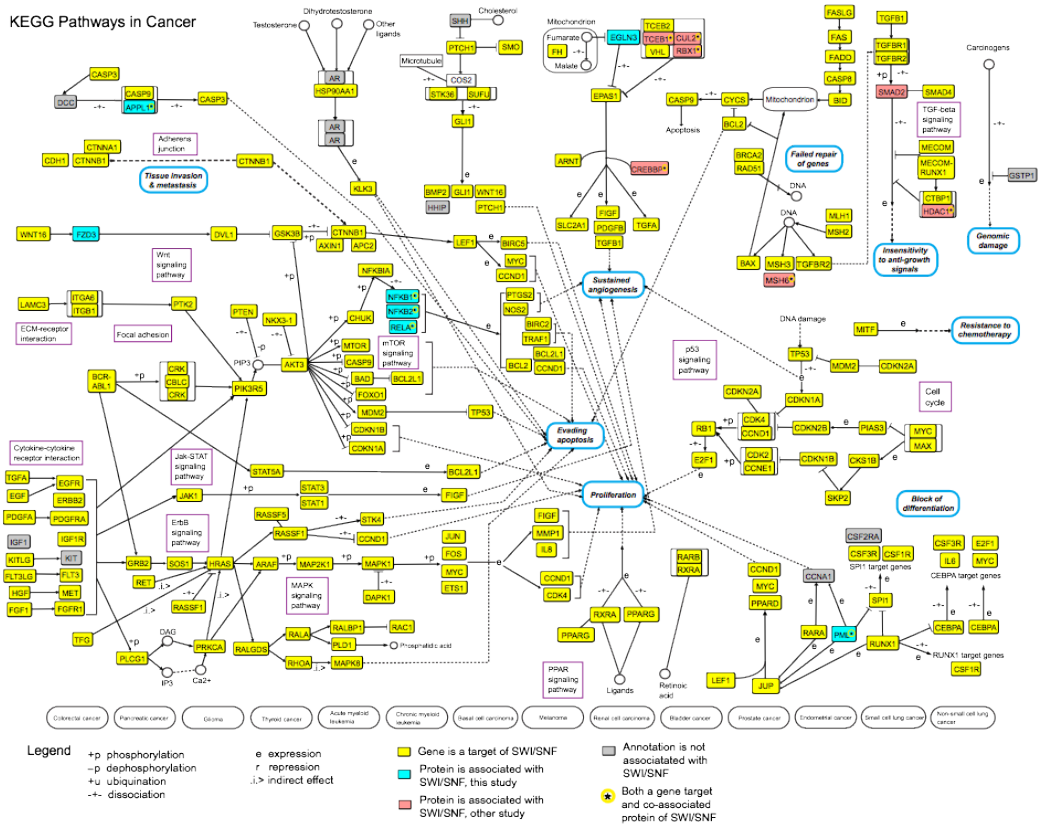


Figure B.2 SWI/SNF ChIP-Seq targets and interacting proteins superimposed on KEGG 'Pathways in Cancer'. The KEGG 'Pathways in Cancer' network was among those pathways overrepresented using our 49,555 SWI/SNF high-confidence union regions (Benjamini adjusted p-value < 4.7 x 10-8). SWI/SNF ChIP-Seq targets are highlighted in yellow and SWI/SNF co-purifying proteins detected in our IP-mass spectrometry experiments are highlighted in blue. SWI/SNF co-purifying proteins reported in other studies (Table S10) are highlighted in red. Proteins or genes not detected in any known SWI/SNF studies are gray. Starred annotations were detected in both ChIP-Seq and protein interaction studies.

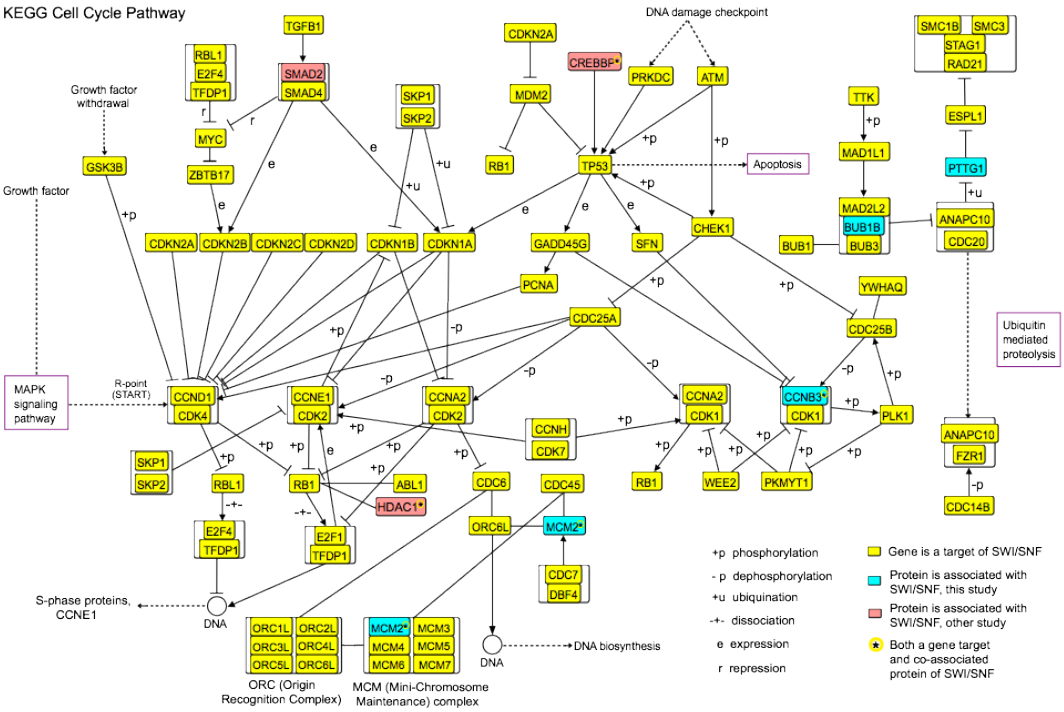


Figure B.3: SWI/SNF ChIP-Seq targets and interacting proteins superimposed on KEGG 'Cell Cycle'. The KEGG 'Cell Cycle' network was among those pathways overrepresented using the 49,555 SWI/SNF high-confidence union regions (Benjamini adjusted p-value < 3.7 x 10-8). SWI/SNF ChIP-Seq targets are highlighted in yellow and SWI/SNF co-purifying proteins detected in our IP-mass spectrometry experiments are highlighted in blue. SWI/SNF co-purifying proteins reported in other studies (Table S10) are highlighted in red. Starred annotations were detected in both ChIP-Seq and protein interaction studies.

B.2 Supporting Tables

Table B.1 Data sources of genomic features used.

Annotation	Number of regions	Platform	Source[a]
Ini1	49,458	ChIP-Seq	this study
Brg1	12,725	ChIP-Seq	this study
BAF155	46,412	ChIP-Seq	this study
BAF170	30,136	ChIP-Seq	this study
RNA Polymerase II	23,320	ChIP-Seq	Rozowsky et al. 2009
IgG control	N/A	ChIP-Seq	Auerbach et al. 2009
SWI/SNF union	69,658	ChIP-Seq	this study
SWI/SNF high-confidence union	49,555	ChIP-Seq	this study
SWI/SNF core (Ini1, BAF155 and BAF170)	9,760	ChIP-Seq	this study
Lamin A/C[b]	1,770	ChIP-chip	this study
Lamin B[b]	1,270	ChIP-chip	this study
H3K27me3	32,704	ChIP-Seq	Cuddapah et al. 2009
CTCF	19,308	ChIP-Seq	Cuddapah et al. 2009
Predicted enhancers[c]	36,562	ChIP-chip	Heintzman et al. 2009
RNA Polymerase III	478	ChIP-Seq	Oler et al. 2010; Barski et al. 2010
RNA-Seq	N/A	RNA-Seq	Morin et al. 2008
Non-canonical small RNAs[d]	48,403	RNA-Seq	Affymetrix ENCODE Transcriptome Project and Cold Spring Harbor Laboratory ENCODE Transcriptome Project (2009)
DNA replication origins [b,e]	282	Short nascent strand-chip	Cadoret et al. 2008

[a]All regions were identified in HeLa cells. Coordinates used are in hg18.
[b]Data are from the ENCODE pilot phase regions.
[c]Enhancers were predicted using a model trained on various chromatin signatures including those for p300/EP300, MED1, CTCF, DNase hypersensitivity sites and histones H3K4me1, H3K27ac and H3K4me3.
[d]Excludes annotated miRNAs, small nucleolar RNAs, repeats and predicted RNA genes.
[e]One of the 283 replication origins from Cadoret et al. was discarded in the process of converting hg17 coordinates to hg18 with the the UCSC 'liftOver' utility

Tables B.2 through B.19 Due to space constraints, only the data sources table is provided as part of t dissertation document. For additional supporting tables related to this work, such as those listing the genomic coordinates of all regions identified by this study, please see the supporting information in [Chapter 2, Reference 91].

Appendix C: Supporting Documentation for Chapter 4

Supplemental materials accompanying the work featured in this chapter can be found in its original form on the *Cell* website (http://www.sciencedirect.com/science/article/pii/S0092867411015170). Supplemental Figures are provided in this dissertation for convenience. In the interest of space and formatting constraints, Supplemental Methods text and Supplemental Tables are omitted from this document but can be found at the above link.

Materials prefaced with the appendix designation (e.g. "Figure C.1) can be found in this section. Materials lacking the appendix designation (e.g. Table S1) can be referenced online.

C.1 Supporting Figures

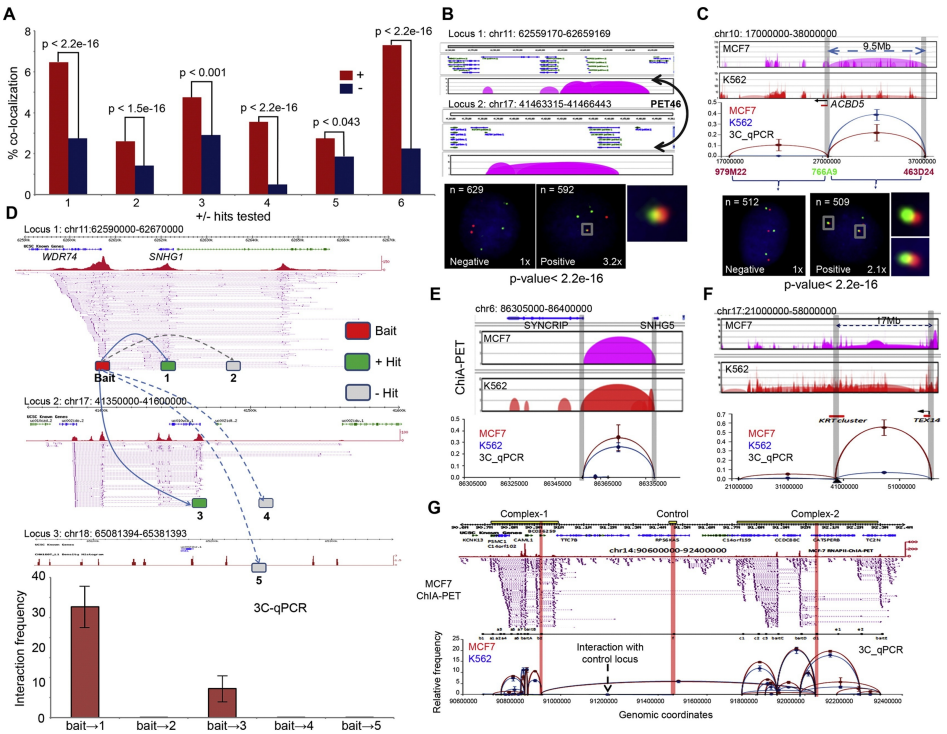


Figure C.1 Validation of Chromatin Interactions in MCF7 and K562 Cells by DNA-FISH and 3C-qPCR, Related to Figure 1.(A) Quantitative DNA FISH data for positive (interaction) and negative (no interaction) hits randomly selected from MCF7 interchromosomal ChIA-PET data. **(B)** An example of a chromatin interaction between chr11 and chr17 in MCF7 cells. This exemplifies that multigene complexes from different chromosomes could further converge to a common active nuclear compartment. **(C–G)** Detailed 3C-qPCR validations for several long-range (up to ~17 Mb) intrachromosomal interactions and an interchromosomal interaction (D). Most of the intrachromosomal interactions are tested in both MCF7 and K562 cell-lines. P values are calculated using binomial test. Panel D, F and G represent local interactions at distant genomic loci converging to each other via long range cis or trans interactions. 3C-qPCR mean values and standard error of means (SEM) from three independent experiments are shown. See also Table S2.

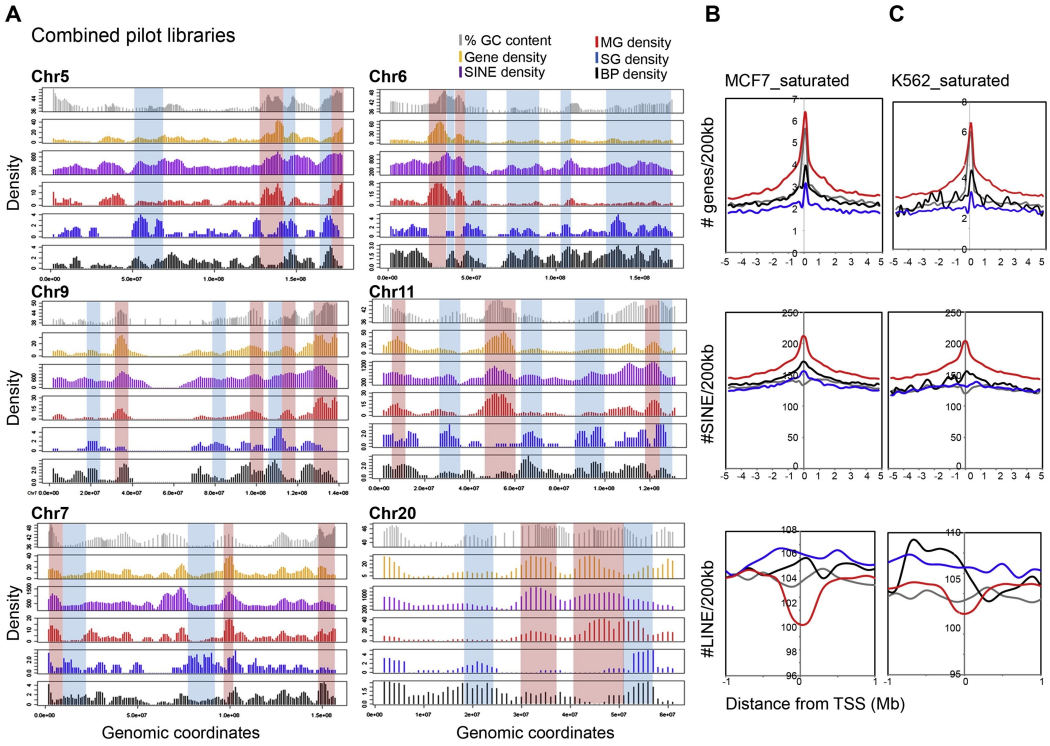


Figure C.2 Detailed Genomic Features of Distinct Chromatin Models, Related to Figure 4.2(A) Detailed examples from 6 different chromosomes illustrating the association of distinct chromatin architectures with genomic descriptors. Density of each descriptor (except %GC, which is measured in isochores) and the interacting anchors in each of our chromatin architectures is measured in each 1 Mb domain across chromosomes and running mean over 5 values are plotted. Certain gene rich domains enriched in multigene (MG) models and depleted in single-gene (SG) models are highlighted in red, while relatively gene-poor domains enriched in SG and depleted in MG are marked in blue. **(B and C)** Genomic features of BP, SG, MG models in MCF7 (B) and K562 (C) saturated libraries. The plots validate our observations on the combined pilot data presented in Figure 2 of the main text.

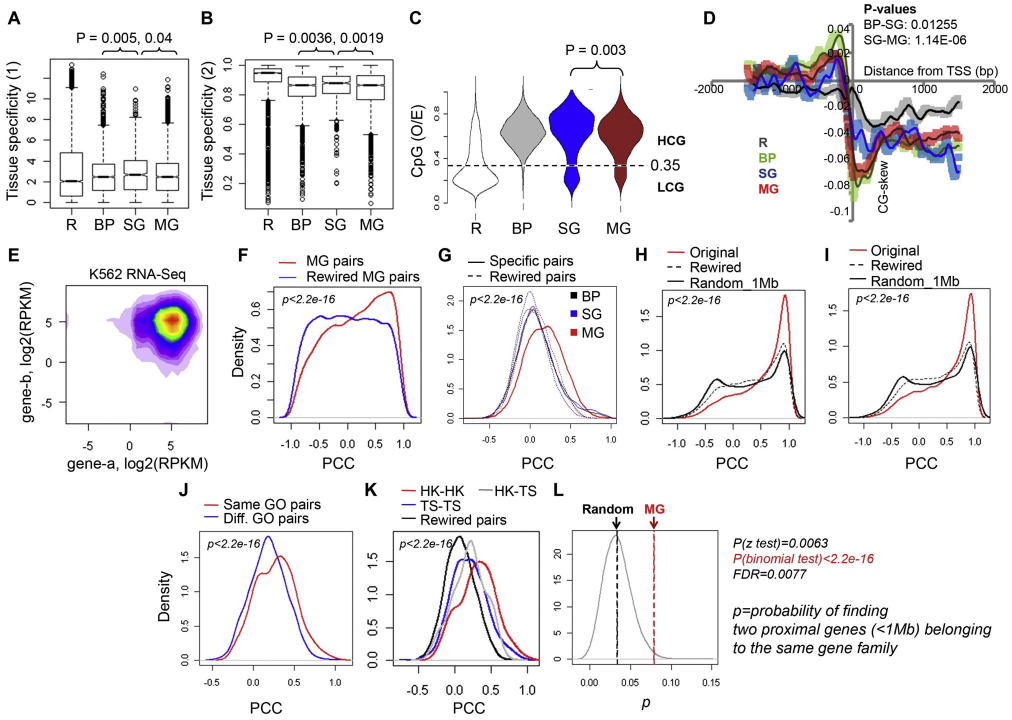


Figure C.3 Promoter Properties and Functional Output (Transcription) of Different Categories of Chromatin Models, Related to Figure 4.3(A and B) Tissue specificity measured by descriptor-1 (A) and descriptor-2 (B). Equations for tissue specificity descriptors are given in the Extended Experimental Procedures. **(C and D)** Normalized CpG content (C) of promoters (±1500 bp to TSS) and strand bias (CG-skew) (D) at promoters of genes in different models. Difference in the representation of High CpG (HCG) promoters (associated with housekeeping genes) and Low CpG (LCG) promoters (associated with tissue specific genes) is found to be significant between SG and MG complexes, while BP model has relatively negligible representation of LCG promoters suggesting their association primarily with housekeeping function. Similarly, CG-skew in (D) shows greater bias (associated with high and housekeeping expression) at promoter sites for BP and MG models, while lower bias (associated with lower and tissue specific expression) for SG model. These predictive measures support our observation in Figure 3D in the main text. **(E)** Coexpression of interacting genes in K562 cells. **(F–K)** Density plots for Pearson's Correlation Coefficient (PCC) values of gene pairs in MG complexes (red), rewired pairs and random gene pairs selected from a control dataset of the same distribution of genomic spans and gene density as MG pairs with an upper limit of 1 Mb. The gene expression datasets analyzed are: (F) E2 induced time course microarray at 6 time points (Fullwood et al., 2009); (G) microarray dataset of 4,787 human samples covering a wide range of diversity in gene expression, like distinct tissues, gender, developmental and differentiation stages etc. (Sahoo et al., 2008). Different controls are selected over genomic spans of BP, SG and MG genes; (H–I) ENCODE RNA-Seq datasets for 5 different cell-lines (K562, MCF7, HeLa, HCT116 and GM12878) for MCF7 and K562 interactions; (J) PCC distribution for MG gene pairs belonging to the same and different functions (GO process); (K) PCC distribution for housekeeping (HK) and tissue specific (TS) gene pairs in MG units. **(L)** Representation of gene families in random (#3383) and MG complexes (#1487) datasets with respect to expected probability "p" of finding 2 proximal genes (within 1 Mb proximity) from the same gene family. The method to compile random control and to calculate the probability of finding two proximal genes from the same gene family is given in the Extended Experimental Procedure. The plot suggests greater enrichment of gene families in multigene complexes. Also see Table S4.

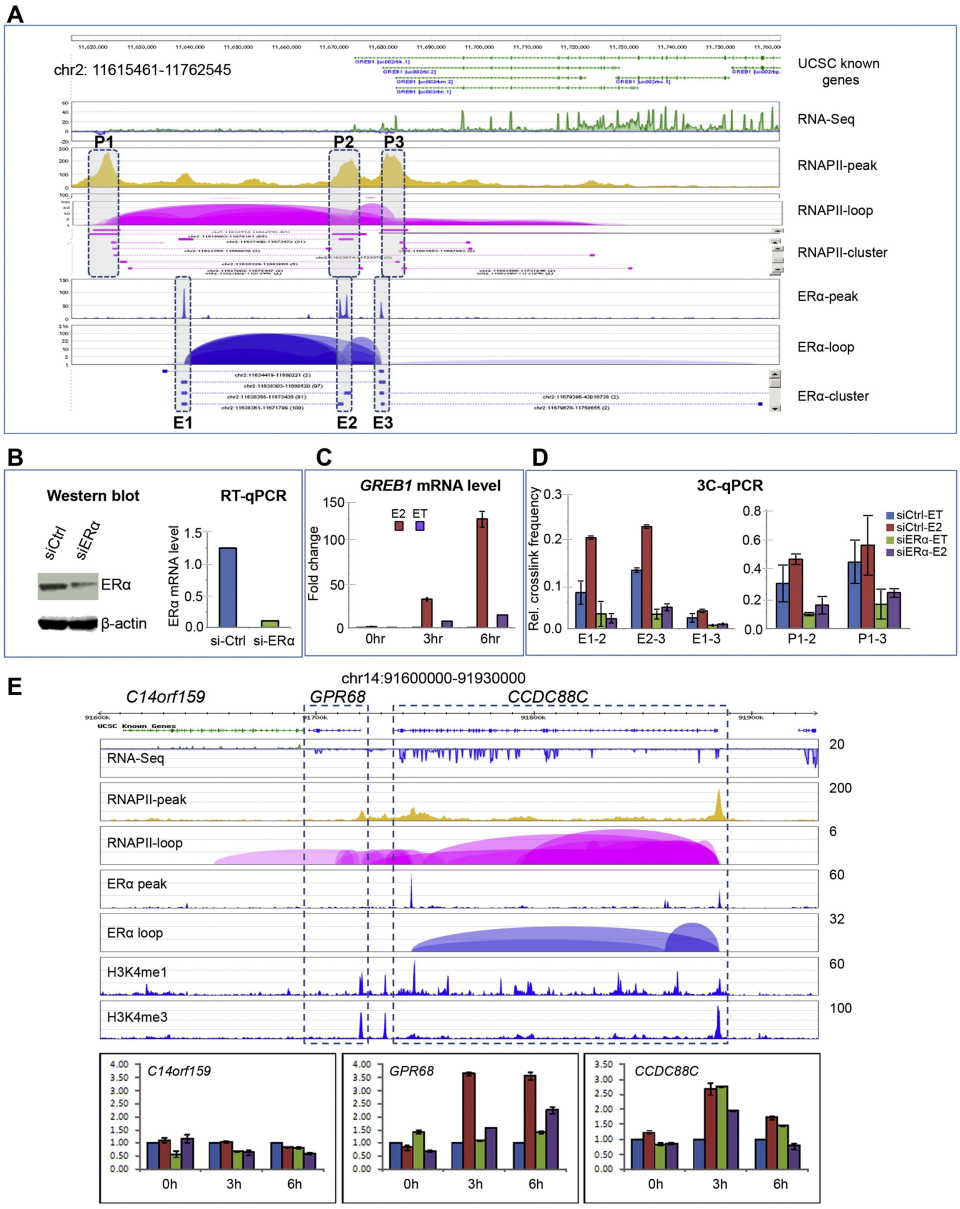


Figure C.4 Chromatin Interactions and Gene Expression Following siERα Transfection in MCF7 Cells, Related to Figure 4.4.(A) Overlap of RNAPII loops with ERα loops at the GREB1 locus. P1, P2, and P3 are RNAPII interacting sites; E1, E2, and E3 are ERα interacting sites. **(B)** ERα knockdown by siERα as tested by Western blot and RT-qPCR. **(C)** GREB1 expression following 0, 3 and 6 hr of ethanol (ET) and estrogen (E2) treatment after siControl and siERα transfections. RT-qPCR mean values and standard deviations (SD) from two independent experiments are shown. **(D)** 3C-qPCR data for chromatin interactions at GREB1 locus following ET and E2 treatment after siControl and siERα transfections. 3C-qPCR mean values and standard error of means (SEM) from three independent experiments are shown. **(E)** Estrogen induction and siERα knockdown led to correlated changes in the expression of interacting genes (CCDC88C and GPR68). ChIA-PET tracks clearly show that interaction between promoters of GPR68 and CCDC88Cis associated with RNAPII, while ERα binds only at promoter and gene-body of CCDC88C. Color codes of the bars are shown in Figure C.4D. RT-qPCR mean values and standard deviations (SD) from two independent experiments are shown.

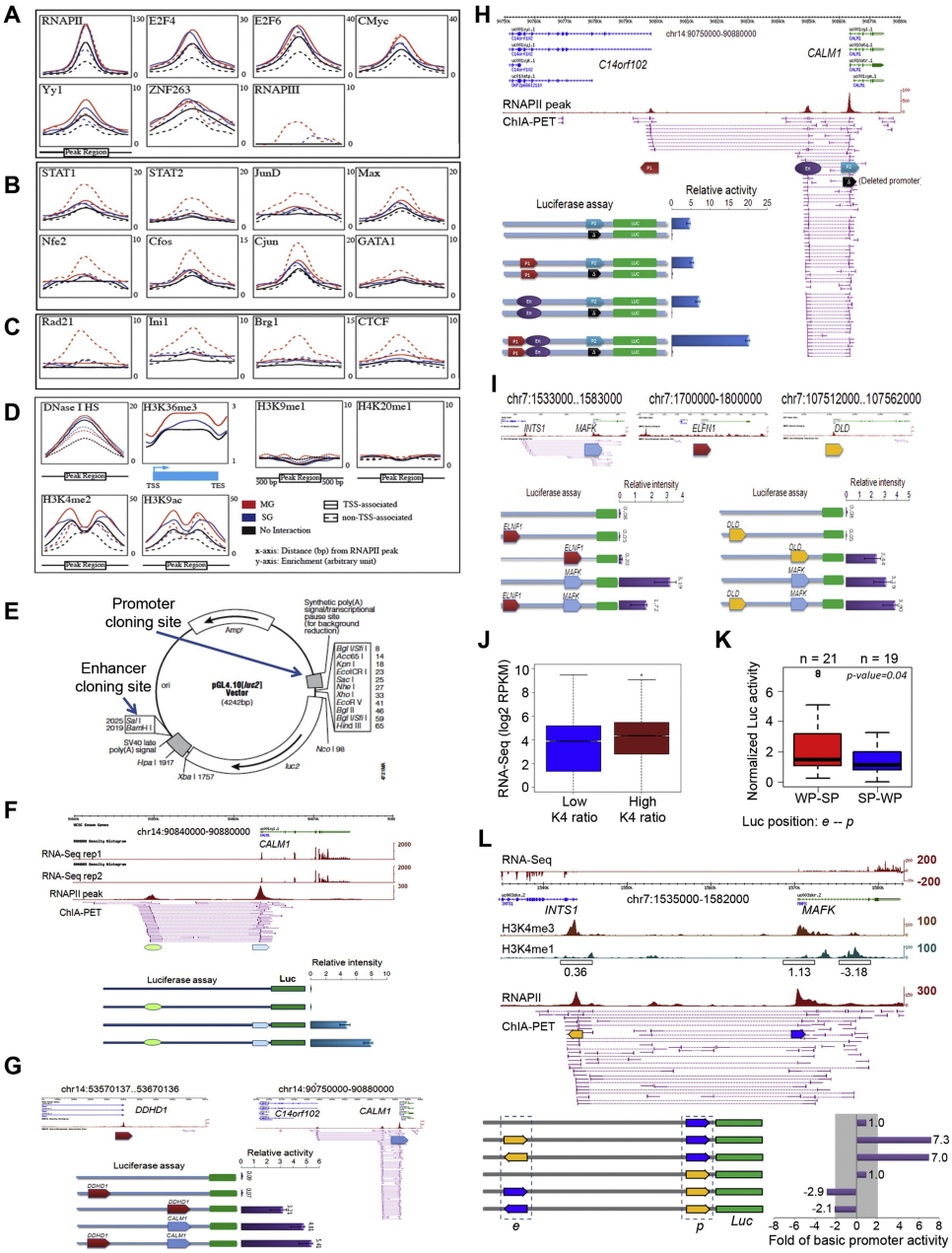


Figure C.5 Enrichment Profiles of Transcription Factors and Histone Modification Marks Centered at the Interaction Anchor Regions of RNAPII-Bound Chromatin Interaction Structures in K562 Cells, and Reporter Gene Assays in MCF7 Cells, Related to Figure 4.5.(A) Aggregation plots of TFs enrichments centered at the RNAPII interaction sites, proximal to TSS (TSS) or distal to TSS (non-TSS). RNAPIII, as a negative control, shows negligible enrichment at the RNAPII interacting sites. y axis: sliding median for ChIP-Seq enrichment in the region. x axis: distance (bp) from RNAPII sites. **(B)** TFs enriched at non-TSS (potential enhancer sites). **(C)** Enrichment profile of chromatin remodeling and chromatin architectural factors. **(D)** Enrichment profile of open chromatin and histone marks around RNAPII interacting sites. Clearly, the open chromatin mark DHS and active histone marks are substantially enriched at the RNAPII interacting sites, while the repressive histone marks show little enrichment. **(E)** Map of pGL4.10 vector and cloning sites of promoters and enhancers for luciferase assays. **(F)** Standard promoter and enhancer reporter assay for elements around the CALM1 locus. The enhancer upstream of CALM1 significantly, but modestly, enhanced the luciferase activity of the CALM1 promoter, which was involved in the typical enhancer-promoter interaction. **(G)** The DDHD1 promoter, which is located on the same chromosome as C14orf102-CALM1 locus and had no interaction with CALM1, showed no significant

enhancement to the CALM1 promoter activity in luciferase assays. **(H)** Deletion promoter reporter assay around CALM1 locus. The TATA box of the CALM1 promoter was deleted (from – 133 bp to +100 bp, black arrow). The reporter construct containing this deletion promoter did not show any promoter activity in luciferase assays by itself or in any combinations with C14orf102. **(I)** Non-MG promoters do not possess enhancer functions. The ELFN1 and DLD promoters, which are located on the same chromosome as for the INTS1-MAFK locus and had no interaction with MAFK, did not enhance the promoter activity of MAFK in luciferase assays. **(J)** Box plots of RNA-Seq data in log2 RPKM for the genes with low and high log ratio of H3K4me3/H3K4me1 in the pairs of interaction sites. The genes with higher log ratio in a pairing relationship have higher RNA-Seq counts on average than the interacting partner with lower log ratio. **(K)** Box plots of normalized luciferase activities when the promoters have a low log ratio of H3K4me3/me1 at the enhancer position of the luciferase constructs, or when the promoters have a high log ratio of H3K4me3/me1 at the enhancer position of the luciferase constructs. The promoters with low log ratio of H3K4me3/me1 at the enhancer position of the luciferase constructs have higher enhancing effects in general. **(L)** Swap of INTS1 and MAFK promoters in positions in reporter gene construct for luciferase assays. The promoter sequence from INTS1 (with lower log ratio of H3K4me3/me1 signals) enhanced the luciferase activity of MAFK (with higher log ratio of H3K4me3/me1 signals). On the reverse, the MAFK promoter showed no enhancer function. The mean values and standard deviations (SD) of luciferase activities from at least three independent experiments are shown.

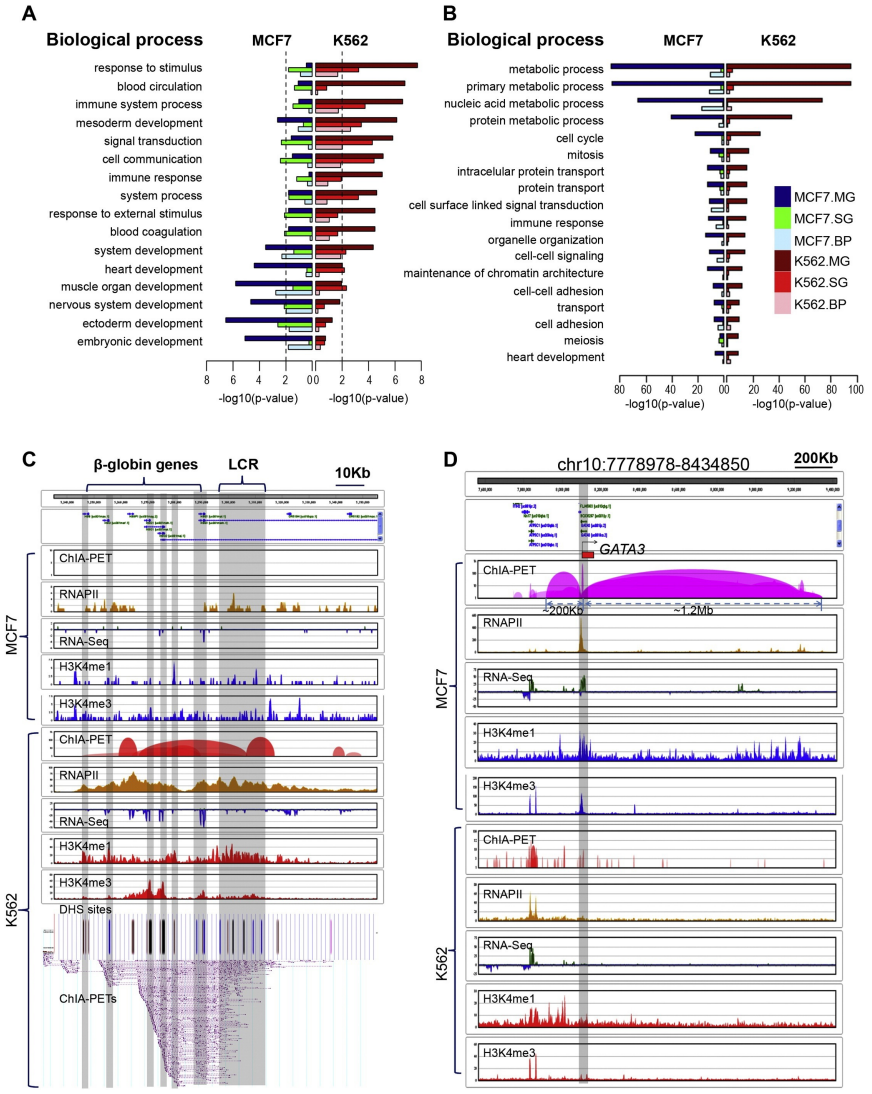


Figure C.6 Cell-Specific Interaction Analysis, Related to Figure 4.6(A) All the Gene Ontology (GO) terms over-represented in the gene sets engaging cell-specific expression and interactions. **(B)** Overrepresented GO terms in the gene set engaged in chromatin interactions common to both MCF7 and K562 cells. The abundance of housekeeping terms is apparent. **(C)** K562-specific interactions around β-globin gene locus on chromosome 11. ChIA-PET loop tracks clearly show that there are chromatin interactions from β-globin genes to the locus control region (LCR) in K562 cells, but not in MCF7 cells. Correspondingly, the RNAPII and RNA-Seq show higher expression of β-globin genes in K562 cells, but not in MCF7 cells. **(D)** MCF7-specific interactions around GATA3 gene locus on chromosome 10. In contrast to the β-globin gene locus, the ChIA-PET loop tracks clearly show that there are chromatin interactions from GATA3gene locus to multiple enhancer sites in MCF7, but not in K562 cells. Correspondingly, the RNAPII binding and RNA-Seq showed high activity in MCF7 and low activity in K562 cells. Especially, one super-long-distance enhancer is about 1.2 Mb away from GATA3 promoter.

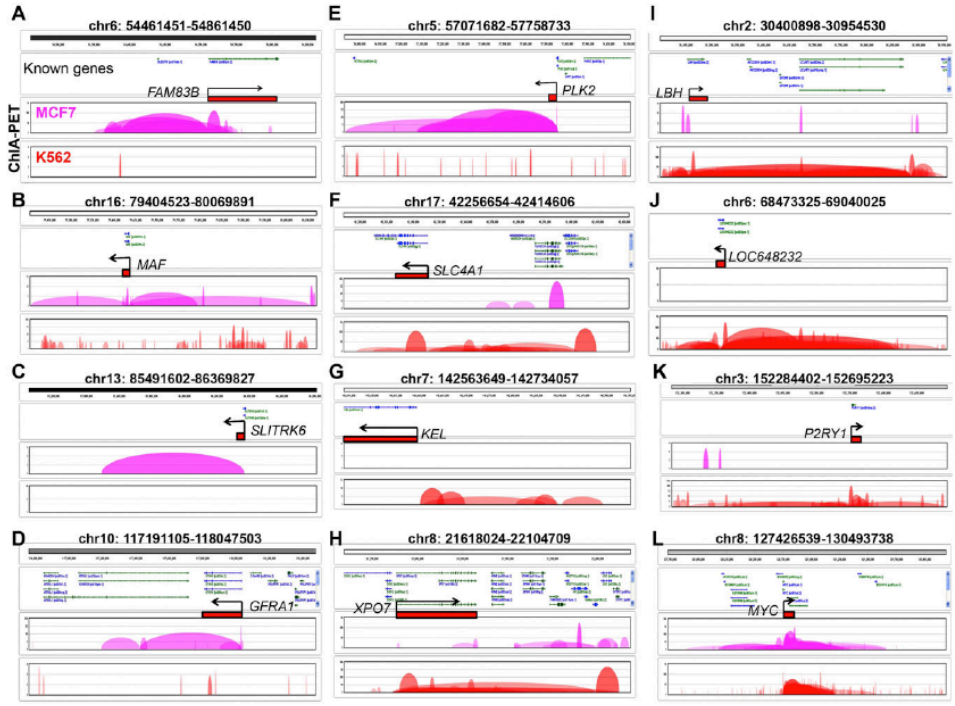


Figure C.7 Examples of Cell-Type-Specific Long-Range Enhancer-Promoter Interactions, Related to Figure 4.6 and Figure 4.7.(A–E) are specific to MCF7 cells, and **(F-K)** are specific to K562 cells. Most of the regulatory sites in these long distance interaction examples are bypassing the nearest promoters and linking to other gene promoters. **(L)** Several distant enhancers converging to MYC gene promoter. Cell-specific alternative usage of certain enhancers can be seen from the interaction loop views from MCF7 and K562 cells.

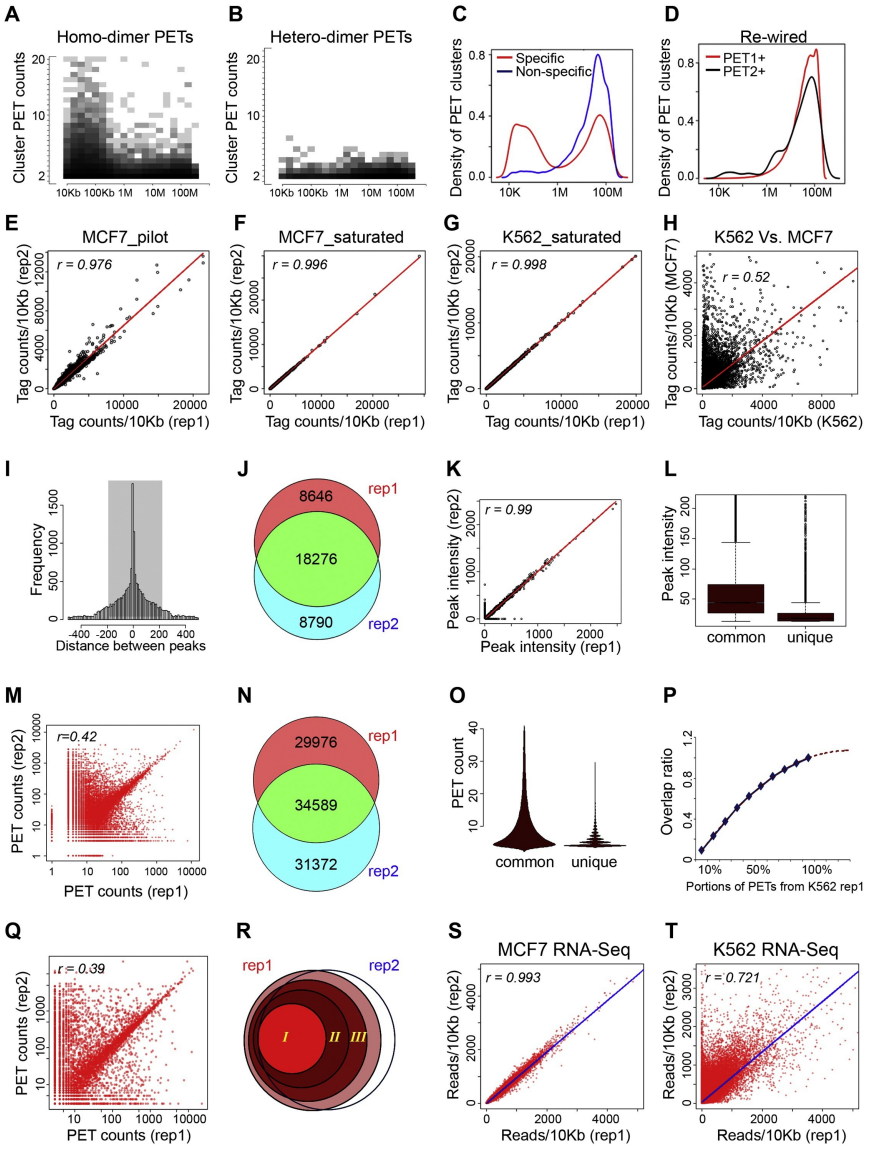


Figure C.8 Assessment of Technical Noise, Library Reproducibility and Saturation Analysis, Related to Experimental Procedures.(A and B) Heatmaps of PET sequence counts versus genomic span for interactions identified from homo-dimer and hetero-dimer PETs from the combined pilot dataset. **(C)** Densities of genomic spans of interactions from homo-dimer PETs and hetero-dimer PETs of the combined pilot data. **(D)** Densities of genomic spans of interactions from rewired PETs. **(E–G)** Scatter plot of sequence reads per 10 kb from RNAPII ChIA-PET replicates: (E) MCF7 pilot datasets, (F) MCF7 saturated and (G) K562 saturated datasets. **(H)** Scatter plot of sequence reads per 10 kb from K562 saturated and MCF7 saturated RNAPII ChIA-PET datasets. **(I–L)** RNAPII binding site reproducibility of K562 saturated replicates. (I) Histogram of genomic distances between RNAPII peaks from replicates. (J) Venn diagram of RNAPII peak overlap between replicates. (K) Scatter plot of RNAPII peak intensities of replicates. (L) Box plot of peak intensities of RNAPII peaks common and unique in replicate 1. (M–O) RNAPII interaction reproducibility of K562 saturated replicates. **(M)** Scatter plot of interaction PET counts between replicates. **(N)** Venn diagram of interaction overlaps between replicates. **(O)** Violin plot of interaction PET counts from common and unique interactions from replicate 1. **(P)** Saturation assessment of chromatin interactions from K562 saturated RNAPII ChIA-PET replicates. The overlap ratio between replicates against the proportion of PETs sampled from K562 saturated replicate 1 (more details in the Extended Experimental Procedure; under saturation analysis). **(Q and R)** RNAPII interaction region reproducibility of K562 saturated replicates. (Q) Scatter plot of interaction region PET counts between replicates. (R) Venn diagram of interaction region overlaps from replicates. I, II, and III for the top 25%, 50% and 75% interaction

regions from K562 saturated replicate 1. **(S and T)** Scatter plots of RNA-Seq reads per 10 kb in replicates from MCF7 (S) and K562 (T).

Appendix D: Supporting Documentation for Chapter 5

D.1 Link to Supporting Documents

Supporting documents including sample input and output files, UML class and activity diagrams for CAPE, and the Users' Guide can be found at http://cape.gersteinlab.org. Some files are reproduced here for convenience.

D.2 Documents Included in this appendix

This appendix includes the CAPE Users Guide, Use Cases, and UML diagrams. Updated versions are always available at the CAPE website.

Coupled Analysis of Polymerase Binding and Expression (CAPE) Users Guide

Version 1.0

Raymond Auerbach, Arif Harmanci, Joel Rozowsky, and Mark Gerstein

Gerstein Laboratory

Yale University

Note: The latest version of the Users Guide can always be found at http://cape.gersteinlab.org.

Introduction

CAPE Is designed with two primary goals in mind. The first is to associate transcription factor binding site data from next-generation sequencing experiments such as ChIP-Seq with a set of genomic features such as transcription start sites, transcription end sites, etc. The second aim is to classify transcripts based on levels of transcription factor binding at the promoter vs. the expression level of the transcript. This tool is designed to work with many of the formats employed by large consortia such as NHGRI's ENCODE and modENCODE consortia.

Program Requirements

CAPE is written in Java and requires Java version 1.6 or higher. The Java runtime environment (JRE) must be installed. If the JRE is not installed (i.e. typing "java –version" at the command line does not produce a version number or the version number is < 1.6), an updated version of the Java Runtime Environment can be downloaded from http://www.oracle.com/technetwork/java/index.html. Please see that website for instructions.

Obtaining the program

The CAPE programs are distributed as self-contained, executable java archives (a .jar file) and can be downloaded from http://cape.gersteinlab.org. If you wish to download the source code and associated first-party libraries for this tool, they can also be found at the same address. CAPE is distributed as-is as open-source software.

CAPE-analyze

Generating Transcription and Binding Reports using CAPE-analyze

Once the JRE is installed and the jar file is downloaded, the program can be run from a terminal window. From the command prompt, type "java –jar CAPE-analyze.jar <arguments>". See the section "Command line arguments below for a detailed description of possible arguments. Long-form arguments are given in the format "--longFormOptionName=value" and short-form arguments are given in the format "-shortFormOptionName value". For larger data sets, we recommend increasing the

maximum amount of memory allocated to the Java heap to improve system performance. Depending upon your hardware configuration and the number of transcripts to be analyzed, the java heap size may be increased using Java's "-Xmx" option (e.g. "-Xmx512M" will set the heap size to use 512 MB of RAM, "-Xmx1G" will use 1 GB of RAM, etc.). This option should appear between "java" and "-jar" in the example command given above.

CAPE-analyze Command-Line Arguments

Program Mode Arguments

One of the following arguments may be specified to select the mode of operation. If no argument is specified, CAPE will run in transcript mode by default.

Long-Form	Short-Form	Description
--binding	-b	Run in binding mode
--transcript	-t	Run in transcript mode (default)

If both arguments are specified, CAPE will exit with an error.

Binding Mode Arguments

The following arguments are required when running in binding mode:

Long-Form	Short-Form	Description
--signalfile	-s	Signal file in bigWig or bigBed format. Extension must be .bw, .bigWig, .bb, or .bigBed
--transcriptfile	-tf	File with transcript information. Must be GTF format
--peakfile	-p	Peak file. Extension must be .bed or .narrowPeak
--outputfile	-o	Filename to use for output

Additionally, the following optional parameters may be specified:

Long-Form	Short-Form	Description
--upstreampad	-up	The upstream overlap window in bp to use for feature association. Default=1000 bp.
--downstreampad	-dp	The downstream overlap window in bp to use for feature association. Default=1000 bp.

Transcript mode arguments

The following arguments are required when running in transcript mode:

Long-Form	Short-Form	Description
--signalfile	-s	Signal file in bigWig or bigBed format. Extension must be .bw, .bigWig, .bb, or .bigBed
--transcriptfile	-tf	File with transcript information. Must be in GTF format
--outputfile	-o	Filename to use for output

Additionally, the following optional parameters may be specified:

Long-Form	Short-Form	Description
--aggpad	-ap	The initial window size in bp on each side of a feature to use for the aggregation step. Default=1000 bp (note: this results in +/- 1000 bp from start)
--aggoverride	-ao	In transcript mode, the pad value to use on each side of a feature for aggregation. Value must be >= 50. Skips window size detection if set.
--rseqtoolsfile	-r	File created by the rSeqTools program mrfQuantifier containing expression values for all transcripts (the transcript ID appears in column 1 and the RPKM appears in column 2)
--cufflinksfile	-c	File created by cufflinks containing expression values for all transcripts (the transcript id appears in column 2 and the FPKM appears in column 10)
--GFFkey	-k	The feature type to use from the third column of a valid GFF or GTF file. Defaults to "transcript". Change if you are analyzing genes, etc.
--percentileoverride	-po	A comma-separated list of percentile cutoffs for low- and high-value binding and expression. Default=25,75

--expressionoverride	-eso	A comma separated list of expression value cutoffs for low- and high-value expression. Given in real units (RPKM, FPKM, etc). Masks the --percentileoverride option for expression.
--bindingoverride	-bso	A comma separated list of expression value cutoffs for low- and high-value expression. Given in real units (RPKM, FPKM, etc). Masks the -percentileoverride option for binding.

Description of Algorithm - Binding Mode

Given a transcript file, a signal file, and a peaks file, binding mode will identify the position within each peak where the signal is the highest. Using this summit position to represent the peak region, overlap analysis is then run against the transcript list and the closest TSS and TTS is reported as well as the distance to each. Distances are reported in a strand-specific manner where negative values correspond to the summit position being upstream of a feature. Conversely, a positive distance value indicates that the summit occurs downstream of a feature. Each peak will be assigned an association of “TSS”, “TTS”, “Neither” or “Ambiguous” based on the distance to each feature. The values of --upstreampad and --downstream pad are used to assign a peak to a category. For example, the default value of 1000 bp means that a peak that falls within +/- 1000 bp of a TSS will be assigned a TSS association. If a peak falls halfway between a TSS and a TTS and both distances fall within the cutoff, a value of “ambiguous” is reported. Peaks falling outside the pad range will be reported as “Neither.” Note that for peak files in narrowPeak format, the summit position given in the file for each peak is used as the summit position. For files in bed format, the maximum signal in the region is determined and if this region spans multiple base pairs, the midpoint of the range is reported as the summit value.

In the event that there is more than one maximum value, the peak is reported as "multimodal" and an association is not calculated. Please note that binding mode is still under active development at this time and that the crux of CAPE-analyze is transcript mode.

Description of Algorithm - Transcript Mode

Given a transcript file, a signal file, and a file with expression values, the tool begins by aggregating the TF binding signal around TSSs given the window size of +/- the --aggPad argument (default: 1000 bp). The results from this aggregation are used to determine an ideal size for the promoter region as follows: for the aggregation values, the global maximum and minimum positions are determined. From the position of the global maximum, the closest positions where the value crosses below the value of (global minimum + ((global maximum - global minimum) * .10) are determined in both the upstream and downstream directions and the largest values used as the pad. For example, if the above criteria are met at -500 and +300 bp from the global maximum, then an "ideal" pad size of +/- 500 bp from the TSS is used. Using the ideal pad size to represent the promoter region, the average signal level in the promoter and the gene body are determined for each transcript and the ratio between these values is reported. Note that currently, there is no size restriction on transcripts. If a transcript is smaller than the ideal pad for a promoter, no signal over the transcript body is calculated and the ratio will be "N/A". For gene expression, values are read directly from an expression file (typically in RPKM or FPKM) or can be read from the annotation file. Please see the description of input files below for more details.

Once both binding values and expression values are determined for each transcript, low- and high-value cutoffs are determined. This process occurs separately for binding and expression but the steps are the same. We will discuss in terms of binding in this example. Using only the binding signal values that are non-zero, we determine the values that correspond to the percentile cutoffs specified by --percentileoverride (default= 25th and 75th percentile). All transcripts with a binding value below the lower cutoff value will be deemed as "low binding," any transcript with a binding value above the upper cutoff will be deemed "high binding." Transcripts with binding values in the intermediate range are deemed "normal." Alternatively, we expect there will be some cases where a researcher will want to use actual data values instead of percentiles to set the lower and upper cutoffs. These values can be set with the --expressionoverride and --bindingoverride arguments (e.g. --bindingoverride=20,80 will set the low and high binding cutoffs to be below the 20th percentile and above the 80th percentile, respectively. These arguments will override the percentile cutoffs for their respective data types. After all transcripts are categorized, a list of transcripts that fall into each category is written to the file named in the --outputfile argument.

Input File Formats

CAPE-analyze supports a variety of different input formats for peak files, signal files, and expression files.

Peak Files

Peak files are used in binding mode and can be provided in three or four column, tab-delimited bed format. The first column corresponds to the chromosome, the second column gives the start position, the third column gives the end position, and the optional fourth column is a single character denoting the strand ("+" or "-"). If strand is omitted, all features will be considered to be on the forward strand. Please note that bed files must end in the extension ".bed". For more information about bed format, please see http://genome.ucsc.edu/FAQ/FAQformat#format1. For example:

```
chr1    500    1000    +
chr2    100    300     -
```

Alternatively, peak files can also be provided in ENCODE narrowPeak format to allow for direct download from the ENCODE data repository at UCSC. For a description of the ENCODE narrowPeak format, please see http://genome.ucsc.edu/FAQ/FAQformat#format12. ENCODE narrowPeak files must end in the extension ".narrowPeak".

Signal Files

Signal files must be provided in bigWig or bigBed formats. These files are indexed, allow for random access, and reduce the storage requirements for large signal files such as those produced from whole-genome analyses in human cells. Other formats such as bedGraph and wiggle can be easily converted to bigBed or bigWig formats using Jim Kent's toolkit at UCSC (binaries available at http://hgdownload.cse.ucsc.edu/admin/exe/. Please see the documentation in the source distribution of Jim Kent's utilities for more information. The relevant programs are bedGraphToBigWig, wigToBigWig, and bedToBigBed. Alternatively, Galaxy may also

be used to convert other signal file formats to bigWig and bigBed (http://galaxy.psu.edu/). For more information about the bigWig and bigBed formats, please see http://genome.ucsc.edu/FAQ/FAQformat#format1.5 and http://genome.ucsc.edu/FAQ/FAQformat#format6.1.

Transcript Files

Transcript files are accepted in either GTF (preferred) or GFF3 files. A full description of these files formats can be found at http://genome.ucsc.edu/FAQ/FAQformat#format4 (GTF) and http://genome.ucsc.edu/FAQ/FAQformat#format3 (GFF3). An example in GTF format is given below:

```
II	modENCODE_TX	gene	8651057	8658766	.	+	.	RPKM "109.609215"; gene_id "pyr-1"
II	modENCODE_TX	gene	5399522	5405988	.	+	.	RPKM "94.070177"; gene_id "mog-5"
II	modENCODE_TX	gene	13670567	13694711	.	-	RPKM "61.110324"; gene_id "Y48E1A.1"
```

To illustrate some of the features of CAPE-analyze, let's use the above snippet as an example. This GTF file targets genes instead of transcripts. The default behavior of CAPE-analyze is to look for transcripts, but this can be changed by setting the "--GFFkey=gene" parameter at the command line. This will tell CAPE-analyze to look for genes and to key off the gene_id value given in the ninth column.

In both GFF3 and GTF, the ninth column is a "catch-all" column where the id and, occasionally, gene expression data will appear. Running in default mode, CAPE-analyze will look for records with "transcript" in the third column and "transcript_id" in the ninth

column. If the "--GFFkey=gene" option is set, CAPE-analyze will instead analyze records with "gene" in the third column and "gene_id" in the ninth column.

Also shown above is the expression value given in the ninth column ("RPKM"). CAPE-analyze will use this field if no other expression files are provided. CAPE-analyze is designed to look for RPKM or FPKM tags in the ninth column. Multiple instances of these tags can exist (e.g. RPKM1 and RPKM2 in the case of multiple replicates being described in the same file). In these cases, all values will be averaged to calculate the overall expression value to be used for the transcript.

Expression Files

CAPE-analyze accepts expression values in any of three possible formats:

- rSeqTools - A tab-delimited file where the transcript ID occurs in column 1 and the expression value occurs in column 2.
- cufflinks - A tab-delimited file where the transcript ID occurs in column 2 and the expression value occurs in column 10. Note that users of Cufflinks 2.0+ should use the GTF option below if a final GTF file is produced.
- GTF - a GTF or GFF file with one or more of the following tags present in the final field for each transcript: RPKM or FPKM. In cases where tags are present for multiple replicates (ex: RPKM1 and RPKM2), the average expression value will be used.

For rSeqTools and cufflinks files, the file should be specified on the command-line using the "--expressionfile" option. For GTF format files, one must only specify the GTF file using the "--signalfile" option (the "--expressionfile" option should be omitted). NOTE: Using a cufflinks file or an rSeqTools file will override any expression values given in the GTF file. In transcript mode, the program will exit with an error if expression

values are omitted completely. Binding mode will still function, but expression values will be omitted from the final reports.

CAPE-analyze Output File

CAPE-analyze will produce a tab-delimited text file that can be easily viewed in Excel or a text editor of your choice. A snippet of a CAPE-analyze report for *C. elegans* follows:

#Transcript File: /emb_study/worm/total-RNA_20_degree_celsius_N2_Early_Embryos_RNA-seq.gtf
#Signal File: /emb_study/worm/2435_Snyder_N2_POLII_eemb_combined.bw
#Mode: transcript
#Promoter Binding Low Percentile Threshold: 25.0 (Binding <= 12.771
#Promoter Binding High Percentile Threshold: 75.0 (Binding >= 27.305
#Expression Low Percentile Threshold: 25.0 (Expression <= 43.015
#Expression High Percentile Threshold: 75.0 (Expression >= 80.253

#A. Transcripts with low promoter binding and high expression: 32
#B. Transcripts with high promoter binding and low expression: 45
#C. Transcripts with no promoter binding and no expression: 0
#D. Transcripts with no promoter binding: 6
#E. Transcripts with no expression: 0
#F. "Normal" transcripts: 937

#Category	Transcript ID		Chromosome		Start	End	Strand	PromoterSignal	BodySignal
	Ratio (Stalling Index)		Expression Value						
A	ZK858.6	I	9138260	9145625	-	8.889	24.151	0.368	88.639
A	rpt-5	I	5741868	5743576	-	12.163	22.264	0.546	106.276
A	sup-17	I	8792767	8797287	+	12.737	19.403	0.656	90.673
A	ngp-1	I	8394722	8398622	+	12.222	17.766	0.688	86.28
…									
…									
…									

The top of the report shows the files used for analysis as well as the parameters and cutoffs used by CAPE-analyze to categorize transcripts (or in this particular analysis, genes). This is followed by a breakdown of the six classifications used by CAPE-analyze as well as summary counts. The table then shows the values calculated for each gene or transcript as well as identifying characteristics such as chromosome, position, and strand. Raw values for promoter and body signals are also provided as well as the expression value from the expression file. The stalling index is the ratio of signal in the promoter vs

the gene body. For genes that are smaller than the average window size being used for aggregation, the bodySignal value will be 0 and the Stalling Index will be “NA”.

CAPE-compare

Comparing Results Using CAPE-compare

Often, researchers will want to compare transcripts from different samples or organisms to identify changes between ortholog groups or compare multiple samples from the same organism (e.g. healthy vs diseased cells, replicates, etc). The CAPE-compare tool takes two or more reports generated using the transcript mode of CAPE-analyze as well as an optional tab-delimited ortholog file and generates a tabular file combining the results into a single file. This file can easily be viewed in Excel or any text editor. Additionally, if two, three, or four reports are specified as input, an HTML file summarizing the overlaps within each transcription/binding category is also produced in addition to a ready-to-run R script that will produce Venn diagrams comparing samples across each category and save them as TIFF files. Please note that the free “VennDiagram” R package must be installed within R to use the script produced by CAPE-compare (see Acknowledgement of External Libraries section). We envision CAPE-compare to be useful in two different scenarios: comparing between organisms and comparing within an organism.

Comparing Results Between Organisms

For comparing reports generated between different organisms or annotation sources (e.g. worm, fly, and human), an ortholog file must also be provided to CAPE-

compare and specified on the command line with the "--orthologs" option. The ortholog file is a tab-delimited text file with the number of columns equal to the number of organisms being compared where each column contains the gene or transcript IDs used in the report Each row represents an ortholog set. For example:

```
humanOrtholog1    flyOrtholog1    wormOrtholog1
humanOrtholog2    flyOrtholog2    wormOrtholog2
humanOrtholog3    flyOrtholog3    wormOrtholog3
...
...
...
etc...
```

Only transcripts for which a complete list of orthologs exists should be provided (e.g. if three samples are being used, a value must exist in all three columns in each row).

When comparing different organisms, the following should be noted:

1. The IDs in the ortholog file must exactly match those in the report file.
2. The order of the filenames given to the --input option must match the order of the columns in the ortholog file (e.g. for the above example, the --input option must be "--input=humanReport.txt,flyReport.txt,wormReport.txt").
3. Only transcripts specified in the orthologs file will be written to the output files and considered in the output files.
4. When comparing two, three, or four CAPE-analyze reports, the HTML files and R script for Venn diagrams will be produced. In these files, only transcripts for which data exists for all samples being compared will be used. In the raw data output from CAPE-analyze, all transcripts will be listed with those missing data being classified as "NA".

5. Venn diagrams will only be produced for categories containing at least one data point.
6.

Comparing Reports from the Same Organism

Researchers may also wish to compare CAPE reports from the same organism generated under different experimental conditions, as different experimental replicates, etc. In this case, an ortholog file is not necessary and CAPE-analyze will assume that the reports share the same set of transcript IDs (for cases where transcript IDs may differ, such as when using different annotation versions, use the "Comparing Reports Between Organisms" workflow described above). Alternatively, an ortholog file can also be specified when a researcher wishes to limit the subset of transcripts being examined. In this case, the "Comparing Reports Between Organisms" workflow should be used. A tab-delimited text file or raw data generated from combining the reports will be generated. An HTML summary file and an R script to generate associated Venn diagrams will also be generated when comparing two, three, or four CAPE-analyze reports. These files are as described above.

Running CAPE-compare

CAPE-compare is run in the same manner as CAPE-analyze. From the command-line use: java -jar CAPE-compare.jar <argument list>

CAPE-compare Command Line Arguments

The following arguments are required for CAPE-compare:

Long-Form	Short-Form	Description
--input	-i	Comma-delimited list of CAPE reports to compare
--prefix	-p	the prefix to use for output files

Additionally, the following arguments may also be specified:

Long-Form	Short-Form	Description
--labels	-l	comma-delimited list of labels to use in the reports. Order must correspond to the order given for the --input option. File names will be used as labels if not specified
--orthologs	-o	the ortholog file to use

CAPE-compare File Formats

Input Files

Input files are in the output format produced by CAPE-analyze and should be specified as a comma-delimited list at the command line using the input option. For example, "--input=file1.txt,file2.txt".

Labels

We recommend giving each file a descriptive label, as it will make the reports easier to read. These labels should be provided in a comma-delimited list with the "--labels" parameter. For example: "--labels=worm,fly". Please note that the order of the labels must coincide with the order of the input files passed to the --input argument.

Output Files

All output files will contain the prefix specified by the "--prefix" argument. For example, for "--prefix=testOutput" with three input files, CAPE will produce the output files testOutput.txt, testOutput.html, and testOutput.r.

Ortholog File

The ortholog file is described in the section "Comparing Results Between Organisms" above and is optional.

Raw Data Output

Regardless of the number of CAPE-analyze input reports to compare, the raw data text file will be produced. This file is tab-delimited and an example follows:

```
#A. Transcripts with low promoter binding and high expression
#B. Transcripts with high promoter binding and low expression
#C. Transcripts with no promoter binding and no expression
#D. Transcripts with no promoter binding
#E. Transcripts with no expression
#F. "Normal" transcripts

#Worm Feature    Fly Feature        Human Feature      Worm State    Fly State  Human State
dhc-1    FBgn0261797    ENSG00000197102    F    F    NA
prp-8    FBgn0033688    ENSG00000174231    F    F    NA
ama-1    FBgn0003277    ENSG00000181222    F    F    NA
sma-1    FBgn0004167    ENSG00000137877    F    F    F
…
…
…
```

In the above example, some genes could not be analyzed by CAPE-analyze (in this case, the quality data for some transcripts were subpar and as a result, these genes were omitted from the human GTF annotation file). Genes and transcripts for which data are missing will still be shown in the raw data text file but will be given the category "NA". However, the entire record for the corresponding ortholog set will be omitted from both the html files and the Venn diagram calculations, should these be generated.

HTML Summary File

In cases where two, three, or four CAPE-analyze reports are being compared, a summary file in HTML format will be produced. This file will contain a breakdown of transcript counts by category and sample combination as well as a list of the orthologous features comprising each category. The HTML file links the numbers in the summary tables to the corresponding feature list. A sample can be viewed as part of the Use Case document available at http://cape.gersteinlab.org.

R Script for Venn Diagrams

In cases where two, three, or four CAPE-analyze reports are being compared, a ready-to-run R script will be produced. When run, this script will create Venn diagrams for each category, showing the distribution of orthologs across sample combinations (e.g. Worm Only, Worm and Fly Only, etc). Please note that the R script requires that the free VennDiagram R package be installed (http://cran.r-project.org/web/packages/VennDiagram/index.html). Sample Venn diagrams can be viewed as part of the Use Case document available at http://cape.gersteinlab.org.

Acknowledgement of External Libraries

CAPE-analyze makes use of the following external, publicly-available libraries: Google Guava, Apache Commons Math, Apache Commons CLI, JavaPlot, and the Broad Institute's IGV BigFile. CAPE-compare utilizes the Apache Commons CLI external

library. These libraries are packaged as part of the respective executable jar file in their original, unaltered forms.

For more information on each library, please see the following links:

Broad IGV BigFile: http://code.google.com/p/bigwig/

Apache Commons Math: http://commons.apache.org/math/

Apache Commons CLI: http://commons.apache.org/cli/

Google Guava: http://code.google.com/p/guava-libraries/

JavaPlot: http://gnujavaplot.sourceforge.net/JavaPlot/About.html

Additionally, the following free R library is required to run the Venn diagram script file generated by CAPE-compare:

VennDiagram: http://cran.r-project.org/web/packages/VennDiagram/index.htm

Frequently Asked Questions:

Q. In CAPE-analyze, I don't know what percentile I want to use but I do know which specific cutoff values I want to use for my expression and binding cutoffs. Can I specify raw values?

A. You bet! We designed CAPE-analyze to allow for either percentiles or raw values to be used. If you want to define the low RPKM cutoff at ≤ 1 and the high RPKM cutoff at ≥ 5, for example, the following argument should be used at the command line: "--expressionoverride=1,5". The "--bindingoverride" option can be used to set binding

cutoffs in the same manner. Please note that the cutoffs should be given in ascending order. The CAPE-analyze report will reflect your chosen cutoffs in the file header.

Q. CAPE-analyze is giving me very strange results for the ideal window size determined by aggregation. What should I do?

A. Although we have tried to tune our algorithm to choose an ideal window size, sometimes a larger window than expected can be selected if ChIP-Seq signal is particularly broad or if the annotation is not specific (e.g. in worm, for example, there are splice leaders that are not always removed from an annotation. In these cases the actual RNAPII signal may be offset from the annotation start site). For these cases, we suggest setting the "--aggoverride" option to a reasonable setting for your organism. This will skip the initial aggregation step and use your value to define promoter size. For example, "--aggoverride=500" will set the promoter region as +/- 500 bp from the start sites in the annotation file.

Q. Are there any sample runs available?

A. Yes. Sample runs with links to data, logs, and program output are included in the Use Case document available at http://cape.gersteinlab.org. Use cases are given for both transcript and binding modes.

CAPE Use Cases

The following document describes several use cases illustrating how the Coupled Polymerase Binding and Expression Tool (CAPE) can be used to relate matched RNA polymerase II (RNAPII) ChIP-Seq and RNA-seq experiments, identify features with

unusual levels of RNAPII binding vs. mRNA abundance, and compare these transcripts across samples or organisms. Current use cases may always be found at http://cape.gersteinlab.org.

Use Case 1: Comparing Transcription between Worm, Fly, and Human Embryos

Our first use case uses publicly available data from the ENCODE and modENCODE consortia to show how CAPE can be used to compare orthologs between different organisms. In this case, we are comparing embryos from worm, fly, and human. Also, no additional expression files are needed as RPKM values are contained in the gtf files.

Sample Data

The following sample data will be used:

Worm RNAPII ChIP-Seq Signal Data (early embryo):

http://archive.gersteinlab.org/proj/CAPE/usecases/data/2435_Snyder_N2_POLII_eemb_combined.bw

(Note: converted to bigWig format from the public modMine file available at http://submit.modencode.org/submit/public/get_file/2435/extracted/Snyder_N2_POLII_eemb_combined.wig)

Fly RNAPII ChIP-Seq Signal Data (embryo):

http://archive.gersteinlab.org/proj/CAPE/usecases/data/3251_ON_PolII.bw

(Note: converted to bigWig format from the public modMine file available at http://submit.modencode.org/submit/public/get_file/3251/extracted/ON_PolII.wig)

Human RNAPII ChIP-Seq Signal Data (H1 embryonic stem cells):

Available from the ENCODE public data repository at

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs/wgEncodeHaibTfbsH1hescPol2V0416102RawRep1.bigWig

Worm Annotation and Expression Data:

http://archive.gersteinlab.org/proj/CAPE/usecases/data/wormEmbryo.gtf

(Note: Adapted from public data available as part of the modENCODE project. This file only includes entries for known orthologous genes. Expression values are also given in this file).

Excerpt:

```
II      modENCODE_TX    gene    8651057 8658766 .       +       .       RPKM "109.609215"; gene_id "pyr-1"
II      modENCODE_TX    gene    5399522 5405988 .       +       .       RPKM "94.070177"; gene_id "mog-5"
II      modENCODE_TX    gene    13670567        13694711        .       -       .       RPKM "61.110324"; gene_id "Y48E1A.1"
II      modENCODE_TX    gene    11661516        11675863        .       -       .       RPKM "94.316146"; gene_id "trr-1"
```

Fly Annotation and Expression Data:

http://archive.gersteinlab.org/proj/CAPE/usecases/data/flyEmbryo.gtf

(Note: Adapted from public data available as part of the modENCODE project. This file only includes entries for known orthologous genes. Expression values are also given in this file).

Human Annotation and Expression Data:

http://archive.gersteinlab.org/proj/CAPE/usecases/data/humanEmbryo.gtf

(Note: Adapted from public data available as part of the ENCODE/GENCODE projects. This file only includes entries for known orthologous genes. Expression values are also given in this file).

Ortholog File:

http://archive.gersteinlab.org/proj/CAPE/usecases/data/wfhOrthologList.txt

(adapted from the pairwise MIT-Broad Ortholog Project files at

http://compbio.mit.edu/modencode/orthologs/modencode-orths-2012-01-30/ensembl-

v65/modencode.merged.orth.txt.gz)

Excerpt:

```
dhc-1   FBgn0261797    ENSG00000197102
prp-8   FBgn0033688    ENSG00000174231
ama-1   FBgn0003277    ENSG00000181222
sma-1   FBgn0004167    ENSG00000137877
pyr-1   FBgn0003189    ENSG00000084774
F33H2.5 FBgn0020756    ENSG00000177084
T08A11.2       FBgn0031266    ENSG00000115524
rme-8   FBgn0015477    ENSG00000138246
```

Generating CAPE-analyze reports for each organism

The first step is to generate individual reports for each organism using CAPE-analyze in transcript mode. We will now show the text of an example session at the commend line to generate these reports in addition to the output that was produced at each step. Please note that the directory names in the script below are tailored to our test system and should be changed if trying to reproduce these results. Also for this example, CAPE-analyze was run using Java's default heap size on Mac OSX.

Worm Embryo Analysis

```
bash-3.2$ java -jar CAPE-analyze.jar --
signalfile=/emb_study/usecase/data/2435_Snyder_N2_POLII_eemb_combined.bw --
transcriptfile=/emb_study/usecase/data/wormEmbryo.gtf --
outputfile=/emb_study/usecase/wormReport.txt --GFFkey=gene --aggoverride=500
Running in transcript mode...
Signal File: /emb_study/usecase/data/2435_Snyder_N2_POLII_eemb_combined.bw
Transcript File: /emb_study/usecase/data/wormEmbryo.gtf


Loading transcript file...
Performing aggregation with user-defined window size +/- 500 bp...
Processing 1020 features...
Processed 1000 features...
Processed 1020 features...

Determining binding and expression cutoffs...
Updating expression states for all transcripts...
Determining binding and expression cutoffs...
Updating RNAPII promoter binding states for all transcripts...
Writing output to /emb_study/usecase/wormReport.txt...
Program completed in 4 seconds
```

Fly Embryo Analysis

```
bash-3.2$ java -jar CAPE-analyze.jar --
signalfile=/emb_study/usecase/data/3251_ON_PolII.bw --
transcriptfile=/emb_study/usecase/data/flyEmbryo.gtf --
outputfile=/emb_study/usecase/flyReport.txt --GFFkey=gene
Running in transcript mode...
Signal File: /emb_study/usecase/data/3251_ON_PolII.bw
Transcript File: /emb_study/usecase/data/flyEmbryo.gtf

Loading transcript file...
Determining ideal window size with starting value +/- 1000 bp...
Processing 1015 features...
Processed 1000 features...
Processed 1015 features...

Performing aggregation with ideal peak window size +/- 250 bp...
Processing 1015 features...
Processed 1000 features...
Processed 1015 features...

Determining binding and expression cutoffs...
Updating expression states for all transcripts...
Determining binding and expression cutoffs...
Updating RNAPII promoter binding states for all transcripts...
Writing output to /emb_study/usecase/flyReport.txt...
Program completed in 4 seconds
```

Human H1 Embryonic Stem Cells Analysis

```
bash-3.2$ java -jar CAPE-analyze.jar --
signalfile=/emb_study/usecase/data/wgEncodeHaibTfbsH1hescPol2V0416102RawRep1.bigWig --
transcriptfile=/emb_study/usecase/data/humanEmbryo.gtf --
outputfile=/emb_study/usecase/humanReport.txt --GFFkey=gene
Running in transcript mode...
Signal File: /emb_study/usecase/data/wgEncodeHaibTfbsH1hescPol2V0416102RawRep1.bigWig
Transcript File: /emb_study/usecase/data/humanEmbryo.gtf

Loading transcript file...
Determining ideal window size with starting value +/- 1000 bp...
Processing 999 features...
Processed 999 features...

Performing aggregation with ideal peak window size +/- 500 bp...
Processing 999 features...
Processed 999 features...

Determining binding and expression cutoffs...
Updating expression states for all transcripts...
Determining binding and expression cutoffs...
Updating RNAPII promoter binding states for all transcripts...
Writing output to /emb_study/usecase/humanReport.txt...
Program completed in 8 seconds



bash-3.2$ java -jar CAPE-analyze.jar --
signalfile=/emb_study/usecase/data/wgEncodeHaibTfbsH1hescPol2V0416102RawRep1.bigWig --
transcriptfile=/emb_study/usecase/data/humanEmbryo.gtf --
outputfile=/emb_study/usecase/humanReport2.txt --GFFkey=gene --expressionoverride=.1,.3
Running in transcript mode...
Using expression signal thresholds of 0.1 and 0.3...
Signal File: /emb_study/usecase/data/wgEncodeHaibTfbsH1hescPol2V0416102RawRep1.bigWig
Transcript File: /emb_study/usecase/data/humanEmbryo.gtf

Loading transcript file...
Determining ideal window size with starting value +/- 1000 bp...
Processing 999 features...
Processed 999 features...
```

```
Performing aggregation with ideal peak window size +/- 500 bp...
Processing 999 features...
Processed 999 features...

Determining binding and expression cutoffs...
Updating expression states for all transcripts...
Determining binding and expression cutoffs...
Updating RNAPII promoter binding states for all transcripts...
Writing output to /emb_study/usecase/humanReport2.txt...
Program completed in 8 seconds
```

The above runs also show several of the options available to customize analyzes in CAPE. For example, when analyzing the worm data we used the "—aggoverride=500" command line option to tell CAPE to use a window size of +/- 500 bp around the start position given in the annotation file. Worm transcripts have what are called splice leaders that can be included in the annotation files, resulting in a shift of the polymerase-binding site from the annotated start position. CAPE will detect the maximum regardless of its position inside the initial aggregation window, but this "play" in the annotation can produce an overly broad aggregation profile and hence, CAPE-analyze would choose a larger ideal window size. We chose to set a manual window size in this instance. We also performed two different analyses on human. The first uses CAPE's default boundaries for low and high cutoffs for binding and expression (the 25th and 75th percentiles). The second run overrides the expression cutoffs with defined RPKM values using the –expressionoverride option. This was done as an exercise to show that one can refine CAPE-analyze cutoffs using either percentile or raw data values.

CAPE-analyze Output

CAPE-analyze produces four output files from the above runs. For convenience, these files can be obtained at the following links.

Fly CAPE-analyze Report:

http://archive.gersteinlab.org/proj/CAPE/usecases/reports/flyReport.txt

```
#Transcript File: /emb_study/usecase/data/flyEmbryo.gtf
#Signal File: /emb_study/usecase/data/3251_ON_PolII.bw
#Mode: transcript
#Promoter Binding Low Percentile Threshold: 25.0 (Binding <= 1.972
#Promoter Binding High Percentile Threshold: 75.0 (Binding >= 15.744
#Expression Low Percentile Threshold: 25.0 (Expression <= 25.209
#Expression High Percentile Threshold: 75.0 (Expression >= 62.315

#A. Transcripts with low promoter binding and high expression: 25
#B. Transcripts with high promoter binding and low expression: 26
#C. Transcripts with no promoter binding and no expression: 0
#D. Transcripts with no promoter binding: 0
#E. Transcripts with no expression: 0
#F. "Normal" transcripts: 964

#Category      Transcript ID  Chromosome     Start   End     Strand  PromoterSignal
        BodySignal     Ratio (Stalling Index) Expression Value
A       FBgn0004603    2R      1868785 1900039 +       -5.832  3.463   -1.684  65.285
A       FBgn0033062    2R      1968333 1973125 -       -0.213  9.449   -0.023  81.819
A       FBgn0016697    2R      13300269        13301274        +       1.262   8.521   0.148
        111.608
A       FBgn0035046    2R      20551921        20552962        +       1.683   10.89   0.155
        65.167
```

Worm CAPE-analyze Report:

http://archive.gersteinlab.org/proj/CAPE/usecases/reports/wormReport.txt

Human CAPE-analyze Report:

http://archive.gersteinlab.org/proj/CAPE/usecases/reports/humanReport.txt

Human CAPE-analyze Report (using manual expression cutoffs):

http://archive.gersteinlab.org/proj/CAPE/usecases/reports/humanReport-expOverride.txt

Combining and Visualizing Results Using CAPE-compare

CAPE-compare is used to combine and visualize CAPE-analyze reports into a summary format containing information about all organisms. Since we are exploring different organisms in this use case, all with different gene/transcript ID nomenclatures, an ortholog file must be used for the comparison. If we were comparing organisms mapped

against the same annotation set (such as diseased human cells vs. healthy human cells), an ortholog file would not be necessary.

A sample run of CAPE-compare follows, using the CAPE-analyze reports generated above. Note that for human H1 embryonic stem cells, we are using the report generated using CAPE's default options:

```
bash-3.2$ java -jar CAPE-compare.jar -i
/emb_study/usecase/wormReport.txt,/emb_study/usecase/flyReport.txt,/emb_study/usecase/hum
anReport.txt -l Worm,Fly,Human -p /emb_study/usecase/output/wfhEmbryoComparison -o
/emb_study/usecase/data/wfhOrthologList.txt

Running CAPE-compare on the following label/file pairs:
Worm    /emb_study/usecase/wormReport.txt
Fly     /emb_study/usecase/flyReport.txt
Human   /emb_study/usecase/humanReport.txt

Initializing variables...
Parsing reports...
Comparing lists and writing results...
Using ortholog file /emb_study/usecase/data/wfhOrthologList.txt...
Writing raw comparison data to /emb_study/usecase/output/wfhEmbryoComparison.txt...
Writing summary tables to /emb_study/usecase/output/wfhEmbryoComparison.html...
Writing R script to generate Venn Diagrams to
/emb_study/usecase/output/wfhEmbryoComparison.r...
Complete!
```

CAPE-compare Output

Up to three files will be generated for each CAPE-compare run. In all cases, a tab-delimited text file containing the category breakdown for each transcript in an ortholog set will be produced. In cases where two, three, or four CAPE-analyze reports are being compared, two additional files will also be generated. The first is a summary breakdown of transcripts shared between organisms within each CAPE-analyze category. Clicking the numbers in the table will take you the corresponding list of feature IDs. The third file is a ready-to-run R script that will generate Venn diagrams for each CAPE-analyze category showing the breakdown by organism. Note that the free VennDiagram R package must be installed to use the R script (see Users Guide) and that Venn diagrams

will only be produced for categories with at least one data point. Links to the files generated by this run appear below, as well as the Venn diagrams produced by R.

CAPE-compare tab-delimited report (raw data):

http://archive.gersteinlab.org/proj/CAPE/usecases/reports/wfhEmbryoComparison.txt

Excerpt:

```
#A. Transcripts with low promoter binding and high expression
#B. Transcripts with high promoter binding and low expression
#C. Transcripts with no promoter binding and no expression
#D. Transcripts with no promoter binding
#E. Transcripts with no expression
#F. "Normal" transcripts
#NA. No data available in CAPE-analyze report file

#Worm Feature  Fly Feature    Human Feature  Worm State     Fly State      Human State
dhc-1   FBgn0261797    ENSG00000197102        F       F       F
prp-8   FBgn0033688    ENSG00000174231        F       F       F
ama-1   FBgn0003277    ENSG00000181222        F       F       F
sma-1   FBgn0004167    ENSG00000137877        F       F       D
pyr-1   FBgn0003189    ENSG00000084774        F       F       F
F33H2.5 FBgn0020756    ENSG00000177084        F       F       A
…
…
…
```

CAPE-compare HTML Summary Report:

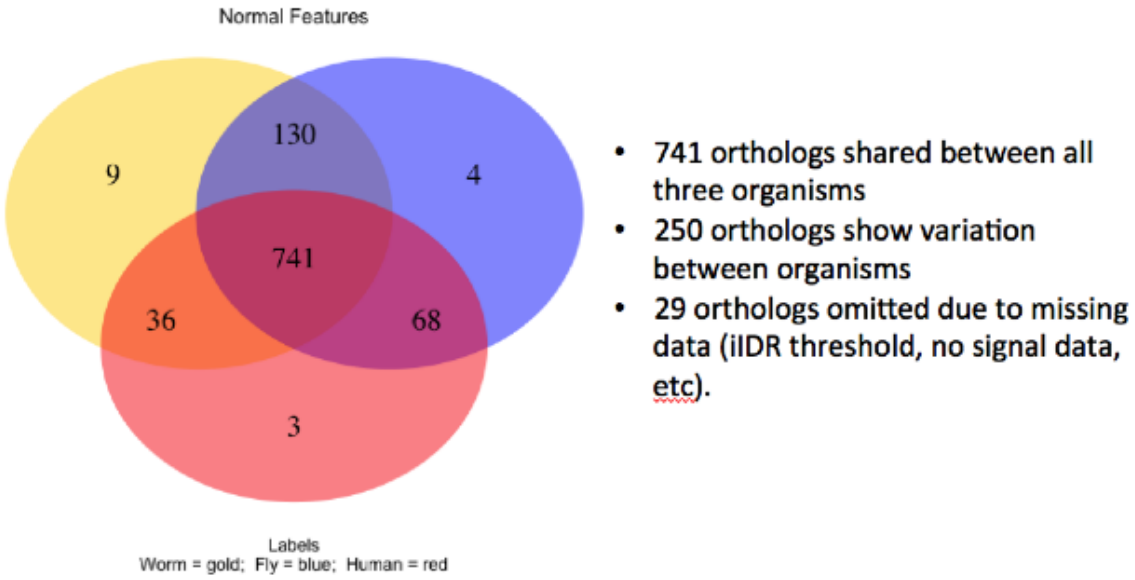
http://archive.gersteinlab.org/proj/CAPE/usecases/reports/wfhEmbryoComparison.html

CAPE-compare R Script (Produces the Venn Diagrams):

http://archive.gersteinlab.org/proj/CAPE/usecases/reports/wfhEmbryoComparison.r

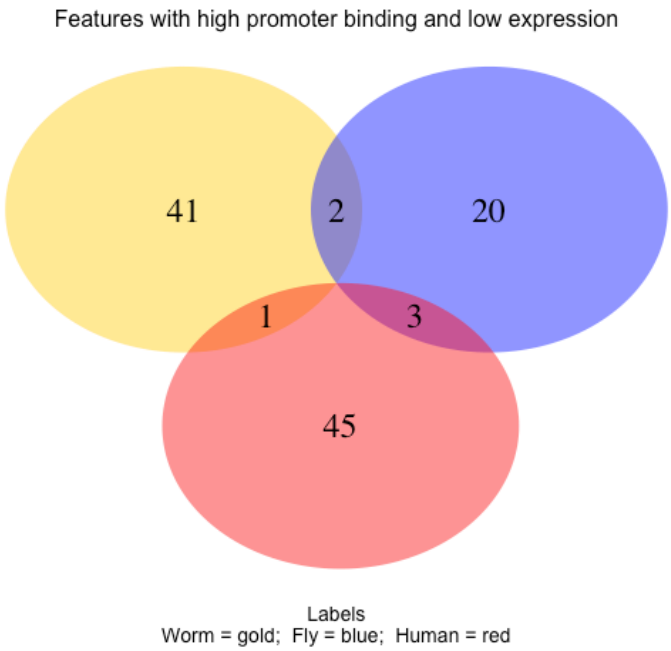
When comparing 1,010 possible orthologous genes across worm early embryo, fly embryo, and human H1 embryonic stem cells, we see that most transcripts fall in the "normal" category for all three organisms. That is, 741 orthologs do not show an extreme difference between the degrees of mRNA abundance and RNAPII binding. 250 orthologs show an extreme case in at least one organism. 29 orthologs were not included in the comparison due to missing data in at least one organism (in most cases, this was due to

the human ortholog not meeting an iIDR quality cutoff of > 1). In cases where data is missing for at least one member of an ortholog set, the entire set will be ignored by CAPE-compare.

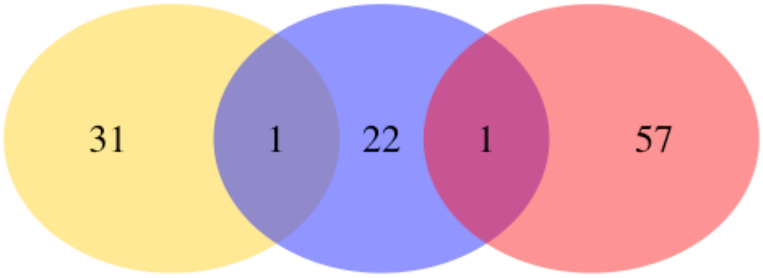


The 250 orthologs that are classified differently in at least one organism are interesting, as this difference may indicate differential regulation between worm, fly, and human embryos. Examining both genes with a stalled polymerase (high RNAPII binding, low expression) and genes that are undergoing a burst of transcription or not transcribed by RNAPII (high expression, low RNAPII binding), we find that affected genes are predominantly organism-specific. 37 orthologs did not have RNAPII binding data for human H1 embryonic stem cells from ChIP-Seq, either due to these promoters falling in unmappable regions or due to a genuine lack of RNAPII binding. No orthologs in this analysis fell into the “No expression” or the “No binding, no expression” categories.

Gene IDs for each category and grouping can be found in the CAPE-compare HTML summary report linked above.

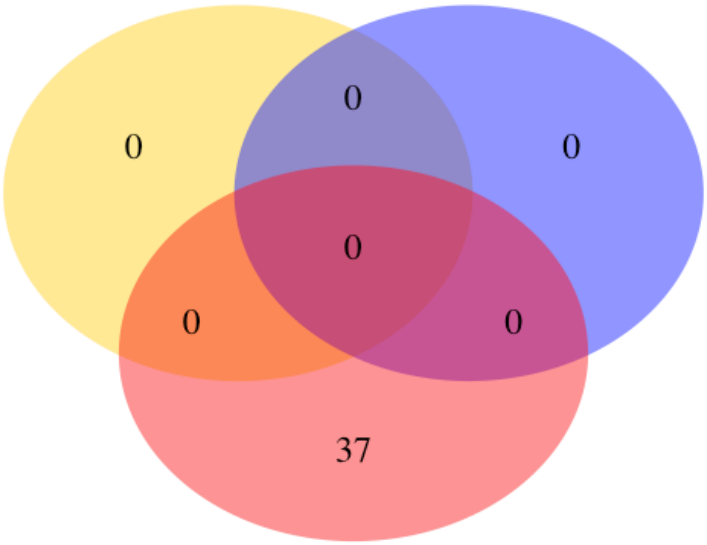


Features with low promoter binding and high expression



Labels
Worm = gold; Fly = blue; Human = red

Features with no promoter binding



Labels
Worm = gold; Fly = blue; Human = red

Use Case 2: Classifying ChIP-Seq Peaks Using CAPE-analyze in Binding Mode

This use case will demonstrate CAPE's binding mode, an additional mode that can help researchers better annotate ChIP-Seq peaks. The functionality of this mode is similar to that of BedTools' closestBed program (http://code.google.com/p/bedtools/), but supports ENCODE formats such as bigwig, bigBed, and narrowPeak files. This mode combines ChIP-Seq peak annotations with transcript or gene annotations and mRNA abundance from RNA-seq. The output is a table of each peak that identifies the nearest feature within a user-defined window, and whether a peak should be classified as associated with a transcription start site (TSS), a transcription termination site (TTS), or neither. For more information about binding mode, please see the Users Guide.

This use case uses publicly available ChIP-Seq signal and peak files for RNA polymerase II from the ENCODE consortium. Annotations were produced by the GENCODE Project using expression data from the ENCODE consortium. All data were generated from the K562 human cell line.

Sample Data

Signal File:

http://archive.gersteinlab.org/proj/CAPE/usecases/data/wgEncodeSydhTfbsK562Pol2StdSig.bigWig

(Note: this file is unaltered from its original version available from the ENCODE public data repository at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/wgEncodeSydhTfbsK562Pol2StdSig.bigWig)

Transcript File:

http://archive.gersteinlab.org/proj/CAPE/usecases/data/K562-HighIDRTranscripts.gtf

(Note: this file is adapted from the GENCODE project file by taking only those transcripts with iIDR [a quality metric] >= 1.0. The original file is available at http://genome.crg.es/~jlagarde//encode/pre-DCC/wgEncodeCshlLongRnaSeq//20120220_long_quantifications_gencodev10_cufflinks_cshl_NOT_SUBMITTED/LID16629-LID16630_TranscriptGencV10IAcuff.gtf.

Peak File:

http://archive.gersteinlab.org/proj/CAPE/usecases/data/wgEncodeSydhTfbsK562Pol2StdPk.narrowPeak

(Note: this file is unaltered from its original version available from the ENCODE public data repository at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/wgEncodeSydhTfbsK562Pol2StdPk.narrowPeak.gz).

Generating a CAPE-analyze Report in Binding Mode

A sample run of CAPE-analyze in binding mode follows. Please note that the "—binding" flag must be specified to run the tool in binding mode, as CAPE's default behavior is to run in transcript mode. Please note that directories are specific to our test system and should be changed if trying to reproduce this result.

```
bash-3.2$ java -jar CAPE-analyze.jar --
signalfile=/ptemp/wgEncodeSydhTfbsK562Pol2StdSig.bigWig --transcriptfile=/ptemp/
gtf/K562-HighIDRTranscripts.gtf --outputfile=/ptemp/CAPE-K562BindingReport.xls --
peakfile=/ptemp/ChIP-Seq/peaks/wgEncodeSydhTfbsK562Pol2StdPk.narrowPeak --binding
Running in binding mode...
Loading ChIP-Seq peak file...
Loading transcript file...
Finding maxima and corresponding signal levels...
Finding nearest TSS and TTS for maxima...
Writing output to /ptemp/CAPE-K562BindingReport.xls...
Program completed in 38 seconds
```

Sample Output

Output from CAPE-analyze running in binding mode:

http://archive.gersteinlab.org/proj/CAPE/usecases/reports/CAPE-K562BindingReport.xls

Excerpt:

```
#Transcript File: ptemp/NewApproach/gtf/K562-HighIDRTranscripts.gtf

#Signal File: ptemp/wgEncodeSydhTfbsK562Pol2StdSig.bigWig
#Peak File: /ptemp/NewApproach/ChIP-Seq/peaks/wgEncodeSydhTfbsK562Pol2StdPk.narrowPeak
#Upstream pad = 1000 bp; Downstream pad = 1000
#Mode: binding

#Pad values used: 1000bp upstream, 1000bp downstream

#Chromosome    Start   End     PeakPosition   PeakScore      DistanceToNearestTSS
       TSSTranscriptID       TSS_RPKM       DistanceToNearestTTS   TTSTranscriptID
       TTS_RPKM       Association
chr1   713770  714492  713983  514.7   23      ENST00000428504.1      2.328   3588
       ENST00000457084.1      0.278   TSS
chr1   762552  763294  762819  163.9   83      ENST00000473798.1      0.415   1233
       ENST00000473798.1      0.415   TSS
chr1   839851  840391  840212  90.4    42228   ENST00000483767.1      0.112   39372
       ENST00000327044.6      5.482   Neither
chr1   878395  878889  878597  35.1    3843    ENST00000483767.1      0.112   987
       ENST00000327044.6      5.482   TTS
chr1   894411  894815  894624  185.8   12      ENST00000469563.1      0.902   998
       ENST00000469563.1      0.902   TSS
chr1   901209  902557  902316  112.5   3384    ENST00000481067.1      0.455   1221
       ENST00000338591.3      1.553   Neither
chr1   935246  936536  935441  136.5   111     ENST00000428771.2      1.113   1099
       ENST00000428771.2      1.113   TSS
```

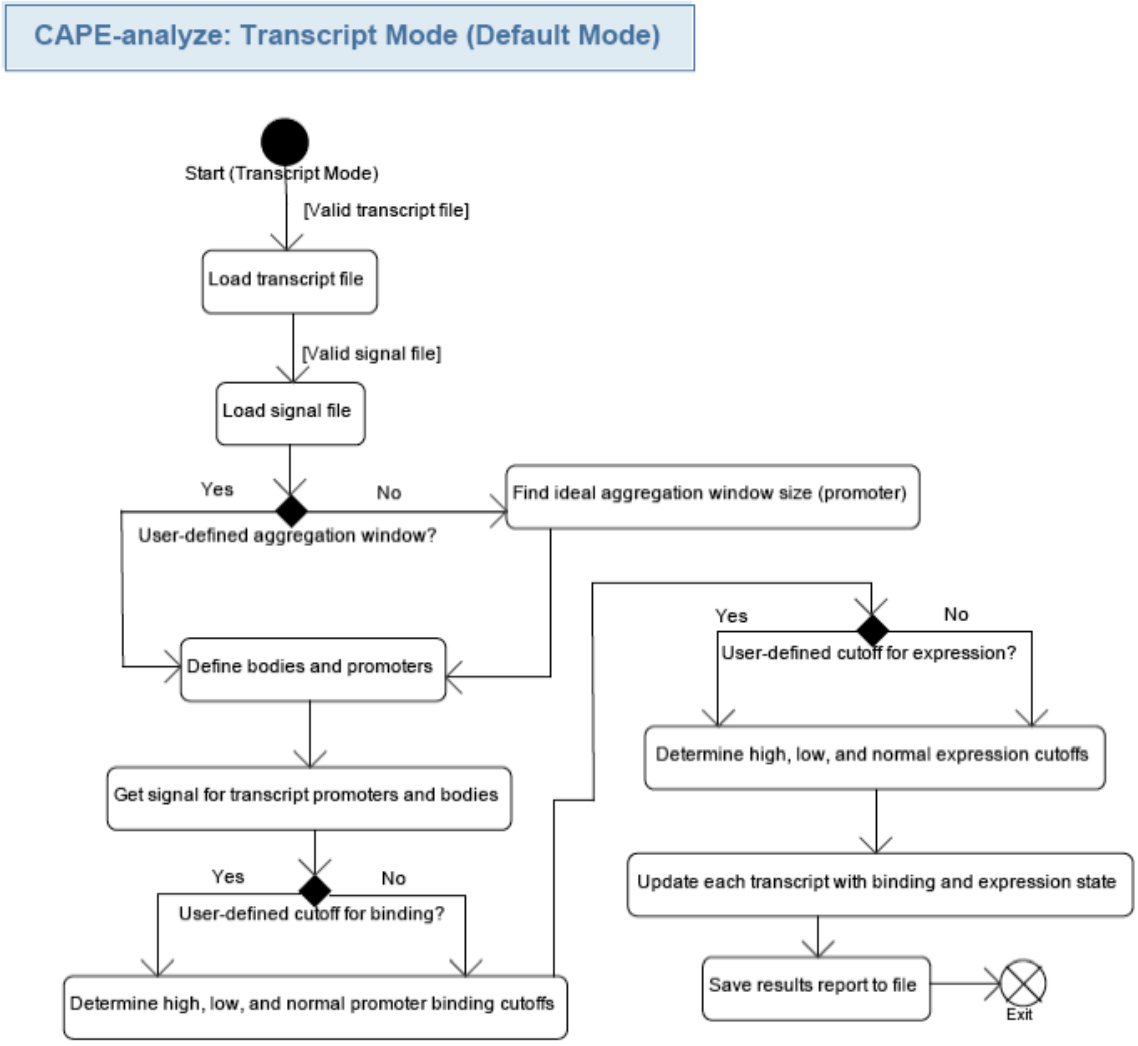
UML Diagrams for CAPE and AnnotationLibrary

Unified Modeling Language (UML) is a common method to show relationships between different classes and functionalities in object-oriented analysis and design. This section contains full UML documentation at the class and activity levels. These graphics are reproduced from http://cape.gersteinlab.org.

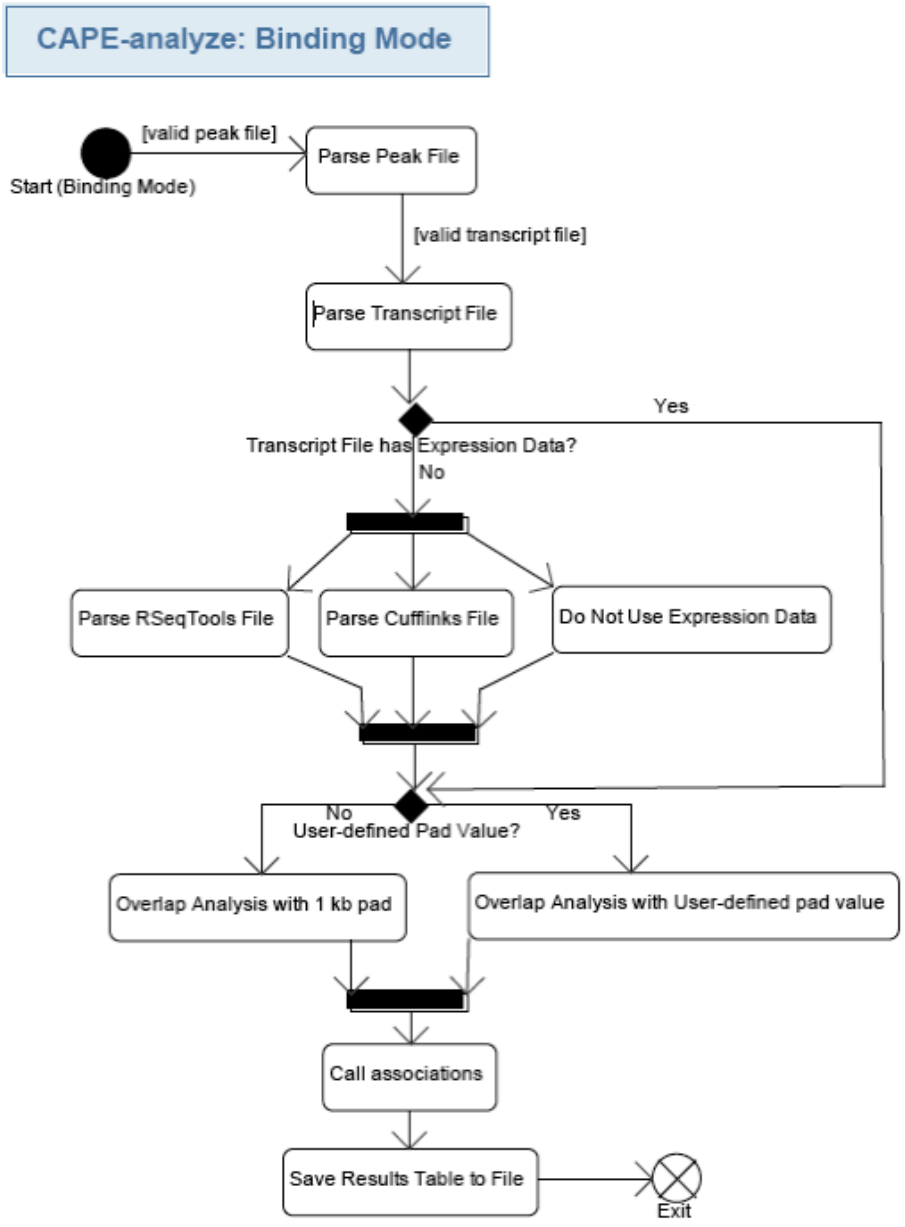
Activity Diagrams

Activity diagrams are essentially flow charts that illustrate the general flow of a program's logic. The following are more specialized versions of Figure 1 in the manuscript. All diagrams were made using the TopCoder UML Tool.

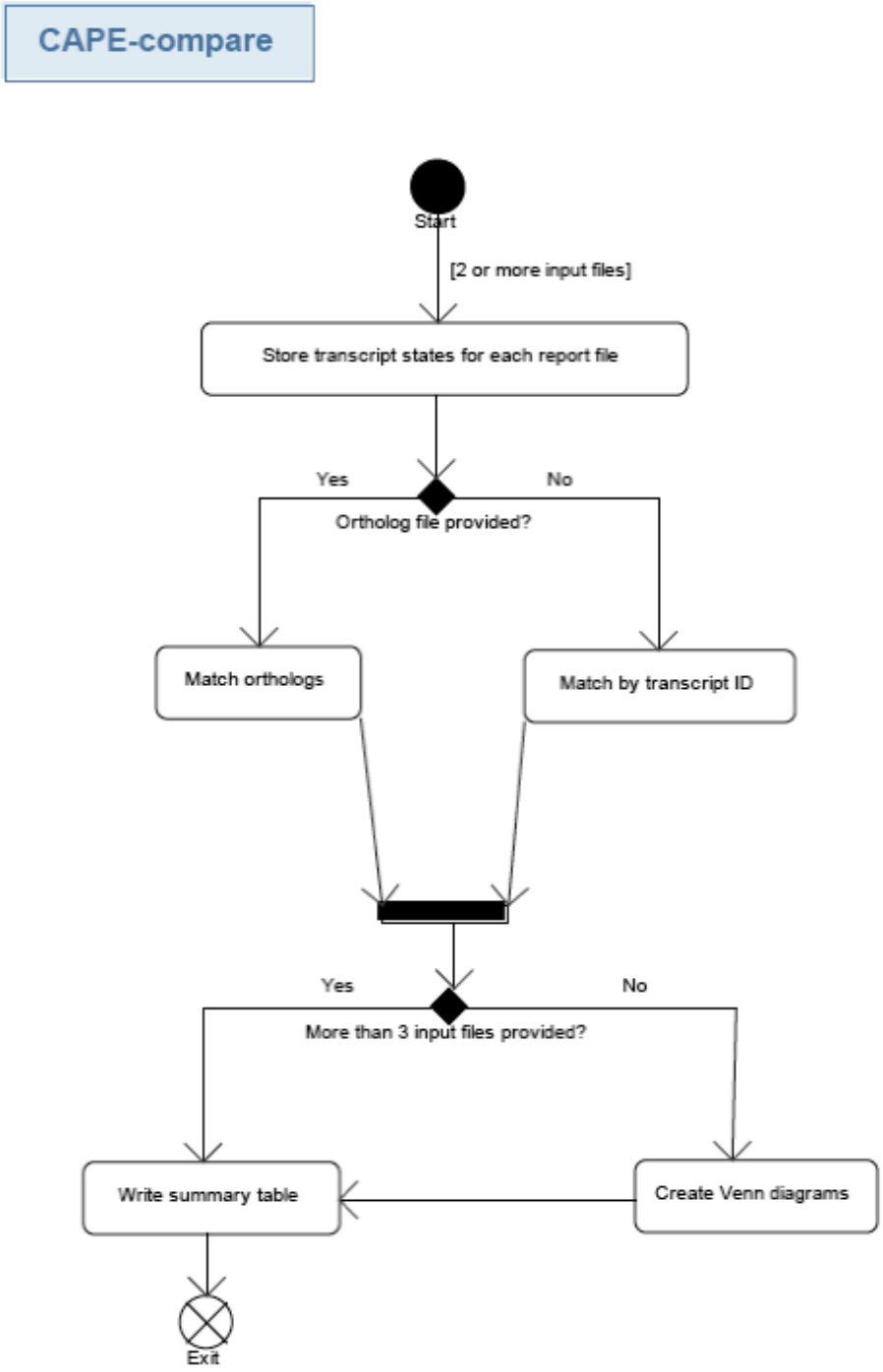
CAPE-analyze (Transcript Mode)



CAPE-analyze (Binding Mode)



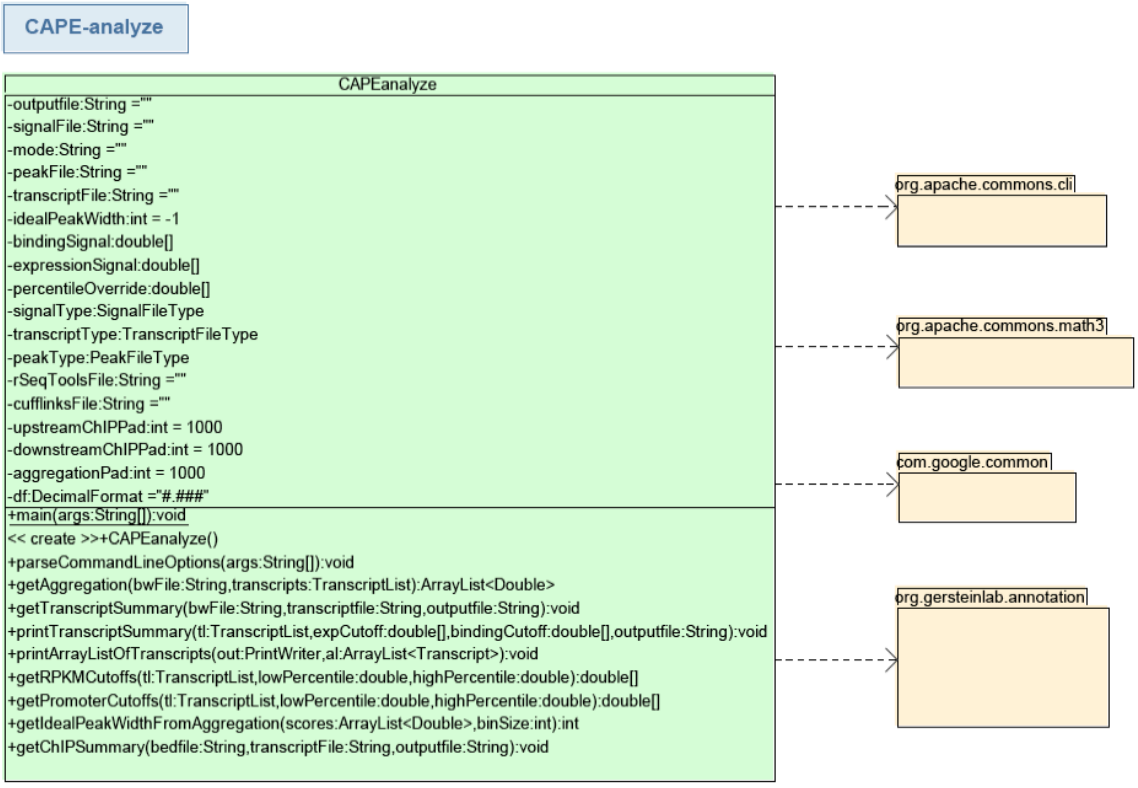
CAPE-compare



Class Diagrams

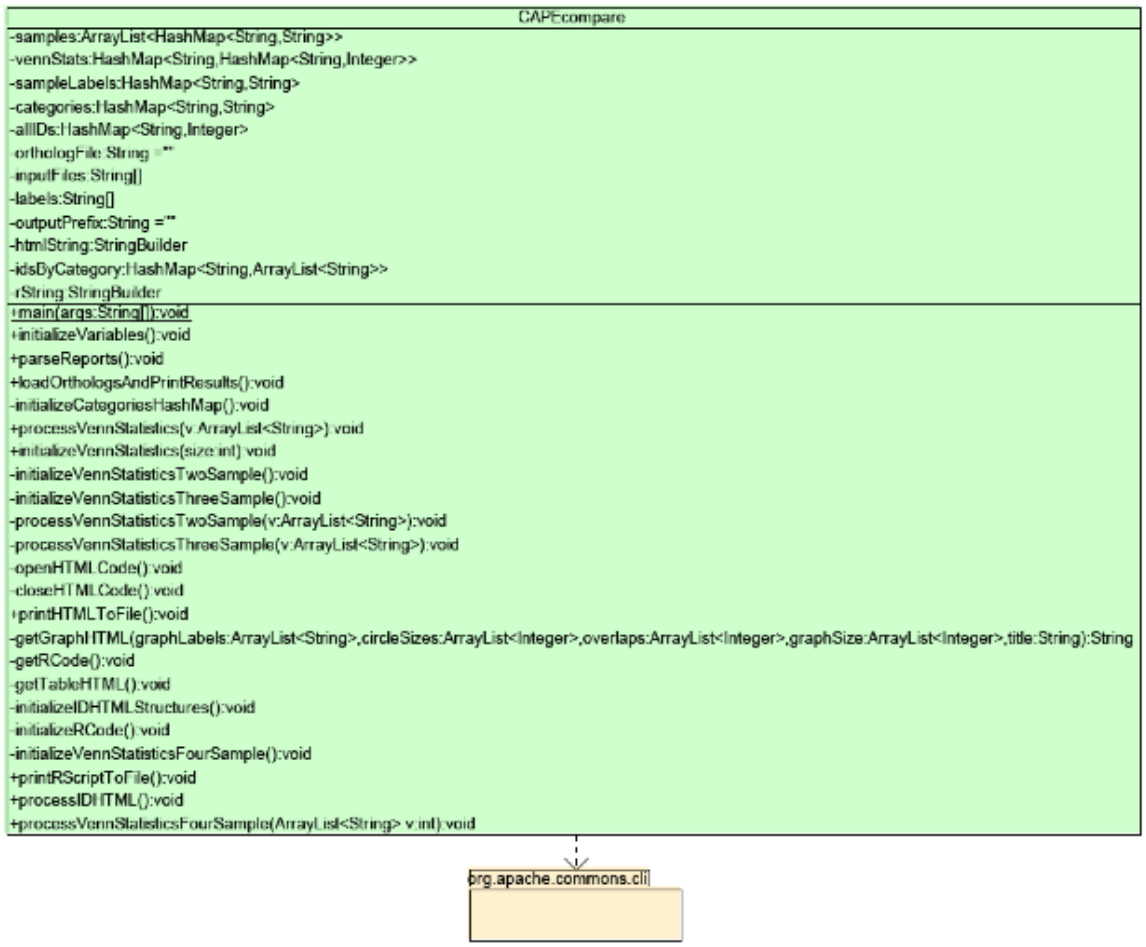
Class diagrams show the internal structure of a class, describing attributes, methods, and parameters. Class diagrams were produced for all AnnotationLibrary classes as well as the main CAPE-analyze and CAPE-compare classes.

CAPE-analyze

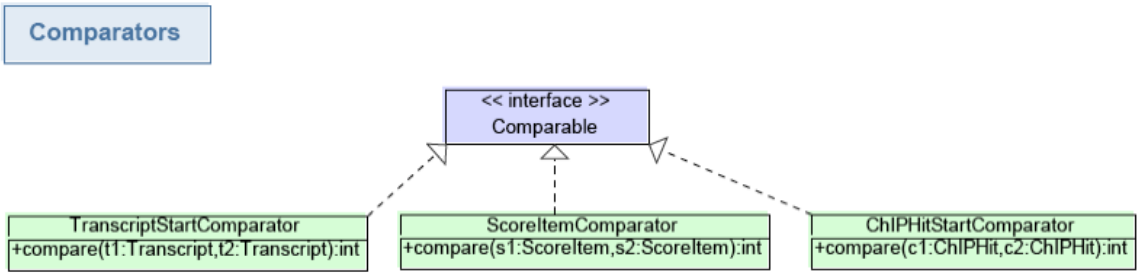


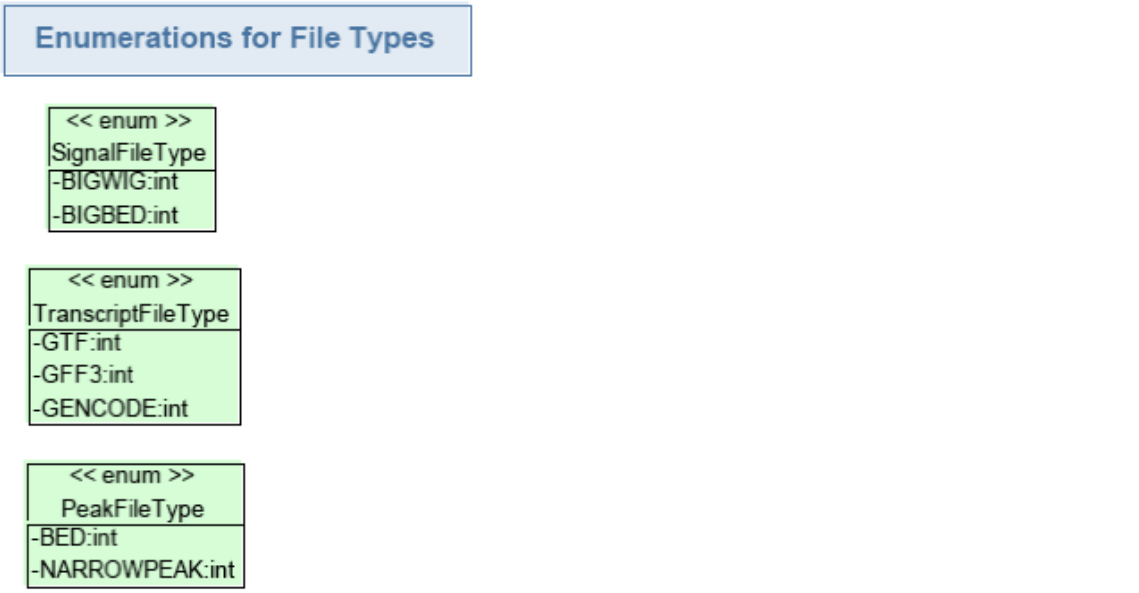
CAPE-compare

CAPE-compare

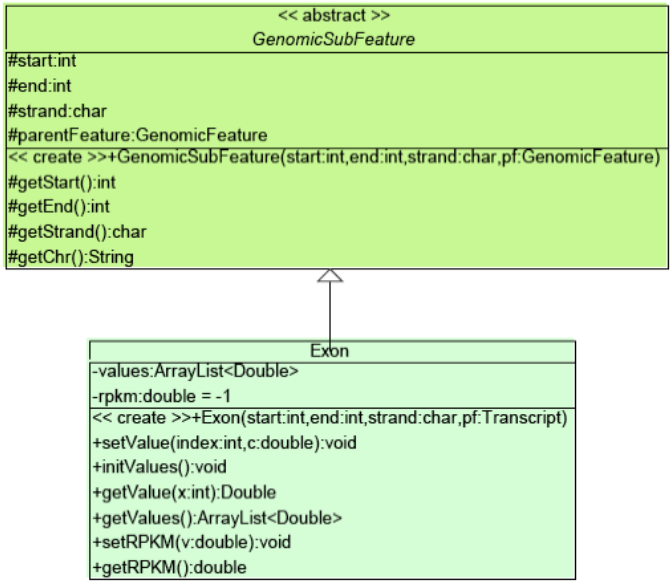


AnnotationLibrary Classes



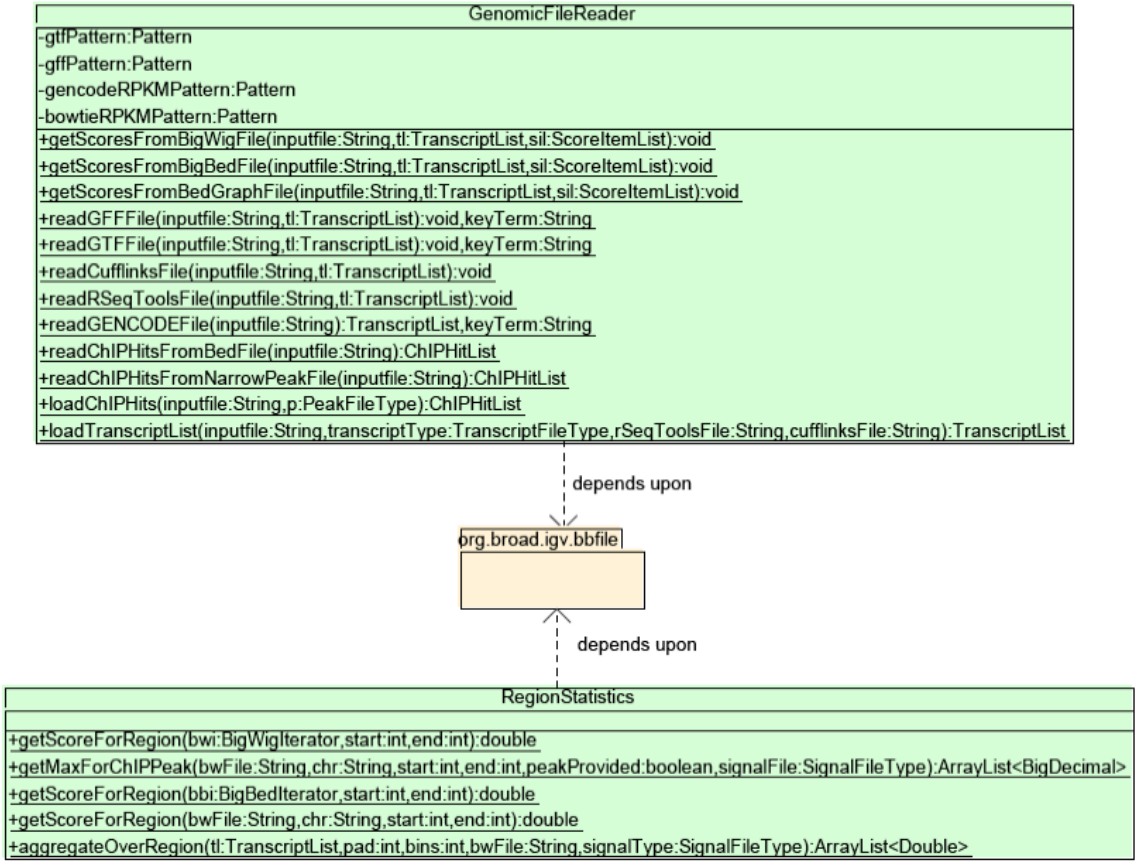


Exons and Genomic Subfeatures



Genomic Features

File Readers and Data Analysis



Classes for Encapsulating Scores