The Dissertation Committee for Daechan Park certifies that this is the approved version of the following dissertation:

Genome-wide Approaches To Explore Transcriptional Regulation In Eukaryotes

Committee:

Vishwanath R. Iyer, Supervisor

Edward M. Marcotte

Tanya T. Paull

Kyle M. Miller

Scott W. Stevens

Genome-wide Approaches To Explore Transcriptional Regulation In Eukaryotes

by

Daechan Park, B.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2014

Dedication

I dedicate this work

to my parents who always love their unruly son,

to my parents-in-law who always support their humble son-in-law,

to my wife, Sanghee, who is the reason I live.

Acknowledgements

I would like to thank my advisor, Dr. Vishwanath R. Iyer, for his patience, support, advice, and generosity, which have allowed me to enjoy science and live a happy life in and outside the lab during graduate school.

I am grateful to the former and current members of the Iyer lab. I thank Drs. Akshay Bhinge, Patrick Killion, Bum-Kyu Lee, and Zheng Liu for teaching me basic and essential skills when I was a baby in the lab. I also thank Drs. Yunyun Ni, Adam Morris, soon to be Drs. Damon Polioudakis, Dia Bagchi, Yaelim Lee, Amelia Hall, and Haridha Shivram, Nathan Abell for many fruitful discussions and collaborations. I especially thank Anna Battenhouse for helping with computational analysis, managing sequencing data, and editing my writings. Also, I want to especially thank Dia Bagchi for editing several of my articles throughout my graduate school and being a nice friend.

I thank my dissertation committee members: Dr. Edward Marcotte, Dr. Tanya Paull, Dr. Kyle Miller, and Dr. Scott Stevens for their time, discussions, and helpful comments. I would like to thank Caitlin Sanford, Amelia Hall, Dia Bagchi, Nathan Abell, and Anna Battenhouse for editorial assistance with this dissertation.

I would like to thank my lovely family. I have never been a good son and brother, but they always support and love me; I am so lucky to have been born into this family. I also thank my parents-in-law for their assistance and for always encouraging me in my studies. Lastly, I thank my wife, Sanghee, for understanding her hard-working husband, supporting my dreams, being my best friend, and helping me learn the meaning of life and family.

Genome-wide Approaches To Explore Transcriptional Regulation In Eukaryotes

Daechan Park, Ph.D.

The University of Texas at Austin, 2014

Supervisor: Vishwanath R. Iyer

Transcriptional regulation is a complicated process controlled by numerous factors such as transcription factors (TFs), chromatin remodeling enzymes, nucleosomes, post-transcriptional machineries, and *cis*-acting DNA sequence. I explored the complex transcriptional regulation in eukaryotes through three distinct studies to comprehensively understand the functional genomics at various steps.

Although a variety of high throughput approaches have been developed to understand this complex system on a genome wide scale with high resolution, a lack of accurate and comprehensive annotation transcription start sites (TSS) and polyadenylation sites (PAS) has hindered precise analyses even in *Saccharomyces cerevisiae*, one of the simplest eukaryotes. We developed Simultaneous Mapping Of RNA Ends by sequencing (SMORE-seq) and identified the strongest TSS and PAS of over 90% of yeast genes with single nucleotide resolution. Owing to the high accuracy of TSS identified by SMORE-seq, we detected possibly mis-annotated 150 genes that have

a TSS downstream of the annotated start codon. Furthermore, SMORE-seq showed that 5'-capped non-coding RNAs were highly transcribed divergently from TATA-less promoters in wild-type cells under normal conditions.

Mapping of DNA-protein interactions is essential to understanding the role of TFs in transcriptional regulation. ChIP-seq is the most widely used method for this purpose. However, careful attention has not been given to technical bias reflected in final target calling due to many experimental steps of ChIP-seq including fixation and shearing of chromatin, immunoprecipitation, sequencing library construction, and computational analysis. While analyzing large-scale ChIP-seq data, we observed that unrelated proteins appeared to bind to the gene bodies of highly transcribed genes across datasets. Control experiments including input, IgG ChIP in untagged cells, and the Golgi factor Mnn10 ChIP also showed the strong binding at the same loci, indicating that the signals were obviously derived from bias that is devoid of biological meaning. In addition, the appearance of nucleosomal periodicity in ChIP-seq data for proteins localizing to gene bodies is another bias that can be mistaken for false interactions with nucleosomes. We alleviated these biases by correcting data with proper negative controls, but the biases could not be completely removed. Therefore, caution is warranted in interpreting the results from ChIP-seq.

Nucleosome positioning is another critical mechanism of transcriptional regulation. Global mapping of nucleosome occupancy in *S. cerevisiae* strains deleted for chromatin remodeling complexes has elucidated the role of these complexes on a genome wide scale. In this study, loss of chromodomain helicase DNA binding protein 1 (Chd1)

resulted in severe disorganization of nucleosome positioning. Despite the difficulties of performing ChIP-seq for chromatin remodeling complexes due to their transient and dynamic localization on chromatin, we successfully mapped the genome-wide occupancy of Chd1 and quantitatively showed that Chd1 co-localizes with early transcription elongation factors, but not late transcription elongation factors. Interestingly, Chd1 occupancy was independent of the methylation levels at H3K36, indicating the necessity of a new working model describing Chd1 localization.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 From DNA to Life

1.1.1 Central Dogma of Molecular Biology

Deoxyribonucleic acid (DNA) is a molecule that encodes inheritable information in every living organism. It is composed of a sugar (deoxyribose) and a base, which is one of adenine (A), thymine (T), guanine (G), and cytosine (C). Long polymers of DNA bases are called DNA sequences, and there are rules that determine the functionality of the genetic codes embedded in the DNA sequences. Transcription refers to a process by which genetic information is transferred from DNA to ribonucleic acid (RNA) molecules by RNA polymerases. Ribosomes bind to messenger RNA (mRNA) and interpret RNA sequence by synthesizing polymers of amino acids, called translation. Finally, mature proteins produced via translation act as structural building blocks, transporting vehicles, energy storages, immune response molecules, and enzymes. This process, called the central dogma of molecular biology, was suggested in 1970 by Francis Crick [1]. This dogma represents a main conduit of genetic information in cells, but modern biology has discovered additional mechanisms by which DNA can determine the characteristics of life. For example, reverse transcription, from RNA to DNA, is a widespread process in many viruses, in particular retroviruses [2]. Another recent example is the identification of functional non-protein coding RNAs such as microRNAs (miRNA), long non coding RNAs (lncRNA), enhancer RNAs (eRNA) [3-5]. These examples of novel genetic coding

open the possibility that a new dogma of molecular biology may yet be discovered.

1.1.2 Control of Gene Expression

Every step in the process from DNA to protein is tightly controlled, and the regulatory mechanisms at each layer allow biological systems to precisely maintain cellular homeostasis and efficiently respond to external conditions. Overall, the control system is comprised of six layers: 1) transcriptional control, 2) RNA processing control, 3) RNA transport and localization control, 4) translation control, 5) mRNA degradation control, 6) protein activity control [6]. This dissertation represents research on the first step, transcriptional control, using multiple genome wide approaches, and further detail is introduced in the next sub-section. In brief, this earliest stage signifies cellular decisions regarding RNA synthesis and the quantity of RNA to be synthesized. The process of RNA synthesis involves many proteins including general and sequence-specific transcription factors (TFs), chromatin modifying enzymes, splicing machineries, and auxiliary machineries. These proteins collaborate to initiate, elongate, and terminate transcription in service to the needs of cells that differentiate and dynamically adapt to environmental variation. For instance, cell type and tissue specific transcription has been extensively connected to cell and tissue phenotypes [7, 8]. More importantly, transcription control is a master regulatory step in cell fate determination, reflected by the fact that expression of four TFs (Oct4, Sox2, c-Myc, and Klf4) can generate induced pluripotent stem (iPS) cells from fully differentiated somatic cells [9]. Moreover, dynamic control of transcription against changes in environmental conditions is another

necessary mechanism for cells or organism to survive. Cells have evolutionarily acquired stress responsive transcription factors such as HSF1 and TP53 to protect against heat shock and DNA damage, respectively [10, 11]. Mutations in key transcription factors lead to mis-regulation of transcription, and subsequently cause a number of diseases including many cancers, autoimmune disorders, and developmental abnormalities [12-15]. Therefore, research on gene expression control is vital for not only understanding basic mechanisms of transcription but also identifying potential therapeutic targets in the treatment of human diseases.

1.2 Transcriptional Regulation

1.2.1 DNA Binding Proteins

Yeast and human RNA polymerase II (RNAPII) is made of 12 subunits, and is recruited by conserved general TFs (TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH) that bind to promoters [16, 17]. TATA-Binding-Protein (TBP) in TFIID recognizes TATA box elements at promoters. Interestingly, TATA-containing genes in yeast tend to be highly regulated by stress conditions [18]. Genome-wide high-resolution ChIP-seq in combination with exonuclease treatment (ChIP-exo) in yeast revealed that previously known TATA-less promoters also contain TATA-like elements, which contain one or two mismatches to canonical TATA box consensus sequences [19]. Additionally, this study showed that pre-initiation complexes (PIC) at TATA-like elements containing genes are well-aligned with +1 nucleosomes, suggesting that well-positioned

nucleosomes proximal to nucleosome depleted regions (NDR) could be indicators of PIC presence.

Sequence specific TFs are functionally more specialized TFs that bind to promoters and enhancers using DNA-binding domains, and *trans*-acting domains recruit co-activators or co-repressors. Through forming DNA loop, TFs can regulate gene expression despite binding to remote enhancers [20]. In other words, TFs, other than RNAPII-associated TFs, are unlikely to randomly bind to gene bodies, because protein-coding regions only constitute about 1.5% of the human genome [8, 21]. Since open chromatin structure generally correlates with high DNA accessibility for TFs [20], it is possible that non-functional stochastic binding of TF could be observed at highly transcribed genes. This topic is discussed in chapter 2.

1.2.2 Chromatin Structure

The nucleosome is a basic unit of eukaryotic chromatin. It consists of about 147 bp DNA wrapped around a histone octamer that is made of two copies each of H2A, H2B, H3, and H4 [22]. DNA that connects two nucleosomes is called "linker" DNA. Eukaryotic chromatin is highly compact, but maintains a more open structure at regions of transcriptional activity. When nucleosomes block TF access to DNA by localizing at TF binding sites, transcription is inhibited even in euchromatin [23]. In other words, nucleosome occupancy is highly dynamic in order to regulate transcription levels. For example, acute heat shock evicted nucleosomes at the promoter and gene bodies of heat-activated genes [24]. This dynamic re-positioning of nucleosomes is executed by ATP-

dependent chromatin remodelers [25]. The functionality of yeast chromatin remodelers is redundant, as the deletion of a single chromatin remodeler gene has little effect on nucleosome organization [26, 27]. Interestingly, loss of Chd1 led to severe nucleosome disruption in yeast [26, 28, 29]. However, loss of Chd1 paradoxically has negligible effect on the levels of transcription based on previous reports of the roles of nucleosome positioning in transcription regulation [26, 30-32]. Therefore, the relationship between transcription and nucleosome positioning deserves further study.

1.2.3 Post-transcriptional Control

The localization, stability, and translation efficiency of mRNAs are highly regulated, with not only a number of proteins but also a variety of non-coding RNAs involved in post-transcriptional regulation. In higher eukaryotic cells, the RNA interference pathway is one well-studied example [33]. Briefly, long double-strand RNAs are processed and cleaved in the nucleus by Drosha and Dicer, respectively, then subsequently exported to the cytoplasm. The RNA-induced silencing complex (RISC) incorporates a single strand of the exported RNA, which guides the RISC to sequence complementary to the short RNA sequence in the 3' untranslated region (UTR) of mRNAs, resulting in either mRNA cleavage and degradation or translational inhibition [34]. The key subunit that has catalytic activity in RISC is Argonaute 2. Meanwhile, bacteria and *Saccharomyces cerevisiae* do not have Argonaute proteins, so they were thought to have relatively simple post-transcriptional controls by ncRNA. However, high resolution tiling arrays and deep sequencing technologies have identified several classes

of non-coding RNAs in yeast, including stable unannotated transcripts (SUTs), cryptic unstable transcripts (CUTs), and Xrn1-sensitive unstable transcripts (XUTs) [35, 36]. Most of XUTs are antisense to open reading frames, and antisense XUTs are associated with transcriptional gene silencing [36]. Therefore, post-transcriptional control via ncRNAs is also widespread in yeast, but a vast majority of newly identified ncRNAs have unknown functions.

1.3 Genome-wide approaches

1.3.1 Microarray

The advent of microarray technology enabled biologists to study gene regulatory networks in a systematic manner. The microarray is a hybridization-based high throughput tool to detect nucleic acids. Labeled nucleic acids are hybridized to DNA probes attached to the surface of a slide, and subsequent automatic scanning allows the identification and quantification of nucleic acid abundance. Depending on how DNA probes are attached to an array slide, microarrays are classified into two types: spotted microarrays and in-situ synthesized oligonucleotides arrays. Correspondingly, there are two different types of probes: cDNA and oligonucleotide [37]. For highest density, resolution, and consistency, in-situ synthesized oligonucleotides arrays are more widely used, and oligo length resides between 25 to 75 nucleotides (nt). However, it is impractical to make whole genome microarray with high-resolution (< 10 bp) for higher eukaryotes due to the limitations of scanner resolution and slide size. At present, the

highest resolution of yeast whole genome microarrays is 4 bp [38].

1.3.2 Next generation sequencing

Relative to "first generation" Sanger sequencing, new sequencing methods are called "Next-generation sequencing (NGS)". The terms "deep sequencing" or "massive parallel sequencing" are also referred to as NGS. In NGS, each heterogeneous short DNA template is clonally amplified, and then a sequence from each population of identical templates is optically detected with fluorescence-based methods. The chemistry of each step varies by NGS platform. For instance, emulsion PCR or bridge PCR can be used for clonal amplification, and the types of fluorescence incorporation methods include cyclic reversible termination, single-nucleotide addition, and dinucleotide ligation [39]. Because each heterogeneous DNA template is sequenced in parallel, NGS of large genomes and transcriptomes is enormously faster and cheaper than Sanger sequencing. Additionally, a key advantage over microarrays is single nucleotide resolution. Therefore, NGS has dominated genomics research over Sanger sequencing and microarray in the last 5 years.

1.3.3 ChIP-seq

Chromatin immunoprecipitation followed by deep sequencing is called "ChIP-seq". The ChIP technique was developed approximately 30 years ago [40]. Since then, its usage has broadened as new DNA detection methods have been developed: ChIP-PCR, ChIP-qPCR, ChIP-chip, and ChIP-seq. Currently, ChIP-seq is the most comprehensive method to map protein binding sites on chromatin. There are three key steps in ChIP, each with its own possible limitations. The first step is the crosslinking of DNA and

protein by formaldehyde. It has been shown that proteins need to stay on chromatin at least 5 seconds to be effectively fixed by formaldehyde [41], so ChIP for dynamically moving chromatin remodelers is notoriously difficult [42]. Although formaldehyde fixation is generally used, it has been recently reported that fixation-free ChIP reduced false positive peak calling [43]. Second, chromatin is randomly sheared using an ultra-sound sonicator, which is necessary for effective immunoprecipitation. However, open chromatin (e.g. highly transcribed loci and linker regions) tends to be more susceptible to shearing, meaning that shearing is not random [44, 45]. In order to computationally correct for background chromatin structure, input or mock ChIP sequencing is commonly performed as a negative control. Third, DNA-protein complexes are incubated with antibodies to pull down specific proteins and their bound DNA. This is the most experimentally difficult step because the availability of high-quality antibodies determines the specificity and efficiency of immunoprecipitation. If an antibody to a native protein is unavailable, tagging the protein of interest can be an alternative method [46]. In *S. cerevisiae*, about 80% of endogenous proteins have been tagged by tandem affinity purification (TAP) at their C-terminus, and the TAP-tag strain library is commercially available [47]. Because TAP has strong affinity to IgG, proteins of interest can be easily immunoprecipitated by IgG conjugated beads.

1.3.4 MNase-seq

Micrococcal nuclease (MNase) is a deoxyribonuclease that digests linker DNA. Due to the nature of MNase, DNA wrapping around nucleosomes is isolated through MNase

digestion followed by size-selection for ~147 bp in a gel [48]. This mononucleosomal DNA can be sequenced with NGS, an experimental protocol called MNase-seq, and the regions where sequencing reads map represent the genomic location of nucleosomes. Based upon several publicly available MNase-seq datasets, the number of nucleosomes in yeast was estimated to be approximately 50,000 [49]. This estimation was correlated with sequencing depth: the more sequencing reads, the higher number of nucleosomes estimated [49]. To quantitatively measure the extent of nucleosome re-positioning, the centers of individual nucleosomes were compared after nucleosome calling [50]. The idea of nucleosome calling is adapted from the concept of peak calling in ChIP-seq [51]. However, nucleosomal peaks are too numerous and too low relative to background, compared to peaks observed in ChIP-seq. Also, this approach to studying nucleosome dynamics is accurate only when the MNase-seq datasets have similar sequence depths, because sequence depth changes the number of identified nucleosomes [49]. Therefore, the comparison of nucleosome occupancy shape is more reasonable approach, and the method is described in chapter 4.

1.3.5 RNA-seq

RNA sequencing (RNA-seq) refers to the sequencing of whole RNA molecules with NGS. Although direct RNA sequencing with NGS has been developed using the Helicos platform [52], prior cDNA synthesis from the RNA sample is required for most NGS platforms. Currently, there are over 20 different RNA-seq library construction protocols, but two methods dominate for strand specific RNA-seq: RNA ligation and deoxy Uridine

Triphosphate (dUTP) methods. In the RNA ligation method, RNA is fragmented, and sequencing adaptors are directly ligated to the RNA. Then, cDNA is synthesized and sequenced. In the latter method, in contrast, cDNA is first synthesized in the presence of dUTP, and then adaptors are ligated to the DNA/RNA duplex. Subsequently, treatment with uracil specific excision reagent (USER/dUTPase) removes the U residues in the cDNA strand, so strand information can be preserved [53]. Many variations of RNA-seq library construction protocols have been developed to obtain more specific information from the whole RNA population. For example, the cap analysis of gene expression followed by deep sequencing (deepCAGE) protocol maps the 5' end of capped RNAs [54], and the 3' region extraction and deep sequencing (3'READ) protocol maps the 3'end of polyadenylated RNAs [55]. Beyond the simple quantification of RNA abundance, structural characterization of RNA is also achievable using RNA-seq methodologies [56], thus novel development of RNA-seq protocols is crucial to improving the efficiency of current protocols, and to the deep exploration of RNA biology on a genome wide scale.

Chapter 2 Simultaneous Mapping of RNA Ends by Sequencing

2.1 Abstract

Understanding the relationships between regulatory factor binding, chromatin structure, *cis*-regulatory elements and RNA regulation mechanisms relies on accurate information about transcription start sites (TSS) and polyadenylation sites (PAS). Although several approaches have identified transcript ends in yeast, limitations of resolution and coverage have remained, and definitive identification of TSS and PAS with single-nucleotide resolution has not yet been achieved. We developed SMORE-seq (Simultaneous Mapping Of RNA Ends by sequencing) and used it to simultaneously identify the strongest TSS for 5207 (90%) genes and PAS for 5277 (91%) genes. The new transcript annotations identified by SMORE-seq showed improved distance relationships with TATA-like regulatory elements, nucleosome positions and active RNA polymerase. We found 150 genes whose TSS were downstream of the annotated start codon, and additional analysis of evolutionary conservation and ribosome footprinting suggests that these protein coding sequences are likely to be mis-annotated. SMORE-seq detected short non-coding RNAs transcribed divergently from more than a thousand promoters in wild-type cells under normal conditions. These divergent non-coding RNAs

---

This work was published in Park D., Morris A.R., Battenhouse A. & Iyer V.R. Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements (2014) *Nucleic Acids Res.* 42(6): 3736-49. ARM, DP, and VRI conceived and designed the experiments. ARM and DP performed the experiments. DP, ARM, and AB analyzed the data. All authors wrote the manuscript. Permission to adapt the contents of the publication was acquired from the co-authors.

were less evident at promoters containing canonical TATA boxes, suggesting a model where transcription initiation at promoters by RNAPII is bidirectional, with TATA elements serving to constrain the directionality of initiation.

2.2 Introduction

Transcription initiation depends on interactions between general transcription factors and RNA polymerase with promoter sequences and nucleosomes near the transcription start site (TSS) [18], while posttranscriptional regulation typically depends on sequences in 5' and 3' untranslated regions (UTRs) of mRNAs [57]. Understanding the overall relationships between these aspects of gene regulation requires knowledge of TSS and transcript ends, or polyadenylation sites (PAS), of mRNAs genome-wide at single-nucleotide resolution. Although genes often have multiple TSS and PAS, identifying the most prominent transcript ends is useful for revealing their relationships to cis elements like TATA boxes, polyadenylation control sequences, and other features like positioned nucleosomes. Definitive annotation of transcript ends is also critical for accurate mapping of reads generated by next-generation sequencing (NGS) to reference transcriptomes. High-resolution tiling microarrays and NGS methods are increasingly used for transcript analysis, but even for a well-studied model organism like *Saccharomyces cerevisiae*, currently available and commonly used transcript annotations remain inaccurate and potentially obscure relationships between the above-mentioned aspects of gene regulation.

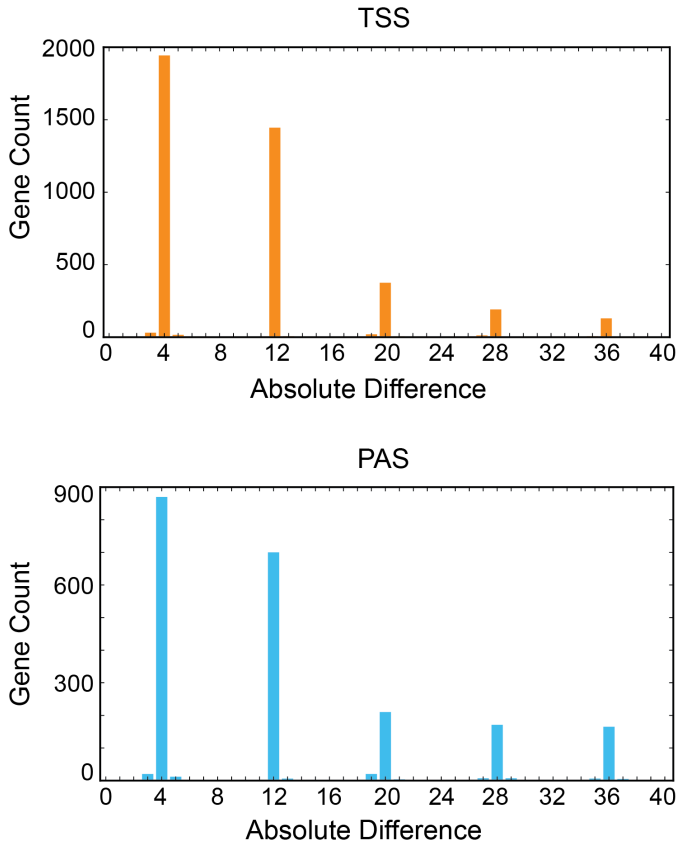

Figure 2.1 Absolute difference between tiling microarray annotations

Absolute difference between TSS coordinates reported in previous analyses by David et al. and Xu et al. shows 8 nt periodicity which corresponds to the maximum resolution of the tiling microarray platform used in these studies [35, 38].

S. cerevisiae has many qualities that make it an ideal model organism for studying gene expression and chromatin architecture, including a relatively small number of genes and a compact genome. The first high-throughput TSS identification in yeast was based on Sanger sequencing of 5' end tags from cDNAs, and mapped 2231 TSS with single-nucleotide resolution [58]. A subsequent study used a "vector-capping" approach with Sanger sequencing to identify TSS, but coverage was limited to only about 60% of all genes [59]. More importantly, although the Sanger sequencing provided single-nucleotide

resolution, the number of sequence tags counting towards a given TSS was low. This inherently low sampling of ends with Sanger sequencing makes it difficult to assign one prominent TSS for a gene with high confidence, especially for genes with low transcript levels.

Subsequent approaches used tiling oligonucleotide microarrays to study the yeast transcriptome at high resolution and defined TSS of mRNAs and non-coding RNAs (ncRNAs) [35, 38]. However, in these studies, which are the highest resolution microarray analyses of transcripts carried out to date in any organism, the resolution was limited to 8 nucleotides (nt), the distance between adjacent probes interrogating transcripts from each strand of genomic DNA. This 8 nt resolution is apparent for both TSS and PAS, in a comparison of independently published datasets using the same microarray platform (Figure 2.1). Although these TSS and PAS have been used in many recent landmark analyses of transcription factor and nucleosome localization datasets [19, 42], the 8 nt resolution remains a limitation. RNA-seq can potentially identify TSS and PAS at single-nucleotide resolution [60]. However, RNA-seq signals are complex and do not necessarily show an easily identifiable boundary corresponding to transcript ends. In addition, this strategy will tend to identify the most distal 5' or 3' ends, which may not be the site most frequently used in vivo.

In order to overcome these limitations, refinements of NGS-based methods have been developed to map TSS and PAS. One approach involves the use of tobacco acid pyrophosphatase (TAP) to remove the 5' cap and allow ligation of a sequencing adapter

specifically to the 5' end of the RNA [61-63]. To map PAS, methods based on oligo(dT) priming or poly(A) capture have been used [55, 64-69]. Existing methods work well to map TSS or PAS but only identify one end of transcripts. The recently introduced TIF-seq method can be utilized to simultaneously map TSS and PAS [70], but this study focused on the diversity of transcript isoforms in yeast and did not define canonical TSS and PAS. Thus, none of these methods has been employed to identify a definitive set of TSS and PAS in yeast, which has the most extensive complementary data on the location of the general transcription machinery [19, 71] and nucleosome positions [42].

Here, we describe SMORE-seq (Simultaneous Mapping Of RNA Ends with sequencing), a method for identifying both mRNA TSS and PAS from a common set of experimental data with single-nucleotide resolution. We demonstrate that SMORE-seq maps TSS and PAS more accurately and efficiently than existing methods. The improved annotations of transcript ends revealed a significant fraction of likely mis-annotated protein coding sequences in the genome, and showed sharper relationships between cis-regulatory elements, chromatin features and transcript ends. SMORE-seq also revealed pervasive bidirectional transcription from most promoters, and our analysis suggests that the TATA element serves to constrain the direction of transcription initiation by RNA polymerase.

2.3 Materials and Methods

2.3.1 Yeast growth and RNA preparation

The *S. cerevisiae* strain used in this study was BY4741, and cells were grown in yeast extract-peptone-dextrose (YPD, Difco) at 30°C to an A600 OD of 0.8. We harvested the cells by centrifugation at 3000 rcf for 5 min, and the cell pellets were frozen in liquid nitrogen after discarding supernatant. Total RNA was extracted with a standard hot phenol method [72].

2.3.2 Construction of SMORE-seq libraries

Poly(A)+ RNA was purified from yeast total RNA using the MicroPoly(A) Purist kit from Life Technologies. 500 ng poly(A) RNA was mixed with 5 units (1 µl) Tobacco Acid Pyrophosphatase (TAP) (Epicentre) and 2.5 μl 10x TAP buffer in a 25 µl total volume. A parallel reaction without TAP enzyme was also performed. TAP reactions were carried out at 37 °C for one hour, followed by heat inactivation at 65 °C for 5 minutes. RNA was purified with the RNEasy MinElute kit (Qiagen) and eluted in 26 µl of water. 23.5 µl of this RNA was combined with 1 µl of a 1/2 dilution of 5' SR Adaptor, 3 µl 10x Ligation Reaction Buffer, and 2.5 ul 5' Ligase Enzyme Mix (for descriptions of these components see NEBNext Small RNA Library Prep Set for Illumina). This reaction was incubated one hour at 25 °C, followed by purification with Agencourt AmPure XP beads (Beckman Coulter) following manufacturer's instructions at a 1.5x concentration and elution in 18 µl water. This RNA was then fragmented for 4 minutes at 94 °C using NEB fragmentation reagent, followed by cleanup with AmPure XP (1.8x) and elution in 10 µl of water. This RNA was then ligated to a 3' sequencing adapter as described in the manufacturer's protocol (NEBNext Small RNA Library Prep Set for Illumina), followed

by reverse transcription and 10 cycles of PCR according to manufacturer's instructions. PCR products of ~250 bp were selected by E-gel (Invitrogen) and subjected to another 8 cycles of PCR. The resulting libraries were verified on an Agilent Bioanalyzer and sequenced on an Illumina HiSeq 2000 with single-end or paired-end 100 base reads.

2.3.3 Analysis of sequencing reads

Alignment of sequencing reads was performed with bwa (version 0.6.2) using default options for paired end or single end libraries, as appropriate [73]. The reference genome was sacCer3 (April 2011) from UCSC, derived from the Saccharomyces Genome Database. The 100 bp read sequences were trimmed to 50 bp before alignment. Aligned R1 (5' reads) were extracted from the resulting BAM files using samtools (version 0.1.18) [74] and merged for the three TAP+ and TAP- replicates, respectively. Reads that mapped to snRNA and rRNA were removed. Plus (Watson) and minus (Crick) strand aligned reads were then extracted and processed separately for TSS calling.

2.3.4 TSS calling algorithm

According to previous studies that mapped TSS in yeast, the estimated median 5'-UTR length is 50-60 bp, and approximately 90% of 5'-UTRs are shorter than 300 bp [38, 58, 60, 75]. For each verified and uncharacterized gene, we searched for TSS within a window ranging from 300 bp upstream of the annotated ORF to the midpoint of the ORF downstream from its annotated start codon. In order to correct TAP+ by TAP-, Gaussian kernel density estimation was utilized for peak calling, and a bandwidth of 5 and a read threshold of 2 were applied. When TAP+ peaks were present within ±50bp of TAP-

peaks, the peaks were corrected. Then, the base position with the highest read stack within the highest corrected peak was called as the TSS. Manual curation was mainly aimed at calling TSS for the genes with a 5'-UTR longer than 300 bp. In addition to recovering TSS with long 5'-UTRs, potential TSS that showed the following examples were dropped during manual curation: evenly distributed peaks with a low number of reads, TSS adjacent to tRNA, snRNA, or rRNA, TSS overlapping with a neighboring gene, and TSS in close proximity to a neighboring gene.

2.3.5 TATA element data processing

The ChIP-exo technique previously identified the TATA box as well as TATA-like elements at "TATA-less" promoters [19]. In this study, a canonical TATA represents a TATA-box with no mismatches, and TATA-elements include canonical TATA with 0,1, or 2 mismatches. TATA element data for sacCer3 were downloaded from the SGD Genome Browser (http://browse.yeastgenome.org/fgb2/gbrowse/scgenome/).

2.3.6 High resolution tiling array data processing

Although these data are available in SGD, the data were downloaded from the journals where the papers were originally published because the authors assigned gene names but SGD provided only segment information [35, 38]. The data were lifted over into sacCer3 from the genome version the authors originally used.

2.3.7 RNAPII Ser 5-P and nucleosome localization

150 ml of WT cells were grown in YPD, and harvested at 0.8 A600 OD for each sample. Cells were cross-linked with formaldehyde to a final concentration of 1% for 15

min, then quenched with glycine to a final concentration of 125 mM. For ChIP, cells were resuspended in 2 ml of lysis buffer, and were lysed by glass bead beating for 9 min. Chromatin was sheared with a probe-sonicator to 150 bp – 200 bp fragments. After pre-clearing with protein A-agarose beads (Roche), the fragmented chromatin was incubated with 8 μg of RNAPII Ser 5-P specific antibody (Abcam, cat.# ab5131) overnight, then further incubated with 100 μl protein A beads. Serial washing was performed, and finally DNA was reverse-crosslinked at 65 °C overnight, then collected by ethanol precipitation. For mononucleosome isolation, we followed the protocol described in [24]. Briefly, cells were resuspended 20 ml of zymolyase buffer, and spheroplasts were made with 250 μg of zymolyase. The spheroplasts were spun down and resuspended in 2 ml NP buffer. Then, micrococcal nuclease (MNase) was added at a concentration from 40 U-100 U for 10 min at 37 °C. Digested chromatin was reverse-crosslinked with Proteinase K in 1% SDS and 10 mM EDTA solution at 65 °C overnight. After RNase A treatment, DNA was purified by phenol chloroform extraction followed by ethanol precipitation. Finally, DNA fragments of ~147 bp were size-selected with an E-gel system (Invitrogen). Sequencing libraries for both ChIP and mono-nucleosomes were prepared using NEB Library Prep Kit and Bioo multiplex adapter for Illumina, then sequenced by paired-end sequencing. In order to profile occupancy, coordinates of mapped reads were shifted toward the center of the insert DNA by a distance equal to half of the insert size, then reads were counted in bins of 5 bp.

2.3.8 Conservation and ribosome footprinting analysis

WIG files of conservation scores were downloaded from the UCSC Genome browser (http://hgdownload.cse.ucsc.edu/goldenPath/sacCer3/phastCons7way/). Using a customized python script, we extracted base-by-base conservation scores near annotated start codons of all genes and internal TSS genes.

Raw sequencing data of ribosome footprinting in rich media were downloaded from Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) (accession number GSE13750) [76]. Only the first 21 nucleotides were mapped onto sacCer3 with bwa using default options, and for any given gene, only reads that mapped to the sense strand were considered.

2.3.9 Polyadenylation site analysis

The sequenced read fastq files from both TAP+ and TAP- (three replicates each) were first processed to remove 3' adapter sequences with cutadapt (version 1.2.1). Each resulting sequence set was filtered, retaining only the R1 sequences with at least 35 bases followed by a stretch of at least 8 A bases within 5 bp of the adapter-trimmed 3' end. For each resulting poly(A) selected sequence set, a corresponding trimmed version was created such that only bases 5' of the poly(A) stretch were retained. The poly(A) selected full length and poly(A) selected trimmed fastq files were then single-end aligned to sacCer3 with bwa as described above (Table 2.1).

Polyadenylation site data in sacCer3 were downloaded from the SGD Genome Browser [66]. Since Ozsolak et al. provided the genomic coordinates of the read clusters and the scores of the clusters as the read counts that support the highest peak, we

processed the data to call one poly(A) site per gene. In order to process the data in the same way as the SMORE-seq, we assigned the clusters into ranges from annotated stop codons to 300 bp downstream. Among the clusters per gene, the position that had the highest read count was defined as the poly(A) site for the gene.

2.3.10 Accession number

The SMORE-seq data from this study have been deposited in NCBI GEO under accession number GSE49026. The MNase-seq data for nucleosome mapping is also available from GEO under accession number GSE52355.

Sample	Selected reads	Mapped reads Before trimming	Mapped reads After trimming	Trim-only reads	PAS reads
TAP+	402,728	33,444 (8.30%)	355,881 (88.37%)	322,457	304,035
TAP-	1,066,647	69,440 (6.51%)	975,941 (91.50%)	906,526	859,568

Table 2.1 Counts of poly(A) selected reads and PAS reads

From raw sequencing reads, poly(A)-containing reads were selected. The reads were mapped before and after poly(A) trimming. The reads that were mapped after trimming but unmapped before trimming were called "Trim-only reads". The reads that included non-M in the CIGAR string in the bam file were excluded in PAS reads.

2.4 Results

2.4.1 5' cap sites with single-nucleotide resolution in SMORE-seq

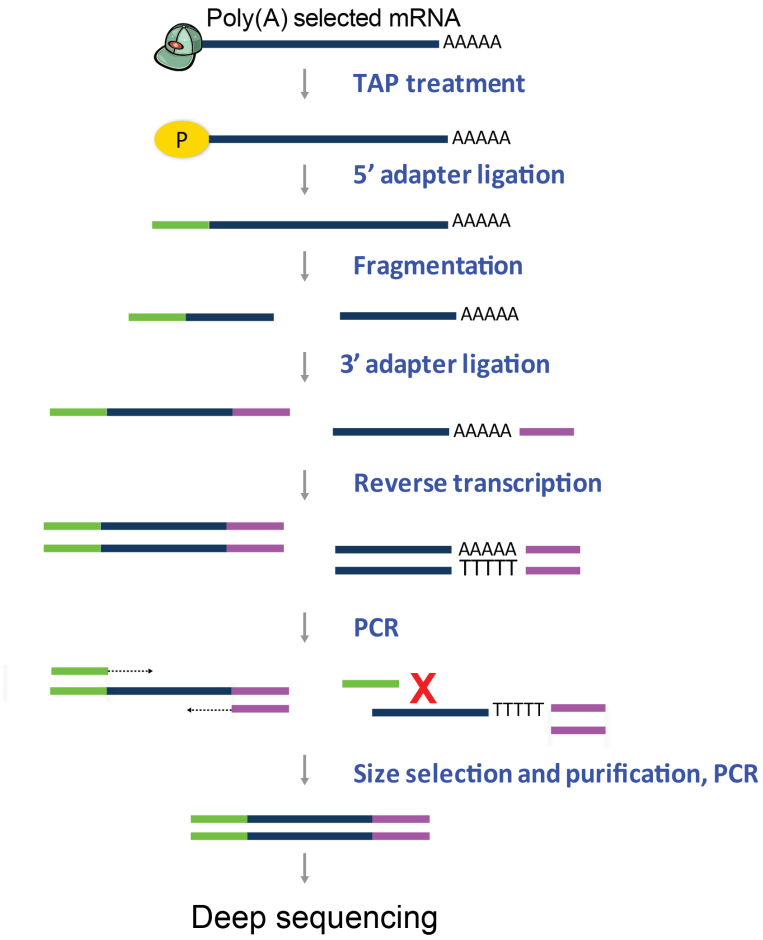

Figure 2.2 Overview of the SMORE-seq method.

TAP enzyme is used to convert mRNA 5' caps into phosphates, followed by 5' adapter ligation, fragmentation, 3' adapter ligation, RT-PCR, size selection, and additional PCR.

We constructed SMORE-seq libraries according to the flowchart shown in Figure 2.2. Two technical replicate libraries and one biological replicate library were prepared, with a control library that was not treated with the TAP enzyme prepared in parallel for each sample. In total, we produced 12,652,059 and 11,161,171 single-end, 100 base reads for the TAP+ and TAP- samples respectively, after filtering reads mapping to snRNA and rRNA regions. 7,622,443 (60.2%) reads in the TAP+ libraries were mapped within 300

bp region upstream of ORF start codons, whereas only 890,128 (8.0%) reads were mapped to those regions in the TAP- libraries. Most reads mapped to or near genes in both TAP+ and TAP-, and the strongest read signals were observed just upstream of annotated start codons in TAP+, whereas relatively few reads in TAP- mapped to these locations (Figure 2.3). This difference in the read pattern in the TAP+ and TAP- libraries suggests that our TAP+ library was selective for the TSS.

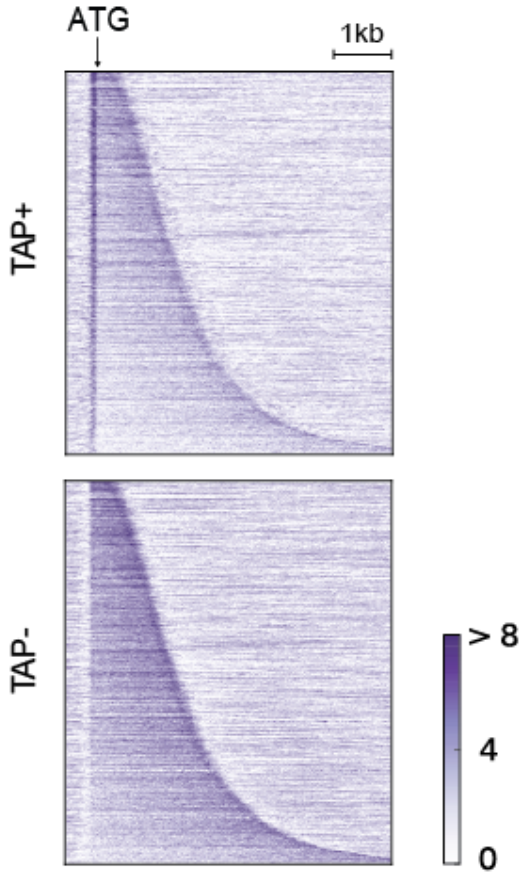

Figure 2.3 Heat map representation of SMORE-seq read data.

Genes are sorted by ORF length. The arrow represents the positions of start codons in SGD annotation, and genes are aligned by the start codon. Color scale is read count per 10 bp.

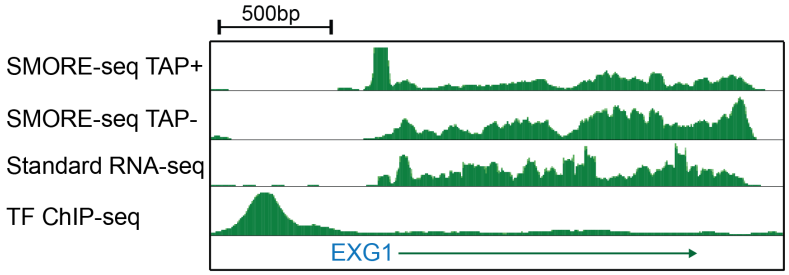

Figure 2.4 Comparison of SMORE-seq to standard RNA-seq and ChIP-seq

The data are visualized in the UCSC Genome Browser mirror. TAP+ has similar background signal as TAP-, including appreciable signal at the 3' end, indicating that correction by TAP- is necessary for TSS identification. The peak shape of transcription factor (TF) binding sites in ChIP-seq is different from that of TSS peaks in TAP+, thus adjustments to peak-calling algorithms used in ChIP-seq were required for analysis of SMORE-seq data.

To identify candidate TSS, we employed a modified version of our peak-finding algorithm based on Gaussian kernel density estimation, followed by correction of the TAP+ data by the TAP- control. Although TAP+ reads were highly enriched in 5'-UTRs, there was appreciable background signal within ORFs and 3'-UTRs, necessitating its correction by TAP- in order to reduce false positives. We adapted the parameters of our peak finding algorithm to exploit the characteristics of the SMORE-seq data, which was distinct from standard RNA-seq and ChIP-seq data in terms of its strand-specificity, localization relative to ORFs, and sharpness (Figure 2.4). This procedure resulted in the identification of 138,352 candidate TSS that were defined by 2 or more reads. To identify the most prominent TSS for a gene, we assigned the corrected peaks at 5' ends to genes, then determined the position of the most abundant read stack within the most significant peak for each gene. By doing so, we obtained the most frequently used TSS for a gene rather than the most upstream TSS identified in previous studies [35, 38]. We applied this

procedure independently to the three replicates, and ascertained that the identification of TSS with single base resolution was highly reproducible across replicates (Figure 2.5A).

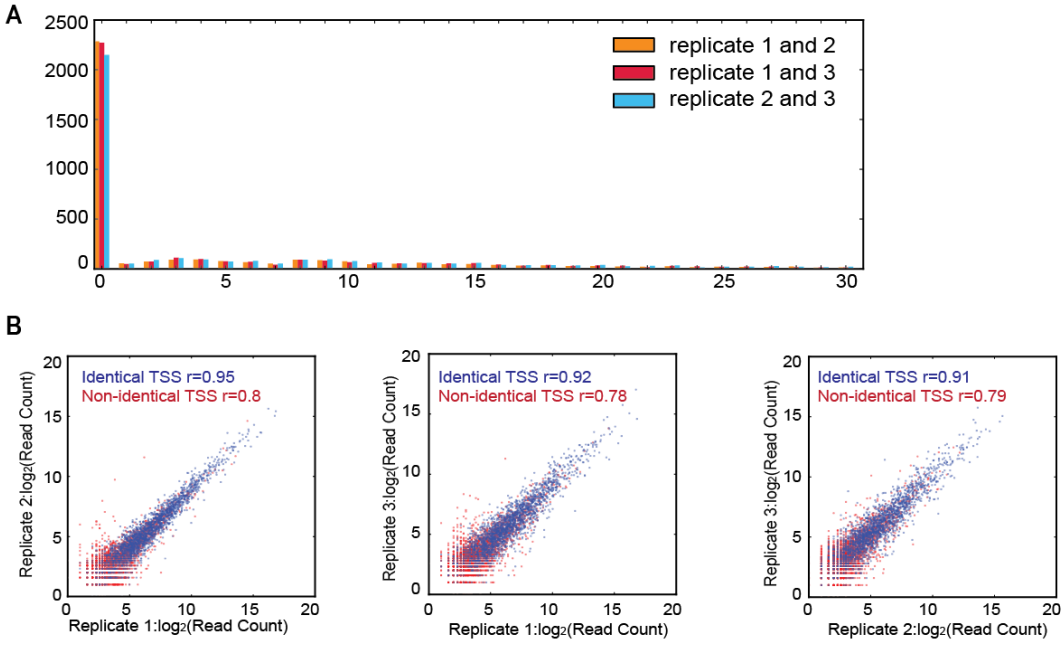

Figure 2.5 Highly reproducible SMORE-seq

(A) Absolute difference between TSS coordinates among SMORE-seq replicates shows high reproducibility of TSS calls at single nucleotide resolution. (B) Correlation of read counts at TSS between SMORE-seq replicates. Identical TSS have higher read counts and correlation, indicating that non-identical TSS are caused by low read counts of low abundance mRNAs.

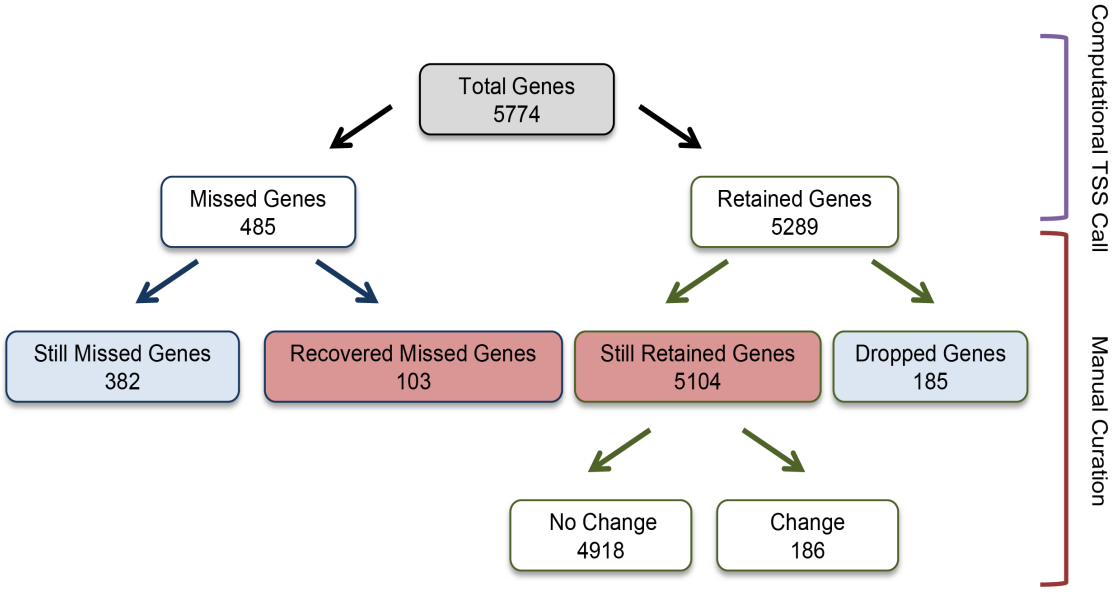

Figure 2.6 Flowchart of TSS calling

Out of 5774 verified and uncharacterized genes, TSS was computationally defined for 5289 genes. All 5774 genes were visually inspected for possible manual adjustments if necessary. We were not able to call TSS for 382 genes due to low coverage or correction by TAP-. TSS for 103 genes were manually assigned mainly due to long 5'-UTRs > 300 nt. 185 genes were dropped because they had few reads (<10) that were evenly distributed across the promoter and ORF or because proximal/overlapping neighboring genes led to ambiguity in TSS calling and assignment. The TSS coordinate of 186 genes was corrected during manual curation, with incorrect TSS calls mostly due to either long 5'-UTRs or proximity to other genes.

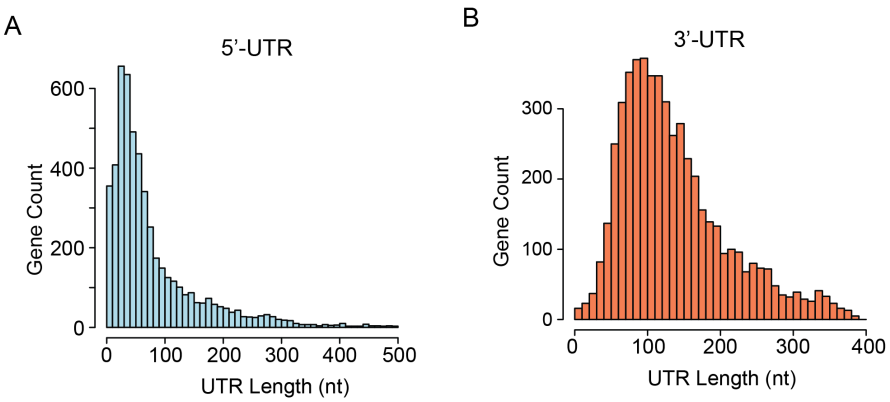

Figure 2.7 Histogram of 5'-UTR and 3'-UTR lengths estimated from S-TSS and PAS

The median and mean 5'-UTR lengths are 52 and 84 nt, respectively (n= 5203). The number of genes with 5'-UTR longer than 500 nt is 41. The median and mean 3'-UTR lengths are 120 and 137 nt, respectively (n= 5277).

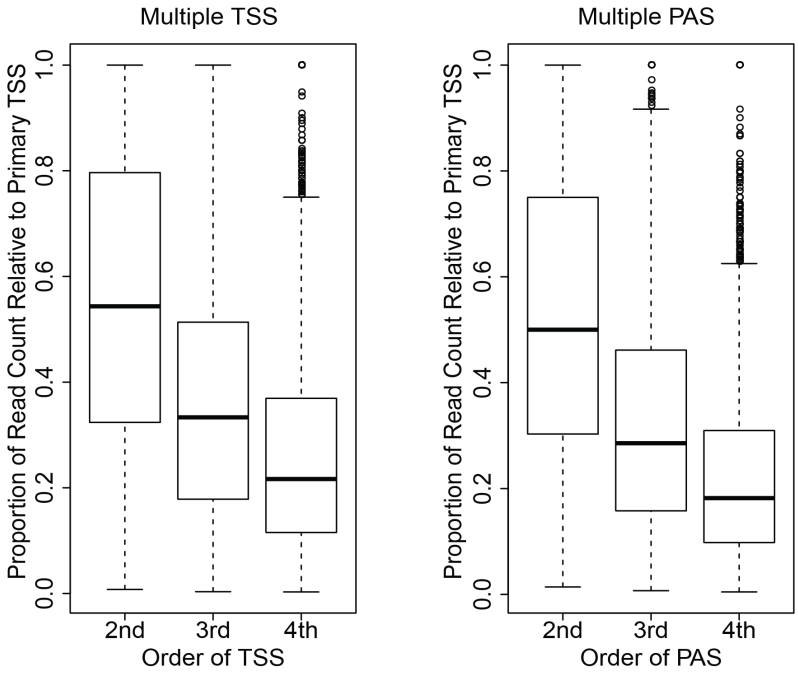

Figure 2.8 Relative utilization of multiple (alternative) TSS and PAS

The relative utilization of 2nd, 3rd or 4th strongest TSS (or PAS) compared to the primary (1st) TSS (or PAS) is plotted in the standard box plot in terms of read counts. The Y axis shows the proportion of read counts in the alternative TSS (or PAS) compared to the primary TSS (or PAS).

The major cause of non-identical TSS calls between replicates was low read coverage in genes with low expression (Figure 2.5B); therefore to increase coverage, we combined all replicates, then identified TSS again as described above using the combined datasets. These computationally-defined TSS were further manually curated by visual inspection of the raw read data in our UCSC genome browser mirror. For a small fraction of cases (289/5207, or 5.6%), our computational procedure had missed the TSS that was evident by visual inspection of the data; these were therefore manually corrected (Figure 2.6). This rate of manual correction is significantly lower than in previous studies [35], and could potentially be reduced further by incorporating steps in our algorithms tailored to address the main reasons for erroneous assignment that we observed during curation. Based on our TSS annotations, we determined that the median and mean 5'-UTR lengths in yeast are 52 and 84 nt, respectively (n= 5203, Figure 2.7A). Thus, SMORE-seq provides a systematic framework to reproducibly identify TSS with single-nucleotide resolution in a largely automated manner. These and all subsequent analyses in this study are based on the primary TSS that we identified for each gene. However, we also used our data to determine the extent of utilization of additional TSS for a given gene. As expected, secondary TSS tend to be used less than the primary TSS, but their distribution of relative utilization varied over a broad range, indicating that for some genes, the secondary TSS are used at rates comparable to the primary one (Figure 2.8).

2.4.2 SMORE-seq TSS and other transcriptional features

Currently, the most complete and widely utilized yeast TSS annotations are based

on the study of Xu et al [35, 49, 77, 78], because the data are strand specific, manually curated, replicated several times, generated under various perturbation conditions, and cover almost all genes. We therefore assessed the accuracy of the TSS from SMORE-seq (S-TSS) by comparison to the TSS from Xu et al (X-TSS). The S-TSS and X-TSS were generally in agreement, with 80% of the S-TSS located within 40 bp of the X-TSS (Figures 2.9).

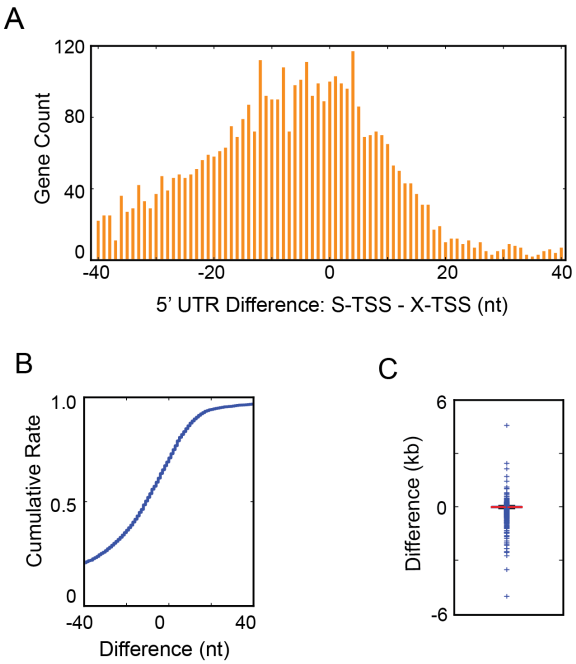

Figure 2.9 Comparison of SMORE-seq TSS with the commonly referenced TSS

SMORE-seq TSS (S-TSS) is compared with the commonly referenced TSS coordinates reported by Xu et al (X-TSS) [35] by histogram (A), cumulative distribution (B) and box plot (C), demonstrating that S-TSS and X-TSS are in high agreement. Overall, 5'-UTRs of S-TSS are shorter (S-TSS are more downstream).

Globally, 5'-UTRs from S-TSS were shorter than X-TSS by a median of 11 nucleotides. Our algorithm was designed to pick the nucleotide position with the

strongest read signal as the TSS whereas Xu et al picked the upstream coordinate of the 8 nt tile containing the most upstream signal for a given gene as its TSS. Because of this systematic difference, the finding that our S-TSS were closer to the start codon with a median difference of 11 nt is likely due to the improved accuracy of our TSS calls. However, to independently verify the accuracy of S-TSS, we evaluated both sets of TSS calls with regard to TATA element positions, consensus sequences at TSS, nucleosome positions near the TSS, and localization of active RNAPII phosphorylated at Serine 5.

Interaction between TATA-boxes or TATA-like elements in promoters and general transcription factors serves to recruit RNAPII and initiate transcription some distance away [79]. Distances between the TSS and canonical TATA boxes are believed to be distributed in a narrow range of 45-125 bp for most yeast genes [58, 80]. Compared to X-TSS, S-TSS showed a narrower distance distribution from canonical TATA boxes (n=716) (Figure 2.10A and 2.10B). A similar pattern was also observed for the distance between TATA-like elements and the TSS in TATA-less genes (n=4065) (Figure 2.10C and 2.10D). Together, these results suggest that initiation of transcription within a narrow distance window from TATA elements generates the sharper distribution of distances between TATA elements and S-TSS, supporting the higher accuracy of SMORE-seq. A consensus sequence of PyA has previously been identified at TSS in yeast [58, 80]. This sequence was readily identifiable in TSS identified by SMORE-seq, but could not be identified using X-TSS coordinates or a typical RNA-seq data set (Figure 2.11A).

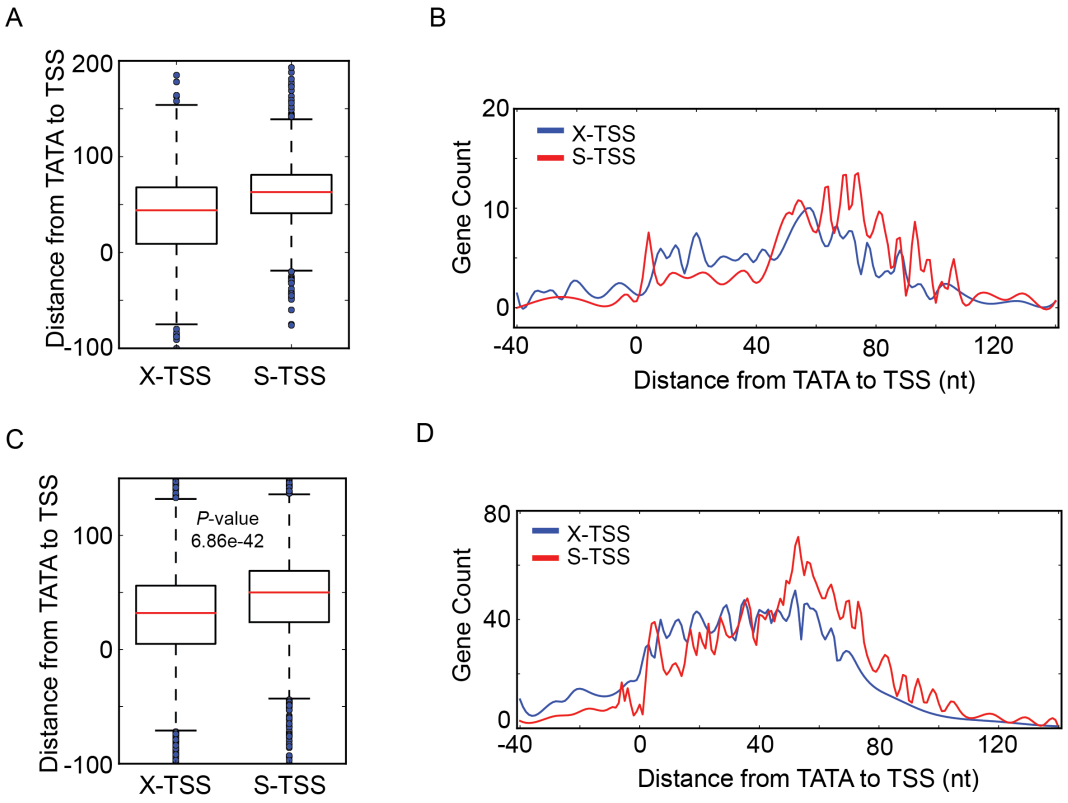

Figure 2.10 Distance between TATA elements and either S-TSS or X-TSS

(A-B) Distance between canonical TATA box motifs (n=716) [19] and either S-TSS or X-TSS demonstrates a narrower distance distribution from TATA boxes to S-TSS. (C-D) Distance between TATA-like elements in TATA-less genes (n=4065) [19] and S-TSS or X-TSS. S-TSS shows a narrower distribution with a larger average distance.

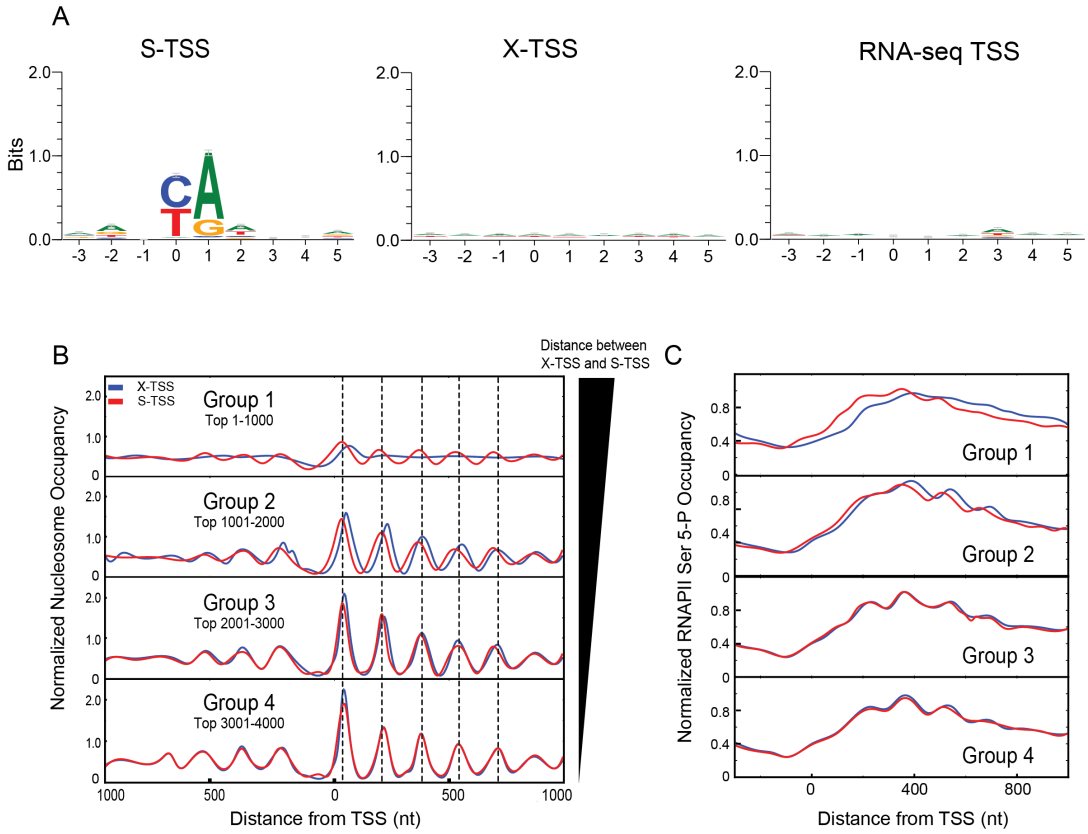

Figure 2.11 High resolution and accuracy of S-TSS

(A) Consensus sequence around TSS identified by SMORE-seq, Xu et al., and RNA-seq data [35, 60], visualized by WebLogo [81]. The consensus motif identified in S-TSS matches what has been previously described [58]. (B) Nucleosome occupancy profiles relative to the TSS in each of 4 groups of 1000 yeast genes, arranged by the distance between S-TSS and X-TSS in descending order. Nucleosome positions relative to X-TSS (blue line) differ between the groups whereas their positions relative to S-TSS (red line) are constant. Nucleosomes also show the expected periodicity in group 1 (top) relative to S-TSS but not X-TSS. (C) Localization of RNAPII Ser 5-P in the same groups as in 'B'.

A core promoter in yeast is situated within a nucleosome-depleted region (NDR) and is followed by a well-aligned array of nucleosomes starting from the TSS [82]. This property of nucleosome organization allowed us to test the accuracy of TSS calls by examining their relationship to nucleosome profiles. We generated nucleosome

occupancy maps using MNase-seq and plotted their profiles for each of four groups of 1000 genes formed in descending order of absolute difference between S-TSS and X-TSS coordinates (Figure 2.11B). Interestingly, the centers of the nucleosomes relative to the TSS did not change between these gene groups when using S-TSS coordinates. In contrast, when using X-TSS coordinates, nucleosomes appeared to be shifted downstream with decreasing rank of the gene groups. Because the rank of the groups had no prior relationship to nucleosome positions, the S-TSS coordinates, which yielded a similar nucleosome occupancy pattern across all four groups are likely to be more accurate. Moreover, the nucleosome profile in group 1 was flatter and poorly defined relative to X-TSS, whereas the S-TSS coordinates showed a more characteristic NDR and nucleosome periodicity. Thus, the inaccuracy of X-TSS leads to lower definition and offset of nucleosome occupancy profiles for a subset of genes.

Phosphorylation of RNAPII at Serine 5 (Ser 5-P) is a marker of transcription initiation and early elongation [83]. RNAPII Ser 5-P occupancy is therefore expected to start at the TSS and increase toward mid-ORF. We used ChIP-seq to measure the localization of RNAPII Ser 5-P relative to S-TSS and X-TSS, in the same four groups of genes as for the nucleosome analysis above. RNAPII Ser 5-P occupancy increased from the TSS to 200 bp downstream in all four groups, but as with the nucleosome profiles, its pattern of occupancy relative to S-TSS was more invariant than the occupancy relative to X-TSS (Figure 2.11C). Thus, S-TSS shows a more consistent relationship with a genome-wide mark of transcription initiation. Taken together, these analyses show that

the genome-wide TSS identified by SMORE-seq are not merely more downstream than TSS identified by other global methods, but show more clear-cut relationships to biological features of transcription initiation and are therefore likely to be more accurate.

2.4.3 SMORE-seq identifies mis-annotated start codons

We identified 222 genes with TSS downstream of their annotated ATG start codons: we refer to these as internal TSS. We defined the putative start codon of these genes as the first ATG downstream of the TSS. Of the 222 genes, 127 had a putative start codon in frame with the annotated ORF, 91 had a putative start codon out of frame with the annotated ORF, and 4 had no start codon between the TSS and the annotated stop codon. We reexamined these 95 genes with an out of frame or no start codon and flagged 72 genes because they either had a secondary, well-represented upstream TSS that agreed with the SGD start codon, an apparently incorrect TSS call, or low, ambiguous signal that prevented a confident TSS call. The 23 genes that were not flagged were grouped with the 127 that contained an in frame start codon, and the 123 verified genes out of this combined group of 150 were used for further analysis.

Although previous studies have reported internal TSS and confirmed several by quantitative PCR (qPCR) and primer extension assays [38, 58, 60], the veracity of such internal TSS, which would be indicative of potentially mis-annotated protein coding regions, has not been systematically evaluated. We evaluated whether these internal TSS were indeed the bonafide TSS by examining the propensity of such genes to have an alternative start codon downstream of the annotated start codon, evolutionary

conservation, ribosome footprinting profiles, and presence of a preferred Kozak consensus sequence for translation initiation.

We observed that internal TSS genes tended to have an in-frame methionine codon just downstream of the internal TSS (Figure 2.12). For most genes, the likelihood of having an internal methionine downstream of the annotated start codon is expected to increase monotonically with increasing distance from the start codon. Indeed, all verified genes showed this expected pattern (Figure 2.13A). However, internal TSS genes showed a markedly steeper increase, indicative of a higher likelihood of having another methionine shortly downstream of the annotated start codon. This distinctive behavior suggests that the internal TSS of this subset of genes could be the true TSS, with translation initiating from an internal methionine to generate a polypeptide that is truncated at the N-terminus relative to the currently annotated protein coding sequence.

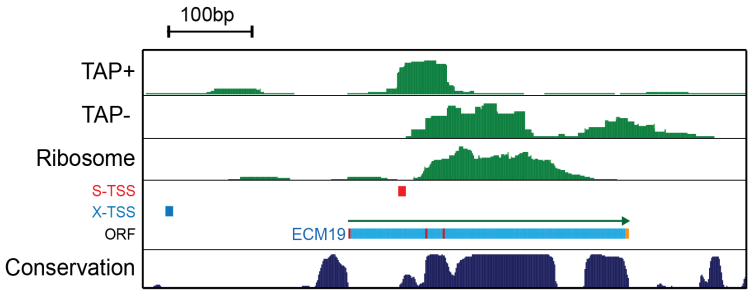

Figure 2.12 Example of an internal TSS downstream of the annotated start codon

Ribosome footprinting and conservation score are visualized in the UCSC Genome Browser mirror [76, 84]. In-frame methionine codons are indicated in red within the ORF track.

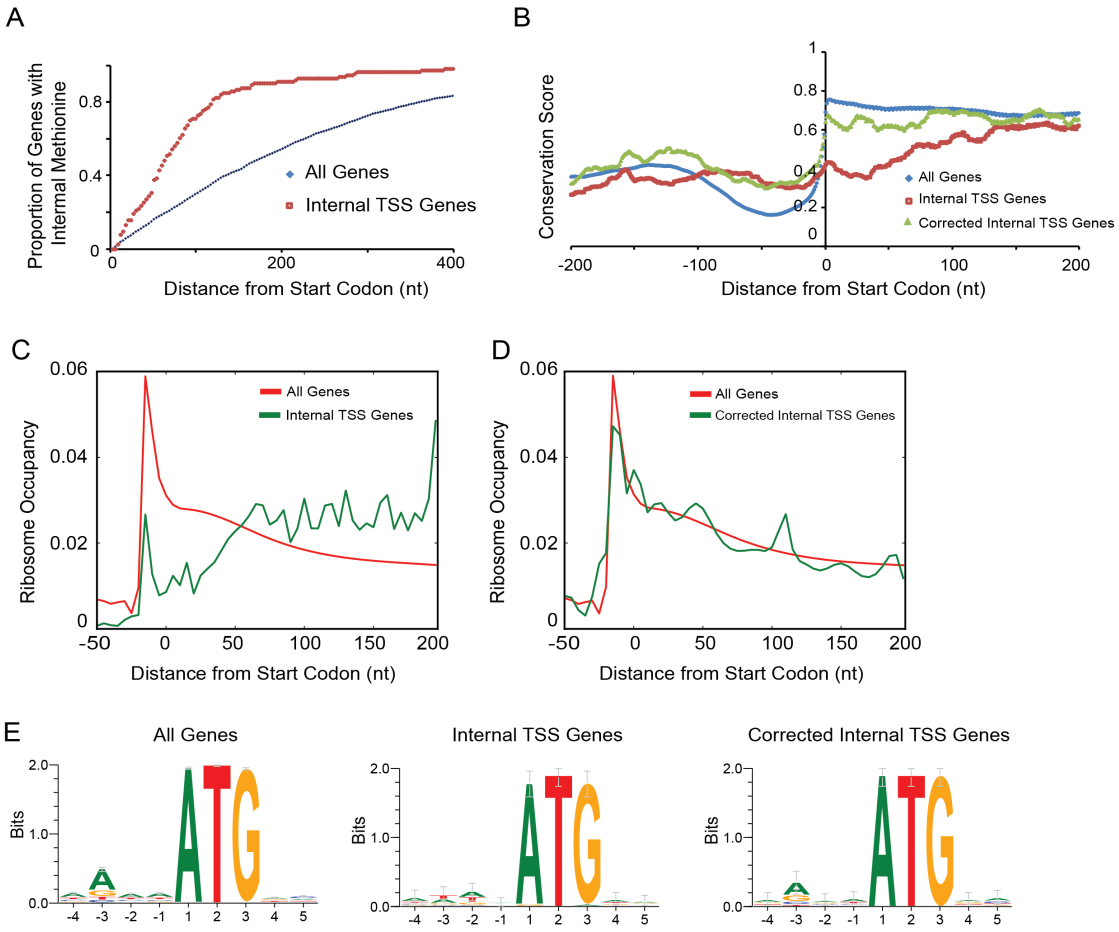

Figure 2.13 Mis-annotated start codons identified by SMORE-seq

(A) Cumulative proportion of genes that have an in-frame methionine at the indicated distance from the SGD annotated start codon, for each of the indicated groups. (B) Seven-species yeast conservation near start codons of all verified genes according to SGD annotations (All Genes), internal TSS genes according to SGD (Internal TSS Genes), and internal TSS genes with start codon predicted based on SMORE-seq (Corrected Internal TSS Genes). (C-D) Ribosome profiles near start codons as predicted in 'C'. Ribosome profiling data was taken from [76] and plotted as the average proportion of reads. (E) Consensus sequence upstream of the start codon of the indicated gene sets, where the start codon used was as described in 'C'.

Protein-coding regions of yeast genes show significantly higher evolutionary conservation than non-coding regions [84-86]. To determine if this conservation could shed light on the potential usage of internal TSS, we analyzed conservation around the start codon between seven yeast species. The set of all genes showed a sharp increase in conservation downstream of the start codon. This increase in conservation was not seen in the internal TSS genes when using the SGD start codon (Figure 2.13B). However, if we used the first methionine downstream of our internal S-TSS as the start codon, conservation just downstream of the start was restored for this set of genes. This data strongly suggests that the internal methionine downstream of the internal S-TSS is the true start of the protein coding region for these genes, rather than the currently annotated start codon.

Next, we analyzed published genome-wide ribosome footprinting data to obtain experimental evidence regarding translation at either annotated or internal start codons [76]. Ribosome footprinting measures occupancy of ribosomes along mRNAs, and has shown that there is high ribosome occupancy 12-13 nt upstream of start codons [76]. We analyzed ribosome footprints from the previously published study in the three groups of genes described above. The set of all genes showed a strong ribosome occupancy peak 12-13 nt upstream of the start codon. This peak was largely absent near the SGD-annotated start codons of internal TSS genes (Figure 2.13C), but was clearly restored when we used start codons downstream of the internal TSS predicted by SMORE-seq (Figure 2.13D). This analysis provides strong evidence of the accuracy of SMORE-seq

TSS coordinates and start codon predictions for internal TSS genes.

Consensus sequence analysis of the regions near annotated start codons for all genes showed strong enrichment of A residues at the -3 position relative to the ATG start codon, which is a characteristic of the Kozak consensus sequence in yeast [87, 88] (Figure 2.13E). In contrast, enrichment of A at the -3 position was not observed for internal TSS genes, indicating that the annotated start codons are unlikely to be used for translation initiation. Strikingly, the Kozak consensus sequence was restored at the corrected, internal start codon for the internal TSS genes. Thus, start codons predicted by SMORE-seq for internal TSS genes have a more appropriate sequence context for initiation of translation than the current SGD annotations.

2.4.4 SMORE-seq identifies polyadenylation sites

Visual inspection of SMORE-seq data indicated a large number of reads mapped to 3' regions of mRNAs, near ORF stop codons (Figure 2.3 and 2.15). Because of the 3' bias of these reads and their abundance in both TAP+ and TAP- samples, we hypothesized that these reads originated from mRNA degradation products. One of the main mRNA degradation pathways in eukaryotes starts with shortening of poly(A) tails to ~10-15 A bases, followed by decapping and 5'-3' exonuclease-mediated degradation [89]. These degradation products have a 5' phosphate that is amenable to ligation during the SMORE-seq protocol, and thus would be represented in both TAP+ and TAP- samples. Because some of these reads would also be expected to contain the PAS, the site where the mRNA is cleaved and an untemplated stretch of A residues is added, we

reasoned that these reads could be used to map PAS.

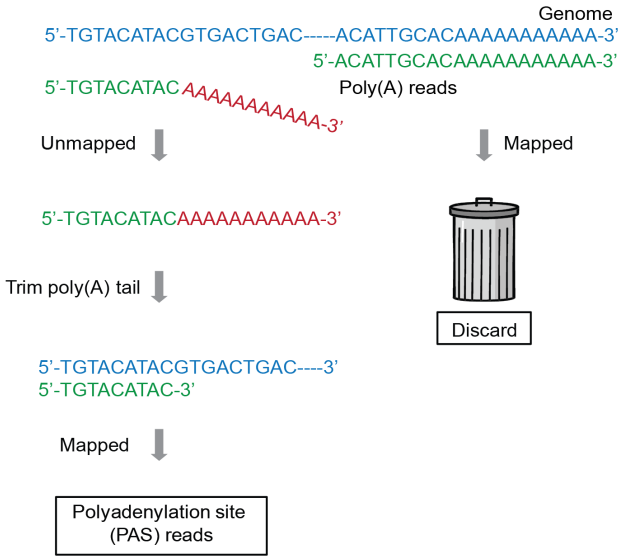

Figure 2.14 Strategy used to extract PAS containing reads

Strategy used to extract PAS containing reads, those with a 3' stretch of untemplated As, from SMORE-seq data. Reads that ended in a stretch of A residues were selected, and those that mapped to the genome only after removal of the 3' poly(A) stretch were retained as PAS reads.

We used a simple but effective workflow to obtain reads representing potential PAS in our data (Figure 2.14). We first selected all reads ending in a string of As (see Methods). We then mapped these reads to the yeast genome and sorted the results into unmapped or mapped groups, with the expectation that reads with an untemplated stretch of As, representing a potential PAS, would be unmapped, whereas those reads that mapped represented a genomic poly(A) stretch and should be discarded. We then trimmed the poly(A) stretch off the unmapped reads and mapped these trimmed reads again, with the expectation that the reads that mapped after trimming represented PAS. This set of reads, which we called PAS reads, mapped almost exclusively to likely 3'-

UTR regions of mRNAs (Figure 2.15), indicating that our strategy was effective in identifying PAS. This procedure yielded a total of 55,419 candidate PAS where each PAS was defined by at least 2 reads. In order to identify a dominant PAS for each gene, we determined the base position with the highest read stack in the PAS reads in the range from the gene's stop codon to 300 bases downstream. We were able to identify a PAS for 5,277 (91%) yeast genes using this strategy. Based on SMORE-seq PAS annotations, the median and mean 3'-UTR lengths in yeast are 120 and 137 nt, respectively (n= 5277, Figure 2.7B).

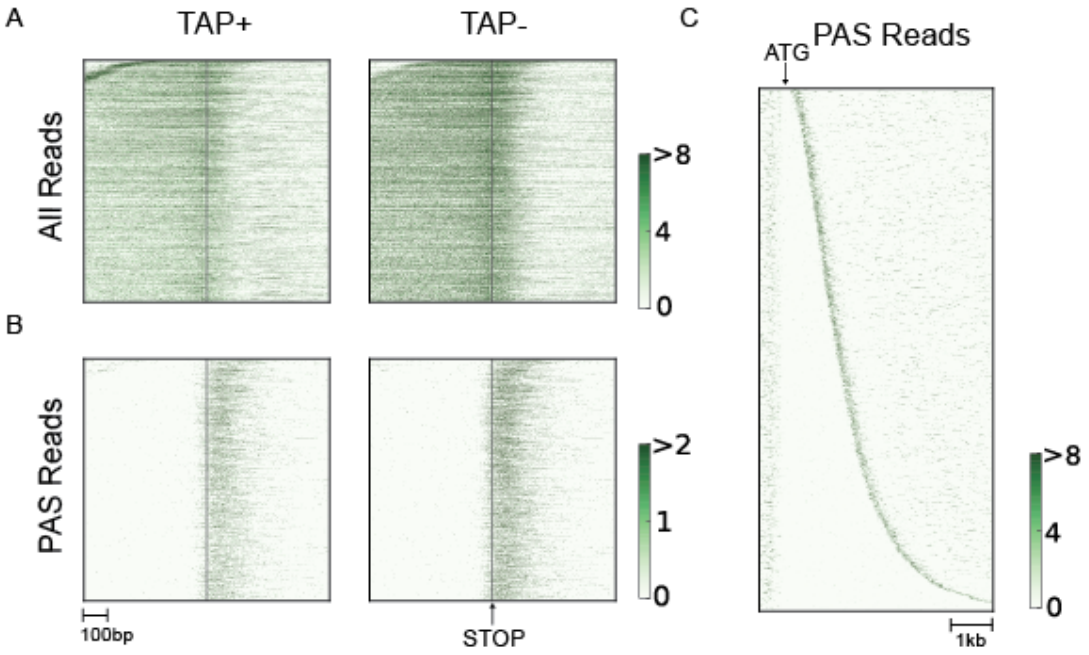

Figure 2.15 Heat map representation of PAS reads from SMORE-seq

SMORE-seq reads near gene stop codons (vertical line) before and after applying filtering described in A. PAS reads mostly mapped just downstream of stop codons. (D) PAS reads in all genes sorted by ORF length and aligned by start codon (arrow), demonstrating that few PAS reads mapped within ORFs.

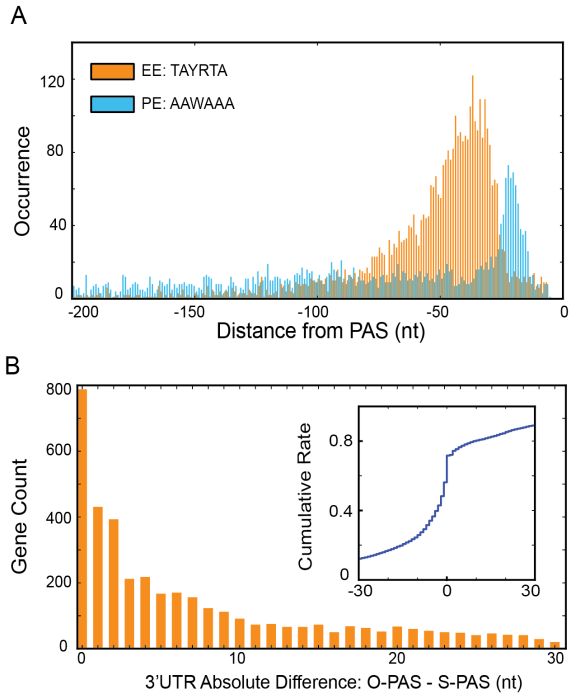

Figure 2.16 High resolution and accuracy of S-PAS

(A) Occurrence of the polyadenylation efficiency element (EE) and positioning element (PE), elements utilized for PAS selection, relative to PAS identified by SMORE-seq. (B) Difference between SMORE-seq PAS and those identified by Ozsolak et al. [66] using Helicos NGS-based method. The inset shows the cumulative difference profile.

Sequence elements that contribute to PAS selection have been discovered in yeast, and although these elements are less conserved and less well-defined than in higher eukaryotes, a positioning element (PE) with sequence AAWAAA and an efficiency element (EE) with sequence TAYRTA have been identified about 10-30 nt and 25-75 nt upstream of PAS, respectively [90]. A search for these elements in the sequences surrounding PAS as determined by SMORE-seq revealed enrichment of these sequences with expected positioning relative to PAS, indicating that SMORE-seq was successful in determining correct PAS (Figure 2.16A).

Polyadenylation sites have been previously measured in yeast with a specialized deep-sequencing based strategy [66]. To further verify the accuracy of SMORE-seq PAS, we compared our results to this study. In order to define PAS with single-nucleotide resolution from the published data, which reported PAS regions rather than a single base position, we downloaded their data and found the position with the highest read stack as described above (see Methods). We could identify a PAS for 5,314 genes in the published data, and of these genes, 5,119 also had a PAS identified by SMORE-seq. There was striking agreement between PAS identified by the two methods, with almost 80% of PAS within 30 bases and almost 800 genes showing an identical PAS between samples (Figure 2.16B). Thus, SMORE-seq can accurately map both TSS and PAS from the same sequencing dataset with single-nucleotide resolution. Similar to TSS, many genes also showed alternative PAS, which were used at rates lower than the primary PAS (Figure 2.8).

2.4.5 SMORE-seq reveals widespread bidirectional transcription

We observed more than a thousand regions where reads aligning in the opposite direction of the coding strand were concentrated in a region 50-300 bp upstream of the S-TSS, indicating non-coding RNA (ncRNA) transcripts resulting from bidirectional promoters. Previous studies have reported ncRNAs at bidirectional promoters only in strains deleted for genes associated with gene looping or the nuclear exosome [35, 91], as the directionality of transcription was thought to be tightly regulated and antisense ncRNAs rapidly degraded in wild-type (WT) strains. For example, the promoter

associated ncRNA at the bidirectional promoter between *OPY1* and *SHE3* was previously identified only in an *ssu72-2* mutant and therefore interpreted as arising due to disruption of a gene loop [91]. However, this RNA was readily identifiable by SMORE-seq in a wild-type strain, likely due to the higher sensitivity of our method (Figure 2.17A). SMORE-seq identified more than a thousand new bidirectional promoter-associated ncRNAs (Figure 2.17A). Here, we refer to the antisense ncRNAs detected by SMORE-seq at promoters as bncRNAs (bidirectional non-coding RNAs).

In order to visualize the prevalence of bncRNAs in WT cells under normal growth conditions, we separately plotted the SMORE-seq reads aligning to each strand near promoters, split according to the orientation of two adjacent genes. The terms "same" and "opposite" for read directionality are defined with respect to the downstream gene, and "tandem" and "divergent" define the orientation of the upstream gene (Figure 2.17B and 2.17C). Interestingly, opposite reads showed a strong signal 50-300 bp upstream from the S-TSS of the downstream genes (Figure 2.17B). In particular, the widespread signal from opposite reads in tandem genes, where this signal is unequivocally independent from the TSS of the upstream gene, shows that products of bidirectional transcription are much more pervasive than previously appreciated in WT yeast cells.

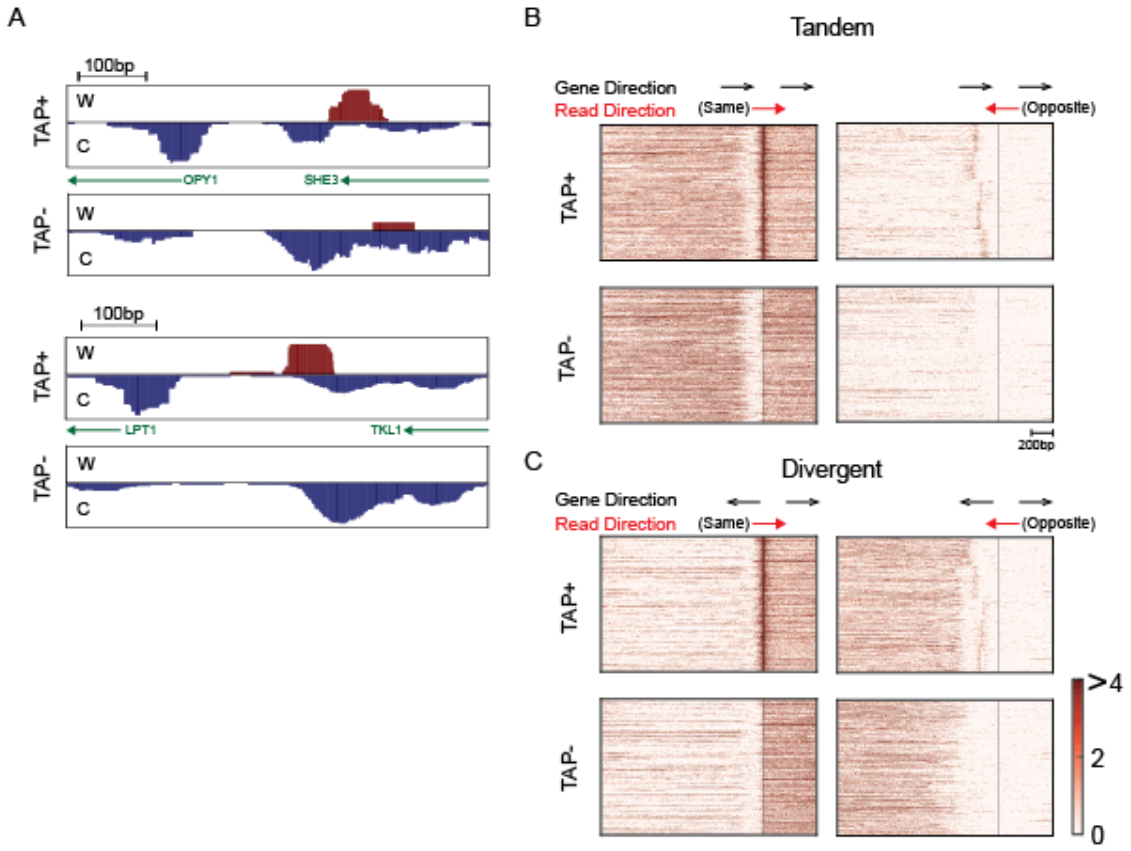

Figure 2.17 Widespread occurrence of bncRNAs

(A) A previously known ssu72-restricted transcript (SRT) in the promoter of *OPY1* is detected by SMORE-seq in WT cells under normal growth conditions [91] (top 2 panels). A novel antisense ncRNA that may share a bidirectional promoter with *LPT1* is shown below. (B,C) Widespread occurrence of bncRNAs (antisense ncRNAs at bidirectional promoters). Genes were clustered by K-means clustering (k=5, repeat=1000) of bnc signal in a range 300 bp to 50 bp upstream of TSS. 'B' shows genes in the indicated tandem arrangement, and 'C', in the divergent arrangement. Divergent genes whose TSS are closer than 300 bp are excluded in 'C' The vertical line represents the TSS of downstream genes. The number of tandem genes and divergent genes in this heat map are 2401 and 1635, respectively.

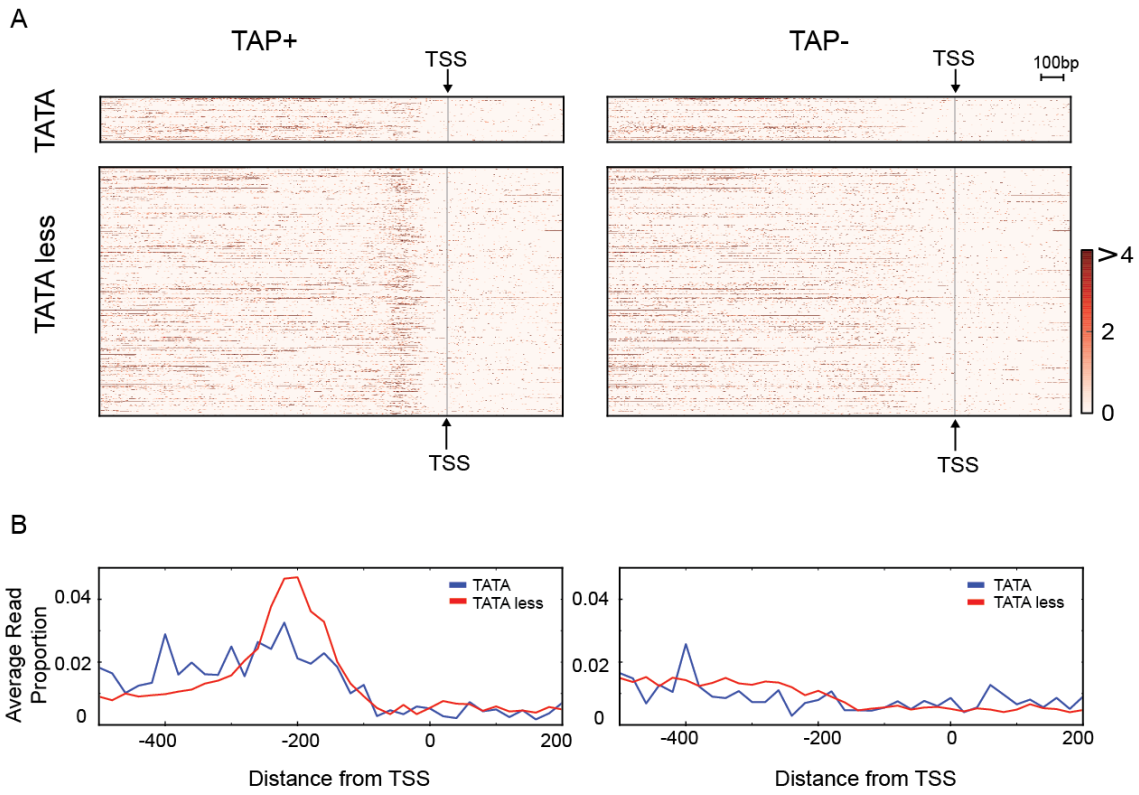

Figure 2.18 A canonical TATA-box element suppresses bidirectional transcription

(A) Opposite reads in tandem genes grouped by presence or absence of a canonical TATA-box in the gene's promoter. TATA-less tandem genes (n=2031) show stronger bncRNA signal than TATA-box containing tandem genes (n=370). (B) Average proportion of reads in this window, demonstrating that TATA-less genes have higher bncRNA expression. The *P*-value for the difference in bnc signal between TATA (blue) and TATA-less (red) signals at -200 was 3.67 X 10^{-7} by Welch's t-test.

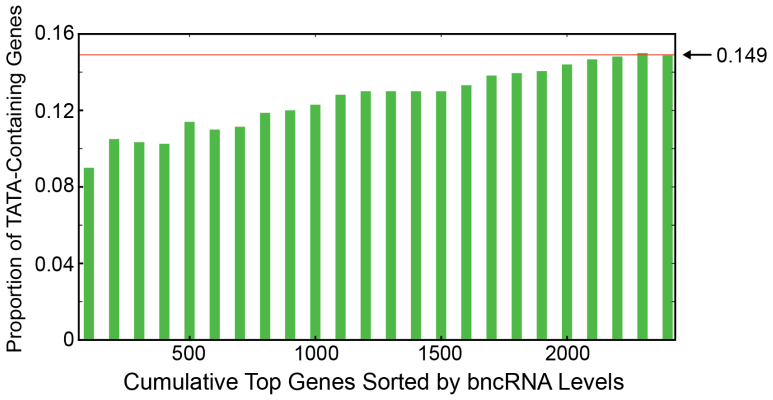

Figure 2.19 The proportion of TATA-containing genes related to levels of bncRNA

The proportion of TATA-containing genes is inversely related to levels of bncRNA transcription. Tandem genes were sorted by the strength of their bncRNA transcription read counts between -50 and -300 (X-axis). The proportion of TATA-containing genes in each group of 100 genes in this ranked set is plotted on the Y-axis. The overall proportion of TATA-containing genes among all tandem genes is 14.9 % (far right).

2.4.6 A canonical TATA-box element suppresses bidirectional transcription

Previous studies reporting the expression of promoter-associated ncRNAs in mutants defective in RNA processing have noted that highly expressed genes show higher levels of the promoter-associated non-coding RNAs [91]. In order to assess the correlation between the bncRNAs identified by SMORE-seq and downstream gene expression, we generated heat maps showing bncRNAs with their downstream genes sorted by mRNA abundance [92]. The intensity of the bncRNA signal did not appear to correlate with expression of the downstream gene. The correlation coefficient between levels of bncRNAs and downstream gene expression was close to zero (Spearman rank r = -0.02), indicating that expression of the downstream gene is unrelated to bncRNA levels.

Mutation of the TATA-box in the *TPI1* promoter has been reported to increase antisense transcription from its bidirectional promoter [93]. We hypothesized that the presence of a TATA-box in promoters correlates genome-wide with levels of bncRNAs. To test this hypothesis, we separated tandem genes based on whether the downstream gene contained a canonical TATA-box, then plotted reads arising from the opposite strand in a heat map (Figure 2.18A). The signal from bncRNAs in TATA-box containing genes was significantly lower compared to TATA-less genes (Figure 2.18B, $P = 3.67 \text{ X } 10^{-7}$ by Welch's t-test). Moreover, the proportion of TATA-containing genes was lower for genes with higher levels of bncRNA transcription (Figure 2.19). Thus, promoters lacking a canonical TATA box, or TATA-less promoters, have a higher chance of giving rise to a bidirectional non-coding RNA in the opposite direction. Additional evidence in favor of the TATA-box model for bncRNA transcription comes from nucleosome localization data. A well-positioned +1 nucleosome is believed to help form the pre-initiation complex and recruit RNAPII at TATA-less promoters [19, 94]. If bncRNA transcription uses the same mechanism as normal initiation, the -1 nucleosome with respect to sense genes could act as the +1 nucleosome with respect to bncRNA, and similarly facilitate bncRNA transcription. Supporting this hypothesis, TATA-less genes, which have high bncRNA expression, have well-positioned -1 nucleosomes (Figure 2.20A), and the highly expressed bncRNAs have a more well defined +1 nucleosome (Figure 2.20B).

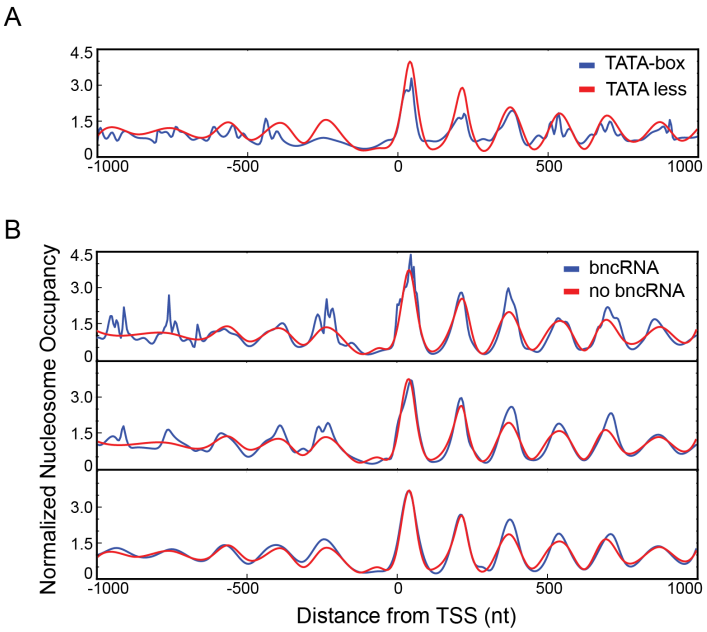

Figure 2.20 Average nucleosome profile of tandem genes

Y-axis represents read counts per million reads (bin size=5 bp) (A) TATA box containing versus TATA-less genes. (B) Highly expressed bncRNAs have more well defined -1 nucleosomes with respect to sense genes. bncRNA expression levels were defined as the sum of TAP+ reads between 300 bp upstream and 50 bp upstream of TSS. From top to bottom, cut-offs for defining bncRNAs are 50, 20, and 10 reads. The number of genes with bncRNAs are 287, 566, and 883.

2.5 Discussion

We have shown that the high accuracy and sensitivity of mapping transcript 5' and 3' ends using SMORE-seq reveals more well-defined relationships of transcript ends with cis-elements and chromatin structure, identifies widespread bidirectional transcriptional initiation, and suggests a novel role for a canonical TATA-elements in orienting transcription initiation. The singular advantage of SMORE-seq is that it can identify TSS and PAS using the same deep sequencing data derived from a single RNA-seq library,

allowing the investigation of transcription initiation as well as termination/polyadenylation in the same RNA sample. Despite this gain in efficiency, SMORE-seq is also a relatively simple method. Other comparable approaches to map TSS using next-generation sequencing, are generally more tedious. For example, CAGE (Cap analysis of gene expression), which has been adapted for deep sequencing, is a relatively cumbersome procedure that involves biotinylated oligos and contains 18-25 major steps spread over 8-14 days to generate a sequencing library [95]. Various NGS-based methods to map PAS have been recently utilized to map PAS [96]. Some PAS mapping methods involve the use of specialized primers, and others require deep sequencing technologies that are not commonly available [55, 66, 97]. While these methods map PAS with single-nucleotide resolution, they provide no data that can be used to map TSS. In contrast, SMORE-seq avoids any specialized primers and can be completed by one researcher in one day using standard reagents and deep sequencing kits. The improved efficiency of SMORE-seq will be valuable in situations where there is a limited amount of material available, such as human patient samples or microbial species that are difficult to propagate.

During preparation of this manuscript, a study using another method to simultaneously map TSS and PAS, TIF-seq, was published [70]. SMORE-seq and TIF-seq generate complementary data, but there are a few noteworthy differences. TIF-seq simultaneously sequences the TSS and PAS of the same mRNA molecule, whereas SMORE-seq identifies TSS and PAS separately for the same population of mRNAs. The

TIF-seq study provided a comprehensive catalog of all transcript ends and isoforms in yeast, but it did not provide a definitive annotation of the most prominent TSS and PAS for each gene, and therefore did not uncover the same biological insights about transcriptional regulation that we were able to with SMORE-seq. Although the two methods use a similar strategy to ligate a 5' adapter at mRNA cap sites, TIF-seq follows this step with reverse transcription using a modified oligo(dT) primer. This may result in several potential complications: 1) efficiency of reverse transcription will be biased toward shorter RNA molecules, resulting in overrepresentation of shorter mRNAs and under-representation of longer mRNAs in final libraries, 2) mRNAs with a high degree of secondary structure may not be efficiently reverse transcribed and therefore under-represented, 3) mis-priming with the modified oligo(dT) primer may result in improper PAS calls, and 4) intact full length mRNAs are likely to be rare in partially degraded RNA samples, such as those from human patient material. Points 1, 2, and 3 are addressed in SMORE-seq by direct ligation of sequencing adapters to both 5' and 3' ends of RNA molecules, whereas point 4 is a weakness of both methods. This weakness can be easily addressed in SMORE-seq by using ribosomal RNA depletion rather than poly(A) selection in the first step, and although the data would be noisier and contain more ncRNA signal, this could largely be addressed through deeper sequencing. Another minor weakness of the TIF-seq method is that 30 total cycles of PCR were necessary compared to just 18 cycles in SMORE-seq, likely due to the additional steps in the TIF-seq protocol. However, TIF-seq provides single molecule data that SMORE-seq cannot. We

compared transcript annotations generated by SMORE-seq with the major TSS and PAS sites identified in the TIF-seq study and found strong concordance between both methods (Figure 2.21A). This study also identified a set of genes with TSS downstream of the annotated start codon, similar to what we reported (Figure 2.13). There is strong and significant overlap of the two sets of genes with internal TSS genes (Figure 2.21B). We believe that the existence of these complementary methods will assist researchers by allowing them to choose the one best suited to their research goals and conditions.

It is noteworthy that the dominant TSS of at least 150 genes is downstream of the annotated start codon, resulting in protein sequences that differ from SGD annotations. In 127 of these genes the start codon predicted by SMORE-seq is in frame with the annotated start codon, resulting in truncation of the encoded proteins at the N-terminus, with implications for protein function and construction of N-terminal fusion derivatives in experimental studies. For 22 genes our predicted start codon is not in frame with the annotated start codon, resulting in either a protein with a completely different sequence or a short ORF that is unlikely to encode a functional protein. Interestingly, the TSS and predicted start codon are very close in many of these genes, which may prevent the ribosome from binding to this ATG and allowing initiation of translation at a downstream ATG that is in frame with the annotated protein. Another possibility is that these loci encode non-coding RNAs with regulatory, enzymatic, or structural function.

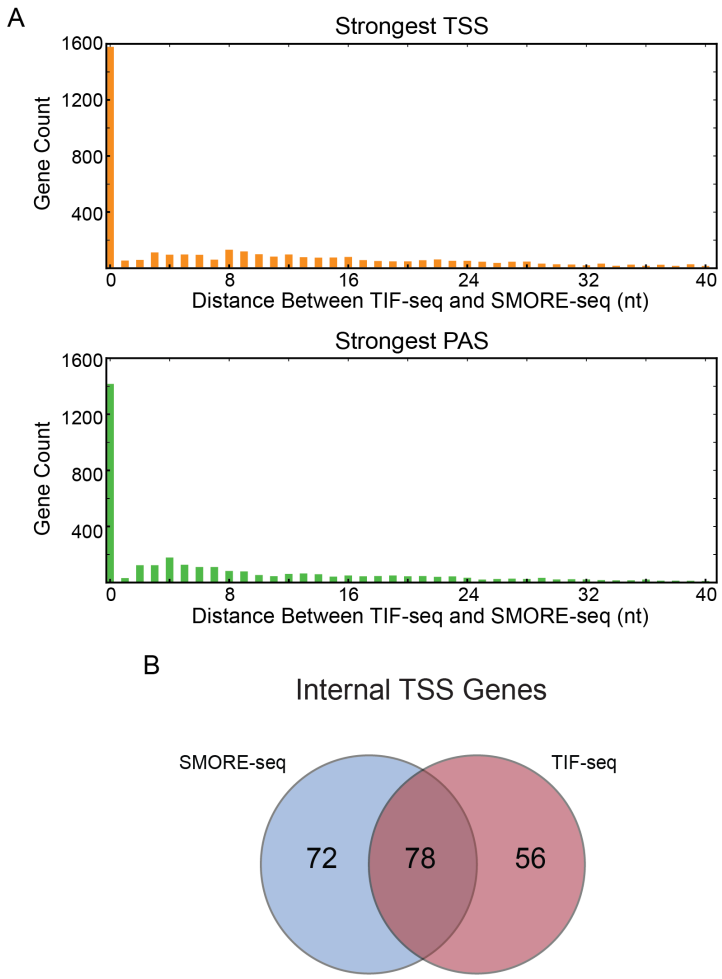

Figure 2.21 Comparison of SMORE-seq annotations with TIF-seq.

(A) Data for major transcript isoforms (mTIFs) covering one intact ORF from the TIF-seq data [70] was obtained and the most prominent TSS and PAS from this data was identified. The threshold of mTIF reads was set at ≥ 2. The plot shows the histogram of differences between SMORE-seq TSS or PAS and the corresponding mTIF coordinates. (B) Overlap of internal TSS genes. The figure shows the overlap between the 150 genes identified by SMORE-seq as internal TSS genes, and the 134 genes identified by TIF-seq where the TSS inside the ORF is at least 50% stronger than the TSS upstream of the ORF based on read counts, in YPD-grown cells. The P-value for the overlap is 2.4×10^{-95} by hypergeometric test.

The enrichment in SMORE-seq data of reads at the 3' ends of mRNAs likely results from the sequencing of degradation products created by deadenylation and de-capping dependent 5' to 3' degradation. mRNA poly(A) tails are shortened to ~10-20 A residues by the Ccr4-Caf1 deadenylase complex, followed by decapping by Dcp1-Dcp2 and 5' to 3' exonucleolytic degradation by Xrn1 [89]. Although reads resulting from such degradation products might be expected to map along the entire length of the mRNAs, we propose two explanations for the observed 3' enrichment of reads: 1) Short poly(A) tails of degradation products do not support hybridization of long mRNA degradation products to oligo(dT) beads during poly(A) selection, and/or 2) kinetics of degradation result in accumulation of smaller degradation products. Either of these scenarios would result in the observed abundance of reads representing 3' regions and polyadenylation sites of mRNAs. Notably, the presence of these reads in almost all genes indicates that degradation of the vast majority of yeast mRNAs depends at least partially on decapping and 5' to 3' decay, although further experimentation will be needed to confirm this hypothesis. It is also noteworthy that other TSS mapping methods treat RNA with a phosphatase enzyme before TAP [61, 62], but we were able to recover degradation intermediates used to map PAS only because we did not use phosphatase pre-treatment.

Several previous studies have reported antisense ncRNAs [35, 93], but their transcriptional regulatory mechanisms are largely unknown. The observation that bncRNAs were detected in TAP+ samples but not in TAP- (Figure 2.17) strongly indicates that they are 5'-capped. The presence of these RNAs following poly(A)

selection also indicates either that these RNAs had poly(A) tails or that they were recovered via hybridization to sense transcripts during poly(A) selection. A recent study indicates that bidirectionally transcribed, promoter-associated RNAs are indeed polyadenylated in human cells [98], supporting the former possibility. However it is not known whether this is also true in yeast. One study suggested that highly expressed genes also show higher levels of promoter ncRNA transcription, although the evidence for this relationship was modest [91]. Another model suggested that a TATA-box in a sense promoter could suppress antisense transcription [93]. Since highly transcribed genes in yeast generally contain a canonical TATA-box within their promoter [18], these two models are contradictory. We observed no correlation between bncRNA and sense RNA abundance, but we did observe high expression of bncRNAs in TATA-less promoters of sense genes (Figure 2.18), supporting the latter model. The low correlation between bncRNA and sense RNA abundance is consistent with previous studies showing that distinct pre-initiation complexes are responsible for sense and antisense transcription, and that antisense transcripts are independently regulated [19, 99, 100]. The relationships that we observed between TATA elements, nucleosomes and bncRNAs support a model where the presence of a TATA-box strongly influences the directionality of transcription. We anticipate that the use of SMORE-seq in conjunction with other genomic assays of chromatin structure in different species and cellular states will shed new light on the genome-wide mechanisms of transcriptional control.

Chapter 3 Widespread Misinterpretable ChIP-seq Bias

3.1 Abstract

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is widely used to detect genome-wide interactions between a protein of interest and DNA *in vivo*. Loci showing strong enrichment over adjacent background regions are typically considered to be sites of binding. Insufficient attention has been given to systematic artifacts inherent to the ChIP-seq procedure that might generate a misleading picture of protein binding to certain loci. We show here that unrelated transcription factors appear to consistently bind to the gene bodies of highly transcribed genes in yeast. Strikingly, several types of negative control experiments, including a protein that is not expected to bind chromatin, also showed similar patterns of strong binding within gene bodies. These false positive signals were evident across sequencing platforms and immunoprecipitation protocols, as well as in previously published datasets from other labs. We show that these false positive signals derive from high rates of transcription, and are inherent to the ChIP procedure, although they are exacerbated by sequencing library construction procedures. This expression bias is strong enough that a known transcriptional repressor like Tup1 can erroneously appear to be an activator. Another type of background bias stems from

---

This work was published in Park D., Lee Y., Bhupindersingh G. & Iyer V.R. Widespread Misinterpretable ChIP-seq Bias in Yeast (2013) *PLoS One* 8(12): e83506. DP, YL, and VRI conceived and designed the experiments. YL performed ChIP-seq under 30°/39°C conditions and Hsf1 ChIP-seq, and YL and GB conducted mock ChIP qPCR. DP performed all the other experiments and analyzed the data. YL wrote the figure legends for mock ChIP qPCR data and "Materials and Methods", except for the subsections of "Deep sequencing data analysis" and "Mock and input comparison". DP and VRI wrote the rest of the manuscript. Permission to adapt the contents of the publication was acquired from the co-authors.

the inherent nucleosomal structure of chromatin, and can potentially make it seem like certain factors bind nucleosomes even when they don't. Our analysis suggests that a mock ChIP sample offers a better normalization control for the expression bias, whereas the ChIP input is more appropriate for the nucleosomal periodicity bias. While these controls alleviate the effect of the biases to some extent, they are unable to eliminate it completely. Caution is therefore warranted regarding the interpretation of data that seemingly show the association of various transcription and chromatin factors with highly transcribed genes in yeast.

3.2 Introduction

The genome-wide mapping of protein localization on chromatin at high resolution is crucial for understanding the molecular mechanisms of transcription *in vivo*. Chromatin immunoprecipitation (ChIP) followed by deep sequencing (ChIP-seq) is currently the preferred and widespread method to accomplish this [51, 101, 102]. Because of the power of the ChIP assay, the Encyclopedia of DNA Elements (ENCODE) and Roadmap Epigenome Projects have adopted ChIP-seq to map the genomic locations of many transcription factors, histone marks, and DNA modifications in both cell lines and model organisms [103-106]. Because the localization of chromatin-associated factors is dependent on cell type and environmental conditions [8, 107], ChIP-seq is being increasingly used to explore hundreds of DNA-binding proteins in different types of cells and under different conditions.

Yeast is the first and only eukaryote for which nearly every transcription factor has been ChIP-ed and for which the resulting immunoprecipitated DNA has been mapped on a genome-wide scale using microarrays [108, 109]. With the advent of deep sequencing technology, ChIP-seq also has been broadly applied to yeast genomics [42, 71, 110]. Yeast is ideal for comprehensive studies on protein-DNA interactions due to its relatively small genome, the resulting low cost of experiments, and the availability of a tandem affinity purification (TAP)-tagged collection for 80% of its proteins [47]. This latter benefit is of particular importance, as TAP-tagged strains do not suffer from the same non-uniform quality as antibodies, whose variability can affect the efficiency of ChIP.

Several algorithms have been developed to computationally identify peaks of enrichment in ChIP-seq data, indicative of protein binding locations, and to distinguish such peaks from background reads [101, 111]. Experimentally and computationally, the background signal is typically defined using either a parallel input sample which has not been subject to the immunoprecipitation step, after reversal of crosslinks, or a mock ChIP sample (where a non-specific IgG antibody, or pre-immune serum, or an untagged strain is used).

In the course of carrying out ChIP-seq experiments for various yeast transcription-related proteins, we unexpectedly found strong enrichment signals suggestive of proteins binding to genomic loci where genes were highly transcribed, regardless of which protein was being analyzed. The functions of the genes exhibiting this universally high protein occupancy however did not always align with the established roles of the proteins

apparently binding to them. Moreover, the enrichment for proteins binding to highly-transcribed genes was observed even in controls like mock ChIP-seq data, which points to an overall bias that could contaminate any ChIP-seq data with false positives. A secondary bias of nucleosomal periodicity was also commonly observed across ChIP-seq datasets and contributed additional false positives in which proteins falsely appeared to interact with nucleosomes. We present our analysis of this phenomenon, and suggest ways in which these artifacts can be ameliorated by the proper choice of control experiments. Our data suggest however that the enrichment bias at highly transcribed genes could be an intrinsic characteristic of ChIP-seq experiments, and caution is therefore warranted in interpreting the results of ongoing and published results purporting to show the association of many proteins with the transcribed regions of genes.

3.3 Materials and Methods

3.3.1 Yeast strains and culture conditions

The yeast strain used in this study as a WT was BY4741 (*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*). For ChIP, the TAP-tagged yeast strains including SWI6-TAP, TUP1-TAP, RSC2-TAP and MNN10-TAP strains were obtained from the yeast TAP-fusion collection (Open Biosystems) [47]. We generated the HSF1-TAP strain from BY4741 by integrating the TAP-HIS3MX6 cassette into the 3'-end of *HSF1* through homologous recombination, enabling the expression of C-terminal TAP-tagged Hsf1. Using the same scheme, we also generated a SWI6-13XMYC strain from BY4741. For gene expression

profiling, we used the *TUP1* deletion strain from the yeast deletion collection (Open Biosystems) [112]. The identity of all engineered strains was verified by genomic PCR. Normal growth conditions were 30°C in YPD media with shaking at 250 rpm. Yeast cells were grown to mid-log phase (O.D 600 nm of 0.6 to 0.8), fixed with formaldehyde and collected for ChIP; or, were collected without fixation for gene expression profiling. For heat shock, mid-log phase yeast cells were collected and re-suspended in pre-warmed 39°C YPD media, then incubated for 15 min at 39°C. For rapamycin treatment, either DMSO or rapamycin was added to mid-log phase yeast cells and incubated for 30 min at 30°C. Since DMSO is a solvent for rapamycin, control and rapamycin-added cells were treated with DMSO and rapamycin to be a final concentration of 0.1% and 100 nM, respectively.

3.3.2 Chromatin immunoprecipitation

Proteins were cross-linked to DNA by adding formaldehyde to the culture (final concentration of 1%) and the cross-linking reaction was quenched with glycine (final concentration of 0.125 M). Yeast cells were re-suspended with lysis buffer and disrupted by agitation with glass beads using a Bead beater (BioSpec Products). The cell lysates were sheared using a Branson Sonifier (Emerson Industrial Automation), and immunoprecipitated using the following beads or anti-body: IgG Sepharose 6 Fast Flow (GE Healthcare Life Sciences) to pull-down all TAP-tagged proteins used in this study, anti-Myc conjugated agarose bead (Sigma Aldrich, cat.# E6654) to pull-down Swi6 in the SWI6-13XMyc strains, and RNAPII Ser5P antibody (Abcam, cat.# ab5131) to pull-down

active RNAPII. Mock ChIP DNA was prepared by immunoprecipitation with IgG Sepharose in the wild type strain with no TAP-tagged protein expression. Input DNA was prepared in parallel with the SWI6-TAP ChIP sample but leaving out the immunoprecipitation step. The crosslinks were reversed and the immunoprecipitated DNA was purified using UltraPure Phenol:Chloroform:Isoamyl alcohol (25:24:1 v/v, Invitrogen).

3.3.3 Sequencing library preparation

Sequencing library preparation with ChIP-ed DNA and input DNA was carried out by following either the NEB ChIP-seq library preparation for Illumina (New England Biolabs) or the SOLiD V3 barcoded fragment library preparation protocol (Life Technologies). Sequencing was performed through either Illumina HiSeq 2000 or SOLiD V4 at the University of Texas at Austin Genome Sequencing and Analysis Facility (UT GSAF).

3.3.4 Gene expression profiling

The collected yeast cells were re-suspended with AE buffer (50 mM Sodium Acetate pH 5.2, 10 mM EDTA) containing 1.7% SDS, and total RNA was extracted with a hot acid phenol method [72]. Double-stranded cDNA was synthesized from total RNA, and labeled with Cy3 using the NimbleGen One-Color DNA labeling kits (Roche NimbleGen). The labeled cDNA was hybridized onto a NimbleGen *S. cerevisiae* HX12 array (Roche NimbleGen), and the array was washed and scanned with a GenePix 4000A scanner (Molecular Devices). The scanned image was processed using NimbleScan for

quantification of signal intensities and Robust Multi-array Average normalization with a large set of other NimbleGen array datasets in our lab (Roche NimbleGen). Differentially expressed genes in *tup1Δ* relative to WT were identified with Bioconductor limma package version 3.14.4**.**

3.3.5 Quantitative PCR

Three high TR genes (*CCW12*, *TDH3*, and *PDC1*) and three low TR genes (*PDR8*, *HKR1*, and *BIT61*) were selected. Two control primers used for normalization were designed from the tail-to-tail intergenic regions between YHL004W (*MRP4*) and YHL003C, stated as iYHL004W, and between YCR023C and YCR024C, described as iYCR024C. Primer pairs used in qPCR were designed to amplify 80-100 bp regions within the respective ORFs. qPCR was performed using Power SYBR Green PCR Master Mix (Applied Biosystems) on a ViiA7 Real Time PCR System (Life Technologies). For relative quantification of target DNA compared to control DNA, qPCR data was analyzed through a standard curve-based method.

3.3.6 Deep sequencing data analysis

Deep sequencing data were mapped onto the sacCer3 reference using BWA (Version: 0.5.9-r16) with default options [73]. Non-uniquely mapped reads were filtered out in order to remove reads with low mapping quality. Wig files of sequencing data were loaded in a local mirror of the UCSC Genome Browser for snapshots [113]. For average read profiles, reads were counted by bin size 10 bp within 1.5 kb from transcription start sites (unpublished data), and counts were divided by the total number of mapped reads

and multiplied by 1 million. The graphs were drawn with Python module matplotlib. Peak calling was performed with MACS2 (version: 2.0.9) [114]. Cse4 and untagged control ChIP-seq were downloaded from Gene Expression Omnibus database (GEO) Series accession number GSE13322 and GSE20870 [110, 115], respectively. We also downloaded histone MNase ChIP-seq data from NCBI Sequence Read Archive accession number SRA012303 [77]. These published datasets were processed with the same analysis pipeline as above.

3.3.7 Mock and input comparison

We executed the MACS2 module (version: 2.0.9) for 4 different experimental pairs: 1) DMSO Tup1 ChIP and DMSO input, 2) DMSO Tup1 ChIP and DMSO mock ChIP, 3) Rap Tup1 ChIP and Rap input, and 4) Rap Tup1 ChIP and Rap mock ChIP. Also, two thresholds ($\log_{10}$[q-value] = 2 and 20) were chosen to compare the efficiency of a threshold to eliminate expression bias peaks based upon stringency (Table 3.1). Then, MAnorm was utilized to identify DBTs from the MACS generated data [116]. MAnorm allowed us to ignore regions where the control showed higher signals than the treated sample. Thus, by using different controls in MACS followed by MAnorm analysis, we were able to test the effect of controls on the removal of background signals based on the number of DBTs and the percentage of DBTs within gene bodies. We transferred the MACS peak data from experimental pairs 1 and 3 (see above) to MAnorm and repeated for experimental pairs 2 and 4 (see above). We applied the same cut-off p-value ($-\log_{10}$(q-value) = 5) for DBTs to the MAnorm results.

3.3.8 Accession number

Sequencing data reported in this manuscript are available from NCBI GEO as GSE51251 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51251).

3.4 Results

3.4.1 Common enrichment signals in ChIP-seq datasets

In order to study the targets of chromatin binding proteins in response to transcriptional perturbations, we performed ChIP-seq against multiple chromatin-associated factors after treatment of cells with rapamycin (with DMSO treatment serving as control) or heat shock (at 39°C, with growth at 30°C serving as control). Included among these experiments were two unrelated transcription factors, Swi6 and Tup1, and various negative controls. One type of control was a mock ChIP-seq, in which immunoglobulin G (IgG)-conjugated sepharose beads were incubated with wild-type (WT) yeast chromatin. In another control, the input of a Swi6 ChIP sample (the sheared chromatin from a SWI6 TAP-tagged strain) was sequenced. Finally, we also ChIP-ed a subunit of Golgi mannosyltransferase complex Mnn10; as a cytoplasmic complex, Mnn10 is unlikely to associate with chromatin and thus was not expected to pull down any DNA.

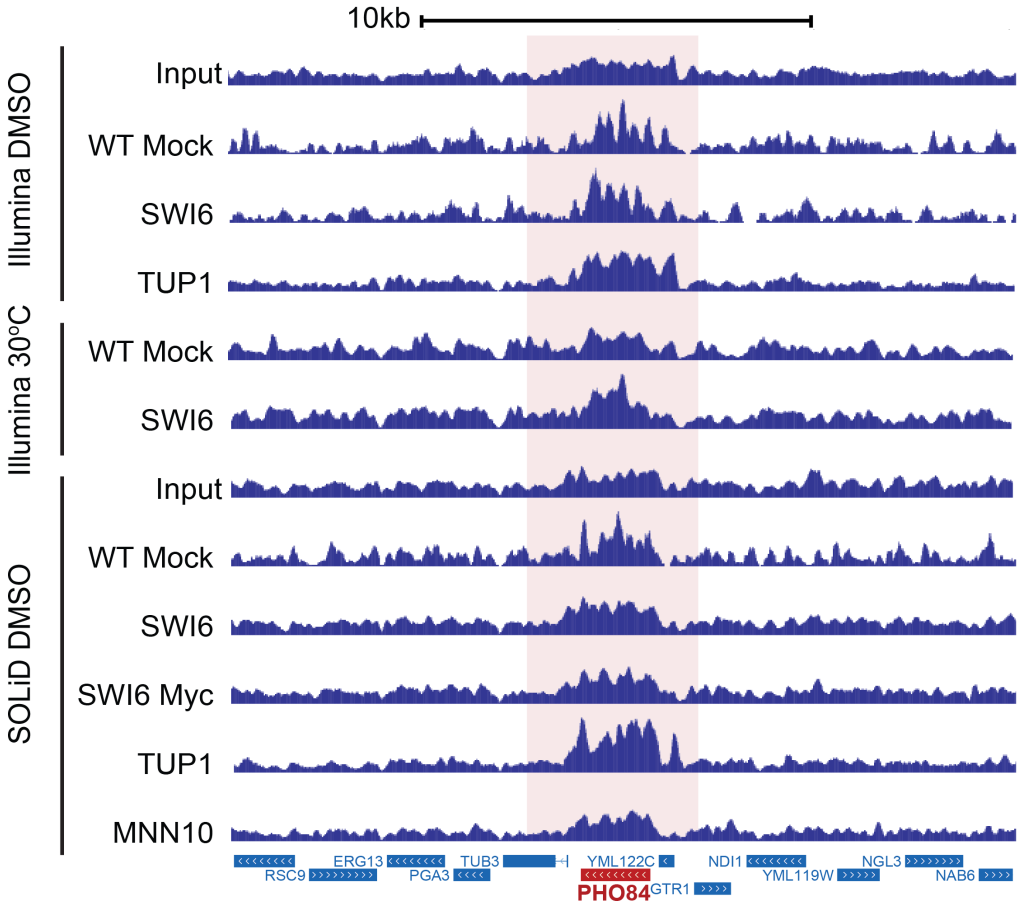

Figure 3.1 Example of high background signal across multiple datasets

Sequencing datasets from different factors, controls, epitope tags, transcription factors and growth conditions as indicated ares represented in a browser view. Based on the read counts normalized by transcript lengths from RNA-seq data [36], *PHO84* is the 82nd most highly expressed gene under normal conditions in WT yeast

We noticed that surprisingly, common targets were enriched across several data sets, including Mnn10 ChIP (Figure 3.1). Such peaks were observed across different sequencing platforms (Illumina or SOLiD), epitope tags (SWI6 TAP-tagged or SWI6 13XMyc tagged), bead types (IgG-tagged sepharose beads or c-Myc antibody-conjugated agarose beads), and immunoprecipitated factors (Swi6 or Tup1) (Figure 3.2), indicating

that the shared signals were not derived from the use of a specific protocol or reagent. Perhaps most significantly, the targets were shared between the standard mock ChIP and input control experiments, suggesting that these shared targets represented non-random false positives.

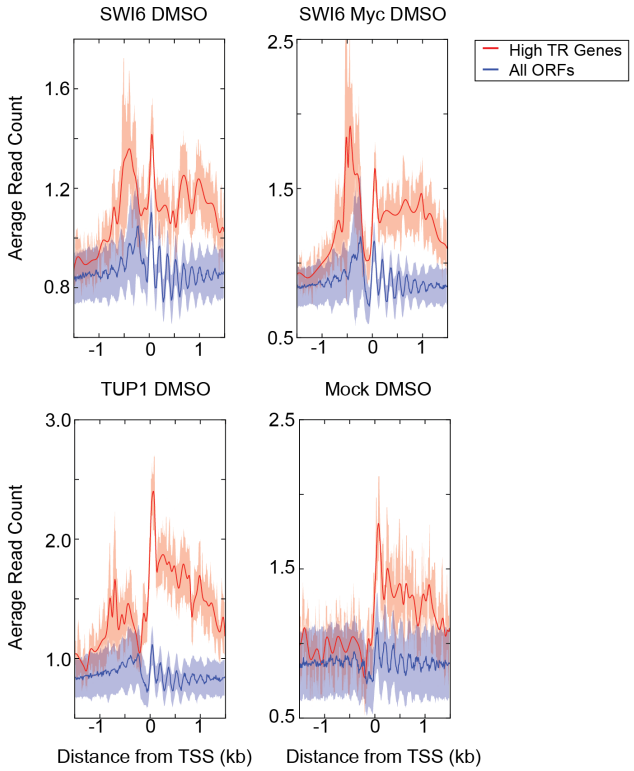

Figure 3.2 High background signals at high TR genes in SOLiD sequencing data

SWI6 Myc indicates ChIP against 13XMyc tagged Swi6 using c-Myc antibody conjugated agarose beads. We pulled down TAP tagged proteins for other ChIPs. The expression bias in *TUP1* was the highest in SOLiD, and mock ChIP showed expression bias comparable to Swi6 ChIP.

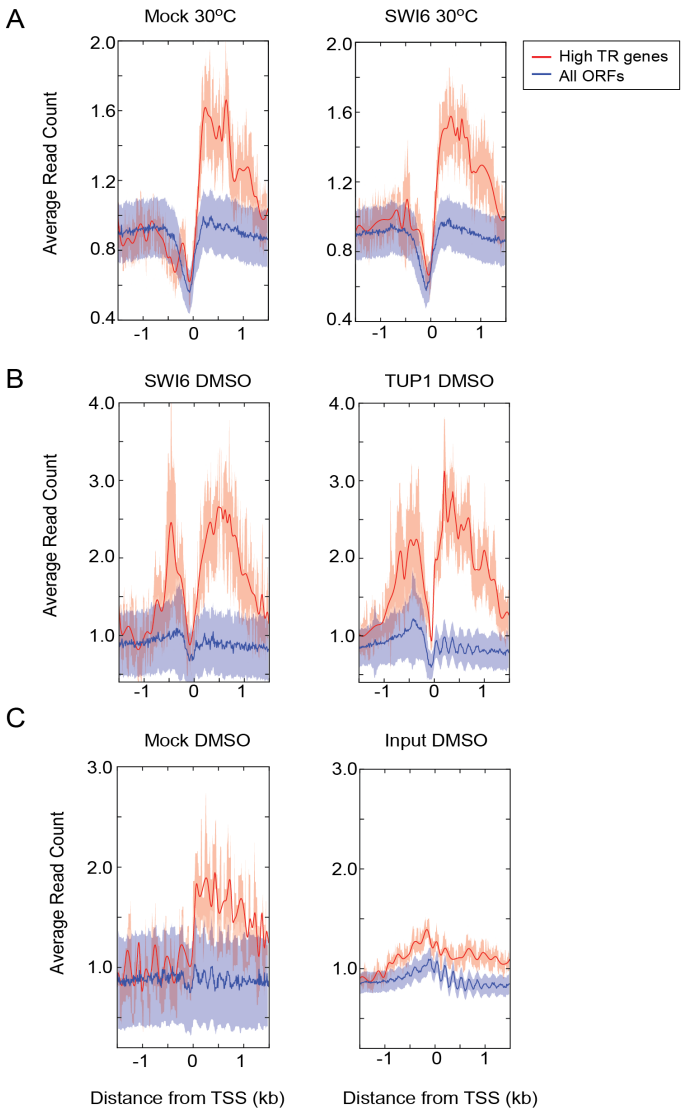

Figure 3.3 Genes with high transcription rates (TR) have high average read counts

Lines show average read counts in 10 bp bins for the indicated groups of genes, which are either the 100 most highly transcribed genes based on RNAPII Ser5P occupancy as described in the text (High TR genes, red line) or all the other genes (All ORFs, blue line). The shaded bands represent the 95% confidence interval of the data. All ChIP samples in this figure were sequenced using the Illumina platform. (A) Under normal growth conditions (30°C in YPD), mock ChIP had comparable bias to Swi6 ChIP. (B) Both SWI6 (an activator) and TUP1 (a repressor) show comparable high levels of the expression bias at high TR genes. (C) Input has a lower expression bias than mock ChIP. For (B) and (C) cells were treated with DMSO, which was a control for rapamycin treatment.

3.4.2 Highly expressed genes demonstrate widespread, strong ChIP-seq signals

We next examined whether the phenomenon described above was generally observable genome-wide. We observed two features among the strong false positive signals. First, the signals were present within gene bodies and second, the strongest signals derived from yeast genes that are known to be highly expressed. Thus, we termed this artifact an "expression bias". In order to better define the set of highly transcribed genes, we performed ChIP-seq against active RNA polymerase II under the same conditions. The occupancy of RNAPII phosphorylated at serine 5 of its C-terminal domain repeats (RNAPII Ser5P) is a better indicator of transcription rate than steady state RNA levels [117]. We defined the top 100 open reading frames (ORFs) in terms of RNAPII Ser5P occupancy (after normalizing for gene length and sequencing depth) as high transcription rate (high TR) genes.

Read counts over genes in several ChIP-seq and control experiments were strongly enriched for high TR genes compared to other genes (Figure 3.3). Consistent with the example shown in Figure 3.1, the expression bias was a recurrent artifact in all ChIP-seq data, although the degree of expression bias varied from factor to factor. To examine if the expression bias was an artifact specific to ChIP-seq data from our lab, we downloaded previously published ChIP-seq data from other labs and analyzed them using the same pipeline [110, 115]. Specifically, we compared ChIP for a centromere binding protein, Cse4, [110] and an independent mock ChIP that had been used as a negative control for the association of the transcription factor Tbf1 [115]. Cse4 in particular is a

centromere-specific histone H3 variant that is not expected to occupy transcribed regions. Both of these published datasets exhibited the same artifacts as we describe above (Figure 3.4), suggesting that the expression bias seen for high TR genes is a commonly occurring phenomenon in yeast ChIP-seq data and could confound the interpretation of many types of experiments.

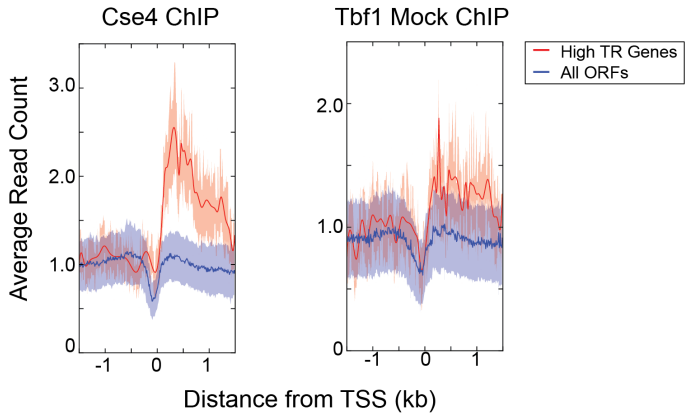

Figure 3.4 The expression bias in two independent, previously published datasets

We downloaded two previously published ChIP-seq datasets and ran our pipeline. 13XMyc tagged Cse4 was immunoprecipitated with the same beads as used in 13XMyc Swi6 ChIP in Figure S1 [110]. As a negative control ChIP for 13XMyc Tbf1 ChIP, monoclonal anti-Myc antibody was incubated with untagged W303-1A strain [115]. Both ChIP-ed DNA samples were sequenced using the Illumina platform.

3.4.3 Human CTCF ChIP-seq has the expression bias

Recent studies on the co-localization of TFs showed that the regions highly co-occupied regions by TFs were associated with high levels of RNAPII occupancy [118, 119], implying that expression bias could also be present in human ChIP-seq. In order to examine the expression bias in human ChIP-seq, we chose CTCF ChIP-seq data because

the ENCODE project indicated that CTCF target sites were associated with both gene activation and repression [8]. The GM12892 cell line was used to conduct the ChIP experiments, and the total number of mapped reads was 148,671,491. To calculate gene expression levels in GM12892, we downloaded RNA-seq data from the ENCODE project data archive. Human genes were sorted by the Fragments Per Kilobase of exon per Million fragments mapped (FPKM). After excluding genes that encode miRNAs, ribosomal proteins, and snoRNAs, the 1000 most highly transcribed genes were defined as "high TR genes". Surprisingly, as observed in yeast TFs, CTCF showed similar high background signals at the high TR genes, suggesting that the expression bias could be universal in ChIP-seq data (Figure 3.5).

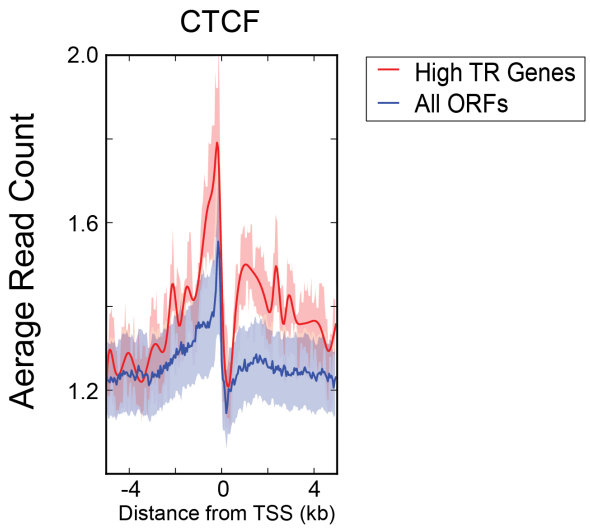

Figure 3.5 The expression bias in human CTCF ChIP-seq

CTCF in GM12892 cells shows high levels of the expression bias at the 1000 most highly transcribed genes.

3.4.4 Expression bias of ChIP-seq by condition-specific transcriptional activation

The transcript levels of stress-responsive genes are dramatically altered by rapamycin treatment and heat shock [120, 121]. Given the fact that upregulated genes under stress conditions show comparable transcription rates to high TR genes under normal conditions, we wondered whether the expression bias in ChIP would similarly be detectable in upregulated genes specifically under stress conditions. To answer this question, we first measured the condition-specific occupancy of active RNAPII on chromatin by ChIP-seq after rapamycin treatment and heat shock. The top 100 ORFs showing increased occupancy after treatment relative to normal were defined as transcriptionally upregulated genes in response to rapamycin and heat shock (or "Rap Up" genes and "Heat Up" genes), respectively.

As the cell cycle is arrested at G1 by heat shock [122], we reasoned that Swi6, a well-known transcription activator of the G1/S transition [123, 124], would not bind strongly to heat shock-induced genes. Surprisingly, we found that Swi6 bound strongly to the transcribed regions of Heat Up genes specifically after heat shock (Figure 3.6A). A mock ChIP control sample for this experiment showed similar enrichment at Heat Up genes. While this illustrated the expression bias as manifested for differentially expressed genes during a perturbation, we investigated a different stress condition to rule out the possibility that the expression bias was specific to heat shock or to Swi6. We performed ChIP-seq for Rsc2, a component of the RSC chromatin remodeling complex, and Tup1, a component of the TUP1-CYC8 co-repressor complex, after rapamycin treatment of cells.

Both Rsc2 and Tup1 showed high occupancy over the transcribed regions of Rap Up genes after rapamycin treatment (Figure 3.6A). Thus, unrelated transcription factors appear to show increased binding to the ORFs of genes that are more actively transcribed after different environmental perturbations.

3.4.5 Expression bias can give misleading information

Despite the expression bias observed in mock ChIP and other control experiments above, it is possible that certain transcription factors also truly bind to ORFs as a means of regulating gene expression. For example, occupancy by a transcription factor of the ORFs of high TR genes, or of Heat Up genes specifically after heat shock might suggest a role in activating transcriptional elongation, something that cannot be formally ruled out based on our data for Swi6. However, the case of Tup1 offers a means of testing this notion. The molecular mechanism of the Tup1-Cyc8 complex as a general transcriptional repressor has been well established [125]. In order to confirm that Tup1 does not also serve as a transcriptional activator, we performed gene expression profiling of a *tup1Δ* strain compared to WT. Almost 90% of the differentially expressed genes were repressed by Tup1, showing that Tup1 does not, in fact, activate these genes in wild type cells (Figure 3.6B). Yet, ChIP-seq data for Tup1 suggested just the opposite. Tup1 occupied high TR genes as opposed to the low TR genes one would expect for a repressor. In this instance therefore, occupancy of high TR genes by Tup1 is likely to give a misleading picture regarding its biological function.

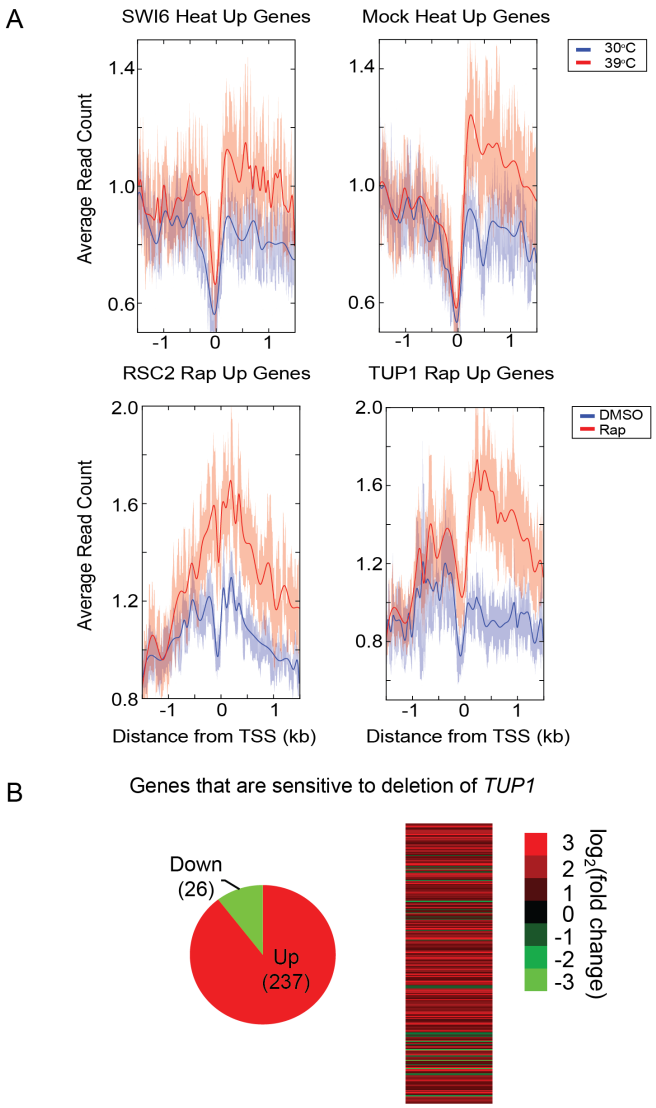

Figure 3.6 Condition-specific expression bias at genes that are transcriptionally activated

(A) ChIP-seq data for an activator (Swi6), corepressor (Tup1), chromatin remodeler (Rsc2), and a mock ChIP control for genes that are transcriptionally activated by the indicated treatment. "Heat Up" are genes activated by heat shock, and "Rap Up" are genes activated by rapamycin treatment. Red lines show data after treatment (39°C or rapamycin), while blue lines show data before treatment (30°C or DMSO) for the same set of genes. (B) Differentially expressed genes comparing a WT strain to *tup1Δ*. The majority of genes were activated upon deletion of *TUP1*, demonstrating that Tup1 is primarily a transcriptional repressor.

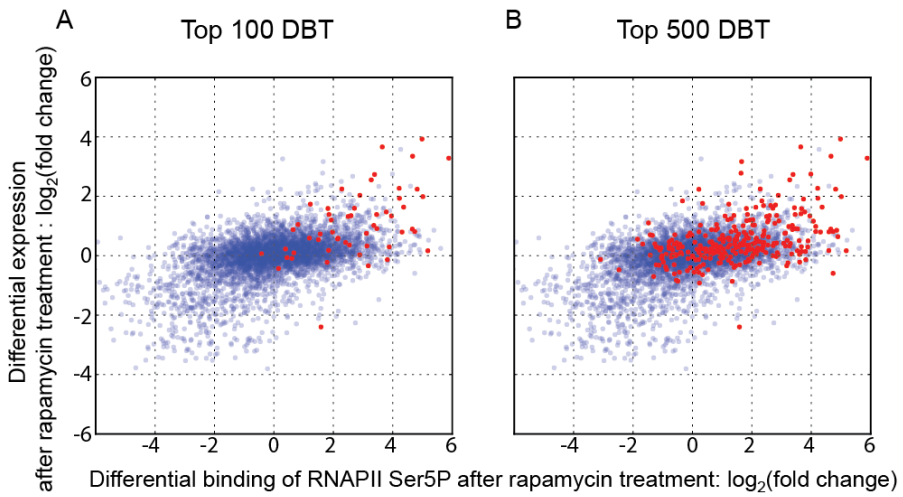

Figure 3.7 Misleading pictures showing that Tup1 is primarily a transcriptional activator

Scatter plots show the differential transcriptional activation after rapamycin treatment as blue points. Differential RNAPII Ser5P occupancy before and after rapamycin treatment was measured by ChIP-seq and plotted on the X-axis. Differential mRNA expression levels in the same cultures were measured using microarrays and plotted on the Y-axis, in scatter plots showing 4929 genes. We used MACS to identify differential binding targets (DBTs) of Tup1 as described in the text and plotted them on the same plots in red. (A) The top 100 DBT peaks ranked by fold change were assigned to 55 ORFs, which are plotted in red. (B) The top 500 DBT peaks were assigned to 295 ORFs, which are plotted in red. Tup1 DBT ORFs tended to be upregulated genes in response to rapamycin. Uncorrected Tup1 differential binding targets misleadingly indicate that Tup1 is primarily a transcriptional activator.

A common use of ChIP-seq is to examine binding of a given factor under different growth conditions or backgrounds. Since only a single variable is changed (the experimental or growth condition), it might be assumed that comparing binding under different conditions offers a reliable means of identifying biologically relevant targets, with most background artifacts being normalized out. We wondered whether the expression bias we noted earlier could nevertheless confound the interpretation of such experiments. We used the MACS algorithm to identify targets showing increased binding

of Tup1 in response to rapamycin treatment [114]. We used vehicle (DMSO) treated cells as the control and rapamycin treated cells as the experimental sample, and used MACS to identify differential binding targets (DBTs) from the ChIP-seq data for Tup1 under these two parallel conditions. 57 of the top 100 and 322 of the top 500 DBTs identified by MACS were in ORFs. Strikingly, the majority of these DBT ORFs were ORFs that were transcriptionally activated by rapamycin treatment. When superimposed on a scatterplot of gene expression versus RNAPII Ser5P occupancy, the Tup1 DBT ORFs were concentrated in the upper right quadrant (Figure 3.7). In the absence of other knowledge about Tup1 function, one would misinterpret this data to mean that Tup1, since it associates with the ORFs of rapamycin-upregulated genes after rapamycin treatment, likely functions in the activation of those genes. These results therefore raised the question of what type of normalization controls might be appropriate for minimizing false positives in ChIP-seq data, even when analyzing differential binding under different conditions.

3.4.6 Mock ChIP is a better control for expression bias

We observed that mock ChIP-seq data exhibited a stronger expression bias than the corresponding input samples (Figure 3.3C), and therefore hypothesized that correction by mock ChIP (normalization) would more effectively reduce the false-positives exemplified by Tup1 DBT ORFs than normalization by input. To test this hypothesis, we first used MACS to normalize each condition specific ChIP-seq dataset to either its corresponding input or mock ChIP-seq sample. We used low and high

stringency thresholds to compare their effectiveness in minimizing false positives (Table 3.1). We then used MAnorm to identify DBTs from this MACS-normalized data [116]. At a given p-value threshold, fewer Tup1 DBTs were identified when using mock ChIP-seq data as the normalization control (Table 3.1)

Category		Cut-off Stringency	Input Correction	Mock Correction
MACS2 Peak Calling	Control	Low	2478	726
		High	1120	296
	Rapamycin	Low	2419	845
		High	725	309
Rapamycin-specific targets by MAnorm		Low	770	407
		High	384	165
Rapamycin-specific targets within gene bodies		Low	379 (49.2%)	149 (36.6%)
		High	139 (36.2%)	32 (19.4%)

Table 3.1 Rapamycin-specific Tup1 peaks

Rapamycin-specific Tup1 peaks were identified by using MACS followed by MAnorm analysis. Low and high stringency cut-offs were $-\log_{10}$(q-value) = 2 and 20, respectively. Rapamycin-specific targets were those differential binding peaks found by MAnorm with $-\log_{10}$(q-value) > 5.

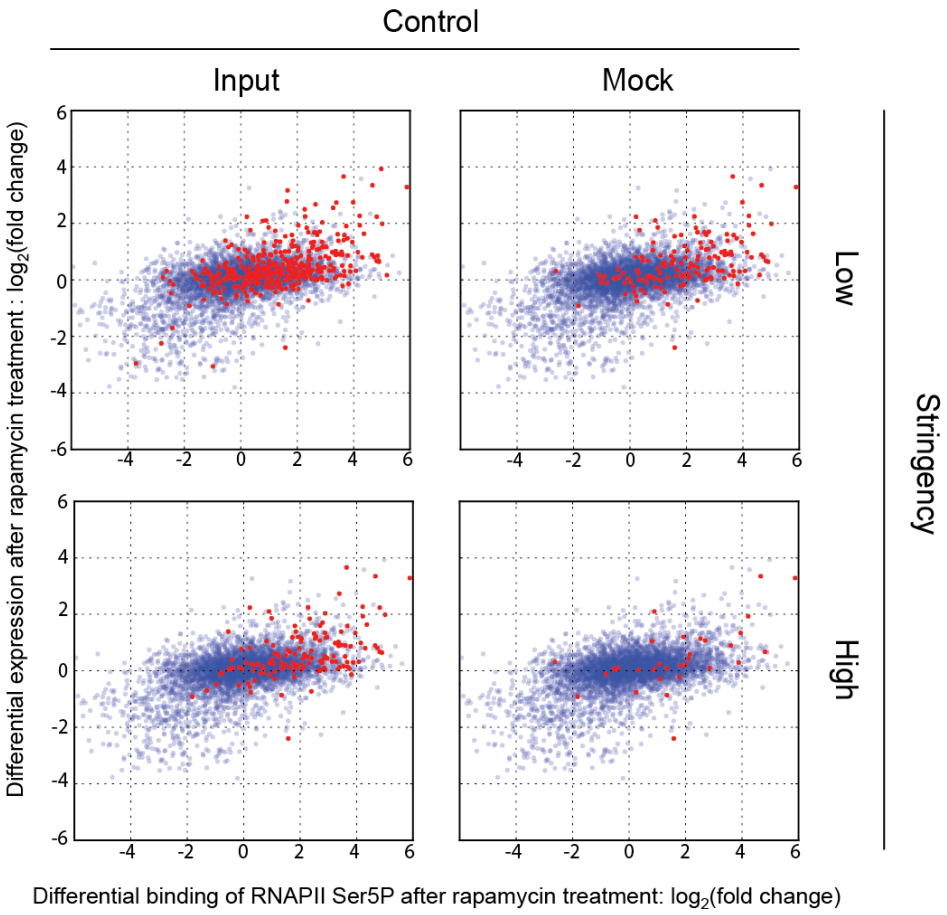

Figure 3.8 Mock ChIP is a better control for minimizing false positive ChIP-seq targets

Either ChIP input or mock-ChIP was used as a control, at two q-value thresholds to obtain high and low significant peaks (see text and Materials and Methods). The scatter plots were drawn as described in Figure 3.7, and the numbers of Tup1 DBT ORFs (red) were as follows: input low stringency=379, input high stringency=139, mock low stringency=149, mock high stringency=32. Mock ChIP is a better normalization control than input for minimizing false positive ChIP-seq targets.

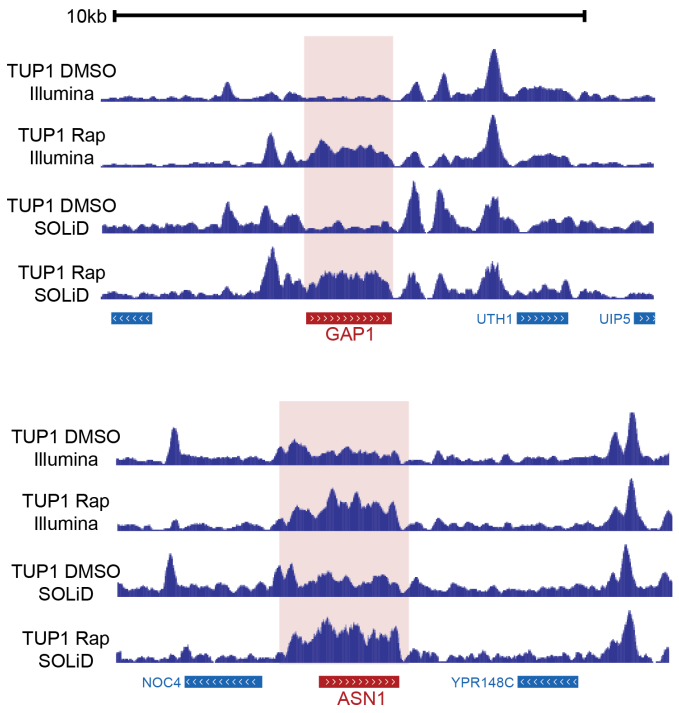

Figure 3.9 Examples of high expression bias in rapamycin-specific targets

Based on differential binding of RNAPII Ser5P (DB) and differential expression by microarray (DE) in response to rapamycin, *GAP1* showed 59 and 10 positive fold-change in terms of DB and DE, respectively, ranking within the top 10 in both measurements. Although the rank of *ASN1* in DE was 1466, the DB was ranked in top 56 as 15 fold change. The gene bodies had strong signals for rapamycin-specific occupancy by Tup1, which could not be corrected by rapamycin-treated mock ChIP.

The use of mock ChIP as a normalization control resulted in a lower proportion of DBT ORFs (49.2% vs 36.6% and 36.2% vs 19.4% in Table 3.1), suggesting that mock ChIP is a more effective normalization control for expression bias than the input sample. The use of a more stringent threshold in conjunction with a mock ChIP normalization control reduced the number of DBT ORFs that were correlated with high transcription rates in an obvious manner (Figure 3.8). However, even this method of minimizing such

likely false positives is not infallible. For example, *GAP1* and *ASN1* were activated by rapamycin and showed Tup1 occupancy signals that were comparable to true peaks (Figure 3.9). *GAP1* expression increased by 3.64 fold in a *tup1Δ* strain compared to WT, strongly suggesting that Tup1 is a repressor, rather than an activator of *GAP1*. Establishing a role for Tup1 in activating these genes in response to rapamycin is therefore non-trivial. Thus, while mock ChIP is a more stringent control for the identification of Tup1 DBTs in response to rapamycin, there is still strong evidence for apparent differential binding to several ORFs, where it is difficult to distinguish between expression bias or true binding with biological significance.

3.4.7 Careful interpretation is required

Unlike sequence-specific transcription factors, ChIP for chromatin remodelers and chromatin-modifying enzymes is inherently difficult because of how transiently these factors bind to chromatin [42]. Many chromatin remodeler ChIPs demonstrate weakly detectable signals to begin with, making it harder to distinguish them from expression bias. To investigate the effect of expression bias in chromatin remodeler ChIPs, we examined MNase ChIP-seq data for the ATP-dependent remodeler Chd1 from a previously published paper reporting that Chd1 associated with the transcribed regions of actively transcribed genes [126]. We mapped these reads with BWA and discarded non-uniquely mapped reads because the paired-end reads had a read length of only 25 bp. We plotted the read profile relative to yeast transcription start sites and observed the nucleosomal periodicity expected for the association of chromatin remodelers with

chromatin. As reported, Chd1 occupancy on high TR genes was higher than other gene groups both before and after input correction (Figure 3.10A and 3.10B). However, the difference in Chd1 occupancy between high TR genes and low TR genes was very small using input correction. When we normalized Chd1 occupancy with our mock ChIP data, however, the correlation with the transcription rate was no longer observed (Figure 3.10C). Thus, the association of Chd1 binding to ORFs and its relationship with transcription rate remains unclear when expression bias is properly accounted for.

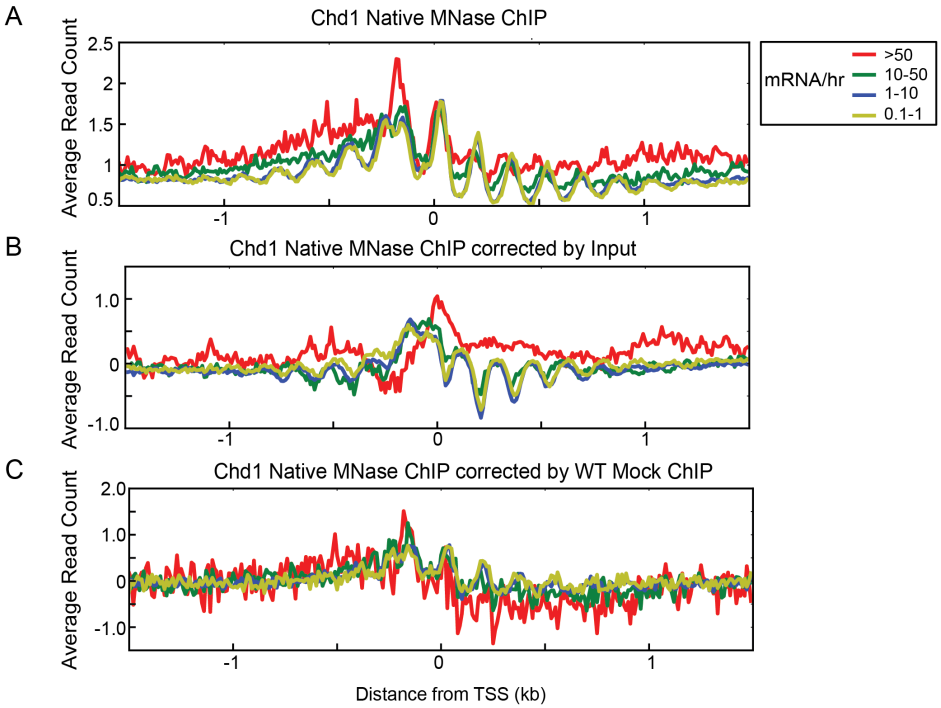

Figure 3.10 A misleading relationship between ORF binding and transcription rate

(A, B) Previously published ChIP-seq data for Chd1 [126] was plotted either uncorrected (A) or corrected by input (B). Occupancy is higher at high TR genes compared to low TR genes, when genes are ranked by mRNA/hr [127]. (C) Same Chd1 ChIP-seq data, after correction by mock ChIP-seq data, no longer shows a strong relationship of Chd1 occupancy with transcription rate

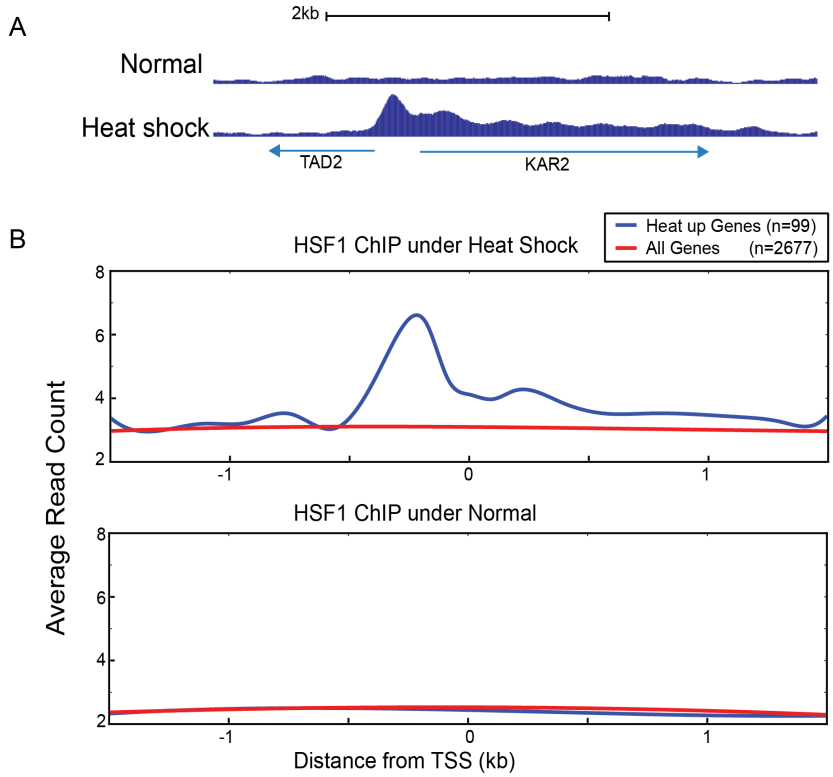

Figure 3.11 ChIP-seq signal from binding of Hsf1 to bidirectional promoters

(A) Hsf1 strongly bound to the shared promoter of *TAD2* and *KAR2*. The binding signals gradually decreased towards the 3' end of *KAR2* which is strongly activated upon heat shock, whereas the signal dropped sharply in the direction of *TAD2* transcription. (B) Average Hsf1 occupancy over the 99 divergent genes out of the top 200 Heat Up genes (red), and all the other divergent genes (blue) under normal and heat shock conditions. In this representation, Heat Up genes were arranged on the right with respect to the genes whose promoter was shared, which reveals that Hsf1 binding decreases gradually over the Heat Up genes.

3.4.8 Expression bias suggests directionality of transcription

When a transcription factor binds to bidirectional (divergently regulated) promoters, it can be difficult to identify which of the two divergent ORFs, if any, is transcriptionally regulated by its binding. We examined ChIP-seq data for Hsf1 to see if expression bias could shed light on this issue. HSF1 is a key regulator of the

transcriptional response to heat shock, strongly binding to the promoters of the Heat Up genes after heat shock [128]. We noticed that the signal for Hsf1 binding was asymmetric across the two divergent ORFs. The peak of Hsf1 binding occurred between the start sites of *TAD2* and *KAR2* but the tail of the Hsf1 ChIP-seq signal extended toward *KAR2*, not *TAD2* (Figure 3.11A). Based upon the differential binding of RNAPII Ser5P after heat shock, *KAR2* was strongly transcriptionally activated, while *TAD2* was not. 99 genes out of the top 200 RNAPII Ser5P heat shock DBTs shared promoters with another divergently transcribed gene. At these genes, the tails of Hsf1 binding stretched toward the DBTs (Figure 3.11B). Thus, the ChIP-seq binding signals over ORFs for transcription factors that strongly regulate gene expression can potentially identify the correct target gene from bidirectionally transcribed ORFs.

3.4.9 The expression bias is amplified during library construction

To establish whether the expression bias is primarily an artifact arising during sequencing library construction procedures or already exists in the immunoprecipitated DNA, we carried out quantitative PCR using ChIP-ed DNA before and after library construction.

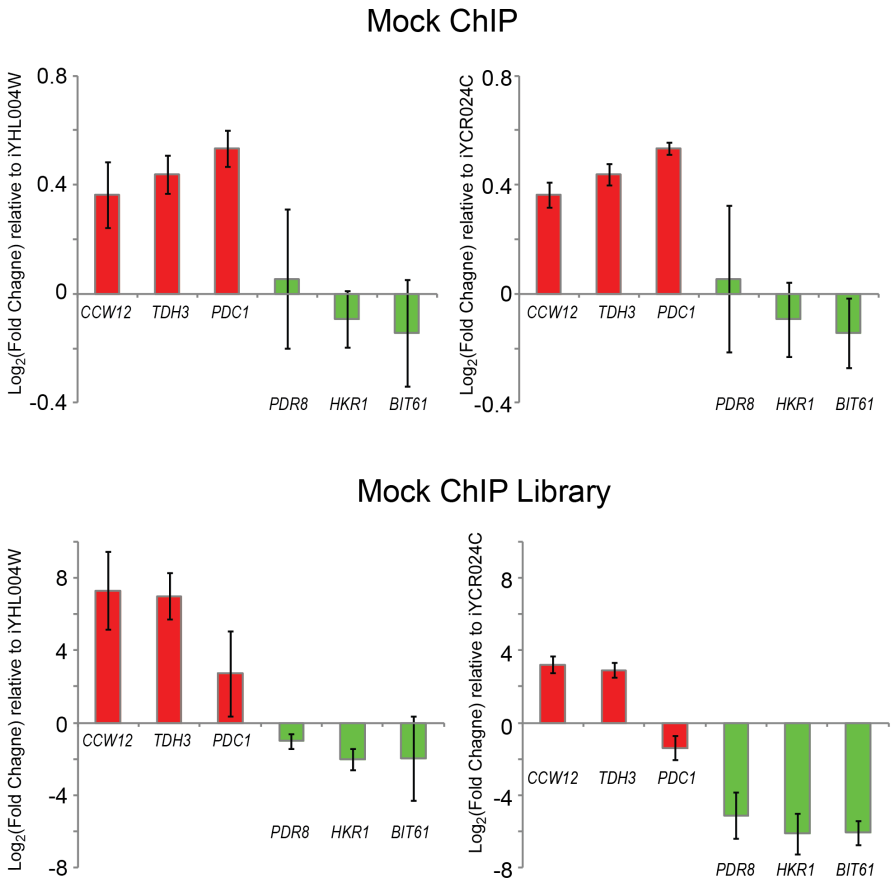

Figure 3.12 qPCR shows higher expression bias in sequencing library than mock ChIP

Three ORFs showing high enrichment of RNAPII Ser5P and high expression levels by RNA-seq were selected as high TR genes, shown in red (*CCW12*, *TDH3*, and *PDC1*). Three ORFs were picked as low TR genes using the same criteria, and are shown in green (*PDR8*, *HKR1*, and *BIT1*). For relative quantification of targets, two different controls were used (iYHL004W, plotted on left and iYCR024C, plotted on right), and fold-changes were calculated by dividing the mean of target quantities by the mean of control quantities. Three biological replicates were carried out with two independently prepared mock ChIP samples, one of which was used for sequencing. Error bars represent the standard deviation of three log2-transformed fold change values from the replicate experiments.

As examples of genes showing the expression bias, we chose three genes, *CCW12*, *TDH3*, and *PDC1*, which had the highest expression bias based on the mock ChIP read counts and also ranked within the top 20 most highly expressed genes based on read counts from RNA-seq and RNAPII Ser5P ChIP-seq. As negative targets, we selected *PDR8*, *HKR1*, and *BIT6* as they had low read counts in mock ChIP, RNA-seq, and RNAPII Ser5P ChIP-seq. In mock ChIP DNA, the genes showing high expression bias were overrepresented, whereas the genes showing no expression bias were underrepresented, indicating that the expression bias was present even before sequencing libraries were made (Figure 3.12). In the sequencing libraries, these differences in representation were magnified (Figure 3.12), indicating that amplification during sequencing library construction could result in the over-representation of high TR genomic regions in sequencing results.

3.4.10 Nucleosomal periodicity of RNAPII Ser5P ChIP

We observed that many ChIP-seq profile plots showed a periodicity of mean read counts over regions devoid of strong peaks (Figure 3.2 and 3.3). This periodicity within gene bodies, which was identical to nucleosomal periodicity, was also present in RNAPII Ser5P ChIP and especially noticeable for low TR genes (Figure 9A).

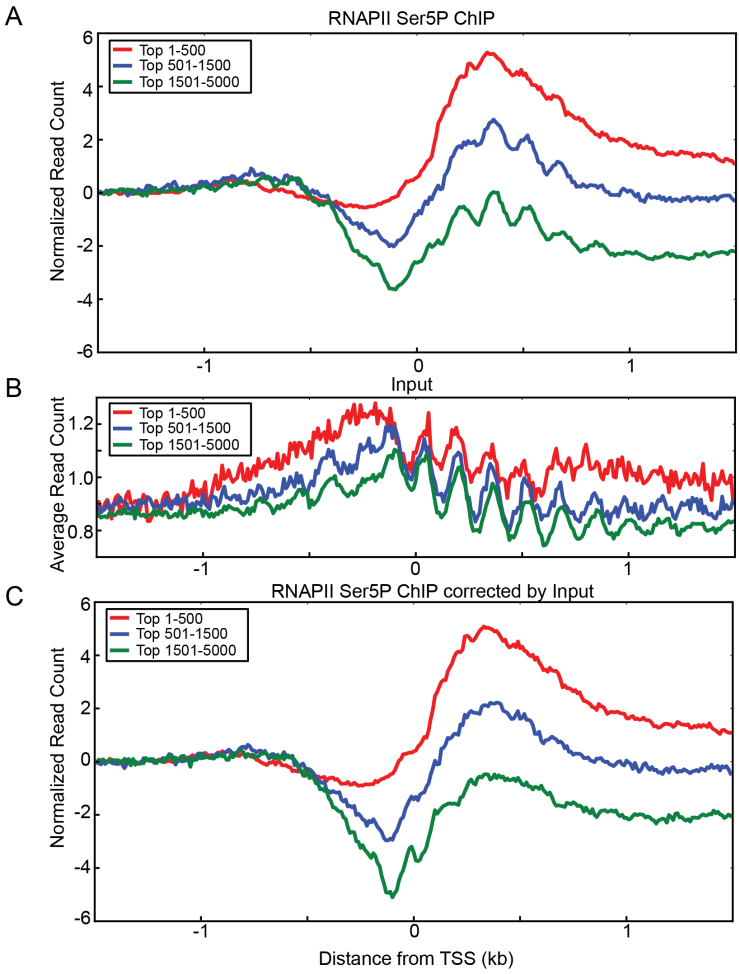

Figure 3.13 Nucleosomal periodicity in a ChIP-seq dataset

ORFs were grouped by RNAPII Ser5P occupancy into 3 categories as indicated, and normalized sequencing reads of the 3 categories are shown on the Y-axis. (A) In the RNAPII Ser5P ChIP-seq read profiles, low expressed genes (blue and green lines) exhibited nucleosomal periodicity. Occupancy within each set was independently scaled and the profiles were set to start at the zero position on the Y-axis. (B) Input shows strong nucleosomal periodicity although average signal intensity is low (C) By subtracting input signal from RNAPII Ser5P signals, the apparent nucleosomal periodicity of RNAPII Ser5P was greatly reduced. Scaling factor for each group of genes was the same as in A

The naïve interpretation of these data would be that active RNAPII binds to individual nucleosomes and/or that RNAPII stalls at the center of nucleosomes during transcription. However, this interpretation, solely based on this observation would be misleading because even input exhibited similar strong periodicity (Figure 3.13B), as did Tup1 and Swi6 (Figure 3.2 and 3.3). When we normalized the RNAPII Ser5P read counts by the input read counts for each corresponding gene, the nucleosomal periodicity of the RNAPII Ser5P ChIP-seq was eliminated (Figure 3.13C), indicating that this periodicity was not a true signal but rather another artifact.

3.5 Discussion

In the analysis of ChIP-seq data, two types of normalization or correction controls are commonly used: mock ChIP and input DNA. The input sample has the advantage that all the regions of the genome are well represented, the sample concentration is ample and stable for constructing sequencing libraries, and the same sample can potentially serve as the control for several related experiments. The input generates a baseline signal for reads across the genome, factoring in sequence mappability and copy number differences relative to the reference genome. For these reasons, input has been suggested as a more effective control [129]. However, our results show that a background signal deriving from expression bias, namely genes transcribed at high rates, is not adequately represented in the input (Figure 3.3C). A mock ChIP sample processed in parallel through the immunoprecipitation and subsequent steps better reflects the

background enrichment from highly transcribed genes and therefore is a better control for minimizing the appearance of occupancy signal over transcribed regions. However the DNA yield after a mock ChIP step is typically lower and likely to be more variable from experiment to experiment.

It is often assumed that measuring the binding of a transcription factor under two different conditions and identifying the differentially bound targets (DBTs) offers the most reliable way to identify targets of biological significance. This assumes that most sources of background signal are canceled out between the two samples in such an experimental strategy. Our results indicate that this assumption is risky. Because the expression bias derives directly from actively transcribed genes, and transcription will differ between the two conditions, it will appear as if the factor under study shows differential binding when in fact it is the background expression bias that is differently represented in the two conditions. We suggest therefore that even in these cases, the ChIP data from each condition has to be properly corrected by the corresponding mock ChIP data to minimize false positives.

The expression bias we demonstrate has the potential to skew ChIP-seq data into representing any chromatin-associated protein as being associated with gene bodies or ORFs in yeast, regardless of the protein's true role. In particular, this misinterpretation is easy to arrive at when the proteins of interest are ones that often show low signal strength in ChIP-seq experiments, such as chromatin remodelers, histone modifying or associated factors, or components of the general transcription machinery [126, 130, 131].

It is beyond the scope of this study to definitively identify the source and mechanism of this background expression bias in ChIP-seq data. However, given that it is most strongly observed at highly transcribed genes, we speculate that in many cases it arises from direct or indirect non-specific interactions of the immunoprecipitated protein with DNA in open chromatin at highly transcribed regions, trapped by the crosslinking process. It is unclear why the phenomenon exists even in mock ChIP datasets, where there is no expected interaction between the non-specific antibody and any cellular protein that might interact with DNA. Here, it is possible that even low level non-specific interactions between the antibody and cross-reacting cellular proteins contribute to this phenomenon, or that open chromatin shows preferential recovery through the immunoprecipitation process. Indeed, the latter property underlies methods such as FAIRE and Sono-seq, which are aimed at globally recovering open chromatin regions [44, 132].

This pattern of the Hsf1 ChIP-seq signal is informative with regard to how background peaks derived from expression bias might be related to true occupancy in some cases. The strong background starts just downstream of the true Hsf1 binding site and gradually tapers off toward the 3' end of the gene (Figure 3.11). This tail structure suggests a model in which high TR genes that are opened by the transcription process facilitate the expression bias. The transcription machinery and co-factors are recruited onto the open chromatin of heat-activated genes upon heat shock in conjunction with Hsf1 recruitment. The close proximity of Hsf1 to this transcription machinery can allow

them to be cross-linked and co-immunoprecipitated. We speculate that this proximity effect of HSF1 around open chromatin generates the tail structure observed. The expression bias in the other ChIP data may similarly be derived from these open chromatin interactions. Importantly, to the extent that the expression bias is always related to transcriptional activity, and will be observed most strongly when a transcription factor capable of interacting with chromatin is immunoprecipitated after crosslinking, this background is essentially indistinguishable from true "biological" targets, especially when the true targets are seen at low levels. Our data address this phenomenon only in yeast ChIP-seq data, but conceivably, this could extend to ChIP-seq experiments in other eukaryotes as well. For example in mammals, cell-type or tissue-specific open chromatin is known to occur at promoters and enhancers [133]. A similar phenomenon as we described here for yeast could in part explain observations of hotspots of transcription factor binding and instances of neutral transcription factor binding, where such apparent binding has no biological meaning [134, 135].

The nucleosomal periodicity observed in input and non-target regions from ChIP may be the result of the high susceptibility of linker DNA to shearing. Linker DNA is not protected by histones and may be easier to break by shearing. As a result, the ends of sheared DNA even in the input are more likely to be in linker DNA and have a higher chance of being ligated by sequencing adapters. The resulting sequenced fragments would show the nucleosomal periodicity that is typically observed in MNase-seq experiments (Figure 3.14).

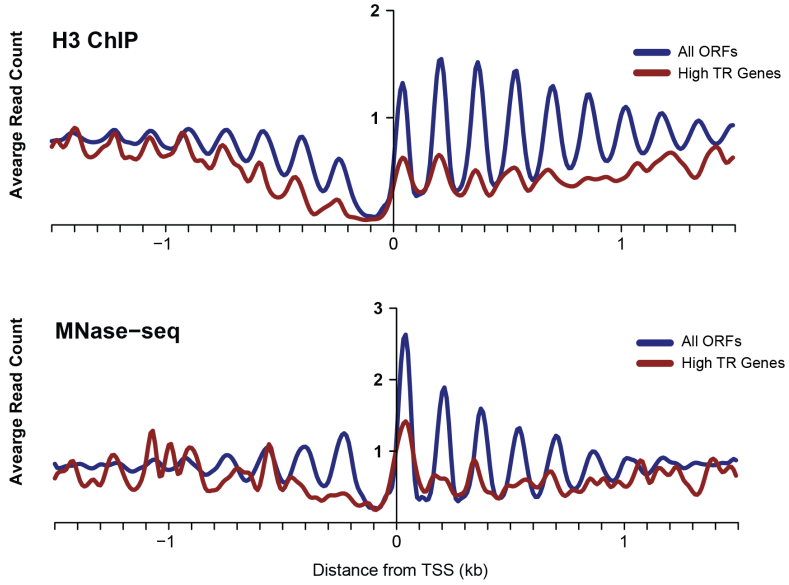

Figure 3.14 Transcription depletes nucleosomes

Both H3 MNase ChIP [77] and MNase-seq from our lab showed lower nucleosome occupancy in the top 100 highly transcribed genes under normal growth conditions

These low-level nucleosomal periodicity signals are not typically of concern in transcription factor ChIP because these experiments usually focus on stronger peaks at regulatory elements. However, the background nucleosomal periodicity may give a misleading picture when analyzing ChIP-seq against proteins that are localized within gene bodies, such as RNAPII-associated factors or chromatin remodelers, which do in fact associate with nucleosomes and/or demonstrate peaks in a similarly low range to the nucleosomal background. Our findings urge careful choice of ChIP-seq normalization controls and call for caution in interpreting the signals from ChIP-seq datasets showing transcription dependent occupancy of proteins over coding regions.

Chapter 4 Genome-wide Chd1 Co-occupancy with Early Transcription Elongation Factors

4.1 Abstract

Chromatin in yeast consists of well-organized nucleosomes that are controlled by adenosine triphosphate (ATP)-dependent chromatin remodeling complexes. One remodeler, chromodomain helicase DNA binding protein 1 (Chd1), plays an integral role in nucleosomal organization as the loss of Chd1 is recognized to cause widespread disruption. Despite its importance, the functional and physical localization of Chd1 on chromatin remains largely unknown and controversial. Here, we quantitatively showed that the deletion of *CHD1* significantly disrupted nucleosome arrays within the gene bodies of highly transcribed genes. Further, using ChIP-seq followed by a quantitative comparison of peak shapes, we found that the structure of the Chd1 occupancy signal for gene bodies was highly similar to that of RNAPII Ser 5-P, not RNAPII Ser 2-P. Follow-up experiments revealed that local RNAPII Ser 5-P occupancy was altered in the *chd1Δ* strain whereas the deletion did not affect RNAPII Ser 2-P occupancy, suggesting that Chd1 is associated with early transcription elongation. Previous studies had suggested that Chd1 is associated with a methylated histone, H3K36me3, found in highly transcribed gene bodies. To investigate this possibility, we mapped genome-wide Chd1 occupancy in a strain lacking the histone methyltransferase for H3K36 (i.e. *set2Δ*). Unexpectedly, deletion of *SET2* did not appear to affect either nucleosome positioning or

Chd1 occupancy. Therefore, it is reasonable to conclude that Chd1 is recruited onto the gene bodies of highly transcribed genes in a Set2-independent manner.

4.2 Introduction

A nucleosome represents the basic unit of chromatin and is typically composed of ~147 bp DNA wrapped around a histone octamer. The biochemical modification and physical position of individual nucleosomes play a critical role in both the structure and transcriptional regulation of chromatin [136]. The advent of microarray and deep sequencing technology has allowed us to comprehensively map nucleosomes and has revealed that nucleosome arrays are well organized *in vivo* [24, 137, 138]. The conserved organization is comprised of two main structures: (i) a nucleosome depleted region (NDR) flanked by -1 and +1 nucleosomes and (ii) well-positioned nucleosomes separated at regular distances by linker DNA. *In vitro* DNA-histone reconstitution assays and *in vivo* micrococcal nuclease (MNase) digestion experiments have shown that preferable DNA sequences and structural features on nucleosomes determine nucleosomal organization [138-140]. In addition, ATP-dependent chromatin remodeling complexes were also revealed to be key determinants of nucleosome organization [141]. High-resolution mapping of chromatin remodelers on chromatin showed that the complexes demonstrate position specificity to nucleosomes relative to the transcription start site (TSS) [42].

Chromatin remodeling complexes control nucleosome turnover, sliding, and

spacing, so it is somewhat surprising that the deletion of a single complex does not necessarily result in catastrophic disruption. That is, global nucleosome positions are not typically altered by a single deletion yet tend to be significantly disrupted by double or triple deletions [26, 27], which suggests that chromatin remodeling complexes operate with redundant functionality. Exceptions to this trend however can be observed. For example, in contrast to other chromatin remodelers, the singular loss of Chd1 severely disrupts well-organized nucleosome arrays in yeast [26, 28, 29].

Originally, high-throughput experiments exploring the functional localization of Chd1 in *Schizosaccharomyces pombe* reported that nucleosome arrays in a strain deleted for *CHD1* were more disorganized at highly transcribed genes [29]. A more recent high-throughput study however showed that genes with high and low transcription rates are equally disrupted [28]. Though conflicting in their interpretations, the two papers actually reported very similar nucleosome profiles. Their divergence highlights lacks of a definitive quantitative method for the comparison of nucleosomal periodicities.

Low-throughput studies regarding the physical localization of Chd1 showed that Chd1 localizes on highly transcribed genes and interacts with transcription elongation factors [142, 143]. Consistent with these observations, Chd1 ChIP-seq confirmed the localization of Chd1 within gene bodies and with high enrichment at highly transcribed genes [26]. Interestingly, the average nucleosome profile of *chd1Δ* showed that the extent of disruption was particularly strong at +2 and later nucleosomes, implying that Chd1 works in non-promoter regions [26, 28, 29]. Although a conflicting report was recently

published showing Chd1 binding to promoters [126], the Chd1 association with transcription elongation continues to be observed. For example, the loss of transcription elongation factors leads to serious nucleosome disruption, and transcription elongation factor mutants display patterns of disorganization very similar to those of *chd1Δ* [144, 145]. While evidence of the physical and functional association of Chd1 with transcription elongation factors continues to accumulate, the exact transcription elongation step involved remains unknown.

Chd1 has two chromodomains that are known to interact with H3K4me3 [146]. In fact, recent mass spectrometry experiments following H3K36me3 IP from mononucleosomes linked Chd1 to H3K36me3 [147]. Two additional independent studies clarified that deletion of *CHD1* does not affect the levels of H3K36me3 but rather moves the distribution of H3K36me3, not H3K4me3, upstream in the gene bodies [31, 147]. This is significant because it suggests that Chd1 plays a role in maintaining the positioning of H3K36me3.

In this study, we first quantitatively compared the shapes of nucleosomal peaks between the WT and *chd1Δ* strains by mathematically smoothing signals and then calculating the Pearson correlations for each gene. This novel approach (termed 'shapeDiff') revealed that Chd1 function is more important at highly transcribed genes. Next, to determine where within these highly transcribed gene bodies Chd1 localizes, we mapped the occupancy of the initiating and elongating forms of RNA polymerase, RNAPII Ser 5-P and Ser 2-P, respectively. Then, we compared these peak shapes with

those of Chd1. A strong similarity between Chd1 and RNAPII Ser 5-P peak shapes suggested that Chd1 co-occupied with early transcription elongation factors. We also found that the local RNAPII Ser 5-P peak shapes were altered when *CHD1* was deleted. Lastly, we tested the possibility that methylated H3K36 may determine Chd1 occupancy. The mapping of Chd1 in *set2Δ* revealed that methylation levels at H3K36 have no effect on Chd1 occupancy and nucleosome organization.

4.3 Materials and Methods

4.3.1 Yeast strains and cell culture

The *S. cerevisiae* BY4741 (MATa *his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*) was used as a wild type strain and background genotype for *chd1Δ* and *set2Δ*. *chd1Δ* strain in Figure 4.1B was obtained from the yeast deletion collection (Open Biosystems) [112]. For *chd1Δ* strain in Figure 4.1A, we deleted *CHD1* by replacing the protein coding region of *CHD1* with the His3MX6 cassette.

All cells were cultured in YPD (yeast extract, peptone, dextrose) media at 30 °C to an A600 OD of 0.8 with shaking 250 rpm. For a heat shock sample, harvested cells were incubated in 39 °C water bath for 15 mins, treated with formaldehyde, and then stored at -80 °C. To tag endogenous Chd1, 13Myc-His3MX6 cassettes were amplified from pFA6a-13Myc-His3MX6, and transformed into WT or *set2Δ* [148]. The cassettes were integrated into the *CHD1* stop codon, and 13XMyc at the C terminal of Chd1 was confirmed by PCR and western blot.

4.3.2 Western Blot

Whole cell extracts were prepared from 30 ml culture of 0.8 OD WT and *set2Δ* cells carrying endogenous 13XMyc tagged Chd1. 30 ul of each extract was run on a 4-20 % gradient SDS-polyacrylamide gel and transferred to a PVDF membrane. In order to confirm 13XMyc tagging of Chd1 and compare the levels of Chd1 expression between WT and *set2Δ,* we detected Chd1 using HRP-conjugated c-Myc antibody (Santa Cruz Biotechnology, 9E10, cat.# sc-40). Reduced levels of H3K36me3 in *set2Δ* were probed with anti-Histone H3 (tri methyl K36) antibody (Abcam, cat# ab9050), and GAPDH antibody (Santa Cruz Biotechnology, FL-335, cat.#sc-25778) was used to visualize loading control proteins.

4.3.3 Chromatin Immunoprecipitation

150 ml cells were treated with formaldehyde to be a final concentration 1 % for 15 min, then quenched with glycine to a final concentration of 125 mM for 5 min. The DNA-protein complexes were sheared by ultra-sound sonication, then incubated overnight with 100 µl of anti-Myc conjugated agarose beads (Sigma Aldrich, cat.# E6654), 8 µg of RNAPII Ser 5-P specific antibody (Abcam, cat.# ab5131), and 8 µg of RNAPII Ser 2-P specific antibody (Abcam, cat.# ab5095) to pull down Chd1, RNAPII Ser 5-P and, Ser 2-P, respectively. Then, for RNAPII ChIP, 100 µl of pre-washed protein A beads were added and incubated for 4 hours. After serial wash steps, immunoprecipitated DNA was recovered with overnight incubation at 65°C water bath followed by ethanol precipitation. Subsequently, sequencing libraries were prepared

using NEBNext® ChIP-Seq Library Prep Master Mix Set (cat.# E6240L) and Bioo multiplex adapter for Illumina, then sequenced in Illumina HiSeq 2000.

4.3.4 Mononucleosome isolation

We followed the mononucleosome isolation protocol described in [24]. Briefly, cells were prepared as described above for ChIP by the quenching step and resuspended in 20 ml of zymolyase buffer. 250 μg of zymolyase (MP Biomedicals, cat.# IC320921) were added to make spheroplasts, then resuspended in 2 ml NP buffer. The spheroplasts were treated with MNase (Worthington Biochemical Corp., cat.# LS004797) at a concentration from 40 U-100 U for 10 min at 37 °C. The DNA-protein complexes were reverse-crosslinked in 10 mM EDTA and 1% SDS buffer with Proteinase K at 65 °C overnight. RNA was removed by RNase A treatment, then DNA was extracted with phenol-chloroform and purified by ethanol precipitation. Finally, DNA was run on an E-gene (Invitrogen), and ~147 bp DNA fragments were size-selected. Library preparation and sequencing were performed as described above for ChIP-seq.

4.3.5 Bioinformatics Analysis

Sequencing reads were mapped onto the sacCer3 reference genome using bwa (version 0.6.2) with default options [73]. Wig files were generated from the bam files and loaded on the UCSC Genome Browser mirror to take snap shots. For shapeDiff analysis, genomic regions between TSS and PAS were divided into bins of 10 bp, and reads were counted. Then, the counts were smoothened using a built-in spline function in R with default parameters (R version 3.0.2). For a given gene, Pearson correlation coefficient

was calculated for the smoothened value of counts between two samples. This process was iterated for every gene that has TSS and PAS coordinate [149].

4.3.6 Accession number

The ChIP-seq data from this study have been deposited in the Gene Expression Omnibus (GEO) database under accession number GSE56061. The MNase-seq data are also available from GEO as accession number GSE56095.

4.4 Results and Discussion

4.4.1 A correlation-based comparison of nucleosome positioning

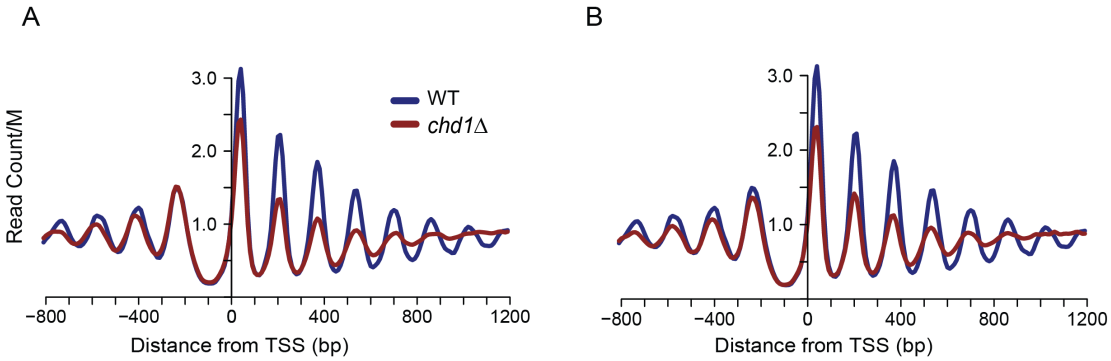

Figure 4.1 Average nucleosome profile of *chd1Δ*

Average nucleosome profiles for all genes show that nucleosome densities are reduced at the gene bodies of *chd1Δ*. The changes are stronger at later nucleosomes than early nucleosomes. Y-axis represents average read counts per million reads (M) (A) *CHD1* is deleted using a His3MX6 cassette (B) Nucleosome disruption is consistently observed in a strain deleted for *CHD1* by a G418 resistance marker

As previously observed in both budding and fission yeast [26, 28, 29], the loss of Chd1 disrupted nucleosome organization within gene bodies (Figure 4.1A). We further confirmed this phenotype in the *chd1Δ* strain with a different resistance marker

(Materials and Methods, Figure 4.1B). Although studies revealed that global nucleosome occupancy is altered, the challenge of identifying the specific genomic loci exhibiting a high level of disruption remained. Here, we developed a simple but powerful approach.

MNase digestion followed by deep sequencing (MNase-seq) has been widely used in studies seeking (i) to define the positions of individual nucleosomes on a genome-wide scale as well as (ii) to investigate changes in nucleosome occupancy. The analysis pipeline adopted the use of peak calling algorithm implemented in ChIP-seq analysis [24, 49, 50]. However, this approach remains three problems. First, the low and dense nucleosomal peaks in MNase-seq make analysis for challenging; traditional peak calling methods rely upon the ability to measure high, individual, and disperse peaks. Second, total number of nucleosome calling is correlated with sequencing depth [49], thus different sequencing depth can be mis-interpreted as nucleosome depletion or acquisition when two MNase-seq data sets are compared. Third, peak height can be varying by artifacts that occur as the result of different MNase digest concentrations [150], but this widespread artifacts have little effect on nucleosomal periodicity. Therefore, studies instead chose to use a different analytical tool – the Pearson correlation – as it conveniently offered a solution to these disadvantages.

The Pearson correlation is a computational technique that can be applied to quantitatively compare two sets of MNase-seq data for levels of nucleosome occupancies at a given loci [150, 151]. It compares correlation coefficients and in doing so observes trends between data sets; by ignoring individual peak data in favor of comparative trends,

the Pearson correlation mitigates the effect of MNase artifacts. Moreover, the Pearson correlation between two MNase-seq data sets at given genomic window is a more accurate quantification to compare nucleosome positioning as peak shape is a more informative indicator to measure nucleosome dynamics than the peak calling utilized with ChIP-seq.

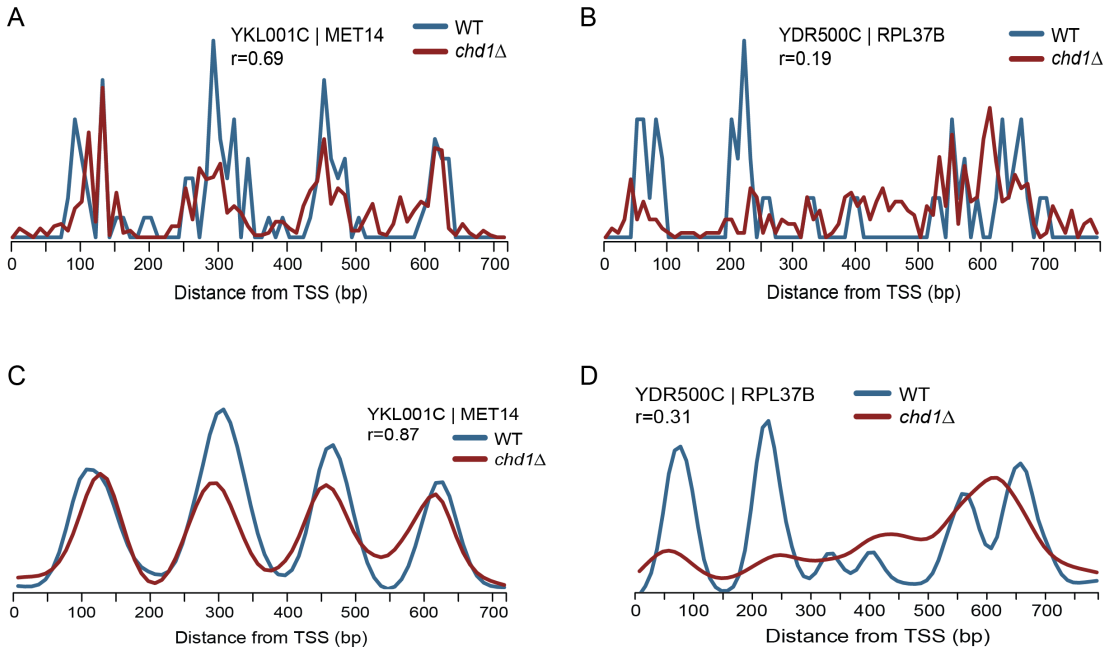

Figure 4.2 Examples of shapeDiff analysis

Y-axis represents normalized read counts (A-B) Before performance of smoothing with a spline function: Spline smoothing removes the noisiness possibly led by low read coverage and produces the smooth pictures of nucleosome occupancy in 'C' and 'D'. (C) *MET14* has a high correlation coefficient due to the high similarity of nucleosome occupancy. (D) shapeDiff estimates the nucleosome occupancy similarity as 0.31 between WT and *chd1Δ* at *RPL37B* gene body.

Thus far, for our experiments MNase-seq analyzed via the Pearson correlation appeared to be the strongest plan, with one exception. Sequencing depth could pose a

critical problem as the noisy signals characterized by low sequencing depth could subsequently result in low correlations. In order to resolve this issue, we implemented a novel method called "shapeDiff analysis" wherein we smoothed nucleosome occupancy signals using a spline function as a preliminary step before the Pearson correlation (Figure 4.2A and 4.2B). Fortunately, due to the nature of the Pearson correlation, the process of normalizing sequencing depth (i.e. multiplying or dividing peaks by a scaling factor) should not affect any correlation calculations.

In our first experiments using shapeDiff analysis, we began by comparing WT and *chd1Δ* strains, focusing on a window ranging from the transcription start site (TSS) to the polyadenylation site (PAS) for all genes with TSS and PAS annotation [149]. For example, *MET14* has 4 distinct nucleosomes within the gene body, both for the WT and the mutant (Figure 4.2C); in accordance with the small nucleosome shifts observed between the two strains, shapeDiff measured a high correlation for the nucleosomes. In contrast, the densities of the +1 and +2 nucleosomes for *RPL37B* were visibly dramatically reduced in *chd1Δ* and shapeDiff showed that nucleosomal periodicity disappeared altogether at the 3' end of the gene (Figure 4.2D).

Globally, we observed that when comparing WT and *chd1Δ* nucleosome positioning the Pearson correlation is significantly lower at highly transcribed genes, indicating that the mutant nucleosome arrays experience greater disorganization at loci with high transcription rates (Figure 4.3B).

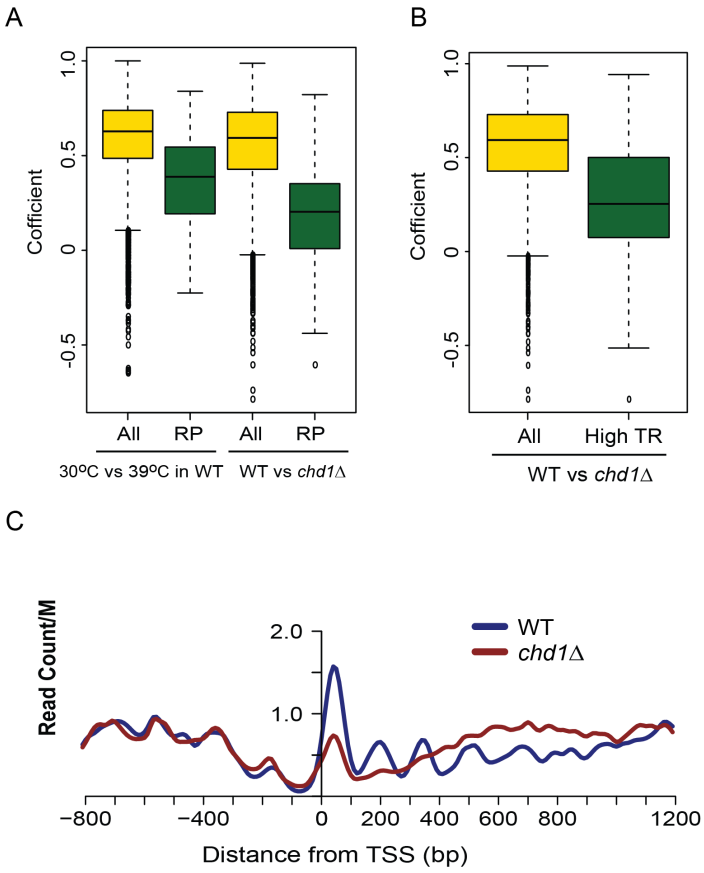

Figure 4.3 Functional Chd1 localization at highly transcribed genes

(A) Upon heat shock, RP genes are highly down-regulated [120] and the nucleosome arrays are also significantly disorganized [24]. The disorganization extents at RP gene bodies by the loss of Chd1 are severer than that those by heat shock. (B) Nucleosome arrays are significantly disrupted at the gene bodies of highly transcribed genes in *chd1Δ* (D) Average nucleosome profile for RP genes confirms severe nucleosome disruption at RP genes by deletion of *CHD1*

In order to understand the relative value of this *chd1Δ* disruption, we compared it to heat shock conditions – a biological phenomenon also known to disrupt the nucleosome organization of highly transcribed genes. Specifically, we compared (i) the correlation between WT ribosomal protein (i.e. RP) genes under normal conditions and heat shock and (ii) the correlation between RP genes in WT and *chd1Δ* strains. RP genes

serve as a good comparative measure for nucleosome occupancy affected by Chd1 deletion because (i) they exhibit high transcription levels, (ii) they are dramatically repressed by heat shock conditions, and (iii) their nucleosomes are significantly depleted [24, 120]. For RP genes, the median correlation of nucleosome occupancies between normal and heat shock conditions in WT was 0.39, whereas the median correlation between WT and *chd1Δ* was 0.20. This suggests that the deletion of *CHD1* more strongly depleted nucleosome arrays at highly transcribed genes than acute heat shock did (Figure 4.3A and 4.3C).

4.4.2 The Chd1 binding peak shape is similar to RNAPII Ser 5-P

The first low-throughput experiments performing Chd1 ChIP followed by qPCR showed that Chd1 localizes within gene bodies [142, 152]. More recently, two high-throughput studies characterized genome-wide Chd1 occupancy [26, 126], but they drew conflicting conclusions with respect to global Chd1 localization. One study performed native MNase ChIP-seq for Chd1 that was tagged with 3 FLAG repeats [126]; the authors showed that Chd1 bound to promoters. Another study used formaldehyde-treated ChIP-seq after tagging Chd1 with 13 Myc repeats (i.e. Chd1-13XMyc) [26]; their results revealed that Chd1 localized within gene bodies. For our experiments, we repeated the latter approach, generating Chd1-13XMyc and immunoprecipitating Chd1 after formaldehyde treatment. Consistent with the latter study, we observed Chd1 occupancy within gene bodies (Figure 4.4A). Additionally, we observed that Chd1 occupancy appeared similar to RNAPII signals, which are associated with various steps in

transcription. As such, we pulled down two different active elongating RNAPII (RNAPII Ser 5-P and RNAPII Ser 2-P) and performed deep sequencing to investigate whether Chd1 is associated with the early or late elongation step in transcription. From the initial wide view, Chd1 occupancy appeared indistinguishable between the published Chd1 data, our own Chd1 data, RNAPII Ser 5-P occupancy, and RNAPII Ser 2-P occupancy (Figure 4.4A). When we zoomed in however and examined the Chd1 peak shapes at individual genes, surprisingly the Chd1 peaks appeared similar to the peaks of RNAPII Ser 5-P but not to those of RNAPII Ser 2-P (Figure 4.4B). To quantify this perceived similarity in peak shape we applied shapeDiff analysis for Chd1 and RNAPII (just as done before with nucleosome occupancy) as calculating the correlation for peak shapes demanded a more sophisticated approach than the simple measurement of peak height. Once again, the selected windows spanned from TSS to PAS for each gene. The correlation distribution of Chd1 and RNAPII Ser 5-P skewed towards a positive correlation (median = 0.54) but the distribution of Chd1 and RNAPII Ser 2-P was centered on 0 (median = 0.04) (Figure 4.4C). The peak shape correlations confirmed that Chd1 is co-localized with early transcription elongation factors and not late transcription elongation factors.

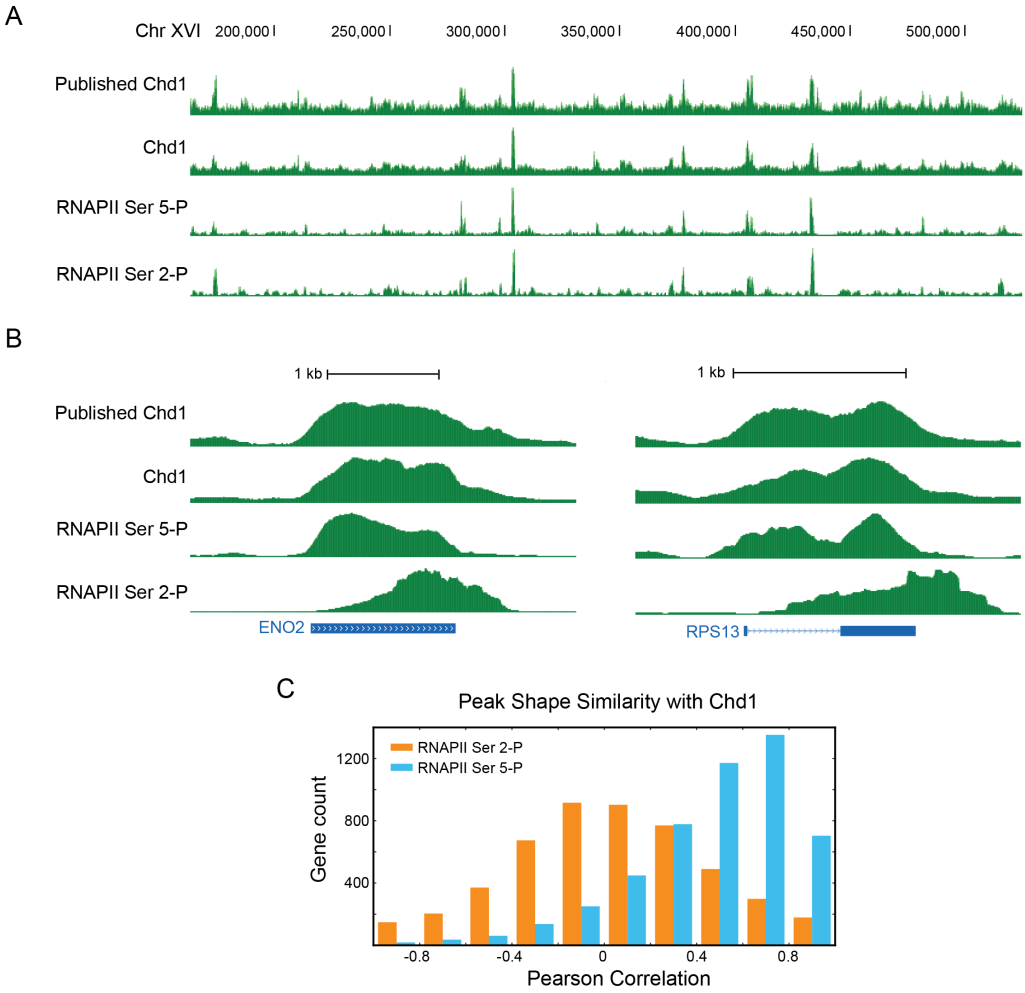

Figure 4.4 Chd1 co-occupancy with RNAPII Ser 5-P

Chd1 co-localizes with elongating RNAPII, and the binding peak shapes of Chd1 are highly similar to those of RNAPII Ser 5-P. (A) In a wide view, Chd1 occupancy is similar to localization of RNAPII Ser 5-P and Ser 2-P. (B) Zoomed-in snap shots for some highly expressed genes reveal that Chd1 peak shapes appear similar RNAPII Ser 5-P peak shapes, but not Ser 2-P shapes. (C) The peak shapes of Chd1 are compared with the peaks of either RNAPII Ser5-P or Ser 2-P. Histogram of correlation coefficients shows that RNAPII Ser 2-P has no correlation with Chd1 at level of local peak shapes on a genome wide scale. In contrast, 2056 genes have over correlation coefficient 0.6 in the peak shape comparison between Chd1 and RNAPII Ser 5-P.

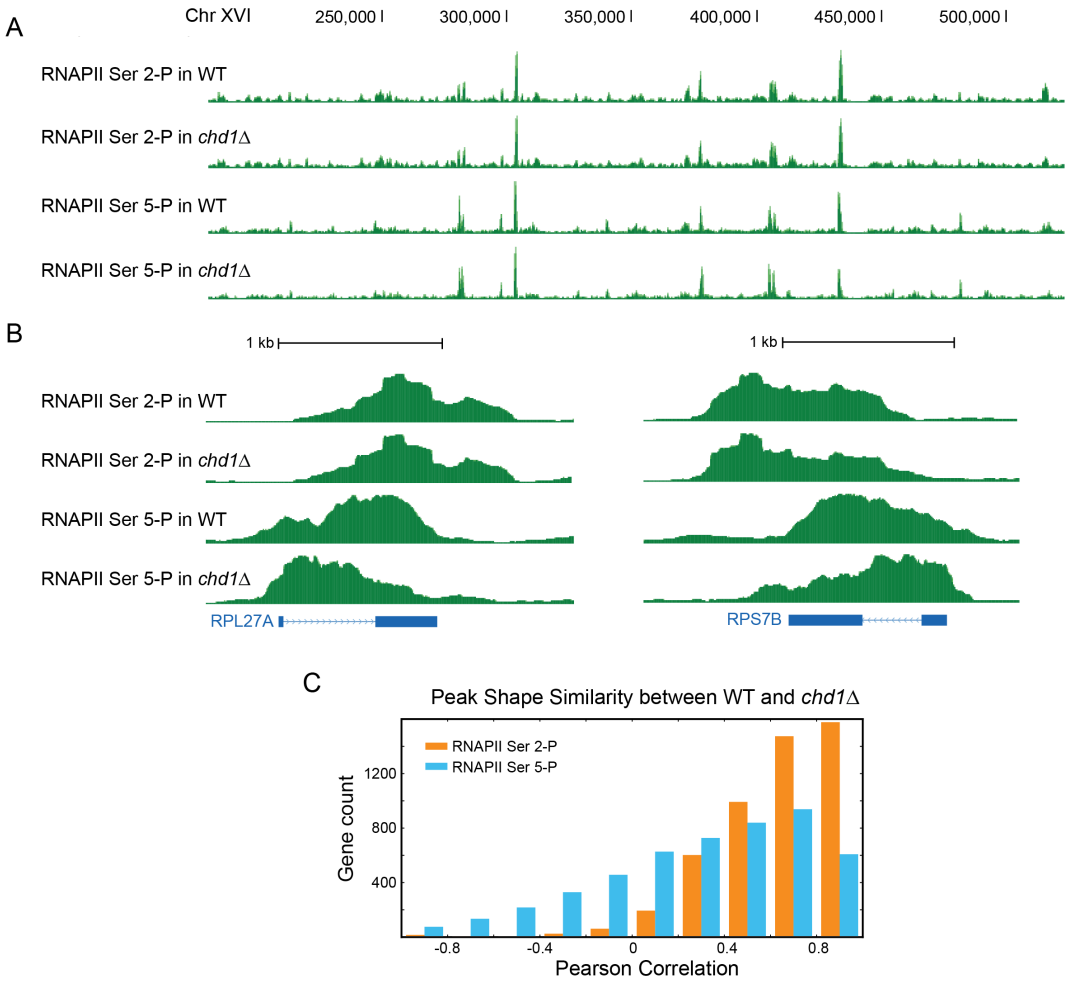

Figure 4.5 Loss of Chd1 leads to changes in local occupancy of RNAPII Ser 5-P

(A) Zoomed-out snap shot cannot discriminate elongating RNAPII localization between WT and *chd1Δ*. (B) In a narrow view, the peak shapes of RNAPII Ser 5-P shift upstream at highly transcribed genes in *chd1Δ*, but RNAPII Ser 2-P peaks appear identical between WT and *chd1Δ*. (C) shapeDiff analysis quantifies the peak shape similarity of either RNAPII Ser 5-P or RNAPII Ser 2-P between WT and *chd1Δ*. The effect of *CHD1* deletion on changes in peak shape is much stronger for RNAPII Ser 5-P

4.4.3 The loss of Chd1 alters the peak shapes of RNAPII Ser 5-P

Gene expression levels were only slightly altered by the loss of Chd1 [30, 32], but cryptic and antisense transcription were notably increased in the absence of Chd1 [28, 29]. Based on these observations, we hypothesized that the loss of Chd1 does not lead to the complete delocalization of elongating RNAPII but rather alters its local positioning. To test this hypothesis, we mapped the occupancy of RNAPII Ser 5-P and RNAPII Ser 2-P separately in a strain carrying *chd1Δ*. In the wide view of our results we did not observe changes in either elongating RNAPII occupancy (Figure 4.5A). When we examined individual peaks however, the peak shapes of RNAPII Ser 5-P were disrupted at some highly transcribed genes, especially RP genes (Figure 4.5B). In order to measure the peak shape similarities of elongating RNAPII between WT and *chd1Δ* strains on a genome-wide scale, we performed shapeDiff to generate correlations for each gene. Interestingly, RNAPII Ser 2-P was unaffected by the loss of Chd1 (median=0.69), but relative to WT the correlation in peak shapes for RNAPII Ser 5-P decreased by deleting *CHD1* (median=0.38) (Figure 4.5C). This implies that Chd1 may regulate the processivity of early transcription elongation machinery but not late stage.

4.4.4 The deletion of *SET2* does not appear to affect Chd1 occupancy or nucleosome positioning

Although Chd1 chromodomains interact with H3K4me3 [146], some evidence has suggested that Chd1 may functionally and physically associate with H3K36 methyl groups. First, Chd1 was shown to be associated with H3K36me3 by mass spectrometry

[147]. Second, H3K36me3 was shifted upstream by the loss of Chd1 whereas H3K4me3 showed no change [31, 147]. Third, nucleosome disruption in *chd1Δ* occurred mainly at the +2 nucleosome as well as later nucleosomes where H3K36 methylation is more abundant than H3K4 methylation [26]. Fourth, H3K36me3 is a well-known mark within the gene bodies of highly transcribed genes during transcription elongation while we separately showed in our experiments that Chd1 localized within gene bodies as well [153-155]. Fifth, RNAPII Ser 5-P was co-purified with the H3K36 methyltransferase Set2 [156].

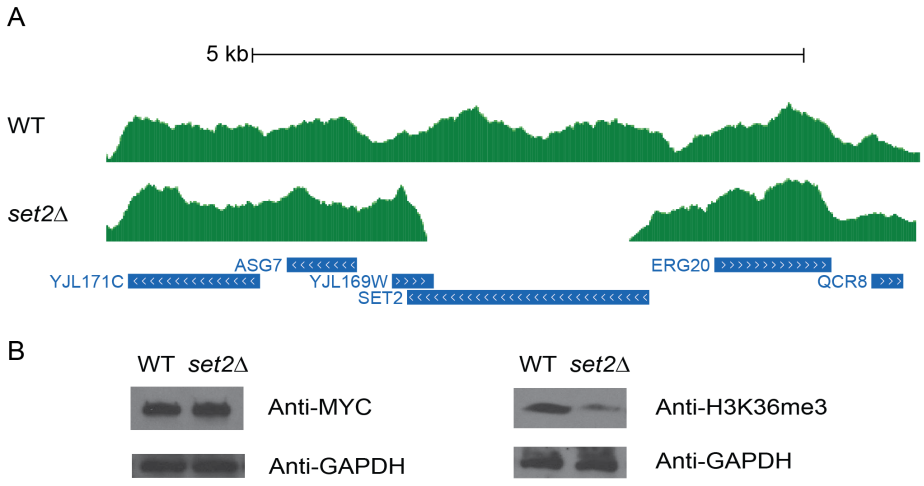

Figure 4.6 Confirmation of *SET2* deletion and H3K36me3 levels

(A) Chd1 ChIP-seq in WT and *set2Δ*. Almost no read on the *SET2* genomic region confirms that successful deletion of *SET2* in the Chd1 ChIP-seq data. (B) Chd1 is tagged with 13 repeats of Myc. Deletion of *SET2* does not affect the expression levels of Chd1, but the loss of Set2 significantly reduces H3K36me3

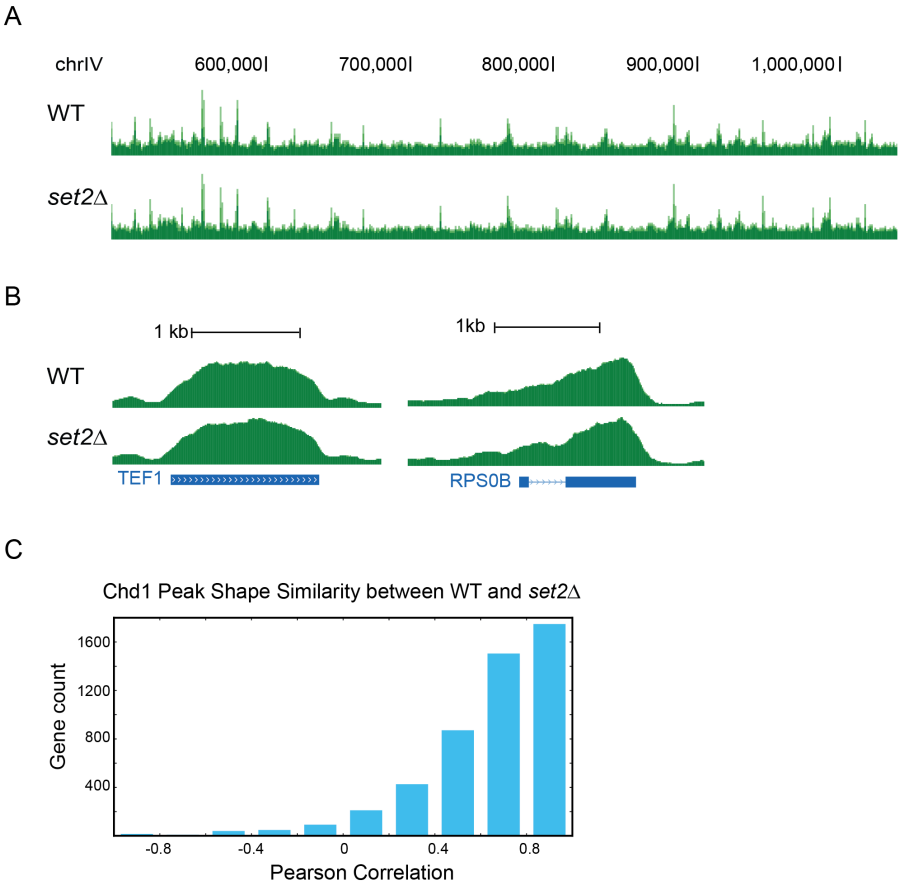

Figure 4.7 Chd1 localization on chromatin is Set2-independent

(A, B) The peaks shapes of Chd1 binding are indistinguishable between WT and *set2Δ*. (C) shapeDiff analysis re-confirms that *set2Δ* has no effect on Chd1 occupancy on a genome wide scale.

Based on this evidence, we hypothesized that Chd1 is recruited onto chromatin via the recognition of H3K36me3. To test this hypothesis, we generated a Chd1-13XMyc strain carrying a *SET2* deletion and measured the Chd1 occupancy in the mutant. Sequencing data confirmed the successful deletion of *SET2* in that only a few reads were mapped to the *SET2* gene body (Figure 4.6A). Immunoblotting revealed that the levels of H3K36me3 were greatly reduced by loss of Set2, however the loss did not alter

expression levels of Chd1 (Figure 4.6B). Despite the evidence in support of our hypothesis, global Chd1 occupancy in *set2Δ* appeared identical to that in WT (Figure 4.7A). Even looking more closely at a few individual genes, no changes were observed in the Chd1 peak shapes (Figure 4.7B). Additionally, shapeDiff analysis also revealed a high similarity between Chd1 occupancies in WT and *set2Δ* (median=0.71) (Figure 4.7C). To place this in context, the shapeDiff analysis of two independent replicates WT Chd1 ChIP-seq yielded a median value of 0.65. Thus, we concluded that loss of Set2 had no effect on Chd1 occupancy, suggesting that Chd1 occupancy within gene bodies is Set2-independent. This observation was also supported by the fact that *set2Δ* has normal nucleosome organization (Figure 4.8A and 4.8B). The average nucleosome profile of *set2Δ* displayed well-organized nucleosome arrays similar to WT as well as no difference in shapeDiff analysis. Therefore, we conclude that Chd1 is recruited onto chromatin in a H3K36 methylation-independent manner and that methylation at H3K36 has no effect on well-organized nucleosome arrays.

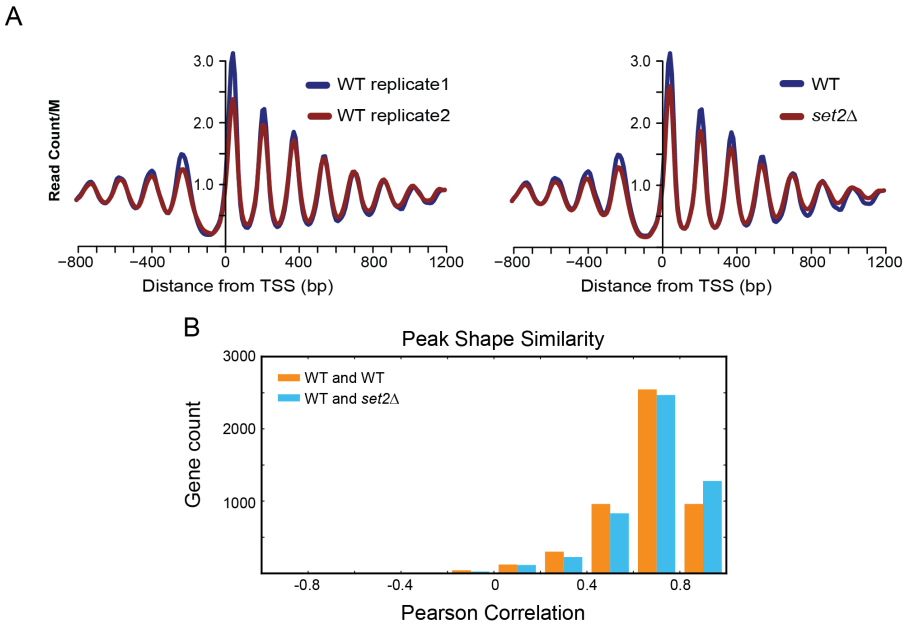

Figure 4.8 Loss of Set2 has no effect on nucleosome organization

(A) Average nucleosome profile represents that high degree of overlap between WT replicates, and the similarity between WT and *set2Δ* is as close as the WT replicates similarity. (B) Peak shape comparison by shapeDiff shows that nucleosome occupancy in *set2Δ* has no difference from that in WT on a genome wide scale.

Chapter 5 Summary and Future Direction

Using SMORE-seq described in chapter 2, we comprehensively mapped TSS and PAS at single-nucleotide resolution in yeast and identified potentially mis-annotated genes that have TSS within their internal coding regions. However, we cannot rule out the possibility that those genes are isoforms and have multiple start codons with different frequency of usage. One of the internal TSS genes is *ISW1*, which is a well-studied and characterized chromatin remodeler [147, 157]. If the ATG downstream of our reported internal TSS is the start codon of a new isoform, functional comparison between the two isoform Isw1 proteins for chromatin remodeling could prove illuminating. If the internal TSS is the only TSS for *ISW1*, the current SGD annotation is wrong and should be corrected as soon as possible for the community.

Another key finding in chapter 2 is widespread bidirectional transcription in WT. Although the bidirectional nature of transcription has been previously reported [98, 158, 159], the role of bidirectional transcripts in cells is largely unknown, and mechanistic models by which the ncRNA are controlled are poorly proposed. SMORE-seq in a variety of yeast strains that have slightly different promoters can elucidate how promoter sequences affect levels of bidirectional transcription. Additionally, SMORE-seq in TF-deletion strains can provide clues for understanding the role of promoter-associated proteins in transcriptional directionality.

In chapter 3, we showed that unrelated TFs appear to bind to highly transcribed loci in ChIP-seq, and that this artifact was also present in ChIP-seq from some human cell lines. Recent large-scale analyses of co-occupancy showed that a large number of TFs co-localize with RNAPII at highly transcribed genes [118, 119]. It is possible that this observation could be the ChIP-seq expression bias without any biological meaning. We also cautiously speculate that a portion of the signals reported at super-enhancers may be explained by the ChIP-seq expression bias because super-enhancers are bound by many TFs and are highly transcribed [160, 161]. Further and careful investigation is necessary to discriminate biologically meaningful signals from technically biased ones. To achieve this goal, affinity-purified naturally isolated chromatin (ORGANIC) profiling was proposed and shown to remove ChIP signals at known, potentially artifactual hotspots [43]. However, the native ChIP-seq data in Figure 3.10 of this dissertation were produced by the ORGANIC method, and the ChIP-seq expression bias was still observed [126], suggesting that the performance of ORGANIC could vary from sample to sample. Therefore, development of robust experimental and computational methods is critical to accurately map protein localization on chromatin on a genome wide scale.

Another bias we showed in chapter 3 was a nucleosomal periodicity in ChIP-seq signals. For proteins that bind to gene bodies, this bias can suggest a misleading interaction of the immunoprecipitated proteins with nucleosomes. Therefore, input data

that are prepared in parallel (with the same extent of cross-linking and sonication) is an essential control to correctly analyze ChIP-seq data.

From the chapter about the effect of Chd1 on nucleosomes, we saw that Chd1 binds to early transcription elongation regions in a H3K36 methylation independent manner. This study together with previous reports [26, 28, 30-32, 147] leaves three challenging questions: i) Why does loss of Chd1 have little effect on transcription and growth defects although nucleosomes are severely disrupted? ii) Why does loss of histone marks (e.g. H3K36me3) not affect either nucleosome positioning or occupancy of transcription elongation machinery? iii) How is Chd1 recruited onto chromatin? To speculate on the first question, it is important to note that nucleosome positioning regulates transcription [22], but conversely transcription elongation also influences nucleosome positioning [144, 162]. The latter observation suggests that loss of Chd1 as a transcription elongation factor might slow down transcription elongation but not lead to significant change in the abundance of transcripts. Instead, a reduced transcription rate could disorganize nucleosome arrays. Regarding the second question, ChIP-seq for transcription elongation factors in the mutants and MNase-seq in the mutants may not be sensitive enough to capture the dynamic changes from ensembles of millions of cells. High-resolution biochemical methods could be more appropriate methods for monitoring the dynamics of chromatin-associated proteins at single molecule levels in single cells. Lastly, it is possible that transcription elongation factors recruit Chd1 onto chromatin based on the

observation that Chd1 directly interacts with the PAF complex and that occupancy patterns of Chd1 appear nearly identical to those of early transcription elongation factors. Therefore, Chd1 ChIP-seq in strains deleted for transcription elongation factors will be important experiments to examine to determine whether interaction of Chd1 with transcription elongation factors is essential for its localization on chromatin.

References

1. Crick, F., *Central dogma of molecular biology*. Nature, 1970. **227**(5258): p. 561-3.
2. Goff, S.P., *Host factors exploited by retroviruses*. Nat Rev Microbiol, 2007. **5**(4): p. 253-63.
3. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14*. Cell, 1993. **75**(5): p. 843-54.
4. Mercer, T.R., M.E. Dinger, and J.S. Mattick, *Long non-coding RNAs: insights into functions*. Nat Rev Genet, 2009. **10**(3): p. 155-9.
5. Kim, T.K., et al., *Widespread transcription at neuronal activity-regulated enhancers*. Nature, 2010. **465**(7295): p. 182-7.
6. Alberts, B., *Molecular biology of the cell*. 5th ed. 2008, New York: Garland Science.
7. Marstrand, T.T. and J.D. Storey, *Identifying and mapping cell-type-specific chromatin programming of gene expression*. Proc Natl Acad Sci U S A, 2014. **111**(6): p. E645-54.
8. Lee, B.K., et al., *Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells*. Genome Res, 2012. **22**(1): p. 9-24.
9. Takahashi, K. and S. Yamanaka, *Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors*. Cell, 2006. **126**(4): p. 663-76.
10. Guo, Y., et al., *Evidence for a mechanism of repression of heat shock factor 1 transcriptional activity by a multichaperone complex*. J Biol Chem, 2001. **276**(49): p. 45791-9.
11. Meek, D.W., *Tumour suppression by p53: a role for the DNA damage response?* Nat Rev Cancer, 2009. **9**(10): p. 714-23.
12. Lubbe, S.J., et al., *The 14q22.2 colorectal cancer variant rs4444235 shows cis-acting regulation of BMP4*. Oncogene, 2012. **31**(33): p. 3777-84.
13. Huang, F.W., et al., *Highly recurrent TERT promoter mutations in human melanoma*. Science, 2013. **339**(6122): p. 957-9.
14. Bose, T. and J.L. Gerton, *Cohesinopathies, gene expression, and chromatin organization*. J Cell Biol, 2010. **189**(2): p. 201-10.
15. Akirav, E.M., N.H. Ruddle, and K.C. Herold, *The role of AIRE in human autoimmune disease*. Nat Rev Endocrinol, 2011. **7**(1): p. 25-33.
16. Myer, V.E. and R.A. Young, *RNA polymerase II holoenzymes and subcomplexes*. J Biol Chem, 1998. **273**(43): p. 27757-60.
17. Orphanides, G., T. Lagrange, and D. Reinberg, *The general transcription factors of RNA polymerase II*. Genes Dev, 1996. **10**(21): p. 2657-83.
18. Basehoar, A.D., S.J. Zanton, and B.F. Pugh, *Identification and distinct regulation*

of yeast TATA box-containing genes. Cell, 2004. **116**(5): p. 699-709.
19. Rhee, H.S. and B.F. Pugh, *Genome-wide structure and organization of eukaryotic pre-initiation complexes*. Nature, 2012. **483**(7389): p. 295-301.
20. Spitz, F. and E.E. Furlong, *Transcription factors: from enhancer binding to developmental control*. Nat Rev Genet, 2012. **13**(9): p. 613-26.
21. Lander, E.S., et al., *Initial sequencing and analysis of the human genome*. Nature, 2001. **409**(6822): p. 860-921.
22. Jiang, C. and B.F. Pugh, *Nucleosome positioning and gene regulation: advances through genomics*. Nat Rev Genet, 2009. **10**(3): p. 161-72.
23. Luo, Y., et al., *Nucleosomes accelerate transcription factor dissociation*. Nucleic Acids Res, 2014. **42**(5): p. 3017-27.
24. Shivaswamy, S., et al., *Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation*. PLoS Biol, 2008. **6**(3): p. e65.
25. Varga-Weisz, P., *ATP-dependent chromatin remodeling factors: nucleosome shufflers with many missions*. Oncogene, 2001. **20**(24): p. 3076-85.
26. Gkikopoulos, T., et al., *A role for Snf2-related nucleosome-spacing enzymes in genome-wide nucleosome organization*. Science, 2011. **333**(6050): p. 1758-60.
27. Tirosh, I., N. Sigal, and N. Barkai, *Widespread remodeling of mid-coding sequence nucleosomes by Isw1*. Genome Biol, 2010. **11**(5): p. R49.
28. Hennig, B.P., et al., *Chd1 chromatin remodelers maintain nucleosome organization and repress cryptic transcription*. EMBO Rep, 2012. **13**(11): p. 997-1003.
29. Shim, Y.S., et al., *Hrp3 controls nucleosome positioning to suppress non-coding transcription in eu- and heterochromatin*. EMBO J, 2012. **31**(23): p. 4375-87.
30. Tran, H.G., et al., *The chromo domain protein chd1p from budding yeast is an ATP-dependent chromatin-modifying factor*. EMBO J, 2000. **19**(10): p. 2323-31.
31. Radman-Livaja, M., et al., *A key role for Chd1 in histone H3 dynamics at the 3' ends of long genes in yeast*. PLoS Genet, 2012. **8**(7): p. e1002811.
32. Lee, J.S., et al., *Codependency of H2B monoubiquitination and nucleosome reassembly on Chd1*. Genes Dev, 2012. **26**(9): p. 914-9.
33. Kim, V.N. and J.W. Nam, *Genomics of microRNA*. Trends Genet, 2006. **22**(3): p. 165-73.
34. Wilson, R.C. and J.A. Doudna, *Molecular mechanisms of RNA interference*. Annu Rev Biophys, 2013. **42**: p. 217-39.
35. Xu, Z., et al., *Bidirectional promoters generate pervasive transcription in yeast*. Nature, 2009. **457**(7232): p. 1033-7.
36. van Dijk, E.L., et al., *XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast*. Nature, 2011. **475**(7354): p. 114-7.
37. Stekel, D., *Microarray bioinformatics*. 2003, Cambridge ; New York: Cambridge University Press. xiv, 263 p., 8 p. of plates.
38. David, L., et al., *A high-resolution map of transcription in the yeast genome*. Proc Natl Acad Sci U S A, 2006. **103**(14): p. 5320-5.

39. Metzker, M.L., *Sequencing technologies - the next generation*. Nat Rev Genet, 2010. **11**(1): p. 31-46.
40. Gilmour, D.S. and J.T. Lis, *In vivo interactions of RNA polymerase II with genes of Drosophila melanogaster*. Mol Cell Biol, 1985. **5**(8): p. 2009-18.
41. Schmiedeberg, L., et al., *A temporal threshold for formaldehyde crosslinking and fixation*. PLoS One, 2009. **4**(2): p. e4636.
42. Yen, K., et al., *Genome-wide nucleosome specificity and directionality of chromatin remodelers*. Cell, 2012. **149**(7): p. 1461-73.
43. Kasinathan, S., et al., *High-resolution mapping of transcription factor binding sites on native chromatin*. Nat Methods, 2014. **11**(2): p. 203-9.
44. Auerbach, R.K., et al., *Mapping accessible chromatin regions using Sono-Seq*. Proc Natl Acad Sci U S A, 2009. **106**(35): p. 14926-31.
45. Park, D., et al., *Widespread misinterpretable ChIP-seq bias in yeast*. PLoS One, 2013. **8**(12): p. e83506.
46. Kim, J., et al., *Use of in vivo biotinylation to study protein-protein and protein-DNA interactions in mouse embryonic stem cells*. Nat Protoc, 2009. **4**(4): p. 506-17.
47. Ghaemmaghami, S., et al., *Global analysis of protein expression in yeast*. Nature, 2003. **425**(6959): p. 737-41.
48. Radman-Livaja, M. and O.J. Rando, *Nucleosome positioning: how is it established, and why does it matter?* Dev Biol, 2010. **339**(2): p. 258-66.
49. Jiang, C. and B.F. Pugh, *A compiled and systematic reference map of nucleosome positions across the Saccharomyces cerevisiae genome*. Genome Biol, 2009. **10**(10): p. R109.
50. Chen, K., et al., *Stabilization of the promoter nucleosomes in nucleosome-free regions by the yeast Cyc8-Tup1 corepressor*. Genome Res, 2013. **23**(2): p. 312-22.
51. Park, P.J., *ChIP-seq: advantages and challenges of a maturing technology*. Nat Rev Genet, 2009. **10**(10): p. 669-80.
52. Ozsolak, F., et al., *Direct RNA sequencing*. Nature, 2009. **461**(7265): p. 814-8.
53. Levin, J.Z., et al., *Comprehensive comparative analysis of strand-specific RNA sequencing methods*. Nat Methods, 2010. **7**(9): p. 709-15.
54. Valen, E., et al., *Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE*. Genome Res, 2009. **19**(2): p. 255-65.
55. Hoque, M., et al., *Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing*. Nat Methods, 2013. **10**(2): p. 133-9.
56. Wan, Y., et al., *Landscape and variation of RNA secondary structure across the human transcriptome*. Nature, 2014. **505**(7485): p. 706-9.
57. Mignone, F., et al., *Untranslated regions of mRNAs*. Genome Biol, 2002. **3**(3): p. REVIEWS0004.
58. Zhang, Z. and F.S. Dietrich, *Mapping of transcription start sites in Saccharomyces cerevisiae using 5' SAGE*. Nucleic Acids Res, 2005. **33**(9): p. 2838-51.

59. Miura, F., et al., *A large-scale full-length cDNA analysis to explore the budding yeast transcriptome*. Proc Natl Acad Sci U S A, 2006. **103**(47): p. 17846-51.
60. Nagalakshmi, U., et al., *The transcriptional landscape of the yeast genome defined by RNA sequencing*. Science, 2008. **320**(5881): p. 1344-9.
61. Yamashita, R., et al., *Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis*. Genome Res, 2011. **21**(5): p. 775-89.
62. Gu, W., et al., *CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as C. elegans piRNA precursors*. Cell, 2012. **151**(7): p. 1488-500.
63. Olivarius, S., C. Plessy, and P. Carninci, *High-throughput verification of transcriptional starting sites by Deep-RACE*. Biotechniques, 2009. **46**(2): p. 130-2.
64. Jenal, M., et al., *The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites*. Cell, 2012. **149**(3): p. 538-53.
65. Derti, A., et al., *A quantitative atlas of polyadenylation in five mammals*. Genome Res, 2012. **22**(6): p. 1173-83.
66. Ozsolak, F., et al., *Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation*. Cell, 2010. **143**(6): p. 1018-29.
67. Fu, Y., et al., *Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing*. Genome Res, 2011. **21**(5): p. 741-7.
68. Jan, C.H., et al., *Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs*. Nature, 2011. **469**(7328): p. 97-101.
69. Shepard, P.J., et al., *Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq*. RNA, 2011. **17**(4): p. 761-72.
70. Pelechano, V., W. Wei, and L.M. Steinmetz, *Extensive transcriptional heterogeneity revealed by isoform profiling*. Nature, 2013. **497**(7447): p. 127-31.
71. Rhee, H.S. and B.F. Pugh, *Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution*. Cell, 2011. **147**(6): p. 1408-19.
72. Iyer, V. and K. Struhl, *Absolute mRNA levels and transcriptional initiation rates in Saccharomyces cerevisiae*. Proc Natl Acad Sci U S A, 1996. **93**(11): p. 5208-12.
73. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform*. Bioinformatics, 2009. **25**(14): p. 1754-60.
74. Li, H., et al., *The Sequence Alignment/Map format and SAMtools*. Bioinformatics, 2009. **25**(16): p. 2078-9.
75. Lipson, D., et al., *Quantification of the yeast transcriptome by single-molecule sequencing*. Nat Biotechnol, 2009. **27**(7): p. 652-8.
76. Ingolia, N.T., et al., *Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling*. Science, 2009. **324**(5924): p. 218-23.
77. Zhang, L., H. Ma, and B.F. Pugh, *Stable and dynamic nucleosome states during a meiotic developmental process*. Genome Res, 2011. **21**(6): p. 875-84.

78. Fan, X., et al., *Nucleosome depletion at yeast terminators is not intrinsic and can occur by a transcriptional mechanism linked to 3'-end formation*. Proc Natl Acad Sci U S A, 2010. **107**(42): p. 17945-50.
79. Lemon, B. and R. Tjian, *Orchestrated response: a symphony of transcription factors for gene control*. Genes Dev, 2000. **14**(20): p. 2551-69.
80. Hampsey, M., *Molecular genetics of the RNA polymerase II general transcriptional machinery*. Microbiol Mol Biol Rev, 1998. **62**(2): p. 465-503.
81. Crooks, G.E., et al., *WebLogo: a sequence logo generator*. Genome Res, 2004. **14**(6): p. 1188-90.
82. Iyer, V.R., *Nucleosome positioning: bringing order to the eukaryotic genome*. Trends Cell Biol, 2012. **22**(5): p. 250-6.
83. Kim, H., et al., *Gene-specific RNA polymerase II phosphorylation and the CTD code*. Nat Struct Mol Biol, 2010. **17**(10): p. 1279-86.
84. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes*. Genome Res, 2005. **15**(8): p. 1034-50.
85. Kellis, M., et al., *Sequencing and comparison of yeast species to identify genes and regulatory elements*. Nature, 2003. **423**(6937): p. 241-54.
86. Cliften, P., et al., *Finding functional features in Saccharomyces genomes by phylogenetic footprinting*. Science, 2003. **301**(5629): p. 71-6.
87. Gingold, H. and Y. Pilpel, *Determinants of translation efficiency and accuracy*. Mol Syst Biol, 2011. **7**: p. 481.
88. Kozak, M., *Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes*. Cell, 1986. **44**(2): p. 283-92.
89. Chen, C.Y. and A.B. Shyu, *Mechanisms of deadenylation-dependent decay*. Wiley Interdiscip Rev RNA, 2011. **2**(2): p. 167-83.
90. Zhao, J., L. Hyman, and C. Moore, *Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis*. Microbiol Mol Biol Rev, 1999. **63**(2): p. 405-45.
91. Tan-Wong, S.M., et al., *Gene loops enhance transcriptional directionality*. Science, 2012. **338**(6107): p. 671-5.
92. van Dijk, E.L., et al., *Suppl_ XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast*. Nature, 2011. **475**: p. 114-7.
93. Neil, H., et al., *Widespread bidirectional promoters are the major source of cryptic transcripts in yeast*. Nature, 2009. **457**(7232): p. 1038-42.
94. Nock, A., et al., *Mediator-regulated transcription through the +1 nucleosome*. Mol Cell, 2012. **48**(6): p. 837-48.
95. Itoh, M., et al., *Automated workflow for preparation of cDNA for cap analysis of gene expression on a single molecule sequencer*. PLoS One, 2012. **7**(1): p. e30809.
96. Tian, B. and J.L. Manley, *Alternative cleavage and polyadenylation: the long and short of it*. Trends Biochem Sci, 2013. **38**(6): p. 312-20.
97. Moqtaderi, Z., et al., *Species-specific factors mediate extensive heterogeneity of mRNA 3' ends in yeasts*. Proc Natl Acad Sci U S A, 2013. **110**(27): p. 11073-8.

98. Almada, A.E., et al., *Promoter directionality is controlled by U1 snRNP and polyadenylation signals*. Nature, 2013. **499**(7458): p. 360-3.
99. Yassour, M., et al., *Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species*. Genome Biol, 2010. **11**(8): p. R87.
100. Murray, S.C., et al., *A pre-initiation complex at the 3'-end of genes drives antisense transcription independent of divergent sense transcription*. Nucleic Acids Res, 2012. **40**(6): p. 2432-44.
101. Furey, T.S., *ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions*. Nat Rev Genet, 2012. **13**(12): p. 840-52.
102. Rougemont, J. and F. Naef, *Computational analysis of protein-DNA interactions from ChIP-seq data*. Methods Mol Biol, 2012. **786**: p. 263-73.
103. Gerstein, M.B., et al., *Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project*. Science, 2010. **330**(6012): p. 1775-87.
104. mod, E.C., et al., *Identification of functional elements and regulatory circuits by Drosophila modENCODE*. Science, 2010. **330**(6012): p. 1787-97.
105. Consortium, E.P., et al., *An integrated encyclopedia of DNA elements in the human genome*. Nature, 2012. **489**(7414): p. 57-74.
106. Chadwick, L.H., *The NIH Roadmap Epigenomics Program data resource*. Epigenomics, 2012. **4**(3): p. 317-24.
107. Workman, C.T., et al., *A systems approach to mapping DNA damage response pathways*. Science, 2006. **312**(5776): p. 1054-9.
108. Harbison, C.T., et al., *Transcriptional regulatory code of a eukaryotic genome*. Nature, 2004. **431**(7004): p. 99-104.
109. Venters, B.J., et al., *A comprehensive genomic binding map of gene and chromatin regulatory proteins in Saccharomyces*. Mol Cell, 2011. **41**(4): p. 480-92.
110. Lefrancois, P., et al., *Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing*. BMC Genomics, 2009. **10**: p. 37.
111. Wilbanks, E.G. and M.T. Facciotti, *Evaluation of algorithm performance in ChIP-seq peak detection*. PLoS One, 2010. **5**(7): p. e11471.
112. Winzeler, E.A., et al., *Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis*. Science, 1999. **285**(5429): p. 901-6.
113. Kent, W.J., et al., *The human genome browser at UCSC*. Genome Res, 2002. **12**(6): p. 996-1006.
114. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS)*. Genome Biol, 2008. **9**(9): p. R137.
115. Preti, M., et al., *The telomere-binding protein Tbf1 demarcates snoRNA gene promoters in Saccharomyces cerevisiae*. Mol Cell, 2010. **38**(4): p. 614-20.
116. Shao, Z., et al., *MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets*. Genome Biol, 2012. **13**(3): p. R16.
117. Tippmann, S.C., et al., *Chromatin measurements reveal contributions of synthesis*

and decay to steady-state mRNA levels. Mol Syst Biol, 2012. **8**: p. 593.
118. Xie, D., et al., *Dynamic trans-acting factor colocalization in human cells*. Cell, 2013. **155**(3): p. 713-24.
119. Foley, J.W. and A. Sidow, *Transcription-factor occupancy at HOT regions quantitatively predicts RNA polymerase recruitment in five human cell lines*. BMC Genomics, 2013. **14**: p. 720.
120. Shivaswamy, S. and V.R. Iyer, *Stress-dependent dynamics of global chromatin remodeling in yeast: dual role for SWI/SNF in the heat shock stress response*. Mol Cell Biol, 2008. **28**(7): p. 2221-34.
121. Hardwick, J.S., et al., *Rapamycin-modulated transcription defines the subset of nutrient-sensitive signaling pathways directly controlled by the Tor proteins*. Proc Natl Acad Sci U S A, 1999. **96**(26): p. 14866-70.
122. Li, X. and M. Cai, *Recovery of the yeast cell cycle from heat shock-induced G(1) arrest involves a positive regulation of G(1) cyclin expression by the S phase cyclin Clb5*. J Biol Chem, 1999. **274**(34): p. 24220-31.
123. Horak, C.E., et al., *Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae*. Genes Dev, 2002. **16**(23): p. 3017-33.
124. Iyer, V.R., et al., *Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF*. Nature, 2001. **409**(6819): p. 533-8.
125. Smith, R.L. and A.D. Johnson, *Turning genes off by Ssn6-Tup1: a conserved system of transcriptional repression in eukaryotes*. Trends Biochem Sci, 2000. **25**(7): p. 325-30.
126. Zentner, G.E., T. Tsukiyama, and S. Henikoff, *ISWI and CHD chromatin remodelers bind promoters but act in gene bodies*. PLoS Genet, 2013. **9**(2): p. e1003317.
127. Holstege, F.C., et al., *Dissecting the regulatory circuitry of a eukaryotic genome*. Cell, 1998. **95**(5): p. 717-28.
128. Hahn, J.S., et al., *Genome-wide analysis of the biology of stress responses through heat shock transcription factor*. Mol Cell Biol, 2004. **24**(12): p. 5249-56.
129. Kidder, B.L., G. Hu, and K. Zhao, *ChIP-Seq: technical considerations for obtaining high-quality data*. Nat Immunol, 2011. **12**(10): p. 918-22.
130. Fan, X. and K. Struhl, *Where does mediator bind in vivo?* PLoS One, 2009. **4**(4): p. e5029.
131. Zhu, X., et al., *Genome-wide occupancy profile of mediator and the Srb8-11 module reveals interactions with coding regions*. Mol Cell, 2006. **22**(2): p. 169-78.
132. Giresi, P.G., et al., *FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin*. Genome Res, 2007. **17**(6): p. 877-85.
133. Thurman, R.E., et al., *The accessible chromatin landscape of the human genome*. Nature, 2012. **489**(7414): p. 75-82.
134. Moorman, C., et al., *Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster*. Proc Natl Acad Sci U S A, 2006. **103**(32): p. 12027-

32.
135. Li, X.Y., et al., *Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm*. PLoS Biol, 2008. **6**(2): p. e27.
136. Kornberg, R.D. and Y. Lorch, *Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome*. Cell, 1999. **98**(3): p. 285-94.
137. Yuan, G.C., et al., *Genome-scale identification of nucleosome positions in S. cerevisiae*. Science, 2005. **309**(5734): p. 626-30.
138. Lee, W., et al., *A high-resolution atlas of nucleosome occupancy in yeast*. Nat Genet, 2007. **39**(10): p. 1235-44.
139. Kaplan, N., et al., *The DNA-encoded nucleosome organization of a eukaryotic genome*. Nature, 2009. **458**(7236): p. 362-6.
140. Segal, E., et al., *A genomic code for nucleosome positioning*. Nature, 2006. **442**(7104): p. 772-8.
141. Zhang, Z., et al., *A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome*. Science, 2011. **332**(6032): p. 977-80.
142. Simic, R., et al., *Chromatin remodeling protein Chd1 interacts with transcription elongation factors and localizes to transcribed genes*. EMBO J, 2003. **22**(8): p. 1846-56.
143. Warner, M.H., K.L. Roinick, and K.M. Arndt, *Rtf1 is a multifunctional component of the Paf1 complex that regulates gene expression by directing cotranscriptional histone modification*. Mol Cell Biol, 2007. **27**(17): p. 6103-15.
144. DeGennaro, C.M., et al., *Spt6 regulates intragenic and antisense transcription, nucleosome positioning, and histone modifications genome-wide in fission yeast*. Mol Cell Biol, 2013. **33**(24): p. 4779-92.
145. van Bakel, H., et al., *A compendium of nucleosome and transcript profiles reveals determinants of chromatin architecture and transcription*. PLoS Genet, 2013. **9**(5): p. e1003479.
146. Pray-Grant, M.G., et al., *Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation*. Nature, 2005. **433**(7024): p. 434-8.
147. Smolle, M., et al., *Chromatin remodelers Isw1 and Chd1 maintain chromatin structure during transcription by preventing histone exchange*. Nat Struct Mol Biol, 2012. **19**(9): p. 884-92.
148. Longtine, M.S., et al., *Additional modules for versatile and economical PCR-based gene deletion and modification in Saccharomyces cerevisiae*. Yeast, 1998. **14**(10): p. 953-61.
149. Park, D., et al., *Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements*. Nucleic Acids Res, 2014. **42**(6): p. 3736-49.
150. Rizzo, J.M., J.E. Bard, and M.J. Buck, *Standardized collection of MNase-seq experiments enables unbiased dataset comparisons*. BMC Mol Biol, 2012. **13**: p. 15.
151. Rizzo, J.M., P.A. Mieczkowski, and M.J. Buck, *Tup1 stabilizes promoter nucleosome positioning and occupancy at transcriptionally plastic genes*. Nucleic

Acids Res, 2011. **39**(20): p. 8803-19.
152. Quan, T.K. and G.A. Hartzog, *Histone H3K4 and K36 methylation, Chd1 and Rpd3S oppose the functions of Saccharomyces cerevisiae Spt4-Spt5 in transcription*. Genetics, 2010. **184**(2): p. 321-34.
153. Carrozza, M.J., et al., *Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription*. Cell, 2005. **123**(4): p. 581-92.
154. Joshi, A.A. and K. Struhl, *Eaf3 chromodomain interaction with methylated H3-K36 links histone deacetylation to Pol II elongation*. Mol Cell, 2005. **20**(6): p. 971-8.
155. Keogh, M.C., et al., *Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex*. Cell, 2005. **123**(4): p. 593-605.
156. Schaft, D., et al., *The histone 3 lysine 36 methyltransferase, SET2, is involved in transcriptional elongation*. Nucleic Acids Res, 2003. **31**(10): p. 2475-82.
157. Morillon, A., et al., *Isw1 chromatin remodeling ATPase coordinates transcription elongation and termination by RNA polymerase II*. Cell, 2003. **115**(4): p. 425-35.
158. Seila, A.C., et al., *Divergent transcription from active promoters*. Science, 2008. **322**(5909): p. 1849-51.
159. Chen, R.A., et al., *The landscape of RNA polymerase II transcription initiation in C. elegans reveals promoter and enhancer architectures*. Genome Res, 2013. **23**(8): p. 1339-47.
160. Hnisz, D., et al., *Super-enhancers in the control of cell identity and disease*. Cell, 2013. **155**(4): p. 934-47.
161. Whyte, W.A., et al., *Master transcription factors and mediator establish super-enhancers at key cell identity genes*. Cell, 2013. **153**(2): p. 307-19.
162. Hughes, A.L., et al., *A functional evolutionary approach to identify determinants of nucleosome positioning: a unifying model for establishing the genome-wide pattern*. Mol Cell, 2012. **48**(1): p. 5-15.

Vita

Daechan Park attended Yangjung High School in Busan, Korea. He earned a Bachelor of Science in Biological Sciences from Seoul National University, Seoul, Korea, in 2008. In fall of 2008, he started graduate school at the University of Texas at Austin as a Ph.D. student in the Institute for Cell and Molecular Biology.

During graduate school he wrote 4 manuscripts as the first or co-first author. Two have been peer-reviewed and published, and the other two works are currently being prepared.

Park D., Morris A.R., Battenhouse A. & Iyer V.R. Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements (2014) *Nucleic Acids Res*. 42(6): p. 3736-49.

Park D., Lee Y., Bhupindersingh G. & Iyer V.R. Widespread Misinterpretable ChIP-seq Bias in Yeast (2013) *PLoS One* 8(12): e83506

Email: daechan.park@utexas.edu

This dissertation was typed by Daechan Park.