Mechanism and Scaling of Eukaryotic Transcription Activation

Thesis by
Porfirio Quintero Cadena

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2020
Defended May 15, 2020

Porfirio Quintero Cadena
ORCID: 0000-0003-0067-5844

ACKNOWLEDGEMENTS

I believe that I will remember my time in Caltech as one of the most formative, interesting, and intellectually stimulating experiences of my life. I feel incredibly privileged to have had the opportunity to meet and interact with such talented, kind, and supportive colleagues and friends throughout my PhD.

I would like to thank my advisor Paul Sternberg for his unyielding patience, support, and enthusiasm. Paul believed in me even when the direction of my research branched off both our comfort zones. I will always be grateful for the intellectual freedom and support to pursue the problems I found interesting in his laboratory. Although often challenging, I believe I come out of this experience with a humbling yet more accurate sense of what it means to try to understand nature through scientific paradigms.

I would also like to thank my collaborator Tineke Lenstra for her considerable contribution from the other side of the ocean. Without any obligation, she kindly shared reagents and expertise that were instrumental to my research.

To my thesis committee, thank you for your support. Matt Thomson's encouragement to concretely write down my perspective of transcription for computer simulations was pivotal in consolidating my work. Ellen Rothenberg's incisiveness was a constant source of reason in our meetings. Mitch Guttman provided insightful feedback well beyond committee meetings, and I am extremely grateful for his support, mentorship, and generosity throughout my PhD.

Coming to Caltech was a dream I did not expect would come true, and several people made significant contributions that made it possible. I would like to thank my undergraduate mentors Claudio Moreno, for the inspiring stories and conversations that motivated me to stay in biology, and Ruben Morones, for believing in me and opening doors that I did not know how to find. Claudio also introduced me to Eric Davidson's work. Eric, in turn, accepted my invitation to talk in my undergraduate institution and motivated me to apply to Caltech. I am profoundly thankful for Eric's early show of faith and for its transformative impact on my academic path.

To the long list of people from whom I learned in Caltech, thank you. A few examples include Justin Bois, whose excellent course material was an invaluable data analysis resource; Heun Jin Lee taught me how to use and build a microscope; in the Sternberg group, Han Wang was a constant source of feedback and support,

Carmie Robinson always open to stimulating conversation, Margaret Ho welcomed me with a friendly face and showed me the ropes of worm research, and each of my labmates provided a warm, welcoming environment.

I am also very proud to be part of the small and exceptional Mexican community at Caltech, and extremely grateful for their friendship. In particular, Manuel Razo has been a dear friend and inspiration since we started graduate school together, an every week I looked forward to lunch with my friends Alejandro Granados, Jorge Castellanos, and Emmanuel Garza.

To my partner Ellen Yan, thank you for your support in my transition out of the PhD. Your timely arrival to my life made my last year in graduate school immensely enjoyable, and I feel especially lucky to have you in my life during the challenging times of this pandemic.

Finally, I would like to thank my parents Martha Luisa Cadena Pale and Porfirio Quintero Gómez. They instilled in me the value of an education, nurtured and supported my ambitions, and never stopped believing in me. Their hard, relentless work put me in the privileged position that allowed me to be where I am today, and I will be forever grateful.

ABSTRACT

Transcription activation is a universal process by which living cells adapt. Decades of work in this field have produced an intelligible paradigm of transcription activation that provides fundamental insights into its underlying molecular mechanisms. This thesis attempts to extend such paradigm to explain how transcription activation can be implemented across the diversity of molecular environments found in eukaryotic nuclei. Specifically, this diversity calls for an explanation of how this process scales throughout a range of genome sizes that spans five orders of magnitude, and of how to think about this subject in the increasingly relevant context of liquid-liquid phase-separation. We leverage data from RNA-seq, smFISH, growth-rate, fluorescence microscopy, computer simulations and literature to identify an appropriate and useful level of abstraction in which to grow our current paradigm. We propose scaling and phase-separation, two seemingly disparate aspects of transcription, are explained and intrinsically linked by a novel molecular state in which multiple RNA polymerases can bind the transcription complex. We provide support and rationale for this addition to the transcription model, and generate testable hypotheses that may further clarify the mechanism and evolution of eukaryotic transcription activation.

PUBLISHED CONTENT AND CONTRIBUTIONS

Quintero-Cadena, P., Lenstra, T. L., & Sternberg, P. W. (2020). RNA Pol II Length and Disorder Enable Cooperative Scaling of Transcriptional Bursting. *Molecular Cell*. doi:https://doi.org/10.1016/j.molcel.2020.05.030.
P.Q.C. conceived the project, designed and performed experiments, analyzed the data, and participated in writing the manuscript.

Quintero-Cadena, P. & Sternberg, P. W. (2016). Enhancer Sharing Promotes Neighborhoods of Transcriptional Regulation Across Eukaryotes. *G3 (Bethesda, Md.) 6*(12), 4167–4174. doi:https://doi.org/10.1534/g3.116.036228.
P.Q.C. conceived the project, designed and performed experiments, analyzed the data, and participated in writing the manuscript.

CONTENTS

Chapter 1

INTRODUCTION TO TRANSCRIPTION ACTIVATION

1.1 Paradigm of Eukaryotic Transcription

Every living cell expresses a set of genes that largely defines its identity. Of the available repertoire, this set is activated in response to initial and environmental conditions (Davidson & Peter, 2015); constant fluctuations in these conditions are followed by adaptation through gene expression, in turn enabling differentiation, communication or survival.

The process of turning on a gene is universal to all cells. It is thus not unreasonable to expect for its fundamental mechanisms to be conserved throughout life. Decades of work in this field have indeed produced an intelligible paradigm of transcription activation. This perspective articulates the process in a reductive yet useful cartoon that has been successfully applied to the interpretation of experimental data, from single genes to entire genomes.

In this cartoon, a set of transcription factors (TF, proteins involved in activating transcription) forms a preinitiation complex that facilitates recruitment of the RNA Polymerase II (Roeder, 1996; Hahn, 2004). This multi-subunit enzyme transcribes the information genetically encoded as DNA into an mRNA molecule, which can diffuse away to be translated into a protein or fulfill other roles (Figure 1.1).

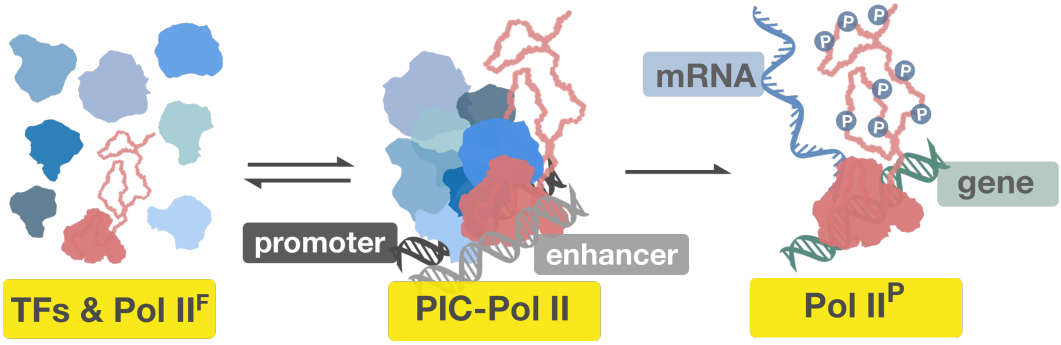

Figure 1.1: **Simplified cartoon of eukaryotic transcription activation.** In response to a stimulus, transcription factors (TFs) bind the enhancer and the promoter, assembling into a preinitiation complex (PIC) that recruits a free RNA polymerase II (Pol II^F). A functional polymerase can be released by phosphorylation (Pol II^P) to initiate transcription of the downstream gene.

In eukaryotes, the molecular assembly of such transcription complex typically occurs on a scaffold comprised of two DNA loci: the promoter, located at the start of the gene (Fuda, Ardehali, & Lis, 2009), and the enhancer, located generally in physical proximity within the same DNA molecule (Plank & Dean, 2014; Pombo & Dillon, 2015; Long, Prescott, & Wysocka, 2016; Furlong & Levine, 2018). Enhancers usually bind gene-specific TFs, providing an avenue for pathways responsive to environmental conditions to induce corresponding changes in gene expression (Shlyueva, Stampfel, & Stark, 2014). On the other hand, promoters tend to be bound by a TF set that is common to many genes (Roeder, 1996). These proteins include the components of the Mediator complex, which constitutes a physical bridge between enhancers and promoters (Allen & Taatjes, 2015; Plaschka et al., 2015; Robinson et al., 2015; Petrenko et al., 2016; Jeronimo & Robert, 2017), general transcription factors that interact with components of the RNA Polymerase II, and the resulting preinitiation complex (PIC), from which a functional polymerase molecule is released to start transcription (Thompson, Koleske, Chao, & Young, 1993; Kim, Björklund, Li, Sayre, & Kornberg, 1994; Guidi et al., 2004; Takagi & Kornberg, 2006; Esnault et al., 2008; Malik, Molina, & Xue, 2017; Wong, Jin, & Struhl, 2014).

1.2 Dynamics of Transcription

The temporal behavior of transcription is not intuitive from the static cartoon described above. Transcription activation is a dynamic process involving over fifty distinct proteins (Cramer, 2019), which gather on a two-part DNA scaffold. Given this multi-component, rare molecular event precedes the production of an mRNA molecule, the naive expectation is that populations of mRNAs per cell would follow a Poisson distribution. Instead, single molecule measurements yield distributions that consistently deviate from this expectation (Raj, Peskin, Tranchina, Vargas, & Tyagi, 2006; Tunnacliffe & Chubb, 2020).

Experimentally visualizing the process of transcription in living cells (Figure 1.2A) reveals that mRNA molecules are not typically produced one by one but in bursts of activity (Golding, Paulsson, Zawilski, and Cox, 2005; Chubb, Trcek, Shenoy, and Singer, 2006; Figure 1.2B). Quantitatively, a model that includes an inactive state, in which no mRNA molecules can be produced, is sufficient to recapitulate experimental examples of bursting and mRNA distributions (Raj et al., 2006).

The molecular origin of bursting is not clear and may be multifactorial (Nicolas,

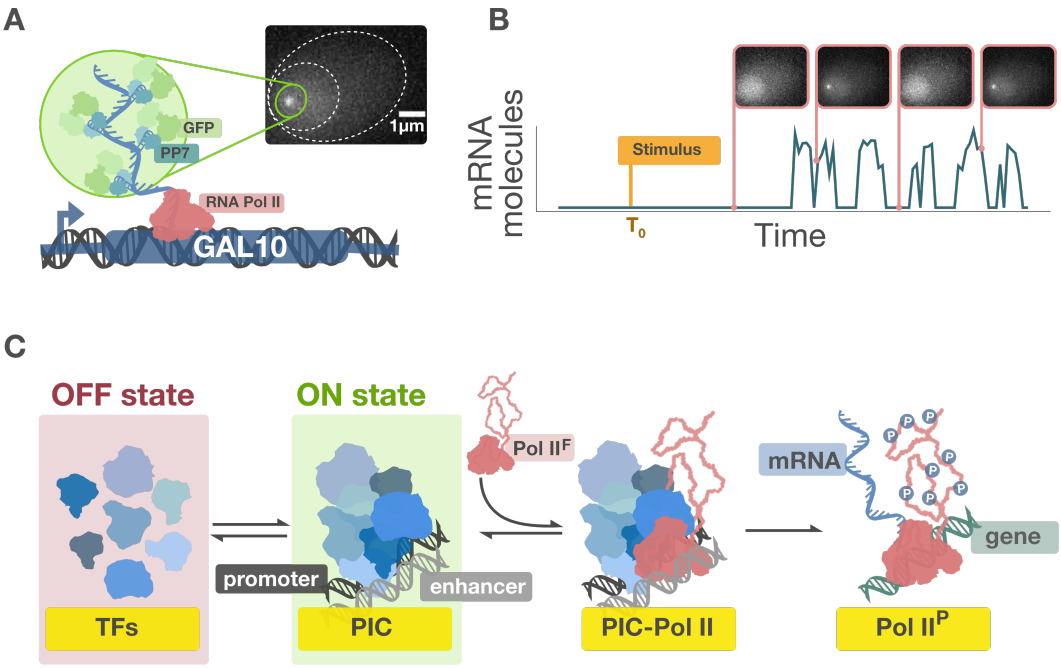

Figure 1.2: **Bursty transcription can be generated by a two-state model.** (A) Experimental strategy used to observe transcription dynamics in live cells. A transcript is tagged with RNA hairpins that are bound by nuclear expressed PP7 fused to GFP. Because the transcription site concentrates mRNA molecules, PP7-GFP binding results in a fluorescent spot in the cell nucleus upon transcription activation, whose fluorescence through time (B) reveals the bursty dynamics of transcription. This behavior can be quantitatively recapitulated by a two-state model; this cartoon (C) depicts a potential molecular mechanism underlying each of these and subsequent states.

Phillips, & Naef, 2017; Tunnacliffe & Chubb, 2020). One compelling possibility, inspired in the biochemical cartoon of transcription, is that the transcription machinery can remain assembled through multiple rounds of polymerase binding and release, each of these events resulting in an mRNA molecule; disassembly of this complex would lead to inactivity periods, after which the cycle repeats (Figure 1.2C).

Support for this interpretation can be found in live, simultaneous imaging of enhancer-promoter interactions and transcription: burst initiation is correlated with events of physical proximity between these two DNA loci (Fukaya, Lim, & Levine, 2016; Chen et al., 2018), which is presumably linked to TF binding and formation of the transcription complex (Donovan et al., 2019; Stavreva et al., 2019). The statistics of transcription data also provide informative clues. Burst arrival behaves like a Poisson process (Chubb et al., 2006; Larson, Zenklusen, Wu, Chao, & Singer,

2011), matching the original expectation of a rare assembly event. The numbers of mRNA molecules per burst appear to be geometrically distributed (Raj et al., 2006), which is consistent with the story of several rounds of polymerase loading until complex dissassembly. Transcriptome-wide, enhancers appear to shape burst frequency, while promoter sequences modulate burst size (Larsson et al., 2019).

1.3 Scaling of Transcription Activation

Assembling the transcription machinery requires available TFs and accessible DNA loci. These features provide major opportunities for regulating gene activation: TFs can be produced, or their binding facilitated by post-translational modifications (Zabidi & Stark, 2016); DNA can be made accessible by chromatin remodeling, which must precede initial TF binding to allow assembly (Fuda et al., 2009; Lorch & Kornberg, 2015); genome architecture can be modified to bring linearly distant DNA loci into close physical proximity (Pombo & Dillon, 2015; Catarino & Stark, 2018).

The assembly of an enhancer-promoter scaffold represents a concrete conceptual challenge for transcription activation. The probability of interaction between two DNA loci in the same molecule, using polymer chain statistics as the expectation, should decay exponentially with the distance between them (Ringrose, Chabanis, Angrand, Woodroofe, & Stewart, 1999). However, particularly in large genomes, transcription from distant enhancers is not orders of magnitude less common.

Along with genome sizes, nucleotide (nt) distances over which enhancer-promoter interactions occur in eukaryotes vary across several orders of magnitude. While yeast cells have a genome of 10^7 nt, with interactions typically occurring over DNA stretches of only 10^2 nt (Dobi & Winston, 2007), human cells have a genome of $3x10^9$ nt, where enhancers and promoters are commonly separated by distances of 10^5 nt (Sanyal, Lajoie, Jain, & Dekker, 2012).

How is transcription scaled in genomes to cope with such order-of-magnitude variation? Part of the answer lies in genome organization, whose measurements have revealed structures that facilitate these seemingly unlikely enhancer-promoter interactions.

Different eukaryotic species exhibit arrangements of nuclear DNA that may partially compensate for large variations in genome size (Szabo, Bantignies, & Cavalli, 2019). Generally, so-called topologically-associated domains (Dixon et al., 2012; Nora et al., 2012) form clustered chromatin neighborhoods, such that *in vivo* probabilities of

pairwise interactions decay with a length-scale that is significantly larger than *in vitro* measurements (Lieberman-Aiden et al., 2009). Pairwise interactions are presumably driven by extruding DNA through protein rings anchored at the base of topological domains (Sanborn et al., 2015; Fudenberg et al., 2016). This dynamic process temporarily brings linearly distant pieces of DNA into a range of physical proximity in which protein-mediated physical enhancer-promoter interactions become possible (Pombo & Dillon, 2015). However, even with the significant rearrangements that take place in large genomes, pairwise interactions between distant genomic loci are rare (Rao et al., 2014), leaving room for additional mechanisms to compensate for lower frequencies of enhancer-promoter interactions.

1.4 Liquid-Liquid Phase-separation in Transcription

As a coherent picture of transcription emerges, connecting the rich biochemistry of the transcription complex with the quantitative and dynamic measurements of mRNA production, a revolutionizing perspective becomes increasingly relevant across biology: liquid-liquid phase-separation (LLPS).

The phenomenon of LLPS manifests as cellular structures that concentrate certain biological molecules, including nucleic acids and proteins, into fluid, membraneless compartments that dynamically emerge, merge and dissolve (Banani, Lee, Hyman, & Rosen, 2017; Shin & Brangwynne, 2017). Under this light, the cell nucleus becomes an undeniably dynamic organelle, hosting a range of environments in which seemingly unlikely interactions and reactions can occur.

Unstructured protein domains, pervasive in nuclear proteins, with residues free to form weakly specific inter-molecular interactions, have been discovered to promote the formation of LLPS droplets (Banani et al., 2017; Shin & Brangwynne, 2017). These events can be facilitated by nucleic acid scaffolds (Banani et al., 2016; Jain & Vale, 2017; Banani et al., 2017), which constitute droplet nucleation sites with spatial coordinates that can be exploited for biological function.

Evidence supporting the relevance of LLPS in transcription is growing. LLPS droplets have been observed in live cells at highly transcribed loci, enriched in RNA polymerases and major TFs (Cho et al., 2018; Chong et al., 2018; Sabari et al., 2018; Shin et al., 2018; Figure 1.3). The RNA Polymerase II itself contains an unstructured C-terminal domain (CTD) that can dynamically form and bind these droplets (Kwon et al., 2013; Boehning et al., 2018). Moreover, phosphorylation can modulate protein entry and release from these droplets (Kwon et al., 2013; Chong

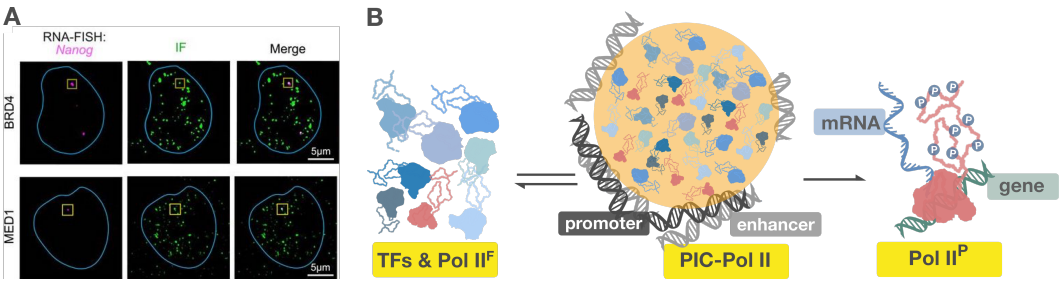

Figure 1.3: **Transcription factors and the RNA Polymerase II form droplets through their disordered domains.** (A) Transcription factors BRD4 and MED1 form puncta, detected by immunofluorescence imaging of mouse embryonic stem cells, that colocalize with the nascent mRNA of Nanog, detected by RNA fluorescence *in situ* hybridization. Figure from Sabari et al., 2018. (B) Cartoon of phase separated droplets presumably underlying these puncta. These droplets form through the disordered domains of transcription factors (TFs) and free RNA Pol II (Pol II^F) at super-enhancers, which are linked to highly transcribed genes.

et al., 2018; Boehning et al., 2018; Cho et al., 2018; Nair et al., 2019). Generally, mechanisms modulating protein-protein interactions could conceivably be exploited to regulate droplet dynamics (Cramer, 2019).

LLPS brings a completely new perspective to biology that could radically change the way we think about transcription. Are highly transcribed loci fundamentally different from the average gene, or does transcription generally occur in a droplet? Is the mechanistic cartoon of transcription an accurate representation of the underlying process? How do we reconcile LLPS and the transcription paradigm described above?

1.5 Thesis overview

This thesis addresses two questions that arise when considering enhancer-promoter interactions that drive transcription activation in eukaryotes.

1) If Mediator and a common set of TFs generally bind the promoter, how can enhancers specifically drive transcription from target promoters? Chapter II argues this specificity is mostly determined by physical proximity, by showing that neighboring genes tend to be more correlated in expression than expected by chance. This correlation occurs regardless of genome size and decays exponentially with distance between gene pairs. A corollary of this observation is that enhancer-promoter distances are an additional layer of information to modulate the genetic influence of enhancers.

2) What are the compensatory mechanisms for the reduced frequency of enhancer-promoter interactions expected in large genomes? While changes in genome organization are necessary and contribute significantly, Chapter III explores how cross-species variations in the RNA Polymerase II may influence the scaling of transcriptional dynamics. By providing a quantitative argument to explain the role of this variation, a framework to connect the canonical transcription paradigm with the emerging perspective of LLPS is proposed.

Finally, the last chapter offers some concluding remarks to summarize lessons learned and communicate experimental propositions that could falsify or build on the paradigm and ideas described throughout this thesis.

BIBLIOGRAPHY

Allen, B. L. & Taatjes, D. J. (2015). The Mediator complex: a central integrator of transcription. *Nature reviews. Molecular cell biology*, *16*(3), 155–166. doi:10.1038/nrm3951

Banani, S. F., Lee, H. O., Hyman, A. A., & Rosen, M. K. (2017). Biomolecular condensates: Organizers of cellular biochemistry. *Nature Reviews Molecular Cell Biology*, *18*, 285–298. doi:10.1038/nrm.2017.7

Banani, S. F., Rice, A. M., Peeples, W. B., Lin, Y., Jain, S., Parker, R., & Rosen, M. K. (2016). Compositional Control of Phase-Separated Cellular Bodies. *Cell*. doi:10.1016/j.cell.2016.06.010

Boehning, M., Dugast-Darzacq, C., Rankovic, M., Hansen, A. S., Yu, T., Marie-Nelly, H., . . . Zweckstetter, M. (2018). RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nature Structural and Molecular Biology*, *25*(9), 833–840. doi:10.1038/s41594-018-0112-y

Catarino, R. R. & Stark, A. (2018). Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. doi:10.1101/gad.310367.117

Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J. B., & Gregor, T. (2018). Dynamic interplay between enhancer–promoter topology and gene activity. *Nature Genetics*, *50*, 1296–1303. doi:10.1038/s41588-018-0175-z

Cho, W.-K., Spille, J.-H., Hecht, M., Lee, C., Li, C., Grube, V., & Cisse, I. I. (2018). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science*, *361*(6400), 412–415. doi:10.1126/science.aar4199

Chong, S., Dugast-Darzacq, C., Liu, Z., Dong, P., Dailey, G. M., Cattoglio, C., . . . Tjian, R. (2018). Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science*, *361*(6400), eaar2555. doi:10.1126/science.aar2555

Chubb, J. R., Trcek, T., Shenoy, S. M., & Singer, R. H. (2006). Transcriptional Pulsing of a Developmental Gene. *Current Biology*, *16*(10), 1018–1025. doi:10.1016/j.cub.2006.03.092

Cramer, P. (2019). Organization and regulation of gene transcription. doi:10.1038/s41586-019-1517-4

Davidson, E. H. & Peter, I. S. (2015). Chapter 1 - the genome in development. In E. H. Davidson & I. S. Peter (Eds.), *Genomic control process* (pp. 1–40). Oxford: Academic Press. doi:http://dx.doi.org/10.1016/B978-0-12-404729-7.00001-0

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., . . . Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. doi:10.1038/nature11082

Dobi, K. C. & Winston, F. (2007). Analysis of Transcriptional Activation at a Distance in Saccharomyces cerevisiae. *Molecular and Cellular Biology*, *27*(15), 5575–5586. doi:10.1128/MCB.00459-07

Donovan, B. T., Huynh, A., Ball, D. A., Patel, H. P., Poirier, M. G., Larson, D. R., . . . Lenstra, T. L. (2019). Live-cell imaging reveals the interplay between transcription factors, nucleosomes, and bursting. *The EMBO Journal*. doi:10.15252/embj.2018100809

Esnault, C., Ghavi-Helm, Y., Brun, S., Soutourina, J., Van Berkum, N., Boschiero, C., . . . Werner, M. (2008). Mediator-Dependent Recruitment of TFIIH Modules in Preinitiation Complex. *Molecular Cell*. doi:10.1016/j.molcel.2008.06.021

Fuda, N. J., Ardehali, M. B., & Lis, J. T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. doi:10.1038/nature08449

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., & Mirny, L. A. (2016). Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports*. doi:10.1016/j.celrep.2016.04.085

Fukaya, T., Lim, B., & Levine, M. (2016). Enhancer Control of Transcriptional Bursting. *Cell*, *166*(2), 358–368. doi:10.1016/j.cell.2016.05.025

Furlong, E. E. & Levine, M. (2018). Developmental enhancers and chromosome topology. doi:10.1126/science.aau0320

Golding, I., Paulsson, J., Zawilski, S. M., & Cox, E. C. (2005). Real-time kinetics of gene activity in individual bacteria. *Cell*. doi:10.1016/j.cell.2005.09.031

Guidi, B. W., Bjornsdottir, G., Hopkins, D. C., Lacomis, L., Erdjument-Bromage, H., Tempst, P., & Myers, L. C. (2004). Mutual targeting of Mediator and the TFIIH kinase Kin28. *Journal of Biological Chemistry*. doi:10.1074/jbc.M404426200

Hahn, S. (2004). Structure and mechanism of the RNA polymerase II transcription machinery. doi:10.1038/nsmb763

Jain, A. & Vale, R. D. (2017). RNA phase transitions in repeat expansion disorders. *Nature*. doi:10.1038/nature22386

Jeronimo, C. & Robert, F. (2017). The Mediator Complex: At the Nexus of RNA Polymerase II Transcription. doi:10.1016/j.tcb.2017.07.001

Kim, Y. J., Björklund, S., Li, Y., Sayre, M. H., & Kornberg, R. D. (1994). A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. *Cell*. doi:10.1016/0092-8674(94)90221-6

Kwon, I., Kato, M., Xiang, S., Wu, L., Theodoropoulos, P., Mirzaei, H., ... McKnight, S. L. (2013). Phosphorylation-Regulated Binding of RNA Polymerase II to Fibrous Polymers of Low-Complexity Domains. *Cell*, *155*(5), 1049–1060. doi:10.1016/j.cell.2013.10.033

Larson, D. R., Zenklusen, D., Wu, B., Chao, J. A., & Singer, R. H. (2011). Real-Time Observation of Transcription Initiation and Elongation on an Endogenous Yeast Gene. *Science*, *332*(6028), 475–478. doi:10.1126/science.1202142. arXiv: NIHMS150003

Larsson, A. J., Johnsson, P., Hagemann-Jensen, M., Hartmanis, L., Faridani, O. R., Reinius, B., ... Sandberg, R. (2019). Genomic encoding of transcriptional burst kinetics. doi:10.1038/s41586-018-0836-1

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., ... Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, *326*(5950), 289–293. doi:10.1126/science.1181369

Long, H. K., Prescott, S. L., & Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. doi:10.1016/j.cell.2016.09.018

Lorch, Y. & Kornberg, R. D. (2015). Chromatin-remodeling and the initiation of transcription. doi:10.1017/S0033583515000116

Malik, S., Molina, H., & Xue, Z. (2017). PIC Activation through Functional Interplay between Mediator and TFIIH. *Journal of Molecular Biology*. doi:10.1016/j.jmb.2016.11.026

Nair, S. J., Yang, L., Meluzzi, D., Oh, S., Yang, F., Friedman, M. J., ... Rosenfeld, M. G. (2019). Phase separation of ligand-activated enhancers licenses cooperative chromosomal enhancer assembly. *Nature Structural & Molecular Biology*, *26*(3), 193–203. doi:10.1038/s41594-019-0190-5

Nicolas, D., Phillips, N. E., & Naef, F. (2017). What shapes eukaryotic transcriptional bursting? *Molecular BioSystems*, *13*(7), 1280–1290. doi:10.1039/c7mb00154a

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., ... Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*. doi:10.1038/nature11049

Petrenko, N., Jin, Y., Petrenko, N., Jin, Y., Wong, K. H., & Struhl, K. (2016). Mediator Undergoes a Compositional Change during Transcriptional Activation Article Mediator Undergoes a Compositional Change during Transcriptional Activation. *Molecular Cell*, *64*(3), 443–454. doi:10.1016/j.molcel.2016.09.015

Plank, J. L. & Dean, A. (2014). Enhancer function: Mechanistic and genome-wide insights come together. doi:10.1016/j.molcel.2014.06.015

Plaschka, C., Larivière, L., Wenzeck, L., Seizl, M., Hemann, M., Tegunov, D., ... Cramer, P. (2015). Architecture of the RNA polymerase II-Mediator core initiation complex. *Nature*. doi:10.1038/nature14229

Pombo, A. & Dillon, N. (2015). Three-dimensional genome architecture: Players and mechanisms. doi:10.1038/nrm3965

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., & Tyagi, S. (2006). Stochastic mrna synthesis in mammalian cells. *PLOS Biology*, *4*(10), 1–13. doi:10.1371/journal.pbio.0040309

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., ... Aiden, E. L. (2014). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680. doi:10.1016/j.cell.2014.11.021

Ringrose, L., Chabanis, S., Angrand, P. O., Woodroofe, C., & Stewart, A. F. (1999). Quantitative comparison of dna looping in vitro and in vivo: chromatin increases effective dna flexibility at short distances. *EMBO J. 18*(23), 6630–6641. doi:10.1093/emboj/18.23.6630

Robinson, P. J., Trnka, M. J., Pellarin, R., Greenberg, C. H., Bushnell, D. A., Davis, R., ... Kornberg, R. D. (2015). Molecular architecture of the yeast Mediator complex. *eLife*. doi:10.7554/eLife.08719

Roeder, R. G. (1996). The role of general initiation factors in transcription by RNA polymerase II. doi:10.1016/0968-0004(96)10050-5

Sabari, B. R., Dall'Agnese, A., Boija, A., Klein, I. A., Coffey, E. L., Shrinivas, K., ... Young, R. A. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. *Science*, *361*(6400). doi:10.1126/science.aar3958. eprint: https://science.sciencemag.org/content/361/6400/eaar3958.full.pdf

Sanborn, A. L., Rao, S. S., Huang, S. C., Durand, N. C., Huntley, M. H., Jewett, A. I., ... Aiden, E. L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences of the United States of America*. doi:10.1073/pnas.1518552112

Sanyal, A., Lajoie, B. R., Jain, G., & Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, *489*(7414), 109–113. Retrieved from http://dx.doi.org/10.1038/nature11279

Shin, Y. & Brangwynne, C. P. (2017). Liquid phase condensation in cell physiology and disease. *Science*, *357*(6357), eaaf4382. doi:10.1126/science.aaf4382

Shin, Y., Chang, Y.-C., Lee, D. S., Berry, J., Sanders, D. W., Ronceray, P., ... Brangwynne, C. P. (2018). Liquid Nuclear Condensates Mechanically Sense and Restructure the Genome. *Cell*, *175*(6), 1481–1491.e13. doi:10.1016/j.cell.2018.10.057

Shlyueva, D., Stampfel, G., & Stark, A. (2014). Transcriptional enhancers: From properties to genome-wide predictions. doi:10.1038/nrg3682

Stavreva, D. A., Garcia, D. A., Fettweis, G., Gudla, P. R., Zaki, G. F., Soni, V., ... Hager, G. L. (2019). Transcriptional Bursting and Co-bursting Regulation by Steroid Hormone Release Pattern and Transcription Factor Mobility. *Molecular Cell*. doi:10.1016/j.molcel.2019.06.042

Szabo, Q., Bantignies, F., & Cavalli, G. (2019). Principles of genome folding into topologically associating domains. *Science Advances*, *5*(4), eaaw1668. doi:10.1126/sciadv.aaw1668

Takagi, Y. & Kornberg, R. D. (2006). Mediator as a general transcription factor. *Journal of Biological Chemistry*. doi:10.1074/jbc.M508253200

Thompson, C. M., Koleske, A. J., Chao, D. M., & Young, R. A. (1993). A multisubunit complex associated with the RNA polymerase II CTD and TATA-binding protein in yeast. *Cell*. doi:10.1016/0092-8674(93)90362-T

Tunnacliffe, E. & Chubb, J. R. (2020). What Is a Transcriptional Burst? *Trends in Genetics*, 1–10. doi:10.1016/j.tig.2020.01.003

Wong, K. H., Jin, Y., & Struhl, K. (2014). TFIIH Phosphorylation of the Pol II CTD Stimulates Mediator Dissociation from the Preinitiation Complex and Promoter Escape. *Molecular Cell*. doi:10.1016/j.molcel.2014.03.024

Zabidi, M. A. & Stark, A. (2016). Regulatory Enhancer–Core-Promoter Communication via Transcription Factors and Cofactors. doi:10.1016/j.tig.2016.10.003

Chapter 2

ENHANCER SHARING PROMOTES NEIGHBORHOODS OF TRANSCRIPTIONAL REGULATION ACROSS EUKARYOTES

Quintero-Cadena, P. & Sternberg, P. W. (2016). Enhancer Sharing Promotes Neighborhoods of Transcriptional Regulation Across Eukaryotes. *G3 (Bethesda, Md.) 6*(12), 4167–4174. doi:https://doi.org/10.1534/g3.116.036228

ABSTRACT

Enhancers physically interact with transcriptional promoters, looping over distances that can span multiple regulatory elements. Given that enhancer-promoter (EP) interactions generally occur via common protein complexes, it is unclear whether EP pairing is predominantly deterministic or proximity guided. Here we present cross-organismic evidence suggesting that most EP pairs are compatible, largely determined by physical proximity rather than specific interactions. By re-analyzing transcriptome datasets, we find that the transcription of gene neighbors is correlated over distances that scale with genome size. We experimentally show that non-specific EP interactions can explain such correlation, and that EP distance acts as a scaling factor for the transcriptional influence of an enhancer. We propose that enhancer sharing is commonplace among eukaryotes, and that EP distance is an important layer of information in gene regulation.

2.1 Introduction

Enhancers mediate the transcriptional regulation of gene expression, enabling isogenic cells to exhibit remarkable phenotypic diversity (Davidson & Peter, 2015). In complex with transcription factors, they interact with promoters via chromatin looping (Marsman & Horsfield, 2012), finely regulating transcription in time and space. A prevailing view is that most enhancers have a mechanism to selectively loop to a target promoter (van Arensbergen, van Steensel, & Bussemaker, 2014). Examples in this category usually require specific transcription factor binding at both enhancer and promoter sites (Davidson & Peter, 2015), which could explain why some enhancers seem to influence different promoters in varying degrees (Gehrig et al., 2009). On the other hand, EP looping is generally mediated by common protein complexes (Kagey et al., 2010; Malik & Roeder, 2010), conflicting with the specific molecular interactions required by such a model at a larger scale. Examples of non-specific EP pairing also seem to be common (Butler & Kadonaga, 2001). Yet given that this model could result in transcriptional crosstalk, it appears inconsistent with our current paradigm of gene regulation. The predominant EP pairing scheme—specific or non-specific—and its determinants are thus unclear. Here we ask to what extent are pontential EP pairs compatible through a meta-analysis of the genome-wide transcription of gene neighbors in five species. We propose that enhancer sharing occurs widely across eukaryotes, test key aspects of this hypothesis in *C. elegans*, and analyze its implications in other genomic phenomena.

2.2 Materials and Methods

Computational biology

For each analyzed organism, Ensembl (Flicek et al., 2014) protein-coding genes were grouped by chromosome, sorted by position, and paired with the 100 nearest neighbors within the same chromosome. A list of duplicated gene pairs for *H. sapiens* and *M. musculus* was obtained from the Duplicated Genes Database (Ouedraogo et al., 2012) (http://dgd.genouest.org). A list of *C. elegans* genes predicted to be in operons was obtained from Allen, Hillier, Waterston, and Blumenthal, 2011, and gene pairs present in the same operon were removed from the analysis to prevent co-transcriptional bias. Processed RNA-seq data was obtained from multiple sources (Gerstein et al., 2010; Attrill et al., 2016; Ellahi, Thurtle, & Rine, 2015; Consortium, 2012) and converted to transcripts per million (TPM) (Li, Ruotti, Stewart, Thomson, & Dewey, 2010) when necessary. Formatted datasets are available upon request. Genes detected in less than 80% of experiments were discarded. To compute the

Spearman correlation coefficient, TPM values were ranked in each RNA-seq experiment and the pairwise Pearson correlation coefficient was computed on the ranked values according to the following equation:

$$\rho = \frac{cov(\text{gene}_1, \text{gene}_2)}{\sigma_{\text{gene}_1}\sigma_{\text{gene}_2}}$$

where gene_1 and gene_2 are the corresponding ranks of each paired gene in a given RNA-seq experiment, cov their covariance and σ their standard deviation. The list of gene pairs with intergenic distances and correlation coefficients was sorted by increasing intergenic distance, and subsequently smoothed using a sliding median with window size of 1000 gene pairs. The result was then fitted to an exponential decay function of the form:

$$\rho(d) = \rho_0 e^{-\lambda d} + c$$

where ρ_0 is the median Spearman correlation coefficient of the closest neighboring genes, d the intergenic distance and c the baseline correlation. The mean distance at which a pair of genes remain correlated was then computed as:

$$d_{exp} = 1/\lambda$$

To compute the background correlation, each gene was paired with 20 randomly selected genes from a different chromosome and the 95% median confidence interval was computed by bootstrapping with 10,000 samples. A list of genes annotated with RNA *in situ* hybridization data (Tomancak et al., 2007; Hammonds et al., 2013; Tomancak et al., 2002) was obtained from the Berkeley Drosophila Genome Project (http://insitu.fruitfly.org). Insulator ChIP-chip data was obtained from Negre et al., 2010 (GSE16245); the intersection of replicates was used. HiC data was obtained from Rao et al., 2014 (GSE63525, GM12878 primary replicate HiCCUPS looplist). Functional protein classification was conducted using Panther (Mi, Poudel, Muruganujan, Casagrande, & Thomas, 2016). Genomic manipulations were conducted using Bedtools v2.24.0 (Quinlan & Hall, 2010). Data analysis was conducted using Python 2.7.9 and the Scipy library (McKinney, 2010). Plots were generated using Matplotlib 1.5 (Hunter, 2007).

Molecular biology

C. elegans was cultured under standard laboratory conditions (Stiernagle, 2006). For enhancer additivity experiments, transgenic *C. elegans* lines carrying extrachromosomal arrays were generated by injecting each plasmid at 50 ng/μL into

unc-119 mutant animals. The minimal Δ*pes-10* promoter (Fire, Harrison, & Dixon, 1990) and nuclear localized GFP (Lyssenko, Hanna-Rose, & Schlegel, 2007) were used in all constructs. Minimal regions of the *myo-2* and *unc-54* enhancers (Okkema, Harrison, Plunger, Aryana, & Fire, 1993) able to drive tissue specific expression were used. The BWM enhancer was obtained from the upstream region of *F44B9.2*; the BWM/intestine enhancer was obtained from the upstream region of *rps-1*. Animals were imaged at 40x using a GFP filter on a Zeiss Axioskop microscope.

For the enhancer promoter distance and ectopic enhancer experiments, we defined an EP distance of 0 to be the enhancer placed just upstream of the Δ*pes-10* promoter, which is ~350 bp away from the start codon of *gfp*. To ensure neutrality yet maintain a similar GC content as non-coding sequences in *C. elegans*, we used non-overlapping AT-rich DNA spacers obtained from the genome of *Escherichia coli*. Constructs were integrated in single-copy into chromosome IV via CRISPR-Cas9 using plasmids provided as gifts by Dr. Zhiping Wang and Dr. Yishi Jin (unpublished results). Briefly, plasmids containing the following expresssion cassettes were co-injected: reporter and hygromycin resistance genes flanked by homologous arms for recombination-directed repair (10 ng/μL), single-guide RNA (10 ng/μL), Cas9 (10 ng/μL), and extra-chromosomal array reporter for expression of either *rfp* or *gfp* outside the tissue of interest (10 ng/μL). Transformants were selected for using hygromycin at 10 μg/μL, and those not carrying extra-chromosomal transgenes, lacking of *gfp* or *rfp* expression outside the tissue of interest, were subsequently isolated. Animals homozygotic for the insertion were identified by polymerase-chain reaction (PCR) and Sanger sequencing.

Quantitative PCR was carried out as previously described (Ly, Reid, & Snell, 2015) using *pmp-3* as a reference gene (Zhang, Chen, Smith, Zhang, & Pan, 2012). Briefly, third-stage larval (L3) worms, when expression from the test enhancers is maximal according to RNA-seq data, were synchronized at 20° via egg-laying. Fifteen animals were lysed in 1.5 μL of Lysis Buffer (5 mM Tris pH 8.0 (MP Biomedicals), 0.5% Triton X-100, 0.5% Tween 20, 0.25 mM EDTA (Sigma-Aldrich)) with proteinase-K (Roche) at 1.5 μg/μL, incubated at 65° for 10 minutes followed by 85° for one minute. Reverse transcription was carried out using the Maxima H Minus cDNA synthesis kit (Thermo Fisher) by adding 0.3 μL H_2O, 0.6 μL 5x enzyme buffer, 0.15 μL 10mM dNTP mix, 0.15 μL 0.5 μg/μL oligo dT primer, 0.15 μL enzyme mix and 0.15 μL DNAse, and incubated for 2 minutes at 37°, followed by 30 minutes at 50° and finally 2 minutes at 85°. The cDNA solution was diluted to 15 μL and 1 μL was

used for each qPCR reaction, so that on average each well contained the amount of RNA from a single worm. All qPCR reactions were performed with three technical replicates and at least three biological replicates using the Roche LightCycler® 480 SYBR Green I Master in the LightCycler® 480 System. Crossing point-PCR-cycle (Cp) averages were computed for each group of three technical replicates; these values were then substracted from the respective average Cp value of the reference gene.

Data and reagent availability

Strains are available upon request. Relevant DNA sequences, including spacers, enhancers, primers, sgRNA, and homology arms are available in Table S1. Correlation datasets are available in File S1. qPCR data is available in Table S2. Python source code, and links to all expression datasets used in this study, are available for download on the following github repository: https://github.com/WormLabCaltech/QuinteroSternberg2016.git.

2.3 Results and Discussion

Gene neighbors are transcriptionally correlated genome-wide

We reasoned that widespread EP compatibility should result in transcriptional correlation among gene neighbors. Indeed, gene coexpression clusters have been extensively reported in eukaryotic genomes (e.g. Sémon & Duret, September 2006; Roy, Stuart, Lund, & Kim, 2002; Lercher, Urrutia, & Hurst, 2002; Lercher & Hurst, 2006; J. B. E. Williams & Hurst, 2002; Singer, Lloyd, Huminiecki, & Wolfe, 2005; Lercher, Blumenthal, & Hurst, 2003; E. J. Williams & Bowles, 2004; Spellman & Rubin, 2002; Purmann et al., 2007; Zhan, Horrocks, & Lukens, 2006; Boutanaev, Kalmykova, Shevelyov, & Nurminsky, 2002; Kalmykova, Nurminsky, Ryzhov, & Shevelyov, 2005; Caron et al., 2001; Rubin & Green, 2013), in spite of order of magnitude variations in genome size (e.g. ~12 Mb in *S. cerevisiae* vs ~3 Gb in *H. sapiens*). An early informative example is the discovery of chromosomal domains of gene expression in *S. cerevisiae* (Cohen, Mitra, Hughes, & Church, 2000) which exhibit features that strongly support enhancer-sharing, mainly distance-dependence in transcriptional correlation that qualitatively resemble chromosome contact matrices (e.g. Rao et al., 2014), and instances in which a single enhancer seems to be responsible for the coexpression of adjacent gene pairs. The ubiquity of these features across eukaryotes would support the idea that EP interactions are largely determined by physical proximity rather than by specific interactions. Given the

accumulation of transcriptome sequencing data, we decided to investigate the transcriptional correlation of gene neighbors in representative eukaryotic species as a first step to explore the average EP pairing scheme.

We paired every protein-coding gene of five organisms (*Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila melanogaster, Mus musculus* and *Homo sapiens*) with its 100 nearest neighbors within the same chromosome. This yielded lists of around 600,000 (*S. cerevisiae*) and 2 million (each of the rest) gene pairs. We then computed the Spearman correlation coefficient between paired genes across multiple RNA-seq datasets (Gerstein et al., 2010; Attrill et al., 2016; Ellahi et al., 2015; Consortium, 2012) and the intergenic distance between the the start of the 5' untranslated region of the first gene to the start of the second gene in each pair.

We observed that neighboring genes tend to be correlated in transcript abundance genome-wide in all analyzed organisms, and that this correlation decays exponentially with increasing intergenic distance (Figure 2.1a). We thus fitted the data to an exponential decay function to estimate the distance at which a pair of genes remain correlated (d_{exp}). Consistent with the persistence of the correlation pattern across organisms, d_{exp} scaled with genome size, to 1 kilobase in *S. cerevisiae*, ~10 kb in *C. elegans* and *D. melanogaster*, and ~350 kb in *M. musculus* and in *H. sapiens* (Figure 2.4). This trend remained largely the same even after removing duplicated genes pairs (Figure 2.5). Most genes had at least one neighbor closer than d_{exp} in all species (Figure 2.1b), and the representation of gene ontology annotations remained unbiased in correlated gene pairs (Figure 2.6), indicating that the average gene is correlated in expression with its nearest neighbors beyond any particular gene class. In addition, sampled intergenic distances go well beyond d_{exp} (Figure 2.1c), indicating that 100 gene neighbors is a sufficient number to study this effect.

To examine the correlation of gene expression in the spatial domain, we analyzed RNA *in situ* hybridization data for 6053 genes in *D. melanogaster* (Tomancak et al., 2002; Tomancak et al., 2007; Hammonds et al., 2013). We computed the percentage overlap in tissue expression by dividing the number of common tissues over the total number of unique tissues in which genes of any given pair are expressed (Figure 2.7a). This analysis revealed that close neighbors have a tendency to be expressed in the same tissues, and that this overlap also decays exponentially with intergenic distance (Figure 2.7b). However, the correlation extends to a longer mean distance (d_{exp} = 22 compared to 6 kb), suggesting that RNA-seq analysis, which included mostly whole-organism transcriptome averages, resulted in a conservative estimate.

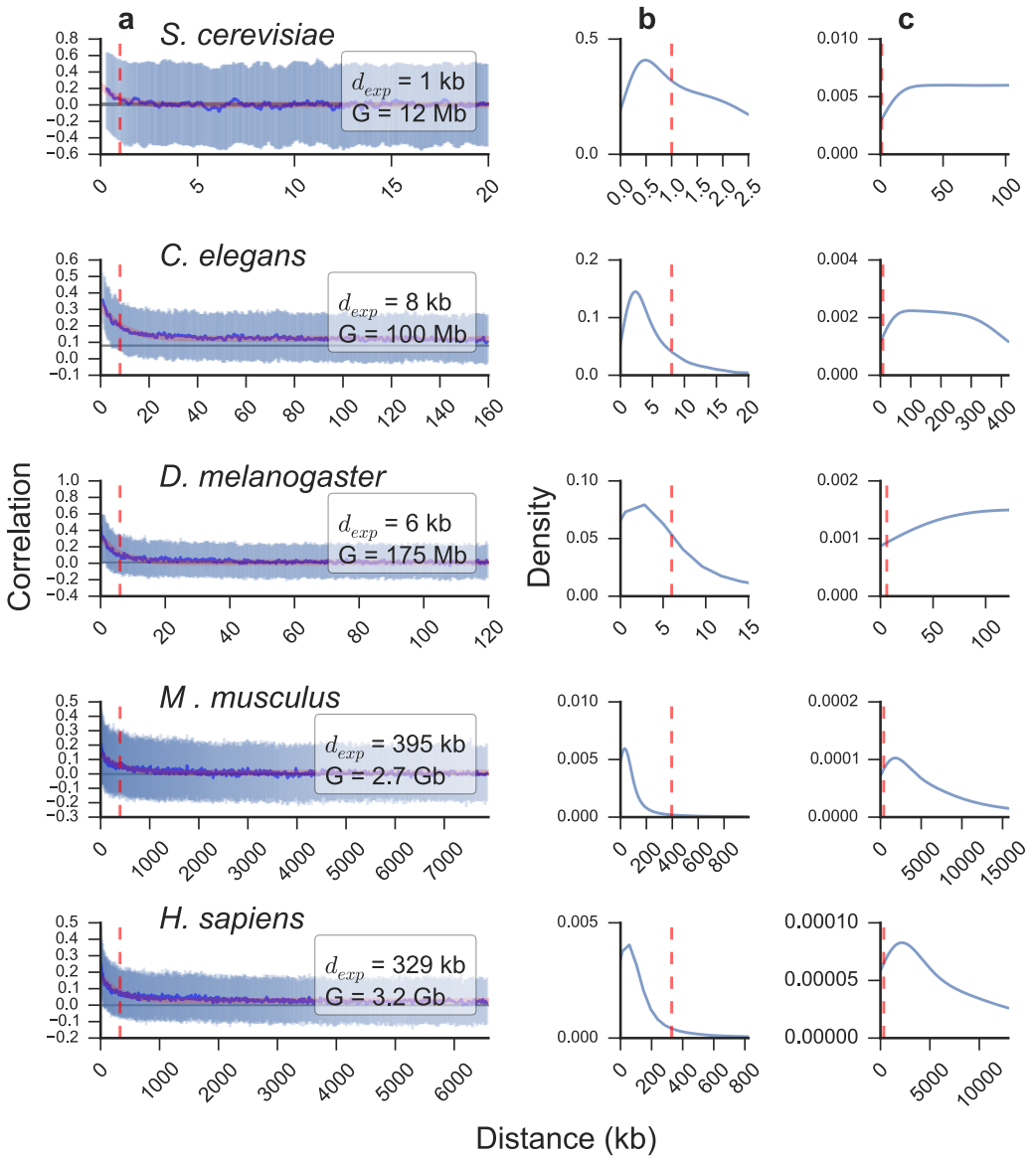

Figure 2.1: **Neighboring genes are transcriptionally correlated genome-wide across eukaryotes.** a) Sliding median of correlations between paired neighbors (blue line) and interquartile range (pale blue) with increasing intergenic distance. Median ± 95% confidence interval of randomly paired genes is shown as a horizontal gray line. Fit to an exponential decay function (red line) was used to compute the mean distance at which gene neighbors remain correlated (d_{exp}, vertical red dashed line). The genome size (G) is displayed for each organism. Distribution of intergenic distances between each gene and its nearest neighbor (b) and all paired genes (c). The organism analyzed in each case is indicated for each group of three subplots.

Given that pairing every gene with 100 proximal genes provides a complete set of distance-dependent correlations between gene pairs, we concluded that gene neigh-

bors have a spatio-temporal correlation in expression that is highly dependent upon the spacing between them. Our meta-analysis unifies the findings of previous reports (reviewed in Michalak, 2008) and highlights the distance-dependence of genome-wide and cross-organismic transcriptional correlations that transcend localized gene coexpression clusters.

Enhancer sharing explains the transcriptional correlation of gene neighbors

The pervasive nature of proximal gene coexpression supported the idea of widespread EP compatibility. This connection is, in turn, supported by several other observations in literature: i) enhancers regulate transcription by making contact with promoters via chromatin looping (Marsman & Horsfield, 2012), whose incidence also decays exponentially as the distance between contacting sites increases (Ringrose, Chabanis, Angrand, Woodroofe, & Stewart, 1999; Rao et al., 2014), with the same pattern as observed here at least in some documented cases (e.g. *H. sapiens*, Figure 2.8) ii) the average distance between a large fraction of studied EP interactions scales with genome size in ranges often consistent with d_{exp}: < 1 kb in *S. cerevisiae*, (Dobi & Winston, 2007); < 10 kb in *C. elegans*, (Araya et al., 2014); and 120 kb in *H. sapiens*, (Sanyal, Lajoie, Jain, & Dekker, 2012) iii) common protein complexes such as the mediator seem to be widely utilized bridges in EP looping (Kagey et al., 2010; Malik & Roeder, 2010) iv) a high frequency of chromatin interactions are observed within topologically associated domains identified through high-resolution Chromosome Conformation Capture (Hi-C) (Rao et al., 2014) and v) studied enhancers often do not show promoter specificity (Butler & Kadonaga, 2001). This line of reasoning suggests a model where, as opposed to only having a specific target gene (Figure 2.2a), the average enhancer has a range of action in which it can influence any active promoter within its reach (Figure 2.2b). A concrete example consistent with this idea is the upregulation of neighboring genes upon enhancer activation by fibroblast growth factor in mammalian cells (Ebisuya, Yamamoto, Nakajima, & Nishida, 2008). Transcriptome analysis could thus provide indirect evidence of genome and condition-wide EP looping that is difficult to access through Hi-C (Rao et al., 2014) due to the low signal-to-noise ratio of short-range interactions.

Because of its compact genome, rapid development and availability of tissue specific enhancers (Corsi, Wightman, & Chalfie, 2015), we decided to use *C. elegans* to test the validity of a non-specific EP pairing model. We first postulated that unrelated enhancers should generally be compatible, showing qualitative additivity when placed upstream of a single promoter. We thus paired the well characterized *myo-2*

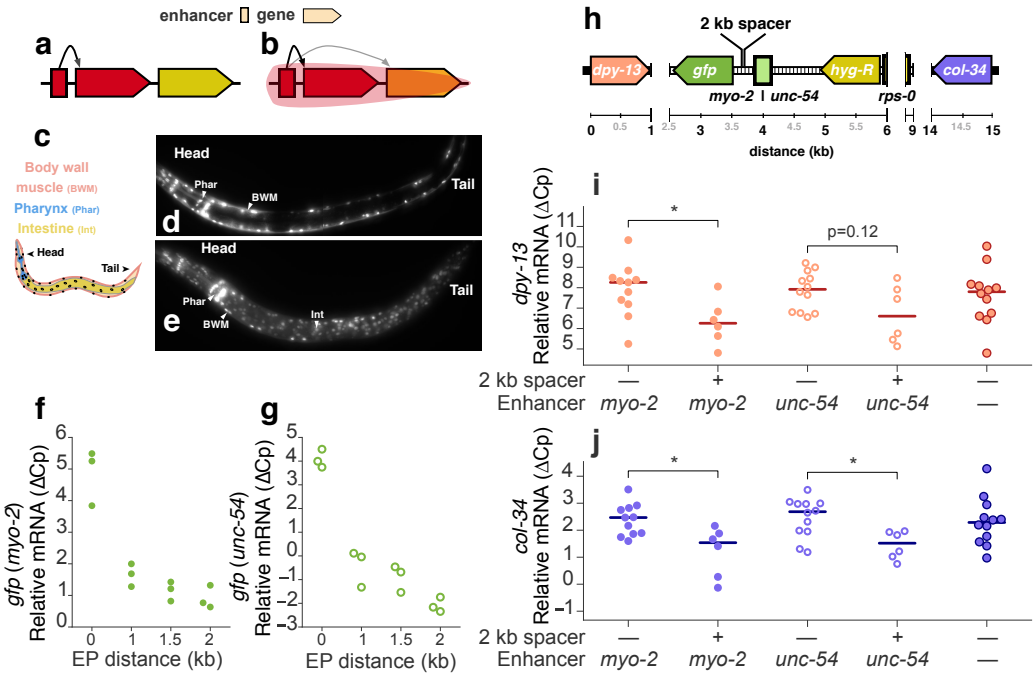

Figure 2.2: **Enhancer sharing explains the transcriptional correlation of gene neighbors.** Two possible models for EP relationship: a) Enhancers have specific target genes and b) enhancers have a range of action in which they influence genes by physical proximity. Tissue specific enhancers (c) are generally compatible. Pharynx and body wall muscle (d) and pharynx, body wall muscle, and intestine (e) enhancers driving nuclear *gfp* expression. mRNA levels of *gfp* with increasing EP distance for lines with *myo-2* (filled circles, f) and *unc-54* (hollow circles, g) enhancers. h) Genomic context of the integration site. The inserted construct is shown over a dashed black line and includes a hygromycin resistance gene (*hyg-R*) regulated by a ribosomal enhancer (*rps-0*) and promoter in addition to the reporter (*gfp*) regulated by either the *myo-2* or *unc-54* enhancers; the native genes *dpy-13* and *col-34* flank the insertion site. Relative mRNA levels of *dpy-13* (i) and *col-34* (j) in wild-type and lines with and without the 2 kb spacer (*two tailed P-val<0.05, Mann-Whitney U test). The difference in crossing point-PCR-cycle (ΔCp) with the reference gene *pmp-3* and the corresponding median for each group of biological replicates is shown for every qPCR experiment.

pharyngeal enhancer with a body wall muscle (BWM) and a BWM plus intestine enhancer, placed them upstream of a minimal promoter and a *gfp* reporter, and examined expression in transgenic animals. In both cases, we observed fluorescence in the corresponding tissues (Figure 2.2c, d, e). This observation is consistent with typical enhancer studies in artificial constructs (Dupuy et al., 2004) and agrees with some EP compatibility studies (Butler & Kadonaga, 2001).

Given that both chromatin looping and expression correlation decay exponentially, we reasoned that transcription of a given gene should also decay exponentially with increasing EP distance if the observed correlation is to be explained by enhancer sharing. To test this hypothesis, we first built a series of genetic constructs with increasing neutral EP distances (0, 1, 1.5, and 2 kb) for two different enhancers, *myo-2* and *unc-54* (~400 and 300 bp, respectively). We then integrated each construct in single copy into the genome of *C. elegans* and used quantitative PCR to i) measure the influence of EP distance on the reporter gene in native chromatin and ii) analyze the impact of the perturbation on the two genes that natively flank the site of transgene insertion (*dpy-13* and *col-34*, Figure 2.2h), which we reasoned should be affected in two counteracting ways. First, the ectopic enhancers should promote their expression. Second, the increased EP distance caused by the addition of spacers should reduce their expression by scaling down the influence of both ectopic and native enhancers (the latter of unknown identity and location) to the opposite side of the spacer.

We found that transcriptional levels of the reporter gene indeed fall rapidly with increasing EP distance with both enhancers (Figure 2.2f, g); this occurred at a rate that seems congruent or faster than the genome-wide correlation decay, likely because of the dramatic separation of every regulatory element at once, as opposed to gradual separation from individual enhancers in a native environment; this dramatic effect suggests complex interactions between multiple EP loops that are disrupted with the insertion of DNA sequences devoid of regulatory elements. Transcription was still well detected even when the enhancers were placed 2 kb away, supporting the hypothesis that EP distance is a scaling factor on the enhancer's influence. Expression of *dpy-13* and *col-34* was reduced with the introduction of the 2 kb spacer when compared to transgenic lines without it (Figure 2.2i, j). On the other hand, spacer-free lines were comparable to wild-type, suggesting the incorporation of ectopic enhancers compensated for the EP distance increase caused by the addition of the genetic construct itself. These observations seem to fit the corollaries of our model, even amid the complexity of a native regulatory environment. However, the distance over which we see an effect on *col-34* falls outside our d_{exp} estimate for *C. elegans* (8 kb). Its expression is impacted by the presence of the 2 kb spacer outside of the interval between the *myo-2*/*unc-54* enhancer, suggesting that enhancers >12 kb away can still influence its expression. As evidenced with the discrepancy in *D. melanogaster* when using *in situ* or RNA-seq data, this observation suggests that d_{exp} is only a rough estimate of the average enhancer rage of action; this is

useful to gain insight into genome-wide mechanisms but not for precise individual predictions.

Chromatin modifications have been shown to have a significant impact on enhancer function (Calo & Wysocka, 2013) and thus likely influence EP pairing. Thus, chromatin features and enhancer sharing might be mutually inclusive rather than stand alone explanations for the observed correlation domains. From this perspective, transcriptionally correlated genes would have similar chromatin states, facilitated by their physical proximity, that make them accessible to enhancer action.

The existence of multiple, independent but similar enhancers is an alternative possible explanation. However, since we are looking at genome-wide averages, this would mean that most gene neighbors have a functionally redundant set of independent enhancers that function through distinct molecular interactions. Although possible, this is a rather intricate explanation.

In agreement with the enhancer sharing hypothesis, it can be argued that the scaling of correlation domains is a result of the ability to connect EP loops over longer distances in larger genomes. Yet in spite of having a compact genome, *D. melanogaster* is able to form many long-range EP interactions (> 50 kb) (Ghavi-Helm et al., 2014), which is considerably different to the range of its estimated d_{exp} (6-22 kb). Additionally, these long-range interactions appear to be particularly specific, with enhancers selectively activating their target promoters (Ghavi-Helm et al., 2014; Kwon et al., 2009). It is thus possible that in compact genomes, long-range EP interactions would need to be specific, whereas nearby interactions would tend to fall in the non-specific pairing scheme, ultimately resulting in the observed correlation domain size.

Enhancer-promoter distance insulates neighboring genes

We next wished to determine the extent to which enhancer sharing could explain other genomic phenomena. Previous reports have suggested that divergent, parallel and convergent gene pairs tend to have distinct correlation profiles (e.g. Chen & Stein, 2006). To explore this hypothesis, we compared the distribution of intergenic distances of gene pairs oriented in parallel, divergent and convergent orientations (Figure 2.3a, Figure 2.9). As expected, divergent gene pairs tend to be closest, followed by parallel and finally convergent genes. We then confirmed that each group appears to have different distributions of correlations *D. melanogaster* (Figure 2.3b, Figure 2.9). To consider the influence of EP distance, we sampled gene pairs from each orientation controlling for intergenic size. This resulted in distributions

of correlations that exactly overlap (Figure 2.3c, Figure 2.9), an observation that is supported by previous reports in specific cases (Ghanbarian & Hurst, 2015; Cohen et al., 2000). We thus conclude that the apparent influence of gene orientation in the transcriptional relationship of neighboring gene pairs is consistent with the enhancer sharing hypothesis. In this scenario, the effect of gene orientation can be simply explained by the different EP distance distributions for each configuration.

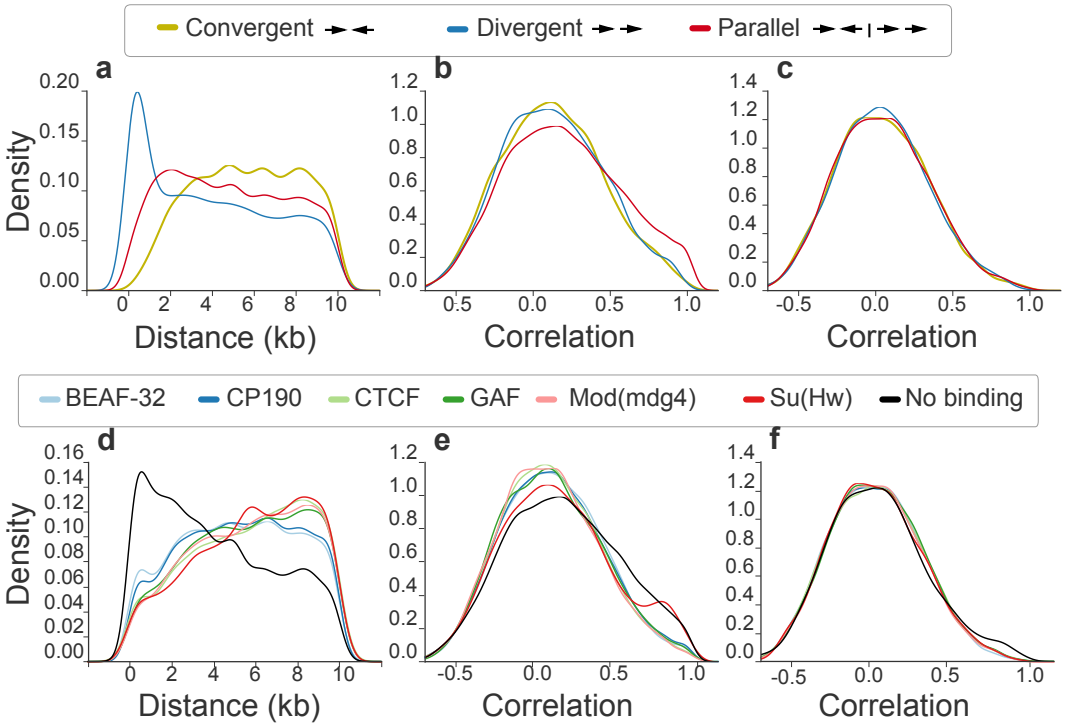

Figure 2.3: **EP distance causes gene orientation-dependent correlation and provides regulatory independence to gene neighbors.** Distribution of intergenic distances below 10 kb of gene pairs in *D. melanogaster* by configuration (~5 to 18 thousand gene pairs for each group, a) and flanking insulator binding sites identified through ChIP-chip (Negre et al., 2010) (~5 to 15 thousand pairs for each group, d). The corresponding distribution of correlations is shown for the same gene pairs (b, e) and pairs with controlled distributions of intergenic distances between 30 and 40 kb (~7 to 14 thousand pairs for gene orientation groups, ~10 to 18 thousand for insulator groups, c, f).

From the regulatory perspective, EP distance provides independence to most gene pairs, as the vast majority have an intergenic distance that puts them in the baseline correlation regime (Figure 2.1c). To study the enhancer-blocking influence of insulators (Bushey, Dorman, & Corces, 2009) genome-wide, we analyzed each group of genes flanked by insulator binding sites, which were previously determined using Chromatin ImmunoPrecipitation coupled with microarrays (ChIP-chip) for six

known insulators in *D. melanogaster*: BEAF-32, CP190, CTCF, GAF, Mod(mdg4) and Su(Hw) (Negre et al., 2010). We observed that gene neighbors closer than 10 kb bound by each of the insulators tend to be less correlated in gene expression than gene pairs not bound by them (Figure 2.3e), supporting their role in genome-wide insulation and agreeing with the observation that insulators tend to separate differentially expressed genes (Negre et al., 2010; Xie et al., 2007). Nevertheless, the same groups of gene pairs also tend to have much larger intergenic distances than genes that are not flanked by insulator binding sites (Figure 2.3d). After controlling for the distribution of intergenic distances, we found very similar correlation distributions between insulator and not insulator flanked gene pairs (Figure 2.3f). This finding agrees with previous reports suggesting that insulators do not block enhancers everywhere they bind, but rather act only on very specific genomic contexts (Schwartz et al., 2012; Liu et al., 2015; Ong & Corces, 2014); it also reconciles the lack of known insulator orthologs in *C. elegans* (Heger, Marin, & Schierenberg, 2009) in the context of local enhancer-blocking. In combination, these studies strongly suggest that EP distance is the general source of transcriptional independence for close gene neighbors.

Previous EP compatibility studies suggest that EP specificity is widespread (Gehrig et al., 2009), while others that it is restricted to a smaller subset of enhancers (Butler & Kadonaga, 2001). Although our results support the latter, views arising from these studies might not be mutually exclusive, as it is likely that enhancers have specificity to promoter classes (Danino, Even, Ideses, & Juven-Gershon, 2015), whose limited number could result in general EP compatibility.

The implications from considering our observations are broadly applicable to gene regulation. Position effects, in which transgene expression levels are influenced by the insertion site (Gierman et al., 2007), are naturally expected from enhancer sharing. Chromosomal translocations and mutations involving regulatory elements likely impact genetic contexts rather than individual genes. Furthermore, enhancer sharing and distance-based scaling of enhancer influence potentially provides an additional layer of information in gene regulation, as the transcriptional output of a given gene would be the result of scaled contributions from multiple shared enhancers. Such a feature could by itself be under selective pressure, leading to a roughly constant size of the correlation domain in number of genes regardless of absolute physical distance, as observed in this study. Our analysis provides a clarifying perspective of gene regulation consistent with both mechanistic and

genome-wide studies.

2.4 Acknowledgments

Our work was supported by the Howard Hughes Medical Institute, of which P.W.S is an investigator. We thank Zhiping Wang and Yishi Jin for plasmids for Crispr-Cas9 single copy insertion, Carmie Robinson for discussions, experimental suggestions and comments on the manuscript, Han Wang for discussions, technical advise and comments on the manuscript, Hillel Schwartz, Mitchell Guttman, Mihoko Kato, David Angeles-Albores, Jonathan Liu, Barbara Wold, Isabelle Peter, and Angelike Stathopoulos for discussions and comments on the manuscript, the Encode and ModEncode consortiums, FlyBase, WormBase, and Ensembl databases, the Wold Lab and the Guigo Lab for data accessibility.

2.5 Contributions

P.Q.C. performed the experiments and analyzed the data. P.Q.C. and P.W.S. designed the experiments and wrote the paper.

2.6 Competing financial interests

The authors declare no competing financial interests.

2.7 Supplementary Figures

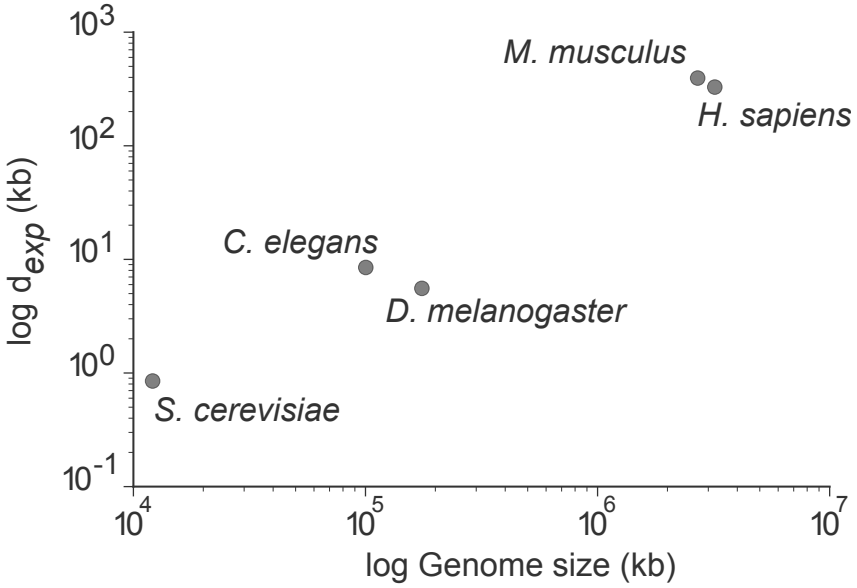

Figure 2.4: **The distance at which a pair of genes remain correlated (d_{exp}) scales with genome size.**

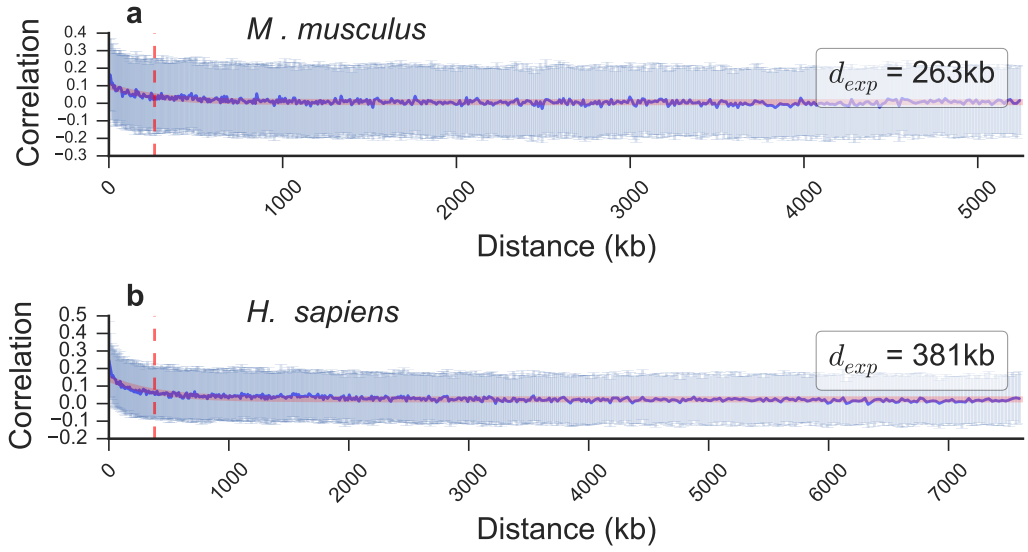

Figure 2.5: **Removing duplicated genes does not affect the overall correlation of gene neighbors.** Sliding median of correlations between paired neighbors (blue line) and interquartile range (pale blue) with increasing intergenic distance in *M. musculus* (a) and *H. sapiens* (b) after removing duplicated gene pairs. Fit to exponential decay function (red line) and corresponding d_{exp} (red dashed line) are shown.

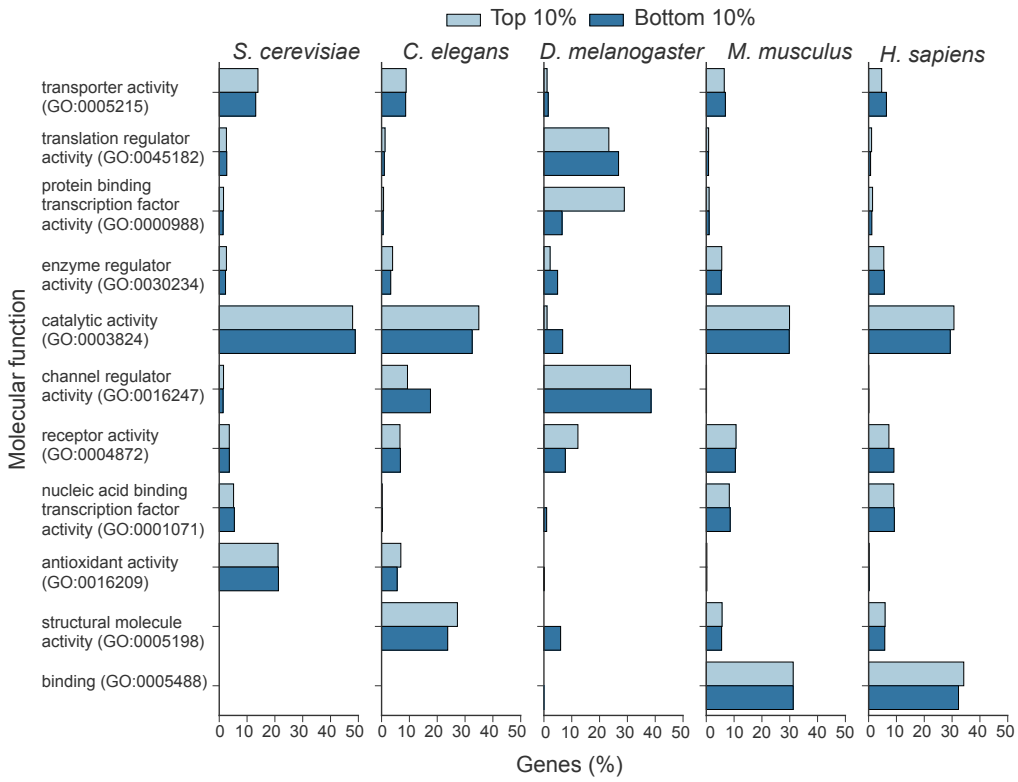

Figure 2.6: **Representation of gene ontology annotations remains unbiased in correlated gene pairs.** The molecular function classification of top and bottom 10% correlated gene pairs with intergenic distance below d_{exp} is shown for each organism.

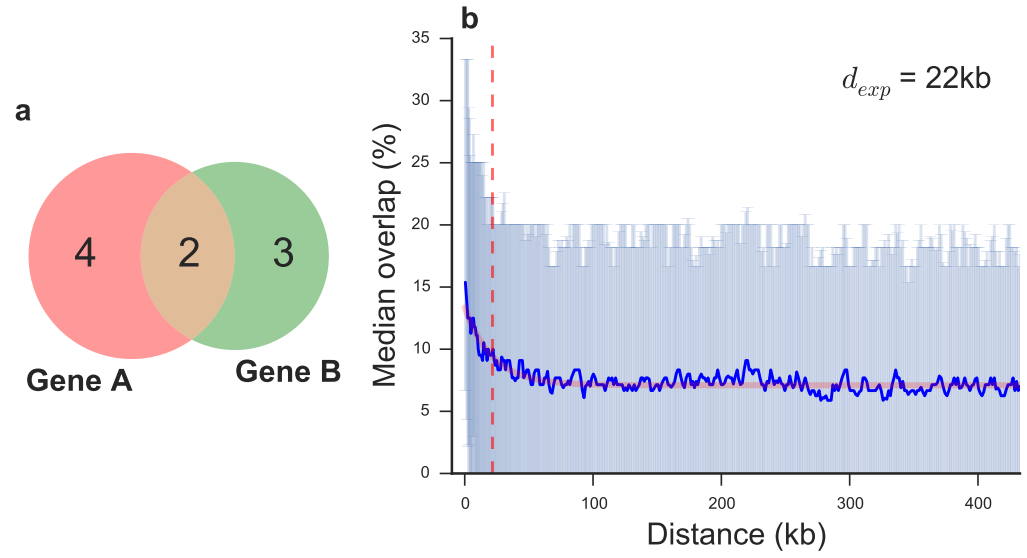

Figure 2.7: **Gene pairs are correlated in spatial expression in *D. melanogaster*.** The size of the intersection between the set of tissues in which each gene of a given pair is expressed was divided over the size of the union of the same sets. An example is shown in (a), where the percentage overlap is 2/(4+3-2)=0.4. b) Sliding median of the percentage overlap in tissue specific expression (blue line) and interquartile range (pale blue) with increasing intergenic distance. Fit to exponential decay function (red line) and corresponding d_{exp} (red dashed line) are shown.

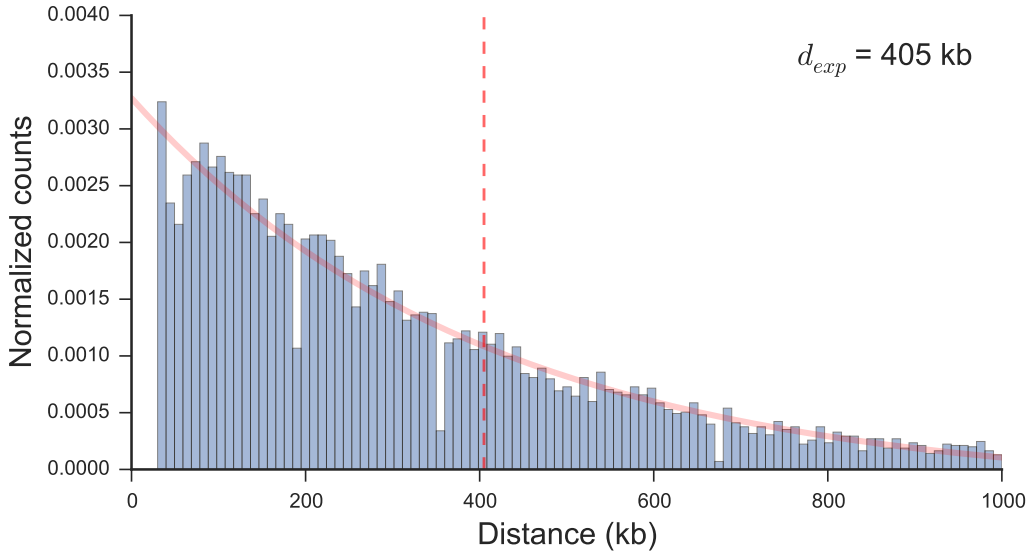

Figure 2.8: **Chromatin looping decreases exponentially with distance in human cell lines.** Normalized count of loops identified through HiC by Rao *et al. 2014* were fit to exponential decay function (red line); the resulting d_{exp} (red dashed line) is shown.

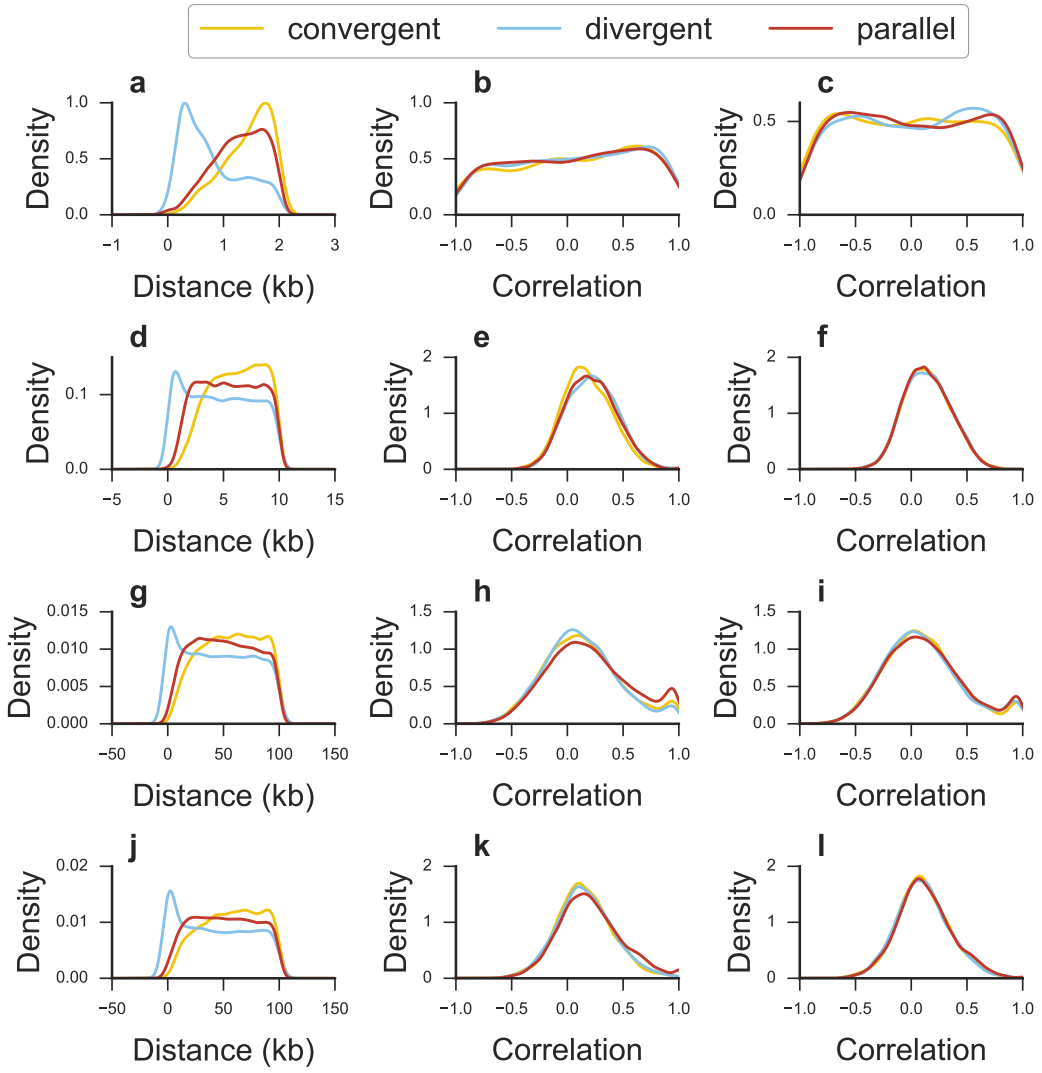

Figure 2.9: **Gene orientation effect in correlation of gene pairs is explained by EP distance.** Distribution of intergenic distances and the corresponding distribution of correlations of gene pairs is shown in the first and second columns, respectively; correlations after controlling for intergenic distance are shown in the third column. The range of distances between paired genes for each plot is as follows: *S. cerevisiae* below 2 kb (a,b) and between 2 and 4 kb (c). *C. elegans* below 10 kb (d,e) and between 10 and 20 kb (f). *H. sapiens* and *M. musculus* below 100 kb (g, h, j, k) and between 100 and 200 kb (i, l).

BIBLIOGRAPHY

Allen, M. A., Hillier, L. W., Waterston, R. H., & Blumenthal, T. (2011). A global analysis of C. elegans trans-splicing. *Genome Res. 21*, 255–264. doi:10.1101/gr.113811.110.The

Araya, C. L., Kawli, T., Kundaje, A., Jiang, L., Wu, B., Vafeados, D., . . . Snyder, M. (2014). Regulatory analysis of the C. elegans genome with spatiotemporal resolution. *Nature*, *512*(7515), 400–405. doi:10.1038/nature13497

Attrill, H., Falls, K., Goodman, J. L., Millburn, G. H., Antonazzo, G., Rey, A. J., . . . the FlyBase consortium. (2016). Flybase: establishing a gene group resource for drosophila melanogaster. *Nucleic Acids Res. 44*(D1), D786–D792. doi:10.1093/nar/gkv1046

Boutanaev, A. M., Kalmykova, A. I., Shevelyov, Y. Y., & Nurminsky, D. I. (2002). Large clusters of co-expressed genes in the drosophila genome. *Nature*, *420*(6916), 666–669. Retrieved from http://dx.doi.org/10.1038/nature01216

Bushey, A. M., Dorman, E. R., & Corces, V. G. (2009). Chromatin insulators:regulatory mechanisms and epigenetic inheritance. *Mol. Cell*, *32*(404), 1–9. doi:10.1016/j.molcel.2008.08.017.Chromatin

Butler, J. E. & Kadonaga, J. T. (2001). Enhancer–promoter specificity mediated by dpe or tata core promoter motifs. *Genes & Development*, *15*(19), 2515–2519. doi:10.1101/gad.924301. eprint: http://genesdev.cshlp.org/content/15/19/2515.full.pdf+html

Calo, E. & Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? *Molecular Cell*, *49*(5), 825–837. doi:http://dx.doi.org/10.1016/j.molcel.2013.01.038

Caron, H., Schaik, B. v., Mee, M. v. d., Baas, F., Riggins, G., Sluis, P. v., . . . Versteeg, R. (2001). The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, *291*(5507), 1289–1292. doi:10.1126/science.1056794. eprint: http://science.sciencemag.org/content/291/5507/1289.full.pdf

Chen, N. & Stein, L. D. (2006). Conservation and functional significance of gene topology in the genome of Caenorhabditis elegans. *Genome Res. 16*(5), 606–617. doi:10.1101/gr.4515306

Cohen, B. A., Mitra, R. D., Hughes, J. D., & Church, G. M. (2000). A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet*, *26*(2), 183–186. Retrieved from http://dx.doi.org/10.1038/79896

Consortium, T. E. P. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, *489*(7414), 57–74.

Corsi, A. K., Wightman, B., & Chalfie, M. (2015). A transparent window into biology: a primer on caenorhabditis elegans. *Genetics*, *200*(2), 387–407. doi:10.1534/genetics.115.176099

Danino, Y. M., Even, D., Ideses, D., & Juven-Gershon, T. (2015). The core promoter: at the heart of gene expression. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1849*(8), 1116–1131. doi:http://dx.doi.org/10.1016/j.bbagrm.2015.04.003

Davidson, E. H. & Peter, I. S. (2015). Chapter 1 - the genome in development. In E. H. Davidson & I. S. Peter (Eds.), *Genomic control process* (pp. 1–40). Oxford: Academic Press. doi:http://dx.doi.org/10.1016/B978-0-12-404729-7.00001-0

Dobi, K. C. & Winston, F. (2007). Analysis of transcriptional activation at a distance in saccharomyces cerevisiae. *Molecular and Cellular Biology*, *27*(15), 5575–5586. doi:10.1128/MCB.00459-07. eprint: http://mcb.asm.org/content/27/15/5575.full.pdf+html

Dupuy, D., Li, Q.-R., Deplancke, B., Boxem, M., Hao, T., Lamesch, P., . . . Vidal, M. (2004). A first version of the caenorhabditis elegans promoterome. *Genome Res. 14*(10b), 2169–2175. doi:10.1101/gr.2497604

Ebisuya, M., Yamamoto, T., Nakajima, M., & Nishida, E. (2008). Ripples from neighbouring transcription. *Nat Cell Biol*, *10*(9), 1106–1113. Retrieved from http://dx.doi.org/10.1038/ncb1771

Ellahi, A., Thurtle, D. M., & Rine, J. (2015). The chromatin and transcriptional landscape of native saccharomyces cerevisiae telomeres and subtelomeric domains. *Genetics*, *2*, 505–521. doi:10.1534/genetics.115.175711

Fire, A., Harrison, S. W., & Dixon, D. (1990). A modular set of lacz fusion vectors for studying gene expression in caenorhabditis elegans. *Gene*, *93*(2), 189–198.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., . . . Searle, S. M. (2014). Ensembl 2014. *Nucleic Acids Res. 42*(D1), D749–D755. doi:10.1093/nar/gkt1196

Gehrig, J., Reischl, M., Kalmar, E., Ferg, M., Hadzhiev, Y., Zaucker, A., . . . Muller, F. (2009). Automated high-throughput mapping of promoter-enhancer interactions in zebrafish embryos. *Nat Meth*, *6*(12), 911–916. Retrieved from http://dx.doi.org/10.1038/nmeth.1396

Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., . . . Waterston, R. H. (2010). Integrative analysis of the caenorhabditis elegans genome by the modencode project. *Science*, *330*(6012), 1775–1787. doi:10.1126/science.1196914

Ghanbarian, A. T. & Hurst, L. D. [Laurence D]. (2015). Neighboring genes show correlated evolution in gene expression. *Mol. Biol. Evol. 32*(7), 1748–1766. doi:10.1093/molbev/msv053

Ghavi-Helm, Y., Klein, F. A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., & Furlong, E. E. M. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*, *512*(7512), 96–100. doi:10.1038/nature13417

Gierman, H. J., Indemans, M. H., Koster, J., Goetze, S., Seppen, J., Geerts, D., ... Versteeg, R. (2007). Domain-wide regulation of gene expression in the human genome. *Genome Res. 17*(9), 1286–1295. doi:10.1101/gr.6276007

Hammonds, A. S., Bristow, C. A., Fisher, W. W., Weiszmann, R., Wu, S., Hartenstein, V., ... Celniker, S. E. (2013). Spatial expression of transcription factors in drosophila embryonic organ development. *Genome Biol. 14*(12), R140. doi:10.1186/gb-2013-14-12-r140

Heger, P., Marin, B., & Schierenberg, E. (2009). Loss of the insulator protein CTCF during nematode evolution. *BMC Mol. Biol. 5*, 1–14. doi:10.1186/1471-2199-10-84

Hunter, J. D. (2007). Matplotlib: a 2d graphics environment. *Computing In Science & Engineering*, *9*(3), 90–95. doi:10.1109/MCSE.2007.55

Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., ... Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, *467*(7314), 430–435.

Kalmykova, A. I., Nurminsky, D. I., Ryzhov, D. V., & Shevelyov, Y. Y. (2005). Regulated chromatin domain comprising cluster of co-expressed genes in drosophila melanogaster. *Nucleic Acids Research*, *33*(5), 1435–1444. doi:10.1093/nar/gki281. eprint: http://nar.oxfordjournals.org/content/33/5/1435.full.pdf+html

Kwon, D., Mucci, D., Langlais, K. K., Americo, J. L., DeVido, S. K., Cheng, Y., & Kassis, J. A. (2009). Enhancer-promoter communication at the drosophila engrailed locus. *Development (Cambridge, England)*, *136*(18), 3067–3075. doi:10.1242/dev.036426

Lercher, M. J., Blumenthal, T., & Hurst, L. D. (2003). Coexpression of neighboring genes in caenorhabditis elegans is mostly due to operons and duplicate genes. *Genome Research*, *13*(2), 238–243. doi:10.1101/gr.553803. eprint: http://genome.cshlp.org/content/13/2/238.full.pdf+html

Lercher, M. J. & Hurst, L. D. [Laurence D.]. (2006). Co-expressed yeast genes cluster over a long range but are not regularly spaced. *Journal of Molecular Biology*, *359*(3), 825–831. doi:http://dx.doi.org/10.1016/j.jmb.2006.03.051

Lercher, M. J., Urrutia, A. O., & Hurst, L. D. (2002). Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet*, *31*(2), 180–183. Retrieved from http://dx.doi.org/10.1038/ng887

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., & Dewey, C. N. (2010). Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, *26*(4), 493–500. doi:10.1093/bioinformatics/btp692. eprint: http://bioinformatics.oxfordjournals.org/content/26/4/493.full.pdf+html

Liu, M., Maurano, M. T., Wang, H., Qi, H., Song, C.-z., Navas, P. A., ... Stamatoyannopoulos, G. (2015). Genomic discovery of potent chromatin insulators for human gene therapy. *Nat. biotechnol. 33*(2), 198–203. doi:10.1038/nbt.3062

Ly, K., Reid, S. J., & Snell, R. G. (2015). Rapid RNA analysis of individual Caenorhabditis elegans. *MethodsX*, *2*, 59–63. doi:10.1016/j.mex.2015.02.002

Lyssenko, N. N., Hanna-Rose, W., & Schlegel, R. A. (2007). Cognate putative nuclear localization signal effects strong nuclear localization of a gfp reporter and facilitates gene expression studies in caenorhabditis elegans. *Biotechniques*, *43*(5), 596–600.

Malik, S. & Roeder, R. G. (2010). The metazoan mediator co-activator complex as an integrative hub for transcriptional regulation. *Nat. Rev. Genet. 11*(11), 761–772.

Marsman, J. & Horsfield, J. A. (2012). Long distance relationships: enhancer–promoter communication and dynamic gene transcription. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1819*(11–12), 1217–1227. doi:http://dx.doi.org/10.1016/j.bbagrm.2012.10.008

McKinney, W. (2010). Data structures for statistical computing in python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th python in science conference* (pp. 51–56).

Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2016). Panther version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res. 44*(D1), D336–D342.

Michalak, P. (2008). Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics*, *91*(3), 243–248. doi:http://dx.doi.org/10.1016/j.ygeno.2007.11.002

Negre, N., Brown, C. D., Shah, P. K., Kheradpour, P., Morrison, C. A., Henikoff, S., ... White, K. P. (2010). A Comprehensive Map of Insulator Elements for the Drosophila Genome. *Plos Genet. 6*(1), e1000814. doi:10.1371/journal.pgen.1000814

Okkema, P. G., Harrison, S. W., Plunger, V., Aryana, A., & Fire, A. (1993). Sequence Requirements for Myosin Gene Expression and Regulation in C elegans. *Genetics*, *135*(Waterston 1988), 385–404.

Ong, C.-T. & Corces, V. G. (2014). Ctcf: an architectural protein bridging genome topology and function. *Nat Rev Genet*, *15*(4), 234–46. doi:10.1038/nrg3663

Ouedraogo, M., Bettembourg, C., Bretaudeau, A., Sallou, O., Diot, C., Demeure, O., & Lecerf, F. (2012). The duplicated genes database: identification and functional annotation of co-localised duplicated genes across genomes. *PLoS ONE*, *7*(11), 1–8. doi:10.1371/journal.pone.0050653

Purmann, A., Toedling, J., Schueler, M., Carninci, P., Lehrach, H., Hayashizaki, Y., ... Sperling, S. (2007). Genomic organization of transcriptomes in mammals: coregulation and cofunctionality. *Genomics*, *89*(5), 580–587. doi:http://dx.doi.org/10.1016/j.ygeno.2007.01.010

Quinlan, A. R. & Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. doi:10.1093/bioinformatics/btq033

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., ... Aiden, E. L. (2014). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680. doi:10.1016/j.cell.2014.11.021

Ringrose, L., Chabanis, S., Angrand, P. O., Woodroofe, C., & Stewart, A. F. (1999). Quantitative comparison of dna looping in vitro and in vivo: chromatin increases effective dna flexibility at short distances. *EMBO J. 18*(23), 6630–6641. doi:10.1093/emboj/18.23.6630

Roy, P. J., Stuart, J. M., Lund, J., & Kim, S. K. (2002). Chromosomal clustering of muscle-expressed genes in caenorhabditis elegans. *Nature*, *418*(6901), 975–979. Retrieved from http://dx.doi.org/10.1038/nature01012

Rubin, A. F. & Green, P. (2013). Expression-based segmentation of the drosophila genome. *BMC Genomics*, *14*(1), 1–8. doi:10.1186/1471-2164-14-812

Sanyal, A., Lajoie, B. R., Jain, G., & Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, *489*(7414), 109–113. Retrieved from http://dx.doi.org/10.1038/nature11279

Schwartz, Y. B., Linder-basso, D., Kharchenko, P. V., Tolstorukov, M. Y., Kim, M., Li, H.-b., ... Karpen, G. H. (2012). Nature and function of insulator protein binding sites in the Drosophila genome. *Genome Res. 11*, 2188–2198. doi:10.1101/gr.138156.112.2188

Sémon, M. & Duret, L. (September 2006). Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Molecular Biology and Evolution*, *23*(9), 1715–1723. Retrieved from http://mbe.oxfordjournals.org/content/23/9/1715

Singer, G. A. C., Lloyd, A. T., Huminiecki, L. B., & Wolfe, K. H. (2005). Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. *Molecular Biology and Evolution*, *22*(3), 767–775. doi:10.1093/molbev/msi062. eprint: http://mbe.oxfordjournals.org/content/22/3/767.full.pdf+html

Spellman, P. T. & Rubin, G. M. (2002). Evidence for large domains of similarly expressed genes in the drosophila genome. *Journal of Biology*, *1*(1), 1–8. doi:10.1186/1475-4924-1-5

Stiernagle, T. (2006). Maintenance of c. elegans. In T. C. elegans Research Community (Ed.), *Wormbook*. WormBook.

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S. E., ... Rubin, G. M. (2002). Systematic determination of patterns of gene expression during drosophila embryogenesis. *Genome Biol. 3*(12), research0088.1–88.14. doi:10.1186/gb-2002-3-12-research0088

Tomancak, P., Berman, B. P., Beaton, A., Weiszmann, R., Kwan, E., Hartenstein, V., ... Rubin, G. M. (2007). Global analysis of patterns of gene expression during drosophila embryogenesis. *Genome Biol. 8*(7), R145. doi:10.1186/gb-2007-8-7-r145

van Arensbergen, J., van Steensel, B., & Bussemaker, H. J. (2014). In search of the determinants of enhancer–promoter interaction specificity. *Trends in cell biology*, *24*(11), 695–702. doi:10.1016/j.tcb.2014.07.004

Williams, E. J. & Bowles, D. J. (2004). Coexpression of neighboring genes in the genome of arabidopsis thaliana. *Genome Research*, *14*(6), 1060–1067. doi:10.1101/gr.2131104. eprint: http://genome.cshlp.org/content/14/6/1060.full.pdf+html

Williams, J. B. E. & Hurst, D. L. (2002). Clustering of tissue-specific genes underlies much of the similarity in rates of protein evolution of linked genes. *Journal of Molecular Evolution*, *54*(4), 511–518. doi:10.1007/s00239-001-0043-8

Xie, X., Mikkelsen, T. S., Gnirke, A., Lindblad-toh, K., Kellis, M., & Lander, E. S. (2007). Systematic discovery of regulatory motifs in conserved regions of the human genome , including thousands of CTCF insulator sites. *Proc. Natl. Acad. Sci. USA*, *104*(17), 7145–7150.

Zhan, S., Horrocks, J., & Lukens, L. N. (2006). Islands of co-expressed neighbouring genes in arabidopsis thaliana suggest higher-order chromosome domains. *The Plant Journal*, *45*(3), 347–357. doi:10.1111/j.1365-313X.2005.02619.x

Zhang, Y., Chen, D., Smith, M. A., Zhang, B., & Pan, X. (2012). Selection of reliable reference genes in caenorhabditis elegans for analysis of nanotoxicity. *PLoS ONE*, *7*(3), 1–7.

Chapter 3

RNA POL II LENGTH AND DISORDER ENABLE COOPERATIVE SCALING OF TRANSCRIPTIONAL BURSTING

Quintero-Cadena, P., Lenstra, T. L., & Sternberg, P. W. (2020). RNA Pol II Length and Disorder Enable Cooperative Scaling of Transcriptional Bursting. *Molecular Cell*. doi:https://doi.org/10.1016/j.molcel.2020.05.030

ABSTRACT

RNA Polymerase II contains a disordered C-terminal domain (CTD) whose length enigmatically correlates with genome size. The CTD is crucial to eukaryotic transcription, yet the functional and evolutionary relevance of this variation remains unclear. Here, we use smFISH, live imaging, and RNA-seq to investigate how CTD length and disorder influence transcription. We find that length modulates the size and frequency of transcriptional bursting. Disorder is highly conserved and mediates CTD-CTD interactions, an ability we show is separable from protein sequence and necessary for efficient transcription. We build a data-driven quantitative model, simulations of which recapitulate experiments and support CTD length promotes initial polymerase recruitment to the promoter but slows down its release from it, and that CTD-CTD interactions enable promoter recruitment of multiple polymerases. Our results reveal how these tunable parameters provide access to a range of transcriptional activity, offering a new perspective for the mechanistic significance of CTD length and disorder in transcription across eukaryotes.

3.1 Introduction

In eukaryotes, the RNA Polymerase II complex that transcribes protein-coding genes is typically composed of 12 subunits (Hantsche & Cramer, 2017). The largest and catalytic subunit RPB1 contains a repetitive and unstructured C-terminal Domain (CTD) that is a major factor for establishing critical protein-protein interactions throughout transcription and downstream processes (Harlen & Churchman, 2017).

Each of the heptad amino acid repeats in the CTD, whose number ranges from 5 in *Plasmodium yoelii* to 60 in *Hydra* (Chapman, Heidemann, Hintermair, & Eick, 2008; Yang & Stiller, 2014), can be subject to post-translational modifications that regulate its physical interactions and consequently RNA Pol II function (Eick & Geyer, 2013; Harlen & Churchman, 2017). The CTD's repetitive nature most likely arose in the last eukaryotic common ancestor (Yang & Stiller, 2014), and its length appears to enigmatically correlate with genome size (Chapman et al., 2008; Yang & Stiller, 2014). What is the role of CTD length variation in transcription?

Truncating the number of CTD repeats impacts cell growth and animal development, with a minimal number required for viability (Nonet, Sweetser, & Young, 1987; Bartolomei, Halden, Cullen, & Corden, 1988; West & Corden, 1995; Gibbs et al., 2017; Lu, Portz, & Gilmour, 2019), and reduces the transcriptional output from enhancer responsive genes (Allison & Ingles, 1989; Scafe, C; Young, 1990; Gerber et al., 1995; Aristizabal et al., 2013). Enhancers physically interact with promoters via protein-protein interactions to activate transcription in bursts of activity (Chubb, Trcek, Shenoy, & Singer, 2006; Bartman, Hsu, Hsiung, Raj, & Blobel, 2016; Chen et al., 2018). Given that mRNA output decays rapidly with increasing separation between enhancers and promoters (Dobi & Winston, 2007; Quintero-Cadena & Sternberg, 2016), an intriguing possibility is that CTD expansion facilitates enhancer function over physical distances to promoters that scale with genome size (Allen & Taatjes, 2015).

Increasingly relevant for the understanding of biological phenomena, liquid-liquid phase separation is an emerging signature of proteins with disordered, repetitive domains (Banani, Lee, Hyman, & Rosen, 2017; Shin & Brangwynne, 2017). Low complexity domains that exhibit this property appear to be abundant in nuclear proteins, including the CTD and major transcription factors (Cho et al., 2018; Chong et al., 2018; Sabari et al., 2018; Shin et al., 2018). CTD length has also been implicated in its ability to form (Boehning et al., 2018) and bind phase-separated droplets, with a minimum threshold that parallels its viability requirement (Kwon

et al., 2013). In addition, the interaction of the CTD with these droplets can be dynamically modulated by phosphorylation (Kwon et al., 2013; Chong et al., 2018; Boehning et al., 2018; Cho et al., 2018; Nair et al., 2019), a major post-translational modification that precedes transcription initiation (Payne, Laybourn, & Dahmus, 1989; Svejstrup et al., 1997). In light of these observations, phase separation provides an appealing framework to explain certain transcriptional phenomena. From this perspective, the CTD could provide a bridge for the RNA polymerase to dynamically participate in multi-molecular assemblies of transcription factors and DNA loci that facilitate the function of highly active enhancers (Hnisz, Shrinivas, Young, Chakraborty, & Sharp, 2017).

Extensive investigations have revealed many roles of CTD sequence and post-translational modifications (Eick & Geyer, 2013; Harlen & Churchman, 2017). On the other hand, the functional and evolutionary relevance of CTD length and the mechanism by which it influences transcription have not been systematically investigated.

Here, by quantitatively analyzing snapshots and dynamics of transcription in budding yeast, we show that CTD length can modulate transcription burst size and frequency. We strengthen the evolutionary relevance of the CTD's long disorder, and provide evidence that its role in transcription is separable from amino acid sequence. Specifically, we demonstrate that the function of the CTD's long disorder can be supplemented by similarly unstructured protein domains. These proteins can interact with and recruit others of their kind, an ability that is necessary for efficient transcription *in vivo*. We use these features, together with known CTD protein-protein interactions, to construct an integrative and quantitative model that explains how CTD length influences the dynamics of eukaryotic transcription.

3.2 Results

CTD is enriched in disordered amino acids and its length inversely correlates with gene density across eukaryotes

The CTD of representative species has been shown to be a random coil (Portz et al., 2017), suggesting this structural feature is relevant for its function and could itself be used to identify it. We sought to learn whether this was a common signature in all known protein sequences. We searched for RPB1 protein sequence homologs, the CTD-bearing catalytic subunit of RNA Pol II (Figure 3.1A, top). We recovered 542 unique sequences from 539 species and 338 genera, whose length varies from

1374 to 3055 amino acids (Figure 3.1B), with some evident bias that likely stems from the limited availability of genome sequences.

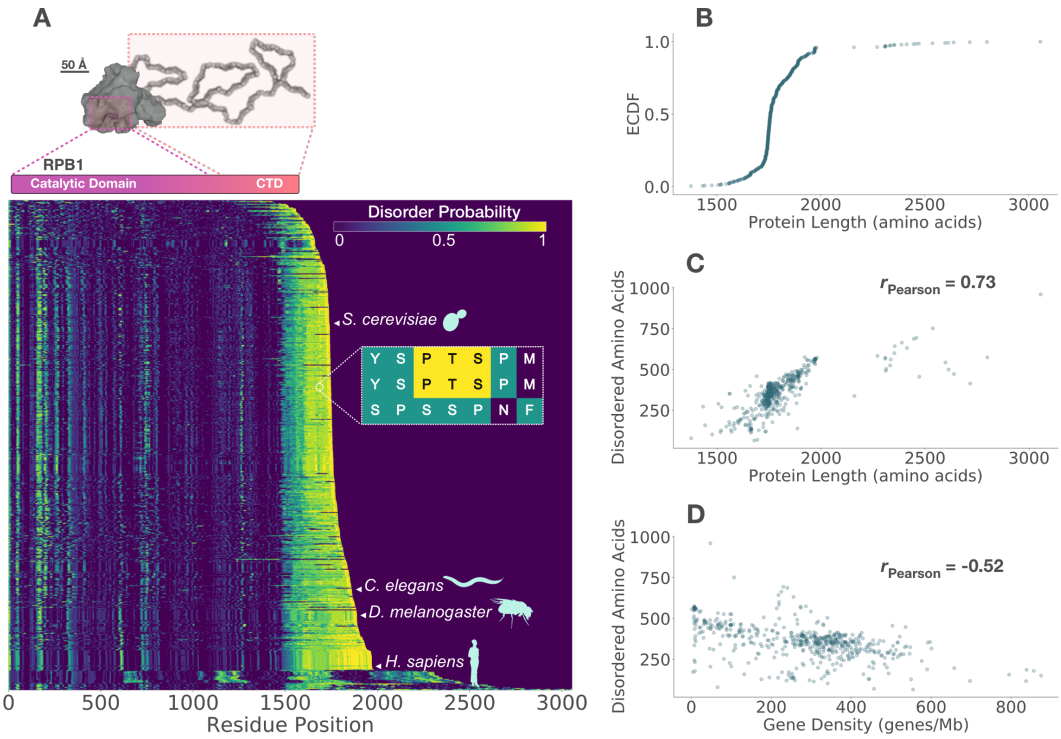

Figure 3.1: **RPB1, the catalytic subunit of RNA Polymerase II, contains an unstructured C-terminal Domain (CTD) whose length correlates with gene density across eukaryotes.** (A) Top: Cartoon of 12 subunit RNA Pol II complex with an unstructured CTD drawn to scale, adapted from 1Y1W.PBD (Kettenberger, Armache, & Cramer, 2004; Portz et al., 2017). RPB1 is highlighted in pink and its CTD in orange. Bottom: distribution of disorder probability along RPB1 sequence homologs sorted by length. Representative species are highlighted. Inset illustrates residue coloring by disorder probability. (B) Empirical cumulative distribution function (ECDF) of RPB1 lengths. (C) RPB1 length positively correlates with number of predicted disordered amino acids and inversely with gene density (D). Pearson correlation coefficient is shown for each pair of variables.

We computed the disorder probability per amino acid along each protein sequence. We found that with few exceptions, the C-termini of RPB1 sequences is enriched in disordered amino acids (Figure 3.1A, bottom). Protein length is positively correlated with the number of disordered amino acids (Figure 3.1C). As noted in previous reports (Chapman et al., 2008; Yang & Stiller, 2014), these C-terminal sequences are enriched in amino acids from the heptad repeat YSPTSPS (Figure 3.7A). Like amino acid content, overall charge, aromaticity, and hydrophobicity have a compact distribution (Figure 3.7B-D).

CTD length has been shown to correlate with genome size for a few representative species that span a wide range of genome sizes (Chapman et al., 2008; Yang & Stiller, 2014), from $1x10^7$ bp in yeast to $3x10^9$ bp in human. To systematically investigate the generality of this phenomenon, we compiled a list of genome sizes and their estimated gene number. We then computed the gene density (genes per megabase of DNA) for each species, in order to account for long stretches of non-coding DNA that are more common in large genomes. The number of disordered amino acids in RPB1 homologs inversely correlated with gene density (Figure 3.1D): sparse genomes tend to have polymerases with longer CTDs.

This systematic characterization of RPB1 sequences builds on previous extensive reports (Chapman et al., 2008; Yang & Stiller, 2014) and highlights three important features of the CTD. First, protein disorder is a highly conserved and likely functionally relevant feature of the CTD. Second, RPB1 length variation mostly originates from the number of disordered amino acids in its C-terminus. Third, CTD length is inversely correlated with gene density.

CTD length modulates transcription burst size and frequency

We sought to understand the role of CTD length using *Saccharomyces cerevisiae* as our model. Wild-type yeast CTD contains 26 heptad repeats (CTDr). We generated strains in which the genomic copy of RPB1 was engineered to have 14, 12, 10, 9, and 8 CTDr, the minimum required for viability in yeast (West & Corden, 1995). Consistent with previous reports (Nonet et al., 1987; West & Corden, 1995), the growth rate of these strains was compromised by CTD truncation. The magnitude of the decrease in growth rate increased with decreasing CTD length (Figure 3.2A); while 14 and 12 CTDr strains have only a subtle growth phenotype, the decrease in growth rate becomes evident with 10 CTDr and progressively larger with 9 and 8 CTDr (Figure 3.2A, inset).

RNA Pol II is mostly present in the nucleus, imported as a fully assembled complex (Boulon et al., 2010; Czeko, Seizl, Augsberger, Mielke, & Cramer, 2011). We hypothesized that nuclear polymerase becoming rate-limiting could explain the observed phenotypes, given that the CTD has been linked to nuclear import (Carre & Shiekhattar, 2011). We fused the fluorescent protein mScarlet to RPB1, and unexpectedly found that CTD truncation increased its nuclear levels (Figure 3.2B). This effect is presumably explained by the requirement of the CTD for ubiquitination (Huibregtse, Yang, & Beaudenon, 1997; Somesh et al., 2005) and the resulting

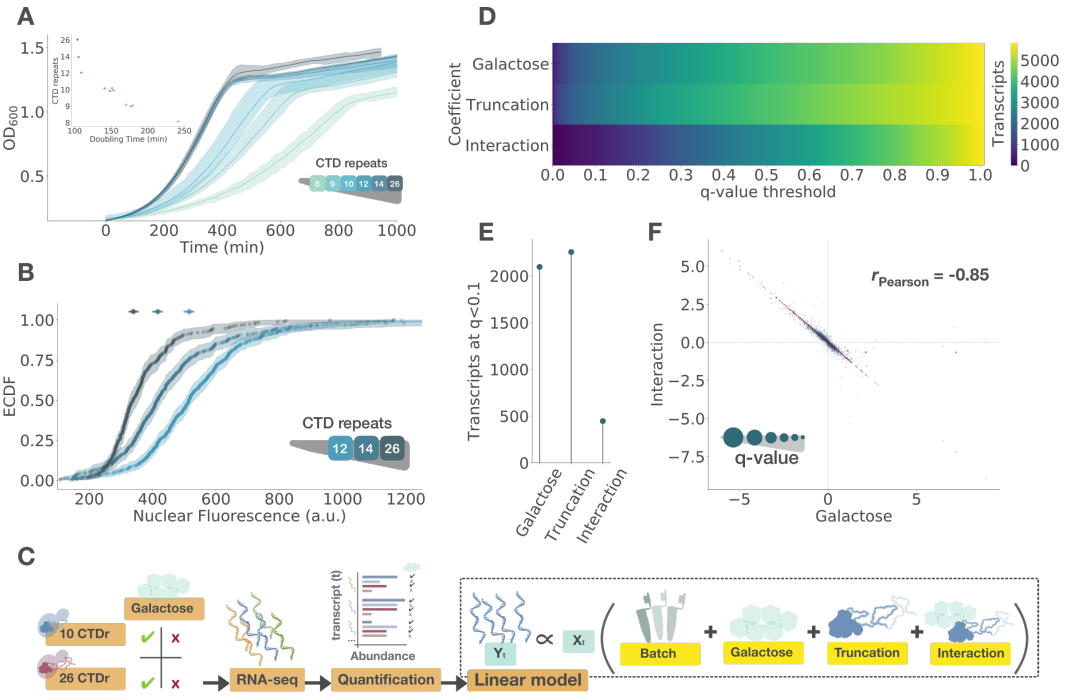

Figure 3.2: **CTD truncation reduces growth rate, increases nuclear Pol II concentration and antagonizes transcription activation genome-wide.** (A) Mean optical density (OD) over time of *S. cerevisiae* strains with 26 (wild-type), 14, 12, 10, 9, and 8 CTD repeats (CTDr). Shaded area shows the range of measured ODs at each time point from three biological replicates per line. Inset shows mean doubling times (DT) with standard error by line. (B) Empirical cumulative distribution function (ECDF) of mScarlet-RPB1 nuclear fluorescence in strains with 26, 14, and 12 CTDr from three biological replicates. Shaded area is bootstrapped 99% confidence interval (CI) and top markers show median with 99% CI. (C) Experimental design to measure the transcriptomic phenotype of CTD truncation and its influence on the transcriptional response after 2 hours of galactose induction via RNA-seq from three biological replicates. A linear model was used to fit the RNA-seq data, where each coefficient estimates the influence on each measured transcript of batch effects, galactose induction, CTD truncation and its interaction with induction. For each coefficient, (D) cumulative distribution, in number of transcripts, of false discovery rates (q-values) and (E) number of differentially expressed transcripts detected at a q-value threshold of 0.1. (F) Comparison of the log fold-change of each transcript resulting from galactose induction and its interaction with CTD truncation. Red points show the positions on the diagonal $x = y$. Marker size of each point is inversely proportional to the q-value of the interaction ($ms = -log(q_{int})$); dotted lines reference no change at zero and the Pearson correlation is indicated. Direct targets of GAL4 listed in Lesurf et al., 2016 are plotted in orange.

accumulation of the protein complex. We were unable to fuse mScarlet to shorter CTD strains; however, this trend suggested the polymerase does not become rate-

limiting upon CTD truncation.

We then asked how CTD length influences transcription. We focused on the galactose transcriptional response, because like other inducible pathways it has been shown to be sensitive to CTD truncations (Allison & Ingles, 1989; Scafe, C; Young, 1990) and involves the differential expression of over 2000 transcripts (Figure 3.2E), about a third of the yeast's transcriptome. We measured the transcriptional phenotype of CTD truncation using RNA-seq, comparing wild-type with the 10CTDr strain with and without galactose (Figure 3.2C). We fitted these data to a linear model that allowed us to estimate the individual contributions, for each measured transcript, of four components in the experiment: batch effects, galactose induction, CTD truncation, and its interaction with galactose induction (Figure 3.2C, right).

We identified over 2000 transcripts affected by CTD truncation at a q-value threshold of 0.1, from a distribution of q-values that indicates a strong transcriptional phenotype (Figure 3.2D,E). A significant proportion of the galactose responding transcripts exhibited a statistical interaction with CTD truncation. This effect revealed a surprisingly specific, globally antagonistic relationship between two components: 1) galactose induction and 2) the interaction of truncation with galactose induction (Figure 3.2F). In other words, CTD truncation reduced the magnitude of change in abundance of most transcripts upon galactose induction.

We sought to understand the source of this antagonism by visualizing transcription dynamics in living cells. We introduced 14 copies of the bacteriophage sequence PP7 in the 5' UTR of GAL10, a strongly galactose responsive gene. Each of the PP7 repeats forms an RNA hairpin that can be bound by a pair of PP7 coat proteins fused to GFP (Coulon et al., 2014; Lenstra, Coulon, Chow, and Larson, 2015, Figure 3.3A, top). This system allowed us to visualize the dynamics of GAL10 transcription upon galactose induction as fluorescence bursts arising from the transcription site (TS; Figure 3.3A, bottom). We found that expressing TS intensity as the ratio of spot to mean nuclear fluorescence could reliably account for the differences in PP7-GFP levels observed between strains (Figure 3.8D-H).

Given the burstiness of transcription, we hypothesized that two parameters could play a role in the diminished transcriptional output, namely burst size and frequency. We measured transcription fluorescence traces for 26 (wild-type), 14, and 12 CTDr strains (Figure 3.3B). CTD truncation decreased the intensity of fluorescence bursts (Figure 3.3C), suggesting a decrease in burst size. Also, truncation increased the time interval between bursts (Figure 3.3D), which is closely related to burst

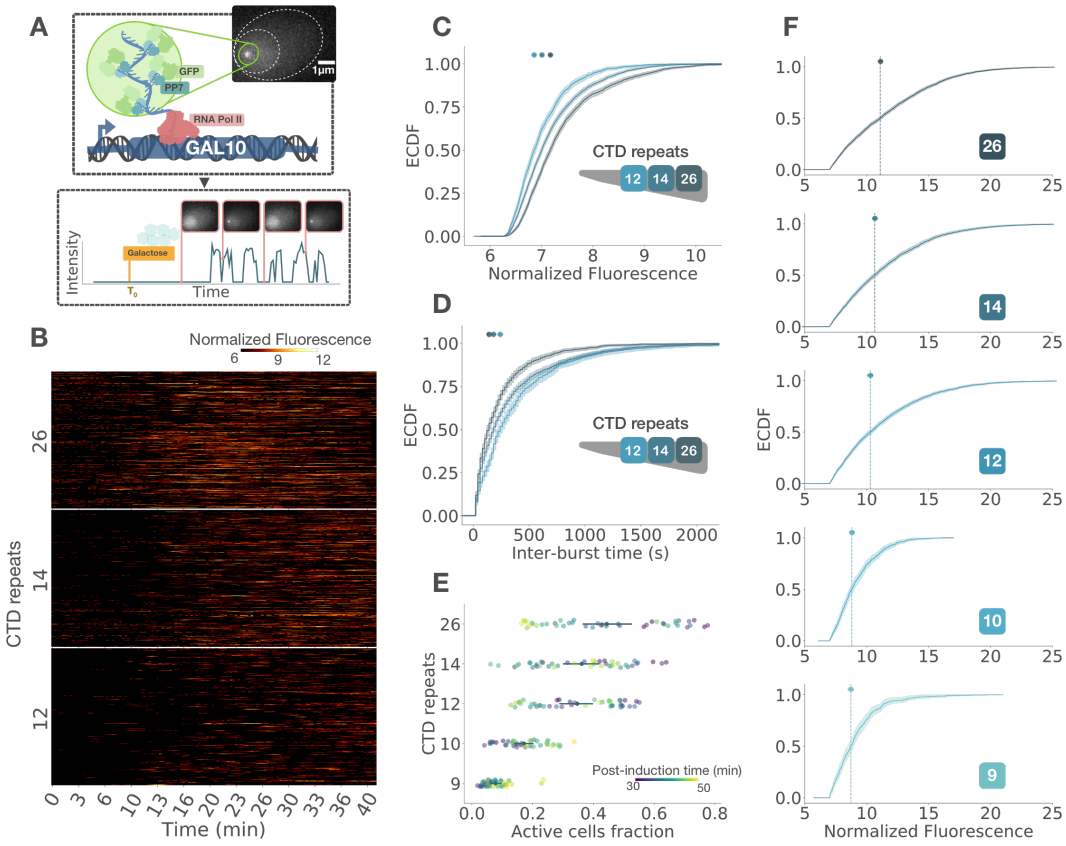

Figure 3.3: **CTD length modulates transcription burst size and frequency.** (A) Experimental strategy used to observe transcription dynamics in live cells. The 5' end of a single allele of galactose-inducible GAL10 is tagged with RNA hairpins that are bound by nuclear expressed GFP-PP7. This protein fusion results in a fluorescent spot in the cell nucleus upon transcription activation, whose fluorescence through time is recorded to investigate transcription dynamics. (B) Transcriptional traces by cell of strains with 26 (wild-type), 14, and 12 CTD repeats (CTDr) from three biological replicates. These are related to burst size and frequency through (C) the empirical cumulative distribution function (ECDF) of transcription site fluorescence intensities and the ECDF of inter-burst times, in seconds (D). Shaded area is bootstrapped 99% confidence interval (CI) and top markers show median with 99% CI. (E) Fraction of active cells per field of view from two biological replicates measured from high-laser-power snapshots of strains with 26, 14, 12, 10, and 9 CTDr. Middle points indicate mean with bootstrapped 99% CI. Color indicates time after galactose induction. (F) ECDFs of normalized fluorescence intensities of transcription bursts from these snapshots. Vertical dotted lines indicate median and shaded area bootstrapped 99% CI. Number of CTD repeats is indicated in the lower right corner of each plot.

frequency. We could similarly observe this frequency decay by looking at whether the average cell was active or inactive over time for each strain (Figure 3.9A).

These average traces also show that burst frequency remained roughly constant after activation, only declining towards the end likely due to photobleaching and mRNA-bound PP7-GFP nuclear export. The autocorrelation of normalized intensity traces increased in amplitude with CTD truncation, similarly supporting a decrease in burst frequency (Figure 3.9B-D). Differences in burst duration measured from this analysis were more subtle. This potential ambiguity could mean that burst duration is independent of size, or that the observed decay in TS intensity was influenced by the decay in burst frequency. Overall, the live transcription measurements suggested CTD length can simultaneously modulate burst size and frequency.

The transcriptional activity in strains with shorter CTDs was too weak to be visualized in these movies without incurring in phototoxic illumination. To circumvent this problem, we took a single snapshot per field of view with maxium laser intensity during a 20 minute window after 30 minutes of galactose induction. This approach additionally allowed us to obtain a better estimate of TS fluorescence. The fraction of active cells per field of view was impacted by CTD truncation (Figure 3.3E). We observed a transition similar to the growth phenotype in this assay, where the magnitude of the effect progressively increased with CTD truncation. We also observed a consistently moderate shift in the distributions of TS fluorescence with CTD truncation (Figure 3.3F). The comparatively small magnitude of this decrease suggested the decay in fraction of active cells is primarily driven by burst frequency. From these measurements, we conclude that CTD length can modulate both the size and the frequency of transcriptional bursting.

Fusion to disordered proteins can rescue the function of a CTD-truncated RNA Pol II

Given the conservation of protein disorder (Figure 3.1A) and recent evidence that the CTD can form and interact with phase separated droplets (Kwon et al., 2013; Chong et al., 2018; Boehning et al., 2018; Cho et al., 2018; Nair et al., 2019), we hypothesized that the function of the CTD's long disorder could be supplied by other proteins of similar chemical and structural features. We tested this idea by fusing the low complexity domains (LCD) of the human proteins FUS and TAF15, which are not present in the yeast genome, to the C-terminus of a 10CTDr truncated RPB1. These LCDs contain neither a known nuclear localization sequence (Gal et al., 2011; Marko, Vlassis, Guialis, & Leichter, 2012) nor ubiquitination sites (Mertins et al., 2013) that could supplement CTD function in a predictable manner; they are similar in amino acid composition, particularly in the frequency of tyrosines, but share little

sequence similarity with the CTD (Figure 3.10A-C).

Astonishingly, strains carrying either protein fusion showed an improved growth rate over the 10CTDr strain. Fusion to FUS LCD progressively rescued growth rates of strains with 9 and 8 CTDr. Furthermore, strains with 7 and 6 CTDr remained viable when fused to FUS, surpassing the minimum requirement of 8 CTDr alone (Figure 3.4A). This supression was particularly striking because the CTD has been extensively mutated, typically with detrimental effects to transcription or downstream processes. This new minimal length is also noticeably close to the four heptad repeats that directly contact Mediator in an assembled preinitiation complex (P. J. J. Robinson, Bushnell, Trnka, Burlingame, & Kornberg, 2012; P. J. Robinson et al., 2016).

We next probed whether the improved growth phenotype originated from a transcriptional rescue. We used RNA-seq to compare the transcriptomes of the FUS and TAF15 rescued strains and their response to galactose induction with that of the wild-type and 10CTDr strains. Using principal component analysis, we observed LCD fusion results in transcriptomes in-between that of the wild-type and truncated strains, under induced and uninduced conditions (Figure 3.11A). As described in Figure 3.2B, we fitted the data to a linear model to identify the contributions to each transcript of galactose induction, CTD truncation, FUS or TAF15 LCD fusion to a 10CTDr truncated polymerase, and their interaction with galactose (Figure 3.4B, top). We found the number of differentially expressed transcripts resulting from CTD truncation decreased from 2256 to 1037 and 883 transcripts at a q-value threshold of 0.1 upon fusion to FUS or TAF15 LCDs, respectively (Figure 3.4C). More generally, the distribution of q-values resulting from CTD truncation shifted towards less significant values (Figure 3.4B, middle), a sign of considerable amelioration in the transcriptional phenotype. These LCD fusions additionally shifted the distribution of q-values of the interaction between CTD truncation and galactose induction (Figure 3.4B, bottom), abolishing the measurable effect at a q-value of less than 0.1 (Figure 3.4C). Moreover, the global antagonism of this interaction term with galactose induction nearly vanished (Figure 3.4D,E).

We interrogated these data further using other linear models, the simplest of which consists of independently measuring the galactose induction in each strain (Figure 3.4F). In this comparison, the galactose response of most transcripts in the rescued strains closely resembles that of the wild-type (Figure 3.4G,H), more than the 10CTDr alone (Figure 3.4I). The FUS and TAF15 transcriptional phenotypes, as

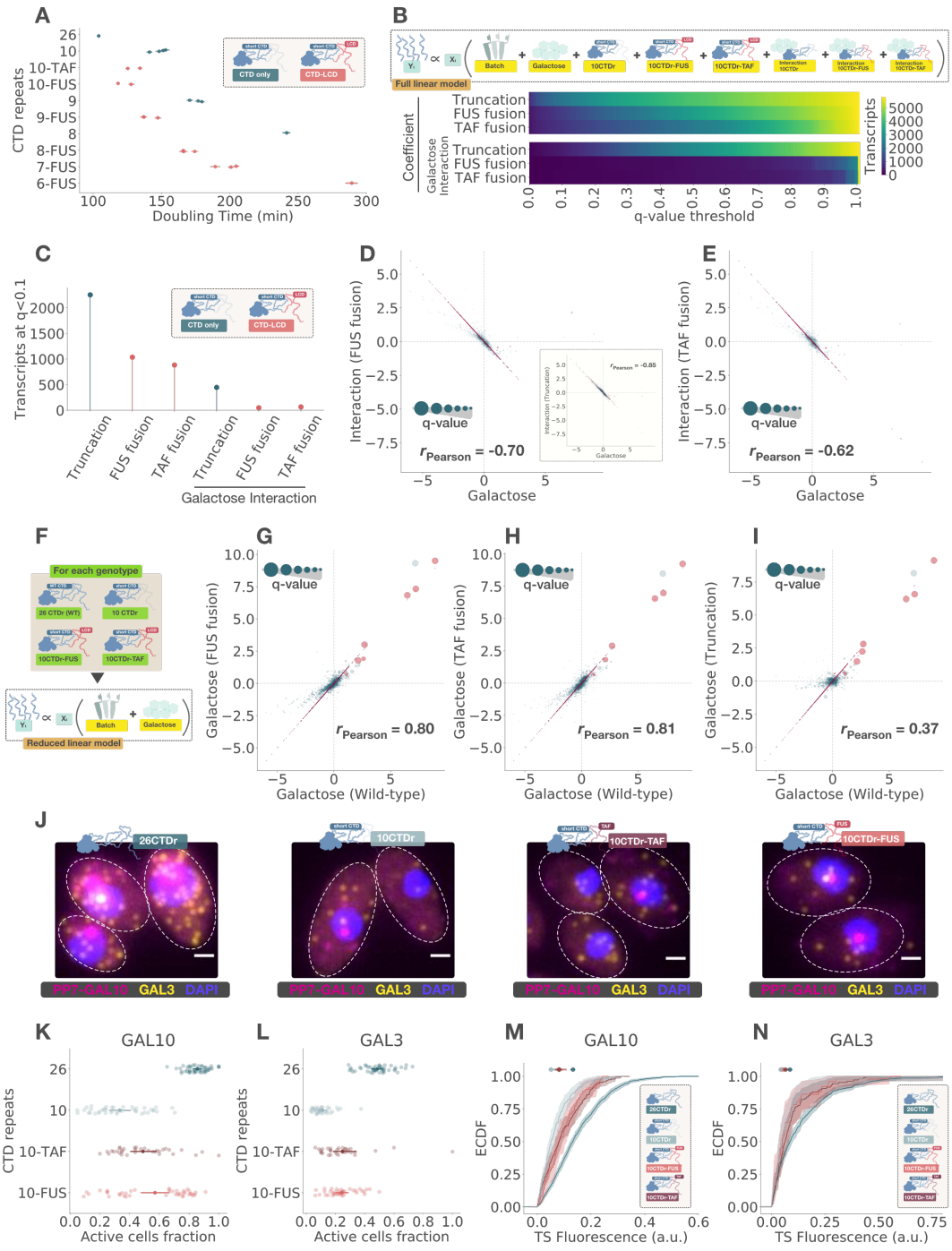

Figure 3.4: **Fusion of the low-complexity domain (LCD) of FUS or TAF15 to a truncated polymerase can rescue its function and reduce the CTD length required to support cell growth.** (A) Comparison of doubling times (DT) of strains with wild-type and decreasing number of CTD repeats (CTDr) with and without fusion to the LCD of FUS or TAF15. Individual points come from independent lines when available and indicate mean DTs with standard error estimated from three biological replicates per line.

Figure 3.4: (B) Top: Linear model used to estimate the effect of truncation to 10CTDr, subsequent LCD fusion, and the interaction of each of these components with galactose induction. Bottom: Resulting cumulative distributions of q-values, in number of transcripts, for each coefficient, excluding galactose and batch effects from three biological replicates. (C) Number differentially expressed transcripts detected at a q-value threshold of 0.1. Comparison of the log fold-change of each transcript induced by galactose and its interaction with 10CTDr fused with FUS (D) and TAF15 (E) LCDs. Inset shows comparison with 10CTDr interaction alone. (F) Reduced linear model used to measure galactose induction in each strain individually. Comparisons of log fold-change of each transcript induced by galactose in wild-type and 10CTDr (G), 10CTDr-FUS (H), and 10CTDr-TAF15 (I). Red points show the positions on the diagonal $x = y$. Marker size of each point is inversely proportional to the q-value of the coefficient in the y-axis ($ms = -log(q_y)$); dotted lines reference no change at zero and the Pearson correlation is indicated. Direct targets of GAL4 listed in Lesurf et al., 2016 are plotted in orange. (J) Representative images of smFISH with probes for a single allele of PP7-GAL10 and both alleles of GAL3 after two hours of galactose induction for 26CTDr (wild-type), 10CTDr, 10CTDr-TAF, and 10CTDr-FUS strains as indicated on top of each image. White dotted contours mark cell outlines. Scale bar is 1 μm. Fraction of active cells per field of view for GAL10 (K) and GAL3 (L) measured from two biological replicates of smFISH. Mean with 99% bootstrapped confidence interval (CI) is shown on top of each group. Corresponding empirical cumulative distribution functions (ECDF) with 99% bootstrapped CI of transcription site intensities of GAL10 (M) and GAL3 (N). Medians with 99% CI are shown on top.

measured using the full model (Figure 3.4B, top), were highly correlated (Figure 3.11B). Based on this observation, we fitted the data to a model in which we consider the two rescued strains as a single LCD group by pooling their transcriptomes together (Figure 3.11C). This model increased the number of transcripts that we can confidently call differentially expressed at a q-value threshold of 0.1 to 1392 (Figure 3.11D). This effect supports the transcriptional rescue occurs through a single pathway, and that there is a CTD sequence-dependent signature that remains shared among the three 10CTDr strains. This signature was evident from the high similarity between the truncation and LCD fusion coefficients (Figure 3.11E). From this experiment, we conclude that the sequence and long disorder of the CTD have separable roles in transcription, the latter of which can be supplemented by the similarly disordered LCDs of FUS or TAF15.

For unknown reasons, our live imaging system did not work with the FUS rescued strains. We circumvented this issue by using two-color Single-Molecule Fluorescence *in situ* Hybridization (smFISH). We used probes against the PP7 repeats,

allowing us to detect mRNA from a single allele of GAL10, and against GAL3, which could detect RNA from both of its alleles (Figure 3.4J). The fraction of active cells consistently increased for both the single allele of GAL10 and the two alleles of GAL3 (Figure 3.4K,L). In addition, we observed an increase in TS fluorescence intensity of both rescued strains over the 10CTDr strain (Figure 3.4L,M), suggesting the fraction of active cells increased specifically because of an increased burst size.

Together, these results show the CTD's long disorder can influence transcription in a way that depends on its chemical and structural properties rather than its precise amino acid sequence.

CTD, FUS, and TAF15 LCDs can self-interact and this ability is necessary for efficient transcription

The CTD can bind the LCD of FUS and more strongly that of TAF15 (Kwon et al., 2013). TAF15 can also interact with components of the Mediator complex (Takahashi et al., 2011). However, the avidity of these interactions does not appear to correlate with the extent of rescue (Figure 3.4A). Other experiments have shown these LCDs are able to phase-separate (Kwon et al., 2013). These phases are thought to form as a result of intermolecular interactions that collectively drive droplet formation (Banani et al., 2017; Shin & Brangwynne, 2017). We hypothesized that FUS, TAF15, and the CTD could be involved in the recruitment of RNA Pol II to the TS via these self-interactions.

We first tested whether FUS and TAF15 variants, with tyrosine to serine misense mutations that make them significantly less able to bind phase-separated droplets *in vitro* (Kwon et al., 2013; data reproduced in Figure 3.10D), would fail to rescue the growth phenotype of a 10CTDr strain.

We observed a correlation between droplet binding rates and the extent of growth rescue upon fusion of these protein variants to the truncated RNA polymerase (Figure 3.5A,B). This rescue also correlated with the compromised ability of these proteins to function as transcription factors when fused to a DNA binding domain (Kwon et al., 2013).

We investigated these mutants more closely by using an assay designed to measure self-interactions *in vivo*, which we define as the ability of a protein to interact with and recruit others of its kind. We speculated that fusing a self-interacting protein to PP7-GFP would lead to brighter spots in our live transcription assay (Figure 3.5C, left). Heterologously expressed FUS and TAF15 can form punctate structures in

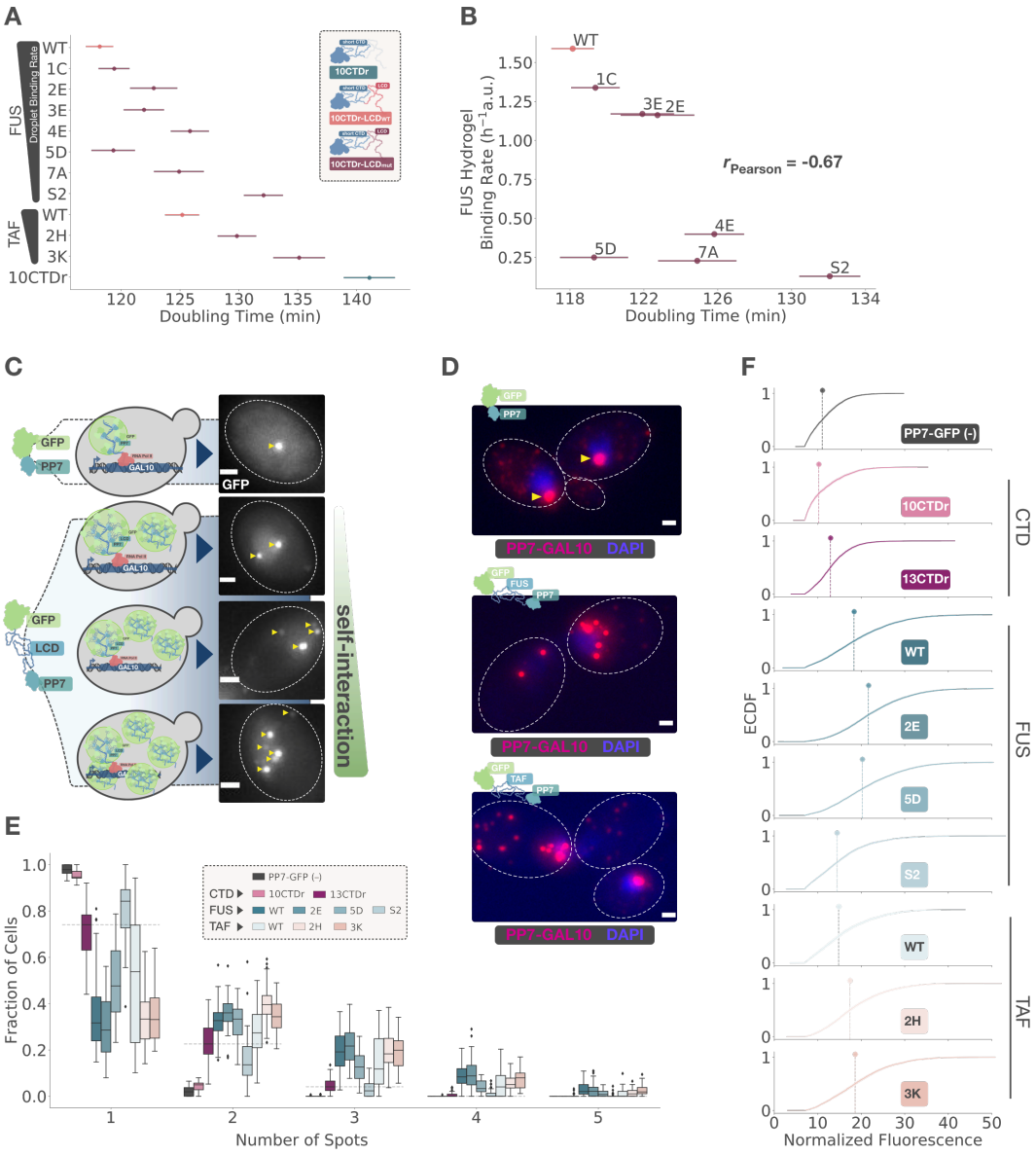

Figure 3.5: **CTD and the low complexity domains (LCD) of FUS and TAF15 can self-interact *in vivo* and this ability is necessary for the function of RNA Pol II.** (A) Doubling times (DT) of 10 CTD repeat (CTDr) strains with and without fusion to FUS or TAF15 wild-type and mutated LCDs sorted by droplet binding rate as reported in Kwon et al., 2013 and comparison of FUS mutants with these rates (B). Pearson correlation is indicated. Color indicates whether fused LCD is present, wild-type or mutant. Individual points come from independent lines when available and indicate mean DTs with standard error estimated from three biological replicates per line.

Figure 3.5: (C) Diagram of self-interaction assay. An LCD is fused to GFP and PP7 coat proteins; GFP-PP7 fusion with no LCD is used as control (top). Upon induction of transcription, these fusion proteins bind mRNA scaffolds. If LCDs can self interact, they increase the brightness of spots by recruiting more proteins to the scaffold, and by bringing more than one mRNA together outside of the active transcription site via LCD-LCD interactions (bottom). Representative snapshots of cells with increasing number of GFP fluorescent spots are shown in the right. (D) smFISH after 30 minutes of galactose induction with probes that bind PP7 hairpins in GAL10 mRNA on strains constitutively expressing PP7-GFP with and without LCD fusion as indicated in the upper left corner of each image. Cells without an LCD have a single nuclear RNA complex per cell corresponding to the transcription site (top, yellow arrowheads). Individual mRNA molecules are visible as dimmer spots. Fusion to FUS (middle) or TAF15 (bottom) LCD leads to the formation of multiple RNA complexes, visible as spots brighter than a single mRNA, in individual cells as a result of intermolecular LCD-LCD interactions. White dotted contours mark cell outlines; scale bar is 1μm. (E) Fraction of cells per field of view that contain each number of GFP spots by strain from three biological replicates. The horizontal blue dotted lines indicate the number of spots in the 13CTDr GFP-PP7 fusion. Protein fused to PP7-GFP is indicated by color. (F) Corresponding empirical cumulative distribution functions (ECDF) of GFP fluorescence intensities from the brightest spot in each cell, typically corresponding to the transcription site. Protein fused to PP7-GFP is indicated by color and in the lower right of each plot.

yeast (Couthouis et al., 2011; Ju et al., 2011). However, in this assay wild-type FUS and TAF15 LCD fusions distributed uniformly across the cell nucleus, presumably due to lower protein concentrations. On the other hand, the spots that formed after galactose induction became brighter with either LCD compared to PP7-GFP alone. Moreover, more than a single spot became visible soon after transcription activation (Figure 3.5C, right).

We characterized this phenomenon via smFISH of galactose induced strains carrying PP7-GFP with and without FUS and TAF15 LCD fusion (Figure 3.5D). This experiment suggested that the increase in the number of spots and their brightness could be a result of 1) indirect recruitment of PP7-GFP-LCD to each mRNA scaffold but mostly 2) LCD-mediated physical interactions between mRNA molecules outside of the TS. We proceeded to use this assay to determine whether a protein can self-interact at physiological concentrations *in vivo*.

We counted the number of bright spots per cell arising 30 minutes after galactose induction in snapshots taken during a 20 minute window with a maximum intensity laser. This number is almost always 1 for PP7-GFP alone, corresponding to the TS,

except for when a cell had just duplicated the GAL10 locus during cell division. A 10CTDr-PP7-GFP protein fusion mostly produced a single spot, similar to the non-self-interacting control. On the other hand, a 13CTDr and all of the FUS and TAF15 variants resulted in a higher fraction of cells with more than a single bright spot (Figure 3.5E).

We also compared the fluorescence intensity of the brightest spot per cell, presumably the TS, to the control expressing PP7-GFP only. Proteins that formed multiple bright spots also increased the intensity of the brightest spot (Figure 3.5F). The 10CTDr construct resembled the negative control more than 13 CTDr. Generally, these measurements support that 13CTDr, FUS, and TAF LCDs can self-interact, and the extent of self-interaction qualitatively recapitulates the transitions observed in our growth and transcription assays.

These results suggest that self-interactions are necessary for the transcriptional rescue of CTD truncation, and support the idea that this ability is a key attribute that the CTD's long disorder contributes to transcription.

An integrative transcription model explains the influence of CTD length

In light of our evidence, we sought to devise a quantitative model for transcription that captured the effects of perturbing CTD length and illuminated the role of disorder-mediated self-interactions. We built upon a model that includes an active and an inactive state, which enables it to produce transcriptional bursting, and specifies that an RNA polymerase molecule can only be recruited during the active state, in agreement with experimental data (Bartman et al., 2019).

The CTD is known to physically interact with Mediator (Thompson, Koleske, Chao, & Young, 1993; Kim, Björklund, Li, Sayre, & Kornberg, 1994), a prevalent component of the preinitiation complex (PIC) (Allen & Taatjes, 2015). Given the repetitive nature of the CTD, we postulated that the number of repeats and hence CTD length could modulate the affinity with which the polymerase binds a Mediator-bearing PIC. Following this logic, we chose to explicitly refer to the active state as the assembled PIC, primed for RNA Pol II binding.

Polymerase release from the PIC is preceded by CTD phosphorylation (Payne et al., 1989; Svejstrup et al., 1997), which disrupts their physical interaction (Jeronimo & Robert, 2014; Wong, Jin, & Struhl, 2014). Assuming each of the CTD repeats contributes to this interaction, we reasoned that their number should correlate with the rate of polymerase release. Our rationale is that given the ratio of unphos-

phorylated to phosphorylated repeats determines the physicochemical state of the CTD and its interaction with Mediator (P. J. Robinson et al., 2016), the number of phosphorylation sites, or CTD length, should be proportional to the time it takes for CTD kinases to reach this threshold.

To recapitulate, we postulated that CTD length influences transcription by enhancing polymerase recruitment rate to the PIC (β) and by decreasing polymerase release rate (ϕ) from the PIC (Figure 3.6A, blue).

An important feature of the current model for transcription is that only a single polymerase molecule can bind the PIC at a given time (Bartman et al., 2019; Figure 3.6A, blue). To incorporate the ability of the CTD to self-interact, we postulated an additional molecular state that allows for the recruitment of more than single polymerase molecule to the extant PIC-Pol II complex via CTD-CTD self-interactions (Figure 3.6A, pink).

Finally, we used this model to assess the transcriptional rescue observed upon fusion of FUS or TAF15 LCDs to a CTD-truncated RNA polymerase. We reasoned that these LCDs would contribute the ability to self-interact, but make it more difficult for the polymerase to be released into the gene because their phosphorylation may not be as efficient. Specifically, we postulated that fusing a self-interacting LCD to a truncated CTD would fix self-recruitment rate ϵ and release rate ϕ, while the rate of initial recruitment β would still be determined by CTD length.

We interrogated the consistency of our models with experimental data using stochastic simulations. For simplicity, we set CTD length (CTD_L) to be a number in the range (0-1), directly proportional to both polymerase recruitment rates β and ϵ, whose complement ($1 - CTD_L$) is proportional to polymerase release rate ϕ. We visualized these simulations as transcriptional traces for each model, including states of PIC assembly, numbers of PIC-bound and phosphorylated polymerases (Figure 3.6B and Figure 3.12), akin to our live transcription imaging data (Figure 3.3B). From these simulations we computed the distributions of burst sizes, inter-burst times and fraction of active cells, and asked how CTD length would affect these parameters. Importantly, to be consistent with prior literature, we define the start and end of a burst to be concomitant with PIC assembly and disassembly, respectively, and only consider those that yield at least one mRNA molecule.

We compared the outcomes of the one-polymerase, the many-polymerases and the rescue models. The identity of the distributions of burst sizes and inter-burst times

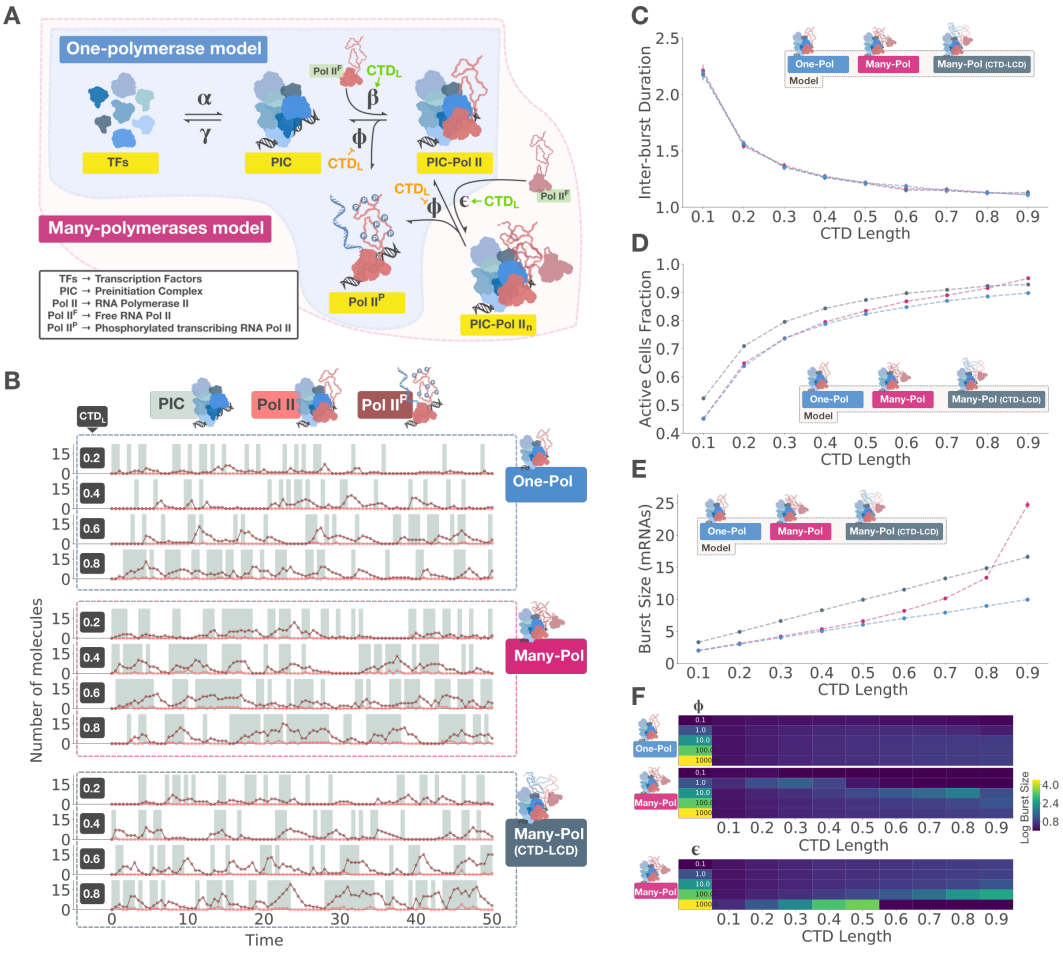

Figure 3.6: **An integrative model for transcription activation explains the role of CTD length.** (A) CTD-centric model for transcription activation. Transcription factors (TFs) assemble to form a preinitiation complex (PIC) to which an RNA polymerase binds. In the one-polymerase model (blue), only a single polymerase is allowed to bind the complex at a time. In the many-polymerases model (pink), more than one polymerase can bind the complex via CTD-CTD interactions. Fusion of a truncated CTD to a self-interacting protein is simulated using the many-polymerases model with fixed ϵ and ϕ. The positive or negative influence of CTD length (CTD_L) on each rate is indicated in green and orange, respectively. (B) Representative traces from stochastic simulations by model, indicated to the right, as a function of CTD_L, indicated in the left side of each plot. Background shows PIC assembly state; number of PIC bound and phosphorylated (transcribing) polymerases are shown in color as indicated in the legend on top. Corresponding mean inter-burst durations (C), mean fractions of active cells (D), and mean burst sizes (E) with 99% bootstrapped confidence interval. (F) Effect of varying ϕ (top) and ϵ (bottom) on the log burst size.

resembled geometric and exponential, respectively, for all the models (Figure 3.13). In addition, because the difference between models lies in the states after the start of a burst and hence does not influence burst frequency, the model choice did not affect the resulting distributions of inter-burst times (Figure 3.6C, 3.13C).

The three models predicted that shortening the CTD would progressively increase the average inter-burst time (Figure 3.6C). This in turn translated into a progressive decrease in the fraction of active cells (Figure 3.6D). Importantly, the magnitude of change in these numbers increased with decreasing CTD length. This effect qualitatively reproduced the observed transitions in growth and transcription phenotypes resulting from CTD truncation (Figure 3.2A, 3.3E).

Although dominated by burst frequency in this regime, the fraction of active cells was also influenced by burst size (Figure 3.14A). Our simulations predicted that such influence would lead to a modest but consistent increase in the fraction of active cells across a range of CTD lengths when comparing rescued with truncated CTD lengths (Figure 3.6D). In contrast to frequency, burst size was significantly influenced by model choice. The rescue model predicted a consistently higher mean burst size at almost every CTD length, except at the extreme of a long CTD (Figure 3.6E). These results are strikingly consistent with our experiments, in that rescued strains show a moderate increase in the fraction of active cells compared to truncated CTDs (Figure 3.4K,L), and TS intensity increased upon LCD fusion (Figure 3.4M,N). These comparisons of simulations with experiments support the existence of a state with more than one polymerase and that LCD fusion specifically rescues burst size via self-interactions.

The one-polymerase and the many-polymerases models produced very similar mean burst sizes with short CTDs. However, burst sizes resulting from each model deviated significantly with longer CTDs (Figure 3.6E). In particular, they increased exponentially when many polymerases were allowed to bind the PIC, but only linearly with a single polymerase. These predictions suggest the impact of binding more than one polymerase may become increasingly relevant as the CTD grows longer.

We also sought to elucidate the relationship between burst size, duration, and TS intensity in our experiments by comparing these three values in our simulations. Because CTD length inhibits release rate in the one-polymerase and the many-polymerases models, burst duration increased faster than burst size (Figure 3.14B); these parameters were linearly proportional only when release rate was fixed in

the case of the rescue model. These results potentially explain why burst duration decays more subtly (Figure 3.9B-D) than TS intensity (Figure 3.3D,F). A potential source of uncertainty in measuring burst size is that the decay in TS intensity could be a result of decreased burst frequency, given that GAL10 is transcribed in highly frequent bursts that could overlap in time and inflate the measured intensity of a burst. We simulated how this intensity signal would vary across CTD lengths and under different elongation rates δ, which determines how long a given mRNA stays and contributes to the TS signal. We compared the fraction of active cells with the difference between this observed TS intensity, influenced by δ and measured as the distributions of peak intensities in the final simulated traces, and the true burst size, which we kept track of as we generated the traces and is independent of δ (Figure 3.14C). The trend in this analysis is that if δ is low, the fraction of active cells would remain high across CTD lengths and TS intensity would overestimate the true burst size because sequential bursts would overlap; if δ is large, the fraction of active cells would remain low, and TS intensity would underestimate the true burst size because intensity would decrease before the end of a burst. Both of these effects were enhanced with longer CTDs. The experimental range of active cells fractions (Figure 3.3E) suggests a scenario where the estimate from TS intensity lies between a slight overestimation to an underestimation of the true burst size; in the latter situation, inferring burst size from these data would be a conservative estimate. In the context of the canonical transcription model (Figure 6A), the polymerase binding rate β intrinsically links burst size and frequency. We thus conclude that the simplest explanation consistent with simulations, experiments and previous literature is that burst size and frequency both decrease with CTD truncation.

Varying each of the model parameters individually (Figure 3.14E,F) offered an additional insight. By modulating the rate of polymerase phosphorylation ϕ – conceivably in local nuclear environments that limit CTD kinase activity– or by increasing the rate of self-interaction ϵ, it is possible to dramatically increase burst size under the many-polymerases model (Figure 3.6F). This effect would be a direct consequence of increased concentrations of unphosphorylated Pol II, akin to recently reported droplets in cells (Chong et al., 2018; Boehning et al., 2018; Cho et al., 2018; Nair et al., 2019). Although not specified in our model, liquid-liquid phase separation may thus naturally emerge from this transcription logic.

Our CTD-centric models helped us understand our experimental observations and integrate them with prior knowledge. By simulating them, we were able to capture

the behavior of transcription upon CTD truncation and subsequent fusion to LCDs, illuminating the role of CTD length and providing support for a novel molecular state where more than a single polymerase can bind the PIC.

3.3 Discussion

A CTD-centric model offers mechanistic insights into transcriptional bursting

RNA Polymerase II is essential and extremely conserved across eukaryotes, yet the amino acid length of its catalytic subunit varies dramatically (Figure 3.1A,B). As discussed above, the number of disordered amino acids in the CTD closely follows this length variation, increasing with genome size as coding sequences become more scattered (Figure 3.1C,D).

The influence of CTD length on transcription is consistent with a simple quantitative model based on known protein-protein interactions with Mediator, CTD phosphorylation and disorder-mediated self-interactions (Figure 3.6A). We explicitly assume that at this level of abstraction, the major, Poissonian source of stochasticity in transcription comes from the multimolecular assembly of the preinitiation complex (PIC) at the enhancer and the promoter, ocurring at the start of each burst and preceding Pol II recruitment. This assumption is motivated by the CTD's influence on both burst size and frequency (Figure 3.3C-F) and by extensive previous experimental evidence. PIC formation is a rate-limiting step in transcription (Kuras & Struhl, 1999; Li, Virbasius, Zhu, & Green, 1999) and transcription bursts are concomitant with enhancer-promoter interactions (Bartman et al., 2016; Chen et al., 2018). Roughly, the likely order of protein recruitment events upon activation is enhancer specific transcription factors, a single Mediator complex (Petrenko et al., 2016) and general transcription factors, and finally RNA polymerase followed by CTD kinases (Bryant & Ptashne, 2003; Krishnamurthy & Hampsey, 2009). In specifying our model, we put forward a view in which bursts are a result of recurrent Pol II binding to the assembled PIC, and inactivity periods a consequence of PIC disassembly (Figure 3.6A). This reductionist framework thus offers an intelligible perspective of the mechanism of eukaryotic transcriptional bursting.

CTD length cooperatively scales transcription

The probability of interaction between two genomic loci is highly dependent on the physical distance between them (Lieberman-Aiden et al., 2009). As a result, order-of-magnitude variation in genome sizes and in the physical spacing between enhancers and promoters represents a challenge: how does the transcription machin-

ery overcome an increasingly infrequent event? Our simulations suggest increasing CTD length can reduce the number of times an assembled PIC fails to recruit RNA Pol II and produce mRNA before disassembly (Figure 3.14E), and that self-interactions can considerably increase burst size (Figure 3.6E,F). By exploiting each rare assembly event, CTD length-enhanced recruitment and self-interactions could contribute to resolve the transcription scaling paradox. This compensatory mechanism would complement changes in genome organization (Szabo, Bantignies, & Cavalli, 2019), without which enhancer-promoter interactions may never occur in the first place.

In this context, CTD length bears an important distinction with the strength of self-interaction and polymerase recruitment rate, determined by interactions with the PIC. While the naive expectation is that these parameters should be correlated, devations from the consensus repeat YSPTSPS may provide a way to modulate them independently. This hypothesis could explain why fruit flies with a CTD of wild-type length that is made up entirely of consensus repeats do not survive, but animals with a yeast CTD remain viable (Lu et al., 2019). Based on this rationale, we speculate CTD length and sequence in eukaryotes coevolves with the physical spacing between enhancers and promoters, primarily determined by genome organization (Lieberman-Aiden et al., 2009; Szabo et al., 2019) and genome size.

CTD-CTD self-interactions link transcription activation to phase-separation

It is possible that FUS and TAF15 LCDs rescue CTD truncation through an alternative recruitment mechanism that does not involve self-interactions, given they can also function as transcription factors when fused to a DNA binding domain (Kwon et al., 2013). This hypothesis would nonetheless be consistent with our inference that CTD length modulates polymerase binding rate to the PIC, but it would not support a many-polymerases state. On the other hand, our data do not suggest rescue is driven by enhanced direct recruitment to the PIC, given the increase in fraction of active cells seems to be predominantly driven by burst size and not frequency (Figure 3.4K-N), while direct recruitment would enhance both parameters. Additionally, we find a correlation between LCD ability to bind liquid droplets (Figure 3.10D) and self-interact with the extent of phenotypic rescue upon fusing them to a truncated RNA Pol II (Figure 3.5A,C,D).

Self-interactions additionally offer a logical connection between the mechanism of transcription activation and liquid-liquid phase separation (LLPS). Our model

predicts that when self-interaction strength is large or the rate of polymerase release is small, large transcription bursts could emerge (Figure 3.6F), implying a high local concentration of unphosphorylated polymerases. This environment has been observed in LLPS droplets at super-enhancers of live cells (Chong et al., 2018; Boehning et al., 2018; Cho et al., 2018; Nair et al., 2019). A corollary of this idea is that the average gene and the super-enhancer gene can both be transcribed using the same mechanisms, but only the latter would manifest LLPS droplets as an epiphenomenon of enhanced polymerase recruitment or kinase exclusion. Super-enhancers would then be at the extreme of the distribution of burst sizes, which is consistent with the observation of only a few droplets per cell whose number does not nearly match the total number of transcribed genes. In this scenario, LLPS could result in emergent behaviors whose understanding would require a different quantitave framework; our model may not apply to these CTD-lengths but could provide a useful expectation to compare them with. In other words, CTD length variation may result in regimes of transcription activation governed by different dynamics.

Self-interactions support a multi-polymerase complex

The key proposition of our model that allows the incorporation of self-interactions is the existence of a molecular complex that can bind more than one RNA Polymerase molecule (Figure 3.6A, pink). Short-lived Pol II clusters observed in mammalian cells that overlap with active transcription sites and whose duration correlates with mRNA output (Cisse et al., 2013; Cho et al., 2016) could be a direct observation of this event. On the other hand, Pol II pausing appears to negatively correlate with transcription initiation (Shao & Zeitlinger, 2017; Gressel et al., 2017), which could suggest that new polymerases may not be able to bind an occupied promoter. Distinguishing the perhaps differential ability of PIC-bound and paused Pol II's CTD to self-interact would be helpful to understand the relationship of this observation with a many-polymerases state.

Pol II is released from the promoter upon CTD phosphorylation (Jeronimo & Robert, 2014; Wong et al., 2014), based on which we argue that CTD length influences release rate. Along this line, depletion of yeast CTD-kinase Kin28 causes an upstream shift in Pol II occupancy along genes (Wong et al., 2014), with a pattern that resembles proximal-promoter accumulation in metazoans (Adelman & Lis, 2012) and is consistent with a defective promoter escape. A conspicuously similar shift was observed upon mutating CTD's serine 5 (Collin, Jeronimo, Poitras, &

Robert, 2019), the specific CTD residue phosphorylated for transcription initiation (Eick & Geyer, 2013; Harlen & Churchman, 2017). We find self-interactions correlate with the efficiency of transcription (Figure 3.5). A sensible interpretation of these experiments is that decreasing CTD-kinases or their activity on the CTD lead to increased RNA Pol II at the promoter by extending the time window for self-interaction mediated recruitment. These observations raise the hypothesis that promoter accumulation of Pol II in metazoans (Adelman & Lis, 2012), congruently not observed in yeast (Steinmetz et al., 2006), could be contingent on a higher phosphorylation release-threshold linked to a long CTD and a multi-polymerase complex. Experiments that directly measure the number of polymerases that can bind the PIC, and how CTD length influences RNA Pol II occupancy profile would be highly informative in this regard.

In summary, our study integrates experimental results and simulations to explain how CTD length influences transcription activation. We revise the current model of transcription by providing evidence that self-interactions are a key feature in this process, intrinsically linked to a state in which multiple polymerases can bind the PIC. This line of reasoning offers a sound connection between a reductionist, concrete transcriptional logic and the emerging perspective of phase-separation, generating testable hypotheses that will further clarify the functional and evolutionary relevance of CTD length variation.

3.4 Acknowledgements

We thank Mitchell Guttman, Matt Thomson and members of the Sternberg laboratory for helpful discussions, Heun J. Lee, Andres Collazo and Giada Spigolon for imaging assistance, Igor Antoshechkin and Vijaya Kumar for RNA-seq experiments, and Steven Mcknight and Masato Kato for reagents. This work was supported by the Howard Hughes Medical Institute with which PWS was an investigator, by the Gordon Ross Medical Foundation and the Benjamin M. Rosen Graduate Fellowships, by the Biological Imaging Center at the Caltech Beckman Institute, and by the Millard and Muriel Jacobs Genetics and Genomics Laboratory.

3.5 Author Contributions

Conceptualization, P.Q.C., P.W.S; Methodology, P.Q.C, P.W.S., T.L.L.; Software, P.Q.C.; Formal analysis, P.Q.C., T.L.L.; Investigation, P.Q.C.; Data Curation, P.Q.C.; Writing – Original Draft, P.Q.C., P.W.S.; Writing – Review & Editing, P.Q.C., P.W.S., T.L.L.; Visualization, P.Q.C.; Supervision, P.W.S.; Funding Acquisition,

P.W.S.

3.6 Declaration of interests

The authors declare no competing interests.

3.7 Methods

Data analysis

Except when indicated, all programming, data extraction, wrangling, calculations and plotting were done using Python 3.7 with standard scientific libraries (Oliphant, 2007; Jones, Oliphant, Peterson, et al., n.d.; Millman & Aivazis, 2011). All scripts used in this paper are available in the following github repository: https://github.com/WormLabCaltech

Image analysis

Maximum-intensity projections were used for all z-stack images, sometimes generated and often visualized using Fiji (Schindelin et al., 2012).

Cells were segmented using local thresholds and the Watershed algorithm. Candidate 2D fluorescent peaks were detected and tracked using Trackpy (Allan, Caswell, Keim, & van der Wel, 2018) with minor adaptations.

For PP7 transcription dynamics imaged with low laser intensity, only the brightest peak per cell per frame was kept. A Gaussian-Process Classifier (GPC) trained with a set of manually classified images was then used to distinguish transcription sites from spurious peaks, only keeping those with a GPC probability of at least 0.5 (Figure 3.8A-C). Transcription intensity was expressed as the fold-change of peak over mean nuclear fluorescence. This metric yielded overlapping intensity distributions of the same strain imaged with different settings (Figure 3.8D-H). Autocorrelation analysis was carried out as previously described (Lenstra & Larson, 2016). Missing timepoints where no peak was detected were imputed using the intensity at the position of the previous spot. For snapshots and smFISH images taken with maximum laser intensity, in which signal-to-noise ratio was greater, manually determined intensity thresholds were used.

RPB1 bioinformatic analysis

RPB1 homologs were retrieved by searching Ensembl database (Zerbino et al., 2018) using the HMMER online tool (Finn, Clements, & Eddy, 2011) with default settings, starting with the yeast RPB1 protein sequence. Amino acid sequences

were analyzed for disorder locally using MobiDB-lite (Necci, Piovesan, Dosztányi, & Tosatto, 2017), which provides a consensus score derived from eight disorder predictors, in a machine running Unix Debian 4.9 and Python 2.7. Genome sizes and gene numbers were scraped from Ensembl websites using a custom script.

Genetic constructs

All constructs used in this paper were built using PCR amplification, Gibson (Gibson et al., 2009) or golden gate (Engler, Kandzia, & Marillonnet, 2008) assembly methods and verified by Sanger sequencing. Plasmids are listed in Table S1.

Wild-type LCDs of FUS (residues 1-214) and TAF15 (residues 1-208) were as previously defined (Kwon et al., 2013) and obtained by PCR amplification from human cell line 293T cDNA. Plasmids with coding sequences of previously reported FUS and TAF15 LCD tyrosine-to-serine mutants (Kwon et al., 2013) were a gift from Steven Mcknight.

The DNA sequence for CTD truncation repair templates was redesigned to facilitate PCR amplification, and together with yeast codon-optimized mScarlet coding sequence, synthesized as an Integrated DNA Technologies (IDT) gBlock and cloned into their respective vectors. sgRNAs were purchased as individual oligos, hybridized and cloned into pWS082 using golden gate assembly.

Strain Engineering

All transformations were carried out using the LiAc/SS Carrier DNA/PEG method (Daniel Gietz & Woods, 2002). Strains are listed in Table S2.

Strain YTL047A (Donovan et al., 2019) was generated by transforming diploid *S. cerevisiae* BY4743 with a PCR product containing the PP7 loop cassette and a loxP-kanMX-loxP marker, which was subsequently removed with Cre recombinase. A single allele of GAL10 was tagged. All strains used in this study are derivatives of YTL047A.

RPB1 modifications were engineered in both alleles using CRISPR-Cas9 with gRNAs of improved stability (Ryan et al., 2014), antibiotic-mediated selection of cells proficient in gap repair (Horwitz et al., 2015), and plamids from the Yeast MoClo Toolkit (Lee, DeLoache, Cervantes, & Dueber, 2015) modified by Tom Ellis's lab. sgRNA sequences are listed in Table S3.

CTD truncations, mScarlet and LCD RPB1 strains were generated by transforming YTL047A with 100 ng of BsmBI linearized and gel-purified Cas9-kanR plasmid

(pWS173), 200 ng of each EcoRV linearized sgRNA vector (pWS082 derivatives), and 2-5 ug linearized repair template, selected for with G418. Correctly modified strains were identified using PCR of zymolyase digested colony scrapes followed by Sanger sequencing.

Strains for live transcription imaging were generated by integrating a single copy of GFP-Envy (Slubowski, Funk, Roesner, Paulissen, & Huang, 2015) fused to PP7 coat protein under an rpl15A promoter and a functional ura3 gene into the ura3Δ0 locus by transforming PacI linearized pTL174 and selected for with plates lacking uracil. Strains for self-recruitment assays (Figure 3.5) were constructed in the same way, except integrating PP7-LCD-GFP fusions. smFISH of self-recruitment assays was done on YTL047A transformed with plasmids pTL092, pQC075 or pQC076 (Figure 3.5D).

Cell growth measurements

Optical density (OD) was measured at an absorbance wavelength of 600 nm for 16 to 24 hours every 15 minutes using 1:100 dilutions of overnight cultures in 150 uL of YPD in a Falcon flat-bottom 96-Well Clear Assay Plate with lid on a Biotek Cytation 3 microplate reader with 1000 rpm shaking at 30C.

Doubling times were estimated non-parametrically from time derivatives of OD measurements with Gaussian processes using previously described software (Swain et al., 2016).

Live fluorescence microscopy

All microscopy experiments were done using early to mid-log cultures (typically 5e6 to 1e7 cells/mL) growing at 30C with 250 RPM shaking.

mScarlet-RPB1 strains were imaged on 2% agarose pads on coverslips at room temperature immediatly after spinning down at 3600 RCF cultures growing in Synthetic Complete (SC) 2% Glucose media on a Zeiss Imager Z2 microscope with an Axiocam 506 Mono camera, 63x oil objective, 150ms exposure time, and 25% laser intensity.

Live transcription and self-interaction imaging was done on concanavalin-A-coated MatTek dishes at 30C as previously described (Lenstra & Larson, 2016) using an Leica DMI 6000 wide-field fluorescence microscope with an Andor Zyla 5.5 or a Hamamatsu Flash 4.0 v3 camera with a 100x oil objective. Cells were induced by adding galactose dissolved in 2 mL SC to 1 mL SC 2% raffinose for a final 3

mL SC 2% galactose and imaged immediately every 20 sec for around 1 hour (live transcription) or after 30 min for 20 min (self-interaction). Live movies were taken with 150 ms exposure, 9 z-stacks every 0.5 μm, and minimal laser intensity to avoid photo-toxicity. Snapshots were imaged once per field-of-view with maximum laser power, 150 ms exposure, and 9-15 manually set z-stacks every 0.5 μm.

RNA-seq

RNA-seq data are available at the Gene Expression Omnibus, accession number GSE140491.

RNA was extracted from mid-log cultures growing in SC 2% raffinose after 2h of 2% galactose or blank induction using Zymo Quick-RNA Fungal/Bacterial Microprep Kit (Catalog # R2010) lysed in an MP Biomedicals FastPrep-24 machine.

RNA integrity was assessed using RNA 6000 Pico Kit for Bioanalyzer (Agilent Technologies #5067-1513) and mRNA was isolated using NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB #E7490). RNA-seq libraries were constructed using NEBNext Ultra II RNA Library Prep Kit for Illumina (NEB #E7770) following manufacturer's instructions. Briefly, mRNA isolated from 1 μg of total RNA was fragmented to the average size of 200 nt by incubating at 94C for 15 min in first strand buffer, cDNA was synthesized using random primers and ProtoScript II Reverse Transcriptase followed by second strand synthesis using NEB Second Strand Synthesis Enzyme Mix. Resulting DNA fragments were end-repaired, dA tailed and ligated to NEBNext hairpin adaptors (NEB #E7335). After ligation, adaptors were converted to the 'Y' shape by treating with USER enzyme and DNA fragments were size selected using Agencourt AMPure XP beads (Beckman Coulter #A63880) to generate fragment sizes between 250 and 350 bp. Adaptor-ligated DNA was PCR amplified followed by AMPure XP bead clean up. Libraries were quantified with Qubit dsDNA HS Kit (ThermoFisher Scientific #Q32854) and the size distribution was confirmed with High Sensitivity DNA Kit for Bioanalyzer (Agilent Technologies #5067- 4626). Libraries were sequenced on Illumina HiSeq2500 in single read mode with the read length of 50 nt following manufacturer's instructions. Base calls were performed with RTA 1.13.48.0 followed by conversion to FASTQ with bcl2fastq 1.8.4.

RNA-seq quantification was performed using Kallisto (Bray, Pimentel, Melsted, & Pachter, 2016) with 200 bootstraps in sigle-end mode with average length of 300 bp and standard deviation of 20 bp. Differential expression analysis was done with

Sleuth (Pimentel, Bray, Puente, Melsted, & Pachter, 2017) using the linear models described in the results and supplementary sections.

smFISH

smFISH experiments were carried out as described previously (Trcek et al., 2012; Lenstra et al., 2015). TYE665 labeled PP7 probes were purchased from IDT. A set of 48 Quasar 570 labeled probes were designed to target the coding sequence of Gal3 and purchased from Biosearch Technologies. Probe sequences are listed in Table S3.

Mid-log yeast cultures were fixed with paraformaldehyde and permeabilized with lyticase. Hybridization solution with 0.1 uM probes, 10% dextran sulfate, 10% formamide, and 2x Sodium Saline Citrate (SSC) was used to hybridize probes in fixed cells for 4 hours at 37 C. Coverslips were washed twice for 30 min with 10% formamide, 2x SSC at 37C, followed by rinses with 2x SSC, and 1x PBS for 5 minutes. PLL-coated 18 mm diameter #1.5 thickness coverslips were purchased from Neuvitro, mounted on microscope slides using ProLong Gold or Glass Antifade Mountant with DAPI (Life Technologies).

smFISH samples were imaged at room temperature with maximum laser power, 300 ms exposure and 9-15 manually set z-stacks every 0.2 μm on the Leica microscope with 100x objective described above.

Stochastic simulations

Stochastic simulations were performed using software described in Bois and Elowitz, 2019 with minor modifications to extract burst start, end and size while generating Gillespie samples. Rates were chosen according to Bartman et al., 2019, with $\alpha = 1$, $\gamma = 3$, $\beta = 30$, $\epsilon = 10$ and $\phi = 100$. For trace visualization purposes, a rate of phosphorylated Pol II removal (elongation rate) $\delta = 1$ was used.

3.8 Supplementary Figures

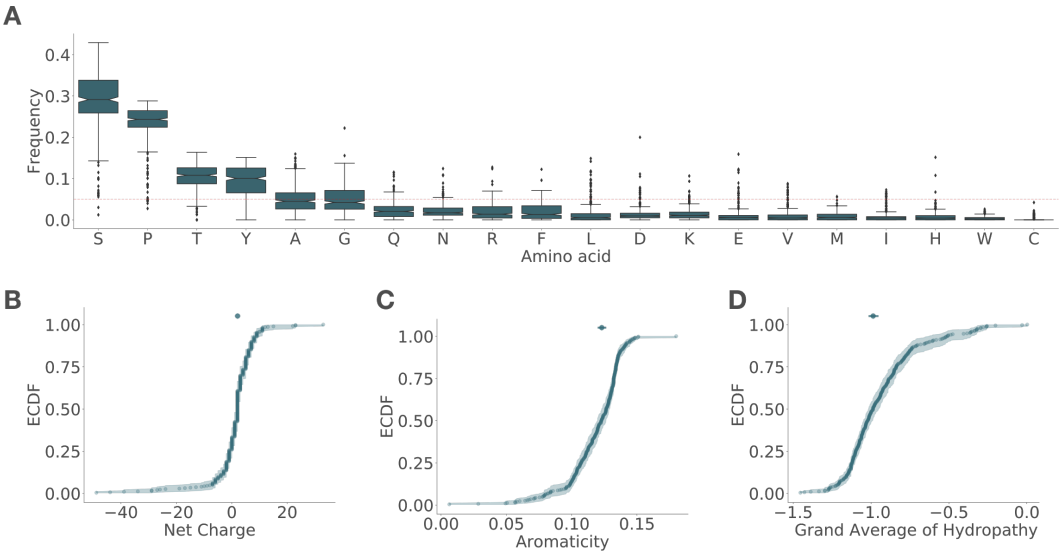

Figure 3.7: **CTDs share amino acid composition. Related to Figure 1.** CTDs were identified as the longest contiguous disordered region in RPB1 sequences. Only the longest protein per genus was considered. (A) Amino acid frequency sorted by mean abundance. Red dotted horizontal line indicates a uniform amino acid frequency of 1/20. Empirical cumulative distributions (ECDF) of net charges (B), aromaticity (C) and hydrophobicities (D) based on the grand average of hydropathy score (Kyte & Doolittle, 1982). Shaded area is bootstrapped 99% confidence interval (CI) and top markers show median with 99% CI.

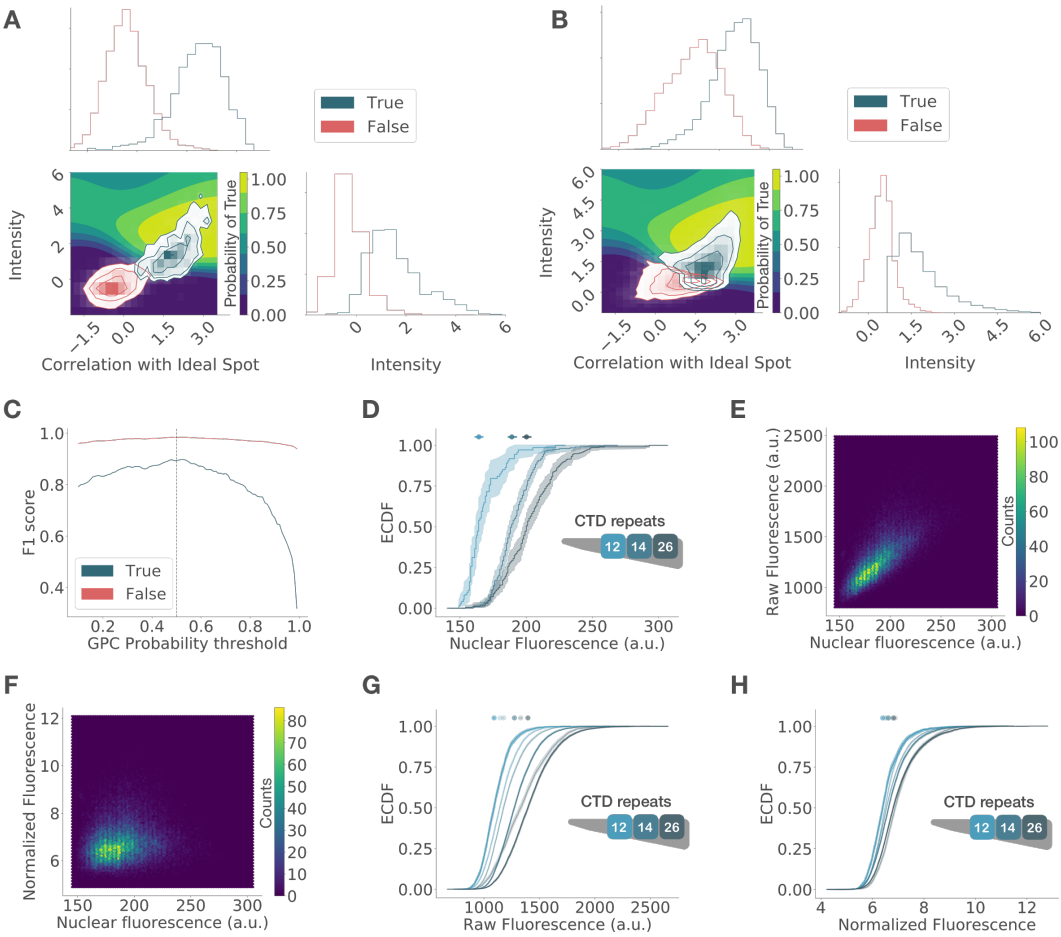

Figure 3.8: **Classification and normalization of PP7-GFP spots enables quantification of transcription dynamics and cross-strain comparisons. Related to figure 3.** Candidate spot images were obtained automatically using Trackpy's (Allan, Caswell, Keim, & van der Wel, 2018) peak detection algorithm. A sample of this image set was manually classified as True or False. For classification, spot images were represented using two features: correlation with an ideal spot (a single light point source blurred with a 2D Gaussian function) and intensity. (A) Histograms show the distribution of correlations (top) and intensities (right) of manually labeled spots. Left corner plot shows the joint distributions. This 2D data set was used to train a Gaussian-Process Classifier (GPC), resulting in the decision surface shown underneath, whose color indicates the probability of being a true spot. Candidate spots with a GPC probability above 0.5 were classified as True (B). This threshold was determined based on the change in the accuracy of classification (C), measured using the F1 score on a test set. The vertical dotted line indicates this probability threshold. Mutant strains show different PP7-GFP expression levels, as seen in the empirical cumulative distribution functions (ECDF) of mean nuclear fluorescence by strain (D). These differences result in a correlation observed in the hexagonal bin plot comparing mean nuclear fluorescence with raw spot fluorescence (E), which is removed after normalization (F). Normalized fluorescence is the ratio of spot fluorescence over mean nuclear fluorescence.

Figure 3.8: The efficacy of normalization can also be seen in the ECDFs of raw burst fluorescence by strain imaged with two laser intensities that artificially shift the intensity distributions of the same strains (G), which overlap after normalization (H). Transparency is used to indicate a different laser intensity. Shaded area is bootstrapped 99% confidence interval (CI) and top markers show median with 99% CI.

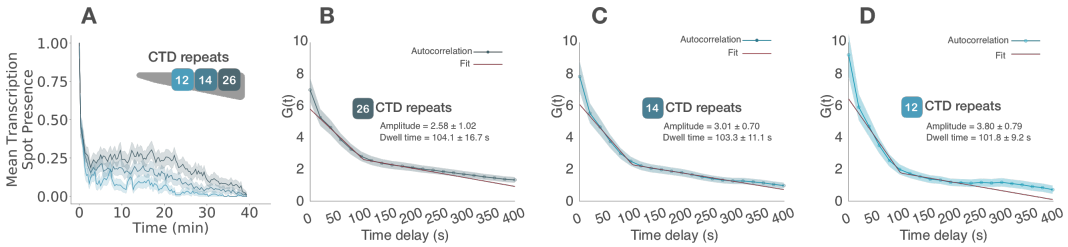

Figure 3.9: **Transcription burst frequency remains constant after activation and decreases with CTD truncation. Related to Figure 3.** (A) Mean aligned GAL10-PP7 boolean transcription traces. Boolean traces were obtained by marking with 1 and 0 the presence or absence of a transcription spot (TS), respectively. These traces were aligned and trimmed to begin with the first appearance of a TS and averaged over time, only considering cells that were active during the movie. These traces show the average frequency remains mostly constant over time and decreases with CTD length. Shaded area is bootstrapped 95% mean confidence interval. Frequency decay is also evident from an increase in amplitude, inversely related to frequency, in the autocorrelation of intensity traces corrected for non-steady-state effects in wild-type (B), 14 (C), and 12 (D) CTDr strains. Shaded area indicates standard error of the mean.

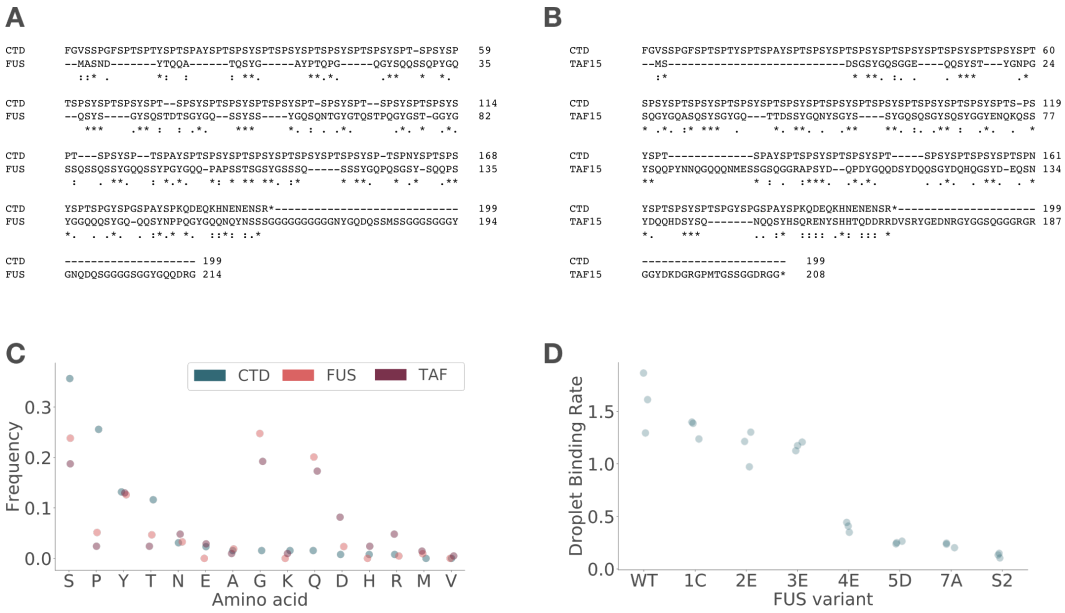

Figure 3.10: **FUS and TAF15 low complexity domains (LCD) are different in sequence but similar in amino acid composition to the CTD. Related to figures 4 and 5**. Protein alignments of FUS (A) and TAF15 (B) LCDs with yeast CTD. (C) Amino acid frequency in each of these proteins, sorted by CTD frequency. Only amino acids present in at least one protein are shown. (D) *In vitro* droplet binding rates of FUS variants used in this study. These numbers are the slopes obtained from a linear regression of LCD-GFP binding to wild-type FUS LCD droplets, measured as droplet fluorescence intensity over time. Each point is an experimental replicate; data are from Kwon et al., 2013.

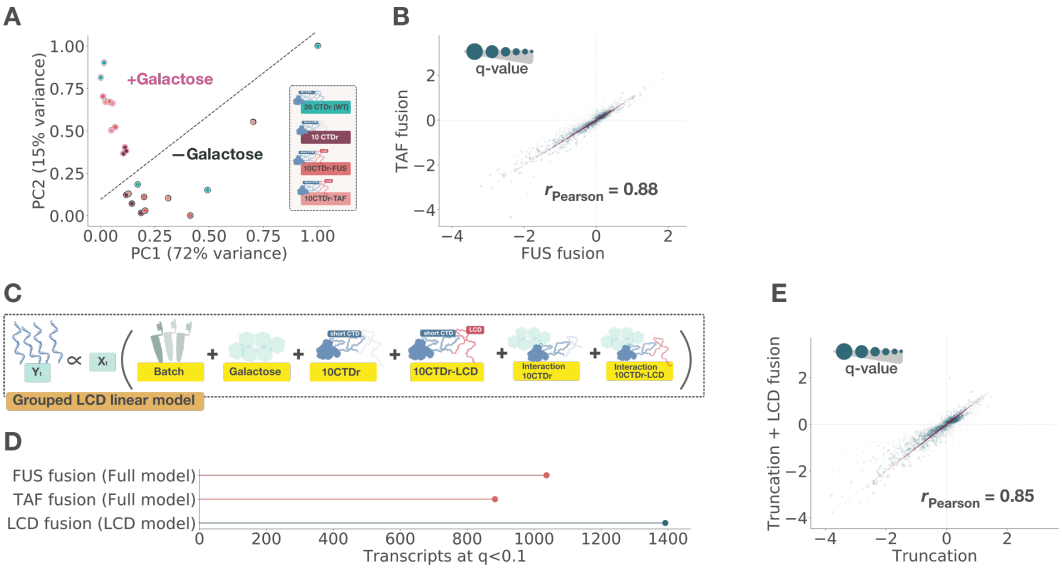

Figure 3.11: **Fusion of a CTD-truncated polymerase to FUS or TAF15 low complexity domains (LCD) results in convergent transcriptomes. Related to figures 2 and 4.** (A) Principal compoment analysis (PCA) with the first two PCs scaled to the range [0,1], which together explain 87% of the variance. Each strain has three biological replicates and two conditions. Marker edge color indicates the presence (pink) or absence (black) of galactose in the media; these groups are also divided by the dotted diagonal line. (B) Comparison of the log fold-change of each transcript resulting from FUS and TAF LCD fusion to 10CTDr truncated RNA Pol II under the full linear model shown in Figure 4B. Red points show the positions on the diagonal $x = y$. Marker size of each point is inversely proportional to the q-value of the interaction ($ms = -log(q_{int})$); dotted lines reference no change at zero and the Pearson correlation is indicated. (C) Alternative linear model where FUS and TAF rescued strains are grouped together. This grouping results in a higher number of genes identified for LCD fusion under a q-value threshold of 0.1 than for individual coefficients (D). Using this model, (E) comparison of the log fold-change of each transcript resulting from truncation with and without LCD fusion.

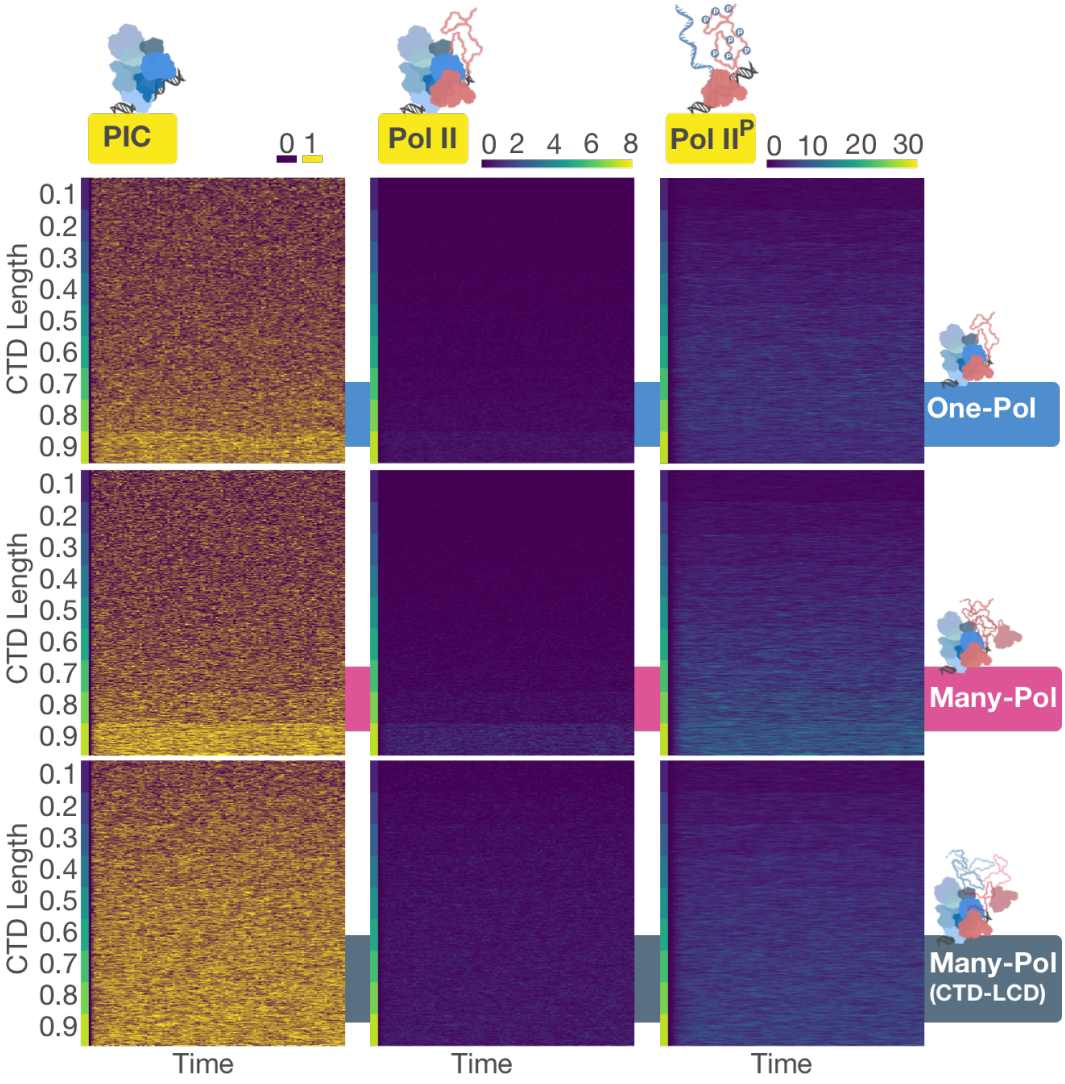

Figure 3.12: **Gillespie simulations yield traces akin to live transcription imaging. Related to figure 6.** Traces from stochastic simulations of PIC assembly states, number of PIC bound and phosphorylated (transcribing) polymerases for each model as a function of CTD_L, indicated with a colorbar to the left of each panel.

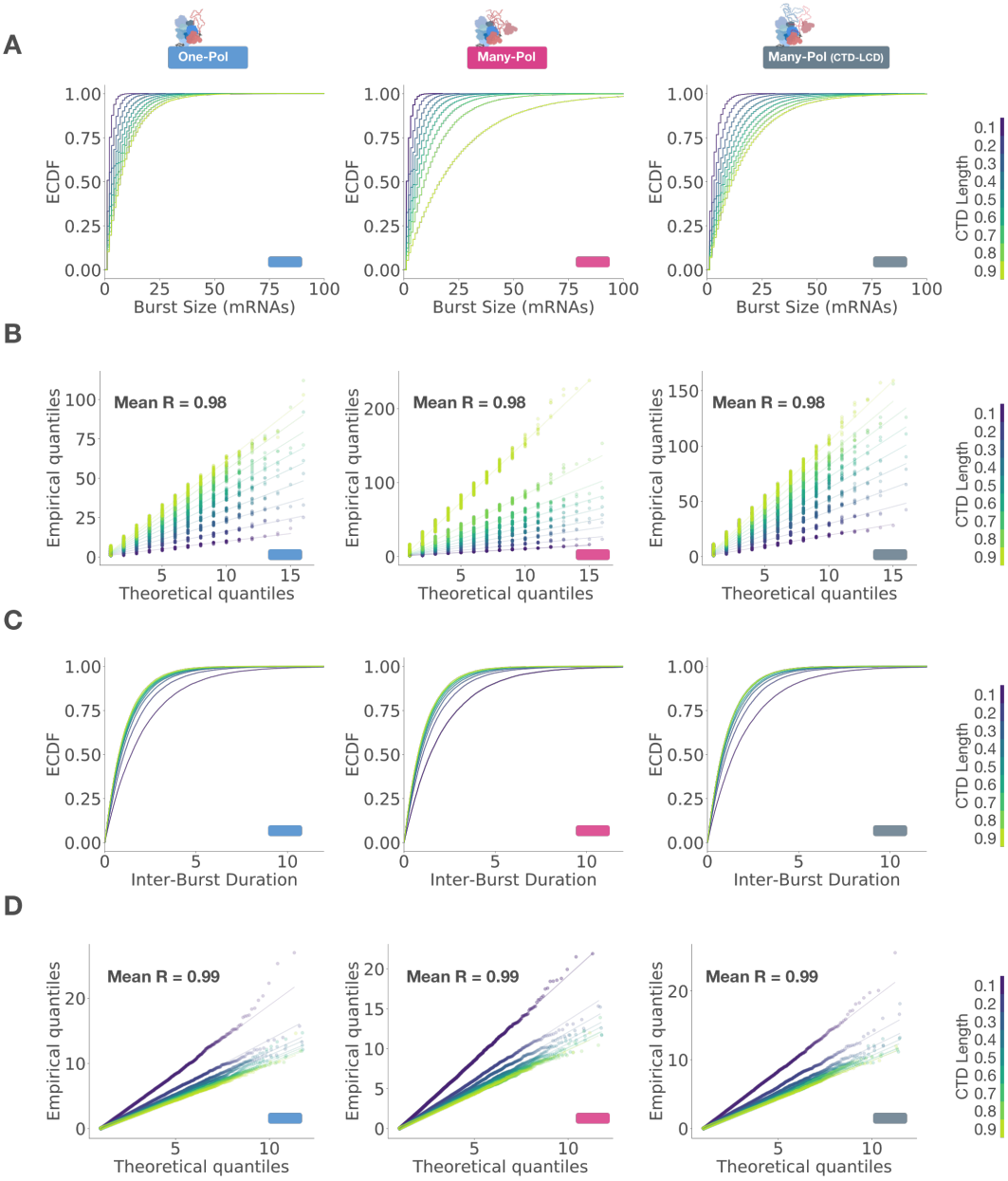

Figure 3.13: **Transcription models produce geometric and exponential distributions of burst sizes and inter-burst durations, respectively. Related to figure 6.** (A) Empirical cumulative distribution functions (ECDF) of burst sizes by CTD length for each model. (B) Q-Q plots comparing quantiles from simulated distributions and a geometric distribution. Similarly, (C) ECDFs of inter-burst durations and (D) Q-Q plots comparing their quantiles with an exponential distribution. Each column comes from the model indicated by the color on top and the lower right corner in each plot. The mean of the square root of the coefficient of determination (R) by model is indicated in each quantile comparison. CTD length is indicated by the color shown to the right of each row.

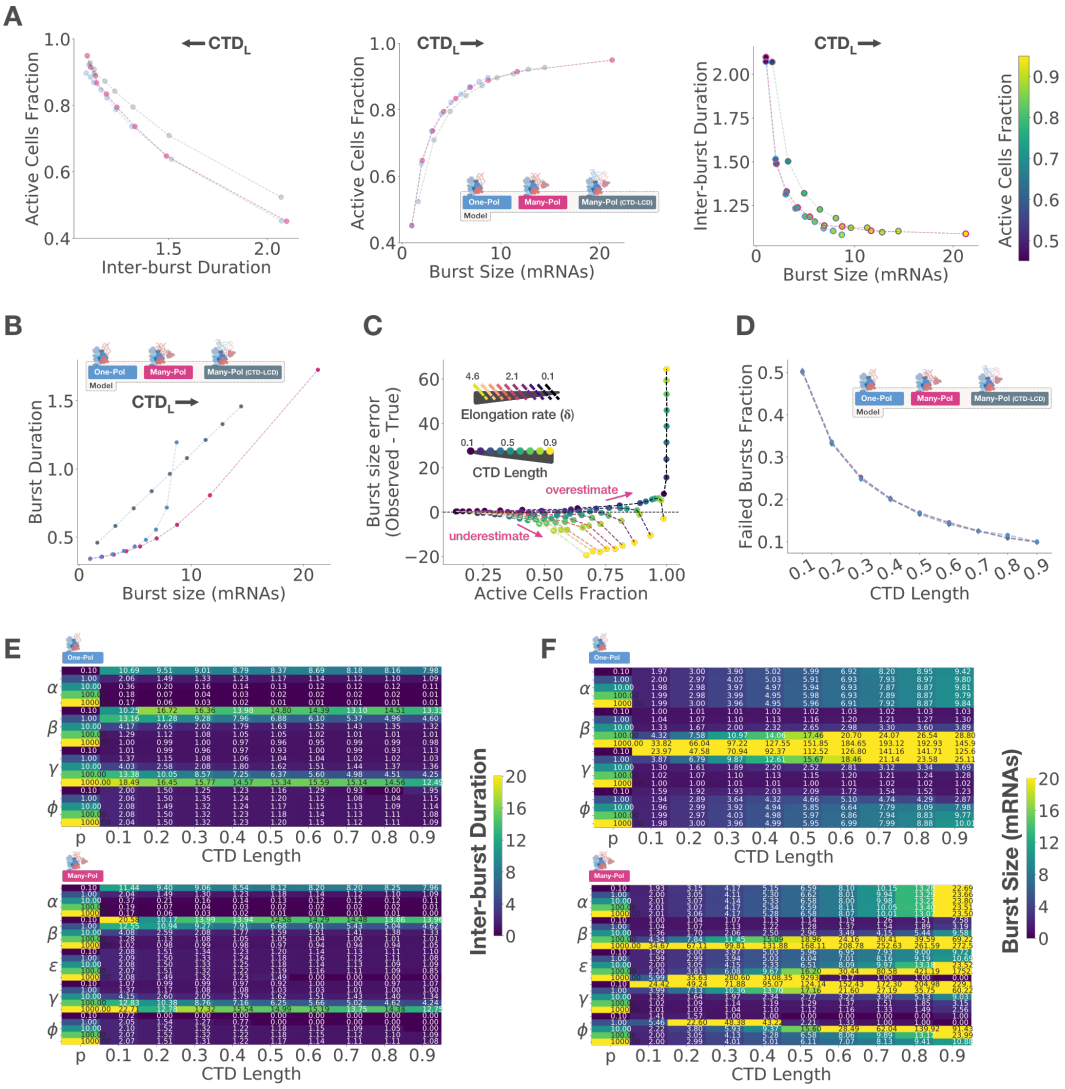

Figure 3.14: **Parameter exploration with stochastic simulations provides insights into experimental observations. Related to figure 6.** (A) Comparison of the mean active cells fraction with means of inter-burst duration (left), burst size (middle) and both of these numbers (right) with increasing CTD length (CTD_L) by model, indicated with color. Direction of CTD increase is indicated with an arrow on top of each plot. (B) Comparison of mean burst duration with mean burst size with increasing CTD length by model. (C) Comparison of the error in burst size estimate, computed as the difference between the means of the observed transcription site intensity and the true burst size, with the fraction of active cells as a function of CTD_L under the many-polymerases model. The elongation rate (δ) determines the time that a given mRNA spends bound to the transcription site and contributes to the observed intensity, thus influencing the fraction of active cells at a given time. (D) Comparison of fraction of failed bursts, where an assembled preinitiation complex produced zero mRNAs before disassembly, as a function of CTD_L by model. Error bars indicate 99% bootstrapped confidence interval. Mean inter-burst duration (E) and burst size (F) as a function of CTD_L and individually varying parameter values, while the others are held constant, as indicated in the first left column of each heatmap. Colormap is artificially fixed to the range [0-20] for visualization purposes and actual numbers are shown in each cell. Model is indicated in the top left corner.

BIBLIOGRAPHY

Adelman, K. & Lis, J. T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature reviews. Genetics*, *13*(10), 720–731. doi:10.1038/nrg3293

Allan, D. B., Caswell, T., Keim, N. C., & van der Wel, C. M. (2018). trackpy: Trackpy v0.4.1 (Version v0.4.1). Retrieved April 21, 2018, from http://doi.org/10.5281/zenodo.1226458

Allen, B. L. & Taatjes, D. J. (2015). The Mediator complex: a central integrator of transcription. *Nature reviews. Molecular cell biology*, *16*(3), 155–166. doi:10.1038/nrm3951

Allison, L. A. & Ingles, C. J. (1989). Mutations in RNA polymerase II enhance or suppress mutations in GAL4. *Proceedings of the National Academy of Sciences*, *86*(8), 2794–2798. doi:10.1073/pnas.86.8.2794

Aristizabal, M. J., Negri, G. L., Benschop, J. J., Holstege, F. C. P., Krogan, N. J., & Kobor, M. S. (2013). High-Throughput Genetic and Gene Expression Analysis of the RNAPII-CTD Reveals Unexpected Connections to SRB10/CDK8. *PLoS Genetics*, *9*(8), e1003758. doi:10.1371/journal.pgen.1003758

Banani, S. F., Lee, H. O., Hyman, A. A., & Rosen, M. K. (2017). Biomolecular condensates: Organizers of cellular biochemistry. *Nature Reviews Molecular Cell Biology*, *18*, 285–298. doi:10.1038/nrm.2017.7

Bartman, C. R., Hamagami, N., Keller, C. A., Giardine, B., Hardison, R. C., Blobel, G. A., & Raj, A. (2019). Transcriptional Burst Initiation and Polymerase Pause Release Are Key Control Points of Transcriptional Regulation. *Molecular Cell*, *73*(3), 519–532.e4. doi:10.1016/j.molcel.2018.11.004

Bartman, C. R., Hsu, S. C., Hsiung, C. C. S., Raj, A., & Blobel, G. A. (2016). Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping. *Molecular Cell*, *62*(2), 237–247. doi:10.1016/j.molcel.2016.03.007

Bartolomei, M. S., Halden, N. F., Cullen, C. R., & Corden, J. L. (1988). Genetic analysis of the repetitive carboxyl-terminal domain of the largest subunit of mouse RNA polymerase II. *Molecular and cellular biology*, *8*(1), 330–9. doi:10.1128/MCB.8.1.330

Boehning, M., Dugast-Darzacq, C., Rankovic, M., Hansen, A. S., Yu, T., Marie-Nelly, H., . . . Zweckstetter, M. (2018). RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nature Structural and Molecular Biology*, *25*(9), 833–840. doi:10.1038/s41594-018-0112-y

Bois, J. S. & Elowitz, M. B. (2019). Stochastic simulation of biological circuits. *Caltech Library*, https://doi.org/10.7907/V8SD–Q741. doi:https://doi.org/10.7907/V8SD-Q741

Boulon, S., Pradet-Balade, B., Verheggen, C., Molle, D., Boireau, S., Georgieva, M., . . . Bertrand, E. (2010). HSP90 and its R2TP/Prefoldin-like cochaperone are involved in the cytoplasmic assembly of RNA polymerase II. *Molecular Cell*, *39*(6), 912–924. doi:10.1016/j.molcel.2010.08.023

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. doi:10.1038/nbt.3519

Bryant, G. O. & Ptashne, M. (2003). Independent Recruitment In Vivo by Gal4 of Two Complexes Required for Transcription. *Molecular Cell*, *11*(5), 1301–1309.

Carre, C. & Shiekhattar, R. (2011). Human GTPases Associate with RNA Polymerase II To Mediate Its Nuclear Import. *Molecular and Cellular Biology*, *31*(19), 3953–3962. doi:10.1128/MCB.05442-11

Chapman, R. D., Heidemann, M., Hintermair, C., & Eick, D. (2008). Molecular evolution of the RNA polymerase II CTD. *Trends in Genetics*, *24*(6), 289–296. doi:10.1016/j.tig.2008.03.010

Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J. B., & Gregor, T. (2018). Dynamic interplay between enhancer–promoter topology and gene activity. *Nature Genetics*, *50*, 1296–1303. doi:10.1038/s41588-018-0175-z

Cho, W.-K., Jayanth, N., English, B. P., Inoue, T., Andrews, J. O., Conway, W., . . . Cisse, I. I. (2016). RNA Polymerase II cluster dynamics predict mRNA output in living cells. *eLife*, *5*, e13617. doi:10.7554/elife.13617

Cho, W.-K., Spille, J.-H., Hecht, M., Lee, C., Li, C., Grube, V., & Cisse, I. I. (2018). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science*, *361*(6400), 412–415. doi:10.1126/science.aar4199

Chong, S., Dugast-Darzacq, C., Liu, Z., Dong, P., Dailey, G. M., Cattoglio, C., . . . Tjian, R. (2018). Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science*, *361*(6400), eaar2555. doi:10.1126/science.aar2555

Chubb, J. R., Trcek, T., Shenoy, S. M., & Singer, R. H. (2006). Transcriptional Pulsing of a Developmental Gene. *Current Biology*, *16*(10), 1018–1025. doi:10.1016/j.cub.2006.03.092

Cisse, I. I., Izeddin, I., Causse, S. Z., Boudarene, L., Senecal, A., Muresan, L., . . . Darzacq, X. (2013). Real-Time Dynamics of RNA Polymerase II Clustering in Live Human Cells. *Science*, *341*(6146), 664–667. doi:10.1126/science.1239053

Collin, P., Jeronimo, C., Poitras, C., & Robert, F. (2019). RNA Polymerase II CTD Tyrosine 1 Is Required for Efficient Termination by the Nrd1-Nab3-Sen1 Pathway. *Molecular Cell*, *73*(4), 655–669.e7. doi:10.1016/j.molcel.2018.12.002

Coulon, A., Ferguson, M. L., de Turris, V., Palangat, M., Chow, C. C., & Larson, D. R. (2014). Kinetic competition during the transcription cycle results in stochastic RNA processing. *eLife*, *3*, e03939. doi:10.7554/eLife.03939

Couthouis, J., Hart, M. P., Shorter, J., DeJesus-Hernandez, M., Erion, R., Oristano, R., ... Gitler, A. D. (2011). A yeast functional screen predicts new candidate ALS disease genes. *Proceedings of the National Academy of Sciences*, *108*(52), 20881–20890. doi:10.1073/pnas.1109434108

Czeko, E., Seizl, M., Augsberger, C., Mielke, T., & Cramer, P. (2011). Iwr1 Directs RNA Polymerase II Nuclear Import. *Molecular Cell*, *42*(2), 261–266. doi:10.1016/j.molcel.2011.02.033

Daniel Gietz, R. & Woods, R. A. (2002). Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. In *Methods in enzymology* (pp. 87–96). doi:10.1016/S0076-6879(02)50957-5

Dobi, K. C. & Winston, F. (2007). Analysis of Transcriptional Activation at a Distance in Saccharomyces cerevisiae. *Molecular and Cellular Biology*, *27*(15), 5575–5586. doi:10.1128/MCB.00459-07

Donovan, B. T., Huynh, A., Ball, D. A., Patel, H. P., Poirier, M. G., Larson, D. R., ... Lenstra, T. L. (2019). Live-cell imaging reveals the interplay between transcription factors, nucleosomes, and bursting. *The EMBO Journal*. doi:10.15252/embj.2018100809

Eick, D. & Geyer, M. (2013). The RNA Polymerase II Carboxy-Terminal Domain (CTD) Code. *Chemical Reviews*, *113*(11), 8456–8490. doi:10.1021/cr400071f

Engler, C., Kandzia, R., & Marillonnet, S. (2008). A One Pot, One Step, Precision Cloning Method with High Throughput Capability. *PLoS ONE*, *3*(11), e3647. doi:10.1371/journal.pone.0003647

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, *39*, W29–W37. doi:10.1093/nar/gkr367

Gal, J., Zhang, J., Kwinter, D. M., Zhai, J., Jia, H., Jia, J., & Zhu, H. (2011). Nuclear localization sequence of FUS and induction of stress granules by ALS mutants. *Neurobiology of Aging*, *32*(12), 2323.e27–2323.e40. doi:10.1016/j.neurobiolaging.2010.06.010

Gerber, H. P., Hagmann, M., Seipel, K., Georgiev, O., West, M. a., Litingtung, Y., ... Corden, J. L. (1995). RNA polymerase II C-terminal domain required for enhancer-driven transcription. *Nature*, *374*(6523), 660–2. doi:10.1038/374660a0

Gibbs, E. B., Lu, F., Portz, B., Fisher, M. J., Medellin, B. P., Laremore, T. N., ... Showalter, S. A. (2017). Phosphorylation induces sequence-specific conformational switches in the RNA polymerase II C-terminal domain. *Nature Communications*, *8*(1), 15233. doi:10.1038/ncomms15233

Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A., & Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods*, *6*(5), 343–345. doi:10.1038/nmeth.1318

Gressel, S., Schwalb, B., Decker, T. M., Qin, W., Leonhardt, H., Eick, D., & Cramer, P. (2017). CDK9-dependent RNA polymerase II pausing controls transcription initiation. *eLife*, *6*, 1–24. doi:10.7554/eLife.29736

Hantsche, M. & Cramer, P. (2017). Conserved RNA polymerase II initiation complex structure. *Current Opinion in Structural Biology*, *47*, 17–22. doi:10.1016/j.sbi.2017.03.013

Harlen, K. M. & Churchman, L. S. (2017). The code and beyond: Transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nature Reviews Molecular Cell Biology*, *18*(4), 263–273. doi:10.1038/nrm.2017.10

Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K., & Sharp, P. A. (2017). A Phase Separation Model for Transcriptional Control. *Cell*, *169*(1), 13–23. doi:10.1016/j.cell.2017.02.007

Horwitz, A. A., Walter, J. M., Schubert, M. G., Kung, S. H., Hawkins, K., Platt, D. M., ... Newman, J. D. (2015). Efficient Multiplexed Integration of Synergistic Alleles and Metabolic Pathways in Yeasts via CRISPR-Cas. *Cell Systems*, *1*(1), 88–96. doi:10.1016/j.cels.2015.02.001

Huibregtse, J. M., Yang, J. C., & Beaudenon, S. L. (1997). The large subunit of RNA polymerase II is a substrate of the Rsp5 ubiquitin-protein ligase. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(8), 3656–61. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9108033%7B%5C%%7D0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC20496

Jeronimo, C. & Robert, F. (2014). Kin28 regulates the transient association of Mediator with core promoters. *Nature Structural & Molecular Biology*, *21*(5), 449–455. doi:10.1038/nsmb.2810

Jones, E., Oliphant, T., Peterson, P., et al. (n.d.). SciPy: Open source scientific tools for Python, http://www.scipy.org/ [Online, accessed 2019–09–30. doi:10.1109/MCSE.2007.58

Ju, S., Tardiff, D. F., Han, H., Divya, K., Zhong, Q., Maquat, L. E., . . . Petsko, G. A. (2011). A Yeast Model of FUS/TLS-Dependent Cytotoxicity. *PLoS Biology*, *9*(4), e1001052. doi:10.1371/journal.pbio.1001052

Kettenberger, H., Armache, K.-J., & Cramer, P. (2004). Complete RNA Polymerase II Elongation Complex Structure and Its Interactions with NTP and TFIIS. *Molecular Cell*, *16*(6), 955–965. doi:10.1016/j.molcel.2004.11.040

Kim, Y. J., Björklund, S., Li, Y., Sayre, M. H., & Kornberg, R. D. (1994). A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. *Cell*. doi:10.1016/0092-8674(94)90221-6

Krishnamurthy, S. & Hampsey, M. (2009). Eukaryotic transcription initiation. *Current Biology*, *19*(4), R153–R156. doi:10.1016/j.cub.2008.11.052

Kuras, L. & Struhl, K. (1999). Binding of TBP to promoters in vivo is stimulated by activators and requires Pol II holoenzyme. *Nature*, *399*(6736), 609–613. doi:10.1038/21239

Kwon, I., Kato, M., Xiang, S., Wu, L., Theodoropoulos, P., Mirzaei, H., . . . McKnight, S. L. (2013). Phosphorylation-Regulated Binding of RNA Polymerase II to Fibrous Polymers of Low-Complexity Domains. *Cell*, *155*(5), 1049–1060. doi:10.1016/j.cell.2013.10.033

Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, *157*(1), 105–132. doi:10.1016/0022-2836(82)90515-0

Lee, M. E., DeLoache, W. C., Cervantes, B., & Dueber, J. E. (2015). A Highly Characterized Yeast Toolkit for Modular, Multipart Assembly. *ACS Synthetic Biology*, *4*(9), 975–986. doi:10.1021/sb500366v

Lenstra, T. L., Coulon, A., Chow, C. C., & Larson, D. R. (2015). Single-Molecule Imaging Reveals a Switch between Spurious and Functional ncRNA Transcription. *Molecular Cell*, *60*(4), 597–610. doi:10.1016/j.molcel.2015.09.028

Lenstra, T. L. & Larson, D. R. (2016). Single-Molecule mRNA Detection in Live Yeast. In *Current protocols in molecular biology* (pp. 14.24.1–14.24.15). Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/0471142727.mb1424s113

Lesurf, R., Cotto, K. C., Wang, G., Griffith, M., Kasaian, K., Jones, S. J. M., . . . Griffith, O. L. (2016). ORegAnno 3.0: a community-driven resource for curated regulatory annotation. *Nucleic Acids Research*, *44*(D1), D126–D132. doi:10.1093/nar/gkv1203

Li, X.-Y., Virbasius, A., Zhu, X., & Green, M. R. (1999). Enhancement of TBP binding by activators and general transcription factors. *Nature*, *399*(6736), 605–609. doi:10.1038/21232

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., ... Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, *326*(5950), 289–293. doi:10.1126/science.1181369

Lu, F., Portz, B., & Gilmour, D. S. (2019). The C-Terminal Domain of RNA Polymerase II Is a Multivalent Targeting Sequence that Supports Drosophila Development with Only Consensus Heptads. *Molecular Cell*, *73*(6), 1232–1242.e4. doi:10.1016/j.molcel.2019.01.008

Marko, M., Vlassis, A., Guialis, A., & Leichter, M. (2012). Domains involved in TAF15 subcellular localisation: Dependence on cell type and ongoing transcription. *Gene*, *506*(2), 331–338. doi:10.1016/j.gene.2012.06.088

Mertins, P., Qiao, J. W., Patel, J., Udeshi, N. D., Clauser, K. R., Mani, D. R., ... Carr, S. A. (2013). Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nature Methods*, *10*(7), 634–637. doi:10.1038/nmeth.2518

Millman, K. J. & Aivazis, M. (2011). Python for Scientists and Engineers. *Computing in Science & Engineering*, *13*(2), 9–12. doi:10.1109/MCSE.2011.36

Nair, S. J., Yang, L., Meluzzi, D., Oh, S., Yang, F., Friedman, M. J., ... Rosenfeld, M. G. (2019). Phase separation of ligand-activated enhancers licenses cooperative chromosomal enhancer assembly. *Nature Structural & Molecular Biology*, *26*(3), 193–203. doi:10.1038/s41594-019-0190-5

Necci, M., Piovesan, D., Dosztányi, Z., & Tosatto, S. C. (2017). MobiDB-lite: Fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics*, *33*(9), 1402–1404. doi:10.1093/bioinformatics/btx015

Nonet, M., Sweetser, D., & Young, R. A. (1987). Functional redundancy and structural polymorphism in the large subunit of RNA polymerase II. *Cell*, *50*(6), 909–915. doi:10.1016/0092-8674(87)90517-4

Oliphant, T. E. (2007). Python for Scientific Computing. *Computing in Science & Engineering*, *9*(3), 10–20. doi:10.1109/MCSE.2007.58

Payne, J. M., Laybourn, P. J., & Dahmus, M. E. (1989). The transition of RNA polymerase II from initiation to elongation is associated with phosphorylation of the carboxyl-terminal domain of subunit IIa. *Journal of Biological Chemistry*, *264*, 19621–19629.

Petrenko, N., Jin, Y., Petrenko, N., Jin, Y., Wong, K. H., & Struhl, K. (2016). Mediator Undergoes a Compositional Change during Transcriptional Activation Article Mediator Undergoes a Compositional Change during Transcriptional Activation. *Molecular Cell*, *64*(3), 443–454. doi:10.1016/j.molcel.2016.09.015

Pimentel, H., Bray, N. L., Puente, S., Melsted, P., & Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods*, *14*(7), 687–690. doi:10.1038/nmeth.4324

Portz, B., Lu, F., Gibbs, E. B., Mayfield, J. E., Rachel Mehaffey, M., Zhang, Y. J., . . . Gilmour, D. S. (2017). Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain. *Nature Communications*, *8*(1), 15231. doi:10.1038/ncomms15231

Quintero-Cadena, P. & Sternberg, P. W. (2016). Enhancer Sharing Promotes Neighborhoods of Transcriptional Regulation Across Eukaryotes. *G3 (Bethesda, Md.) 6*(12), 4167–4174. doi:https://doi.org/10.1534/g3.116.036228

Robinson, P. J. J., Bushnell, D. A., Trnka, M. J., Burlingame, A. L., & Kornberg, R. D. (2012). Structure of the Mediator Head module bound to the carboxy-terminal domain of RNA polymerase II. *Proceedings of the National Academy of Sciences*, *109*(44), 17931–17935. doi:10.1073/pnas.1215241109

Robinson, P. J., Trnka, M. J., Bushnell, D. A., Davis, R. E., Mattei, P. J., Burlingame, A. L., & Kornberg, R. D. (2016). Structure of a Complete Mediator-RNA Polymerase II Pre-Initiation Complex. *Cell*, *166*(6), 1411–1422.e16. doi:10.1016/j.cell.2016.08.050

Ryan, O. W., Skerker, J. M., Maurer, M. J., Li, X., Tsai, J. C., Poddar, S., . . . Cate, J. H. (2014). Selection of chromosomal DNA libraries using a multiplex CRISPR system. *eLife*, *3*, e03703. doi:10.7554/eLife.03703

Sabari, B. R., Dall'Agnese, A., Boija, A., Klein, I. A., Coffey, E. L., Shrinivas, K., . . . Young, R. A. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. *Science*, *361*(6400). doi:10.1126/science.aar3958. eprint: https://science.sciencemag.org/content/361/6400/eaar3958.full.pdf

Scafe, C; Young, R. (1990). RNA polymerase II C-terminal repeat influences response to transcriptional enhancer signals. *Nature*, *374*(18), 685–689. doi:10.1016/0021-9797(80)90501-9. arXiv: NIHMS150003

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., . . . Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. *Nature Methods*, *9*(7), 676–682. doi:10.1038/nmeth.2019

Shao, W. & Zeitlinger, J. (2017). Paused RNA polymerase II inhibits new transcriptional initiation. *Nature genetics*, *49*(7), 1045–1051. doi:10.1038/ng.3867

Shin, Y. & Brangwynne, C. P. (2017). Liquid phase condensation in cell physiology and disease. *Science*, *357*(6357), eaaf4382. doi:10.1126/science.aaf4382

Shin, Y., Chang, Y.-C., Lee, D. S., Berry, J., Sanders, D. W., Ronceray, P., . . . Brangwynne, C. P. (2018). Liquid Nuclear Condensates Mechanically Sense and Restructure the Genome. *Cell*, *175*(6), 1481–1491.e13. doi:10.1016/j.cell.2018.10.057

Slubowski, C. J., Funk, A. D., Roesner, J. M., Paulissen, S. M., & Huang, L. S. (2015). Plasmids for C-terminal tagging in Saccharomyces cerevisiae that contain improved GFP proteins, Envy and Ivy. *Yeast*, *32*(4), 379–387. doi:10.1002/yea.3065

Somesh, B. P., Reid, J., Liu, W. F., Søgaard, T. M. M., Erdjument-Bromage, H., Tempst, P., & Svejstrup, J. Q. (2005). Multiple mechanisms confining RNA polymerase II ubiquitylation to polymerases undergoing transcriptional arrest. *Cell*, *121*(6), 913–923. doi:10.1016/j.cell.2005.04.010

Steinmetz, E. J., Warren, C. L., Kuehner, J. N., Panbehi, B., Ansari, A. Z., & Brow, D. A. (2006). Genome-Wide Distribution of Yeast RNA Polymerase II and Its Control by Sen1 Helicase. *Molecular Cell*, *24*(5), 735–746. doi:10.1016/j.molcel.2006.10.023

Svejstrup, J. Q., Li, Y., Fellows, J., Gnatt, A., Bjorklund, S., & Kornberg, R. D. (1997). Evidence for a mediator cycle at the initiation of transcription. *Proceedings of the National Academy of Sciences*, *94*(12), 6075–6078. doi:10.1073/pnas.94.12.6075

Swain, P. S., Stevenson, K., Leary, A., Montano-Gutierrez, L. F., Clark, I. B., Vogel, J., & Pilizota, T. (2016). Inferring time derivatives including cell growth rates using Gaussian processes. *Nature Communications*, *7*(1), 13766. doi:10.1038/ncomms13766

Szabo, Q., Bantignies, F., & Cavalli, G. (2019). Principles of genome folding into topologically associating domains. *Science Advances*, *5*(4), eaaw1668. doi:10.1126/sciadv.aaw1668

Takahashi, H., Parmely, T. J., Sato, S., Tomomori-Sato, C., Banks, C. A., Kong, S. E., ... Conaway, J. W. (2011). Human Mediator Subunit MED26 Functions as a Docking Site for Transcription Elongation Factors. *Cell*, *146*(1), 92–104. doi:10.1016/j.cell.2011.06.005

Thompson, C. M., Koleske, A. J., Chao, D. M., & Young, R. A. (1993). A multisubunit complex associated with the RNA polymerase II CTD and TATA-binding protein in yeast. *Cell*. doi:10.1016/0092-8674(93)90362-T

Trcek, T., Chao, J. A., Larson, D. R., Park, H. Y., Zenklusen, D., Shenoy, S. M., & Singer, R. H. (2012). Single-mRNA counting using fluorescent in situ hybridization in budding yeast. *Nature Protocols*, *7*(2), 408–419. doi:10.1038/nprot.2011.451

West, M. L. & Corden, J. L. (1995). Construction and analysis of yeast RNA polymerase II CTD deletion and substitution mutations. *Genetics*, *140*(4), 1223–1233.

Wong, K. H., Jin, Y., & Struhl, K. (2014). TFIIH Phosphorylation of the Pol II CTD Stimulates Mediator Dissociation from the Preinitiation Complex and Promoter Escape. *Molecular Cell*. doi:10.1016/j.molcel.2014.03.024

Yang, C. & Stiller, J. W. (2014). Evolutionary diversity and taxon-specific modifications of the RNA polymerase II C-terminal domain. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(16), 5920–5. doi:10.1073/pnas.1323616111

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., ... Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761. doi:10.1093/nar/gkx1098

Chapter 4

CONCLUDING REMARKS

The motivation for this thesis derived from the variation in nucleotide length of eukaryotic genomes spanning five orders of magnitude. We argued this range represents a concrete challenge that organisms must have evolved to overcome: transcription often requires that two or more DNA loci meet (Chen et al., 2018), but the probability of this event becomes exponentially smaller with longer DNA molecules (Ringrose, Chabanis, Angrand, Woodroofe, & Stewart, 1999). Interpreting the observation that gene neighbors are frequently correlated in expression across eukaryotes challenged this expectation (Quintero-Cadena & Sternberg, 2016), and the CTD's correlation with genome size offered a promising hypothesis to explain the paradox of life with long genomes (Allen & Taatjes, 2015; Quintero-Cadena, Lenstra, & Sternberg, 2020). In turn, exploring this hypothesis was a productive avenue to interrogate and better understand the mechanism of transcription.

The CTD has been extensively studied and found to be involved in many aspects of transcription, as a landing pad for protein-protein interactions and more recently as a molecular bridge into phase-separated bodies (Harlen & Churchman, 2017). Using yeast, we discovered that CTD length modulates the size and frequency of transcription bursts, and that different disordered protein domains can supplement the CTD's function at the transcriptional and physiological levels. Fusion to these protein domains reduced the minimum CTD length required for viability. Moreover, we identified that the ability to self-interact, whereby disordered proteins can form weak intermolecular interactions that collectively drive liquid-liquid phase separation, is crucial for the fusion-mediated rescue of CTD truncation (Quintero-Cadena et al., 2020).

We proposed to update the current transcription paradigm with two CTD-centric insights. This update integrates CTD-CTD self-interactions with length-mediated burst size and frequency modulation into an intelligible quantitative model. First, CTD length promotes polymerase recruitment to the promoter but slows down its release from it. Second, we added a novel molecular state, in which self-interactions facilitate secondary, cooperative recruitment of more than a single polymerase molecule. This multi-polymerase complex naturally constitutes a seed for phase

separation.

The repetitiveness of the CTD and the correlation of its length with measurable phenotypes remain a promising avenue to further push our mechanistic understanding of transcription. In particular, the work described herein focused on the short end of the spectrum of CTD lengths. A natural step forward is to test the effect of extending the CTD in yeast beyond the 26 wild-type repeats.

Preliminary results replacing yeast CTD with that of *C. elegans* and *H. sapiens* paint an interesting picture consistent with the model of cooperative recruitment. Galactose induction of the genes GAL10 and GAL3 was unexpectedly less effective with these longer CTDs (Figure 4.1A-D). Because these strains also exhibit slower growth than wild-type (Figure 4.1E), we hypothesized an underlying cause of the compromised galactose induction was a depletion of resources due to a widespread increase in transcription. The result of labeling total mRNA using poly-dT probes supported this hypothesis, with the worm and human CTD strains showing an overall increase of poly-A RNA content (Figure 4.1F, H).

These results are preliminary, should be replicated and interpreted in the context of other experiments. Moreover, interpreting these observations could be confounded by the accompanying change in cell size (Figure 4.1G), which could itself lead to physiological consequences that influence transcription. On the other hand, these results suggest a worm or human CTD in yeast leads to an unsustainable increase in transcription. For this reason, it is possible *in vitro* experiments, where molecular resources are not rate-limiting, could provide valuable complementary insights into the process.

It is currently unclear how precisely a human and a worm CTD differ from the yeast CTD. Evidence described in Chapter III supports that changes in transcription and correlated variables observed upon CTD truncations mostly result from changing two specific biophysical properties of this protein domain: the ability to self-interact, which could eventually but not always lead to liquid-liquid phase separation, and the affinity for other proteins in the transcription complex. Additionally, phosphorylation and potentially other post-translational modifications likely modulate each of these parameters. Specifically dissecting these two physical attributes of the CTD could significantly inform our understanding and interpretation of transcriptomic phenotypes resulting from CTD mutations. Single-molecule experiments designed to directly count the number of polymerases that can bind a transcription complex in the context of these parameters would be illuminating. In the evolutionary context,

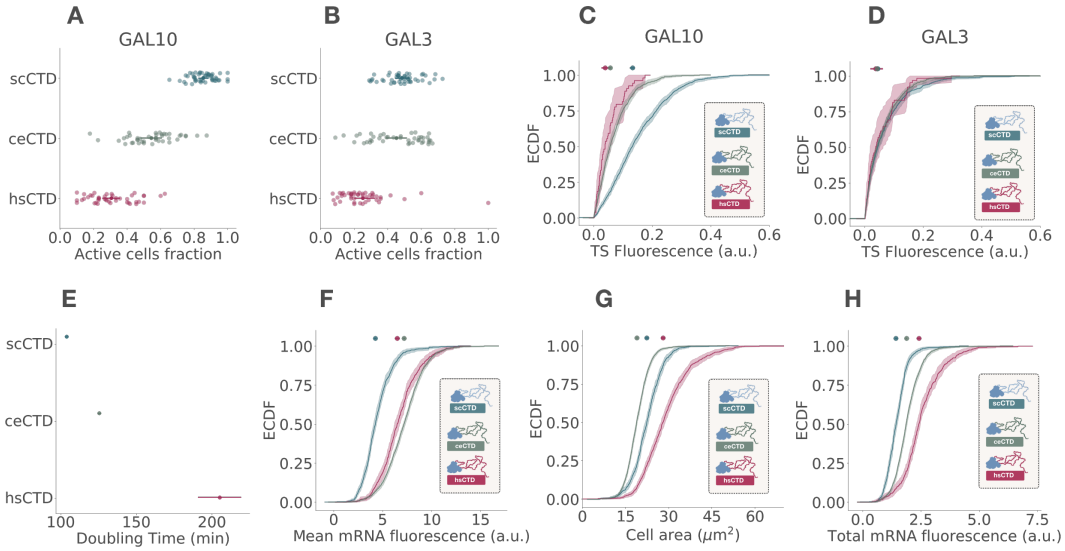

Figure 4.1: **Preliminary results: yeast strains with longer worm and human CTDs exhibit reduced transcriptional induction, slower growth, and increased total mRNA content.** All of the RNA measurements come from a single smFISH experiment and should therefore be replicated. Fraction of active cells per field of view after galactose induction for GAL10 (A) and GAL3 (B) for wild-type *S. cerevisiae* and strains with *C. elegans* and *H. sapiens* CTDs. Mean with 99% bootstrapped confidence interval (CI) is shown on top of each group. Corresponding empirical cumulative distribution functions (ECDF) with 99% bootstrapped CI of transcription site intensities of GAL10 (C) and GAL3 (D). Medians with 99% CI are shown on top. (E) Mean doubling time (DT) with standard error for each strain in YPD. ECDFs of mean poly-A mRNA fluorescence by cell using poly-dT probes (F), (G) of cell areas and (H) of total poly-A mRNA fluorescence summed over entire cells by strain.

the *Plasmodium* genus is a very interesting polymerase-outlier, from the perspective of the emergence of parasitism, amino acid sequence conservation (Kishore, Perkins, Templeton, & Deitsch, 2009) and of disorder conservation (Figure 4.2). If available, measurements of these two parameters and their correlation with transcriptional output could provide a clearer perspective of the mechanism of eukaryotic transcription and the evolutionary forces at play.

The interplay between CTD and CTD-kinases is another interesting avenue to explore the mechanism of transcription. One sensible hypothesis is that the accumulation of Pol II at the promoter in metazoans, commonly referred to as pausing, is a manifestation of long CTDs leading to the recruitment of multiple polymerases and to a delayed phosphorylation-dependent release from the promoter-bound transcription complex. A potentially enlightening experiment in this regard would be

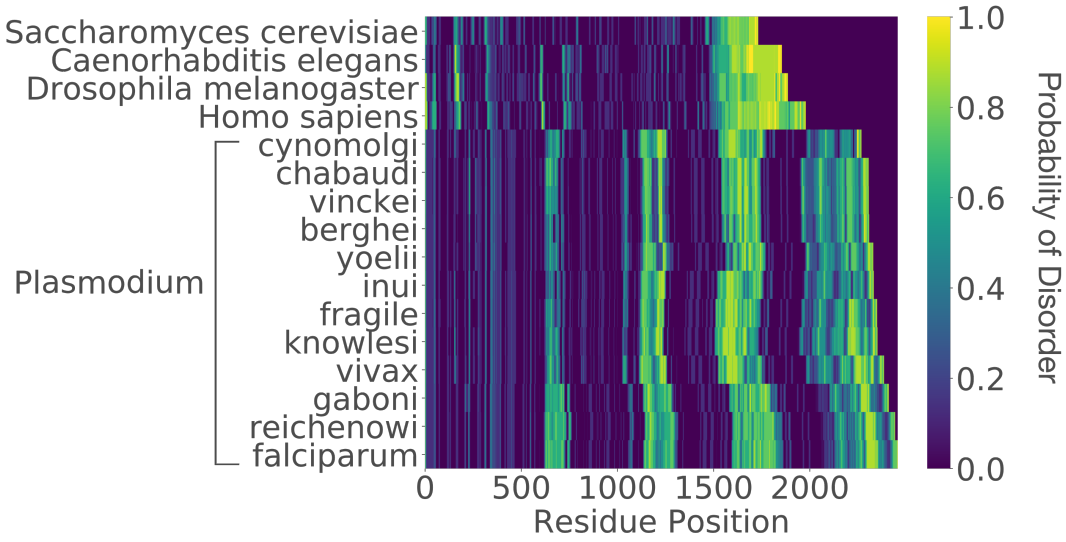

Figure 4.2: ***Plasmodium* RPB1 is an outlier from the perspective of disorder and among the longest in eukaryotes.** Distribution of disorder probability along RPB1 sequences in representative and available species in the *Plasmodium* genus sorted by length.

to quantify Pol II distribution along active genes, comparing long with short CTDs. Correlating the amount of available CTD-kinase with promoter accumulation could further illuminate the origin of pausing. Given its relatively short CTD, its resilience to change in CTD length, and the lack of Pol II pausing, yeast is likely an informative organism in which to conduct these experiments.

We consider the chemical reaction depicted in the last figure of Chapter III the most informative concept put forward in this thesis. Through this model, we attempt to precisely describe, integrate and assess our interpretation of experiments and previous literature. Arguably, its main usefulness lies in that it is falsifiable, by providing the concrete expectations described in this and the discussion section of Chapter III. We hope these expectations can help guide further work. Among the possible long-term goals, two that we considered especially helpful are to explain how one mechanism of transcription has evolved to function in the diversity of biochemical contexts that occur in eukaryotic nuclei, and to reconcile the classical paradigm of transcription, developed and supported by decades of work by the field, with the emerging perspective of phase separation.

BIBLIOGRAPHY

Allen, B. L. & Taatjes, D. J. (2015). The Mediator complex: a central integrator of transcription. *Nature reviews. Molecular cell biology*, *16*(3), 155–166. doi:10.1038/nrm3951

Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J. B., & Gregor, T. (2018). Dynamic interplay between enhancer–promoter topology and gene activity. *Nature Genetics*, *50*, 1296–1303. doi:10.1038/s41588-018-0175-z

Harlen, K. M. & Churchman, L. S. (2017). The code and beyond: Transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nature Reviews Molecular Cell Biology*, *18*(4), 263–273. doi:10.1038/nrm.2017.10

Kishore, S. P., Perkins, S. L., Templeton, T. J., & Deitsch, K. W. (2009). An unusual recent expansion of the c-terminal domain of RNA polymerase II in primate malaria parasites features a motif otherwise found only in mammalian polymerases. *Journal of Molecular Evolution*. doi:10.1007/s00239-009-9245-2

Quintero-Cadena, P., Lenstra, T. L., & Sternberg, P. W. (2020). RNA Pol II Length and Disorder Enable Cooperative Scaling of Transcriptional Bursting. *Molecular Cell*. doi:https://doi.org/10.1016/j.molcel.2020.05.030

Quintero-Cadena, P. & Sternberg, P. W. (2016). Enhancer Sharing Promotes Neighborhoods of Transcriptional Regulation Across Eukaryotes. *G3 (Bethesda, Md.) 6*(12), 4167–4174. doi:https://doi.org/10.1534/g3.116.036228

Ringrose, L., Chabanis, S., Angrand, P. O., Woodroofe, C., & Stewart, A. F. (1999). Quantitative comparison of dna looping in vitro and in vivo: chromatin increases effective dna flexibility at short distances. *EMBO J. 18*(23), 6630–6641. doi:10.1093/emboj/18.23.6630