BOSTON UNIVERSITY

COLLEGE OF ENGINEERING

Dissertation

LOREM IPSUM

by

JAMES CHUANG

B.S., Johns Hopkins University, 2013
M.S., Boston University, 2018

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2019

Approved by

First Reader

Fred Winston, PhD
Professor of Genetics
Harvard Medical School

Second Reader

Ahmad Khalil, PhD
Assistant Professor of Biomedical Engineering

Third Reader

L. Stirling Churchman, PhD
Assistant Professor of Genetics
Harvard Medical School

Fourth Reader

John T. Ngo, PhD
Assistant Professor of Biomedical Engineering

Fifth Reader

Wilson Wong, PhD
Assistant Professor of Biomedical Engineering

Acknowledgments

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

James Chuang

LOREM IPSUM

JAMES CHUANG

Boston University, College of Engineering, 2019

Major Professors: Fred Winston, PhD
Professor of Genetics
Harvard Medical School

Ahmad Khalil, PhD
Assistant Professor of Biomedical Engineering

ABSTRACT

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 A brief introduction to transcription

In eukaryotic cells, transcription of protein-coding genes is carried out by the protein complex RNA polymerase II (Pol II), and broadly occurs in three sequential stages of transcription initiation, elongation, and termination (Shandilya and Roberts, 2012). During each of these stages, the Pol II complex is associated with distinct sets of factors which modulate the activity of Pol II and carry out co-transcriptional processes such as RNA capping, RNA splicing, histone modification, RNA cleavage, and RNA polyadenylation. Given how fundamental transcription is to gene expression, it is unsurprising that every stage of transcription is highly regulated.

To get a rough idea of just how tightly transcription is regulated, it is useful to consider a back-of-the-envelope calculation of the specificity of transcription initiation in the human genome. That is, what proportion of the human genome at which transcription could initiate does transcription initiation actually occur?

The number of positions at which transcription could theoretically initiate is simply the size of the genome: The human genome is approximately three billion base pairs in length (BNID 111378, Weber et al. (2009)), and since each base pair can be transcribed from each of its two strands, there are 6×10^9 available positions.

The number of positions at which transcription *does* initiate can be estimated from the number of genes transcribed by Pol II and the number of positions that Pol II initiates from for each gene. At last count, the human genome contains about twenty thousand protein-coding genes (Consortium et al., 2012). To be conservative in our estimate with regards to specificity, we will assume that all twenty thousand genes are expressed. We also know that protein-coding genes are only a subset of the genes transcribed by Pol II: Pol II also transcribes multiple classes of non-coding genes, including enhancers and long non-coding RNAs (Kaikkonen and Adelman, 2018). Compared to protein-coding genes, the number of non-coding genes is less certain. If we assume that there are five non-coding genes for each coding gene, this brings our estimate of the number of genes transcribed by Pol II to 1.2×10^5 genes.

As you will see from yeast transcription start site data in later chapters, transcription initiation for a single gene generally occurs at multiple nucleotides, generating multiple major transcript isoforms per gene. Assuming that there are, on average, five major transcription start sites (TSSs) per gene, the proportion of the human genome at which transcription initiation occurs is

$$\frac{\left(1.2 \times 10^5 \text{ genes}\right)\left(5 \frac{\text{TSSs}}{\text{gene}}\right)}{\left(6 \times 10^9 \text{ possible TSSs}\right)} = 1 \times 10^{-4}.$$

Our rough estimate says that, when presented with ten thousand positions to choose from, RNA polymerase starts transcription from only one![1]

Many factors are known to contribute to this remarkable specificity. Most notably, transcription initiation requires the presence of specific DNA sequence motifs, which

[1] A similar conclusion is reached by examining ENCODE CAGE-seq data: At the time of writing, ENCODE reports roughly 150,000 TSS peaks across 30 cell types/cell lines. Assuming the signal is concentrated at 5 nucleotides per peak, then $\frac{\left(1.5 \times 10^5 \text{peaks}\right)\left(5 \frac{\text{nt}}{\text{peak}}\right)}{6 \times 10^9 \text{nt}} = \frac{1}{8000}$.

increase the probability of Pol II binding to DNA together with necessary initiation factors (Haberle and Stark, 2018). That factors known to associate with Pol II during transcription initiation control transcription initiation is unsurprising. A less obvious fact is that some transcription *elongation* factors, including histone chaperones and histone modification enzymes, also play a role in determining where transcription initiation is allowed to occur (Cheung et al., 2008; Hennig and Fischer, 2013; Kaplan et al., 2003). Evidence suggests that these elongation factors are likely required to maintain normal chromatin structure over transcribed regions, and that the disruption of normal chromatin structure allows Pol II to initiate transcription in regions which are normally inacessible. Chapters 2 and 3 of this dissertation describe our studies of **Spt6** and **Spt5**, two of the transcription elongation factors involved in this process. One phenotype observed when these factors are disrupted is **intragenic transcription**, transcription appearing to arise from within protein-coding sequences. In chapter 4, I describe our efforts to understand how intragenic transcription might play a role in the cellular response to various stress conditions. The remainder of this introduction provides a brief overview of the considerations taken into account in order to make the data analyses behind this dissertation as transparent and reproducible as possible.

1.2 Reproducible data analysis for genomics

My role in the projects in this dissertation is a mix of **data scientist** and **data engineer**: I build pipelines for processing (usually genomic) datasets, taking raw data through processing, statistical analysis, and data visualization. This mostly entails surveying available tools, selecting the tools most suitable for the task, and coding solutions to problems when existing tools are insufficient.

The analysis of complex datasets like those in genomics presents challenges to achieving transparency and reproducibility when reporting methods and results. In building the data analysis pipelines behind the results of this dissertation, I have tried to meet these challenges by following best practices that would be standards for publication in an ideal world. All of my data analyses are open source (github.com/winston-lab), and are designed to be reproducible by others: For all publications, a self-contained archive is uploaded which includes everything needed to go from raw data to the figures and results of the publication (e.g., https://doi.org/10.5281/zenodo.1409826). This level of accessibility is greatly facilitated by building data analyses using Snakemake (Köster and Rahmann, 2012), one of several available frameworks for workflow management (Di Tommaso et al., 2017; Voss et al., 2017). Snakemake's scalable execution and its ability to specify dependencies in virtual environments allow workflows to truly be reproducible: data analyses can be re-run on personal computers, computing clusters, or cloud environments, and the exact versions of the software used when initially running the data analysis will automagically be deployed.

Open sharing of data and code like this is essential to the scientific process. When analysis pipelines routinely consist of tens of steps with tens of parameters each, seeing the data and code is the only way for those interested to know exactly how the data were handled. Altogether, this allows for more informed evaluation of results from the literature, as well as the possibility of finding and correcting errors in analysis.

1.3 Bibliography

Cheung, V., Chua, G., Batada, N. N., Landry, C. R., Michnick, S. W., Hughes, T. R., and Winston, F. (2008). Chromatin- and transcription-related factors repress transcription from within coding regions throughout the saccharomyces cerevisiae genome. *PLOS Biology*, 6(11):1–13. 1.1

Consortium, T. E. P., Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B.-K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J. M., Song, L., Trinklein, N. D., Altshuler, R. C., Birney, E., Brown, J. B., Cheng, C., Djebali, S., Dong, X., Ernst, J., Furey, T. S., Gerstein, M., Giardine, B., Greven, M., Hardison, R. C., Harris, R. S., Herrero, J., Hoffman, M. M., Iyer, S., Kellis, M., Kheradpour, P., Lassmann, T., Li, Q., Lin, X., Marinov, G. K., Merkel, A., Mortazavi, A., Parker, S. C. J., Reddy, T. E., Rozowsky, J., Schlesinger, F., Thurman, R. E., Wang, J., Ward, L. D., Whitfield, T. W., Wilder, S. P., Wu, W., Xi, H. S., Yip, K. Y., Zhuang, J., Bernstein, B. E., Green, E. D., Gunter, C., Snyder, M., Pazin, M. J., Lowdon, R. F., Dillon, L. A. L., Adams, L. B., Kelly, C. J., Zhang, J., Wexler, J. R., Good, P. J., Feingold, E. A., Crawford, G. E., Dekker, J., Elnitski, L., Farnham, P. J., Giddings, M. C., Gingeras, T. R., Guigó, R., Hubbard, T. J., Kent, W. J., Lieb, J. D., Margulies, E. H., Myers, R. M., Stamatoyannopoulos, J. A., Tenenbaum, S. A., Weng, Z., White, K. P., Wold, B., Yu, Y., Wrobel, J., Risk, B. A., Gunawardena, H. P., Kuiper, H. C., Maier, C. W., Xie, L., Chen, X., Mikkelsen, T. S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M. J., Durham, T., Ku, M., Truong, T., Eaton, M. L., Dobin, A., Tanzer, A., Lagarde, J., Lin, W., Xue, C., Williams, B. A., Zaleski, C., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O. J., Park, E., Preall, J. B., Presaud, K., Ribeca, P., Robyr, D., Ruan, X., Sammeth, M., Sandhu, K. S., Schaeffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Hayashizaki, Y., Reymond, A., Antonarakis, S. E., Hannon, G. J., Ruan, Y., Carninci, P., Sloan, C. A., Learned, K., Malladi, V. S., Wong, M. C., Barber, G. P., Cline, M. S., Dreszer, T. R., Heitner, S. G., Karolchik, D., Kirkup, V. M., Meyer, L. R., Long, J. C., Maddren, M., Raney, B. J., Grasfeder, L. L., Giresi, P. G., Battenhouse, A., Sheffield, N. C., Showers, K. A., London, D., Bhinge, A. A., Shestak, C., Schaner, M. R., Ki Kim, S., Zhang, Z. Z., Mieczkowski, P. A., Mieczkowska, J. O., Liu, Z., McDaniell, R. M., Ni, Y., Rashid, N. U., Kim, M. J., Adar, S., Zhang, Z., Wang, T., Winter, D., Keefe,

D., Iyer, V. R., Zheng, M., Wang, P., Gertz, J., Vielmetter, J., Partridge, E., Varley, K. E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K. M., Anaya, M., Cross, M. K., Muratet, M. A., Newberry, K. M., McCue, K., Nesmith, A. S., Fisher-Aylor, K. I., Pusey, B., DeSalvo, G., Parker, S. L., Balasubramanian, S., Davis, N. S., Meadows, S. K., Eggleston, T., Newberry, J. S., Levy, S. E., Absher, D. M., Wong, W. H., Blow, M. J., Visel, A., Pennachio, L. A., Petrykowska, H. M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Davidson, C., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Gonzalez, J. M., Griffiths, E., Harte, R., Hendrix, D. A., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Leng, J., Lin, M. F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Rodriguez, J. M., Saunders, G., Sboner, A., Searle, S., Sisu, C., Snow, C., Steward, C., Tapanari, E., Tress, M. L., van Baren, M. J., Washietl, S., Wilming, L., Zadissa, A., Zhang, Z., Brent, M., Haussler, D., Valencia, A., Addleman, N., Alexander, R. P., Auerbach, R. K., Balasubramanian, S., Bettinger, K., Bhardwaj, N., Boyle, A. P., Cao, A. R., Cayting, P., Charos, A., Cheng, Y., Eastman, C., Euskirchen, G., Fleming, J. D., Grubert, F., Habegger, L., Hariharan, M., Harmanci, A., Iyengar, S., Jin, V. X., Karczewski, K. J., Kasowski, M., Lacroute, P., Lam, H., Lamarre-Vincent, N., Lian, J., Lindahl-Allen, M., Min, R., Miotto, B., Monahan, H., Moqtaderi, Z., Mu, X. J., O'Geen, H., Ouyang, Z., Patacsil, D., Raha, D., Ramirez, L., Reed, B., Shi, M., Slifer, T., Witt, H., Wu, L., Xu, X., Yan, K.-K., Yang, X., Struhl, K., Weissman, S. M., Penalva, L. O., Karmakar, S., Bhanvadia, R. R., Choudhury, A., Domanus, M., Ma, L., Moran, J., Victorsen, A., Auer, T., Centanin, L., Eichenlaub, M., Gruhl, F., Heermann, S., Hoeckendorf, B., Inoue, D., Kellner, T., Kirchmaier, S., Mueller, C., Reinhardt, R., Schertel, L., Schneider, S., Sinn, R., Wittbrodt, B., Wittbrodt, J., Partridge, E. C., Jain, G., Balasundaram, G., Bates, D. L., Byron, R., Canfield, T. K., Diegel, M. J., Dunn, D., Ebersol, A. K., Frum, T., Garg, K., Gist, E., Hansen, R. S., Boatman, L., Haugen, E., Humbert, R., Johnson, A. K., Johnson, E. M., Kutyavin, T. V., Lee, K., Lotakis, D., Maurano, M. T., Neph, S. J., Neri, F. V., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Rynes, E., Sanchez, M. E., Sandstrom, R. S., Shafer, A. O., Stergachis, A. B., Thomas, S., Vernot, B., Vierstra, J., Vong, S., Wang, H., Weaver, M. A., Yan, Y., Zhang, M., Akey, J. M., Bender, M., Dorschner, M. O., Groudine, M., MacCoss, M. J., Navas, P., Stamatoyannopoulos, G., Beal, K., Brazma, A., Flicek, P., Johnson, N., Lukk, M., Luscombe, N. M., Sobral, D., Vaquerizas, J. M., Batzoglou, S., Sidow, A., Hussami, N., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M. W., Schaub, M. A., Miller, W., Bickel, P. J., Banfai, B., Boley, N. P., Huang, H., Li, J. J., Noble, W. S., Bilmes, J. A., Buske, O. J., Sahu, A. D., Kharchenko, P. V., Park, P. J., Baker, D., Taylor, J., and Lochovsky, L. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489:57 EP –. Article. 1.1

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319. 1.2

Haberle, V. and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*, 19(10):621–637. 1.1

Hennig, B. P. and Fischer, T. (2013). The great repression: chromatin and cryptic transcription. *Transcription*, 4(3):97—101. 1.1

Kaikkonen, M. U. and Adelman, K. (2018). Emerging roles of non-coding rna transcription. *Trends in Biochemical Sciences*, 43(9):654–667. 1.1

Kaplan, C. D., Laprade, L., and Winston, F. (2003). Transcription elongation factors repress transcription initiation from cryptic sites. *Science*, 301(5636):1096–1099. 1.1

Köster, J. and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522. 1.2

Shandilya, J. and Roberts, S. G. (2012). The transcription cycle in eukaryotes: From productive initiation to rna polymerase ii recycling. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1819(5):391 – 400. 1.1

Voss, K., Gentry, J., and Van der Auwera, G. (2017). Full-stack genomics pipelining with gatk4 + wdl + cromwell. In *18th Annual Bioinformatics Open Source Conference (BOSC 2017)*. 1.2

Weber, G., Springer, M., Jorgensen, P., Milo, R., and Moran, U. (2009). Bionumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Research*, 38(suppl_1):D750–D753. 1.1

Chapter 2

Genomics of transcription elongation factor Spt6

2.1 Collaborators

Steve Doris	optimized TSS-seq and ChIP-nexus protocols
	generated TSS-seq and ChIP-nexus libraries
Olga Viktorovskaya	generated MNase-seq libraries
Magdalena Murawska	generated NET-seq libraries
Dan Spatt	Northern, Western, and ChIP experiments

2.2 Introduction to Spt6 and intragenic transcription

The conserved transcription elongation factor Spt6 interacts directly with RNA polymerase II (Close et al., 2011; Diebold et al., 2010b; Liu et al., 2011; Sdano et al., 2017; Sun et al., 2010; Yoh et al., 2007), histones (Bortvin and Winston, 1996; McCullough et al., 2015), and another elongation factor called Spn1/Iws1 (Diebold et al., 2010a; Li et al., 2018; McDonald et al., 2010). The classification of Spt6 as a transcription elongation factor is based on its association with elongating Pol II (Andrulis et al., 2000; Ivanovska et al., 2011; Kaplan et al., 2000; Krogan et al., 2002; Mayer et al., 2010), and its ability to enhance elongation both *in vitro* (Endoh et al., 2004) and *in vivo* (Ardehali et al., 2009), though Spt6 has also been shown to regulate initiation in

a small number of cases (Adkins and Tyler, 2006; Ivanovska et al., 2011). Evidence suggests that as Spt6 travels with elongating Pol II, it acts as a histone chaperone, reassembling nucleosomes after their displacement from DNA due to transcription (Duina, 2011; Ivanovska et al., 2011). Consistent with its histone chaperone function, Spt6 influences chromatin structure (Bortvin and Winston, 1996; DeGennaro et al., 2013; Ivanovska et al., 2011; Jeronimo et al., 2015; Kaplan et al., 2003; Perales et al., 2013; van Bakel et al., 2013); Spt6 is also required for some histone modifications, including H3K36 methylation (Carrozza et al., 2005; Chu et al., 2006; Yoh et al., 2008; Youdell et al., 2008), and, in some organisms, H3K4 and H3K27 methylation (Begum et al., 2012; Chen et al., 2012; DeGennaro et al., 2013; Wang et al., 2017, 2013).

Studies in the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* have previously examined the requirement for Spt6 in normal transcription (Cheung et al., 2008; DeGennaro et al., 2013; Kaplan et al., 2003; Pathak et al., 2018; Uwimana et al., 2017; van Bakel et al., 2013). As Spt6 is essential for viability in *S. cerevisiae*, many of these studies use the same temperature-sensitive *spt6* mutant used in this project, ***spt6-1004***, which encodes an in-frame deletion of a helix-hairpin-helix domain within Spt6 (Kaplan et al., 2003). When *spt6-1004* cells are shifted from 30°C to 37°C for 80 minutes, bulk Spt6 protein levels are depleted to about 20% of wild-type levels, though cells are still viable (Figure 2.1, (Ka-

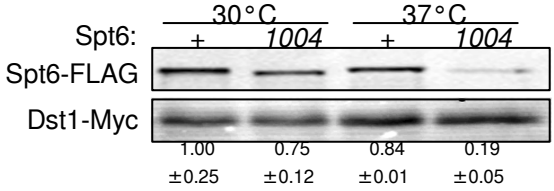

Figure 2.1: Western blot for Spt6 in wild-type and *spt6-1004* cells, at 30°C and after 80 minutes at 37°C. Spt6 and Dst1 from a spike-in were detected using α-FLAG and α-Myc antibodies, respectively. The mean $\pm$ standard deviation of three blots are shown below each lane.

plan et al., 2003)). A notable phenotype of the *spt6-1004* mutant is the appearance of **intragenic transcripts**, transcripts which appear to arise from within protein-coding sequences in both sense and antisense orientations relative to the coding gene (Figure 2.2) (Cheung et al., 2008; DeGennaro et al., 2013; Kaplan et al., 2003; Uwimana et al., 2017).

Previous genome-wide measurements of transcript levels in *spt6-1004* relied on tiled microarrays (Cheung et al., 2008) and RNA sequencing (DeGennaro et al., 2013; Uwimana et al., 2017). Studying intragenic transcription is difficult with these methods, since the signal for an intragenic transcript in the same orientation as the gene it overlaps is convoluted with the signal from the full-length 'genic' transcript (Figures 2.2, 2.3) (Cheung et al., 2008; Lickwar et al., 2009). Identification of intragenic transcription has thus relied on identifying cases where the signal towards the 3′ end of a transcript is greater than the signal towards the 5′ end. However, this leads to both false positives, due to the inherent variability of the signal over a transcript, as well as false negatives, due to the requirement of the intragenic transcript to be well-expressed relative to its corresponding genic transcript in order to be identified. Additionally, these methods are assays of steady-state RNA levels, which makes them unable to distinguish whether the intragenic transcripts observed in *spt6-1004* result from: A) new intragenic transcription initiation in the mutant, B) reduced decay of intragenic transcripts which are rapidly degraded in wild-type, or C) processing of full-length protein-coding RNAs. New transcription initiation has been shown

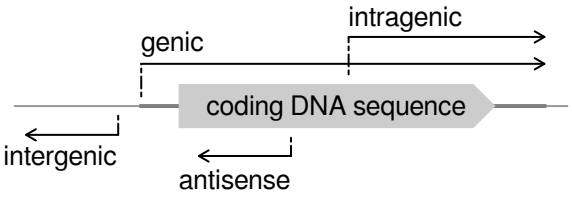

Figure 2.2: Diagram of transcript orientation with respect to coding DNA sequences, for the categories of transcripts referred to in this document.

to be responsible for individual cases of intragenic initiation (Kaplan et al., 2003), but this has not previously been studied on a genome-wide scale.

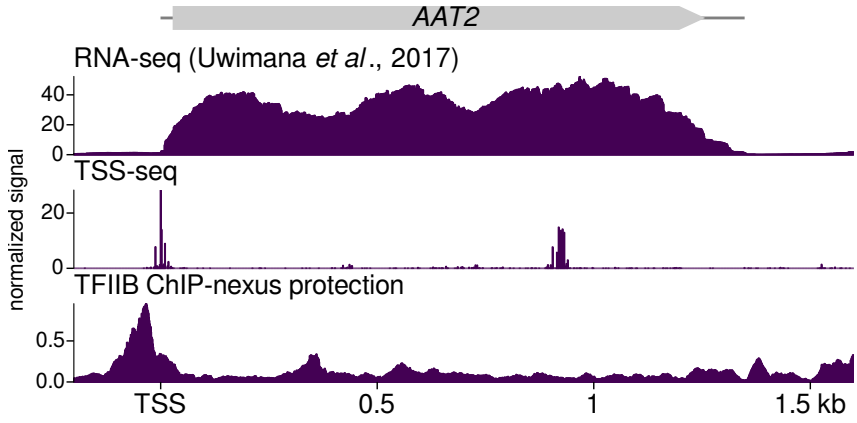

Figure 2.3: Sense strand RNA-seq signal, sense strand TSS-seq signal, and TFIIB ChIP-nexus protection at the *AAT2* gene, in *spt6-1004* after 80 minutes at 37°C.

To address these challenges to studying intragenic transcription, we applied two genomic assays to *spt6-1004*: transcription start-site sequencing (**TSS-seq**), and **ChIP-nexus of TFIIB**, a component of the RNA polymerase II pre-initiation complex (PIC). TSS-seq sequences the 5′ end of capped and polyadenylated RNAs (Arribere and Gilbert, 2013; Malabat et al., 2015), allowing separation of intragenic from genic RNA signals and identification of intragenic transcript starts with single-nucleotide resolution (Figure 2.3). ChIP-nexus is a high-resolution chromatin immunoprecipitation technique, in which the immunoprecipitated DNA is exonuclease digested up to the bases crosslinked with the protein of interest before sequencing (He et al., 2015). When applied to the PIC component TFIIB, ChIP-nexus reports where transcription initiation is occurring, thus allowing us to determine if intragenic transcripts in *spt6-1004* result from new transcription initiation.

2.3 TSS-seq and TFIIB ChIP-nexus results for *spt6-1004*

To study the relationship between Spt6 and transcription, TSS-seq and TFIIB ChIP-nexus libraries were prepared from wild-type and *spt6-1004* cells, after cultures were shifted from 30°C to 37°C for 80 minutes. In wild-type cells, TSS-seq and TFIIB ChIP-nexus recapitulate their expected distributions over the genome: Most TSS signal is restricted to annotated genic TSSs, while most TFIIB signal is localized just upstream of the TSS (Figures 2.4, 2.5). In *spt6-1004*, the signal for both assays infiltrates gene bodies, reflecting widespread intragenic expression of capped and polyadenylated transcripts, and suggesting that new transcription initiation contributes to the intragenic transcription phenotype. Notably, sense strand TSS-seq signal in *spt6-1004* tends to occur towards the 3′ end of genes, while antisense strand TSS-seq signal tends to occur towards the 5′ end of genes.

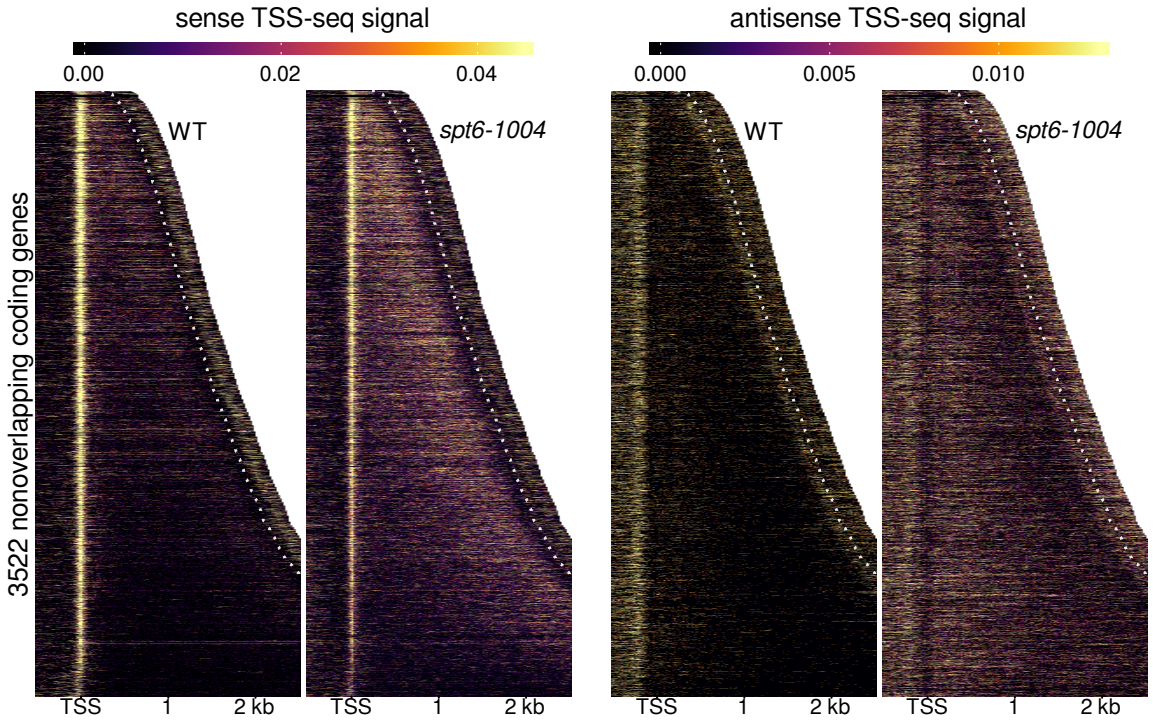

Figure 2.4: Heatmaps of sense and antisense TSS-seq signal from wild-type and *spt6-1004* cells, over 3522 non-overlapping coding genes aligned by wild-type genic TSS and sorted by annotated transcript length. Data are shown for each gene up to 300 nucleotides 3′ of the cleavage and polyadenylation site (CPS), indicated by the white dotted line. Values are the mean of spike-in normalized coverage in non-overlapping 20 nucleotide bins, averaged over two replicates. Values above the 92nd percentile are set to the 92nd percentile for visualization.

The TSS-seq data were quantified by peak calling and differential expression analysis, and classified into genomic categories based on their position relative to coding genes. As suggested by the heatmap visualization (Figure 2.4), we detect significant induction of over 4000 intragenic and antisense TSSs in *spt6-1004* (Figure 2.6). Compared to previous studies identifying *spt6-1004* intragenic transcription by tiled microarray and RNA-seq (Cheung et al., 2008; Uwimana et al., 2017), we identify intragenic transcription at over 1000 additional genes (Figure 2.7), with the

additional information of exact start sites for all identified TSSs.

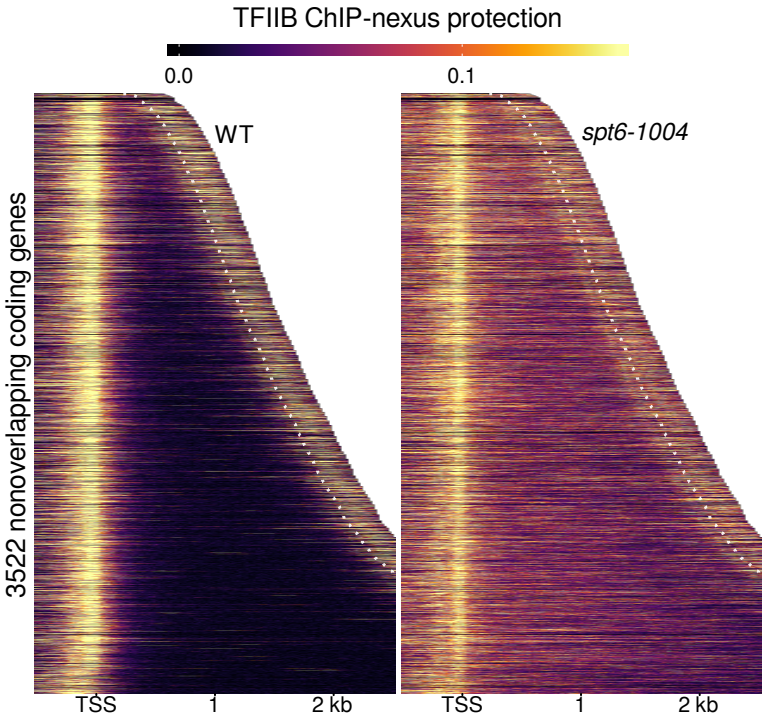

Figure 2.5: Heatmaps of TFIIB binding measured by ChIP-nexus, over the same regions shown in Figure 2.4. Values are the mean of library-size normalized coverage in non-overlapping 20 bp bins, averaged over two replicates. Values above the 85th percentile are set to the 85th percentile for visualization.

The TSS-seq data also revealed an unexpected downregulation of most genic TSSs: In this experiment, we detected a significant downregulation to levels below 67% of wild-type levels at 75% (3579/4792) of genic TSSs (Figure 2.6). As a result of intragenic/antisense induction and genic repression, expression levels in *spt6-1004* of all classes of transcripts become similar to one another (Figure 2.8).

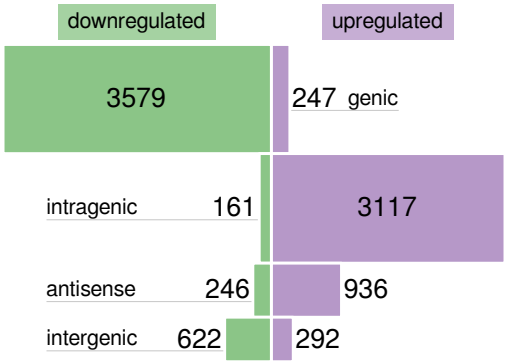

Figure 2.6: Bar plot of the number of TSS-seq peaks differentially expressed in *spt6-1004* versus wild-type, both after 80 minutes at 37°C. The height of each bar is proportional to the total number of peaks in the category, including those not found to be significantly differentially expressed.

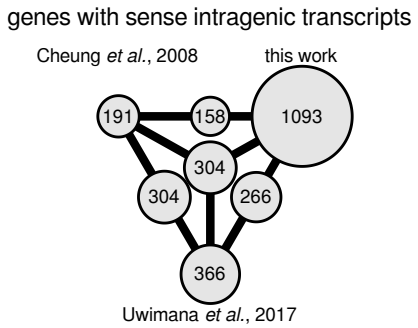

Figure 2.7: Set diagram of the number of genes reported to have *spt6-1004*-induced intragenic transcripts using tiled arrays (Cheung et al., 2008), RNA-seq (Uwimana et al., 2017), and TSS-seq (this work).

The changes in transcript levels in *spt6-1004* observed by TSS-seq correspond with substantial differences in the pattern of TFIIB binding on the genome. While TFIIB in wild-type binds in discrete peaks within promoter regions, TFIIB in *spt6-1004* binds much more promiscuously, with many loci having TFIIB signal spread over broad regions of the genome (Figure 2.9). This difference in binding pattern makes peak calling ineffective for quantifying TFIIB signal in this

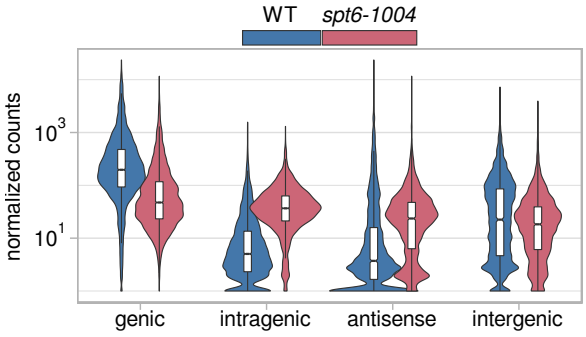

Figure 2.8: Violin plots of expression level distributions for genomic classes of TSS-seq peaks in wild-type and *spt6-1004*, both after 80 minutes at 37°C. Normalized counts are the mean of spike-in size factor normalized counts from two replicates.

case: ChIP-seq peak callers generally use different algorithms for calling 'narrow' peaks (e.g. for sequence-specific transcription factors) and 'broad' peaks (e.g. for histone modifications), meaning that a single algorithm is unable to call a unified set of peaks that is meaningful for differential binding analyses between wild-type and *spt6-1004*. Therefore, to see if changes in transcript levels in *spt6-1004* correspond to changes in transcription initiation, we compared the change in TSS-seq signal at TSS-seq peaks in *spt6-1004* to the change in TFIIB ChIP-nexus signal in the window extending 200 bp upstream of the TSS-seq peak. Changes in TSS-seq signal in *spt6-1004* are associated with a change in TFIIB signal of the same sign at over 82% of TSSs of any genomic class (Figure 2.10), indicating that the increase in intragenic transcript levels and decrease in genic transcript levels observed in *spt6-1004* are in large part explained by changes in transcription initiation.

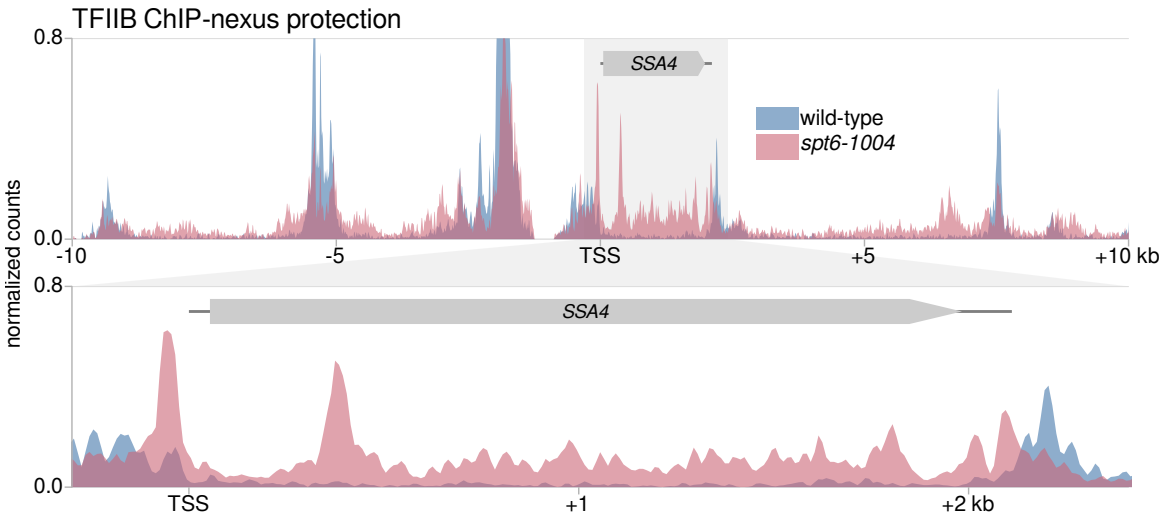

Figure 2.9:

top) TFIIB ChIP-nexus protection in wild-type and *spt6-1004*, over 20 kb of chromosome II flanking the *SSA4* gene.

bottom) Expanded view of TFIIB protection over the *SSA4* gene.

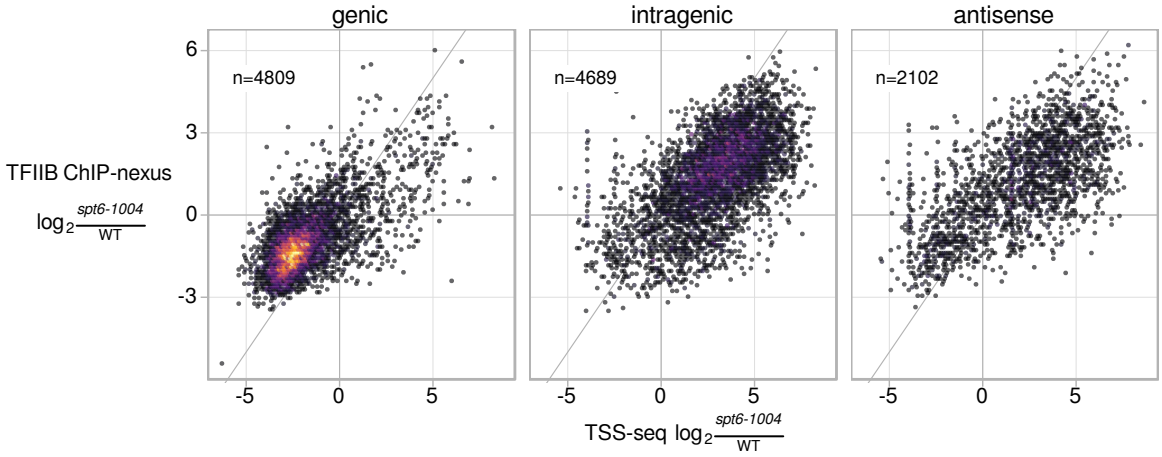

Figure 2.10: Scatterplots of fold-change in *spt6-1004* over wild-type, comparing TSS-seq and TFIIB ChIP-nexus. Each dot represents a TSS-seq peak paired with the window extending 200 bp upstream of the TSS-seq peak summit for quantification of TFIIB ChIP-nexus signal. Fold-changes are regularized fold-change estimates from DESeq2, with size factors determined from the *S. pombe* spike-in (TSS-seq), or *S. cerevisiae* counts (ChIP-nexus).

2.4 MNase-seq results from *spt6-1004*

Because a primary function of Spt6 is to act as histone chaperone that reassembles nucleosomes in the wake of transcription (Duina, 2011), it is reasonable to expect that the transcriptional changes seen in *spt6-1004* would be associated with changes in chromatin structure. The requirement for Spt6 in maintaining normal chromatin structure has been demonstrated in previous studies (Bortvin and Winston, 1996; DeGennaro et al., 2013; Ivanovska et al., 2011; Jeronimo et al., 2015; Kaplan et al., 2003; Perales et al., 2013; van Bakel et al., 2013). To re-examine this requirement in higher resolution, we assayed nucleosome protection genome-wide using micrococcal nuclease digestion of chromatin followed by sequencing (MNase-seq).

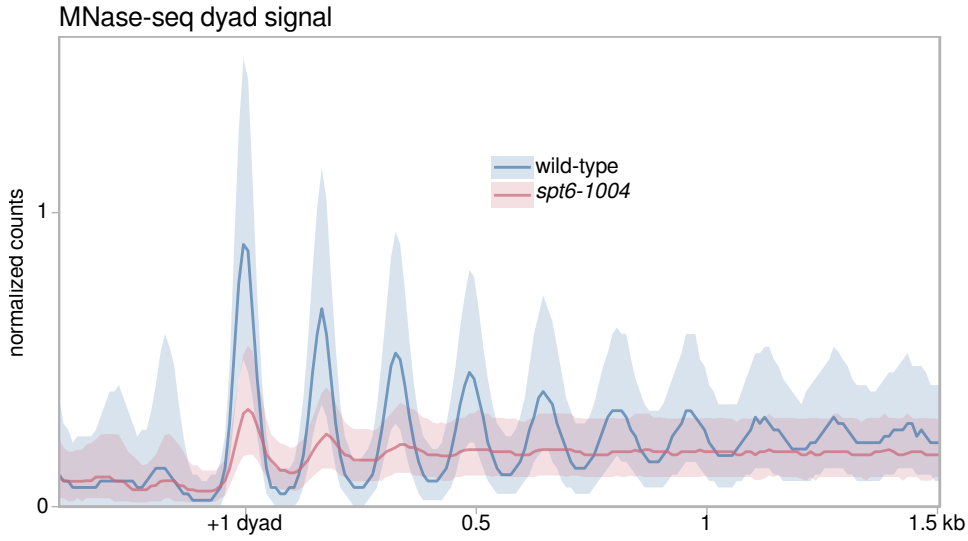

Figure 2.11: Average MNase-seq dyad signal in wild-type and *spt6-1004*, over 3522 non-overlapping coding genes aligned by wild-type +1 nucleosome dyad. The solid line and shading are the median and inter-quartile range of the mean spike-in normalized coverage over two replicates (*spt6-1004*) or one experiment (wild-type), in non-overlapping 20 bp bins.

In wild-type, the MNase-seq data recapitulate the expected signature over genes, with a nucleosome-depleted region upstream of a strongly positioned '+1' nucleosome, and a regularly phased array of nucleosomes over the gene body (Figure 2.11). In *spt6-1004*, nucleosome signal is severely reduced at canonical nucleosome positions and spreads into inter-nucleosome regions. Changes in aggregate nucleosome signal such as those observed in Figure 2.11 are the combination of changes to nucleosome occupancy (the number of reads assigned to a nucleosome), fuzziness (the standard deviation of read positions for a nucleosome), and position (the coordinate with the maximum reads for a nucleosome) (Chen et al., 2013). Using DANPOS2 (Chen et al., 2013), we called nucleosome positions and quantified these metrics for wild-type and *spt6-1004*. Wild-type nucleosomes span a relatively wide range of occupancy and fuzziness space, with highly occupied nucleosomes tending to be less fuzzy (i.e., more well-positioned) (Figure 2.12). In *spt6-1004*, the population of nucleosomes is much more homogeneous: nucleosome occupancy is decreased globally, and nucleosome fuzziness is restricted to the high end of the wild-type distribution.

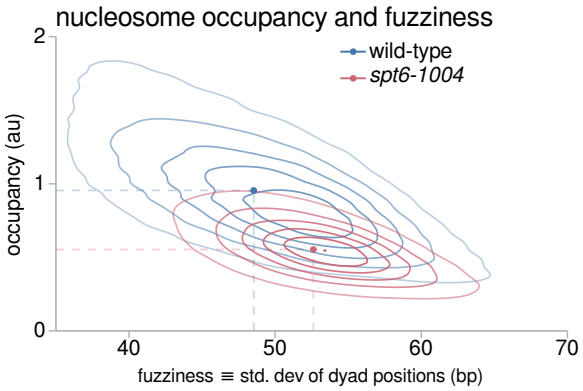

Figure 2.12: Contour plot of the distribution of nucleosome occupancy and fuzziness in wild-type and *spt6-1004*. Dashed lines indicate median values.

Previous studies observed two trends: 1) In wild-type cells, nucleosome positioning is weaker over highly transcribed genes than over moderately transcribed genes (Shivaswamy et al., 2008), and 2) In *spt6-1004* cells, the decrease in nucleosome occupancy is greater for highly transcribed genes (Ivanovska et al., 2011). To re-

examine these trends, we looked at the MNase-seq data in the context of NET-seq data, which reports the position of actively transcribing RNAPII and reflects a gene's level of transcription (Figure 2.13) (Churchman and Weissman, 2012). The data support the first trend: in wild-type, genes with the strongest NET-seq signal have weak patterning of MNase-seq signal. However, we find no obvious relationship between transcription level and the nucleosome occupancy changes observed in *spt6-1004* (Figure 2.13): Genes with the greatest transcription do tend to have lower MNase-seq signal in *spt6-1004*, but this is expected since these genes also have lower MNase-seq signal in wild-type. The discrepancy with prior work might be explained by the greater resolution and breadth of MNase-seq versus MNase and microarray of chromosome III (Ivanovska et al., 2011).

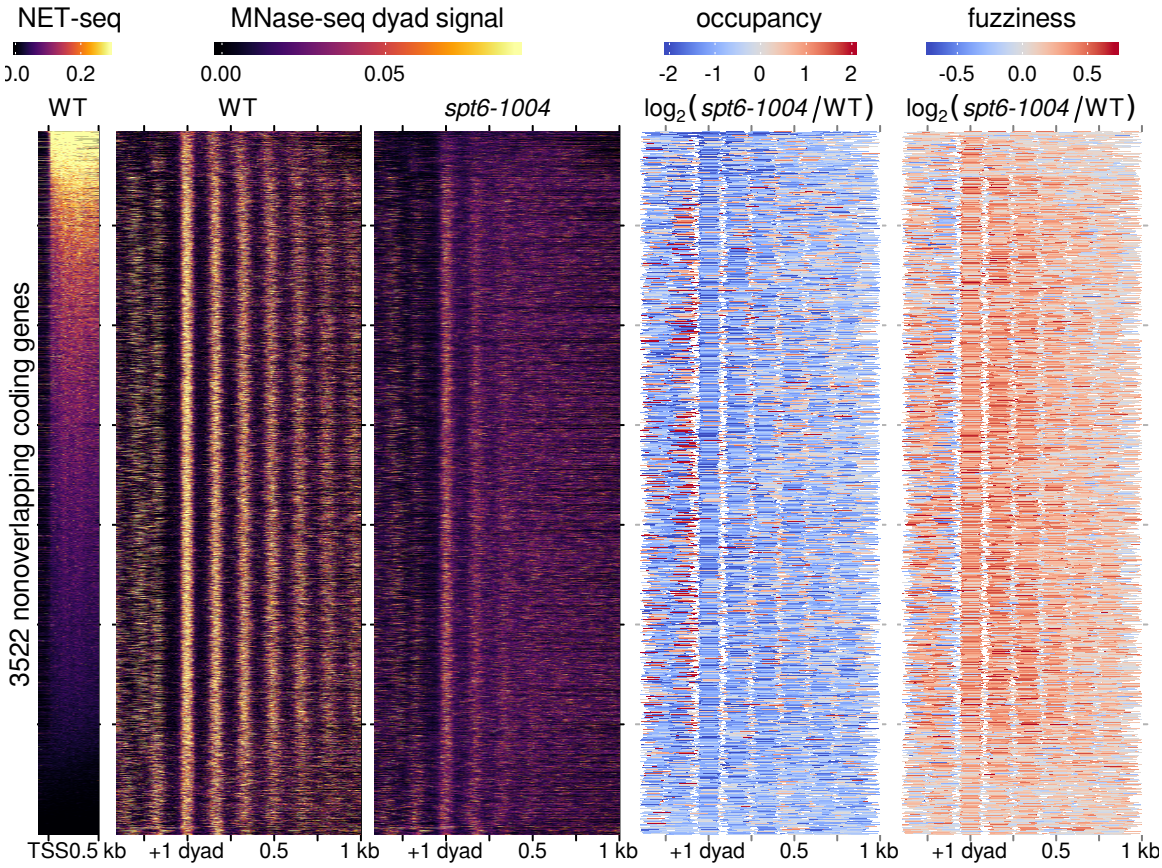

Figure 2.13:

left) Heatmap of sense strand NET-seq signal for 3522 non-overlapping genes, aligned by genic TSS and sorted by total sense strand NET-seq signal in the window extending 500 nt downstream from the genic TSS. Values are the mean of library-size normalized coverage in non-overlapping 20 nt bins, averaged over two replicates.

middle) Heatmaps of MNase-seq dyad signal in wild-type and *spt6-1004* for the same genes, aligned by wild-type +1 nucleosome dyad and arranged by sense NET-seq signal as in the leftmost panel. Values are the mean of spike-in normalized coverage in non-overlapping 20 bp bins, averaged over two replicates (*spt6-1004*) or one experiment (wild-type).

right) Heatmaps of fold-change in nucleosome occupancy and fuzziness for the same genes, aligned by wild-type +1 nucleosome dyad and arranged by sense NET-seq signal as in the leftmost panel.

2.4.1 Clustering of MNase-seq profiles at *spt6-1004*-induced intragenic TSSs

The aggregate MNase-seq dyad signal around all *spt6-1004* intragenic TSSs is aperiodic (Figure 2.15, top left panel), which occurs as a result of destructive interference from offset nucleosome phasing patterns. To discover these phasing patterns, we used the wild-type and *spt6-1004* MNase-seq data flanking intragenic TSSs to train a self-organizing map to assign TSSs with similar MNase-seq patterns to nearby nodes in a rectangular grid (Figure 2.14). This allowed us to see that, although there is considerable diversity in the nucleosome pattern surrounding intragenic TSSs, most intragenic TSSs occur in areas between the positions of nucleosome dyads. By hierarchically clustering the nodes of the self-organizing map, we further grouped intragenic TSSs into three major clusters differing primarily by the phasing of the nucleosome array relative to the TSS, as shown in Figure 2.15. In all three clusters, nucleosomes are disrupted to similar levels in *spt6-1004*.

Because GC-poor DNA sequences are nucleosome disfavoring and are known to occur in promoter regions (Iyer and Struhl, 1995; Kaplan et al., 2008; Tillo and Hughes, 2009; Zhang et al., 2009), we also examined the GC content surrounding the three clusters of intragenic TSSs. For all three clusters, the GC content of the DNA drops just upstream of the TSS to a slightly lesser degree than for genic TSSs (Figure 2.15).

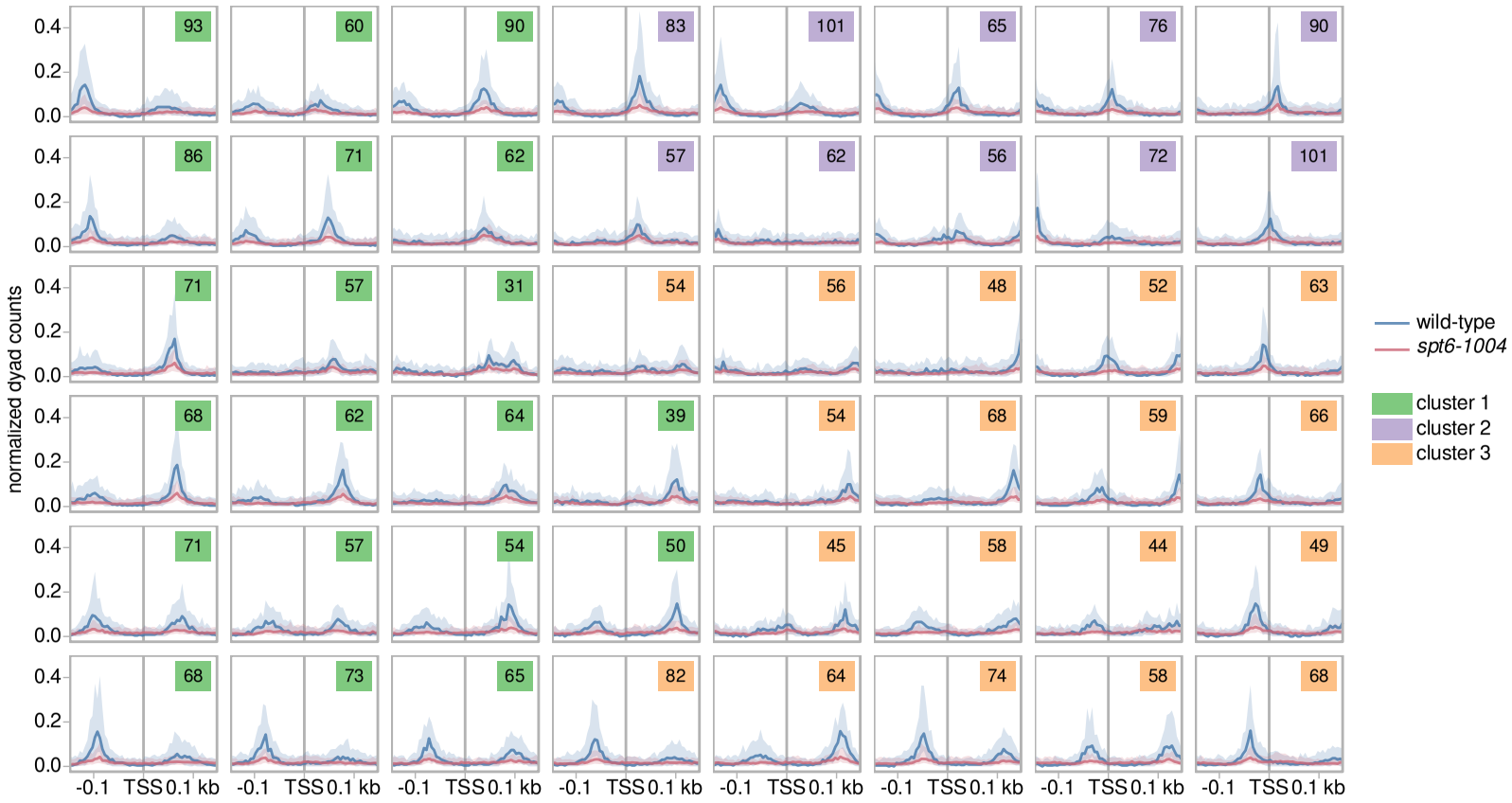

Figure 2.14: Average MNase-seq dyad signal around all *spt6-1004*-induced intragenic TSSs, grouped by assignment to nodes of a 6x8 super-organizing map (SOM). The number of TSSs assigned to each node is shown in the upper right of each panel, and is shaded by the node's assignment to a cluster determined by agglomerative hierarchical clustering of the nodes. The solid line and shading are the median and inter-quartile range of the mean spike-in normalized coverage over two replicates (*spt6-1004*) or one experiment (wild-type), in non-overlapping 5 bp bins.

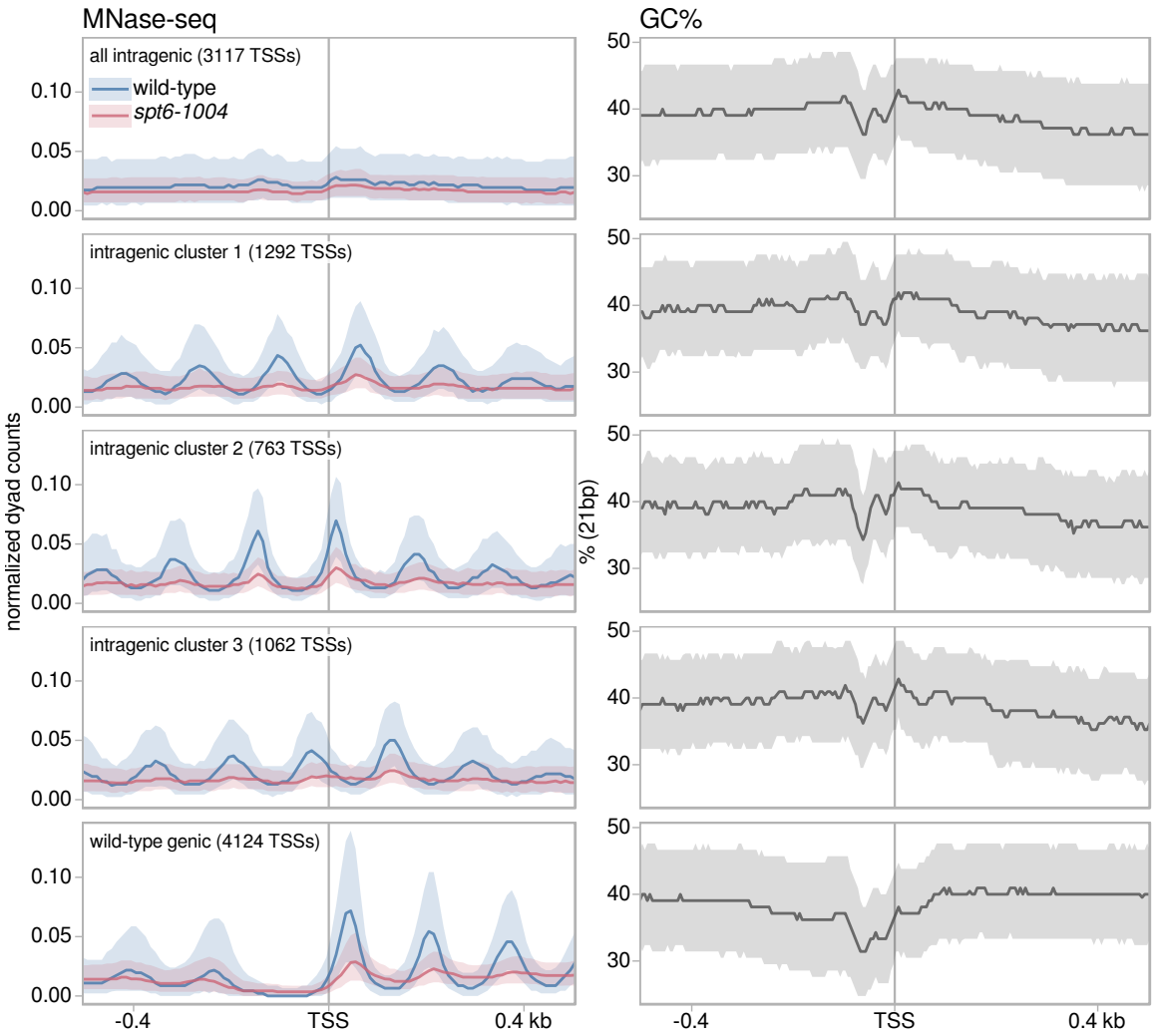

Figure 2.15:

left column) Average MNase-seq dyad signal for *spt6-1004* intragenic TSSs, both aggregated and grouped into three clusters by the wild-type and *spt6-1004* MNase-seq dyad signal flanking the TSS, as well as all genic TSSs detected in wild-type. Values are the mean of spike-in normalized dyad coverage in non-overlapping 10 bp bins, averaged over two replicates (*spt6-1004*) or one experiment (wild-type). The solid line and shading are the median and inter-quartile range.

right column) Average GC content of the DNA sequence in a 21 bp window, as above.

2.5 Other features of *spt6-1004* intragenic promoters

MNase-seq indicates that nucleosomes are lost across the entire genome in *spt6-1004*. However, TSSs observed in *spt6-1004* occur in specific locations, suggesting that loss of nucleosomes is necessary but not sufficient for intragenic transcription, and that additional features such as the drop in GC content at intragenic TSSs (Figure 2.15) may be required. The resolution with which we were able to identify intragenic TSSs allowed us to closely examine sequence features that might contribute to intragenic transcription.

2.5.1 Information content and sequence preference

To examine the DNA sequence preference of TSSs in *spt6-1004*, we aligned the sequences of all TSS-seq reads overlapping TSS-seq peaks of each class, and calculated the information content and sequence distribution for each class. Intragenic TSSs have a sequence preference almost identical to the previously observed sequence preference of genic TSSs (Figure 2.16) (Malabat et al., 2015), suggesting that RNA polymerase initiates transcription similarly at genic and intragenic TSSs.

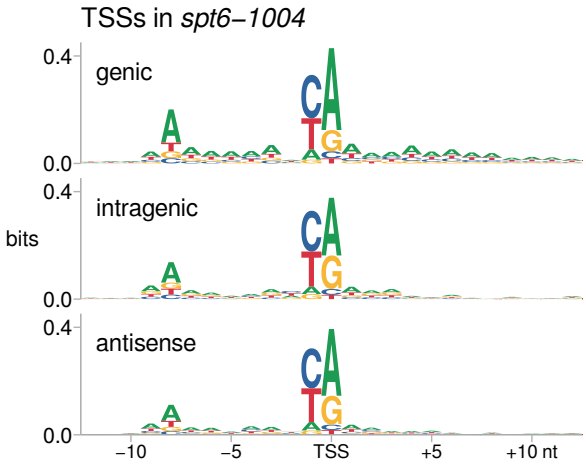

Figure 2.16: Sequence logos depicting information content and sequence preference of TSS-seq reads overlapping genic and intragenic TSS-seq peaks in *spt6-1004*.

2.5.2 Enrichment of the TATA box

A characteristic feature of canonical genic promoters is the presence of a TATA box or TATA-like DNA element which allows for the recruitment of Pol II and general transcription factors via binding of the TFIID complex, which includes TATA-binding protein (Rhee and Pugh, 2012). To examine whether the presence of TATA elements might contribute to *spt6-1004* intragenic transcription, we looked for exact matches to the TATA consensus sequence TATAWAWR in the window extending 200 nucleotides upstream of *spt6-1004* TSSs, finding matches at 13.7% of regions upstream of intragenic TSSs and 24.7% for antisense TSSs, versus 24.4% for all genic TSSs and 8.9% for random locations in the genome. Moreover, the TATA elements found near intragenic and antisense TSSs are highly concentrated in the region 50 to 100 nucleotides upstream of the TSS, where TATA elements are most often found for genic TSSs (Figure 2.17). This further supports the model that *spt6-1004* intragenic promoters are sequences similar to canonical genic promoters, which become accessible for transcription initiation when the normal chromatin state is disturbed.

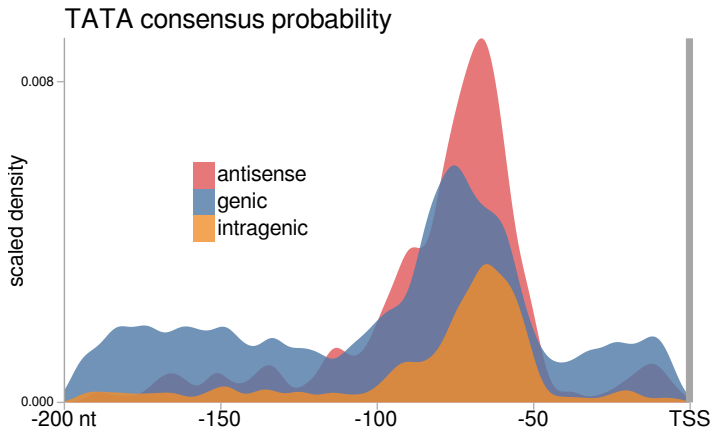

Figure 2.17: Scaled density of occurrences of exact matches to the motif TATAWAWR upstream of TSSs. For each category, a Gaussian kernel density estimate of the positions of motif occurrences is scaled by the number of motif occurrences per region.

2.5.3 Sequence motifs discovered

To discover additional sequence features of *spt6-1004* intragenic promoters, we performed *de novo* motif discovery using MEME-ChIP (Machanick and Bailey, 2011) for the regions -100 to +30 nucleotides relative to TSS summits. The most enriched motif found by MEME at both intragenic and antisense *spt6-1004* TSSs is, with respect to sense genic transcription, a GA-rich motif with 3-nucleotide periodicity (Figure 2.18). This motif occurs at only a small subset of intragenic TSSs, but is highly unlikely to occur by chance (compare the expected to observed number of occurrences in Figure 2.18). The motif is not enriched at genic TSSs upregulated in *spt6-1004*, and is not an obvious match to a DNA-binding factor in the databases searched (de Boer and Hughes, 2011; MacIsaac et al., 2006; Newburger and Bulyk, 2008; Ozonov et al., 2012; Teixeira et al., 2017; Weirauch et al., 2014; Zhu and Zhang, 1999). If this motif is directly related to intragenic transcription, we speculate that it might create a DNA structure favorable for transcription initiation.

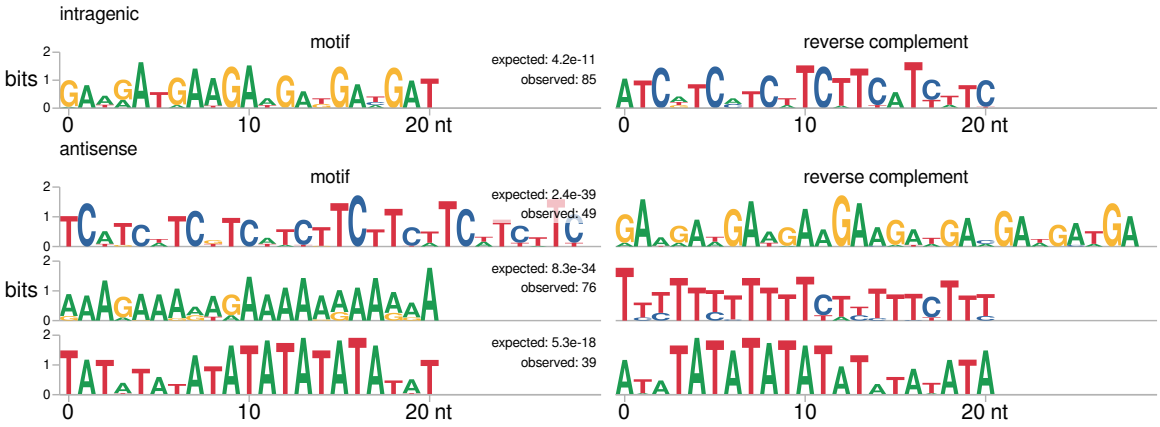

Figure 2.18: Sequence logos of motifs discovered by MEME (Bailey et al., 2015) in the window -100 to +30 bp relative to *spt6-1004* intragenic and antisense TSSs. For each motif, the observed number of occurrences and the expected number of occurrences if the input sequences were scrambled are shown.

2.6 Discussion

In this work, we integrated multiple quantitative genomic approaches to study the conserved transcription elongation factor Spt6. Our TSS-seq and TFIIB ChIP-nexus results reveal the full extent of intragenic and antisense transcript expression in *spt6-1004*, and show that these transcripts are largely explained by new RNA Pol II transcription initiation. Our MNase-seq results show that this new transcription initiation happens in the context of a global depletion and disordering of nucleosomes from chromatin. We speculate that this dramatic decrease in nucleosome protection of the genome leads to intragenic transcription by allowing initiation factors to access normally inaccessible promoter-like sequences within coding sequences. This model is supported by the similarities we observe between genic and intragenic promoters in DNA GC content, initiation motif, and TATA element frequency. This may also explain the unexpected decrease in transcription initiation we see at almost all genic promoters in *spt6-1004*: Assuming that the pool of transcription initiation factors in the cell is limiting, then making thousands of additional binding sites available to the initiation machinery would decrease the frequency at which the initiation machinery finds its correct targets at genic promoters.

2.7 Methods

2.7.1 Yeast strain construction and growth conditions

All yeast strains were constructed by standard yeast transformation or crosses. The *spt6-1004* and wild-type strains were grown as previously described (Cheung et al., 2008): Cells were grown in YPD at 30 °C to a density of approximately 1×10^7 cells/ml ($OD_{600} = 0.6$), at which point an equal volume of YPD medium pre-warmed to 44 °C was added, and the cultures were shifted to 37 °C for 80 minutes.

2.7.2 Western blotting

The protocols for western blotting and quantification are described in Doris et al. (2018).

2.7.3 Sequencing library preparation (TSS-seq, ChIP-nexus, MNase-seq, NET-seq)

All library preparation methods are detailed in Doris et al. (2018).

2.7.4 Genome builds

The genome build used for *S. cerevisiae* was R64-2-1 (Engel et al., 2014), and the genome build used for *S. pombe* was ASM294v2 (Wood et al., 2002).

2.7.5 TSS-seq data analysis

An up-to-date version of the Snakemake (Köster and Rahmann, 2012) workflow used to process TSS-seq libraries is maintained at github.com/winston-lab/tss-seq. At the time of writing, removal of adapter sequences and random hexamer sequences from the 3′ end of the read and 3′ quality trimming were performed using cutadapt (Martin,

2011). The random hexamer molecular barcode on the 5′ end of the read was then removed and processed using a custom Python script (adapted from Mayer et al. (2015)). Reads were aligned to the combined *S. cerevisiae* and *S. pombe* reference genomes using Tophat2 (Kim et al., 2013) without a reference transcriptome, and uniquely mapping reads were selected using SAMtools (Subgroup et al., 2009). Reads mapping to the same location as another read with the same molecular barcode were identified as PCR duplicates and removed using a custom Python script (adapted from Mayer et al. (2015)). Coverage of the 5′-most base, corresponding to the TSS, was extracted using bedtools genomecov (Quinlan and Hall, 2010) and normalized to the total number of uniquely mapping, non-duplicate *S. pombe* alignments. Quality statistics of raw, cleaned, non-aligning, and uniquely aligning non-duplicate reads were assessed using FastQC (Andrews, 2010).

The pipeline additionally performs TSS-seq peak calling, differential expression, classification of peaks into genomic categories, sequence logo visualization, motif enrichment analysis, *de novo* motif discovery, gene ontology analysis (Young et al., 2010), and data visualization with the option to separate data into clusters of similar signal.

2.7.5.1 Reannotation of *S. cerevisiae* TSSs using TSS-seq data

TSS-seq coverage from two replicates of a wild-type *S. cerevisiae* strain grown at 30 °C in YPD was averaged and used to adjust the 5′ ends of an annotation of major transcript isoforms based on TIF-seq data (Pelechano et al., 2013). The 5′ end of the original annotation was changed to the position of maximum TSS-seq signal in a window ± 250 nt of the original 5′ end if the maximum TSS-seq signal was greater than the 95th percentile of all non-zero TSS-seq signal.

2.7.5.2 TSS-seq peak calling

TSS-seq data representing transcription from a single promoter tends to occur as a cluster of signal distributed over a range of positions, rather than a single nucleotide (Arribere and Gilbert, 2013; Malabat et al., 2015). It is reasonable to consider such a cluster of TSS-seq signal as a single entity, because the signals within the cluster are usually highly correlated to one another across different conditions. Therefore, to identify TSSs from TSS-seq data and quantify them for downstream analyses such as differential expression, it is necessary to annotate these groups of TSS-seq signal by using the data to perform peak-calling.

At the time of writing, TSS-seq peak calling for a given experimental group was performed by 1-D watershed segmentation of the data for each sample in the group, followed by filtering for reproducibility within the group by the Irreproducible Discovery Rate (IDR) method (Li et al., 2011). First, a smoothed version of the TSS-seq coverage is generated for each sample using an adaptive two-stage kernel density estimation with a discretized Gaussian kernel (Silverman, 1986). For a given nucleotide, the adaptive kernel bandwidth, σ_{adaptive}, is given by

$$\sigma_{\text{adaptive}} = \sigma_{\text{pilot}} \left(\frac{\rho_{\text{pilot}}}{g} \right)^{-\alpha},$$

where σ_{pilot} is the standard, fixed bandwidth of a Gaussian kernel used to calculate the pilot signal density ρ_{pilot} at that nucleotide, g is the geometric mean of ρ_{pilot} over the whole genome, and α is a parameter in $[0, 1]$ that determines the degree to which the pilot density ρ_{pilot} affects σ_{adaptive}. The adaptive kernel adjusts the kernel bandwidth to be smaller in regions of high signal density and larger in regions of lower signal density, allowing the smoother to better accommodate both 'sharp' TSSs where the

signal is distributed over a relatively small window, as well as 'broad' TSSs where the signal is more dispersed. For all analyses in this document, adaptive smoothing was performed with $\sigma_{\text{pilot}} = 10$ and $\alpha = 0.2$.

Following smoothing, an initial set of peaks is formed by assigning all nonzero signal in the original, unsmoothed coverage to the nearest local maximum of the smoothed coverage, and taking the minimum and maximum genomic coordinates of the original coverage as the peak boundaries for each local maximum of the smoothed coverage. Peaks are then trimmed to the smallest genomic interval that includes 95% of the original coverage, and the probability of the peak begin generated by noise is estimated by a Poisson model where λ, the expected coverage, is the maximum of the expected coverage over the chromosome and the expected coverage in the 2 kb window upstream of the peak (à la the ChIP-seq peak caller MACS2 (Zhang et al., 2008)). Finally, peaks are ranked by their significance under the Poisson model, and a final list of peaks for the group is generated using the IDR method (IDR $= 0.1$) (Li et al., 2011). In brief, IDR compares ranked lists of regions in order to set a cutoff, beyond which the regions are no longer consistent between replicates.

The python script used for 1-D watershed segmentation of TSS-seq data is available as part of the TSS-seq pipeline, and the IDR implementation used in the pipeline is also available on GitHub.

2.7.5.3 TSS differential expression analysis

For TSS-seq differential expression analysis, TSS-seq peak-calling was performed as described above for both *S. cerevisiae* and the *S. pombe* spike-in. The read counts for each peak in each condition were used as the input to differential expression analysis by DESeq2 (Love et al., 2014), with the alternative hypothesis

$|\log_2(\text{fold-change})| > 1.5$ and a false discovery rate of 0.1. To normalize by spike-in, the size factors of the *S. pombe* spike-in counts were used as the size factors for *S. cerevisiae*, although we note that due to the median of ratios normalization used in DESeq2, the major TSS-seq results of this work are still observed when *S. cerevisiae* size factors are used.

2.7.5.4 Classification of TSS-seq peaks into genomic categories

TSS-seq peaks were assigned to genomic categories based on their position relative to the transcript annotation described above and an annotation of all verified open reading frames (ORFs) and blocked reading frames in *S. cerevisiae* (Engel et al., 2014). First, 'genic' regions were defined as follows: If a gene was present in both the transcript and ORF annotations, the genic region was defined as the interval (annotated TSS-30 nt, start codon). If gene was present in the transcript annotation but not the ORF annotation, the genic region was defined as the interval (annotated TSS - 30 nt, annotated TSS + 30 nt). If a gene was present only in the ORF annotation, the genic region was defined as the interval (start codon - 30 nt, start codon). For the purposes of peak classification, regions were considered overlapping if they had at least one base of overlap. TSS-seq peaks were classified as genic if they overlapped a genic region on the same strand. Peaks were classified as intragenic if they were not classified as a genic peak, and their summit position overlapped an open or closed reading frame on the same strand. Peaks were classified as antisense if their summit position overlapped a transcript on the opposite strand. Finally, peaks were classified as intergenic if they did not overlap a transcript, reading frame, or genic region on either strand.

2.7.5.5 TSS information content and sequence composition

TSS-seq alignments were pooled for all replicates in a condition, and the DNA sequence flanking the position of every read overlapping TSS-seq peaks of a particular genomic category was extracted using SAMtools (Subgroup et al., 2009) and bedtools (Quinlan and Hall, 2010). The information content and sequence composition of the sequences was quantified using WebLogo (Crooks et al., 2004), with the zeroth-order Markov model of the *S. cerevisiae* genomic sequence as the background composition. Sequence logos were plotting using helper functions from ggseqlogo (Wagih, 2017).

2.7.5.6 Enrichment of the TATA box

An up-to-date version of the Snakemake (Köster and Rahmann, 2012) workflow used to test the enrichment of motifs is maintained at github.com/winston-lab/motif-enrichment. To test for enrichment of consensus TATA boxes, FIMO (Grant et al., 2011) was used to search the *S.cerevisiae* genome for matches to the query motif TATAWAWR (where the ambiguous bases are equiprobable) at a p-value threshold of 6×10^{-4}. Regions extending 200 nucleotides upstream of TSS summits were merged if they were overlapping, and were considered to contain a consensus TATA box if the entire motif was overlapping the region on the same strand. The frequency of motif occurrences in the regions of interest was compared to the frequency of occurrences in the regions upstream of 6000 randomly chosen locations in the genome, using Fisher's exact test.

2.7.5.7 *De novo* motif discovery

De novo motif discovery for the regions around TSSs was performed by running MEME-ChIP (Machanick and Bailey, 2011) on the DNA sequence -100 to +30 nucleotides from the TSS summits of the genomic classes of TSSs significantly upregulated in *spt6-1004* versus wild-type.

2.7.6 ChIP-nexus data analysis

An up-to-date version of the Snakemake (Köster and Rahmann, 2012) workflow used to process ChIP-nexus libraries is maintained at github.com/winston-lab/chip-nexus. At the time of writing, filtering for reads containing the constant region of the adapter on the 5′ end of the read, 3′ adapter removal, and 3′ quality trimming were performed using cutadapt (Martin, 2011). The random pentamer molecular barcode on the 5′ end of the read was then removed and processed using a custom Python script modified from Mayer et al. (2015). Reads were aligned to the combined *S. cerevisiae* and *S. pombe* genomes using Bowtie 2 (Langmead and Salzberg, 2012), and uniquely mapping alignments were selected using SAMtools (Subgroup et al., 2009). Reads mapping to the same location as another read with the same molecular barcode were identified as PCR duplicates and removed using a custom Python script modified from Mayer et al. (2015). Coverage of the 5′-most base, corresponding to the point of crosslinking, was extracted using bedtools genomecov (Quinlan and Hall, 2010). The median fragment size estimated by MACS2 (Zhang et al., 2008) over all samples was used to generate coverage of factor protection and fragment midpoints, by extending reads to the fragment size, or by shifting reads by half the fragment size, respectively. Coverage was normalized to the total number of reads uniquely mapping to *S. cerevisiae* (the *S. pombe* spike-in was not used for normalization due to the low num-

ber of reads mapping to *S. pombe*). Quality statistics of raw, cleaned, non-aligning, and uniquely aligning non-duplicate reads were assessed using FastQC (Andrews, 2010).

The ChIP-nexus pipeline additionally performs peak calling, differential occupancy analysis, and data visualization with the option to separate data into clusters of similar signal.

An second Snakemake workflow for TFIIB-specific analyses is maintained at github.com/winston-lab/chip-nexus-tfiib, and performs classification of TFIIB peaks into genomic categories, motif enrichment analysis, and gene ontology analysis.

2.7.6.1 ChIP-nexus peak calling

A number of tools have been created specifically for peak-calling using data from high-resolution ChIP techniques such as ChIP-nexus and ChIP-exo (Hansen et al., 2016; Wang et al., 2014). When applied to our TFIIB ChIP-nexus data, these tools tended to split what appeared to be a single TFIIB binding event into multiple peaks. This may be because TFIIB crosslinks to DNA at multiple points (Rhee and Pugh, 2012), suggesting that while these tools may work well for factors that bind symmetrically to DNA with a single crosslinking point on either side, there is room for improvement for factors with more complex crosslinking patterns.

The ChIP-nexus pipeline currently performs peak calling for a condition using the standard ChIP-seq peak caller MACS2 (Zhang et al., 2008), followed by filtering for reproduciblity by the Irreproducible Discovery Rate (IDR) method (IDR = 0.1 for all analyses in this chapter) (Li et al., 2011).

2.7.6.2 TFIIB ChIP-nexus differential occupancy analysis

For TFIIB ChIP-nexus differential binding analysis, TFIIB peaks were called by MACS2 and IDR filtering as described above. A non-redundant list of peaks called in the condition and control groups being compared was generated using bedtools multiinter (Quinlan and Hall, 2010), and the counts of fragment midpoints for each peak in each sample were used as the input to differential binding analysis by DESeq2 (Love et al., 2014), with the alternative hypothesis $|\log_2(\text{fold-change})| > 1.5$ and a false discovery rate of 0.1. For estimation of change in TFIIB binding upstream of TSS-seq peaks, TFIIB fragment midpoint counts in the window extending 200 bp upstream of the TSS-seq peak summit were used as the input to DESeq2. *S. cerevisiae* counts were used for size factor calculation.

2.7.6.3 Classification of TFIIB ChIP-nexus peaks into genomic categories

As for TSS-seq peaks, TFIIB ChIP-nexus peaks were assigned to genomic categories based on their position relative to the transcript annotation described above, an annotation of all verified open reading frames (ORFs) and blocked reading frames (Engel et al., 2014), and an annotation of 'genic' regions derived from the transcript and ORF annotations. TFIIB ChIP-nexus peaks were classified as genic if they overlapped a genic region. Peaks were classified as intragenic if they were not classified as a genic peak, and the entire peak overlapped an open or closed reading frame. Finally, peaks were classified as intergenic if they did not overlap a transcript, reading frame, or genic region.

2.7.7 Comparison of TSS-seq to TFIIB ChIP-nexus

An up-to-date version of the Snakemake (Köster and Rahmann, 2012) workflow used to compare TSS-seq data to TFIIB ChIP-nexus data is maintained at github.com/winston-lab/tss-seq-vs-tfiib-nexus. The pipeline matches and compares peaks from the two assays, and also performs the TFIIB differential occupancy analysis over windows upstream of TSS-seq peaks shown in section 2.3 and described in section 2.7.6.2.

2.7.8 MNase-seq data analysis

An up-to-date version of the Snakemake (Köster and Rahmann, 2012) workflow used to demultiplex paired-end MNase-seq libraries is maintained at github.com/winston-lab/demultiplex-paired-end. At the time of writing, demultiplexing was performed using fastq-multx (Aronesty, 2013), allowing one mismatch to the barcode, followed by filtering for and removal of the barcode on read 2 using cutadapt (Martin, 2011).

An up-to-date version of the Snakemake (Köster and Rahmann, 2012) workflow used to process MNase-seq libraries is maintained at github.com/winston-lab/mnase-seq. At the time of writing, 3′ quality trimming was performed using cutadapt (Martin, 2011). Reads were aligned to the combined *S. cerevisiae* and *S. pombe* genome using Bowtie 1 (Langmead et al., 2009), and correctly paired alignments were selected using SAMtools (Subgroup et al., 2009). Coverage of nucleosome protection and nucleosome dyads were extracted using bedtools (Quinlan and Hall, 2010) and custom shell scripts to get the entire fragment or the midpoint of the fragment, respectively. Smoothed nucleosome dyad coverage was generated by smoothing dyad coverage with a Gaussian kernel of 20 bp bandwidth. Coverage was normalized to the total number of correctly paired *S. pombe* fragments. Quality statistics of raw,

cleaned, non-aligning, and correctly paired reads were assessed using FastQC (Andrews, 2010).

The MNase-seq pipeline additionally performs quantification of nucleosome properties, and data visualization with the option to separate data into clusters of similar signal.

2.7.8.1 Quantification of nucleosome properties

Quantification of nucleosome occupancy, fuzziness, and position shifts were calculated using DANPOS2 (Chen et al., 2013), with spike-in normalization by scaling the total counts in condition group libraries by

$$\frac{\text{mean observed percent spike-in in condition libraries}}{\text{mean observed percent spike-in in control libraries}}.$$

2.7.8.2 Clustering of MNase-seq signal at *spt6-1004* intragenic TSSs

The Snakemake (Köster and Rahmann, 2012) workflow for clustering MNase-seq data by self/super-organizing map and hierarchical clustering is maintained at github.com/winston-lab/cluster-mnase-seq. To cluster *spt6-1004* intragenic TSSs based on surrounding MNase-seq signal, spike-in normalized MNase-seq dyad signal in the window $\pm$150 bp of the TSS-seq peak summit of all intragenic TSS-seq peaks significantly upregulated in *spt6-1004* was binned by taking the mean signal in non-overlapping 5 bp bins, and then averaged taking the mean of two replicates (*spt6-1004*) or one experiment (wild-type). The data were then standardized over each TSS, and the wild-type and *spt6-1004* data were used as equally weighted input layers to a super-organizing map (Wehrens and Buydens, 2007) trained on the input data to assign similar MNase-seq observations in 60-dimensional input space

to similar nodes in a 2-dimensional (6 $\times$ 8) rectangular grid. The 48 'code vectors' representing the typical MNase-seq pattern for each node (visualized in Figure 2.14) were then clustered by agglomerative hierarchical clustering using sum of squares distance and Ward linkage. The resulting dendrogram was cut to produce the three clusters of MNase-seq signal shown in Figures 2.14 and 2.15.

2.7.9 NET-seq data analysis

An up-to-date version of the Snakemake (Köster and Rahmann, 2012) workflow used to process NET-seq libraries is maintained at github.com/winston-lab/net-and-rna-seq.

At the time of writing, removal of adapter sequences from the 3′ end of the read and 3′ quality trimming were performed with cutadapt (Martin, 2011). Reads were aligned to the *S. cerevisiae* genome using Tophat2 (Kim et al., 2013) without a reference transcriptome, and uniquely mapping reads were selected using SAMtools (Subgroup et al., 2009). Coverage of the 5′-most base of the read, corresponding to the 3′-most base of the nascent RNA and the active site of elongating RNA polymerase, was extracted using bedtools genomecov (Quinlan and Hall, 2010) and normalized to the total number of uniquely mapped reads. Quality statistics of raw, cleaned, non-aligning, and uniquely aligning reads were assessed using FastQC (Andrews, 2010).

The NET-seq pipeline additionally performs *ab initio* transcript annotation (Pertea et al., 2015), differential expression analysis, and data visualization with the option to split data into clusters of similar signal. For libraries with unique molecular barcodes and/or spike-ins, the pipeline also handles PCR duplicate removal and spike-in normalization, respectively.

2.8 Bibliography

Adkins, M. W. and Tyler, J. K. (2006). Transcriptional activators are dispensable for transcription in the absence of spt6-mediated chromatin reassembly of promoter regions. *Molecular Cell*, 21(3):405 – 416. 2.2

Andrews, S. (2010). Fastqc: A quality control tool for high throughput sequence data. 2.7.5, 2.7.6, 2.7.8, 2.7.9

Andrulis, E. D., Guzmán, E., Döring, P., Werner, J., and Lis, J. T. (2000). High-resolution localization of drosophila spt5 and spt6 at heat shock genes in vivo: roles in promoter proximal pausing and transcription elongation. *Genes & Development*, 14(20):2635–2649. 2.2

Ardehali, M. B., Yao, J., Adelman, K., Fuda, N. J., Petesch, S. J., Webb, W. W., and Lis, J. T. (2009). Spt6 enhances the elongation rate of rna polymerase ii in vivo. *The EMBO Journal*, 28(8):1067–1077. 2.2

Aronesty, E. (2013). Comparison of sequencing utility programs. *The Open Bioinformatics Journal*, 7:1–8. 2.7.8

Arribere, J. A. and Gilbert, W. V. (2013). Roles for transcript leaders in translation and mrna decay revealed by transcript leader sequencing. *Genome Research*, 23(6):977–987. 2.2, 2.7.5.2

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The meme suite. *Nucleic Acids Research*, 43(W1):W39–W49. 2.18

Begum, N. A., Stanlie, A., Nakata, M., Akiyama, H., and Honjo, T. (2012). The histone chaperone spt6 is required for activation-induced cytidine deaminase target determination through h3k4me3 regulation. *Journal of Biological Chemistry*, 287(39):32415–32429. 2.2

Bortvin, A. and Winston, F. (1996). Evidence that spt6p controls chromatin structure by a direct interaction with histones. *Science*, 272(5267):1473–1476. 2.2, 2.4

Carrozza, M. J., Li, B., Florens, L., Suganuma, T., Swanson, S. K., Lee, K. K., Shia, W.-J., Anderson, S., Yates, J., Washburn, M. P., and Workman, J. L. (2005). Histone h3 methylation by set2 directs deacetylation of coding regions by rpd3s to suppress spurious intragenic transcription. *Cell*, 123(4):581 – 592. 2.2

Chen, K., Xi, Y., Pan, X., Li, Z., Kaestner, K., Tyler, J., Dent, S., He, X., and Li, W. (2013). Danpos: Dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Research*, 23(2):341–351. 2.4, 2.7.8.1

Chen, S., Ma, J., Wu, F., Xiong, L.-j., Ma, H., Xu, W., Lv, R., Li, X., Villen, J., Gygi, S. P., Liu, X. S., and Shi, Y. (2012). The histone h3 lys 27 demethylase jmjd3 regulates gene expression by impacting transcriptional elongation. *Genes & Development*, 26(12):1364–1375. 2.2

Cheung, V., Chua, G., Batada, N. N., Landry, C. R., Michnick, S. W., Hughes, T. R., and Winston, F. (2008). Chromatin- and transcription-related factors repress transcription from within coding regions throughout the saccharomyces cerevisiae genome. *PLOS Biology*, 6(11):1–13. (document), 2.2, 2.2, 2.3, 2.7, 2.7.1

Chu, Y., Sutton, A., Sternglanz, R., and Prelich, G. (2006). The bur1 cyclin-dependent protein kinase is required for the normal pattern of histone methylation by set2. *Molecular and Cellular Biology*, 26(8):3029–3038. 2.2

Churchman, L. S. and Weissman, J. S. (2012). Native elongating transcript sequencing (net-seq). *Current Protocols in Molecular Biology*, 98(1):14.4.1–14.4.17. 2.4

Close, D., Johnson, S. J., Sdano, M. A., McDonald, S. M., Robinson, H., Formosa, T., and Hill, C. P. (2011). Crystal structures of the s. cerevisiae spt6 core and c-terminal tandem sh2 domain. *Journal of Molecular Biology*, 408(4):697 – 713. 2.2

Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). Weblogo: A sequence logo generator. *Genome Research*, 14(6):1188–1190. 2.7.5.5

de Boer, C. G. and Hughes, T. R. (2011). YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Research*, 40(D1):D169–D179. 2.5.3

DeGennaro, C. M., Alver, B. H., Marguerat, S., Stepanova, E., Davis, C. P., Bähler, J., Park, P. J., and Winston, F. (2013). Spt6 regulates intragenic and antisense transcription, nucleosome positioning, and histone modifications genome-wide in fission yeast. *Molecular and Cellular Biology*, 33(24):4779–4792. 2.2, 2.2, 2.2, 2.4

Diebold, M.-L., Koch, M., Loeliger, E., Cura, V., Winston, F., Cavarelli, J., and Romier, C. (2010a). The structure of an iws1/spt6 complex reveals an interaction domain conserved in tfiis, elongin a and med26. *The EMBO Journal*, 29(23):3979–3991. 2.2

Diebold, M.-L., Loeliger, E., Koch, M., Winston, F., Cavarelli, J., and Romier, C. (2010b). Noncanonical tandem sh2 enables interaction of elongation factor spt6 with rna polymerase ii. *Journal of Biological Chemistry*, 285(49):38389–38398. 2.2

Doris, S. M., Chuang, J., Viktorovskaya, O., Murawska, M., Spatt, D., Churchman, L. S., and Winston, F. (2018). Spt6 is required for the fidelity of promoter selection. *Molecular Cell*, 72(4):687 – 699.e6. 2.7.2, 2.7.3

Duina, A. A. (2011). Histone chaperones spt6 and fact: Similarities and differences in modes of action at transcribed genes. *Genet Res Int*, 2011:625210. 22567361[pmid]. 2.2, 2.4

Endoh, M., Zhu, W., Hasegawa, J., Watanabe, H., Kim, D.-K., Aida, M., Inukai, N., Narita, T., Yamada, T., Furuya, A., Sato, H., Yamaguchi, Y., Mandal, S. S., Reinberg, D., Wada, T., and Handa, H. (2004). Human spt6 stimulates transcription elongation by rna polymerase ii in vitro. *Molecular and Cellular Biology*, 24(8):3324–3336. 2.2

Engel, S. R., Dietrich, F. S., Fisk, D. G., Binkley, G., Balakrishnan, R., Costanzo, M. C., Dwight, S. S., Hitz, B. C., Karra, K., Nash, R. S., Weng, S., Wong, E. D., Lloyd, P., Skrzypek, M. S., Miyasato, S. R., Simison, M., and Cherry, J. M. (2014). The reference genome sequence of saccharomyces cerevisiae: Then and now. *G3: Genes, Genomes, Genetics*, 4(3):389–398. 2.7.4, 2.7.5.4, 2.7.6.3

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018. 2.7.5.6

Hansen, P., Hecht, J., Ibn-Salem, J., Menkuec, B. S., Roskosch, S., Truss, M., and Robinson, P. N. (2016). Q-nexus: a comprehensive and efficient analysis pipeline designed for chip-nexus. *BMC Genomics*, 17(1):873. 2.7.6.1

He, Q., Johnston, J., and Zeitlinger, J. (2015). Chip-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature Biotechnology*, 33:395 EP –. 2.2

Ivanovska, I., Jacques, P.-�., Rando, O. J., Robert, F., and Winston, F. (2011). Control of chromatin structure by spt6: Different consequences in coding and regulatory regions. *Molecular and Cellular Biology*, 31(3):531–541. 2.2, 2.4, 2.4

Iyer, V. and Struhl, K. (1995). Poly(da:dt), a ubiquitous promoter element that stimulates transcription via its intrinsic dna structure. *The EMBO Journal*, 14(11):2570–2579. 2.4.1

Jeronimo, C., Watanabe, S., Kaplan, C., Peterson, C., and Robert, F. (2015). The histone chaperones fact and spt6 restrict h2a.z from intragenic locations. *Molecular Cell*, 58(6):1113 – 1123. 2.2, 2.4

Kaplan, C. D., Laprade, L., and Winston, F. (2003). Transcription elongation factors repress transcription initiation from cryptic sites. *Science*, 301(5636):1096–1099. 2.2, 2.2, 2.2, 2.4

Kaplan, C. D., Morris, J. R., Wu, C.-t., and Winston, F. (2000). Spt5 and spt6 are associated with active transcription and have characteristics of general elongation factors in d. melanogaster. *Genes & Development*, 14(20):2623–2634. 2.2

Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., Field, Y., LeProust, E. M., Hughes, T. R., Lieb, J. D., Widom, J., and Segal, E. (2008). The dna-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458:362 EP –. 2.4.1

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36. 2.7.5, 2.7.9

Krogan, N. J., Kim, M., Ahn, S. H., Zhong, G., Kobor, M. S., Cagney, G., Emili, A., Shilatifard, A., Buratowski, S., and Greenblatt, J. F. (2002). Rna polymerase ii elongation factors of saccharomyces cerevisiae: a targeted proteomics approach. *Molecular and Cellular Biology*, 22(20):6979–6992. 2.2

Köster, J. and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522. 2.7.5, 2.7.5.6, 2.7.6, 2.7.7, 2.7.8, 2.7.8.2, 2.7.9

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9:357 EP –. 2.7.6

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology*, 10(3):R25. 2.7.8

Li, Q., Brown, J. B., Huang, H., and Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, 5(3):1752–1779. 2.7.5.2, 2.7.6.1

Li, S., Almeida, A. R., Radebaugh, C. A., Zhang, L., Chen, X., Huang, L., Thurston, A. K., Kalashnikova, A. A., Hansen, J. C., Luger, K., and Stargell, L. A. (2018). The elongation factor spn1 is a multi-functional chromatin binding protein. *Nucleic Acids Research*, 46(5):2321–2334. 2.2

Lickwar, C. R., Rao, B., Shabalin, A. A., Nobel, A. B., Strahl, B. D., and Lieb, J. D. (2009). The set2/rpd3s pathway suppresses cryptic transcription without regard to gene length or transcription frequency. *PLOS ONE*, 4(3):1–7. 2.2

Liu, J., Zhang, J., Gong, Q., Xiong, P., Huang, H., Wu, B., Lu, G., Wu, J., and Shi, Y. (2011). Solution structure of tandem sh2 domains from spt6 protein and their binding to the phosphorylated rna polymerase ii c-terminal domain. *Journal of Biological Chemistry*, 286(33):29218–29226. 2.2

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):550. 2.7.5.3, 2.7.6.2

Machanick, P. and Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, 27(12):1696–1697. 2.5.3, 2.7.5.7

MacIsaac, K. D., Wang, T., Gordon, D. B., Gifford, D. K., Stormo, G. D., and Fraenkel, E. (2006). An improved map of conserved regulatory sites for saccharomyces cerevisiae. *BMC Bioinformatics*, 7(1):113. 2.5.3

Malabat, C., Feuerbach, F., Ma, L., Saveanu, C., and Jacquier, A. (2015). Quality control of transcription start site selection by nonsense-mediated-mrna decay. *eLife*, 4:e06722. 2.2, 2.5.1, 2.7.5.2

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12. 2.7.5, 2.7.6, 2.7.8, 2.7.9

Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J. A., and Churchman, L. S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell*, 161(3):541–554. 2.7.5, 2.7.6

Mayer, A., Lidschreiber, M., Siebert, M., Leike, K., Söding, J., and Cramer, P. (2010). Uniform transitions of the general rna polymerase ii transcription complex. *Nature Structural & Molecular Biology*, 17:1272–1278. 2.2

McCullough, L., Connell, Z., Petersen, C., and Formosa, T. (2015). The abundant histone chaperones spt6 and fact collaborate to assemble, inspect, and maintain chromatin structure in saccharomyces cerevisiae. *Genetics*, 201(3):1031–1045. 2.2

McDonald, S. M., Close, D., Xin, H., Formosa, T., and Hill, C. P. (2010). Structure and biological importance of the spn1-spt6 interaction, and its regulatory role in nucleosome binding. *Molecular Cell*, 40(5):725 – 735. 2.2

Newburger, D. E. and Bulyk, M. L. (2008). UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. *Nucleic Acids Research*, 37(suppl_1):D77–D82. 2.5.3

Ozonov, E., Pachkov, M., Arnold, P., Balwierz, P. J., and van Nimwegen, E. (2012). SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Research*, 41(D1):D214–D220. 2.5.3

Pathak, R., Singh, P., Ananthakrishnan, S., Adamczyk, S., Schimmel, O., and Govind, C. K. (2018). Acetylation-dependent recruitment of the fact complex and its role in regulating pol ii occupancy genome-wide in saccharomyces cerevisiae. *Genetics*, 209(3):743–756. 2.2

Pelechano, V., Wei, W., and Steinmetz, L. M. (2013). Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*, 497:127 EP –. 2.7.5.1

Perales, R., Erickson, B., Zhang, L., Kim, H., Valiquett, E., and Bentley, D. (2013). Gene promoters dictate histone occupancy within genes. *The EMBO Journal*, 32(19):2645–2656. 2.2, 2.4

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nature Biotechnology*, 33:290 EP –. 2.7.9

Quinlan, A. R. and Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842. 2.7.5, 2.7.5.5, 2.7.6, 2.7.6.2, 2.7.8, 2.7.9

Rhee, H. S. and Pugh, B. F. (2012). Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*, 483:295 EP –. Article. 2.5.2, 2.7.6.1

Sdano, M. A., Fulcher, J. M., Palani, S., Chandrasekharan, M. B., Parnell, T. J., Whitby, F. G., Formosa, T., and Hill, C. P. (2017). A novel sh2 recognition mechanism recruits spt6 to the doubly phosphorylated rna polymerase ii linker at sites of transcription. *eLife*, 6:e28723. 2.2

Shivaswamy, S., Bhinge, A., Zhao, Y., Jones, S., Hirst, M., and Iyer, V. R. (2008). Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLOS Biology*, 6(3):1–13. 2.4

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, chapter 5.3, pages 100–110. Chapman & Hall. 2.7.5.2

Subgroup, . G. P. D. P., Wysoker, A., Handsaker, B., Marth, G., Abecasis, G., Li, H., Ruan, J., Homer, N., Durbin, R., and Fennell, T. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079. 2.7.5, 2.7.5.5, 2.7.6, 2.7.8, 2.7.9

Sun, M., Larivière, L., Dengl, S., Mayer, A., and Cramer, P. (2010). A tandem sh2 domain in transcription elongation factor spt6 binds the phosphorylated rna polymerase ii c-terminal repeat domain (ctd). *Journal of Biological Chemistry*, 285(53):41597–41603. 2.2

Teixeira, M. C., Monteiro, P. T., Palma, M., Costa, C., Godinho, C. P., Pais, P., Cavalheiro, M., Antunes, M., Lemos, A., Pedreira, T., and Sá-Correia, I. (2017). YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae. *Nucleic Acids Research*, 46(D1):D348–D353. 2.5.3

Tillo, D. and Hughes, T. R. (2009). G+c content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics*, 10(1):442. 2.4.1

Uwimana, N., Collin, P., Jeronimo, C., Haibe-Kains, B., and Robert, F. (2017). Bidirectional terminators in saccharomyces cerevisiae prevent cryptic transcription from invading neighboring genes. *Nucleic Acids Research*, 45(11):6417–6426. (document), 2.2, 2.2, 2.3, 2.7

van Bakel, H., Tsui, K., Gebbia, M., Mnaimneh, S., Hughes, T. R., and Nislow, C. (2013). A compendium of nucleosome and transcript profiles reveals determinants of chromatin architecture and transcription. *PLOS Genetics*, 9(5):1–18. 2.2, 2.2, 2.4

Wagih, O. (2017). ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*, 33(22):3645–3647. 2.7.5.5

Wang, A. H., Juan, A. H., Ko, K. D., Tsai, P.-F., Zare, H., Dell'Orso, S., and Sartorelli, V. (2017). The elongation factor spt6 maintains esc pluripotency by controlling super-enhancers and counteracting polycomb proteins. *Molecular Cell*, 68(2):398 – 413.e6. 2.2

Wang, A. H., Zare, H., Mousavi, K., Wang, C., Moravec, C. E., Sirotkin, H. I., Ge, K., Gutierrez-Cruz, G., and Sartorelli, V. (2013). The histone chaperone spt6 coordinates histone h3k27 demethylation and myogenesis. *The EMBO Journal*, 32(8):1075–1086. 2.2

Wang, L., Chen, J., Wang, C., Uusküla-Reimand, L., Chen, K., Medina-Rivera, A., Young, E. J., Zimmermann, M. T., Yan, H., Sun, Z., Zhang, Y., Wu, S. T., Huang, H., Wilson, M. D., Kocher, J.-P. A., and Li, W. (2014). Mace: model based analysis of chip-exo. *Nucleic Acids Research*, 42(20):e156. 2.7.6.1

Wehrens, R. and Buydens, L. (2007). Self- and super-organizing maps in r: The kohonen package. *Journal of Statistical Software, Articles*, 21(5):1–19. 2.7.8.2

Weirauch, M., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H., Lambert, S., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J., Govindarajan, S., Shaulsky, G., Walhout, A., Bouget, F.-Y., Ratsch, G., Larrondo, L., Ecker, J., and Hughes, T. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6):1431 – 1443. 2.5.3

Wood, V., Gwilliam, R., Rajandream, M.-A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., Basham, D., Bowman, S., Brooks, K., Brown, D., Brown, S., Chillingworth, T., Churcher, C., Collins, M., Connor, R., Cronin, A., Davis, P., Feltwell, T., Fraser, A., Gentles, S., Goble, A., Hamlin, N., Harris, D., Hidalgo, J., Hodgson, G., Holroyd, S., Hornsby, T., Howarth, S., Huckle, E. J., Hunt, S., Jagels, K., James, K., Jones, L., Jones, M., Leather, S., McDonald, S., McLean, J., Mooney, P., Moule, S., Mungall, K., Murphy, L., Niblett, D., Odell, C., Oliver, K., O'Neil, S., Pearson, D., Quail, M. A., Rabbinowitsch, E., Rutherford, K., Rutter, S., Saunders, D., Seeger, K., Sharp, S., Skelton, J., Simmonds, M., Squares, R., Squares, S., Stevens, K., Taylor, K., Taylor, R. G., Tivey, A., Walsh, S., Warren, T., Whitehead, S., Woodward, J., Volckaert, G., Aert, R., Robben, J., Grymonprez, B., Weltjens, I., Vanstreels, E., Rieger, M., Schäfer, M., Müller-Auer, S., Gabel, C., Fuchs, M., Fritzc, C., Holzer, E., Moestl, D., Hilbert, H., Borzym, K., Langer, I., Beck, A., Lehrach, H., Reinhardt, R., Pohl, T. M., Eger, P., Zimmermann, W., Wedler, H., Wambutt, R., Purnelle, B., Goffeau, A., Cadieu, E., Dréano, S., Gloux, S., Lelaure, V., Mottier, S., Galibert, F., Aves, S. J., Xiang, Z., Hunt, C., Moore, K., Hurst, S. M., Lucas, M., Rochet, M., Gaillardin, C., Tallada, V. A., Garzon, A., Thode, G., Daga, R. R., Cruzado, L., Jimenez, J., Sánchez, M., del Rey, F., Benito, J., Domínguez, A., Revuelta, J. L., Moreno, S., Armstrong, J., Forsburg, S. L., Cerrutti, L., Lowe, T., McCombie, W. R., Paulsen, I., Potashkin, J., Shpakovski, G. V., Ussery, D., Barrell, B. G., and Nurse, P. (2002). The genome sequence of schizosaccharomyces pombe. *Nature*, 415(6874):871–880. 2.7.4

Yoh, S. M., Cho, H., Pickle, L., Evans, R. M., and Jones, K. A. (2007). The spt6 sh2 domain binds ser2-p rnapii to direct iws1-dependent mrna splicing and export. *Genes & Development*, 21(2):160–174. 2.2

Yoh, S. M., Lucas, J. S., and Jones, K. A. (2008). The iws1:spt6:ctd complex controls cotranscriptional mrna biosynthesis and hypb/setd2-mediated histone h3k36 methylation. *Genes & Development*, 22(24):3422–3434. 2.2

Youdell, M. L., Kizer, K. O., Kisseleva-Romanova, E., Fuchs, S. M., Duro, E., Strahl, B. D., and Mellor, J. (2008). Roles for ctk1 and spt6 in regulating the different methylation states of histone h3 lysine 36. *Molecular and Cellular Biology*, 28(16):4915–4926. 2.2

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for rna-seq: accounting for selection bias. *Genome Biology*, 11(2):R14. 2.7.5

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of chip-seq (macs). *Genome Biology*, 9(9):R137. 2.7.5.2, 2.7.6, 2.7.6.1

Zhang, Y., Moqtaderi, Z., Rattner, B. P., Euskirchen, G., Snyder, M., Kadonaga, J. T., Liu, X. S., and Struhl, K. (2009). Intrinsic histone-dna interactions are not the major determinant of nucleosome positions in vivo. *Nature Structural &Amp; Molecular Biology*, 16:847 EP –. Article. 2.4.1

Zhu, J. and Zhang, M. Q. (1999). SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics*, 15(7):607–611. 2.5.3

Chapter 3

Genomics of transcription elongation factor Spt5

3.1 Collaborators

Ameet Shetty generated TSS-seq, MNase-seq, NET-seq, RNA-seq, and ChIP-seq libraries

3.2 Introduction to Spt5 and Spt5 depletion

Spt5 is a fundamental component of the transcription elongation complex, with the distinction of being the only RNA polymerase-associated factor known to be conserved across all three domains of life (Hartzog and Fu, 2013; Werner, 2012). In eukaryotes, Spt5 heterodimerizes with the protein Spt4, forming a complex known as DSIF (DRB sensitivity-inducing factor) (Hartzog et al., 1998; Hirtreiter et al., 2010; Schwer et al., 2009; Wada et al., 1998). In metazoans, phosphorylation of DSIF controls the release of the elongation complex from promoter-proximal pausing, a regulatory transition state between transcription initiation and productive elongation (Adelman and Lis, 2012).

Within the elongation complex, biochemical and structural studies place Spt4/5 near the center of the action: Spt5 directly interacts with the noncoding strand of DNA (Crickard et al., 2016; Meyer et al., 2015), the nascent RNA (Blythe et al., 2016;

Crickard et al., 2016; Meyer et al., 2015), and the Pol II clamp domain, which sits above the nucleic acid cleft (Hirtreiter et al., 2010; Martinez-Rucobo et al., 2011; Viktorovskaya et al., 2011; Yamaguchi et al., 1999). Binding of Spt5 to Pol II is likely to stabilize the elongation complex and enhance its processivity (Hirtreiter et al., 2010; Klein et al., 2011; Martinez-Rucobo et al., 2011), consistent with both *in vitro* studies showing that Spt5 reduces pausing of polymerase under nucleotide-limiting conditions (Guo et al., 2000; Wada et al., 1998; Zhu et al., 2007), and *in vivo* studies showing elongation defects upon Spt4/5 mutation or depletion (Diamant et al., 2016a; Kramer et al., 2016; Liu et al., 2012; Mason and Struhl, 2005; Morillon et al., 2003; Quan and Hartzog, 2010; Rondón et al., 2003).

As it travels with elongating Pol II, Spt5 recruits other factors, including the Rpd3S histone deactylase complex (Drouin et al., 2010) and mRNA 3′-end processing factors (Mayer et al., 2012; Stadelmayer et al., 2014; Yamamoto et al., 2014). The recruitment of still other factors to the elongation complex by Spt5 is dependent on the phosphorylation status of the Spt5 C-terminal region (CTR), a domain composed of tandem repeats analogous to the RNA Pol II C-terminal domain. When unphosphorylated, the Spt5 CTR aids in recruitment of the mRNA capping enzyme (Doamekpor et al., 2014, 2015; Schneider et al., 2010; Wen and Shatkin, 1999), while when phosphorylated, the CTR recruits the Paf1 complex (Liu et al., 2009; Mbogning et al., 2013; Wier et al., 2013; Zhou et al., 2009), another complex involved in transcription elongation.

Despite the close relationship of Spt5 to transcription and the transcription-associated processes described above, previous studies which knocked down Spt5 in zebrafish, mice, and HeLa cells observed only mild changes in transcript levels across the genome (Diamant et al., 2016b; Komori et al., 2009; Krishnan et al., 2008; Stanlie et al., 2012),

a result which could potentially be explained by inefficient knockdown of Spt5 and/or the lack of a spike-in control, without which it is impossible to observe a global change over the entire genome (Chen et al., 2016).

To study Spt5 by efficiently disrupting its function, we use a system for conditionally depleting Spt5 protein in the fission yeast *Schizosaccharomyces pombe*. In this system, Spt5 is expressed using the thiamine-repressible *nmt81* promoter, and is tagged with an auxin-inducible degron tag (Kanke et al., 2011), such that addition of thiamine and auxin to the media results in repression of *spt5* transcription and specific degradation of Spt5 protein (Figure 3.1). For all experiments described in this chapter, Spt5-depleted cells are sampled 4.5 hours after the start of depletion, at which point the levels of Spt5 on chromatin are about 12% of non-depleted levels, as measured by ChIP-seq of Spt5 (Figure 3.2, top panel).

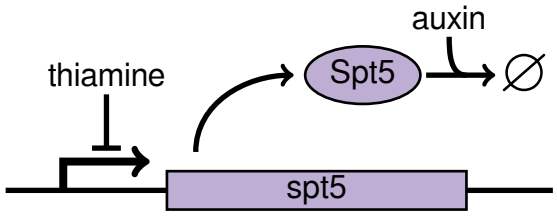

Figure 3.1: Diagram of the dual-shutoff system used to deplete Spt5 from *S. pombe*. Spt5 is expressed from a thiamine-repressible promoter, and tagged with an auxin-inducible degron tag for specific degradation upon addition of auxin.

Using this system, we assayed various aspects of transcription and chromatin state across the genomes of Spt5-depleted and non-depleted cells. The results are presented below, with section 3.3 and the RNA-seq part of section 3.4 describing my reanalyses of data published in Shetty et al. (2017) prior to my involvement in this project.

3.3 RNA Polymerase II in Spt5 depletion

To examine the effects of Spt5 depletion on transcription, we performed Pol II ChIP-seq and NET-seq of Spt5-depleted and non-depleted cells. The data from the two assays paint somewhat different pictures of the changes in Pol II status upon Spt5 depletion (Figure 3.2). ChIP-seq reports that global levels of Pol II on chromatin in Spt5-depleted cells are roughly one-third that of non-depleted cells, with the polymerase remaining after Spt5 depletion tending to be located towards the 5′ ends of genes. By contrast, NET-seq reports that total elongating Pol II is not depleted to an appreciable degree but is redistributed from the 3′ to the 5′ ends of genes, with a greater 5′ bias than that observed by ChIP-seq. We interpret this 5′ shift of Pol II as reflecting a transcription elongation defect upon Spt5 depletion, consistent with previously reported elongation defects upon Spt4/5 disruption (Diamant et al., 2016a; Kramer et al., 2016; Liu et al., 2012; Mason and Struhl, 2005; Morillon et al., 2003; Quan and Hartzog, 2010; Rondón et al., 2003).

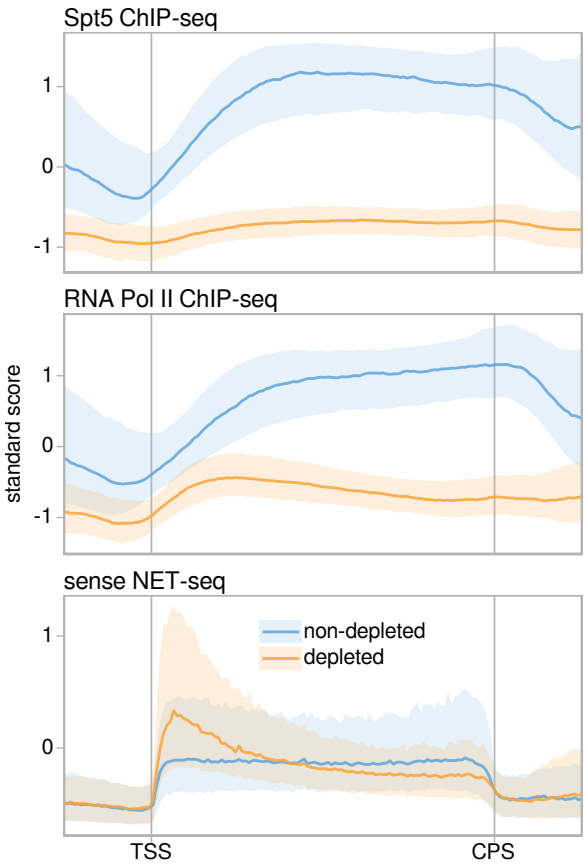

Figure 3.2: Average Spt5 ChIP-seq, RNA Pol II ChIP-seq, and sense NET-seq signal in Spt5 non-depleted and depleted cells, over 1989 non-overlapping coding genes scaled from TSS to CPS. The solid line and shading are the median and inter-quartile range of the mean spike-in normalized coverage over two replicates or one experiment (non-depleted NET-seq), taken in non-overlapping 20 bp bins and standardized per gene.

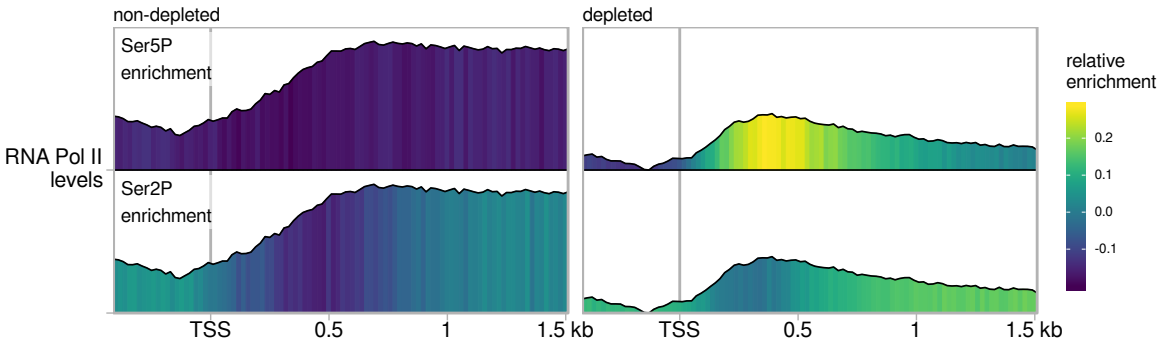

Figure 3.3: Median standardized RNA Pol II ChIP-seq signal and relative enrichment of Pol II phospho-serine 5 and phospho-serine 2 ChIP-seq signal in Spt5 depleted and non-depleted cells, over 1989 non-overlapping coding genes aligned by non-depleted genic TSS. ChIP-seq coverage is spike-in normalized and input-subtracted, and relative enrichment of Pol II modifications is a normalized log-ratio of modification coverage over Pol II coverage.

To learn more about the state of Pol II in Spt5 depletion, we also performed ChIP-seq for two major post-translational modifications of the Pol II C-terminal domain (CTD), namely serine 5 phosphorylation and serine 2 phosphorylation. Looking at the relative enrichment of these modifications over gene bodies, we see that the CTD of the Pol II remaining at the 5′ ends of genes after Spt5 depletion is enriched for phospho-serine 5 and depleted for phospho-serine 2. This is somewhat expected due to the respective tendencies of phospho-serine 5 and 2 to occur towards the 5′ and 3′ ends of genes (Komarnitsky et al., 2000). However, the 5′ enrichment of phospho-serine 5 seen in Spt5 depleted cells is not observed in non-depleted cells (note the uniformity of relative Ser5P enrichment in non-depleted cells in figure 3.3).

One possible explanation for the apparent discrepancy between the ChIP-seq and NET-seq results is the difference in immunoprecipitation strategy between the two techniques. The antibody used to pull down Pol II for ChIP-seq was 8WG16,

which recognizes the Pol II CTD. Reports of this antibody's relative affinity for the various CTD phosphoisoforms vary widely across studies in several species (Zeitlinger et al., 2007). It is conceivable that, for *S. pombe*, the 8WG16 antibody might fail to efficiently pull down a 5′-biased phosphoisoform of Pol II that would be captured by NET-seq, a technique that should theoretically capture all Pol II phosphoisoforms via FLAG pulldown of the Rpb3 subunit of Pol II. If this were the case, it could explain the relative lack of Pol II ChIP-seq signal from Spt5 non-depleted cells over the first 500 base pairs of genes (Figure 3.3). Furthermore, if the levels of this missing CTD phosphoisoform were elevated in Spt5-depleted cells versus non-depleted cells, this could also explain the difference in Spt5-depleted total Pol II levels on chromatin as observed by ChIP-seq and NET-seq.

This missing CTD phosphoisoform is not likely to be serine 5 phosphorylation, because ChIP-seq of this mark looks very much like ChIP-seq of Pol II (again, note the uniformity of relative Ser5P enrichment in non-depleted cells in figure 3.3). One possible candidate is serine 7 phosphorylation, a modification made early in transcription initiation which has been shown in an *in vitro* human system to be a preferred substrate for P-TEFb to carry out subsequent phosphorylation of serine 5 (Czudnochowski et al., 2012).

3.4 The transcriptome in Spt5 depletion

Given the transcriptional changes observed after Spt5 depletion, we performed RNA-seq and TSS-seq to further see how the depletion affects steady-state transcript levels. Changes to the levels of genic transcripts after Spt5 depletion are generally mild, with the median gene being expressed at roughly 68% of non-depleted levels as measured by RNA-seq with an RNA spike-in, or at 104% of non-depleted levels as mea-

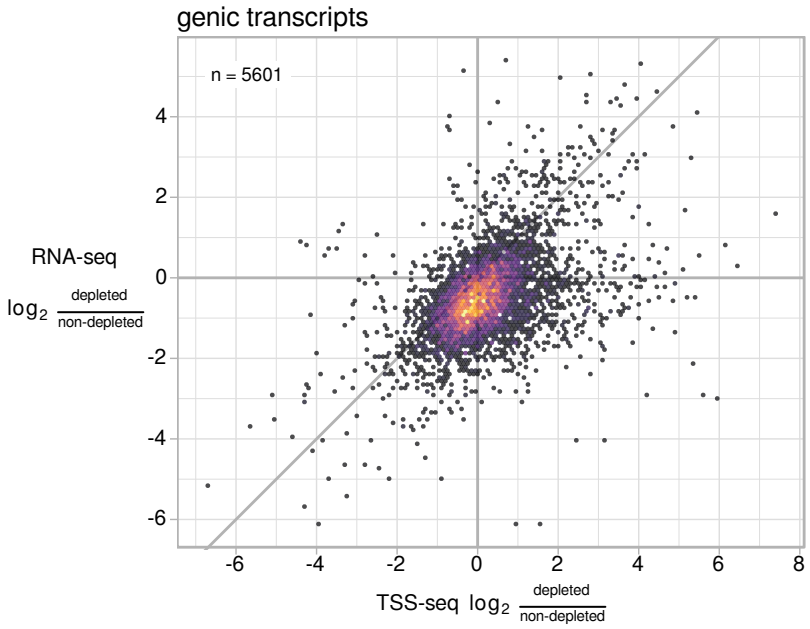

Figure 3.4: Scatterplot of fold-change in Spt5-depleted over non-depleted cells, comparing TSS-seq and RNA-seq. Each dot represents an RNA-seq measurement over an annotated transcript paired with a TSS-seq measurement over a genic peak assigned to the transcript. Fold-changes are regularized fold-change estimates from DESeq2, with size factors determined from ERCC RNA spike-in counts (RNA-seq) or *S. cerevisiae* cell spike-in counts (TSS-seq).

sured by TSS-seq with a cell spike-in (Figure 3.4). The difference in transcript abundance measurements between RNA-seq and TSS-seq can be explained by a change in the distribution of RNA-seq signal over genes in Spt5 depletion: RNA-seq signal is generally reduced over genes, except near the TSS at the very 5′ end of genes (Figure 3.5). We attribute this to defective elongation upon Spt5 depletion, which increases transcriptional pausing, early termination, and the use of intragenic polyadenylation sites, consistent with a previous report in budding yeast (Cui and Denis, 2003).

RNA-seq and TSS-seq also revealed the expression of many novel transcripts in Spt5 depletion, including over 900 antisense transcripts which tend to initiate within

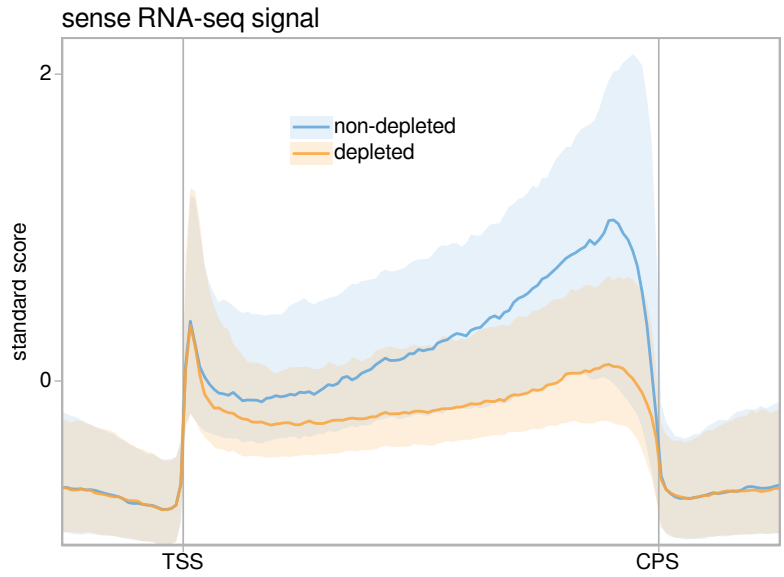

Figure 3.5: Average sense RNA-seq signal in Spt5 non-depleted and depleted cells, over 1989 non-overlapping coding genes scaled from TSS to CPS. The solid line and shading are the median and inter-quartile range of the mean spike-in normalized coverage over two replicates, taken in non-overlapping 20 bp bins and standardized per gene.

the first 500 base pairs downstream of the genic TSS (Figures 3.6, 3.7, 3.8). In general, these Spt5-repressed antisense transcripts are only a few hundred nucleotides in length (Figure 3.7), and are expressed at a lower level than canonical genic transcripts (Figure 3.9). We also find no notable correlation upon Spt5 depletion between changes in antisense transcription and changes in overlapping genic transcription. Interestingly, the most significant motif found by *de novo* motif discovery upstream of Spt5-repressed antisense TSSs is a GA-rich motif with 3-nucleotide periodicity, similar to the most significant motif found upstream of Spt6-repressed intragenic and antisense TSSs in *S. cerevisiae* (Figures 3.10, 2.18).

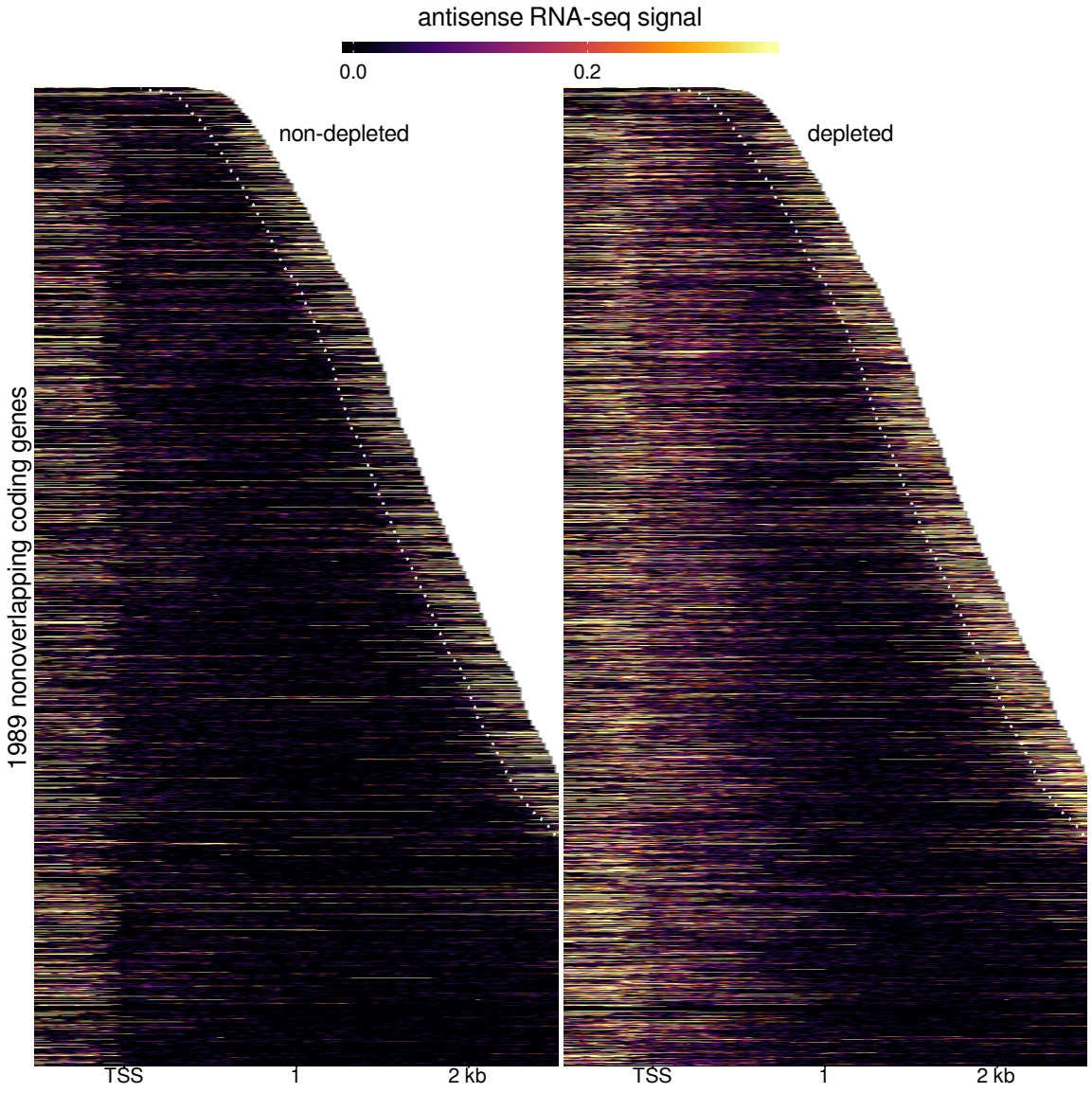

Figure 3.6: Heatmaps of antisense RNA-seq signal from Spt5 depleted and non-depleted cells, over 1989 non-overlapping coding genes aligned by non-depleted genic TSS and sorted by annotated transcript length. Data are shown for each gene up to 300 nucleotides 3′ of the CPS, indicated by the white dotted line. Values are the mean of spike-in normalized coverage in non-overlapping 20 nucleotide bins, averaged over two replicates. Values above the 93rd percentile are set to the 93rd percentile for visualization.

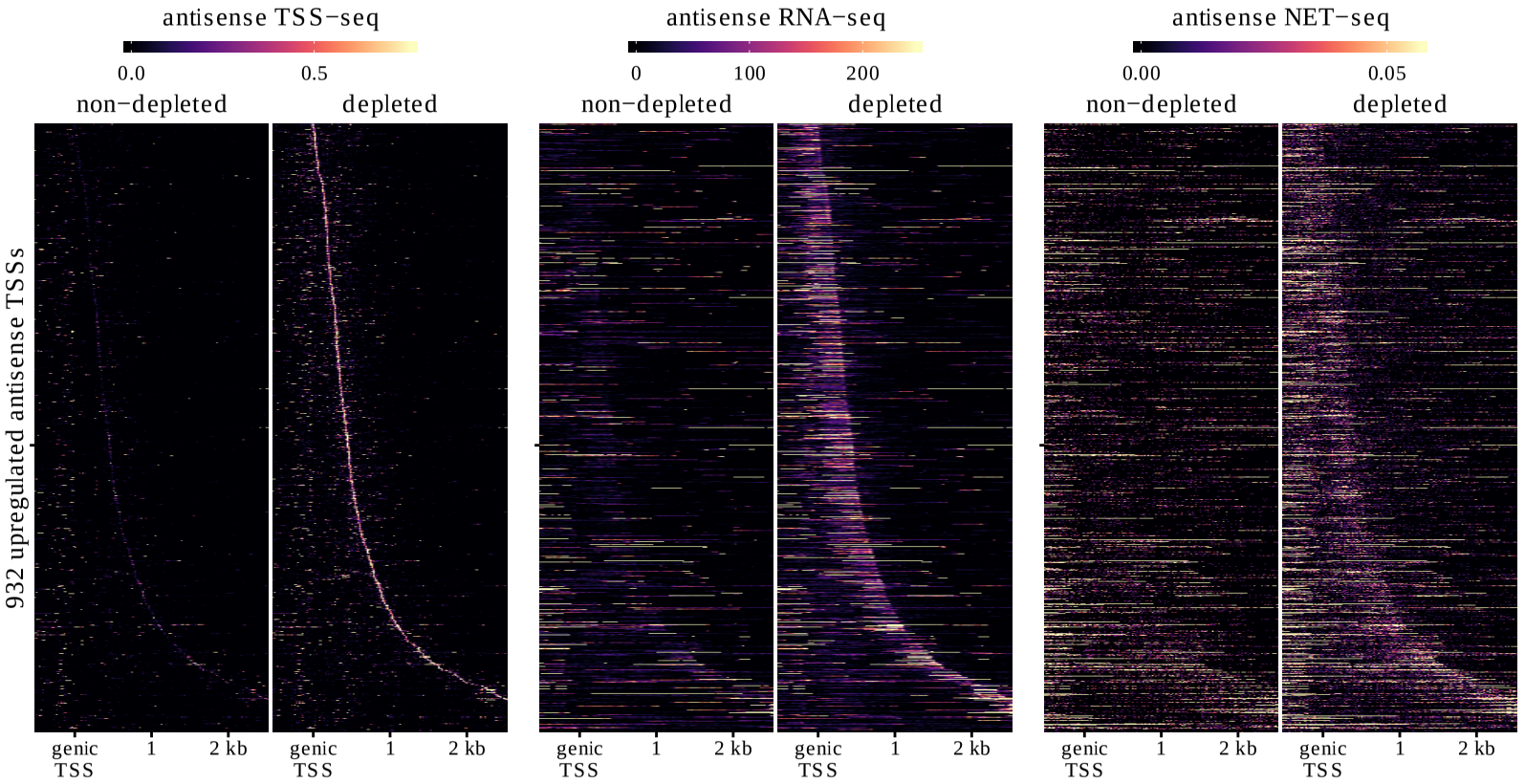

Figure 3.7: Heatmaps of antisense TSS-seq, RNA-seq, and NET-seq signal in Spt5 non-depleted and depleted cells, over all genes overlapping an Spt5-depletion-induced antisense TSS, aligned by the sense, genic TSS and sorted by the distance from the genic TSS to the antisense TSS. Values are the mean of spike-in normalized coverage in non-overlapping 20 nt bins, over one (non-depleted NET-seq) or more experiments. Values above the 0.995 (TSS-seq), 0.98 (RNA-seq), or 0.96 (NET-seq) quantiles are set to their respective quantiles for visualization.

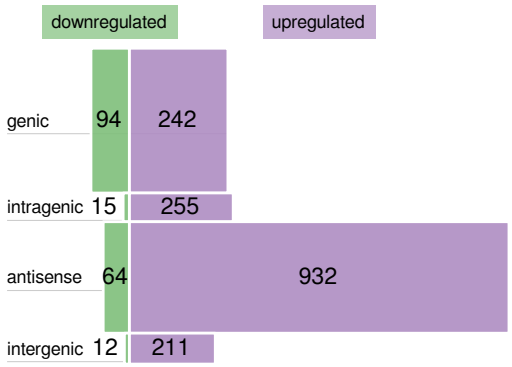

Figure 3.8: Bar plot of the number of TSS-seq peaks differentially expressed in Spt5-depleted versus non-depleted cells. The height of each bar is proportional to the total number of peaks in the category, including those not found to be significantly differentially expressed.

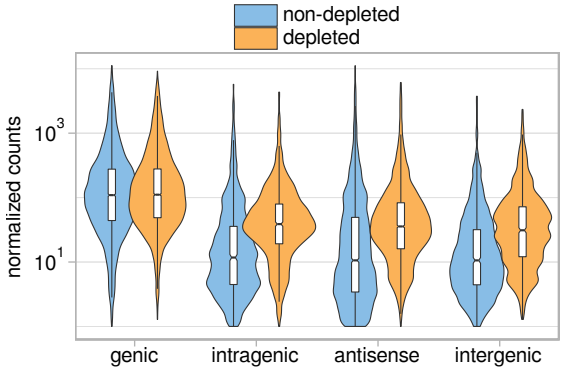

Figure 3.9: Violin plots of expression level distributions for genomic classes of TSS-seq peaks in Spt5-depleted and non-depleted cells. Normalized counts are the mean of spike-in size factor normalized counts from four (non-depleted) or two (depleted) replicates.

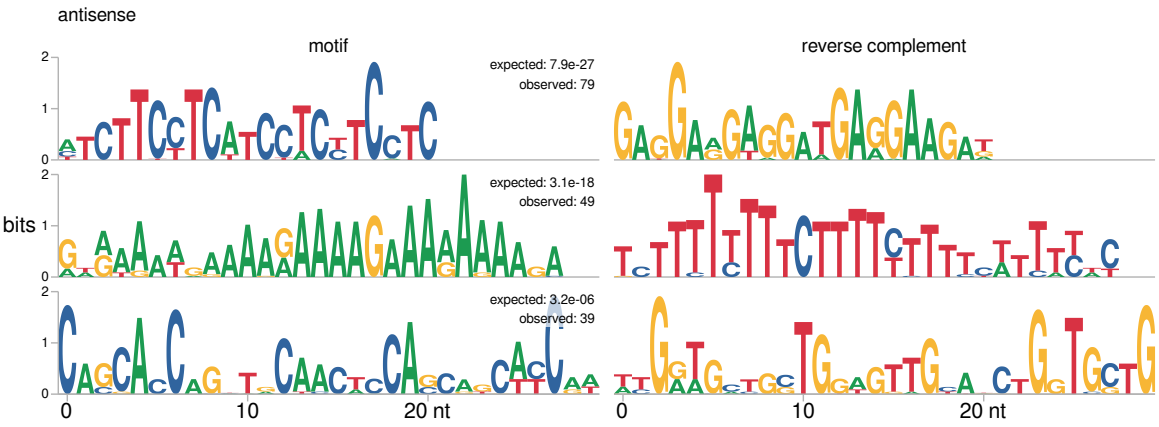

Figure 3.10: Sequence logos of motifs discovered by MEME (Bailey et al., 2015) in the window -100 to +30 bp relative to Spt5-depletion-induced antisense TSSs. For each motif, the observed number of occurrences and the expected number of occurrences if the input sequences were scrambled are shown.

3.5 The chromatin landscape in Spt5 depletion

One hypothesis for why antisense transcripts are expressed upon Spt5 depletion is that changes in chromatin structure create an environment permissive for transcription initiation. To observe possible changes to chromatin, we performed MNase-seq of Spt5-depleted and non-depleted cells (Figure 3.11). Because no spike-in control was included in the experiment, we were unable to use the data to quantify nucleosome occupancy, however, the data do indicate that nucleosomes generally become less well-positioned upon Spt5 depletion (Figure 3.12), and that the severity of these changes increases as one moves downstream from the +1 nucleosome into gene bodies (Figure 3.11).

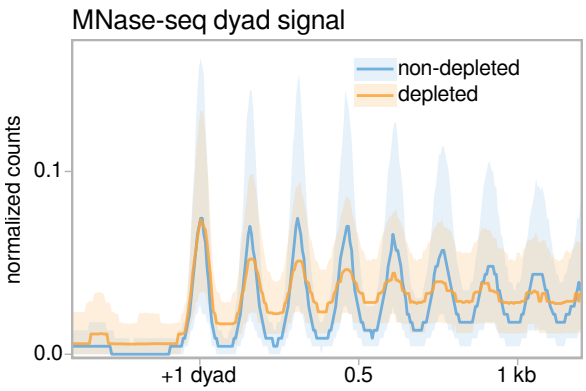

Figure 3.11: Average MNase-seq dyad signal from Spt5-depleted and non-depleted cells, over 1989 non-overlapping coding genes aligned by wild-type +1 nucleosome dyad. The solid line and shading are the median and inter-quartile range of the mean library-size normalized coverage over two (non-depleted) or three (depleted) replicates.

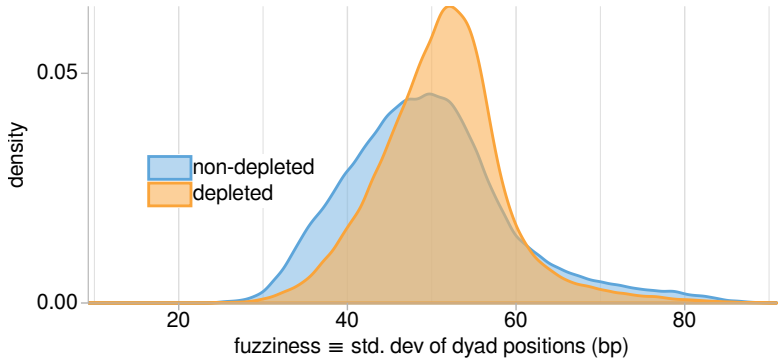

Figure 3.12: Distributions of nucleosome fuzziness in Spt5-depleted and non-depleted cells, quantified by DANPOS2 (Chen et al., 2013).

The data also indicate that Spt5-repressed antisense TSSs generally occur in between the positions of nucleosome dyads, even when viewed as a group (Figure 3.13). Given the tendency of these TSSs to initiate within 500 base pairs downstream of the genic TSS, this is consistent with these TSSs occurring between the +1 and +2, +2 and +3, or +3 and +4 nucleosomes. We do not observe a systematic change in MNase-seq signal around these TSSs upon Spt5 depletion, suggesting that Spt5-repressed antisense TSSs probably do not occur as a result of obvious changes to surrounding nucleosomes. However, it is possible that the increased fuzziness in nucleosome positions upon Spt5 depletion contributes to antisense initiation by creating a chromatin environment favorable for transcription initiation in a subset of the population. We are also unable to rule out a wholesale decrease in nucleosome occupancy after Spt5 depletion, again due to the lack of a spike-in control.

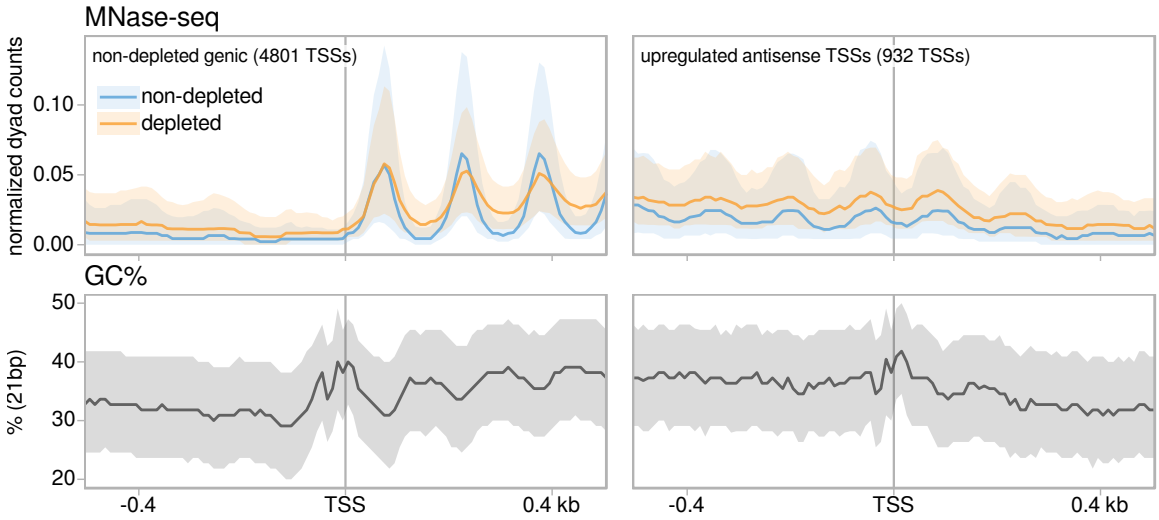

Figure 3.13: Average MNase-seq dyad signal and GC content of DNA for Spt5-depletion-induced antisense TSSs, as well as all genic TSSs detected in non-depleted cells. The solid line and shading are the median and inter-quartile range of the mean library-size normalized dyad coverage, over two (non-depleted) or three (depleted) replicates, in non-overlapping 10 bp bins.

3.6 Discussion

In this work, we integrated multiple quantitative genomic approaches to study the conserved transcription elongation factor Spt5. Our NET-seq and Pol II ChIP-seq results show that, upon Spt5 depletion, Pol II becomes 'stuck' genome-wide at the 5′ ends of genes, consistent with the role of Spt5 in stabilizing and enhancing the processivity of the elongation complex. By TSS-seq and RNA-seq, we see that Spt5 depletion causes mild decreases in steady state RNA signal over gene bodies, but not near the TSS. This is consistent with a model in which a decrease in elongation complex processivity upon Spt5 depletion causes increased pausing of the elongation complex, early termination, and the use of intragenic polyadenylation signals. Our transcriptomic assays also unexpectedly revealed that Spt5 depletion leads to the low-level expression of hundreds of antisense transcripts, primarily initiating within the first 500 base pairs downstream of genic TSSs. To determine if the expression of these antisense transcripts is due to changes in chromatin structure, we performed MNase-seq on Spt5-depleted and non-depleted cells, finding that the antisense transcripts initiate from regions that are already between nucleosomes in non-depleted cells. The full mechanism of how Spt5 normally represses these transcripts remains to be determined, perhaps involving histone modifications or factors recruited to the elongation complex by Spt5.

3.7 Methods

3.7.1 Yeast strain construction and growth conditions

S. pombe strain construction methods are detailed in Shetty et al. (2017). Spt5 depletion was carried out as follows: Cells were grown in EMM at 30°C to a density of

approximately 1×10^7 cells/ml ($OD_{600} \sim 0.5$), at which point thiamine hydrochloride and napthaleneacetic acid (NAA) were added to final concentrations of 100 nM and 0.5 mM, respectively. The cultures were then incubated with shaking for 4.5 hours at 30°C.

3.7.2 Sequencing library preparation (ChIP-seq, NET-seq, RNA-seq, TSS-seq, MNase-seq)

Library preparation methods for ChIP-seq, NET-seq, and RNA-seq are detailed in Shetty et al. (2017). TSS-seq and MNase-seq libraries were prepared as described in Doris et al. (2018), except the experimental species was *S. pombe* and the spike-in species for TSS-seq was *S. cerevisiae*. No spike-in was included in the MNase-seq libraries.

3.7.3 Genome builds

The genome build used for *S. pombe* was ASM294v2 (Wood et al., 2002), and the genome build used for *S. cerevisiae* was R64-2-1 (Engel et al., 2014).

3.7.4 NET-seq data analysis

NET-seq data analysis was performed as described in section 2.7.9, except PCR duplicates were removed using a random hexamer molecular barcode present in the adapter, and spike-in normalization was performed by normalizing to the total number of uniquely mapping, non-duplicate *S. cerevisiae* alignments.

3.7.5 RNA-seq data analysis

RNA-seq data analysis was performed using the Snakemake workflow for NET- and RNA-seq analysis described in section 2.7.9, with the sequences of the ERCC92 synthetic spike-in mix (Thermo Fisher Scientific) as the spike-in genome. No PCR duplicate removal was performed because no molecular barcode was included in the adapter.

3.7.6 ChIP-seq data analysis

An up-to-date version of the Snakemake (Köster and Rahmann, 2012) workflow used to demultiplex single-end ChIP-seq libraries is maintained at github.com/winston-lab/demultiplex-single-end. At the time of writing, demultiplexing was performed using fastq-multx (Aronesty, 2013), allowing one mismatch to the barcode.

An up-to-date version of the Snakemake (Köster and Rahmann, 2012) workflow used to process ChIP-seq libraries is maintained at github.com/winston-lab/chip-seq. At the time of writing, 3′ quality trimming was performed using cutadapt (Martin, 2011). Reads were aligned to the combined *S. pombe* and *S. cerevisiae* genome using Bowtie 2 (Langmead and Salzberg, 2012), and uniquely mapping alignments were selected using SAMtools (Subgroup et al., 2009). The median fragment size estimated by MACS2 (Zhang et al., 2008) over all samples was used to generate coverage of factor protection and fragment midpoints by extending reads to the fragment size, or by shifting reads by half the fragment size, respectively. Input and spike-in normalization was carried out as described below. Quality statistics of raw, cleaned, non-aligning, and uniquely aligning reads were assessed using FastQC (Andrews, 2010).

3.7.6.1 A note on spike-in normalization for ChIP-seq experiments with input samples

While determining how to do spike-in normalization for ChIP-seq experiments with input samples, I discovered the following error in a published spike-in normalization method. Throughout the following explanation, I use 'experimental' and 'spike-in' to refer to the two genomes present in the experiment, e.g., experimental signal and spike-in signal.

The goal when including spike-ins in a ChIP-seq experiment is to be able to normalize the experimental signal, such that the normalized signal is proportional to the absolute abundance of the factor being immunoprecipitated. A straightforward method to accomplish this normalization is to linearly scale the experimental signal of a library by a normalization factor, which we will call α. To calculate α for each library, we can use the fact that a normalized 'spike-in signal' should be the same for all libraries, since the biological state of the spike-in cells is the same in all libraries. The key to correctly determining α is defining exactly what this spike-in signal is.

The measurement we begin with for determination of the spike-in signal of a library is the number of reads in the library which map uniquely to the spike-in genome, R_{spike}. This value will vary based on two factors: the sequencing depth of the library, and the proportion of cells which were spike-in cells, ϕ:

$R_{\text{spike}} \equiv$ the number of reads in the library mapping uniquely to the spike-in genome;

$\phi \equiv$ the proportion of spike-in cells in the sample.

As the derivation of α is more easily understood in terms of absolute cell numbers

rather than ϕ, we will also define the following variables:

$$C_{\text{exp}} \equiv \text{the number of experimental cells used to prepare a library;}$$

$$C_{\text{spike}} \equiv \text{the number of spike-in cells used to prepare a library.}$$

We can express the **number of spike-in reads per spike-in cell** by simply taking the fraction $\frac{R_{\text{spike}}}{C_{\text{spike}}}$. We know that the biological state of a spike-in cell is the same regardless of which sample it belongs to, so $\frac{R_{\text{spike}}}{C_{\text{spike}}}$ is a good candidate for the 'spike-in signal' with which to calculate α. However, this expression does not account for differences in ϕ between samples: We want two libraries representing the same condition and sequenced to the same depth to have equivalent values of spike-in signal, but this does not hold true for $\frac{R_{\text{spike}}}{C_{\text{spike}}}$ if the two libraries differed in the proportion of spike-in added.

The expression for 'spike-in signal' that leads to the correct expression for α is the **number of spike-in reads per spike-in cell *per experimental cell***:

$$\frac{\frac{R_{\text{spike}}}{C_{\text{spike}}}}{C_{\text{exp}}} = \frac{R_{\text{spike}} C_{\text{exp}}}{C_{\text{spike}}}.$$

From here, it's simple to calculate α by setting this value to be equal for all samples. Since the actual value of the spike-in signal doesn't matter as long as it is equal for all libraries, we can arbitrarily set it to 1 for convenience:

$$\alpha \frac{R_{\text{spike}} C_{\text{exp}}}{C_{\text{spike}}} = 1$$

$$\alpha = \frac{C_{\text{spike}}}{R_{\text{spike}} C_{\text{exp}}}.$$

Notice that only the ratio of spike-in to experimental cells is needed to calculate α, and not the absolute number of spike-in and experimental cells. We can rewrite this expression in terms of ϕ, the proportion of the sample that was spike-in cells:

$$\phi = \frac{C_{\text{spike}}}{C_{\text{spike}} + C_{\text{exp}}}$$

$$C_{\text{spike}} = \phi \left(C_{\text{spike}} + C_{\text{exp}} \right)$$

$$C_{\text{spike}} \left(1 - \phi \right) = \phi C_{\text{exp}}$$

$$\frac{C_{\text{spike}}}{C_{\text{exp}}} = \frac{\phi}{1 - \phi}$$

$$\alpha = \frac{C_{\text{spike}}}{R_{\text{spike}} C_{\text{exp}}}$$

$$\alpha = \frac{\phi}{R_{\text{spike}} \left(1 - \phi \right)}.$$

This form for α differs from the one presented in Orlando et al. (2014) with no derivation:

$$\alpha = \frac{\phi}{R_{\text{spike}} \left(1 - \phi \right)} \qquad \alpha_{\text{orlando}} = \frac{\phi}{R_{\text{spike}}}.$$

Working through a few examples with both versions of α reveals that using α_{orlando} leads to incorrect normalization when ϕ is not equivalent for all samples.

In the first example, we will vary sequencing depth between two libraries, keeping everything else constant. Consider a single ChIP library prep in which 20% of the cells were spike-in cells (i.e., $\phi = 0.2$). The library is then unevenly split into two aliquots and sequenced. Library two has four times the reads of library one.

$$R_{\text{spike}_1} = 1 \qquad R_{\text{spike}_2} = 4$$

$$R_{\text{exp}_1} = 4 \qquad R_{\text{exp}_2} = 16$$

$$\alpha_1 = \frac{\phi}{R_{\text{spike}_1}(1-\phi)} \quad \alpha_2 = \frac{\phi}{R_{\text{spike}_2}(1-\phi)} \quad \alpha_{\text{orlando}_1} = \frac{\phi}{R_{\text{spike}_1}} \quad \alpha_{\text{orlando}_2} = \frac{\phi}{R_{\text{spike}_2}}$$
$$\alpha_1 = \frac{0.2}{1(0.8)} \quad \alpha_2 = \frac{0.2}{4(0.8)} \quad \alpha_{\text{orlando}_1} = \frac{0.2}{1} \quad \alpha_{\text{orlando}_2} = \frac{0.2}{4}$$
$$\alpha_1 = \frac{4}{16} \quad \alpha_2 = \frac{1}{16} \quad \alpha_{\text{orlando}_1} = \frac{4}{20} \quad \alpha_{\text{orlando}_2} = \frac{1}{20}.$$

The total levels of spike-in normalized experimental signal can be found for each library by multiplying α by R_{exp}, for our version of α,

$$\text{signal}_1 = \alpha_1 R_{\text{exp}_1} \qquad \text{signal}_2 = \alpha_2 R_{\text{exp}_2}$$
$$\text{signal}_1 = \frac{4}{16}(4) \qquad \text{signal}_2 = \frac{1}{16}(16)$$
$$\text{signal}_1 = 1 \qquad \text{signal}_2 = 1$$

and for α_{orlando}:

$$\text{signal}_{\text{orlando}_1} = \alpha_{\text{orlando}_1} R_{\text{exp}_1} \qquad \text{signal}_{\text{orlando}_2} = \alpha_{\text{orlando}_2} R_{\text{exp}_2}$$
$$\text{signal}_{\text{orlando}_1} = \frac{4}{20}(4) \qquad \text{signal}_{\text{orlando}_2} = \frac{1}{20}(16)$$
$$\text{signal}_{\text{orlando}_1} = 0.8 \qquad \text{signal}_{\text{orlando}_2} = 0.8.$$

Only the relative abundances within normalization methods matter, so in this case both calculations correctly normalize for library size and say that the normalized signal in the two libraries are the same.

Now consider two libraries from two different conditions with $\phi = 0.1$. In condition 2, a global decrease in experimental signal is expected. This time, we will skip the

algebra:

$$R_{\text{spike}_1} = 1 \qquad R_{\text{spike}_2} = 4$$

$$R_{\text{exp}_1} = 9 \qquad R_{\text{exp}_2} = 6$$

$$\alpha_1 = \frac{4}{36} \qquad \alpha_2 = \frac{1}{36} \qquad \alpha_{\text{orlando}_1} = \frac{4}{40} \qquad \alpha_{\text{orlando}_2} = \frac{1}{40}$$

$$\text{signal}_1 = 1 \quad \text{signal}_2 = 1/6 \quad \text{signal}_{\text{orlando}_1} = 0.9 \quad \text{signal}_{\text{orlando}_2} = 0.15$$

Both methods correctly detect that experimental signal levels in library two are 1/6th that of library one.

Finally, consider two libraries from the same condition which were spiked in with different amounts of spike-in cells. Both libraries are sequenced to the same depth. Since the libraries are from the same condition, we expect their total experimental signal to be the same after normalization, even though they had different amounts of spike-in added.

$$\phi_1 = 0.2 \qquad \phi_2 = 0.4$$

$$R_{\text{spike}_1} = 2 \qquad R_{\text{spike}_2} = 4$$

$$R_{\text{exp}_1} = 8 \qquad R_{\text{exp}_2} = 6$$

$$\alpha_1 = \frac{\phi_1}{R_{\text{spike}_1}(1-\phi_1)} \quad \alpha_2 = \frac{\phi_2}{R_{\text{spike}_2}(1-\phi_2)} \quad \alpha_{\text{orlando}_1} = \frac{\phi_1}{R_{\text{spike}_1}} \quad \alpha_{\text{orlando}_2} = \frac{\phi_2}{R_{\text{spike}_2}}$$

$$\alpha_1 = \frac{0.2}{2(0.8)} \quad \alpha_2 = \frac{0.4}{4(0.6)} \quad \alpha_{\text{orlando}_1} = \frac{0.2}{2} \quad \alpha_{\text{orlando}_2} = \frac{0.4}{4}$$

$$\alpha_1 = \frac{3}{24} \quad \alpha_2 = \frac{4}{24} \quad \alpha_{\text{orlando}_1} = \frac{1}{10} \quad \alpha_{\text{orlando}_2} = \frac{1}{10}$$

$$\text{signal}_1 = \alpha_1 R_{\text{exp}_1} \qquad \text{signal}_2 = \alpha_2 R_{\text{exp}_2}$$

$$\text{signal}_1 = \frac{3}{24}(8) \qquad \text{signal}_2 = \frac{4}{24}(6)$$

$$\text{signal}_1 = 1 \qquad \text{signal}_2 = 1$$

$$\text{signal}_{\text{orlando}_1} = \alpha_{\text{orlando}_1} R_{\text{exp}_1} \qquad \text{signal}_{\text{orlando}_2} = \alpha_{\text{orlando}_2} R_{\text{exp}_2}$$

$$\text{signal}_{\text{orlando}_1} = \frac{1}{10}(8) \qquad \text{signal}_{\text{orlando}_2} = \frac{1}{10}(6)$$

$$\text{signal}_{\text{orlando}_1} = 0.8 \qquad \text{signal}_{\text{orlando}_2} = 0.6$$

Here, our method correctly normalizes the two samples to the same total experimental signal while using the Orlando α results in an apparent decrease in signal in library two. This is because the Orlando α fails to account for the fact that when you add more spike-in to a sample, you necessarily decrease the proportion of the sample that is experimental. In most experiments with spike-ins, this isn't an issue because we assume that ϕ is the same for all samples. However, for ChIP-seq experiments that include input samples, if we assume that the experimental and spike-in input sample read counts are proportional to the amounts of experimental and spike-in cells mixed, we can plug these values in for values of ϕ to get a more reliable estimation of experimental signal levels. In this case, it becomes important to use the correct equation for α.

So, putting everything together, here's how I use a spike-in control to normalize an IP ChIP-seq library paired with an input ChIP-seq library.

As stated above, we assume that the experimental and spike-in read counts in the input sample are proportional to the numbers of experimental and spike-in cells used to prepare the library:

$$R_{\text{input}_{\text{exp}}} \propto C_{\text{exp}},$$

$$R_{\text{input}_{\text{spike}}} \propto C_{\text{spike}}.$$

Therefore, we can plug these values in for C for both the input and IP libraries (using the form of α without ϕ):

$$\alpha_{\text{input}} = \frac{C_{\text{input}_{\text{spike}}}}{R_{\text{input}_{\text{spike}}} C_{\text{input}_{\text{exp}}}} \qquad \alpha_{\text{IP}} = \frac{C_{\text{input}_{\text{spike}}}}{R_{\text{IP}_{\text{spike}}} C_{\text{input}_{\text{exp}}}}$$

$$\alpha_{\text{input}} \propto \frac{R_{\text{input}_{\text{spike}}}}{R_{\text{input}_{\text{spike}}} R_{\text{input}_{\text{exp}}}} \qquad \alpha_{\text{IP}} \propto \frac{R_{\text{input}_{\text{spike}}}}{R_{\text{IP}_{\text{spike}}} R_{\text{input}_{\text{exp}}}}$$

$$\alpha_{\text{input}} \propto \frac{1}{R_{\text{input}_{\text{exp}}}}$$

Notice how α_{input} reduces down to normalizing by the experimental library size, with no dependence on the spike-in. This makes sense because the input always represents the same state, regardless of how much spike-in is added to it. The function of the spike-in in the input is only to allow us to estimate abundances in the corresponding IP library. Rewriting α_{IP} in the form

$$\alpha_{\text{IP}} \propto \frac{1}{R_{\text{IP}_{\text{spike}}} \frac{R_{\text{input}_{\text{exp}}}}{R_{\text{input}_{\text{spike}}}}}$$

shows that α_{IP} will basically scale the experimental IP signal to the same scale as the experimental input signal, using the spike-in as a link between the two samples.

This makes it natural to subtract the normalized input signal from the normalized IP signal: since they are on the same scale, the resulting coverage can be interpreted as reporting how much more IP signal was detected than was expected based on the input.

3.7.7 TSS-seq data analysis

TSS-seq data analysis was performed as described in section 2.7.5, except the experimental genome was *S. pombe* and the spike-in genome was *S. cerevisiae*.

Reannotation of the 5′ ends of transcripts was performed as described in section 2.7.5.1, using transcript and ORF annotations from PomBase (Lock et al., 2018), and four replicates of Spt5 non-depleted TSS-seq data.

3.7.8 MNase-seq data analysis

MNase-seq data analysis was performed as described in section 2.7.8, except the *S. pombe* genome was used and no spike-in normalization was performed because no spike-in was included in the experiment.

3.8 Bibliography

Adelman, K. and Lis, J. T. (2012). Promoter-proximal pausing of rna polymerase ii: emerging roles in metazoans. *Nature Reviews Genetics*, 13:720 EP –. Review Article. 3.2

Andrews, S. (2010). Fastqc: A quality control tool for high throughput sequence data. 3.7.6

Aronesty, E. (2013). Comparison of sequencing utility programs. *The Open Bioinformatics Journal*, 7:1–8. 3.7.6

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The meme suite. *Nucleic Acids Research*, 43(W1):W39–W49. 3.10

Blythe, A. J., Yazar-Klosinski, B., Webster, M. W., Chen, E., Vandevenne, M., Bendak, K., Mackay, J. P., Hartzog, G. A., and Vrielink, A. (2016). The yeast transcription elongation factor spt4/5 is a sequence-specific rna binding protein. *Protein Science*, 25(9):1710–1721. 3.2

Chen, K., Hu, Z., Xia, Z., Zhao, D., Li, W., and Tyler, J. K. (2016). The overlooked fact: Fundamental need for spike-in control for virtually all genome-wide analyses. *Molecular and Cellular Biology*, 36(5):662–667. 3.2

Chen, K., Xi, Y., Pan, X., Li, Z., Kaestner, K., Tyler, J., Dent, S., He, X., and Li, W. (2013). Danpos: Dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Research*, 23(2):341–351. 3.12

Crickard, J. B., Fu, J., and Reese, J. C. (2016). Biochemical analysis of yeast suppressor of ty 4/5 (spt4/5) reveals the importance of nucleic acid interactions in the prevention of rna polymerase ii arrest. *Journal of Biological Chemistry*, 291(19):9853–9870. 3.2

Cui, Y. and Denis, C. L. (2003). In vivo evidence that defects in the transcriptional elongation factors rpb2, tfiis, and spt5 enhance upstream poly(a) site utilization. *Molecular and Cellular Biology*, 23(21):7887–7901. 3.4

Czudnochowski, N., Bösken, C. A., and Geyer, M. (2012). Serine-7 but not serine-5 phosphorylation primes rna polymerase ii ctd for p-tefb recognition. *Nature Communications*, 3:842 EP –. Article. 3.3

Diamant, G., Bahat, A., and Dikstein, R. (2016a). The elongation factor spt5 facilitates transcription initiation for rapid induction of inflammatory-response genes. *Nature Communications*, 7:11547 EP –. Article. 3.2, 3.3

Diamant, G., Eisenbaum, T., Leshkowitz, D., and Dikstein, R. (2016b). Analysis of subcellular rna fractions revealed a transcription-independent effect of tumor necrosis factor alpha on splicing, mediated by spt5. *Molecular and Cellular Biology*, 36(9):1342–1353. 3.2

Doamekpor, S. K., Sanchez, A. M., Schwer, B., Shuman, S., and Lima, C. D. (2014). How an mrna capping enzyme reads distinct rna polymerase ii and spt5 ctd phosphorylation codes. *Genes & Development*, 28(12):1323–1336. 3.2

Doamekpor, S. K., Schwer, B., Sanchez, A. M., Shuman, S., and Lima, C. D. (2015). Fission yeast rna triphosphatase reads an spt5 ctd code. *RNA*, 21(1):113–123. 3.2

Doris, S. M., Chuang, J., Viktorovskaya, O., Murawska, M., Spatt, D., Churchman, L. S., and Winston, F. (2018). Spt6 is required for the fidelity of promoter selection. *bioRxiv*. 3.7.2

Drouin, S., Laramée, L., Jacques, P.-�., Forest, A., Bergeron, M., and Robert, F. (2010). Dsif and rna polymerase ii ctd phosphorylation coordinate the recruitment of rpd3s to actively transcribed genes. *PLOS Genetics*, 6(10):1–12. 3.2

Engel, S. R., Dietrich, F. S., Fisk, D. G., Binkley, G., Balakrishnan, R., Costanzo, M. C., Dwight, S. S., Hitz, B. C., Karra, K., Nash, R. S., Weng, S., Wong, E. D., Lloyd, P., Skrzypek, M. S., Miyasato, S. R., Simison, M., and Cherry, J. M. (2014). The reference genome sequence of saccharomyces cerevisiae: Then and now. *G3: Genes, Genomes, Genetics*, 4(3):389–398. 3.7.3

Guo, S., Yamaguchi, Y., Schilbach, S., Wada, T., Lee, J., Goddard, A., French, D., Handa, H., and Rosenthal, A. (2000). A regulator of transcriptional elongation controls vertebrate neuronal development. *Nature*, 408(6810):366–369. 3.2

Hartzog, G. A. and Fu, J. (2013). The spt4–spt5 complex: A multi-faceted regulator of transcription elongation. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1829(1):105 – 115. RNA polymerase II Transcript Elongation. 3.2

Hartzog, G. A., Wada, T., Handa, H., and Winston, F. (1998). Evidence that spt4, spt5, and spt6 control transcription elongation by rna polymerase ii insaccharomyces cerevisiae. *Genes & Development*, 12(3):357–369. 3.2

Hirtreiter, A., Damsma, G. E., Cheung, A. C. M., Klose, D., Grohmann, D., Vojnic, E., Martin, A. C. R., Cramer, P., and Werner, F. (2010). Spt4/5 stimulates transcription elongation through the rna polymerase clamp coiled-coil motif. *Nucleic Acids Research*, 38(12):4040–4051. 3.2

Kanke, M., Nishimura, K., Kanemaki, M., Kakimoto, T., Takahashi, T. S., Nakagawa, T., and Masukata, H. (2011). Auxin-inducible protein depletion system in fission yeast. *BMC Cell Biology*, 12(1):8. 3.2

Klein, B. J., Bose, D., Baker, K. J., Yusoff, Z. M., Zhang, X., and Murakami, K. S. (2011). Rna polymerase and transcription elongation factor spt4/5 complex structure. *Proceedings of the National Academy of Sciences*, 108(2):546–550. 3.2

Komarnitsky, P., Cho, E.-J., and Buratowski, S. (2000). Different phosphorylated forms of rna polymerase ii and associated mrna processing factors during transcription. *Genes & Development*, 14(19):2452–2460. 3.3

Komori, T., Inukai, N., Yamada, T., Yamaguchi, Y., and Handa, H. (2009). Role of human transcription elongation factor dsif in the suppression of senescence and apoptosis. *Genes to Cells*, 14(3):343–354. 3.2

Kramer, N. J., Carlomagno, Y., Zhang, Y.-J., Almeida, S., Cook, C. N., Gendron, T. F., Prudencio, M., Van Blitterswijk, M., Belzil, V., Couthouis, J., Paul, J. W., Goodman, L. D., Daughrity, L., Chew, J., Garrett, A., Pregent, L., Jansen-West, K., Tabassian, L. J., Rademakers, R., Boylan, K., Graff-Radford, N. R., Josephs, K. A., Parisi, J. E., Knopman, D. S., Petersen, R. C., Boeve, B. F., Deng, N., Feng, Y., Cheng, T.-H., Dickson, D. W., Cohen, S. N., Bonini, N. M., Link, C. D., Gao, F.-B., Petrucelli, L., and Gitler, A. D. (2016). Spt4 selectively regulates the expression of c9orf72 sense and antisense mutant transcripts. *Science*, 353(6300):708–712. 3.2, 3.3

Krishnan, K., Salomonis, N., and Guo, S. (2008). Identification of spt5 target genes in zebrafish development reveals its dual activity in vivo. *PLOS ONE*, 3(11):1–13. 3.2

Köster, J. and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522. 3.7.6

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9:357 EP –. 3.7.6

Liu, C.-R., Chang, C.-R., Chern, Y., Wang, T.-H., Hsieh, W.-C., Shen, W.-C., Chang, C.-Y., Chu, I.-C., Deng, N., Cohen, S., and Cheng, T.-H. (2012). Spt4 is selectively required for transcription of extended trinucleotide repeats. *Cell*, 148(4):690–701. 3.2, 3.3

Liu, Y., Warfield, L., Zhang, C., Luo, J., Allen, J., Lang, W. H., Ranish, J., Shokat, K. M., and Hahn, S. (2009). Phosphorylation of the transcription elongation factor spt5 by yeast bur1 kinase stimulates recruitment of the paf complex. *Molecular and Cellular Biology*, 29(17):4852–4863. 3.2

Lock, A., Rutherford, K., Harris, M. A., Hayles, J., Oliver, S. G., Bähler, J., and Wood, V. (2018). PomBase 2018: user-driven reimplementation of the fission yeast database provides rapid and intuitive access to diverse, interconnected information. *Nucleic Acids Research*, 47(D1):D821–D827. 3.7.7

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12. 3.7.6

Martinez-Rucobo, F. W., Sainsbury, S., Cheung, A. C., and Cramer, P. (2011). Architecture of the rna polymerase–spt4/5 complex and basis of universal transcription processivity. *The EMBO Journal*, 30(7):1302–1310. 3.2

Mason, P. B. and Struhl, K. (2005). Distinction and relationship between elongation rate and processivity of rna polymerase ii in vivo. *Molecular Cell*, 17(6):831 – 840. 3.2, 3.3

Mayer, A., Schreieck, A., Lidschreiber, M., Leike, K., Martin, D. E., and Cramer, P. (2012). The spt5 c-terminal region recruits yeast 3' rna cleavage factor i. *Molecular and Cellular Biology*, 32(7):1321–1331. 3.2

Mbogning, J., Nagy, S., Pagé, V., Schwer, B., Shuman, S., Fisher, R. P., and Tanny, J. C. (2013). The paf complex and prf1/rtf1 delineate distinct cdk9-dependent pathways regulating transcription elongation in fission yeast. *PLOS Genetics*, 9(12):1–14. 3.2

Meyer, P. A., Li, S., Zhang, M., Yamada, K., Takagi, Y., Hartzog, G. A., and Fu, J. (2015). Structures and functions of the multiple kow domains of transcription elongation factor spt5. *Molecular and Cellular Biology*, 35(19):3354–3369. 3.2

Morillon, A., Karabetsou, N., O'Sullivan, J., Kent, N., Proudfoot, N., and Mellor, J. (2003). Isw1 chromatin remodeling atpase coordinates transcription elongation and termination by rna polymerase ii. *Cell*, 115(4):425 – 435. 3.2, 3.3

Orlando, D., Chen, M., Brown, V., Solanki, S., Choi, Y., Olson, E., Fritz, C., Bradner, J., and Guenther, M. (2014). Quantitative chip-seq normalization reveals global modulation of the epigenome. *Cell Reports*, 9(3):1163 – 1170. 3.7.6.1

Quan, T. K. and Hartzog, G. A. (2010). Histone h3k4 and k36 methylation, chd1 and rpd3s oppose the functions of saccharomyces cerevisiae spt4–spt5 in transcription. *Genetics*, 184(2):321–334. 3.2, 3.3

Rondón, A. G., García-Rubio, M., González-Barrera, S., and Aguilera, A. (2003). Molecular evidence for a positive role of spt4 in transcription elongation. *The EMBO Journal*, 22(3):612–620. 3.2, 3.3

Schneider, S., Pei, Y., Shuman, S., and Schwer, B. (2010). Separable functions of the fission yeast spt5 carboxyl-terminal domain (ctd) in capping enzyme binding and transcription elongation overlap with those of the rna polymerase ii ctd. *Molecular and Cellular Biology*, 30(10):2353–2364. 3.2

Schwer, B., Schneider, S., Pei, Y., Aronova, A., and Shuman, S. (2009). Characterization of the schizosaccharomyces pombe spt5-spt4 complex. *RNA*, 15(7):1241–1250. 3.2

Shetty, A., Kallgren, S. P., Demel, C., Maier, K. C., Spatt, D., Alver, B. H., Cramer, P., Park, P. J., and Winston, F. (2017). Spt5 plays vital roles in the control of sense and antisense transcription elongation. *Molecular Cell*, 66(1):77 – 88.e5. 3.2, 3.7.1, 3.7.2

Stadelmayer, B., Micas, G., Gamot, A., Martin, P., Malirat, N., Koval, S., Raffel, R., Sobhian, B., Severac, D., Rialle, S., Parrinello, H., Cuvier, O., and Benkirane, M. (2014). Integrator complex regulates nelf-mediated rna polymerase ii pause/release and processivity at coding genes. *Nature Communications*, 5:5531 EP –. Article. 3.2

Stanlie, A., Begum, N. A., Akiyama, H., and Honjo, T. (2012). The dsif subunits spt4 and spt5 have distinct roles at various phases of immunoglobulin class switch recombination. *PLOS Genetics*, 8(4):1–11. 3.2

Subgroup, . G. P. D. P., Wysoker, A., Handsaker, B., Marth, G., Abecasis, G., Li, H., Ruan, J., Homer, N., Durbin, R., and Fennell, T. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079. 3.7.6

Viktorovskaya, O. V., Appling, F. D., and Schneider, D. A. (2011). Yeast transcription elongation factor spt5 associates with rna polymerase i and rna polymerase ii directly. *Journal of Biological Chemistry*. 3.2

Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, A., Imai, T., Hirose, S., Sugimoto, S., Yano, K., Hartzog, G. A., Winston, F., Buratowski, S., and Handa, H. (1998). Dsif, a novel transcription elongation factor that regulates rna polymerase ii processivity, is composed of human spt4 and spt5 homologs. *Genes & Development*, 12(3):343–356. 3.2

Wen, Y. and Shatkin, A. J. (1999). Transcription elongation factor hspt5 stimulates mrna capping. *Genes & Development*, 13(14):1774–1779. 3.2

Werner, F. (2012). A nexus for gene expression—molecular mechanisms of spt5 and nusg in the three domains of life. *Journal of Molecular Biology*, 417(1):13 – 27. 3.2

Wier, A. D., Mayekar, M. K., Héroux, A., Arndt, K. M., and VanDemark, A. P. (2013). Structural basis for spt5-mediated recruitment of the paf1 complex to chromatin. 3.2

Wood, V., Gwilliam, R., Rajandream, M.-A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., Basham, D., Bowman, S., Brooks, K., Brown, D., Brown, S., Chillingworth, T., Churcher, C., Collins, M., Connor, R., Cronin, A., Davis, P., Feltwell, T., Fraser, A., Gentles, S., Goble, A., Hamlin, N., Harris, D., Hidalgo, J., Hodgson, G., Holroyd, S., Hornsby, T., Howarth, S., Huckle, E. J., Hunt, S., Jagels, K., James, K., Jones, L., Jones, M., Leather, S., McDonald, S., McLean, J., Mooney, P., Moule, S., Mungall, K., Murphy, L., Niblett, D., Odell, C., Oliver, K., O'Neil, S., Pearson, D., Quail, M. A., Rabbinowitsch, E., Rutherford, K., Rutter, S., Saunders, D., Seeger, K., Sharp, S., Skelton, J., Simmonds, M., Squares, R., Squares, S., Stevens, K., Taylor, K., Taylor, R. G., Tivey, A., Walsh, S., Warren, T., Whitehead, S., Woodward, J., Volckaert, G., Aert, R., Robben, J., Grymonprez, B., Weltjens, I., Vanstreels, E., Rieger, M., Schäfer, M., Müller-Auer, S., Gabel, C., Fuchs, M., Fritzc, C., Holzer, E., Moestl, D., Hilbert, H., Borzym, K., Langer, I., Beck, A., Lehrach, H., Reinhardt, R., Pohl, T. M., Eger, P., Zimmermann, W., Wedler, H., Wambutt, R., Purnelle, B., Goffeau, A., Cadieu, E., Dréano, S., Gloux, S., Lelaure, V., Mottier, S., Galibert, F., Aves, S. J., Xiang, Z., Hunt, C., Moore, K., Hurst, S. M., Lucas, M., Rochet, M., Gaillardin, C., Tallada, V. A., Garzon, A., Thode, G., Daga, R. R., Cruzado, L., Jimenez, J., Sánchez, M., del Rey, F., Benito, J., Domínguez, A., Revuelta, J. L., Moreno, S., Armstrong, J., Forsburg, S. L., Cerrutti, L., Lowe, T., McCombie, W. R., Paulsen, I., Potashkin, J., Shpakovski, G. V., Ussery, D., Barrell, B. G., and Nurse, P. (2002). The genome sequence of schizosaccharomyces pombe. *Nature*, 415(6874):871–880. 3.7.3

Yamaguchi, Y., Wada, T., Watanabe, D., Takagi, T., Hasegawa, J., and Handa, H. (1999). Structure and function of the human transcription elongation factor dsif. *Journal of Biological Chemistry*, 274(12):8085–8092. 3.2

Yamamoto, J., Hagiwara, Y., Chiba, K., Isobe, T., Narita, T., Handa, H., and Yamaguchi, Y. (2014). Dsif and nelf interact with integrator to specify the correct post-transcriptional fate of snrna genes. *Nature Communications*, 5:4263 EP –. Article. 3.2

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.-W., Nechaev, S., Adelman, K., Levine, M., and Young, R. A. (2007). Rna polymerase stalling at developmental control genes in the drosophila melanogaster embryo. *Nature Genetics*, 39:1512 EP –. 3.3

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E.,

Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of chip-seq (macs). *Genome Biology*, 9(9):R137. 3.7.6

Zhou, K., Kuo, W. H. W., Fillingham, J., and Greenblatt, J. F. (2009). Control of transcriptional elongation and cotranscriptional histone modification by the yeast bur kinase substrate spt5. *Proceedings of the National Academy of Sciences*, 106(17):6956–6961. 3.2

Zhu, W., Wada, T., Okabe, S., Taneda, T., Yamaguchi, Y., and Handa, H. (2007). DSIF contributes to transcriptional activation by DNA-binding activators by preventing pausing during transcription elongation. *Nucleic Acids Research*, 35(12):4064–4075. 3.2

Chapter 4

Stress-responsive intragenic transcription

4.1 Collaborators

Steve Doris	generated TSS-seq and ChIP-nexus libraries
Dan Spatt	polyribosome fractionation, fitness competitions, and other experiments
James Warner	fitness competitions and other experiments

4.2 Possible functions for intragenic transcription in wild-type cells

ASE1 (McKnight et al., 2014). KAR4 (Gammie et al., 1999). ASP3 (Huang et al., 2010).

4.3 Discovery of stress-induced intragenic promoters by TFIIB ChIP-nexus and TSS-seq

To discover cases of stress-induced intragenic transcription initiation, we performed ChIP-nexus of TFIIB in wild-type yeast in conditions of oxidative stress, amino acid stress, and nitrogen stress, along with controls of growth in rich YPD medium and defined SC medium. The genic TFIIB response to each of the stresses either correlated well with the expected transcriptomic response to the stress (Figure 4.1), or was enriched for metabolic pathways consistent with the cellular response to the stress (Figure 4.2), indicating that TFIIB ChIP-nexus is able to We identified 140 intragenic TFIIB peaks significantly induced at least 1.5-fold in at least one stress condition, with some peaks being induced in more than one stress (Figure 4.3).

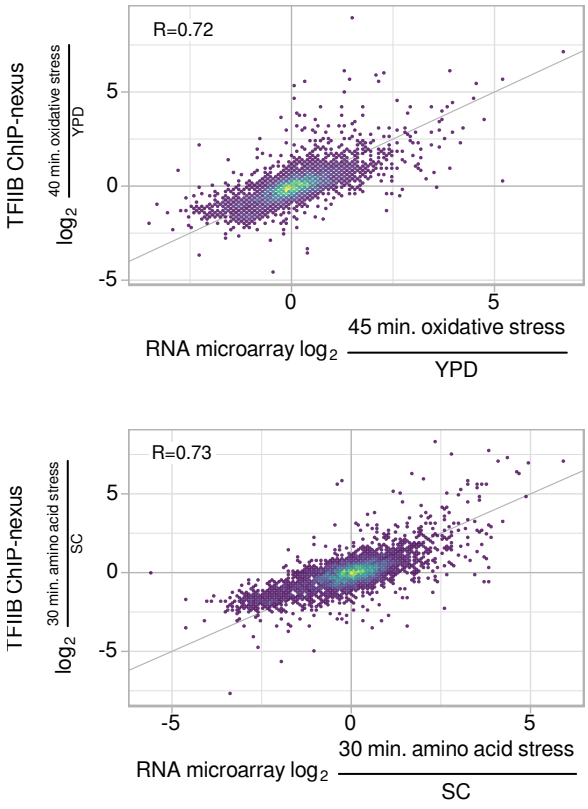

Figure 4.1: Scatterplots comparing change in genic TFIIB signal to change in RNA microarray signal from Gasch et al. (2000), for oxidative and amino acid stresses. The Pearson correlation coefficient is shown for each comparison.

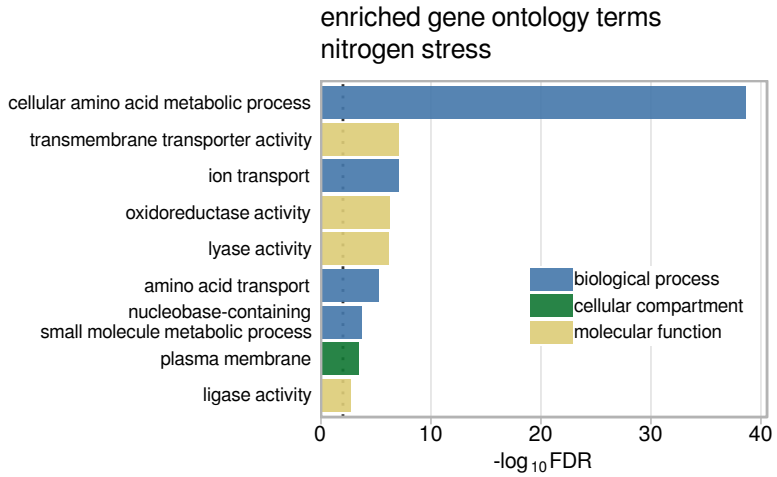

Figure 4.2: Gene ontology terms enriched in genes with significantly upregulated genic TFIIB peaks in nitrogen stress.

4.4 Chromatin landscape of oxidative-stress-induced promoters.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

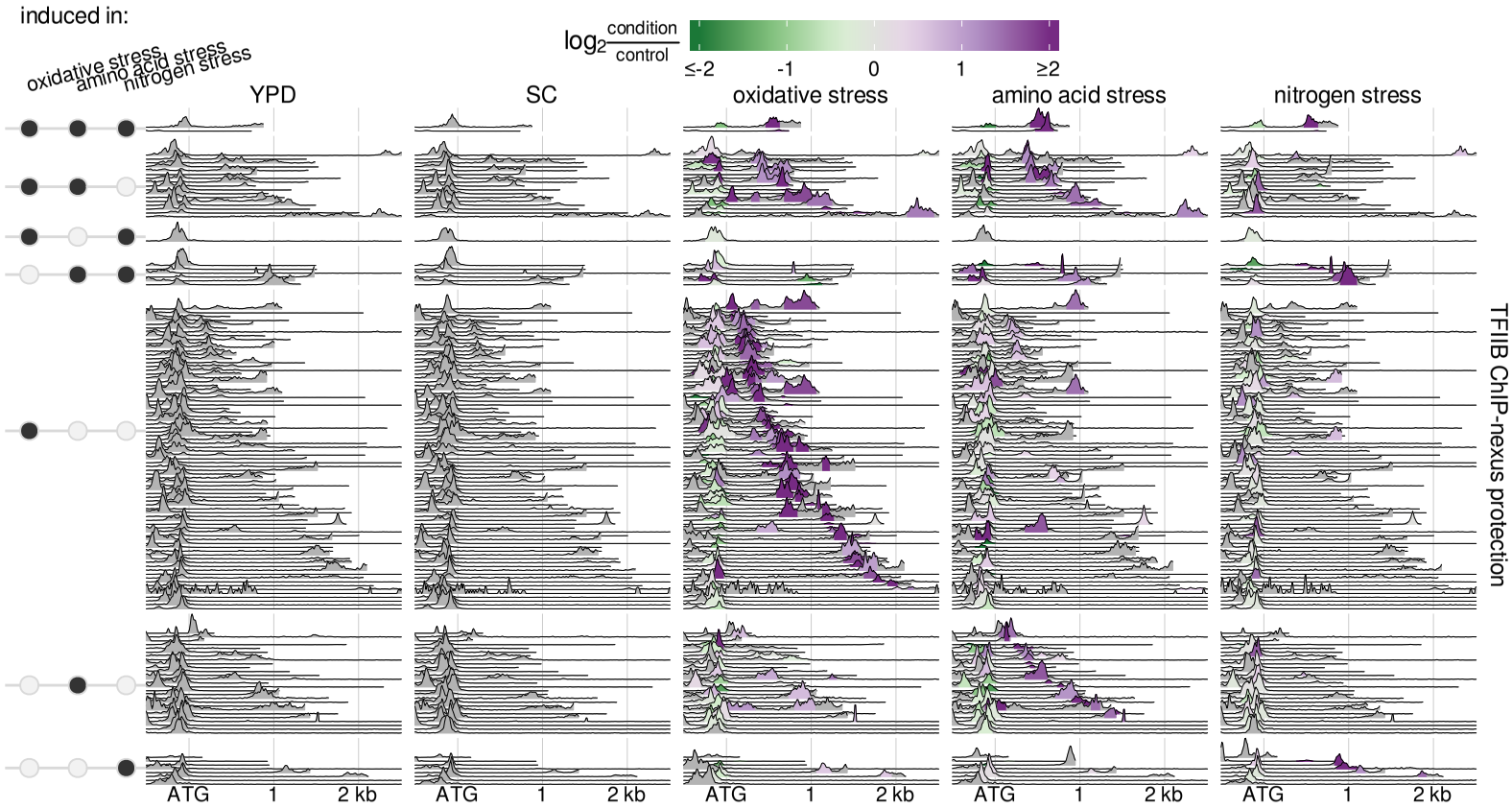

Figure 4.3: Relative TFIIB ChIP-nexus protection over all genes with an intragenic TFIIB peak significantly induced in one or more of the stress conditions tested, as depicted in the left panel. Genes are aligned by start codon, and are sorted within each group by the distance from the start codon to the summit of the induced intragenic TFIIB peak. Data are shown for each gene up to the stop codon of the gene. Regions where TFIIB peaks are called are shaded in the stress conditions according to the fold-change of the peak relative to the corresponding control condition.

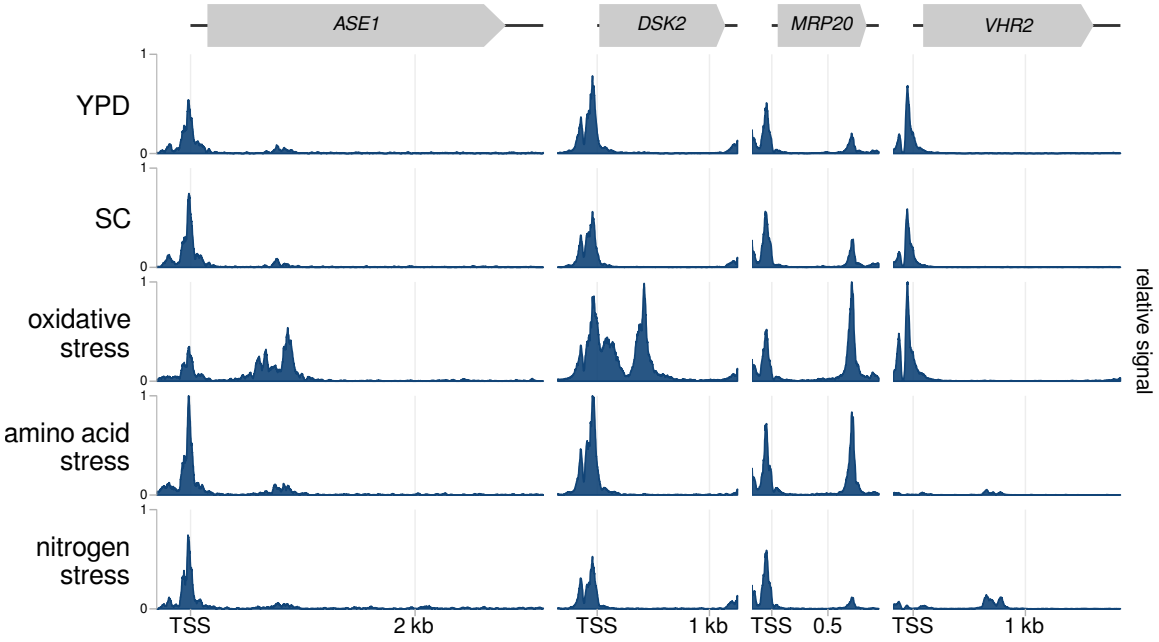

Figure 4.4: Caption asdflkj asldkfjlkj.

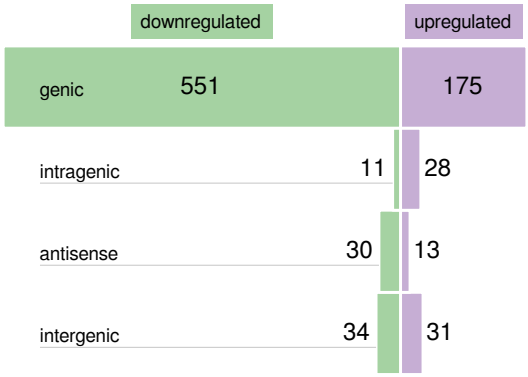

Figure 4.5: Caption dsafklj asldkfjlkj.

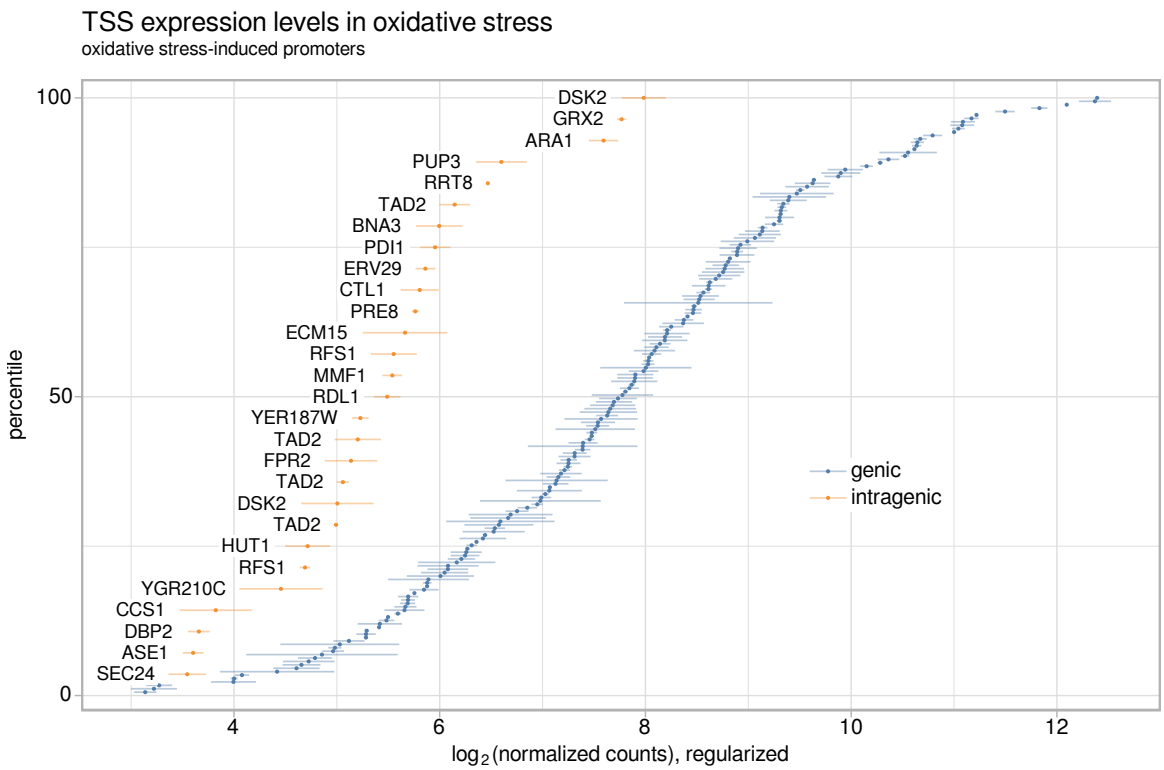

Figure 4.6: Caption dsafklj zzzz.

Figure 4.7: Caption dsafklj .

4.5 Polysome enrichment of oxidative-stress-induced intragenic transcripts

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

4.6 TSS-seq analysis of oxidative stress in *Saccharomyces sensu stricto* species

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis

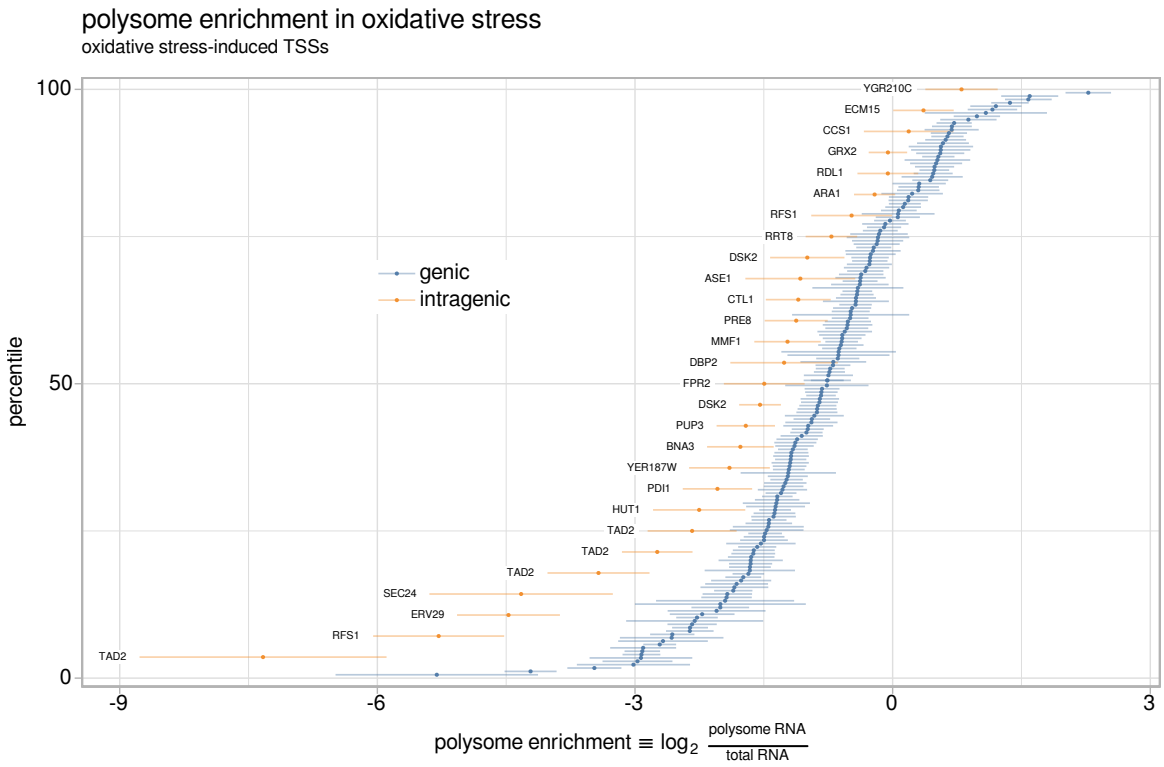

Figure 4.8: Caption wsdasdr zzzz.

Figure 4.9: Caption dsafklj .

Figure 4.10: Caption dsafklj .

nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

4.7 Functions of intragenic DSK2 expression in oxidative stress

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

4.8 Discussion

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu

Figure 4.11: Caption dsafklj .

Figure 4.12: Caption dsafklj .

libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

4.9 Methods

4.9.1 Yeast growth conditions

4.9.2 Genome builds

4.9.3 TFIIB ChIP-nexus data analysis

4.9.4 TSS-seq data analysis

4.9.5 MNase-ChIP-seq data analysis

4.9.6 Sucrose gradient fractionation

4.9.7 Polysome-associated TSS-seq analysis

4.9.8 Multiple genome alignment

4.9.9 Diamide competitive fitness assays

4.10 Bibliography

Gammie, A. E., Stewart, B. G., Scott, C. F., and Rose, M. D. (1999). The two forms of karyogamy transcription factor kar4p are regulated by differential initiation of transcription, translation, and protein turnover. *Molecular and Cellular Biology*, 19(1):817–825. 4.2

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11(12):4241–4257. PMID: 11102521. 4.1

Huang, Y.-C., Chen, H.-T., and Teng, S.-C. (2010). Intragenic transcription of a non-coding rna modulates expression of asp3 in budding yeast. *RNA*, 16(11):2085–2093. 4.2

McKnight, K., Liu, H., and Wang, Y. (2014). Replicative stress induces intragenic transcription of the ase1 gene that negatively regulates ase1 activity. *Current Biology*, 24(10):1101 – 1106. 4.2

Bibliography

Adelman, K. and Lis, J. T. (2012). Promoter-proximal pausing of rna polymerase ii: emerging roles in metazoans. *Nature Reviews Genetics*, 13:720 EP –. Review Article.

Adkins, M. W. and Tyler, J. K. (2006). Transcriptional activators are dispensable for transcription in the absence of spt6-mediated chromatin reassembly of promoter regions. *Molecular Cell*, 21(3):405 – 416.

Andrews, S. (2010). Fastqc: A quality control tool for high throughput sequence data.

Andrulis, E. D., Guzmán, E., Döring, P., Werner, J., and Lis, J. T. (2000). High-resolution localization of drosophila spt5 and spt6 at heat shock genes in vivo: roles in promoter proximal pausing and transcription elongation. *Genes & Development*, 14(20):2635–2649.

Ardehali, M. B., Yao, J., Adelman, K., Fuda, N. J., Petesch, S. J., Webb, W. W., and Lis, J. T. (2009). Spt6 enhances the elongation rate of rna polymerase ii in vivo. *The EMBO Journal*, 28(8):1067–1077.

Aronesty, E. (2013). Comparison of sequencing utility programs. *The Open Bioinformatics Journal*, 7:1–8.

Arribere, J. A. and Gilbert, W. V. (2013). Roles for transcript leaders in translation and mrna decay revealed by transcript leader sequencing. *Genome Research*, 23(6):977–987.

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The meme suite. *Nucleic Acids Research*, 43(W1):W39–W49.

Begum, N. A., Stanlie, A., Nakata, M., Akiyama, H., and Honjo, T. (2012). The histone chaperone spt6 is required for activation-induced cytidine deaminase target determination through h3k4me3 regulation. *Journal of Biological Chemistry*, 287(39):32415–32429.

Blythe, A. J., Yazar-Klosinski, B., Webster, M. W., Chen, E., Vandevenne, M., Bendak, K., Mackay, J. P., Hartzog, G. A., and Vrielink, A. (2016). The yeast transcription elongation factor spt4/5 is a sequence-specific rna binding protein. *Protein Science*, 25(9):1710–1721.

Bortvin, A. and Winston, F. (1996). Evidence that spt6p controls chromatin structure by a direct interaction with histones. *Science*, 272(5267):1473–1476.

Carrozza, M. J., Li, B., Florens, L., Suganuma, T., Swanson, S. K., Lee, K. K., Shia, W.-J., Anderson, S., Yates, J., Washburn, M. P., and Workman, J. L. (2005). Histone h3 methylation by set2 directs deacetylation of coding regions by rpd3s to suppress spurious intragenic transcription. *Cell*, 123(4):581 – 592.

Chen, K., Hu, Z., Xia, Z., Zhao, D., Li, W., and Tyler, J. K. (2016). The overlooked fact: Fundamental need for spike-in control for virtually all genome-wide analyses. *Molecular and Cellular Biology*, 36(5):662–667.

Chen, K., Xi, Y., Pan, X., Li, Z., Kaestner, K., Tyler, J., Dent, S., He, X., and Li, W. (2013). Danpos: Dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Research*, 23(2):341–351.

Chen, S., Ma, J., Wu, F., Xiong, L.-j., Ma, H., Xu, W., Lv, R., Li, X., Villen, J., Gygi, S. P., Liu, X. S., and Shi, Y. (2012). The histone h3 lys 27 demethylase jmjd3 regulates gene expression by impacting transcriptional elongation. *Genes & Development*, 26(12):1364–1375.

Cheung, V., Chua, G., Batada, N. N., Landry, C. R., Michnick, S. W., Hughes, T. R., and Winston, F. (2008). Chromatin- and transcription-related factors repress transcription from within coding regions throughout the saccharomyces cerevisiae genome. *PLOS Biology*, 6(11):1–13.

Chu, Y., Sutton, A., Sternglanz, R., and Prelich, G. (2006). The bur1 cyclin-dependent protein kinase is required for the normal pattern of histone methylation by set2. *Molecular and Cellular Biology*, 26(8):3029–3038.

Churchman, L. S. and Weissman, J. S. (2012). Native elongating transcript sequencing (net-seq). *Current Protocols in Molecular Biology*, 98(1):14.4.1–14.4.17.

Close, D., Johnson, S. J., Sdano, M. A., McDonald, S. M., Robinson, H., Formosa, T., and Hill, C. P. (2011). Crystal structures of the s. cerevisiae spt6 core and c-terminal tandem sh2 domain. *Journal of Molecular Biology*, 408(4):697 – 713.

Consortium, T. E. P., Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie,

B. R., Landt, S. G., Lee, B.-K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J. M., Song, L., Trinklein, N. D., Altshuler, R. C., Birney, E., Brown, J. B., Cheng, C., Djebali, S., Dong, X., Ernst, J., Furey, T. S., Gerstein, M., Giardine, B., Greven, M., Hardison, R. C., Harris, R. S., Herrero, J., Hoffman, M. M., Iyer, S., Kellis, M., Kheradpour, P., Lassmann, T., Li, Q., Lin, X., Marinov, G. K., Merkel, A., Mortazavi, A., Parker, S. C. J., Reddy, T. E., Rozowsky, J., Schlesinger, F., Thurman, R. E., Wang, J., Ward, L. D., Whitfield, T. W., Wilder, S. P., Wu, W., Xi, H. S., Yip, K. Y., Zhuang, J., Bernstein, B. E., Green, E. D., Gunter, C., Snyder, M., Pazin, M. J., Lowdon, R. F., Dillon, L. A. L., Adams, L. B., Kelly, C. J., Zhang, J., Wexler, J. R., Good, P. J., Feingold, E. A., Crawford, G. E., Dekker, J., Elnitski, L., Farnham, P. J., Giddings, M. C., Gingeras, T. R., Guigó, R., Hubbard, T. J., Kent, W. J., Lieb, J. D., Margulies, E. H., Myers, R. M., Stamatoyannopoulos, J. A., Tenenbaum, S. A., Weng, Z., White, K. P., Wold, B., Yu, Y., Wrobel, J., Risk, B. A., Gunawardena, H. P., Kuiper, H. C., Maier, C. W., Xie, L., Chen, X., Mikkelsen, T. S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M. J., Durham, T., Ku, M., Truong, T., Eaton, M. L., Dobin, A., Tanzer, A., Lagarde, J., Lin, W., Xue, C., Williams, B. A., Zaleski, C., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O. J., Park, E., Preall, J. B., Presaud, K., Ribeca, P., Robyr, D., Ruan, X., Sammeth, M., Sandhu, K. S., Schaeffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Hayashizaki, Y., Reymond, A., Antonarakis, S. E., Hannon, G. J., Ruan, Y., Carninci, P., Sloan, C. A., Learned, K., Malladi, V. S., Wong, M. C., Barber, G. P., Cline, M. S., Dreszer, T. R., Heitner, S. G., Karolchik, D., Kirkup, V. M., Meyer, L. R., Long, J. C., Maddren, M., Raney, B. J., Grasfeder, L. L., Giresi, P. G., Battenhouse, A., Sheffield, N. C., Showers, K. A., London, D., Bhinge, A. A., Shestak, C., Schaner, M. R., Ki Kim, S., Zhang, Z. Z., Mieczkowski, P. A., Mieczkowska, J. O., Liu, Z., McDaniell, R. M., Ni, Y., Rashid, N. U., Kim, M. J., Adar, S., Zhang, Z., Wang, T., Winter, D., Keefe, D., Iyer, V. R., Zheng, M., Wang, P., Gertz, J., Vielmetter, J., Partridge, E., Varley, K. E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K. M., Anaya, M., Cross, M. K., Muratet, M. A., Newberry, K. M., McCue, K., Nesmith, A. S., Fisher-Aylor, K. I., Pusey, B., DeSalvo, G., Parker, S. L., Balasubramanian, S., Davis, N. S., Meadows, S. K., Eggleston, T., Newberry, J. S., Levy, S. E., Absher, D. M., Wong, W. H., Blow, M. J., Visel, A., Pennachio, L. A., Petrykowska, H. M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Davidson, C., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Gonzalez, J. M., Griffiths, E., Harte, R., Hendrix,

D. A., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Leng, J., Lin, M. F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Rodriguez, J. M., Saunders, G., Sboner, A., Searle, S., Sisu, C., Snow, C., Steward, C., Tapanari, E., Tress, M. L., van Baren, M. J., Washietl, S., Wilming, L., Zadissa, A., Zhang, Z., Brent, M., Haussler, D., Valencia, A., Addleman, N., Alexander, R. P., Auerbach, R. K., Balasubramanian, S., Bettinger, K., Bhardwaj, N., Boyle, A. P., Cao, A. R., Cayting, P., Charos, A., Cheng, Y., Eastman, C., Euskirchen, G., Fleming, J. D., Grubert, F., Habegger, L., Hariharan, M., Harmanci, A., Iyengar, S., Jin, V. X., Karczewski, K. J., Kasowski, M., Lacroute, P., Lam, H., Lamarre-Vincent, N., Lian, J., Lindahl-Allen, M., Min, R., Miotto, B., Monahan, H., Moqtaderi, Z., Mu, X. J., O'Geen, H., Ouyang, Z., Patacsil, D., Raha, D., Ramirez, L., Reed, B., Shi, M., Slifer, T., Witt, H., Wu, L., Xu, X., Yan, K.-K., Yang, X., Struhl, K., Weissman, S. M., Penalva, L. O., Karmakar, S., Bhanvadia, R. R., Choudhury, A., Domanus, M., Ma, L., Moran, J., Victorsen, A., Auer, T., Centanin, L., Eichenlaub, M., Gruhl, F., Heermann, S., Hoeckendorf, B., Inoue, D., Kellner, T., Kirchmaier, S., Mueller, C., Reinhardt, R., Schertel, L., Schneider, S., Sinn, R., Wittbrodt, B., Wittbrodt, J., Partridge, E. C., Jain, G., Balasundaram, G., Bates, D. L., Byron, R., Canfield, T. K., Diegel, M. J., Dunn, D., Ebersol, A. K., Frum, T., Garg, K., Gist, E., Hansen, R. S., Boatman, L., Haugen, E., Humbert, R., Johnson, A. K., Johnson, E. M., Kutyavin, T. V., Lee, K., Lotakis, D., Maurano, M. T., Neph, S. J., Neri, F. V., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Rynes, E., Sanchez, M. E., Sandstrom, R. S., Shafer, A. O., Stergachis, A. B., Thomas, S., Vernot, B., Vierstra, J., Vong, S., Wang, H., Weaver, M. A., Yan, Y., Zhang, M., Akey, J. M., Bender, M., Dorschner, M. O., Groudine, M., MacCoss, M. J., Navas, P., Stamatoyannopoulos, G., Beal, K., Brazma, A., Flicek, P., Johnson, N., Lukk, M., Luscombe, N. M., Sobral, D., Vaquerizas, J. M., Batzoglou, S., Sidow, A., Hussami, N., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M. W., Schaub, M. A., Miller, W., Bickel, P. J., Banfai, B., Boley, N. P., Huang, H., Li, J. J., Noble, W. S., Bilmes, J. A., Buske, O. J., Sahu, A. D., Kharchenko, P. V., Park, P. J., Baker, D., Taylor, J., and Lochovsky, L. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489:57 EP –. Article.

Crickard, J. B., Fu, J., and Reese, J. C. (2016). Biochemical analysis of yeast suppressor of ty 4/5 (spt4/5) reveals the importance of nucleic acid interactions in the prevention of rna polymerase ii arrest. *Journal of Biological Chemistry*, 291(19):9853–9870.

Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). Weblogo: A sequence logo generator. *Genome Research*, 14(6):1188–1190.

Cui, Y. and Denis, C. L. (2003). In vivo evidence that defects in the transcriptional

elongation factors rpb2, tfiis, and spt5 enhance upstream poly(a) site utilization. *Molecular and Cellular Biology*, 23(21):7887–7901.

Czudnochowski, N., Bösken, C. A., and Geyer, M. (2012). Serine-7 but not serine-5 phosphorylation primes rna polymerase ii ctd for p-tefb recognition. *Nature Communications*, 3:842 EP –. Article.

de Boer, C. G. and Hughes, T. R. (2011). YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Research*, 40(D1):D169–D179.

DeGennaro, C. M., Alver, B. H., Marguerat, S., Stepanova, E., Davis, C. P., Bähler, J., Park, P. J., and Winston, F. (2013). Spt6 regulates intragenic and antisense transcription, nucleosome positioning, and histone modifications genome-wide in fission yeast. *Molecular and Cellular Biology*, 33(24):4779–4792.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319.

Diamant, G., Bahat, A., and Dikstein, R. (2016a). The elongation factor spt5 facilitates transcription initiation for rapid induction of inflammatory-response genes. *Nature Communications*, 7:11547 EP –. Article.

Diamant, G., Eisenbaum, T., Leshkowitz, D., and Dikstein, R. (2016b). Analysis of subcellular rna fractions revealed a transcription-independent effect of tumor necrosis factor alpha on splicing, mediated by spt5. *Molecular and Cellular Biology*, 36(9):1342–1353.

Diebold, M.-L., Koch, M., Loeliger, E., Cura, V., Winston, F., Cavarelli, J., and Romier, C. (2010a). The structure of an iws1/spt6 complex reveals an interaction domain conserved in tfiis, elongin a and med26. *The EMBO Journal*, 29(23):3979–3991.

Diebold, M.-L., Loeliger, E., Koch, M., Winston, F., Cavarelli, J., and Romier, C. (2010b). Noncanonical tandem sh2 enables interaction of elongation factor spt6 with rna polymerase ii. *Journal of Biological Chemistry*, 285(49):38389–38398.

Doamekpor, S. K., Sanchez, A. M., Schwer, B., Shuman, S., and Lima, C. D. (2014). How an mrna capping enzyme reads distinct rna polymerase ii and spt5 ctd phosphorylation codes. *Genes & Development*, 28(12):1323–1336.

Doamekpor, S. K., Schwer, B., Sanchez, A. M., Shuman, S., and Lima, C. D. (2015). Fission yeast rna triphosphatase reads an spt5 ctd code. *RNA*, 21(1):113–123.

Doris, S. M., Chuang, J., Viktorovskaya, O., Murawska, M., Spatt, D., Churchman, L. S., and Winston, F. (2018). Spt6 is required for the fidelity of promoter selection. *Molecular Cell*, 72(4):687 – 699.e6.

Drouin, S., Laramée, L., Jacques, P.-�., Forest, A., Bergeron, M., and Robert, F. (2010). Dsif and rna polymerase ii ctd phosphorylation coordinate the recruitment of rpd3s to actively transcribed genes. *PLOS Genetics*, 6(10):1–12.

Duina, A. A. (2011). Histone chaperones spt6 and fact: Similarities and differences in modes of action at transcribed genes. *Genet Res Int*, 2011:625210. 22567361[pmid].

Endoh, M., Zhu, W., Hasegawa, J., Watanabe, H., Kim, D.-K., Aida, M., Inukai, N., Narita, T., Yamada, T., Furuya, A., Sato, H., Yamaguchi, Y., Mandal, S. S., Reinberg, D., Wada, T., and Handa, H. (2004). Human spt6 stimulates transcription elongation by rna polymerase ii in vitro. *Molecular and Cellular Biology*, 24(8):3324–3336.

Engel, S. R., Dietrich, F. S., Fisk, D. G., Binkley, G., Balakrishnan, R., Costanzo, M. C., Dwight, S. S., Hitz, B. C., Karra, K., Nash, R. S., Weng, S., Wong, E. D., Lloyd, P., Skrzypek, M. S., Miyasato, S. R., Simison, M., and Cherry, J. M. (2014). The reference genome sequence of saccharomyces cerevisiae: Then and now. *G3: Genes, Genomes, Genetics*, 4(3):389–398.

Gammie, A. E., Stewart, B. G., Scott, C. F., and Rose, M. D. (1999). The two forms of karyogamy transcription factor kar4p are regulated by differential initiation of transcription, translation, and protein turnover. *Molecular and Cellular Biology*, 19(1):817–825.

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018.

Guo, S., Yamaguchi, Y., Schilbach, S., Wada, T., Lee, J., Goddard, A., French, D., Handa, H., and Rosenthal, A. (2000). A regulator of transcriptional elongation controls vertebrate neuronal development. *Nature*, 408(6810):366–369.

Haberle, V. and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*, 19(10):621–637.

Hansen, P., Hecht, J., Ibn-Salem, J., Menkuec, B. S., Roskosch, S., Truss, M., and Robinson, P. N. (2016). Q-nexus: a comprehensive and efficient analysis pipeline designed for chip-nexus. *BMC Genomics*, 17(1):873.

Hartzog, G. A. and Fu, J. (2013). The spt4–spt5 complex: A multi-faceted regulator of transcription elongation. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1829(1):105 – 115. RNA polymerase II Transcript Elongation.

Hartzog, G. A., Wada, T., Handa, H., and Winston, F. (1998). Evidence that spt4, spt5, and spt6 control transcription elongation by rna polymerase ii insaccharomyces cerevisiae. *Genes & Development*, 12(3):357–369.

He, Q., Johnston, J., and Zeitlinger, J. (2015). Chip-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature Biotechnology*, 33:395 EP –.

Hennig, B. P. and Fischer, T. (2013). The great repression: chromatin and cryptic transcription. *Transcription*, 4(3):97—101.

Hirtreiter, A., Damsma, G. E., Cheung, A. C. M., Klose, D., Grohmann, D., Vojnic, E., Martin, A. C. R., Cramer, P., and Werner, F. (2010). Spt4/5 stimulates transcription elongation through the rna polymerase clamp coiled-coil motif. *Nucleic Acids Research*, 38(12):4040–4051.

Ivanovska, I., Jacques, P.-�., Rando, O. J., Robert, F., and Winston, F. (2011). Control of chromatin structure by spt6: Different consequences in coding and regulatory regions. *Molecular and Cellular Biology*, 31(3):531–541.

Iyer, V. and Struhl, K. (1995). Poly(da:dt), a ubiquitous promoter element that stimulates transcription via its intrinsic dna structure. *The EMBO Journal*, 14(11):2570–2579.

Jeronimo, C., Watanabe, S., Kaplan, C., Peterson, C., and Robert, F. (2015). The histone chaperones fact and spt6 restrict h2a.z from intragenic locations. *Molecular Cell*, 58(6):1113 – 1123.

Kaikkonen, M. U. and Adelman, K. (2018). Emerging roles of non-coding rna transcription. *Trends in Biochemical Sciences*, 43(9):654–667.

Kanke, M., Nishimura, K., Kanemaki, M., Kakimoto, T., Takahashi, T. S., Nakagawa, T., and Masukata, H. (2011). Auxin-inducible protein depletion system in fission yeast. *BMC Cell Biology*, 12(1):8.

Kaplan, C. D., Laprade, L., and Winston, F. (2003). Transcription elongation factors repress transcription initiation from cryptic sites. *Science*, 301(5636):1096–1099.

Kaplan, C. D., Morris, J. R., Wu, C.-t., and Winston, F. (2000). Spt5 and spt6 are associated with active transcription and have characteristics of general elongation factors in d. melanogaster. *Genes & Development*, 14(20):2623–2634.

Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., Field, Y., LeProust, E. M., Hughes, T. R., Lieb, J. D., Widom, J., and Segal, E. (2008). The dna-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458:362 EP –.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36.

Klein, B. J., Bose, D., Baker, K. J., Yusoff, Z. M., Zhang, X., and Murakami, K. S. (2011). Rna polymerase and transcription elongation factor spt4/5 complex structure. *Proceedings of the National Academy of Sciences*, 108(2):546–550.

Komarnitsky, P., Cho, E.-J., and Buratowski, S. (2000). Different phosphorylated forms of rna polymerase ii and associated mrna processing factors during transcription. *Genes & Development*, 14(19):2452–2460.

Komori, T., Inukai, N., Yamada, T., Yamaguchi, Y., and Handa, H. (2009). Role of human transcription elongation factor dsif in the suppression of senescence and apoptosis. *Genes to Cells*, 14(3):343–354.

Kramer, N. J., Carlomagno, Y., Zhang, Y.-J., Almeida, S., Cook, C. N., Gendron, T. F., Prudencio, M., Van Blitterswijk, M., Belzil, V., Couthouis, J., Paul, J. W., Goodman, L. D., Daughrity, L., Chew, J., Garrett, A., Pregent, L., Jansen-West, K., Tabassian, L. J., Rademakers, R., Boylan, K., Graff-Radford, N. R., Josephs, K. A., Parisi, J. E., Knopman, D. S., Petersen, R. C., Boeve, B. F., Deng, N., Feng, Y., Cheng, T.-H., Dickson, D. W., Cohen, S. N., Bonini, N. M., Link, C. D., Gao, F.-B., Petrucelli, L., and Gitler, A. D. (2016). Spt4 selectively regulates the expression of c9orf72 sense and antisense mutant transcripts. *Science*, 353(6300):708–712.

Krishnan, K., Salomonis, N., and Guo, S. (2008). Identification of spt5 target genes in zebrafish development reveals its dual activity in vivo. *PLOS ONE*, 3(11):1–13.

Krogan, N. J., Kim, M., Ahn, S. H., Zhong, G., Kobor, M. S., Cagney, G., Emili, A., Shilatifard, A., Buratowski, S., and Greenblatt, J. F. (2002). Rna polymerase ii elongation factors of saccharomyces cerevisiae: a targeted proteomics approach. *Molecular and Cellular Biology*, 22(20):6979–6992.

Köster, J. and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9:357 EP –.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology*, 10(3):R25.

Li, Q., Brown, J. B., Huang, H., and Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, 5(3):1752–1779.

Li, S., Almeida, A. R., Radebaugh, C. A., Zhang, L., Chen, X., Huang, L., Thurston, A. K., Kalashnikova, A. A., Hansen, J. C., Luger, K., and Stargell, L. A. (2018). The elongation factor spn1 is a multi-functional chromatin binding protein. *Nucleic Acids Research*, 46(5):2321–2334.

Lickwar, C. R., Rao, B., Shabalin, A. A., Nobel, A. B., Strahl, B. D., and Lieb, J. D. (2009). The set2/rpd3s pathway suppresses cryptic transcription without regard to gene length or transcription frequency. *PLOS ONE*, 4(3):1–7.

Liu, C.-R., Chang, C.-R., Chern, Y., Wang, T.-H., Hsieh, W.-C., Shen, W.-C., Chang, C.-Y., Chu, I.-C., Deng, N., Cohen, S., and Cheng, T.-H. (2012). Spt4 is selectively required for transcription of extended trinucleotide repeats. *Cell*, 148(4):690–701.

Liu, J., Zhang, J., Gong, Q., Xiong, P., Huang, H., Wu, B., Lu, G., Wu, J., and Shi, Y. (2011). Solution structure of tandem sh2 domains from spt6 protein and their binding to the phosphorylated rna polymerase ii c-terminal domain. *Journal of Biological Chemistry*, 286(33):29218–29226.

Liu, Y., Warfield, L., Zhang, C., Luo, J., Allen, J., Lang, W. H., Ranish, J., Shokat, K. M., and Hahn, S. (2009). Phosphorylation of the transcription elongation factor spt5 by yeast bur1 kinase stimulates recruitment of the paf complex. *Molecular and Cellular Biology*, 29(17):4852–4863.

Lock, A., Rutherford, K., Harris, M. A., Hayles, J., Oliver, S. G., Bähler, J., and Wood, V. (2018). PomBase 2018: user-driven reimplementation of the fission yeast database provides rapid and intuitive access to diverse, interconnected information. *Nucleic Acids Research*, 47(D1):D821–D827.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):550.

Machanick, P. and Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, 27(12):1696–1697.

MacIsaac, K. D., Wang, T., Gordon, D. B., Gifford, D. K., Stormo, G. D., and Fraenkel, E. (2006). An improved map of conserved regulatory sites for saccharomyces cerevisiae. *BMC Bioinformatics*, 7(1):113.

Malabat, C., Feuerbach, F., Ma, L., Saveanu, C., and Jacquier, A. (2015). Quality control of transcription start site selection by nonsense-mediated-mrna decay. *eLife*, 4:e06722.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12.

Martinez-Rucobo, F. W., Sainsbury, S., Cheung, A. C., and Cramer, P. (2011). Architecture of the rna polymerase–spt4/5 complex and basis of universal transcription processivity. *The EMBO Journal*, 30(7):1302–1310.

Mason, P. B. and Struhl, K. (2005). Distinction and relationship between elongation rate and processivity of rna polymerase ii in vivo. *Molecular Cell*, 17(6):831 – 840.

Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J. A., and Churchman, L. S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell*, 161(3):541–554.

Mayer, A., Lidschreiber, M., Siebert, M., Leike, K., Söding, J., and Cramer, P. (2010). Uniform transitions of the general rna polymerase ii transcription complex. *Nature Structural & Molecular Biology*, 17:1272–1278.

Mayer, A., Schreieck, A., Lidschreiber, M., Leike, K., Martin, D. E., and Cramer, P. (2012). The spt5 c-terminal region recruits yeast 3' rna cleavage factor i. *Molecular and Cellular Biology*, 32(7):1321–1331.

Mbogning, J., Nagy, S., Pagé, V., Schwer, B., Shuman, S., Fisher, R. P., and Tanny, J. C. (2013). The paf complex and prf1/rtf1 delineate distinct cdk9-dependent pathways regulating transcription elongation in fission yeast. *PLOS Genetics*, 9(12):1–14.

McCullough, L., Connell, Z., Petersen, C., and Formosa, T. (2015). The abundant histone chaperones spt6 and fact collaborate to assemble, inspect, and maintain chromatin structure in saccharomyces cerevisiae. *Genetics*, 201(3):1031–1045.

McDonald, S. M., Close, D., Xin, H., Formosa, T., and Hill, C. P. (2010). Structure and biological importance of the spn1-spt6 interaction, and its regulatory role in nucleosome binding. *Molecular Cell*, 40(5):725 – 735.

McKnight, K., Liu, H., and Wang, Y. (2014). Replicative stress induces intragenic transcription of the ase1 gene that negatively regulates ase1 activity. *Current Biology*, 24(10):1101 – 1106.

Meyer, P. A., Li, S., Zhang, M., Yamada, K., Takagi, Y., Hartzog, G. A., and Fu, J. (2015). Structures and functions of the multiple kow domains of transcription elongation factor spt5. *Molecular and Cellular Biology*, 35(19):3354–3369.

Morillon, A., Karabetsou, N., O'Sullivan, J., Kent, N., Proudfoot, N., and Mellor, J. (2003). Isw1 chromatin remodeling atpase coordinates transcription elongation and termination by rna polymerase ii. *Cell*, 115(4):425 – 435.

Newburger, D. E. and Bulyk, M. L. (2008). UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. *Nucleic Acids Research*, 37(suppl_1):D77–D82.

Orlando, D., Chen, M., Brown, V., Solanki, S., Choi, Y., Olson, E., Fritz, C., Bradner, J., and Guenther, M. (2014). Quantitative chip-seq normalization reveals global modulation of the epigenome. *Cell Reports*, 9(3):1163 – 1170.

Ozonov, E., Pachkov, M., Arnold, P., Balwierz, P. J., and van Nimwegen, E. (2012). SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Research*, 41(D1):D214–D220.

Pathak, R., Singh, P., Ananthakrishnan, S., Adamczyk, S., Schimmel, O., and Govind, C. K. (2018). Acetylation-dependent recruitment of the fact complex and its role in regulating pol ii occupancy genome-wide in saccharomyces cerevisiae. *Genetics*, 209(3):743–756.

Pelechano, V., Wei, W., and Steinmetz, L. M. (2013). Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*, 497:127 EP –.

Perales, R., Erickson, B., Zhang, L., Kim, H., Valiquett, E., and Bentley, D. (2013). Gene promoters dictate histone occupancy within genes. *The EMBO Journal*, 32(19):2645–2656.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nature Biotechnology*, 33:290 EP –.

Quan, T. K. and Hartzog, G. A. (2010). Histone h3k4 and k36 methylation, chd1 and rpd3s oppose the functions of saccharomyces cerevisiae spt4–spt5 in transcription. *Genetics*, 184(2):321–334.

Quinlan, A. R. and Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Rhee, H. S. and Pugh, B. F. (2012). Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*, 483:295 EP –. Article.

Rondón, A. G., García-Rubio, M., González-Barrera, S., and Aguilera, A. (2003). Molecular evidence for a positive role of spt4 in transcription elongation. *The EMBO Journal*, 22(3):612–620.

Schneider, S., Pei, Y., Shuman, S., and Schwer, B. (2010). Separable functions of the fission yeast spt5 carboxyl-terminal domain (ctd) in capping enzyme binding and transcription elongation overlap with those of the rna polymerase ii ctd. *Molecular and Cellular Biology*, 30(10):2353–2364.

Schwer, B., Schneider, S., Pei, Y., Aronova, A., and Shuman, S. (2009). Characterization of the schizosaccharomyces pombe spt5-spt4 complex. *RNA*, 15(7):1241–1250.

Sdano, M. A., Fulcher, J. M., Palani, S., Chandrasekharan, M. B., Parnell, T. J., Whitby, F. G., Formosa, T., and Hill, C. P. (2017). A novel sh2 recognition mechanism recruits spt6 to the doubly phosphorylated rna polymerase ii linker at sites of transcription. *eLife*, 6:e28723.

Shandilya, J. and Roberts, S. G. (2012). The transcription cycle in eukaryotes: From productive initiation to rna polymerase ii recycling. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1819(5):391 – 400.

Shetty, A., Kallgren, S. P., Demel, C., Maier, K. C., Spatt, D., Alver, B. H., Cramer, P., Park, P. J., and Winston, F. (2017). Spt5 plays vital roles in the control of sense and antisense transcription elongation. *Molecular Cell*, 66(1):77 – 88.e5.

Shivaswamy, S., Bhinge, A., Zhao, Y., Jones, S., Hirst, M., and Iyer, V. R. (2008). Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLOS Biology*, 6(3):1–13.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, chapter 5.3, pages 100–110. Chapman & Hall.

Stadelmayer, B., Micas, G., Gamot, A., Martin, P., Malirat, N., Koval, S., Raffel, R., Sobhian, B., Severac, D., Rialle, S., Parrinello, H., Cuvier, O., and Benkirane, M. (2014). Integrator complex regulates nelf-mediated rna polymerase ii pause/release and processivity at coding genes. *Nature Communications*, 5:5531 EP –. Article.

Stanlie, A., Begum, N. A., Akiyama, H., and Honjo, T. (2012). The dsif subunits spt4 and spt5 have distinct roles at various phases of immunoglobulin class switch recombination. *PLOS Genetics*, 8(4):1–11.

Subgroup, . G. P. D. P., Wysoker, A., Handsaker, B., Marth, G., Abecasis, G., Li, H., Ruan, J., Homer, N., Durbin, R., and Fennell, T. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Sun, M., Larivière, L., Dengl, S., Mayer, A., and Cramer, P. (2010). A tandem sh2 domain in transcription elongation factor spt6 binds the phosphorylated rna polymerase ii c-terminal repeat domain (ctd). *Journal of Biological Chemistry*, 285(53):41597–41603.

Teixeira, M. C., Monteiro, P. T., Palma, M., Costa, C., Godinho, C. P., Pais, P., Cavalheiro, M., Antunes, M., Lemos, A., Pedreira, T., and Sá-Correia, I. (2017). YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae. *Nucleic Acids Research*, 46(D1):D348–D353.

Tillo, D. and Hughes, T. R. (2009). G+c content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics*, 10(1):442.

Uwimana, N., Collin, P., Jeronimo, C., Haibe-Kains, B., and Robert, F. (2017). Bidirectional terminators in saccharomyces cerevisiae prevent cryptic transcription from invading neighboring genes. *Nucleic Acids Research*, 45(11):6417–6426.

van Bakel, H., Tsui, K., Gebbia, M., Mnaimneh, S., Hughes, T. R., and Nislow, C. (2013). A compendium of nucleosome and transcript profiles reveals determinants of chromatin architecture and transcription. *PLOS Genetics*, 9(5):1–18.

Viktorovskaya, O. V., Appling, F. D., and Schneider, D. A. (2011). Yeast transcription elongation factor spt5 associates with rna polymerase i and rna polymerase ii directly. *Journal of Biological Chemistry*.

Voss, K., Gentry, J., and Van der Auwera, G. (2017). Full-stack genomics pipelining with gatk4 + wdl + cromwell. In *18th Annual Bioinformatics Open Source Conference (BOSC 2017)*.

Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, A., Imai, T., Hirose, S., Sugimoto, S., Yano, K., Hartzog, G. A., Winston, F., Buratowski, S., and Handa, H. (1998). Dsif, a novel transcription elongation factor that regulates rna polymerase ii processivity, is composed of human spt4 and spt5 homologs. *Genes & Development*, 12(3):343–356.

Wagih, O. (2017). ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*, 33(22):3645–3647.

Wang, A. H., Juan, A. H., Ko, K. D., Tsai, P.-F., Zare, H., Dell'Orso, S., and Sartorelli, V. (2017). The elongation factor spt6 maintains esc pluripotency by controlling

super-enhancers and counteracting polycomb proteins. *Molecular Cell*, 68(2):398 – 413.e6.

Wang, A. H., Zare, H., Mousavi, K., Wang, C., Moravec, C. E., Sirotkin, H. I., Ge, K., Gutierrez-Cruz, G., and Sartorelli, V. (2013). The histone chaperone spt6 coordinates histone h3k27 demethylation and myogenesis. *The EMBO Journal*, 32(8):1075–1086.

Wang, L., Chen, J., Wang, C., Uusküla-Reimand, L., Chen, K., Medina-Rivera, A., Young, E. J., Zimmermann, M. T., Yan, H., Sun, Z., Zhang, Y., Wu, S. T., Huang, H., Wilson, M. D., Kocher, J.-P. A., and Li, W. (2014). Mace: model based analysis of chip-exo. *Nucleic Acids Research*, 42(20):e156.

Weber, G., Springer, M., Jorgensen, P., Milo, R., and Moran, U. (2009). Bionumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Research*, 38(suppl_1):D750–D753.

Wehrens, R. and Buydens, L. (2007). Self- and super-organizing maps in r: The kohonen package. *Journal of Statistical Software, Articles*, 21(5):1–19.

Weirauch, M., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H., Lambert, S., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J., Govindarajan, S., Shaulsky, G., Walhout, A., Bouget, F.-Y., Ratsch, G., Larrondo, L., Ecker, J., and Hughes, T. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6):1431 – 1443.

Wen, Y. and Shatkin, A. J. (1999). Transcription elongation factor hspt5 stimulates mrna capping. *Genes & Development*, 13(14):1774–1779.

Werner, F. (2012). A nexus for gene expression—molecular mechanisms of spt5 and nusg in the three domains of life. *Journal of Molecular Biology*, 417(1):13 – 27.

Wier, A. D., Mayekar, M. K., Héroux, A., Arndt, K. M., and VanDemark, A. P. (2013). Structural basis for spt5-mediated recruitment of the paf1 complex to chromatin.

Wood, V., Gwilliam, R., Rajandream, M.-A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., Basham, D., Bowman, S., Brooks, K., Brown, D., Brown, S., Chillingworth, T., Churcher, C., Collins, M., Connor, R., Cronin, A., Davis, P., Feltwell, T., Fraser, A., Gentles, S., Goble, A., Hamlin, N., Harris, D., Hidalgo, J., Hodgson, G., Holroyd, S., Hornsby, T., Howarth, S., Huckle, E. J., Hunt, S., Jagels, K., James, K., Jones, L., Jones, M., Leather, S., McDonald, S., McLean, J., Mooney, P., Moule, S., Mungall, K., Murphy, L., Niblett, D., Odell, C., Oliver, K., O'Neil, S., Pearson, D., Quail, M. A., Rabbinowitsch, E., Rutherford, K., Rutter,

S., Saunders, D., Seeger, K., Sharp, S., Skelton, J., Simmonds, M., Squares, R., Squares, S., Stevens, K., Taylor, K., Taylor, R. G., Tivey, A., Walsh, S., Warren, T., Whitehead, S., Woodward, J., Volckaert, G., Aert, R., Robben, J., Grymonprez, B., Weltjens, I., Vanstreels, E., Rieger, M., Schäfer, M., Müller-Auer, S., Gabel, C., Fuchs, M., Fritzc, C., Holzer, E., Moestl, D., Hilbert, H., Borzym, K., Langer, I., Beck, A., Lehrach, H., Reinhardt, R., Pohl, T. M., Eger, P., Zimmermann, W., Wedler, H., Wambutt, R., Purnelle, B., Goffeau, A., Cadieu, E., Dréano, S., Gloux, S., Lelaure, V., Mottier, S., Galibert, F., Aves, S. J., Xiang, Z., Hunt, C., Moore, K., Hurst, S. M., Lucas, M., Rochet, M., Gaillardin, C., Tallada, V. A., Garzon, A., Thode, G., Daga, R. R., Cruzado, L., Jimenez, J., Sánchez, M., del Rey, F., Benito, J., Domínguez, A., Revuelta, J. L., Moreno, S., Armstrong, J., Forsburg, S. L., Cerrutti, L., Lowe, T., McCombie, W. R., Paulsen, I., Potashkin, J., Shpakovski, G. V., Ussery, D., Barrell, B. G., and Nurse, P. (2002). The genome sequence of schizosaccharomyces pombe. *Nature*, 415(6874):871–880.

Yamaguchi, Y., Wada, T., Watanabe, D., Takagi, T., Hasegawa, J., and Handa, H. (1999). Structure and function of the human transcription elongation factor dsif. *Journal of Biological Chemistry*, 274(12):8085–8092.

Yamamoto, J., Hagiwara, Y., Chiba, K., Isobe, T., Narita, T., Handa, H., and Yamaguchi, Y. (2014). Dsif and nelf interact with integrator to specify the correct post-transcriptional fate of snrna genes. *Nature Communications*, 5:4263 EP –. Article.

Yoh, S. M., Cho, H., Pickle, L., Evans, R. M., and Jones, K. A. (2007). The spt6 sh2 domain binds ser2-p rnapii to direct iws1-dependent mrna splicing and export. *Genes & Development*, 21(2):160–174.

Yoh, S. M., Lucas, J. S., and Jones, K. A. (2008). The iws1:spt6:ctd complex controls cotranscriptional mrna biosynthesis and hypb/setd2-mediated histone h3k36 methylation. *Genes & Development*, 22(24):3422–3434.

Youdell, M. L., Kizer, K. O., Kisseleva-Romanova, E., Fuchs, S. M., Duro, E., Strahl, B. D., and Mellor, J. (2008). Roles for ctk1 and spt6 in regulating the different methylation states of histone h3 lysine 36. *Molecular and Cellular Biology*, 28(16):4915–4926.

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for rna-seq: accounting for selection bias. *Genome Biology*, 11(2):R14.

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.-W., Nechaev, S., Adelman, K., Levine, M., and Young, R. A. (2007). Rna polymerase stalling at developmental control genes in the drosophila melanogaster embryo. *Nature Genetics*, 39:1512 EP –.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of chip-seq (macs). *Genome Biology*, 9(9):R137.

Zhang, Y., Moqtaderi, Z., Rattner, B. P., Euskirchen, G., Snyder, M., Kadonaga, J. T., Liu, X. S., and Struhl, K. (2009). Intrinsic histone-dna interactions are not the major determinant of nucleosome positions in vivo. *Nature Structural &Amp; Molecular Biology*, 16:847 EP –. Article.

Zhou, K., Kuo, W. H. W., Fillingham, J., and Greenblatt, J. F. (2009). Control of transcriptional elongation and cotranscriptional histone modification by the yeast bur kinase substrate spt5. *Proceedings of the National Academy of Sciences*, 106(17):6956–6961.

Zhu, J. and Zhang, M. Q. (1999). SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics*, 15(7):607–611.

Zhu, W., Wada, T., Okabe, S., Taneda, T., Yamaguchi, Y., and Handa, H. (2007). DSIF contributes to transcriptional activation by DNA-binding activators by preventing pausing during transcription elongation. *Nucleic Acids Research*, 35(12):4064–4075.

Vita

James Chuang

year of birth: 1991

contact address: 77 Avenue Louis Pasteur

Room 239

Boston, MA 02115

Education

2018 MSc, Biomedical Engineering, Boston University

2013 BSc, Biomedical Engineering, Johns Hopkins University

Publications

2018 Doris SM*, **Chuang J***, Viktorovskaya O, Murawska M, Spatt D, Churchman LS, Winston F (2018). Spt6 is required for the fidelity of promoter selection. **Molecular Cell**, doi:10.1016/j.molcel.2018.09.005

2018 **Chuang J**, Boeke JD, Mitchell LA (2018). Coupling Yeast Golden Gate and VEGAS for Efficient Assembly of the Violacein Pathway in *Saccharomyces cerevisiae*. **Synthetic Metabolic Pathways**, doi:10.1007/978-1-4939-7295-1_14

2017 Aquino P, Honda B, Suma Jaini, Lyubetskaya A, Hosur K, Chiu JG, Ekladious I, Hu D, Jin L, Sayeg MK, Stettner AI, Wang J, Wong BG, Wong WS, Alexander SL, Ba C, Bensussen SI, Chou K, **Chuang J**, Gastler DE, Grasso DJ, Greifenberger JS, Guo C, Hawes AK, Israni DV, Jain SR, Kim J, Lei J, Li H, Li D, Li Q, Mancuso CP, Mao N, Masud SF, Meisel CL, Mi J, Nykyforchyn CS, Park M, Peterson HM, Ramirez AK, Reynolds DS, Rim NG, Saffie JC, Su H, Su WR, Su Y, Sun M, Thommes MM, Tu T, Varongchayakul N, Wagner TE, Weinberg BH, Yang R, Yaroslavsky A, Yoon C, Zhao Y, Zollinger AJ, Stringer AM, Foster JW, Wade J, Raman S, Broude N, Wong WW, Galagan JE (2017). Coordinated regulation of acid resistance in *Escherichia coli*. **BMC Systems Biology**, doi:10.1186/s12918-016-0376-y

2015 Mitchell, LA*, **Chuang J***, Agmon N, Khunsriraksakul C, Phillips NA, Cai Y, Truong DM, Veerakumar A, Wang Y, Mayorga M, Blomquist P, Sadda P, Trueheart J, Boeke JD (2015). Versatile genetic assembly system (VEGAS) to assemble pathways for expression in *S. cerevisiae*. **Nucleic Acids Research**, doi:10.1093/nar/gkv466

2015 Agmon N, Mitchell LA, Cai Y, Ikushima S, **Chuang J**, Zheng A, Choi W, Martin JA, Caravelli K, Stracquadanio G, Boeke JD (2015). Yeast Golden Gate (yGG) for the Efficient Assembly of *S. cerevisiae* Transcription Units. **ACS Synthetic Biology**, doi:10.1021/sb500372z

2013 Mitchell LA, Cai Y, Taylor M, Noronha AM, **Chuang J**, Dai L, Boeke JD (2013). Multichange isothermal mutagenesis: a new strategy for multiple site-directed mutations in plasmid DNA. **ACS Synthetic Biology**, doi:10.1021/sb300131w