Dewdrops on the genome:

Regulation of gene expression by biomolecular phase separation

by

Krishna Shrinivas

B. Tech (Honors) in Chemical Engineering,

Indian Institute of Technology Madras (2014)

Submitted to the Department of Chemical Engineering in partial

fulfilment of the requirements for the degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology,

September 2020

© Krishna Shrinivas, 2020, All rights reserved

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

Author:

……..

August 14th 2020

Certified by: ……

Arup K. Chakraborty, *Thesis supervisor*

Robert T. Haslam Professor of Chemical Engineering, Professor of Physics and

Chemistry, and Core Member, Institute for Medical Engineering and Science, MIT

Accepted by:

……

Patrick S. Doyle

Robert T. Haslam Professor of Chemical Engineering, *Graduate Officer*

This doctoral thesis has been examined by a committee of the

Department of Chemical Engineering as follows:

Professor Arup K. Chakraborty, ……

Robert T. Haslam Professor of Chemical Engineering, Professor of Physics and Chemistry, and Core Member, Institute for Medical Engineering and Science, MIT.
Thesis supervisor

Professor Karthish Manthiram, ……

Theodore T. Miller Career Development Chair,
Assistant Professor of Chemical Engineering.
Thesis presider

Professor Mehran Kardar, ……

Francis Friedman Professor of Physics,
Thesis committee member

Professor Phillip A. Sharp, ……

Institute Professor and Professor of Biology;
Member, Koch Institute for Integrative Cancer Research,
Thesis committee member

Dewdrops on the genome

Regulation of gene expression by biomolecular phase separation

Krishna Shrinivas

Submitted to the Department of Chemical Engineering on August 14th, 2020 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Chemical Engineering

Abstract

Human development and physiology depend on the coordinated function of thousands of cell types – for example, neurons, immune cells, and skin cells. Each cell-type contains an identical copy of the genetic material yet performs specialized and diverse functions, in large part, due to the selective expression of particular coding DNA-elements (genes) into RNA. Mutations or dysregulation in control of gene expression underlie many diseased states, including cancer and neurodegenerative disorders. Non-coding DNA elements called enhancers orchestrate the complex biochemical pathways that lead to precise activation of cell-type specific genes. Decades of advances in molecular biology have identified many of the key proteins and their interactions in these pathways. Yet, how dozens of proteins and their complex network of interactions are organized in space and time by enhancers to robustly relay regulatory information to their target genes remains one of the central puzzles of transcriptional control.

In this thesis, I will leverage approaches from statistical physics, simulation, and informatics, in synergy with experimentalists, to gain mechanistic insights into gene control through the lens of biomolecular phase transitions.

Proposal: I will introduce recent evidence that proteins and nucleic acids with certain features phase separate into two liquid phases, like oil from water, to compartmentalize cellular pathways. Employing a simple physical model, I will propose the phase separation of the transcriptional machinery explains established and recently observed puzzles underlying a class of enhancer elements called super-enhancers. Subsequently, I will describe studies performed in collaboration with the Young and Sharp labs that provide direct experimental evidence of transcriptional condensates model *in vivo.*

Mechanism: I then will describe our efforts to identify the mechanisms contributing to the formation of transcriptional condensates. By combining molecular dynamics, informatics, and experimental assays, we identify specific features encoded in DNA that enable spatio-temporally localized formation of condensates. I will discuss implications on the origins of enhancer activity.

Control: Here, we'll combine non-equilibrium models of phase separation, coacervate chemistry, and imaging data in cells to explore the dynamic control of transcription through it eventual outcome i.e. ATP-dependent synthesis of RNA. We propose a dual-feedback mechanism in which low levels of RNA synthesis promote condensate formation and higher levels trigger dissolution. I will close by discussing the ramifications of our model on two enigmatic features of transcription – the pervasive synthesis and degradation of non-coding RNA and discrete and bursty dynamics of mRNA synthesis.

I will conclude with a short summary and brief discussion on future work.

Thesis supervisor: Arup K. Chakraborty,
Title: Robert T. Haslam Professor of Chemical Engineering, Professor of Physics and Chemistry, and Core Member, Institute for Medical Engineering and Science, MIT.

Table of contents

Acknowledgments

My PhD has been challenging, illuminative, and ultimately, rewarding[1]. The last few years have cemented my belief that while scientific fact is observer-independent and rational, the process of discovery is far from it – it is filled with creative leaps, emotional roller-coasters, requires originality and demands unwavering passion[2]. The privilege of being engaged in this scientific pursuit is the greatest adventure of my lifetime and I owe a debt to many that I can't fully express, let alone repay.

Arup, thank you for everything. Your unfettered passion for science, your extraordinary intellect, and your originality are only exceeded by your kindness and generosity of spirit. I am grateful to have had the opportunity to train with you and will dearly miss our many wide-ranging discussions on science, history, and culture (and especially your penchant for citing a timely aphorism[3]).

Rick and Phil, I thank you both for your mentorship and patience in introducing a bumbling theorist to the beautiful world of gene regulation. Rick, your boldness in approaching scientific *terra incognita* has been a great source of inspiration[4]. Phil, your capacity for far-reaching insights and expansive understanding of the literature have left me in awe and have often sparked the ideas that are described in this thesis.

Mehran, attending your lectures on statistical physics were amongst the most fun times I had in grad school. To all my teachers, from high-school to now, your passion for science has fueled mine.

I am forever indebted to the Chakraborty Lab, an eclectic group of warm-hearted scientists[5]. To Dariusz, thanks for being the most patient teacher ever and for being *de facto* cluster admin – our group owes you a great deal for your tireless lead on this. To Renee, Kevin, and Kayla – I will fondly remember our many (sometimes hours-long) coffee breaks and conversations over ice-cream. John, Ang, Assaf, Julia, Sunny, and the many fellow AKC group members over the years - your comraderie and the resultant convivial lab environment have made coming into lab a joy every day. To Halima, Paul, Cecilia, and Pradeep – it was pleasure to serve as a guiding hand on your initial forays in science and I learnt a great deal from each of you. To Michelle and Renee, this lab runs because of your diligence and round-the-clock work, thank you.

[1] Paraphrasing *Nietzche* "the highs were high and the lows were frequent"

[2] Peter Medawar's lovely quote comes to my mind as I write this:
"There is no such thing as a Scientific Mind. Scientists are people of very dissimilar temperaments doing different things in very different ways. Among scientists are collectors, classifiers and compulsive tidiers-up; many are detectives by temperament and many are explorers; some are artists and others artisans. There are poet-scientists and philosopher-scientists and even a few mystics. What sort of mind or temperament can all these people be supposed to have in common?"

[3] My favorite continues to be "There are two kinds of scientists. Those who read the literature, and those who write it."

[4] And draws many interesting parallels with your adventurous globe-trotting exploits.

[5] Not often does one train in a lab where one's colleagues are so diverse, for example - high-energy physicist-turned-virologist and quantum-chemist-turned-immunologist!

To my many collaborators over the years, your honest exchange of ideas and critiques have greatly shaped my understanding of biology and contributed to many of the projects outlined in this thesis. To Ben, I am constantly amazed at your sharp wit and will miss our Sunday evening conversations on the vagaries of condensates in biology[6]. To Jon and Ozgur, our serendipitous adventure into the RNA-feedback project was a fun and close collaboration[7].

To Karthish, Will, Kim, and Ki-Joo, I am thankful to have been part of this amazing teaching team. To the students of 10.302, your passion for science, diversity of ideas, and work ethic renewed my zeal for teaching and research[8] in the midst of a soporific lull in my PhD.

There were numerous adventures that I went on with friends from MIT. To the practice-school team of '16, I had a great time globe-trotting across two continents with you[9] - including many hikes, culinary treats, and driving with our hearts in our hand through narrow roads. To my many seniors at MIT ChemE, especially Ankur and Karthick, thanks for your insights and feedback early in grad school and for our many retreats to the woods and Acadia. To the MIT Communication Lab, and in particular, Jess, Diana, Caitlin and the ChemE team, for creating a fertile forum and serving as excellent sherpas over my explorations on improving scientific communication. To all the wonderful colleagues in MIT ChemE, thank you for the many wonderful TGs, coffee chats, and the constant effort to improve ourselves as a whole.

To my friends, I owe much. To the members of the many different biking, and hiking groups over the years, especially Mihir, Ninad, and Preeti, thanks for the excellent company over our many hours of exploration away from the city. To Rushina, Nigamaaa, Krithika, and Raja – I fondly remember our many chai sessions, "spontaneous" lunches, board-game nights, and music sessions. To GK, thank you for being an excellent friend and a wonderful inspiration – you constantly make me push my boundaries. To Shradhu and Gayathri, you brought a whiff of Madras in the brief time you were here in Cambridge and I loved our ambles to Harvard, our rambling conversations on every imaginable topic, and your humoring of my constant search for better ice-cream. To Sandy, Maddy, Nikhila, Swaroop, and my friends from school days – thank you for opening your homes to me – you are family.

[6] Hopefully our tradition of outlandishly long emails will continue

[7] I especially will remember our many marathon zoom sessions

[8] The student-nominated teaching award that you afforded me will be one of my most treasured honors from graduate school.

[9] And to whom, I owe my discovery of the black poison that is Guinness

To my family, thank you for your constant and unconditional support over the years. Appa, thank you for helping discover my deep passion for classical music. Amma, thank you for making me embrace and reflect on my humanity. Durai, thank you for being a friend, philosopher, and guide on all aspects life-related. Ramanna, you are an amazing sibling and a creative *tour-de-force* and your zen attitude is inspirational, even if not completely emulatable. To Appa, Amma, Ramanna, and Durai, it is your compassion, empathy, and upbringing that make me who I am today.

For the harmony of the world is made manifest in Form and Number, and the heart and soul and all the poetry of Natural Philosophy are embodied in the concept of mathematical beauty.

- Sir D'Arcy Wentworth Thompson

Introduction

"The Moving Finger writes; and, having writ,
Moves on: nor all thy Piety nor Wit
Shall lure it back to cancel half a Line,
Nor all thy Tears wash out a Word of it."

- Omar Khayyám, Rubaiyat

The 20th century harkened seismic shifts in our understanding of world: the discovery of the atom, the invention of the modern computer, and solving the structure of DNA, to list but a few. Many (if not all) of these advancements were made possible by development of new techniques and cross-pollination of expertise across disciplines[1]. The increasing convergence of physical and life sciences in current times, set against the backdrop of remarkable technological advances in sequencing and imaging, has inspired and enabled the work presented in this thesis.

This thesis aspires to make a contribution towards a better understanding of how genes are regulated in health and disease. We'll seek to elucidate how the complex pathways underlying control of genes are organized within the cell – through the lens of biomolecular phase transitions[2]. Towards this end, we'll borrow, modify, and develop (occasionally with success) approaches from statistical physics. Frequent back-and-forth between experiment and theory will be the central *leitmotif* of this work.

The ensuing parts of the introduction will aim to provide primers on the following subjects (whose intersection will be the topic of this thesis):

(1) Gene regulation and cellular identity:

(2) Membraneless organelles or biomolecular condensates, and

(3) Statistical physics of liquid-liquid phase transitions.

The primers will provide brief historical context, background, and describe principal concepts. For the interested reader, links to reviews, papers, and historical scientific literature will be provided for a deeper dive into these topics.

[1] For example, the interplay of Boolean algebra and discrete mathematics, combined with the discovery of the transistor enabled development of modern computing machines. Similarly, bringing together approaches from crystallography, biochemistry, and physical model-building served as the building-blocks of modern molecular biology and facilitated the seminal discoveries of Crick, Watson, Franklin, and Wilkins (amongst others).

[2] Phase transition is the process by a which a state of matter transits into a different (or many different) states. For example, when heating water beyond 100C, water will convert to steam, an example of a liquid-gas phase transition. In this thesis, we will largely focus on liquid-liquid phase transitions (like oil demixing from water solutions).

Armed with this background, we will posit a new model for regulation of gene expression in higher organisms, organized in the following manner.

Hypothesis: A role for phase separation in transcription

Mechanism: Genomic encoding of transcriptional regulation by phase separation

Control: RNA-mediated non-equilibrium feedback regulation of transcriptional condensates

Evolution: Origins of weak cooperative interactions as a dominant mode of biological specificity in higher organisms

Gene regulation and the basis of cellular identity

The development and establishment of molecular biology in the second half of the twentieth century has led to key insights – solving the structure of DNA, deciphering the "code" for protein synthesis (and the central dogma), establishing the structure-function paradigm of molecular machines (proteins and RNAzymes), and many others[3] . More recently, the remarkable advancement in sequencing technologies[4] has enabled "reading"[5] of the genetic code[6] – written in the alphabet of nucleotides (DNA) and spanning many billion letters.

A fundamental question, and one of pertinence to this thesis, lies in understanding how the genetic blue-print encodes the rules for development and physiology, and how mistakes in decoding this blue-print lead to disease and dysregulation. In particular, we'll focus on the biochemical pathways that enable specific, precise, and regulated expression of genes.

Regulation of gene expression in mammals, enhancers

Development and physiology of organisms depend on the coordinated function of distinct cell-types that perform specific yet diverse functions[7]. Cell-type or cellular identity[8], in turn, is dictated in large part by

[3] Please refer (Judson, 1979) for a detailed and accessible overview of this "golden era" of molecular and structural biology.

[4] Cost to sequence DNA nucleotides have been decreasing faster than Moore's Law – which itself is exponential.

[5] Crispr/Cas technologies hold great promise for a suite of techniques that facilitate "writing" (or "re-writing") the genetic code

[6] See (Mukherjee, 2016) for an engaging overview of genetics, genes, and more recent technologies

[7] Even in single-celled organisms, the cell needs to perform many distinct functions that are coordinated by different gene programs

[8] A cell that performs particular tasks and interacts in distinct manners with its surroundings i.e. a unique spatio-temporal phenotype of the cell

the specific parts of its DNA genome that are transcribed into RNA[9] and those that are silenced[10]. Over the years, many of the key players involved in transcription have been uncovered through a combination of molecular, genetic, and structural approaches. An exhaustive overview of these mechanisms, as well as links to the various relevant reviews can be found elsewhere (Cramer, 2019). Below, we summarize the key processes involved in mammalian control of transcription.

Transcription of a particular gene (DNA) into RNA is carried out by enzymes known as RNA polymerases. In multi-cellular organisms, non-coding[11] DNA elements called enhancers[12] orchestrate the spatio-temporal processes that eventually lead to the recruitment of RNA Polymerases to the cell-type specific genes. Enhancers are often physically located far away from genes on the chromosomes (Figure 1) and are bound by transcription factors (TFs), which recognize short and specific DNA sequences. These genomic addresses then serve as platforms for the robust assembly of the dozens of proteins including coactivators and parts of the transcriptional apparatus (Cramer, 2019; Lee and Young, 2013; Levine et al., 2014; Orphanides and Reinberg, 2002; Raser and O'Shea, 2004; Tjian and Maniatis, 1994) which result in activation of the target gene. Imprecision or dysregulation in this process underlie pathological and diseased states, including cancer and neurodegenerative disorders (Lee and Young, 2013; Smith and Shilatifard, 2014).

[9] The process by which DNA is transcribed into RNA, which subsequently is translated into proteins – the "molecular machines" of the cell

[10] The human genome consists of ~20,000 genes (which are the blue-prints of proteins - molecular machines/enzymes that perform particular tasks). These coding genes represent < 2% of our DNA and the vast majority of our genome is non-coding. When specific genes are expressed (and others silenced), a particular combination of functions/tasks are performed, underlying the emergence of a cell's identity. By expressing a different complement of genes, a distinct cellular identity can emerge.

This is analogous to choosing and leaving out different ingredients from the pantry in distinct combinations to make specific dishes. By changing combination of ingredients or removing some, one can make a different dish!

[11] Parts of our genome that are not coding for genes (i.e. they don't go onto to express proteins)

[12] Discovered over 30 years by the labs of Susumu Tomegowa and David Baltimore, amongst others. For a historical perspective, please refer (Schaffner, 2015)

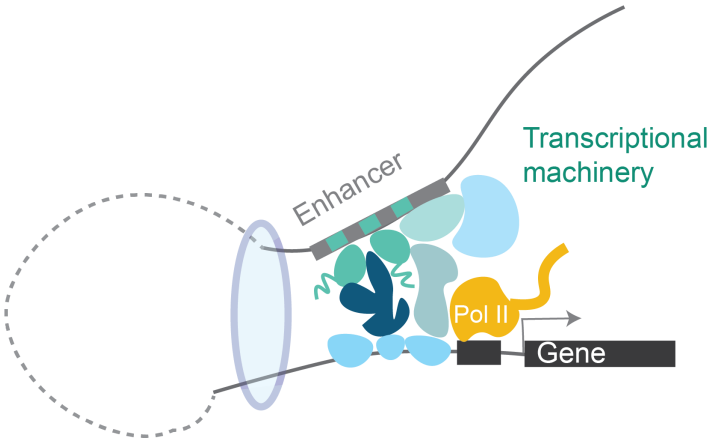


Figure 1 Schematic of gene regulation by enhancers. Enhancers (grey box) are non-coding DNA elements that are bound by transcription factors (green binding sites) and recruit various proteins of the transcriptional apparatus and RNA polymerase (Pol II) to activate the target gene that are often physically distant.

Advances over the last few decades in single-cell genomics, imaging, and sequencing have shed light on some of the key pieces that contribute to this regulatory mechanism – including investigation of the role of genome topology (most prominently vis-à-vis looping of DNA), epigenetic histone modifications, cooperativity in TF binding, nucleosomal reorganization, chromatin accessibility, sequence-specificities in promoters, and identified many key molecular components and processes[13] (Furlong and Levine, 2018; Haberle and Stark, 2018; Krijger and De Laat, 2016; Lee and Young, 2013; Levine et al., 2014; Long et al., 2016; Ong and Corces, 2011; Reiter et al., 2017; Spitz and Furlong, 2012; Tjian and Maniatis, 1994). Yet, how enhancers organize dozens of molecules and thousands of possible interactions in space and time to relay regulatory information to their target genes is one of the central puzzles of transcriptional regulation (Furlong and Levine, 2018).

[13] These references more or less point to reviews which then provide much greater depth on individual topics

In this thesis, we will seek to further our understanding of gene control and enhancer function through the lens of phase separation and biomolecular condensates (primers on both subjects are described below). We will provide begin by exploring a phase separation model in regulation of a class of enhancers (Chapter 1), investigate and characterize the mechanisms of enhancer assembly (Chapter 2), and investigate the role of RNA synthesis as a regulatory axes of feedback control on gene expression (Chapter 3).

Membraneless organelles

Essential to life is the organization and control of material through time and space. A central example is the cell (Figure 2), the ubiquitous minimal unit of living systems[14], which organizes all the processes that ultimately enable reproduction, heritability, and biological function.

[14] *Modulo* viruses

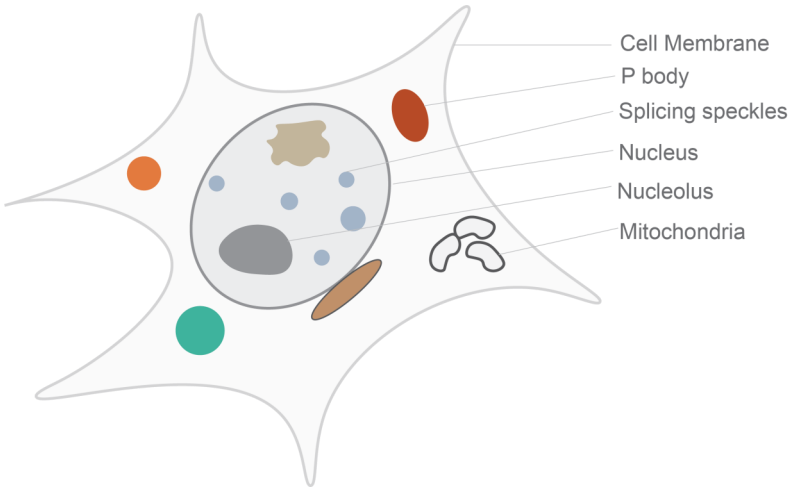


Figure 2 Schematic of a typical mammalian cell. Membranes are highlighted with solid lines around the interface of structures (cell membrane, around the nucleus and mitochondria) and membraneless compartments lack a physical interface (as is case with the P body, nucleolus and the splicing speckles)

Intra-cellular organization

Cells must organize and coordinate thousands of biomolecular species and reaction pathways within their interiors[15] . A prevalent modality of intra-cellular control is sub-compartmentation of related enzymatic pathways within organelles. Membrane-bound compartments are classic examples – sequestration of the genetic material to the nucleus, energy production pathways to the mitochondria or the "powerhouse", and the cellular contents themselves from the outside through the cell membrane. Membranes, which are made of lipids, serve as impermeable physical barriers and permit selective transport through specialized machinery (Figure 2). However, many organelles and cellular compartments (often labeled as bodies, puncta, inclusions, factories, or granules) lack membranes yet concentrate specific proteins and nucleic acids (Figure 2). In fact, membraneless organelles have been documented for nearly two centuries –from the description of the nucleolus in the 1830s by bright-field microscopy[16] (Wagner, 1835), cytological studies around the turn of the 19th century by Balbiani, Montgomery, and Cajal[17] (Balbiani, 1881; Cajal and others, 1903; Montgomery, 1900), and many other examples in the nucleus and the cytoplasm across different organisms (Dellaire et al., 2006; Gall, 2000; Mao et al., 2011; Nizami et al., 2010; Pederson, 2011; Spector and Lamond, 2011). Despite their widespread prevalence (membraneless organelles are found in nearly all eukaryotic organisms) a mechanistic framework to describe their behavior, formation, and function has remained elusive until this past decade.

[15]The interiors of a cell are far from simply being a "watery bag of enzymes"- (Mathews, 1993) discusses this topic in-depth

[16] See (Pederson, 2011) for a thorough history of the nucleolus

[17]See (Gall, 2000) for a thorough history of the Cajal body

A breakthrough in our understanding of membraneless organelles stemmed from a series of original and creative experiments probing P granules in the popular model organism *C. elegans* (Brangwynne et al., 2009). P granules were found to behave like liquids – exhibiting characteristics of liquid droplets suspended in the cellular cytoplasm– including dynamic exchange of constituents, flow under shearing, fusion upon contact, and spherical rounding to minimize surface tension. Subsequently, many other membraneless organelles have been characterized to exhibit many of the same liquid-like characteristics[18] (reviewed extensively in Banani et al., 2017; Bergeron-Sandoval et al., 2016; Shin and Brangwynne, 2017). Membraneless organelles, which are collective ensembles of biological matter, often resemble liquid-like droplets that are suspended in another liquid (the cyto or nucleoplasm, for example). In fact, the proposition that the cell's interiors are micro-compartmented into immiscible phases is not a new one - early speculations by interdisciplinary biologist EB Wilson (Wilson, 1899) on the structure of the protoplasm, propositions by polymaths Oparin and JBS Haldane on the coacervate model of "origin of life"(Oparin, 1953), and more recent theoretical letters on micro-phase separation (Sear, 2001; Walter and Brooks, 1995) have invoked similar concepts. These compartments are increasingly understood to be formed through phase separation of the underlying moieties from the cellular milieu. Advances in imaging, biochemical, and sequencing technologies, first leveraged by Brangwynne, Hyman, and Rosen in the last decade (Brangwynne et al., 2009; Li et al., 2012), have begun describing such membraneless compartments through the unifying principles of phase transitions.

How these condensed phases, or biomolecular condensates, form and what their properties are will be discussed below. In the next section, we will describe the basic physical and mathematical frameworks required to investigate and probe phase behavior.

[18] In the next section, we will see that while originally "liquidity" was an identifying metric of these collectives of matter, assemblies that condense out of the milieu can exhibit diverse physical properties ranging from simple liquids to amorphous gels.

[19] This is a fascinating article published in *Science* in *1899* with reflections that are stunningly prescient, and even more remarkably, forward-looking. Below is an evergreen quote:

"It is especially important in this field of biological inquiry to distinguish clearly between theory and observed fact, for theories of protoplasmic structure have always far outrun the actual achievements of observation"

Properties and features of biomolecular condensates

Membraneless organelles, or biomolecular condensates, are increasingly recognized to be formed through, in particular through liquid-liquid phase separation[20] and/or formation of networked (percolated) assem-

[20] There are many common examples of phase-separated or de-mixed liquids – such as oil and water or oil and vinegar.

blies. It has been long recognized that polymeric solutions (Asherie et al., 1996; Cohen and Benedek, 1982; Flory, 1942; Huggins, 1941; Overbeek and Voorn, 1957; Semenov and Rubinstein, 1998), such as those containing biological molecules, exhibit a rich phase behavior and transition between distinct phases dependent on variables such as concentration, temperature, or solvent quality[21]. A universal feature of phase transitions is their ultra-sensitivity – for example, solutions of RNA/protein will spontaneously condense into liquid-droplets beyond sharp thresholds in total concentration or when the temperature drops below a value – a phenomenon that is uniquely sensitive to the nature of intermolecular interactions. This represents a powerful axes of control in biology – a switch to toggle assemblies through their interactions.

A key property of proteins and nucleic acids that promotes phase separation is *multivalency* in interactions (Figure 3). This multivalency can be encoded as repeated modular and often structured domains (Figure 3, for example (Banani et al., 2016; Li et al., 2012)), through intrinsically disordered regions (IDRs) of proteins that interact through multiple weak interactions (Figure 3 see (Brangwynne et al., 2015; van der Lee et al., 2014) for detailed reviews on IDRs), or through combinations or extensions of above[22]. Unlike the classical lock-key paradigm for interactions between structured domains[23], IDRs interact through numerous biochemical modes - electrostatics, aromatics, π-cation, short linear-motifs (or SLiMs), to name a few prominent ones[24]. How these molecular interaction networks specify phase behavior in general is a broadly interesting and largely unsolved problem[25].

Biomolecular condensates are also typically far from *simple* liquids[26] and consist of many different components, possess a range of physical properties, and exhibit a range of morphologies. For example, most condensates contain tens to hundreds of constituents (Banani et al., 2017) which are enriched at different levels over the cellular background and are depleted in many other molecules. This allows for the formation of a particular micro-environment conducive to certain biochemical reactions. Condensates have been observed to also exhibit a wide-range of material properties spanning viscous liquids (Elbaum-Garfinkle et al., 2015; Feric et al., 2016; Hondele et al., 2019), gels (Banjade and Rosen, 2014; Lin et al., 2015; Schmit et al., 2019; Su et al., 2016), anisotropic or

[21] In the next section on physical principles, we will include a discussion of the basic framework to discuss the nature of phase transitions.

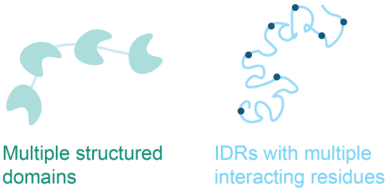


Figure 3 Examples of multivalency in proteins; Left panel shows a protein witb multiple structured domains and right panel depicts an intrinsically disordered protein (IDR) with multiple interacting residues (darker spots).

[22] For example through branched polymers or through globular domains with solvent-accessible interaction domains ("patchy" colloids)

[23] Initially postulated by Emil Fischer long before the structural evidence for lock-key type interactions was discovered

[24] See (Choi et al., 2020; Kim et al., 2019; Lin et al.; Pak et al., 2016; Posey et al., 2018; Sherry et al., 2017; Vernon et al., 2018) for details

[25] Some recent work on this topic (Choi et al., 2020; Lin et al., 2019b; Sanders et al., 2020; Wang et al., 2018) has shown promise in connecting the molecular features of these interactions to their macroscopic phase behavior

[26] Like oil

crystalline liquids (Freeman Rosenzweig et al., 2017; Yu et al., 2020; Zhang et al., 2018), and solid or amyloid-like substances[27] (Boke and Mitchison, 2017; Khan et al., 2018; Patel et al., 2015). Unlike simple fluids, the internal morphology of a condensate can exhibit structure – from vacuoles (Banerjee et al., 2017; Johnson and Jones, 1967) and "Russian-doll" like nested multi-layer phases (Feric et al., 2016) to ring-shell structures (Fox et al., 2018; Yu et al., 2020). Additionally, unlike their equilibrium counter-parts in the test-tubes, these collective assemblies of molecules are constantly subject to various out-of-equilibrium driving forces – including ATP-dependent hydrolysis, chemical reactions, molecular motors and protein-remodelers, and dynamically varying concentrations of constituents. Finally, condensates exhibit rich dynamical and spatial characteristics - including organization at specific cellular addresses (particular DNA elements or near other organelles for example (Ma and Mayr, 2018; Sabari et al., 2018, 2020; Shrinivas et al., 2019)) and regulation across the cell-cycle (for e.g. many condensates dissolve during cell division and reform in the daughter cells (Rai et al., 2018)).

[27] While solids are thought to be more likely to be pathological states, this is not the norm. In fact, there are many examples of solid (or amyloid) like physiological condensates (Boke and Mitchison, 2017)

Understanding and connecting the various molecular-scale processes (interactions, concentrations, reactions) to the emergent physical, material, compositional, and importantly, biologically relevant functional behavior of condensates is a central goal of this exciting field requiring cross-disciplinary expertise amongst engineers, physicists and biologists.

In this thesis, we will touch on physical, compositional, and functional relevance of a class of condensates that regulate an important biological phenomenon – the control of gene expression. We will review and link the diverse features of transcriptional control that are consistent with a phase separation model (Chapter 1), characterize the various features of DNA & protein that drive localized condensation of transcriptional condensates (Chapter 2), and investigate a role for non-equilibrium feedback regulation mediated by synthesis of RNA i.e. the transcript.

Introduction to statistical physics of liquid-liquid phase separation

Statistical physics provides a mathematical framework to characterize the collective behavior of many interacting objects. Initially developed by Boltzmann (Boltzmann, 1910) to connect the thermodynamics (or mechanics) of gaseous molecules to their macroscopic properties, this branch of physics has since made significant contributions to thermodynamics, phase transitions, universality, condensed matter physics, active matter, and more recently – deep learning, amongst other topics. Describing how heterogeneous states of matter co-exist and transition from one state to another has been a mainstay of thermodynamics for nearly three centuries - work of chemists such as Jospeh Black[28] in mid-eighteenth century Britain to luminaries such as Boltzmann, Clausius, and Gibbs[29] at the turn of the nineteenth century, and coupled with great advances in statistical physics in the twentieth century on magnetism, universality, and the renormalization group, to name but a few. A lengthy introduction to this topic is beyond the scope of this thesis, but the interested reader may choose to dig deeper here[30].

In the next subsection, a short introduction to the description of phase transitions will be provided in the language of statistical thermodynamics. Subsequently, we'll explore the development of models that are relevant to understanding and probing the emergent properties of biomolecular condensates.

[28]Who coined the term "sensible" or "latent" heat

[29]They first invoked the concept of available energy and entropy to describe the state of matter. See (Gibbs and Willard, 1879) for an original treatise on treating the multiple phases of matter

[30]There are far too numerous references to suggest –a few key text-books are listed (Chandler, 1987; Goldenfeld, 1992; Kardar, 2007).

Introduction to thermodynamics of liquid-liquid phase transitions

Phase transitions represent processes where a material changes state of matter to reach a free-energy minimum – driven by a competition between entropy and enthalpy. In general, entropy is maximized by increased disorder or more available configurations, which is always favored in phase where all constituents are well-mixed. However, interactions between particular moieties counter-act this drive to mix, by providing an advantage to demixing. Phase separation typically occurs when the total energetic advantage offsets the entropic costs of demixing[31]. The magnitude of the energetic advantage depends on the

[31] Thermodynamic systems reach a free-energy (F) minimum at equilibrium:

$$F = E - TS$$

Where E is the internal energy (which captures the intermolecular interactions),

T is the temperature, and

S is the entropy, which counts the total available configurations. Mixed states have more configurations than demixed ones.

range and strength of intermolecular interactions. The resultant phases are both considered "liquid-like" when they exhibit rapid internal rearrangement (i.e. molecules are not caged – as they are in solids) and there is no long-range order present within either phase. Nearly a century of work on polymeric systems[32] has provided a mathematical/physical framework that can be leveraged to describe the phase behavior of biological molecules.

[32] An incomplete list of key advances are presented here (Cohen and Benedek, 1982; Flory, 1942; Huggins, 1941; Overbeek and Voorn, 1957; Semenov and Rubinstein, 1998)

The general criteria for co-existence between two (or more) phases are:

1. Chemical equilibrium: The thermodynamic potential to drive fluxes of chemicals across the interface between the two phases should be equal (α,β) i.e. the chemical potential should for each species (i) should the same. ($\mu^{\alpha}_{i}=\mu^{\beta}_{i}$)

2. Mechanical equilibrium: The net difference in pressure between the two phases must be equivalent. ($p^{\alpha}=p^{\beta}$)

3. Thermal equilibrium: The two phases must have net zero enthalpic (or temperature potential) across the phases, giving the familiar relation that ($T^{\alpha}=T^{\beta}$)[33].

[33] This is why the temperature remains constant at 100C when water boils to steam

Physical principles underlying phase behavior of biomolecular condensates

A general mathematical framework to describe the molecular basis of biomolecular phase separation is the topic of much investigation. The most widely used coarse-grained model is the Flory-Huggins framework (Flory, 1942; Huggins, 1941), which is agnostic to the molecular interaction potentials as long as they are "short-ranged". This model importantly accounts for configurational entropy[34] of polymeric molecules but is formally defined only in the case of homo-polymers (which most biomolecules are not). This framework has shown modest promise in obtaining zeroth-order estimates of interaction potentials for certain proteins (Brady et al., 2017) and in probing qualitative features of multicomponent condensates such as layered morphologies (Mao et al., 2019; Riback et al., 2020; Sanders et al., 2020). However, understanding the generic phase behavior and statistical mechanics of many component (N>>1) systems continues to remain a challenging problem and an ac-

[34] i.e. the fact that polymeric molecules have much reduced entropy cost to demix than if the constituent monomers were free in solution. This is intuitive to understand as if the monomers were not constrained by connectivity, they could explore many more configurations in the system

tive area of ongoing research (Jacobs and Frenkel, 2017; Mao et al., 2019; Sear and Cuesta, 2003).

An orthogonal class of approaches seek a more mechanistic view of predicting phase behavior i.e. tying the molecular features of biological heteropolymers to their phase behavior. At the most coarse-grained level, adaptations of classic models of self-associating polymers and related simulations (Boeynaems et al., 2019; Choi et al., 2020; Cohen and Benedek, 1982; Semenov and Rubinstein, 1998; Wang et al., 2018) have shown promise in predicting threshold concentrations (or saturation concentrations) for individual proteins through their compositional trends. Recently, more sophisticated analytical and numerical approaches that have leveraged field-theoretic models to treat long-ranged electrostatic interactions as well as higher-order interactions, which in turn, enable connection of single-polymer sequence to phase behavior (Firman and Ghosh, 2018; Huihui et al., 2018; Lin et al., 2019a, 2019b; McCarty et al., 2019; Shen and Wang, 2018; Wei et al., 2017). In parallel, computer simulation methods have been developed to probe quantitative physical properties and dynamics of particular biomolecules – but here the extensive numerical costs and accuracy of empirical parameters continue to be a strong limitation on the generalizability of these techniques (Dignon et al., 2018). Overall, a combination of the above methods hold potential to explain many emergent properties of condensates in terms of the molecular features of their principal constituents[35]. Simultaneously, experimental advances that leverage optical and genetic techniques *in vivo* (Bracha et al., 2018; Sanders et al., 2020) and spectroscopic/biochemical methods *in vitro* (Brady et al., 2017; Kim et al., 2019) provide fertile ground for iteration between experiment and theory to refine models.

[35]Though these approaches are limited to systems with only 1 or 2 polymers

The physical principles described above help in connecting the molecular interaction networks to the expected equilibrium phase behavior (as determined by the nature of the underlying free-energy landscape). However, most biomolecular condensates are not at equilibrium in cells. There have been limited but important studies on understanding the coupling between active processes and phase behavior (reviewed in (Berry et al., 2018; Weber et al.)) – mostly in the context of two-species (polymer+solvent) systems with chemical reactions included. These studies typically write down dynamical models, following (Hohenberg

and Halperin, 1977), that couple the fluxes of individual species to the underlying free-energy landscape (or chemical potentials) as well as other active processes[36]. These studies have made a number of important predictions[37] for non-trivial distributions of droplet sizes at steady-state[38], hydrodynamic instabilities that trigger droplet division[39] (Seyboldt and Jülicher, 2018; Zwicker et al., 2016), and dynamics of condensate growth and nucleation (Zwicker et al., 2015). The interplay between active biological processes such as ATP-dependent remodeling, multi-component reactions, molecular motors etc. and condensate behavior in general, and more specifically contextualized in regards to specific biological functions, can shed light on the emergent behavior and interplay of condensates.

In this thesis, we will deploy a kinetic model of self-association to probe the relevance of phase separation to transcriptional control (Chapter 1). Motivated by the relevant biological parameters, and in iteration with experiment, we'll develop coarse-grained simulations of DNA and transcriptional proteins to study the multi-component phase behavior (Chapter 2). Finally, we'll develop a non-equilibrium description that couples the active process of RNA synthesis to the phase-behavior of RNA-protein coacervates to study their interplay (Chapter 3)

[36] Similar to frameworks used to explore reactive multi-component polymeric mixtures as in (Glotzer et al., 1994)

[37] Most of which still lack concrete evidence in biological contexts

[38] At equilibrium, all droplets would eventually come together to form one macroscopically different phase – like what happens after shaking a tube with oil and water)

[39] With intriguing ramifications on "origin of life" models of proto-cells. (Zwicker et al., 2016) discusses this topic in some detail.

Thesis overview and statements of collaboration

In this thesis, we will explore the role of phase separation in control of gene expression. The work presented in this thesis has emerged from synergistic exploration and synthesis of ideas with many collaborators[40]. I will provide a brief overview below.

[40] With the lion's share of experimental methods, analyses, and techniques from the Young lab.

In Chapter 1, we will propose that condensation of transcriptional molecules by phase separation underlies formation and function of a class of regulatory elements called super-enhancers. We will discuss experimental evidence that supports this view. This work was carried out in collaboration with Dr. Denes Hnisz and jointly supervised by Profs. Arup K. Chakraborty, Phillip A. Sharp, and Richard A. Young.

In Chapter 2, we will leverage the physical principles of phase transitions to explore how transcriptional condensates are organized across the genome. Through exploring the mechanistic forces driving condensate assembly, we will shed light on enhancer function and genomic encoding of phase behavior. Chapter 2 arose from a fruitful collaboration with the Young Laboratory. In particular, Dr. Benjamin Sabari was the *tour-de-force* who devised most of the experimental assays described.

Chapter 3 emerged from a serendipitous convergence of ideas on RNA-mediated control (experimental side) and non-equilibrium regulation (theoretical side) of transcriptional condensates. This body of work[41] was performed in close collaboration with Drs. Jon Henninger and Ozgur Oksuz at the Young Lab, who performed all the experimental analyses described.

[41] Under review for publication currently

Chapter 4 is the one deviation from the above theme, in that, it is a purely theoretical exercise. This work was inspired by the seminal essays and books on facilitated variation by cell biologist extraordinaire Marc Kircshner and John Gerhart (Kirschner et al., 1998, 2006). This work was carried out in collaboration with Dr. Ang Gao, a phenomenal post-doctoral candidate in the lab, and Profs. Phillip Sharp and Arup K. Chakraborty.

Bibliography

Asherie, N., Lomakin, A., and Benedek, G.B. (1996). Phase Diagram of Colloidal Solutions. Phys. Rev. Lett. *77*, 4832–4835.

Balbiani, E.G. (1881). Sur la structure du noyau des cellules salivaires chez les larves de Chironomus. Zool. Anz *4*, 662–667.

Banani, S.F., Rice, A.M., Peeples, W.B., Lin, Y., Jain, S., Parker, R., and Rosen, M.K. (2016). Compositional Control of Phase-Separated Cellular Bodies. Cell *166*, 651–663.

Banani, S.F., Lee, H.O., Hyman, A.A., and Rosen, M.K. (2017). Biomolecular condensates: organizers of cellular biochemistry. *18*, 285–298.

Banerjee, P.R., Milin, A.N., Moosa, M.M., Onuchic, P.L., and Deniz, A.A. (2017). Reentrant Phase Transition Drives Dynamic Substructure Formation in Ribonucleoprotein Droplets. Angew. Chemie Int. Ed. *56*, 11354–11359.

Banjade, S., and Rosen, M.K. (2014). Phase transitions of multivalent proteins can promote clustering of membrane receptors. Elife *3*.

Bergeron-Sandoval, L.-P.P., Safaee, N., and Michnick, S.W. (2016). Mechanisms and Consequences of Macromolecular Phase Separation (Cell Press).

Berry, J., Brangwynne, C.P., and Haataja, M. (2018). Physical principles of intracellular organization via active and passive phase transitions Charge pattern matching as a "fuzzy" mode of molecular recognition for the functional phase separations of intrinsically disordered proteins Droplet ripening in concentration gradients Physical principles of intracellular organization via active and passive phase transitions. Reports Prog. Phys.

Boeynaems, S., Holehouse, A.S., Weinhardt, V., Kovacs, D., Van Lindt, J., Larabell, C., Bosch, L. Van Den, Das, R., Tompa, P.S., Pappu, R. V., et al. (2019). Spontaneous driving forces give rise to protein–RNA condensates with coexisting phases and complex material properties. Proc. Natl. Acad. Sci. U. S. A. *116*, 7889–7898.

Boke, E., and Mitchison, T.J. (2017). The balbiani body and the concept of physiological amyloids. Cell Cycle *16*, 153–154.

Boltzmann, L. (1910). Vorlesungen über gastheorie (JA Barth).

Bracha, D., Walls, M.T., Wei, M.-T., Zhu, L., Kurian, M., Avalos, J.L., Toettcher, J.E., and Brangwynne, C.P. (2018). Mapping Local and Global Liquid Phase Behavior in Living Cells Using Photo-Oligomerizable Seeds. Cell *175*, 1467-1480.e13.

Brady, J.P., Farber, P.J., Sekhar, A., Lin, Y.-H., Huang, R., Bah, A., Nott, T.J., Chan, H.S., Baldwin, A.J., Forman-Kay, J.D., et al. (2017). Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. Proc. Natl. Acad. Sci. U. S. A. *114*, E8194–E8203.

Brangwynne, C.P., Eckmann, C.R., Courson, D.S., Rybarska, A., Hoege, C., Gharakhani, J., Jülicher, F., and Hyman, A.A. (2009). Germline P granules are liquid droplets that localize by controlled dissolution/condensation. Science *324*, 1729–1732.

Brangwynne, C.P., Tompa, P., and Pappu, R. V. (2015). Polymer physics of intracellular phase transitions. Nat. Phys. *11*, 899–904.

Cajal, S.R., and others (1903). Un sencillo metodo de coloracion seletiva del reticulo protoplasmatico y sus efectos en los diversos organos nerviosos de vertebrados e invertebrados. Trab. Lab. Invest. Biol.(Madrid) *2*, 129–221.

Chandler, D. (1987). Introduction to modern statistical mechanics (Oxford University Press).

Choi, J.-M., Holehouse, A.S., and Pappu, R. V. (2020). Physical Principles Underlying the Complex Biology of Intracellular Phase Transitions. Annu. Rev. Biophys. *49*.

Cohen, R.J., and Benedek, G.B. (1982). Equilibrium and kinetic theory of polymerization and the sol-gel transition. J. Phys. Chem. *86*, 3696–3714.

Cramer, P. (2019). Organization and regulation of gene transcription. Nature *573*, 45–54.

Dellaire, G., Ching, R.W., Dehghani, H., Ren, Y., and Bazett-Jones, D.P. (2006). Journal of Cell Science. J. Cell Sci. *118*, 847–854.

Dignon, G.L., Zheng, W., Best, R.B., Kim, Y.C., and Mittal, J. (2018). Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. Proc. Natl. Acad. Sci. U. S. A. *115*, 9929–9934.

Elbaum-Garfinkle, S., Kim, Y., Szczepaniak, K., Chen, C.C.H., Eckmann, C.R., Myong, S., and Brangwynne, C.P. (2015). The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. Proc. Natl. Acad. Sci. U. S. A. *112*, 7189–7194.

Feric, M., Vaidya, N., Harmon, T.S., Mitrea, D.M., Zhu, L., Richardson, T.M., Kriwacki, R.W., Pappu, R. V., and Brangwynne, C.P. (2016). Coexisting Liquid Phases Underlie Nucleolar Subcompartments. Cell *165*, 1686–1697.

Firman, T., and Ghosh, K. (2018). Sequence charge decoration dictates coil-globule transition in intrinsically disordered proteins A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins Sequence charge decoration dictates coil-globul. J. Chem. Phys. J. Chem. Phys. J. Chem. Phys. J. Chem. Phys. *148*, 123305–123303.

Flory, P.J. (1942). Thermodynamics of High Polymer Solutions. J. Chem. Phys. *10*, 51–61.

Fox, A.H., Nakagawa, S., Hirose, T., and Bond, C.S. (2018). Paraspeckles: Where Long Noncoding RNA Meets Phase Separation. Trends Biochem. Sci. *43*, 124–135.

Freeman Rosenzweig, E.S., Xu, B., Kuhn Cuellar, L., Martinez-Sanchez, A., Schaffer, M., Strauss, M., Cartwright, H.N., Ronceray, P., Plitzko, J.M., Förster, F., et al. (2017). The Eukaryotic CO2-Concentrating Organelle Is Liquid-like and Exhibits Dynamic Reorganization. Cell *171*, 148-162.e19.

Furlong, E.E.M., and Levine, M. (2018). Developmental enhancers and chromosome topology. Science *361*, 1341–1345.

Gall, J.G. (2000). Cajal Bodies: The First 100 Years. Annu. Rev. Cell Dev. Biol. *16*, 273–300.

Gibbs, A.:, and Willard, J. (1879). Heidelberger Texte zur Mathematikgeschichte On the Equilibrium of Heterogeneous Substances.

Glotzer, S.C., Stauffer, D., and Jan, N. (1994). Monte Carlo simulations of phase separation in chemically reactive binary mixtures. Phys. Rev. Lett. *72*, 4109–4112.

Goldenfeld, N. (1992). Lectures on phase transitions and the renormalization group (Addison-Wesley, Advanced Book Program, Reading).

Haberle, V., and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. Nat. Rev. Mol. Cell Biol. 1.

Hohenberg, P.C., and Halperin, B.I. (1977). Theory of dynamic critical phenomena. Rev. Mod. Phys *49*, 435–479.

Hondele, M., Sachdev, R., Heinrich, S., Wang, J., Vallotton, P., Fontoura, B.M.A., and Weis, K. (2019). DEAD-box ATPases are global regulators of phase-separated organelles. Nat. 2019 1–5.

Huggins, M.L. (1941). Solutions of Long Chain Compounds. J. Chem. Phys. *9*, 440–440.

Huihui, J., Firman, T., and Ghosh, K. (2018). Modulating charge patterning and ionic strength as a strategy to induce conformational changes in intrinsically disordered proteins. J. Chem. Phys. *149*, 085101.

Jacobs, W.M., and Frenkel, D. (2017). Phase Transitions in Biological Systems with Many Components. Biophys. J. *112*, 683–691.

Johnson, J.M., and Jones, L.E. (1967). BEHAVIOR OF NUCLEOLI AND CONTRACTING NUCLEOLAR VACUOLES IN TOBACCO CELLS GROWING IN MICROCULTURE. Am. J. Bot. *54*, 189–198.

Judson, H.F. (1979). The eighth day of creation. New York 550.

Kardar, M. (2007). Statistical physics of fields (Cambridge University Press).

Khan, T., Kandola, T.S., Wu, J., Venkatesan, S., Ketter, E., Lange, J.J., Rodríguez Gama, A., Box, A., Unruh, J.R., Cook, M., et al. (2018). Quantifying Nucleation In Vivo Reveals the Physical Basis of Prion-like Phase Behavior. Mol. Cell *71*, 155-168.e7.

Kim, T.H., Tsang, B., Vernon, R.M., Sonenberg, N., Kay, L.E., and Forman-Kay, J.D. (2019). Phospho-dependent phase separation of FMRP and CAPRIN1 recapitulates regulation of translation and deadenylation. Science *365*, 825–829.

Kirschner, M., Gerhart, J., Otey, C.R., and Arnold, F.H. (1998). Evolvability. Proc. Natl. Acad. Sci. U. S. A. *95*, 8420–8427.

Kirschner, M., Gerhart, J.C., and Norton, J. (2006). The Plausibility of Life: Resolving Darwin's Dilemma (Yale University Press).

Krijger, P.H.L., and De Laat, W. (2016). Regulation of disease-associated gene expression in the 3D genome. Nat. Rev. Mol. Cell Biol. *17*, 771–782.

Lee, T.I., and Young, R.A. (2013). Transcriptional Regulation and Its Misregulation in Disease. Cell *152*, 1237–1251.

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T., et al. (2014). Classification of Intrinsically Disordered Regions and Proteins. Chem. Rev. *114*, 6589–6631.

Levine, M., Cattoglio, C., and Tjian, R. (2014). No Title (Cell Press).

Li, P., Banjade, S., Cheng, H.-C.C., Kim, S., Chen, B., Guo, L., Llaguno, M., Hollingsworth, J. V., King, D.S., Banani, S.F., et al. (2012). Phase transitions in the assembly of multivalent signalling proteins. Nature *483*, 336–340.

Lin, Y.-H., Forman-Kay, J.D., and Chan, H.S. Sequence-Specific Polyampholyte Phase Separation in Membraneless Organelles.

Lin, Y.-H., Brady, J.P., Chan, H.S., and Ghosh, K. (2019a). A unified analytical theory of heteropolymers for sequence-specific phase behaviors of polyelectrolytes and polyampholytes.

Lin, Y., Protter, D.S.W., Rosen, M.K., and Parker, R. (2015). Formation and Maturation of Phase Separated Liquid Droplets by RNA Binding Proteins. Mol. Cell *60*, 208–219.

Lin, Y., McCarty, J., Rauch, J.N., Delaney, K.T., Kosik, K.S., Fredrickson, G.H., Shea, J.-E.E., and Han, S. (2019b). Narrow equilibrium window for complex coacervation of tau and RNA under cellular conditions. Elife *8*, 1–31.

Long, H.K., Prescott, S.L., and Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell *167*, 1170–1187.

Ma, W., and Mayr, C. (2018). A Membraneless Organelle Associated with the Endoplasmic Reticulum Enables 3'UTR-Mediated Protein-Protein Interactions. Cell *175*, 1492-1506.e19.

Mao, S., Kuldinow, D., Haataja, M.P., and Košmrlj, A. (2019). Phase behavior and morphology of multicomponent liquid mixtures. Soft Matter *15*, 1297–1311.

Mao, Y.S., Zhang, B., and Spector, D.L. (2011). Biogenesis and function of nuclear bodies. Trends Genet. *27*, 295–306.

Mathews, C.K. (1993). The cell - Bag of enzymes or network of channels? J. Bacteriol. *175*, 6377–6381.

McCarty, J., Delaney, K.T., Danielsen, S.P.O., Fredrickson, G.H., and Shea, J.-E. (2019). Complete Phase Diagram for Liquid–Liquid Phase Separation of Intrinsically Disordered Proteins. J. Phys. Chem. Lett. *10*, 1644–1652.

Montgomery, T.H. (1900). Comparative cytological studies with especial regard to the morphology of the nucleolus (Ginn).

Mukherjee, S. (2016). The gene : an intimate history.

Nizami, Z., Deryusheva, S., and Gall, J.G. (2010). The Cajal Body and Histone Locus Body. Cold Spring Harb. Perspect. Biol. *2*, a000653.

Ong, C.T., and Corces, V.G. (2011). Enhancer function: New insights into the regulation of tissue-specific gene expression. Nat. Rev. Genet. *12*, 283–293.

Oparin, A. (1953). The origin of life.

Orphanides, G., and Reinberg, D. (2002). A unified theory of gene expression (Cell Press).

Overbeek, J.T.G., and Voorn, M.J. (1957). Phase separation in polyelectrolyte solutions. Theory of complex coacervation. J. Cell. Comp. Physiol. *49*, 7–26.

Pak, C.W., Kosno, M., Holehouse, A.S., Padrick, S.B., Mittal, A., Ali, R., Yunus, A.A., Liu, D.R., Pappu, R. V., and Rosen, M.K. (2016). Sequence Determinants of Intracellular Phase Separation by Complex Coacervation of a Disordered Protein. Mol. Cell *63*, 72–85.

Patel, A., Lee, H.O., Jawerth, L., Maharana, S., Jahnel, M., Hein, M.Y., Stoynov, S., Mahamid, J., Saha, S., Franzmann, T.M., et al. (2015). A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. Cell *162*, 1066–1077.

Pederson, T. (2011). The nucleolus. Cold Spring Harb. Perspect. Biol. *3*, a000638.

Posey, A.E., Holehouse, A.S., and Pappu, R. V. (2018). Phase Separation of Intrinsically Disordered Proteins (Elsevier Inc.).

Rai, A.K., Chen, J.-X., Selbach, M., and Pelkmans, L. (2018). Kinase-controlled phase transition of membraneless organelles in mitosis. Nature 1.

Raser, J.M., and O'Shea, E.K. (2004). Control of stochasticity in eukaryotic gene expression. Science *304*, 1811–1814.

Reiter, F., Wienerroither, S., and Stark, A. (2017). Combinatorial function of transcription factors and cofactors. Curr. Opin. Genet. Dev. *43*, 73–81.

Riback, J.A., Zhu, L., Ferrolino, M.C., Tolbert, M., Mitrea, D.M., Sanders, D.W., Wei, M.T., Kriwacki, R.W., and Brangwynne, C.P. (2020). Composition-dependent thermodynamics of intracellular phase separation. Nature *581*, 209–214.

Sabari, B.R., Dall'Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A. V., Manteiga, J.C., et al. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. Science *361*.

Sabari, B.R., Dall'Agnese, A., and Young, R.A. (2020). Biomolecular Condensates in the Nucleus. Trends Biochem. Sci. *0*.

Sanders, D.W., Kedersha, N., Lee, D.S.W., Strom, A.R., Drake, V., Riback, J.A., Bracha, D., Eeftens, J.M., Iwanicki, A., Wang, A., et al. (2020). Competing Protein-RNA Interaction Networks Control Multiphase Intracellular Organization. Cell *181*, 306-324.e28.

Schaffner, W. (2015). Enhancers, enhancers - From their discovery to today's universe of transcription enhancers. Biol. Chem.

Schmit, J.D., Bouchard, J.J., Martin, E.W., and Mittag, T. (2019). Protein network structure enables switching between liquid and gel states. BioRxiv 754952.

Sear, R.P. (2001). Phase Separation in Mixtures of Colloids and Long Ideal Polymer Coils. Phys. Rev. Lett. *86*, 4696–4699.

Sear, R.P., and Cuesta, J.A. (2003). Instabilities in Complex Mixtures with a Large Number of Components. Phys. Rev. Lett. *91*, 245701.

Semenov, A.N., and Rubinstein, M. (1998). Thermoreversible Gelation in Solutions of Associative Polymers. 1. Statics. Macromolecules *31*, 1373–1385.

Seyboldt, R., and Jülicher, F. (2018). Role of hydrodynamic flows in chemically driven droplet division.

Shen, K., and Wang, Z.-G. (2018). Polyelectrolyte Chain Structure and Solution Phase Behavior. Macromolecules *51*, 1706–1717.

Sherry, K.P., Das, R.K., Pappu, R. V, and Barrick, D. (2017). Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the Notch receptor. Proc. Natl. Acad. Sci. U. S. A. *114*, E9243–E9252.

Shin, Y., and Brangwynne, C.P. (2017). Liquid phase condensation in cell physiology and disease. Science *357*, eaaf4382.

Shrinivas, K., Sabari, B.R., Coffey, E.L., Klein, I.A., Boija, A., Zamudio, A. V., Schuijers, J., Hannett, N.M., Sharp, P.A., Young, R.A., et al. (2019). Enhancer features that drive formation of transcriptional condensates. Mol. Cell *75*, 549-561.e7.

Smith, E., and Shilatifard, A. (2014). Enhancer biology and enhanceropathies. Nat. Struct. Mol. Biol. *21*, 210–219.

Spector, D.L., and Lamond, A.I. (2011). Nuclear speckles. Cold Spring Harb. Perspect. Biol. *3*, 1–12.

Spitz, F., and Furlong, E.E.M.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.

Style, R.W., Sai, T., Fanelli, N., Ijavi, M., Smith-Mannschott, K., Xu, Q., Wilen, L.A., and Dufresne, E.R. (2018). Liquid-Liquid Phase Separation in an Elastic Network. Phys. Rev. X *8*.

Su, X., Ditlev, J.A., Hui, E., Xing, W., Banjade, S., Okrut, J., King, D.S., Taunton, J., Rosen, M.K., and Vale, R.D. (2016). Phase separation of signaling molecules promotes T cell receptor signal transduction. Science *352*, 595–599.

Tjian, R., and Maniatis, T. (1994). Transcriptional activation: A complex puzzle with few easy pieces. Cell *77*, 5–8.

Vernon, R.M., Chong, P.A., Tsang, B., Kim, T.H., Bah, A., Farber, P., Lin, H., and Forman-Kay, J.D. (2018). Pi-Pi contacts are an overlooked protein feature relevant to phase separation. Elife *7*, e31486.

Wagner, R. (1835). Einige bemerkungen und fragen über das keimbläschen (vesicular germinativa). Müller's Arch. Anat Physiol Wiss. Med *268*, 373–377.

Walter, H., and Brooks, D.E. (1995). Phase separation in cytoplasm, due to macromolecular crowding, is the basis for microcompartmentation. FEBS Lett.

Wang, J., Choi, J.-M., Holehouse, A.S., Lee, H.O., Zhang, X., Jahnel, M., Maharana, S., Lemaitre, R., Pozniakovsky, A., Drechsel, D., et al. (2018). A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. Cell *174*, 688-699.e16.

Weber, C.A., Zwicker, D., Jülicher, F., and Lee, C.F. Physics of active emulsions. Rep. Prog. Phys *82*, 64601.

Wei, M.-T., Elbaum-Garfinkle, S., Holehouse, A.S., Chen, C.C.-H., Feric, M., Arnold, C.B., Priestley, R.D., Pappu, R. V., Brangwynne, C.P., Chih, C., et al. (2017). Phase behaviour of disordered proteins underlying low density and high permeability of liquid organelles. Nat. Chem. *9*, 1118–1125.

Wilson, E.B. (1899). The structure of protoplasm.*. Science *10*, 33–45.

Yu, H., Lu, S., Gasior, K., Singh, D., Tapia, O., Vazquez-Sanchez, S., Toprani, D., Beccari, M., Yates, J.R., Cruz, S. Da, et al. (2020). TDP-43 and HSP70 phase separate into anisotropic, intranuclear liquid spherical annuli. BioRxiv 2020.03.28.985986.

Zhang, L., Köhler, S., Rillo-Bohn, R., and Dernburg, A.F. (2018). A compartmentalized signaling network mediates crossover control in meiosis. Elife *7*.

Zwicker, D., Hyman, A.A., and Jülicher, F. (2015). Suppression of Ostwald ripening in active emulsions. Phys. Rev. E *92*.

Zwicker, D., Seyboldt, R., Weber, C.A., Hyman, A.A., and Jülicher, F. (2016). Growth and division of active droplets provides a model for protocells. Nat. Phys. *13*, 408–413.

Chapter 1: *Hypothesis*

A phase separation model for transcriptional control

"*The real voyage of discovery consists not in seeking new landscapes, but in having new eyes*"

Marcel Proust, Remembrance of Things Past

This chapter is primarily based on work published in
"A Phase Separation Model for Transcriptional Control"
Hnisz, D=., **Shrinivas, K=.,** Young, R.A.[c], Chakraborty, A.K.[c], and Sharp, P.A.[c]
(2017). Cell 169, 13–23.

Supporting data is included from
"Coactivator condensation at super-enhancers links phase separation and gene control"
Sabari, B.R. =, Dall'Agnese, A. =, Boija, A., Klein, I.A., Coffey, E.L., **Shrinivas, K.**, Abraham, B.J., Hannett, N.M., Zamudio, A. V., Manteiga, J.C., et al. (2018).
Science 361.
(=equal contributions,[c] co-corresponding authors)

Phase-separated molecular assemblies provide a general regulatory mechanism to compartmentalize biochemical reactions within cells. In this chapter, we propose that a phase separation model explains established and recently described features of transcriptional control. These features include the formation of super-enhancers, the sensitivity of super-enhancers to perturbation, the transcriptional bursting patterns of enhancers and the ability of an enhancer to produce simultaneous activation at multiple genes. This model provides a conceptual framework to further explore principles of gene control in mammals. We end this chapter with experimental data that provide support for the phase separation model *in vivo*.

Introduction

Recent studies of transcriptional regulation have revealed several puzzling observations that as yet lack quantitative description, but whose further understanding would likely afford new and valuable insights into gene control during development and disease. For example, although thousands of enhancer elements control the activity of thousands of genes in any given human cell type, several hundred clusters of enhancers, called super-enhancers (SEs), control genes that have especially prominent roles in cell-type-specific processes (Dunham et al., 2012; Hnisz et al., 2013; Lovén et al., 2013; Parker et al., 2013; Roadmap Epigenomics Consortium et al., 2015; Whyte et al., 2013). Cancer cells acquire super- enhancers to drive expression of prominent oncogenes, so SEs play key roles in both development and disease (Chapuy et al., 2013; Lovén et al., 2013). Super- enhancers are occupied by an unusually high density of interacting factors, are able to drive higher levels of transcription than typical enhancers, and are exceptionally vulnerable to perturbation of components that are commonly associated with most enhancers (Chapuy et al., 2013; Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013).

Another puzzling[1] observation that has emerged from recent studies is that a single enhancer is able to simultaneously activate multiple proximal genes (Fukaya et al., 2016). Enhancers physically contact the promoters of the genes they activate, and early studies using

[1] Recent advances in imaging provide direct evidence (Chen et al., 2018; Heist et al., 2019) that enhancers are separated by more than "molecular bridging" distances from their target genes. Even more intriguingly, gene activation under certain conditions is characterized by increased separation (Benabdallah et al.) between enhancers and promoters. Together, these models suggest that direct physical contact may not be predictive, or even required, for gene activation by enhancers. Phase separation provides one possible mechanism for this "spooky action at a distance".

chromatin contact mapping techniques (e.g. at the β-globin locus) found that at any given time, enhancers activate only one of the several globin genes within the locus (Palstra et al., 2003; Tolhuis et al., 2002). However, more recent work using quantitative imaging at a high temporal resolution revealed that enhancers typically activate genes in bursts, and that two gene promoters can exhibit synchronous bursting when activated by the same enhancer (Fukaya et al., 2016).

Previous models of transcriptional control have provided important insights into principles of gene regulation. A key feature of most previous transcriptional control models is that the underlying regulatory interactions occur in a step-wise manner dictated by biochemical rules that are probabilistic in nature (Chen and Larson, 2016; Elowitz et al., 2002; Levine et al., 2014; Orphanides and Reinberg, 2002; Raser and O'Shea, 2004; Spitz and Furlong, 2012; Suter et al., 2011; Zoller et al., 2015). Such kinetic models predict that gene activation on a single gene level is a stochastic, noisy process, and also provide insights into how multi-step regulatory processes can suppress intrinsic noise and result in bursting[2]. These models do not shed light on the mechanisms underlying the formation, function, and properties of SEs or explain puzzles such as how two gene promoters exhibit synchronous bursting when activated by the same enhancer.

In this perspective, we propose and explore a model that may explain the puzzles described above. This model is based on principles involving phase separation of multi-molecular assemblies.

[2]Advances in imaging have enabled newer and more refined models – still largely agnostic to mechanisms and framed with effective rates of "recruitment" and "release" of Polymerases – but are increasingly supportive of a model of transient recruitment of clusters of Polymerases (Chen and Larson, 2016; Forero-Quintero et al., 2020; Larsson et al., 2019; Quintero-Cadena et al., 2020; Rodriguez and Larson, 2020)

Cooperativity in transcriptional control

Since the discovery of enhancers over 30 years ago, studies have attempted to describe functional properties of enhancers in a quantitative manner, and these efforts have mostly relied on the concept of co-operative interactions between enhancer component. Classically, enhancers have been defined as elements that can increase transcription from a target gene promoter when inserted in either orientation at various distances upstream or downstream of the promoter (Banerji et al., 1981; Benoist and Chambon, 1981; Gruss et al., 1981).

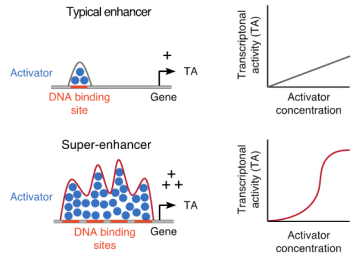


Figure 1 Schematic depiction of the classic model of co-operativity exemplified for typical enhancers and super-enhancers. The higher density of transcriptional regulators (referred to as "activators") through co-operative binding to DNA binding sites is thought to contribute to both higher transcriptional output and increased sensitivity to activator concentration at super-enhancers. Image adapted from (Lovén et al., 2013).

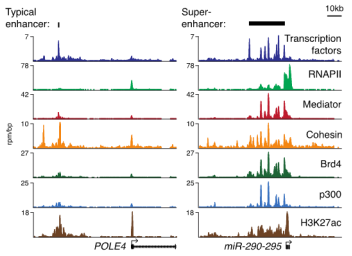


Figure 2 ChIP-seq binding profiles for RNA polymerase II (RNAPII) and the indicated transcriptional cofactors and chromatin regulators at the *POLE4* and *miR-290-295* loci in murine embryonic stem cells. The transcription factor binding profile is a merged ChIP-seq binding profile of the TFs Oct4, Sox2, and Nanog. rpm/bp, reads per million per base pair. Image adapted from (Hnisz et al., 2013).

Enhancers typically consist of hundreds of base-pairs of DNA and are bound by multiple transcription factor (TF) molecules in a co-operative manner (Bulger and Groudine, 2011; Levine et al., 2014; Malik and Roeder, 2010; Ong and Corces, 2011; Spitz and Furlong, 2012). Classically, co-operative binding describes the phenomenon that the binding of one TF molecule to DNA impacts the binding of another TF molecule (Figure 1) (Carey, 1998; Kim and Maniatis, 1997; Thanos and Maniatis, 1995; Tjian and Maniatis, 1994). Co-operative binding of transcription factors at enhancers has been proposed to be due to the effects of TFs on DNA bending (Falvo et al., 1995), interactions between TFs (Johnson et al., 1979) and combinatorial recruitment of large cofactor complexes by TFs (Merikaet al., 1998).

Super-enhancers exhibit highly co-operative properties

Several hundred clusters of enhancers, called super-enhancers (SEs), control genes that have especially prominent roles in cell-type-specific processes (Hnisz et al., 2013; Whyte et al., 2013). Three key features of SEs indicate that co-operative properties are especially important for their formation and function: 1) SEs are occupied by an unusually high density of interacting factors; 2) SEs can be formed by a single nucleation event; and 3) SEs are exceptionally vulnerable to perturbation of some components that are commonly associated with most enhancers.

SEs are occupied by an unusually high density of enhancer-associated factors, including transcription factors, co-factors, chromatin regulators, RNA polymerase II, and non-coding RNA (Hnisz et al., 2013). The non-coding RNA (enhancer RNA or eRNA), produced by divergent transcription at transcription factor binding sites within SEs (Hah et al., 2015; Sigova et al., 2013), can contribute to enhancer activity and the expression of the nearby gene *in cis* (Dimitrova et al., 2014; Engreitz et al., 2016; Lai et al., 2013; Pefanis et al., 2015). The density of the protein factors and eRNAs at SEs has been estimated to be approximately 10-fold the density of the same set of components at typical enhancers in the genome (Figure 2) (Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013). Chromatin

contact mapping methods indicate that the clusters of enhancers within SEs are in close physical contact with one another and with the promoter region of the gene they activate (Figure 3) (Dowen et al., 2014; Hnisz et al., 2016; Ji et al., 2016; Kieffer-Kwon et al., 2013).

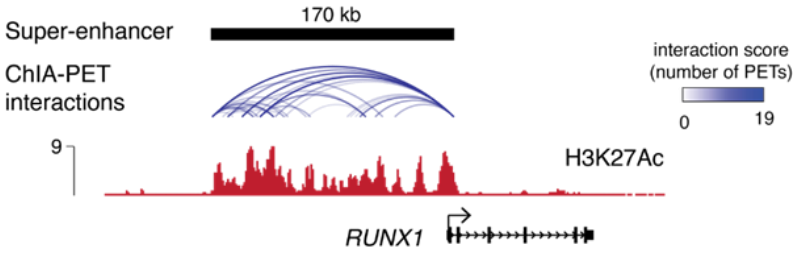


Figure 3 ChIA-PET interactions at the RUNX1 locus displayed above the ChIP-Seq profiles of H3K27Ac in human T-cells. The ChIA-PET interactions indicate frequent physical contact between the H3K27Ac occupied regions within the super-enhancer and the promoter of RUNX1.

SEs can be formed as a consequence of introducing a single transcription factor binding site into a region of DNA that has the potential to bind additional factors. In T cell leukemias, a small (2-12bp) mono-allelic insertion nucleates the formation of an entire SE by creating a binding site for the master transcription factor MYB, leading to the recruitment of additional transcriptional regulators to adjacent binding sitesand assembly of a host of factors spread over an 8 kb domain whose features are typical of a SE (Mansour et al., 2014). Inflammatory stimulation also leads to rapid formation of SEs in endothelial cells; here again, the formation of a SE is apparently nucleated by a single binding event of a transcription factor responsive to inflammatory stimulation (Brown et al., 2014).

Entire super-enhancers spanning tens of thousands of base-pairs can collapse as a unit when their co-factors are perturbed, and genetic deletion of constituent enhancers within an SE can compromise the function of other constituents. For example, the co-activator BRD4 binds acetylated chromatin at SEs, typical enhancers and promoters, but SEs are far more sensitive to drugs blocking the binding of BRD4 to acetylated chromatin (Chapuy et al., 2013; Lovén et al., 2013). A similar hypersensitivity of SEs to inhibition of the cyclin-dependent kinase CDK7 has also been observed in multiple studies (Chipumuro et al., 2014; Kwiatkowski et al., 2014; Wang et al., 2015). This kinase is critical for initiation of transcription by RNA Polymerase II (RNAPII) and phosphorylates its repetitive C-terminal domain (CTD) (Larochelle et al., 2012). Furthermore, genetic deletion of constituent enhancers within SEs can compromise the activities of other constituents within the super- enhancer (Hnisz et al., 2015;

Jiang et al., 2016; Proudhon et al., 2016; Shin et al., 2016), and can lead to the collapse of an entire super-enhancer (Mansour et al., 2014), although this interdependence of constituent enhancers is less apparent for some developmentally regulated super-enhancers (Hay et al., 2016).

In summary, several lines of evidence indicate that the formation and function of SEs involves co-operative processes that bring many constituent enhancers and their bound factors into close spatial proximity. High densities of proteins and nucleic acids – and co-operative interactions among these molecules – have been implicated in the formation of membraneless organelles, called cellular bodies, in eukaryotic cells (Banjade et al., 2015; Bergeron-Sandoval et al., 2016; Brangwynne et al., 2009). Below, we first describe features of the formation of cellular bodies, and then develop a model of super-enhancer formation and function that exploits related concepts.

Formation of membraneless organelles by phase separation

[5] For more up to date reviews, please refer at (Banani et al., 2017; Shin and Brangwynne, 2017; Snead and Gladfelter, 2019). The list of nuclear and cellular bodies thought to be mediated by phase separation continues to grow. However, healthy critiques are demanding better techniques and well-defined metrics to define the extent to which the principles of phase transitions influence *in vivo* function (McSwiggen et al., 2019; Peng and Weber, 2019)

Eukaryotic cells contain membraneless organelles, called cellular bodies, which play essential roles in compartmentalizing essential biochemical reactions within cells[5]. These bodies are formed by phase separation mediated by co-operative interactions between multivalent molecules (Banjade et al., 2015; Bergeron-Sandoval et al., 2016; Brangwynne et al., 2009). Examples of such organelles in the nucleus include nucleoli, which are sites of rRNA biogenesis; Cajal bodies, which serve as an assembly site for small nuclear RNPs; and nuclear speckles, which are storage compartments for mRNA splicing factors (Mao et al., 2011; Zhu and Brangwynne, 2015). These organelles exhibit properties of liquid droplets; for example, they can undergo fission and fusion, and hence their formation has been described as mediated by liquid-liquid phase separation. Mixtures of purified RNA and RNA- binding proteins form these types of phase-separated bodies *in vitro* (Berry et al., 2015; Feric et al., 2016; Kato et al., 2012a; Kwon et al., 2013; Li et al., 2012; Wheeler et al., 2016). Consistent with these observations, past theoretical work indicates

that the formation of a gel is usually accompanied by phase separation (Semenov and Rubinstein, 1998). Thus, a number of studies show that high densities of proteins and nucleic acids – and co-operative interactions among these molecules – are implicated in the formation of phase separated cellular bodies.

As described above, super-enhancers can be in essence considered to be co- operative assemblies of high densities of transcription factors, transcriptional co- factors, chromatin regulators, non-coding RNA and RNA Polymerase II (RNAPII). Furthermore, some transcription factors with low complexity domains have been proposed to create gel-like structures *in vitro* (Han et al., 2012; Kato et al., 2012b; Kwon et al., 2013). We thus hypothesize that phase-separation with formation of a phase separated multi-molecular assembly likely occurs during the formation of SEs and less frequently with typical enhancers (Figure 4).

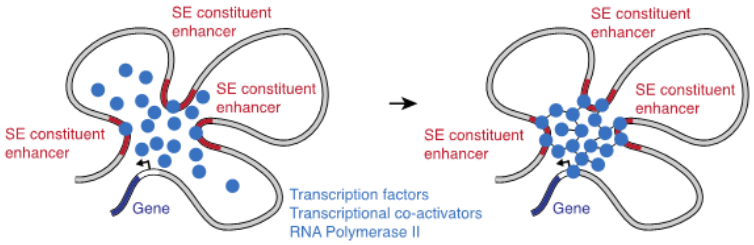


Figure 4 Schematic representation of the biological system that can form the phase-separated multi-molecular complex of transcriptional regulators at a super-enhancer gene locus

We propose a simple model that emphasizes cooperativity in the context of the number and valency of the interacting components, and affinity of interactions between these transcriptional regulators and nucleic acids, to explore the role of a phase separation for SE assembly and function. Computer simulations of this model show that phase separation can explain critical features of SEs, including aspects of their formation, function, and vulnerability. The simulations are also consistent with observed differences between transcriptional bursting patterns driven by weak and strong enhancers, and the simultaneous bursting of genes controlled by a shared single enhancer. We conclude by noting several implications and predictions of the phase separation model that could guide further exploration of this concept of transcriptional control in vertebrates.

A phase separation model of enhancer assembly and function

Many molecules bound at enhancers and SEs, such as transcription factors, transcriptional co-activators (e.g., BRD4), RNAPII and RNA can undergo reversible chemical modifications (e.g., acetylation, phosphorylation) at multiple sites. Upon such modifications, these multivalent molecules are able to interact with multiple other components, thus forming "cross-links" (Figure 4). Here, a cross-link can be defined as any reversible feature, including reversible chemical modification, or any other feature involved in dynamic binding and unbinding interactions. In considering whether phase separation may underlie certain observed features of transcriptional control, a simple model is needed to describe the dependence of phase separation on changes in valences and affinities of the interacting molecules, parameters biologists measure [6]. Below we describe such a model, and explain how the parameters of this model represent characteristics of typical enhancers and super- enhancers.

[6] Improvements by the Pappu Lab of the underlying framework that our model is based on, i.e. Semenov and Rubenstein's study (Semenov and Rubinstein, 1998), have shown promise (Choi et al., 2020; Wang et al., 2018) in predicting *in vitro* phase behavior of certain proteins.

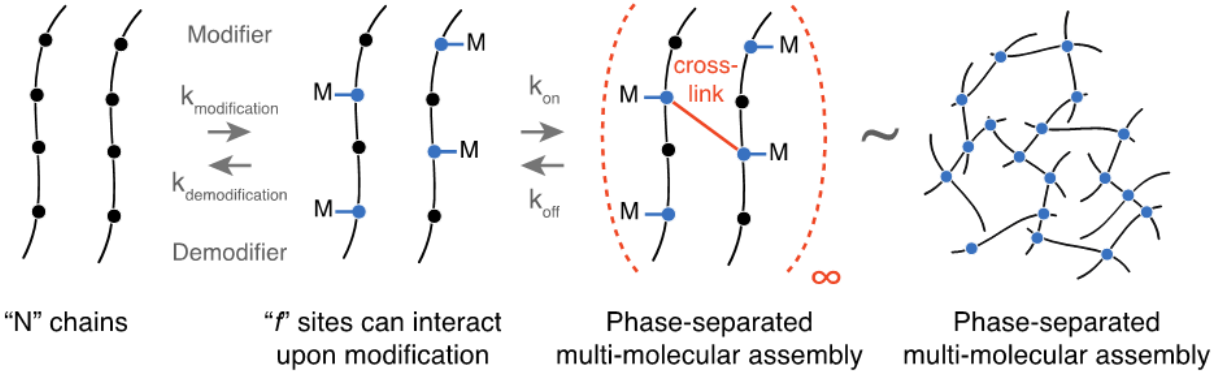


Figure 5 Simplified representation of the biological system, and parameters of the model that could lead to phase separation. "M" denotes modification of residues that are able to form cross-links when modified

In the model, the protein and nucleic acid components of enhancers are represented as chain-like molecules, each of which contains a set of residues that can potentially engage in interactions with other chains (Figure 5). These residues are represented as sites that can undergo reversible chemical modifications, and modification of the residues is associated with their ability to form non-covalent cross-linking interactions between the chains (Figure 5). Numerous enhancer-components, including transcription factors, co-factors, and the heptapeptide repeats of the C-terminal domain (CTD) of

RNA polymerase II are subject to phosphorylation, and are known to bind other proteins based on their phosphorylation status (Phatnani and Greenleaf, 2006). Our model encompasses such phosphorylation or dephosphorylation that can result in binding interactions, as well as interactions of histones and other proteins found at enhancers and transcriptional regulators that are modulated by acetylation, methylation or other types of chemical modifications. For simplicity, we refer to all types of chemical modifications and de-modifications generically as "modification" and "demodification" mediated by "modifiers" and "demodifiers", respectively.

In its simplest form, the model has three parameters: 1) "N" = the number of macromolecules (also referred to as "chains") in the system; this parameter sets the concentration of interacting components – the larger the value of N, the greater the concentration - SEs are considered to have a larger value of N while typical enhancers are modeled as having fewer components. 2) "*f*" = valency, which corresponds to the number of residues in each molecule that can potentially be modified and engage in a cross-link with other chains. Note that in our simplified model, the modification of a residue is required to allow the residue to create a cross- link with another chain. Conceptually, the model works in a similar way if the *demodified* state of a residue is required for cross-link formation, except the enzymatic activities that allow or inhibit cross-link formation are reversed. 3) $K_{eq} = (k_{on}/k_{off})$ the equilibrium constant, defined by the on and off-rates describing the cross-link reaction or interaction (Figure 5).

With a few assumptions, such as large chain length and not allowing intramolecular cross-links or multiple bonds between the same two chains, the equilibrium properties of this model can be obtained analytically (Cohen and Benedek, 1982; Semenov and Rubinstein, 1998). Above a critical concentration of the interacting chains, C*, phase separation occurs creating a multi-molecular assembly[7]. Under these conditions, C* varies as $1/K_{eq}f^2$. Thus the critical concentration for formation of the assembly depends sensitively on valency and less so on the binding constant.

We carried out computer simulations of the model (relaxing some of the assumptions in the equilibrium theories noted above) to explore

[7] In our simplified model, the formation of an assembly is consistent with both gelation/phase separation. In general, the inter-play between these two distinct transitions is an interesting topic, and recent simulation-based studies (Harmon et al., 2017) have shed light on this topic and largely consistent with the early predictions of (Semenov and Rubinstein, 1998).

its dynamic, rather than equilibrium, properties. In dynamic computer simulations of the model, the valency changes between 0 and "*f*" as the residues are modified and de-modified; the rates of the modification and de-modification reactions are not varied in our studies. The modifier to demodifier ratio (e.g., kinase to phosphatase ratio) in the system determines the number of sites on each component that are modified and can be cross-linked, and is varied in our studies.

The model was simulated with N chains in a fixed volume representing the region where various components of the enhancer or SE are concentrated. We considered various values of N. During the simulation, the chains can undergo modifications and de-modifications with kinetic constants $k_{mod}=k_{demod}=0.05$. The modifier and demodifier levels (N_{mod}, N_{demod}) are varied. Cross-link formation and disassociation is simulated with kinetic constants, $k_{on}=k_{off}=0.5, K_{eq}=1$. Only modified residues on different chains were allowed to cross-link - i.e., intra-chain cross-linking reactions are disallowed, but multiple bonds can form between two chains. The simulations were carried out in the limit where every site on every chain is permitted to cross-link with all other sites on other chains(Cohen and Benedek, 1982; Semenov and Rubinstein, 1998) – i.e., while there is an average concentration of interacting sites (determined by N and the number of modified sites); variations in local concentrations within the simulation volume are not considered.

The simulations were carried out using the Gillespie algorithm (Gillespie, 1977), which generates stochastic trajectories of the temporal evolution of the considered dynamic processes (i.e., modifications and cross-linking reactions). Any single trajectory describes the time-evolution of the state of interacting chains, including how they are distributed amongst clusters of varying sizes. All trajectories are initialized with demodified, non-crosslinked chains- i.e., each chain is in a "separate cluster". Simulations are run until steady state is reached, where properties of the system (e.g. average cluster size) are time-invariant. Multiple trajectories (50 replicates) are performed for all calculations to obtain statistically averaged properties when desired.

The proxy for transcriptional activity (TA) in the simulations was defined as the size of the largest cluster of cross-linked chains, scaled by the total number of chains [TA=(size of $Cluster_{max}$) / N]. The approximation of TA with the size of the largest cluster is supported by recent evidence that the concentration of some transcriptional regulators may be rate-limiting for gene activity in mammalian cells (Lin et al., 2012; Nie et al., 2012). When all chains in the system form a single cross-linked cluster (TA≈1), the phase-separated assembly results. This assembly is thought to encompass binding of factors at the enhancer/SE and also at the promoter, which leads to the concentration of components important for enhanced transcription of the gene. We recorded the transcriptional activity generated by the enhancers and SEs as a function of time.

Transcriptional regulation with changes in valency

Modeling transcriptional activity as a function of valency revealed that the formation of SEs involved more pronounced co-operativity than the formation of typical enhancers (Figure 6). In these simulations, SEs were modeled as a system consisting of N=50 molecules, and typical enhancers as a system consisting of N=10 molecules, consistent with an approximately one order of magnitude difference in the density of components at these elements(Hnisz et al., 2013). We then graphed the transcriptional activity (TA) for different valences, while all other parameters remained constant. SEs reached ~90% of the maximum transcriptional activity at a normalized valency value of 2 (i.e. twice the reference value of f=3), while for typical enhancers 90% of the maximum transcriptional activity is attained at a normalized valency value of 5. At a normalized valency value of 2, typical enhancers reached ~40% of the maximum transcriptional activity (Figure 6). These results suggestthat, under identical conditions, SEs consisting of a larger number of components form larger connected clusters (i.e. undergo phase separation) at a lower level of valency than typical enhancers consisting of a smaller number of components. Furthermore, we observed a sharp increase of transcriptional activity at a normalized valency value of ~1.5 for SEs, while increases in valency leads to a more moderate,

smooth increase of transcriptional activity for typical enhancers (Figure 6), in agreement with previous considerations (Figure 1) (Lovén et al., 2013).

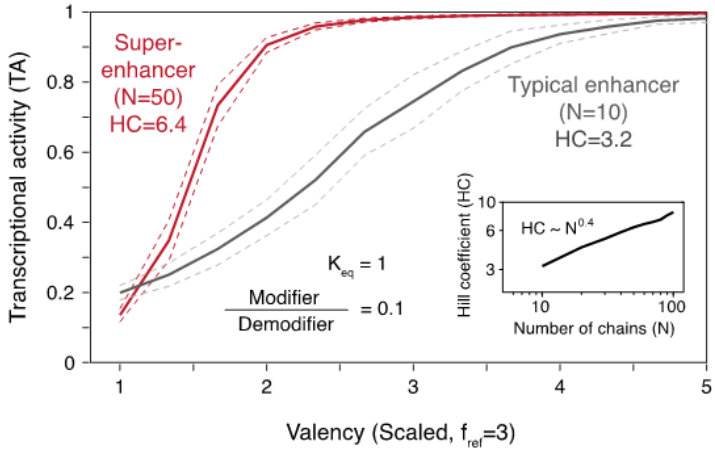


Figure 6 Dependence of transcriptional activity (TA) on the valency parameter for super- enhancers (consisting of N=50 chains), and typical enhancers (consisting of N=10 chains). The proxy for transcriptional activity (TA) is defined as the size of the largest cluster of cross-linked chains, scaled by the total number of chains. The valency is scaled such that the actual valency is divided by a reference number of 3. The solid lines indicate the mean and the dashed lines indicate twice the standard deviation in 50 simulations. The value of Keq and modifier/demodifier ratio was kept constant. HC= Hill coefficient, which is a classic metric to describe co-operative behavior. The inset shows the dependency of the Hill co-efficient on the number of chains, or components, in the system.

The sharper change in transcriptional activity of SEs upon changing the valency of the interacting components due to enhanced co-operativity can be quantified by the Hill coefficient. The behavior of SEs is characterized by a larger value of the Hill coefficient, indicating greater co-operativity and ultrasensitivity to valency changes (Figure 6). Indeed, as the inset in Figure 2C shows, the Hill coefficient increases with the number of components involved in the enhancer as ~ N0.4, over a large range of values of N. Also, as expected, the difference between the transcriptional activity of typical enhancers and SEs correlated with the difference in values of "N" that are used to model them; for a sufficiently large difference in N, the behavior reported in Figure 6 is recapitulated (Supplemental Figure 1).

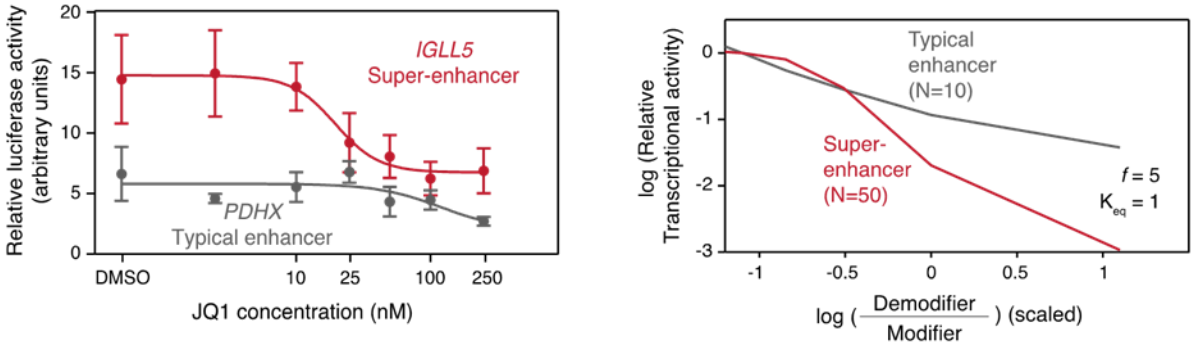
Super-enhancer formation and vulnerability

These predictions of the phase separation model are qualitatively consistent with previously published experimental data. For example, stimulation of endothelial cells by TNFα leads to the formation of SEs at inflammatory genes (Brown et al., 2014). In this study, SE formation was monitored by the genomic occupancy of the transcriptional co-factor BRD4, which is a key component of SEs and typical enhancers. The inflammatory stimulation in these cells result-

ed in a more pronounced recruitment of BRD4 at the SEs of inflammatory genes as compared to typical enhancers at other genes (Brown et al., 2014). Our phase separation model suggests that this is because stimulation by TNFα led to modifications that change the valency of interacting components, and for SEs, phase separation occurs sharply above a lower value of valency compared to typical enhancers, thus resulting in enhanced recruitment of interacting components such as BRD4 (Figure 6).

We next investigated whether the phase separation model explains the unusual vulnerability of SEs to perturbation by inhibitors of common transcriptional co-factors. BRD4 and CDK7 are components of both typical enhancers and SEs, but SEs and their associated genes are much more sensitive to chemical inhibition of BRD4 and CDK7 than typical enhancers (Figure 7) (Chipumuro et al., 2014; Christensen et al., 2014; Kwiatkowski et al., 2014; Lovén et al., 2013). We modeled the effect of BRD4- and CDK7 inhibitors as reducing valency by changing the ratio of Demodifier/Modifier activity in our system, which shifts the balance of modified sites within the interacting molecules. This is because CDK7 is a kinase which acts as a modifier, and BRD4 has a large valency as it can interact with many components, and so inhibiting BRD4 reduces the average valency of the interacting components disproportionately. As shown in Figure 7, SEs (N=50) lose more of their activity sharply at a lower Demodifier/Modifier ratio than typical enhancers (N=10). These results are consistent with the notion that SE activity is very sensitive to variations in valency because phase separation is a co-operative phenomenon that occurs suddenly when a key variable exceeds a threshold value.

Figure 7 (**left**) Enhancer activities of the fragments of the IGLL5 super-enhancer (red) and the PDHX typical enhancer (gray) after treatment with the BRD4 inhibitor JQ1 at the indicated concentrations. Enhancer activity was measured in luciferase reporter assays in human multiple myeloma cells. Note that JQ1 inhibits ~50% of luciferase expression driven by the super-enhancer at a 10-fold lower concentration than luciferase expression driven by the typical enhancer (25nM vs 250nM). Data and image adapted from (Loven et al., 2013). (**right**) Dependence of transcriptional activity (TA) on the demodifier/modifier ratio for super-enhancers (consisting of N=50 chains), and typical enhancers (consisting of N=10 chains). The proxy for transcriptional activity (TA) is defined as the size of the largest cluster of cross-linked chains, scaled by the total number of chains. The solid lines indicate the mean and the dashed lines indicate twice the standard deviation of 50 simulations. Keq and f were kept constant. Note that increasing the demodifier levels is equivalent to inhibiting cross-linking (i.e. reducing valency). TA is normalized to the value at log (demodifier/modifier) = -1.5 and the ordinate shows the normalized TA on a log scale.



Transcriptional bursting

Gene expression in eukaryotes is generally episodic, consisting of transcriptional bursts, and we investigated whether the phase-separation model can predict transcriptional bursting[8]. A recent study using quantitative imaging of transcriptional bursting in live cells suggested that the level of gene expression driven by an enhancer correlates with the frequency of transcriptional bursting (Fukaya et al., 2016). Strong enhancers were found to drive higher frequency bursting than weak enhancers, and above a certain level of strength the bursts were not resolved anymore and resulted in a relatively constant high transcriptional activity (Figure 8). The phase separation model shows that SEs recapitulate the high frequency with low variation (around a relatively constant high transcriptional activity) bursting pattern exhibited by strong enhancers while typical enhancers exhibit more variable bursts with a lower frequency (Figure 8). Once sustained phase separation occurs (TA saturates), fluctuations are quenched, which results in lower variation in TA for SEs. This difference in bursting patterns can be quantified by translating our results to a power spectrum (data not shown). We expect that strong enhancers, in spite of having fewer components (N) than SEs will form stable phase-separated multi- molecular assemblies more readily than typical enhancers because of higher valency cross-links. Therefore, a prediction of our model is that strong enhancers, like SE, should display a different transcriptional bursting pattern-compared to weak or typical enhancers.

[8]A recent study by (Larsson et al., 2019) combines single-cell RNA sequencing with model-building to shed light on genome-wide burst patterns. The interested reader may refer for more details.

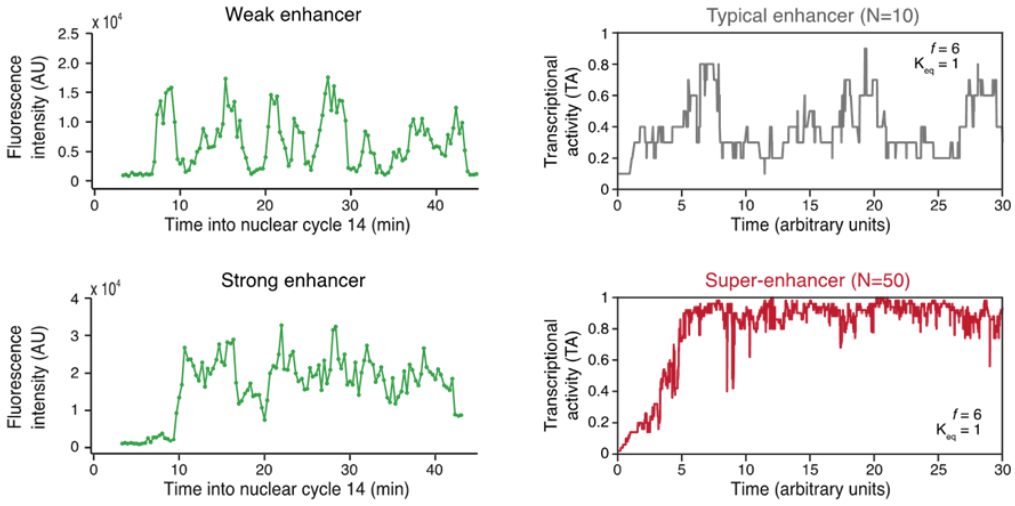


Figure 8 **(left)** Representative traces of transcriptional activity in individual nuclei of Drosophila embryos. Transcriptional activity was measured by visualizing nascent RNAs using fluorescent probes. Top panel shows a representative trace produced by a weak enhancer, the bottom panel shows a representative trace produced by a strong enhancer. Data and image adapted from (Fukaya et al., 2016) **(right)** Simulation of transcriptional activity (TA) of super-enhancers (N=50 chains), and typical enhancers (N=10 chains) over time recapitulates bursting behavior of weak and strong enhancers.

The phase separation model is also broadly consistent with the intriguing observation that two promoters can exhibit synchronous bursting when activated by the same enhancer (Fukaya et al., 2016); in this case the phase-separated assembly incorporates the enhancer and both promoters (Figure 9).

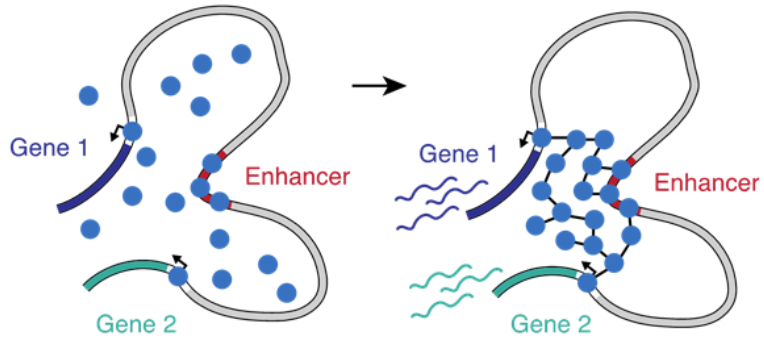


Figure 9 Model of synchronous activation of two gene promoters by a shared enhancer.

Candidate transcriptional regulators forming the phase-separated assembly *in vivo*

In our simplified model, phase separation is mediated by changes in the extent to which residues on the interacting components are modified (or valency), with resulting intermolecular-interactions. In reality, however, enhancers are composed of many diverse factors that could account for such interactions, most of which are subject to reversible chemical modifications (Figure 10).

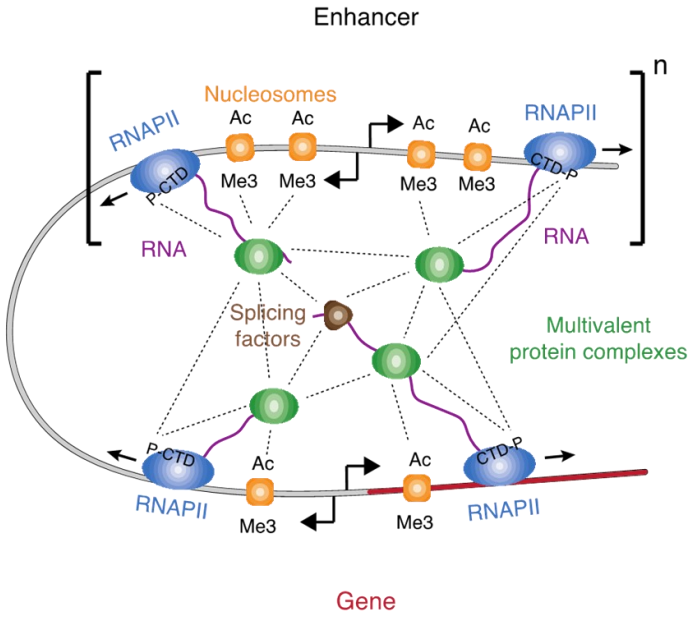


Figure 10 Model of a phase-separated complex at gene regulatory elements. Some of the candidate transcriptional regulators forming complex are highlighted. P-CTD denotes the phosphorylated C-terminal domain of RNA polymerase II (RNAPII). Chemical modifications of nucleosomes (Acetylation: Ac; and Methylation: Me3) are also highlighted. Divergent transcription at enhancers and promoters produces nascent RNAs that can be bound by RNA splicing factors. Potential interactions between the components are displayed as dashed lines.

These components include transcription factors, transcriptional co-activators such as the Mediator complex and BRD4, chromatin regulators (e.g.readers, writers and erasers of histone modifications), cyclin-dependent kinases (e.g. CDK7, CDK8, CDK9, CDK12), non- coding RNAs with RNA-binding proteins and RNA polymerase II (Lai and Shiekhattar, 2014; Lee and Young, 2013; Levine et al., 2014; Malik and Roeder, 2010). Many of these molecules are multivalent, i.e. contain multiple modular domains or interaction motifs, and are thus able to interact with multiple other enhancer components. For example, the large subunit of RNA polymerase II contains 52 repeats of a heptapeptide sequence at its C-terminal domain (CTD) in human cells, and several transcription factors contain repeats of low-complexity domains or repeats of the same amino-acid stretch prone to polymerization (Gemayel et al., 2015; Kwon et al., 2013). The DNA portion of enhancers and many promoters contain binding sites for multiple transcription factors, some of which can bind simultaneously to both DNA and RNA (Sigova et al., 2015). Histone proteins at enhancers are enriched for modifications that can be recognized by chromatin readers, and thus adjacent nucleosomes can be considered as a platform able to interact with multiplechromatin readers. RNA itself can be chemically modified and physically interact with multiple RNA-binding molecules and splicing factors[9]. Many of the residuesinvolved in these interactions can create a "cross -link" (Figure 10).

[9]Indeed, since our original hypothesis was published, numerous studies have provided experimental data that supports the potential of various transcription-associated molecules to phase separate-examples include CTD (Boehning et al., 2018; Guo et al., 2019; Lu et al., 2018), Histone molecules (Gibson et al., 2019), transcriptional activators (Boija et al., 2018; Nair et al.), and coactivators (Sabari et al., 2018).

Possible implications and predictions of the phase separation model

Our simple phase separation model provides a conceptual framework for further exploration of principles of gene control in development and disease. Below we discuss a few examples of phenomena possibly related to assemblies of phase separated multi-molecular complexes in transcriptional control and some testable predictions of the model.

Visualization of phase separated multi-molecular assemblies of transcriptional regulators

A critical test of the model is whether phase separation of multi-molecularassemblies of transcriptional regulators can be directly observed *in vivo*, with the demonstration that phase separation of those complexes is associated with gene activity. Several lines of recent work provide initial insights into these questions. For example, recent studies using high resolution microscopy indicate that signal stimulation leads to the formation of large clusters of RNA polymerase II in living mammalian cells (Cisse et al., 2013) and concordant activation of transcription at a subset of genes (Cho et al., 2016). This, as well as other single molecule technologies (Chen and Larson, 2016; Shin et al., 2017), may thus enable visualization and testing of whether phase separated multi-molecular complexes form in the vicinity of genes regulated by SEs, and whether the simple model we describe here predicts features of transcriptional control[10]. As an example, we hypothesize that the RNAPII C-terminal domain, which consists of 52 heptapeptide repeats, is a key contributor to the valency within this assembly, and in cells that express an RNAPII with a truncated CTD, the clusters would exhibit significantly lower half-lives[11].

[10] (Cho et al., 2018; Sabari et al., 2018) provide visual evidence to support our model

[11] Data in (Lu et al., 2019) is consistent with these predictions, but do not exclude other models.

Signal-dependent gene control

Cells sense and respond to their environment through signal transduction pathways that relay information to genes, but genes responding to a particular signaling pathway may exhibit different amplitudes of activation to the same signal. We have carried out calculations with the hypothesis that once phase separation occurs, the assembly recruits components that are de-modifiers. Under these conditions, transition to and resolution of phase separation, i.e. transcriptional activity, are more distinct for SEs compared to typical enhancers. Interestingly, such simulations suggest that there is a maximum valency and a maximum number of SE components, which if exceeded, does not allow disassembly in a realistic time scale (Supplemental Figure 2). This is because the molecules are so heavily cross-linked that it remains in a metastable state for long periods of time. The prediction of the model is that pathological hyperactivation of cellular signaling could underlie disease states through locking cells in an expression program that - at least transiently – becomes unresponsive to signals that would counteract them under

normal physiological conditions[12]. We speculate that such states can be artificially induced by increasing the valency or number of interacting components.

[12] (Nair et al.) provides evidence for this prediction where pathological hyperactivation of estrogen receptors leads to the formation of "gel-like" arrested assemblies.

Fidelity of transcriptional control

Variability in the transcript levels of genes within isogenic population of cells exposed to the same environmental signals – referred to as transcriptional noise – can have a profound impact on cellular phenotypes (Raj and van Oudenaarden, 2008). The phase separation model indicates that because of the high co-operativity involved in the formation of SEs, transcription occurs when the valency (modulated by the modifier/demodifier ratio, which is in fact similar to the developmental signals being transduced through activation cascades) exceeds a sharply defined threshold(Figure 2C). For the smaller number of components in a typical enhancer, the variation of transcription with the environmental signal is more continuous, potentially leading to "noisier" or more error-prone transcription over a wider range of signal strength. In the vicinity of a phase separation point, there are fluctuations between the two phases (low TA and robust TA in our case). Our model shows that these fluctuations (or noise) is confined to a narrow range of environmental signals for SEs compared to the broad range over which this occurs for a typical enhancer (Supplemental Figure 3). The normalized amplitude of these fluctuations is also smaller for SEs. These results suggest that one reason why SEs have evolved is to enable relatively error free and robust transcription of genes necessary to maintain cell identity. This form of transcriptional fidelity through co-operativity, and not chemical specificity mediated by evolving specific molecules for controlling each gene, may however be co-opted to drive aberrant gene expression in disease states (e.g., SEs in cancer cells).

Resistance to transcriptional inhibition

Small molecule inhibitors of super-enhancer components such as BRD4 arecurrently being tested as anticancer therapeutics in the clinic, where a ubiquitous challenge has been the emergence of tumor cells resistant to the targeted therapeutic agent. Interestingly, recent studies revealed that resistance to JQ1, a drug that inhibits

BRD4, develops without any genetic changes in various tumor cells (Fong et al., 2015; Rathert et al., 2015; Shu et al., 2016). While JQ1 inhibits the interaction of BRD4 with acetylated histones, BRD4 is still recruited to super- enhancers due to its hyper-phosphorylation in JQ1-resistant cells (Shu et al., 2016). This is consistent with a prediction of our model that BRD4 is a high valency component of SEs, and inhibition of its interaction with acetylated histones (i.e. decrease of its valency) may be compensated for by increasing its valency through the activation of kinase pathways targeting BRD4 itself. In our model, super- enhancers are characterized by a high Hill coefficient, i.e. high co-operativity (Figure 2C), which suggests that inhibition of multiple properly chosen SE components might have a synergistic effect SE-driven oncogenes in tumor cells. If this prediction is true, resistance to BRD4 inhibitors may be prevented through combined treatment with additional inhibitors of transcriptional regulators[13].

[13] In (Klein et al., 2020), we explored a version of this question. Rather than combination therapy, can one identify drugs that partition into certain condensates based on their biochemistry and can this be exploited for rational drug design? Evidence from a study from the Hyman Lab (Wheeler et al., 2019) has indeed shown promise for condensate modulators to ameliorate disease phenotypes

Concluding remarks

The essential feature of this phase separation model of transcriptional control is that it considers cooperativity between the interacting components in the context of changes in valency and number of components. This single conceptual framework consistently describes diverse recently observed features of transcriptional control, such as clustering of factors, dynamic changes, hyper-sensitivity of SEs to transcriptional inhibitors, and simultaneous activation of multiple genes by the same enhancer. Cellular signaling pathways could modulate transcription over short time periods by alterations of valency. Selection of cell growth and survival would expand or contract the number of interactions or size of the enhancer over longer times. The model also makes a number of predictions (some noted above) that could be explored in many cellular contexts. Such studies, and others that will be envisaged, will help determine whether a variant of the model we propose underlies transcriptional control in mammals. Also, attractively, this model sets enhancer, and especially super-enhancer -type gene regulation into the broad family of membraneless organelles such as the nucleolus, Cajal bodies

and splicing-speckles in the nucleus, and stress granules and P bodies in the cytoplasm, as results of phase-separated multi-molecular assemblies.

Experimental support for transcriptional condensate formation

To explore whether transcriptional molecules formed condensates in live cells, we generated two engineered murine embryonic stem cell (ESC) lines, using CRISPR/Cas9 to endogenously tag key SE associated proteins, BRD4 and MED1(Sabari et al., 2018)[14]. Live cell microscopy, as well as fixed cell analyses, of these cell-lines revealed the presence of discrete nuclear puncta (Figure 11A-B). To determine if these nuclear puncta colocalized with SEs, we combined Immunofluorescence (IF) for BRD4 and MED1 with DNA and RNA-FISH for several SE genes. We developed an image analysis pipeline to quantify colocalization between IF and FISH[15]. Deploying this pipeline over 100s of cells, we determined that SE genes colocalized with high intensities of BRD4 and MED1, as opposed to typical enhancers or random controls (Figure 11D-G). Importantly, these condensates were associated with RNA, indicating that they are associated with transcription. Live-cell Fluorescence Recovery After Photobleaching (FRAP) experiments showed that the transcriptional condensates have liquid-like properties. We employed bioinformatics analysis to characterize structural disorder and compositional bias in SE components, which are strong biochemical correlates of phase separation (Banani et al., 2017; Brangwynne et al., 2015). This analysis informed and complemented a series of *in vitro* and engineered *in vivo* assays (based on (Shin et al., 2017)), which demonstrated that intrinsically disordered regions (IDRs) of BRD4 and MED1 are capable of driving phase separation on their own. Overall, this study provides concrete evidence that transcriptional coactivators form condensates in cells.

[14] See (Sabari et al., 2018) for detailed experimental characterization and explanation of methods and choice of analyses tools. Key data is reproduced here for the purpose of this thesis.

[15] The code developed can be found at https://github.com/krishna-shrinivas/FISH_IF_colocalization

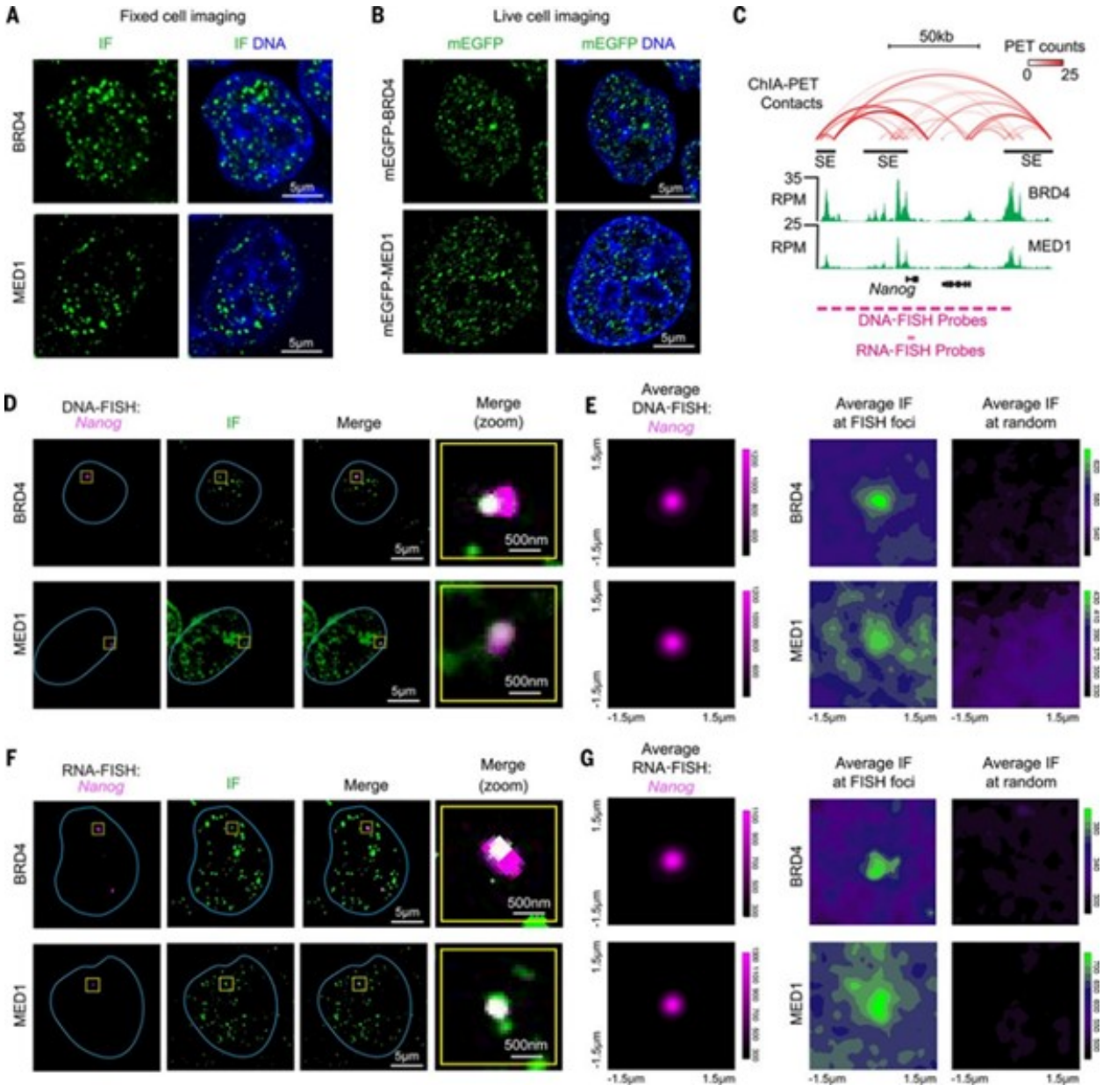


Figure 11 BRD4 and MED1 form puncta at super-enhancers (SEs). (**A**) Immunofluorescence (IF) imaging of BRD4 and MED1 in mouse embryonic stem cells (mESCs). Fluorescence signal is shown alone (left) and merged with Hoechst stain (right). (**B**) Live imaging of endogenously tagged mEGFP-BRD4 and mEGFP-MED1 in mESCs. (**C**) Depiction of *Nanog* locus, associated SEs (black bars), DNA contacts (red arcs), BRD4 and MED1 ChIP-seq (green histograms), and location of FISH probes. ChIA-PET, chromatin interaction analysis with paired-end tag; RPM, reads per million. (**D**) Colocalization between BRD4 or MED1 and the *Nanog* locus by IF and DNA-FISH in fixed mESCs. Separate images of the indicated IF and FISH are shown, along with an image showing the merged channels (overlapping signal in white). The blue line highlights the nuclear periphery, determined by Hoechst staining (not shown). The rightmost column shows the area in the yellow box in greater detail. (**E**) Averaged signal of (left) DNA-FISH for *Nanog* and (right) IF for BRD4 or MED1 centered at *Nanog* DNA-FISH foci or randomly selected nuclear positions. (**F**) Colocalization between BRD4 or MED1 and the nascent RNA of *Nanog*, determined by IF and RNA-FISH in fixed mESCs. Data are shown as in (D). (**G**) Averaged signal of (left) RNA-FISH for *Nanog* and (right) IF for BRD4 or MED1 centered at *Nanog* RNA -FISH foci or randomly selected nuclear positions.

Bibliography

Banani, S.F., Lee, H.O., Hyman, A.A., and Rosen, M.K. (2017). Biomolecular condensates: organizers of cellular biochemistry. Nat. Rev. Mol. Cell Biol. *18*, 285–298.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. Cell *27*, 299–308.

Banjade, S., Wu, Q., Mittal, A., Peeples, W.B., Pappu, R. V., and Rosen, M.K. (2015). Conserved interdomain linker promotes phase separation of the multivalent adaptor protein Nck. Proc. Natl. Acad. Sci. U. S. A. *112*.

Benabdallah, N.S., Williamson, I., Illingworth, R.S., Kane, L., Boyle, S., Sengupta, D., Grimes, G.R., Therizols, P., and Bickmore, W.A. Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation. Mol. Cell *0*.

Benoist, C., and Chambon, P. (1981). In vivo sequence requirements of the SV40 early promoter region. Nature *290*, 304–310.

Bergeron-Sandoval, L.-P.P., Safaee, N., and Michnick, S.W. (2016). Mechanisms and Consequences of Macromolecular Phase Separation (Cell Press).

Berry, J., Weber, S.C., Vaidya, N., Haataja, M., Brangwynne, C.P., and Weitz, D.A. (2015). RNA transcription modulates phase transition-driven nuclear body assembly. Proc. Natl. Acad. Sci. U. S. A. *112*, E5237–E5245.

Boehning, M., Dugast-Darzacq, C., Rankovic, M., Hansen, A.S., Yu, T., Marie-Nelly, H., McSwiggen, D.T., Kokic, G., Dailey, G.M., Cramer, P., et al. (2018). RNA polymerase II clustering through carboxy-terminal domain phase separation. Nat. Struct. Mol. Biol. *25*, 833–840.

Boija, A., Klein, I.A., Sabari, B.R., Dall'Agnese, A., Coffey, E.L., Zamudio, A. V., Li, C.H., Shrinivas, K., Manteiga, J.C., Hannett, N.M., et al. (2018). Transcription factors activate genes through the phase separation capacity of their activation domains. Cell *in press*.

Brangwynne, C.P., Eckmann, C.R., Courson, D.S., Rybarska, A., Hoege, C., Gharakhani, J., Jülicher, F., and Hyman, A.A. (2009). Germline P granules are liquid droplets that localize by controlled dissolution/condensation. Science *324*, 1729–1732.

Brangwynne, C.P., Tompa, P., and Pappu, R. V. (2015). Polymer physics of intracellular phase transitions. Nat. Phys. *11*, 899–904.

Brown, J.D., Lin, C.Y., Duan, Q., Griffin, G., Federation, A.J., Paranal, R.M., Bair, S., Newton, G., Lichtman, A.H., Kung, A.L., et al. (2014). NF-κB directs dynamic super enhancer formation in inflammation and atherogenesis. Mol. Cell *56*, 219–231.

Bulger, M., and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. Cell *144*, 327–339.

Carey, M. (1998). The enhanceosome and transcriptional synergy. Cell *92*, 5–8.

Chapuy, B., McKeown, M.R., Lin, C.Y., Monti, S., Roemer, M.G.M., Qi, J., Rahl, P.B., Sun, H.H., Yeda, K.T., Doench, J.G., et al. (2013). Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma. Cancer Cell *24*, 777–790.

Chen, H., and Larson, D.R. (2016). What have single-molecule studies taught us about gene expression? (Cold Spring Harbor Laboratory Press).

Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J.B., and Gregor, T. (2018). Dynamic interplay between enhancer–promoter topology and gene activity. Nat. Genet. 1.

Chipumuro, E., Marco, E., Christensen, C.L., Kwiatkowski, N., Zhang, T., Hatheway, C.M., Abraham, B.J., Sharma, B., Yeung, C., Altabef, A., et al. (2014). CDK7 inhibition suppresses super-enhancer-linked oncogenic transcription in MYCN-driven cancer. Cell *159*, 1126–1139.

Cho, W.-K., Jayanth, N., English, B.P., Inoue, T., Andrews, J.O., Conway, W., Grimm, J.B., Spille, J.-H., Lavis, L.D., Lionnet, T., et al. (2016). RNA Polymerase II cluster dynamics predict mRNA output in living cells. Elife *5*, 13617.

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

Choi, J.-M., Holehouse, A.S., and Pappu, R. V. (2020). Physical Principles Underlying the Complex Biology of Intracellular Phase Transitions. Annu. Rev. Biophys. *49*.

Christensen, C.L., Kwiatkowski, N., Abraham, B.J., Carretero, J., Al-Shahrour, F., Zhang, T., Chipumuro, E., Herter-Sprie, G.S., Akbay, E.A., Altabef, A., et al. (2014). Targeting Transcriptional Addictions in Small Cell Lung Cancer with a Covalent CDK7 Inhibitor. Cancer Cell *26*, 909–922.

Cisse, I.I., Izeddin, I., Causse, S.Z., Boudarene, L., Senecal, A., Muresan, L., Dugast-Darzacq, C., Hajj, B., Dahan, M., and Darzacq, X. (2013). Real-time dynamics of RNA polymerase II clustering in live human cells. Science *341*, 664–667.

Cohen, R.J., and Benedek, G.B. (1982). Equilibrium and kinetic theory of polymerization and the sol-gel transition. J. Phys. Chem. *86*, 3696–3714.

Dimitrova, N., Zamudio, J.R., Jong, R.M., Soukup, D., Resnick, R., Sarma, K., Ward, A.J., Raj, A., Lee, J.T., Sharp, P.A., et al. (2014). LincRNA-p21 activates p21 in cis to promote Polycomb target gene expression and to enforce the G1/S checkpoint. Mol. Cell *54*, 777–790.

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K., et al. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell *159*, 374–387.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. Science *297*, 1183–1186.

Engreitz, J.M., Haines, J.E., Perez, E.M., Munson, G., Chen, J., Kane, M., McDonel, P.E., Guttman, M., and Lander, E.S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature *539*, 452–455.

Falvo, J. V., Thanos, D., and Maniatis, T. (1995). Reversal of intrinsic DNA bends in the IFNβ gene enhancer by transcription factors and the architectural protein HMG I(Y). Cell *83*, 1101–1111.

Feric, M., Vaidya, N., Harmon, T.S., Mitrea, D.M., Zhu, L., Richardson, T.M., Kriwacki, R.W., Pappu, R. V., and Brangwynne, C.P. (2016). Coexisting Liquid Phases Underlie Nucleolar Subcompartments. Cell *165*, 1686–1697.

Fong, C.Y., Gilan, O., Lam, E.Y.N., Rubin, A.F., Ftouni, S., Tyler, D., Stanley, K., Sinha, D., Yeh, P., Morison, J., et al. (2015). BET inhibitor resistance emerges from leukaemia stem cells. Nature *525*, 538–542.

Forero-Quintero, L., Raymond, W., Handa, T., Saxton, M., Morisaki, T., Kimura, H., Bertrand, E., Munsky, B., and Stasevich, T. (2020). Live-cell imaging reveals the spatiotemporal organization of endogenous RNA polymerase II phosphorylation at a single gene. BioRxiv 2020.04.03.024414.

Fukaya, T., Lim, B., and Levine, M. (2016). Enhancer Control of Transcriptional Bursting. Cell *166*, 358–368.

Gemayel, R., Chavali, S., Pougach, K., Legendre, M., Zhu, B., Boeynaems, S., van der Zande, E., Gevaert, K., Rousseau, F., Schymkowitz, J., et al. (2015). Variable Glutamine-Rich Repeats Modulate Transcription Factor Activity. Mol. Cell *59*, 615–627.

Gibson, B.A., Doolittle, L.K., Schneider, M.W.G., Jensen, L.E., Gamarra, N., Henry, L., Gerlich, D.W., Redding, S., and Rosen, M.K. (2019). Organization of Chromatin by Intrinsic and Regulated Phase Separation. Cell *0*.

Gillespie, D.T. (1977). Exact stochastic simulation of coupled chemical reactions.

Gruss, P., Dhar, R., and Khoury, G. (1981). Simian virus 40 tandem repeated sequences as an element of the early promoter. Proc. Natl. Acad. Sci. U. S. A. *78*, 943–947.

Guo, Y.E., Manteiga, J.C., Henninger, J.E., Sabari, B.R., Dall'Agnese, A., Hannett, N.M., Spille, J.-H.J.-H., Afeyan, L.K., Zamudio, A. V., Shrinivas, K., et al. (2019). Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. Nature *572*, 543–548.

Hah, N., Benner, C., Chong, L.-W.W., Yu, R.T., Downes, M., Evans, R.M., and Evansa, R.M. (2015). Inflammation-sensitive super enhancers form domains of coordinately regulated enhancer RNAs. *112*, E297–E302.

Han, T.W., Kato, M., Xie, S., Wu, L.C., Mirzaei, H., Pei, J., Chen, M., Xie, Y., Allen, J., Xiao, G., et al. (2012). Cell-free Formation of RNA Granules: Bound RNAs Identify Features and Components of Cellular Assemblies. Cell *149*, 768–779.

Harmon, T.S., Holehouse, A.S., Rosen, M.K., and Pappu, R. V (2017). Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. Elife *6*.

Hay, D., Hughes, J.R., Babbs, C., Davies, J.O.J.J., Graham, B.J., Hanssen, L.L.P.P., Kassouf, M.T., Oudelaar, A.M., Sharpe, J.A., Suciu, M.C., et al. (2016). Genetic dissection of the α-globin super-enhancer in vivo. Nat. Genet. *48*, 895–903.

Heist, T., Fukaya, T., and Levine, M. (2019). Large distances separate coregulated genes in living Drosophila embryos. Proc. Natl. Acad. Sci. U. S. A. *116*, 15062–15067.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-Enhancers in the Control of Cell Identity and Disease (Cell Press).

Hnisz, D., Schuijers, J., Lin, C.Y., Weintraub, A.S., Abraham, B.J., Lee, T.I., Bradner, J.E., and Young, R.A. (2015). Convergence of Developmental and Oncogenic Signaling Pathways at Transcriptional Super-Enhancers. Mol. Cell *58*, 362–370.

Hnisz, D., Weintrau, A.S., Day, D.S., Valton, A.L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., et al. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science *351*, 1454–1458.

Ji, X., Dadon, D.B., Powell, B.E., Fan, Z.P., Borges-Rivera, D., Shachar, S., Weintraub, A.S., Hnisz, D., Pegoraro, G., Lee, T.I., et al. (2016). 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. Cell Stem Cell *18*, 262–275.

Jiang, T., Raviram, R., Snetkova, V., Rocha, P.P., Proudhon, C., Badri, S., Bonneau, R., Skok, J.A., and Kluger, Y. (2016). Identification of multi-loci hubs from 4C-seq demonstrates the functional importance of simultaneous interactions. Nucleic Acids Res. *44*, 8714–8725.

Johnson, A.D., Meyer, B.J., and Ptashne, M. (1979). Interactions between DNA-bound repressors govern regulation by the λ phage repressor. Proc. Natl. Acad. Sci. U. S. A. *76*, 5061–5065.

Kato, M., Han, T.W., Xie, S., Shi, K., Du, X., Wu, L.C., Mirzaei, H., Goldsmith, E.J., Longgood, J., Pei, J., et al. (2012a). Cell-free formation of RNA granules: Low complexity sequence domains form dynamic fibers within hydrogels. Cell *149*, 753–767.

Kato, M., Han, T.W., Xie, S., Shi, K., Du, X., Wu, L.C., Mirzaei, H., Goldsmith, E.J., Longgood, J., Pei, J., et al. (2012b). Cell-free Formation of RNA Granules: Low Complexity Sequence Domains Form Dynamic Fibers within Hydrogels. Cell *149*, 753–767.

Kieffer-Kwon, K.-R.R., Tang, Z., Mathe, E., Qian, J., Sung, M.-H.H., Li, G., Resch, W., Baek, S., Pruett, N., Grøntved, L., et al. (2013). Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. Cell *155*, 1507–1520.

Kim, T.K., and Maniatis, T. (1997). The mechanism of transcriptional synergy of an in vitro assembled interferon-β enhanceosome. Mol. Cell *1*, 119–129.

Klein, I.A., Boija, A., Afeyan, L.K., Hawken, S.W., Fan, M., Dall'Agnese, A., Oksuz, O., Henninger, J.E., Shrinivas, K., Sabari, B.R., et al. (2020). Partitioning of cancer therapeutics in nuclear condensates. Science *368*, 1386–1392.

Kwiatkowski, N., Zhang, T., Rahl, P.B., Abraham, B.J., Reddy, J., Ficarro, S.B., Dastur, A., Amzallag, A., Ramaswamy, S., Tesar, B., et al. (2014). Targeting transcription regulation in cancer with a covalent CDK7 inhibitor. Nature *511*, 616–620.

Kwon, I., Kato, M., Xiang, S., Wu, L., Theodoropoulos, P., Mirzaei, H., Han, T., Xie, S., Corden, J.L., and McKnight, S.L. (2013). Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains. Cell *155*, 1049–1060.

Lai, F., and Shiekhattar, R. (2014). Enhancer RNAs: the new molecules of transcription. Curr. Opin. Genet. Dev. *25*, 38–42.

Lai, F., Orom, U.A., Cesaroni, M., Beringer, M., Taatjes, D.J., Blobel, G.A., and Shiekhattar, R. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature *494*, 497–501.

Larochelle, S., Amat, R., Glover-Cutter, K., Sansó, M., Zhang, C., Allen, J.J., Shokat, K.M., Bentley, D.L., and Fisher, R.P. (2012). Cyclin-dependent kinase control of the initiation-to-elongation switch of RNA polymerase II. *19*, 1108–1115.

Larsson, A.J.M., Johnsson, P., Hagemann-Jensen, M., Hartmanis, L., Faridani, O.R., Reinius, B., Segerstolpe, Å., Rivera, C.M., Ren, B., and Sandberg, R. (2019). Genomic encoding of transcriptional burst kinetics. Nature *565*, 251–254.

Lee, T.I., and Young, R.A. (2013). Transcriptional Regulation and Its Misregulation in Disease. Cell *152*, 1237–1251.

Levine, M., Cattoglio, C., and Tjian, R. (2014). Looping back to leap forward: transcription enters a new era. Cell *157*, 13–25.

Li, P., Banjade, S., Cheng, H.-C.C., Kim, S., Chen, B., Guo, L., Llaguno, M., Hollingsworth, J. V., King, D.S., Banani, S.F., et al. (2012). Phase transitions in the assembly of multivalent signalling proteins. Nature *483*, 336–340.

Lin, C.Y., Lovén, J., Rahl, P.B., Paranal, R.M., Burge, C.B., Bradner, J.E., Lee, T.I., and Young, R.A. (2012). Transcriptional amplification in tumor cells with elevated c-Myc. Cell *151*, 56–67.

Long, H.K., Prescott, S.L., and Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell *167*, 1170–1187.

Lovén, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I., and Young, R.A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell *153*, 320–334.

Lu, F., Portz, B., and 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. Mol. Cell *73*, 1232-1242.e4.

Lu, H., Yu, D., Hansen, A.S., Ganguly, S., Liu, R., Heckert, A., Darzacq, X., and Zhou, Q. (2018). Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. Nature *558*, 318–323.

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

Mansour, M.R., Abraham, B.J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A.D., Etchin, J., Lee, L., Sallan, S.E., Silverman, L.B., et al. (2014). An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science *346*, 1373–1377.

Mao, Y.S., Zhang, B., and Spector, D.L. (2011). Biogenesis and function of nuclear bodies. Trends Genet. *27*, 295–306.

McSwiggen, D.T., Mir, M., Darzacq, X., and Tjian, R. (2019). Evaluating phase separation in live cells: diagnosis, caveats, and functional consequences. Genes Dev.

Merika, M., Williams, A.J., Chen, G., Collins, T., and Thanos, D. (1998). Recruitment of CBP/p300 by the IFNβ enhanceosome is required for synergistic activation of transcription. Mol. Cell *1*, 277–287.

Nair, S.J., Yang, L., Meluzzi, D., Oh, S., Yang, F., Friedman, M.J., Wang, S., Suter, T., Alshareedah, I., and Gamliel, A. Phase separation of ligand-activated enhancers licenses cooperative chromosomal enhancer assembly. Nat. Struct. Mol. Biol. *26*, 193–203.

Nie, Z., Hu, G., Wei, G., Cui, K., Yamane, A., Resch, W., Wang, R., Green, D.R., Tessarollo, L., Casellas, R., et al. (2012). c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. Cell *151*, 68–79.

Ong, C.T., and Corces, V.G. (2011). Enhancer function: New insights into the regulation of tissue-specific gene expression. Nat. Rev. Genet. *12*, 283–293.

Orphanides, G., and Reinberg, D. (2002). A unified theory of gene expression (Cell Press).

Palstra, R.-J., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F., and de Laat, W. (2003). The beta-globin nuclear compartment in development and erythroid differentiation. Nat. Genet. *35*, 190–194.

Parker, S.C.J., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J.A., Van Bueren, K.L., Chines, P.S., Narisu, N., Black, B.L., et al. (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc. Natl. Acad. Sci. U. S. A. *110*, 17921–17926.

Pefanis, E., Wang, J., Rothschild, G., Lim, J., Kazadi, D., Sun, J., Federation, A., Chao, J., Elliott, O., Liu, Z.-P., et al. (2015). RNA Exosome-Regulated Long Non-Coding RNA Transcription Controls Super-Enhancer Activity. Cell *161*, 774–789.

Peng, A., and Weber, S.C. (2019). Evidence for and against liquid-liquid phase separation in the nucleus. Non-Coding RNA *5*.

Phatnani, H.P., and Greenleaf, A.L. (2006). Phosphorylation and functions of the RNA polymerase II CTD. Genes Dev. *20*, 2922–2936.

Proudhon, C., Snetkova, V., Raviram, R., Lobry, C., Badri, S., Jiang, T., Hao, B., Trimarchi, T., Kluger, Y., Aifantis, I., et al. (2016). Active and Inactive Enhancers Cooperate to Exert Localized and Long-Range Control of Gene Regulation. Cell Rep. *15*, 2159–2169.

Quintero-Cadena, P., Lenstra, T.L., and Sternberg, P.W. (2020). RNA Pol II Length and Disorder Enable Cooperative Scaling of Transcriptional Bursting. Mol. Cell.

Raj, A., and van Oudenaarden, A. (2008). Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. Cell *135*, 216–226.

Raser, J.M., and O'Shea, E.K. (2004). Control of stochasticity in eukaryotic gene expression. Science *304*, 1811–1814.

Rathert, P., Roth, M., Neumann, T., Muerdter, F., Roe, J.-S.S., Muhar, M., Deswal, S., Cerny-Reiterer, S., Peter, B., Jude, J., et al. (2015). Transcriptional plasticity promotes primary and acquired resistance to BET inhibition. Nature *525*, 543–547.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–329.

Rodriguez, J., and Larson, D.R. (2020). Transcription in Living Cells: Molecular Mechanisms of Bursting. Annu. Rev. Biochem. *89*.

Sabari, B.R., Dall'Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A. V., Manteiga, J.C., et al. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. Science *361*.

Schaffner, W. (2015). Enhancers, enhancers - From their discovery to today's universe of transcription enhancers. Biol. Chem.

Semenov, A.N., and Rubinstein, M. (1998). Thermoreversible Gelation in Solutions of Associative Polymers. 1. Statics. Macromolecules *31*, 1373–1385.

Shin, Y., and Brangwynne, C.P. (2017). Liquid phase condensation in cell physiology and disease. Science *357*, eaaf4382.

Shin, H.Y., Willi, M., Yoo, K.H., Zeng, X., Wang, C., Metser, G., and Hennighausen, L. (2016). Hierarchy within the mammary STAT5-driven Wap super-enhancer. Nat. Genet. *48*, 904–911.

Shin, Y., Berry, J., Pannucci, N., Haataja, M.P., Toettcher, J.E., and Brangwynne, C.P. (2017). Spatiotemporal Control of Intracellular Phase Transitions Using Light-Activated optoDroplets. Cell *168*, 159-171.e14.

Shu, S., Lin, C.Y., He, H.H., Witwicki, R.M., Tabassum, D.P., Roberts, J.M., Janiszewska, M., Jin Huh, S., Liang, Y., Ryan, J., et al. (2016). Response and resistance to BET bromodomain inhibitors in triple-negative breast cancer. Nature *529*, 413–417.

Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C., et al. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc. Natl. Acad. Sci. *110*, 2876–2881.

Sigova, A.A., Abraham, B.J., Ji, X., Molinie, B., Hannett, N.M., Guo, Y.E., Jangi, M., Giallourakis, C.C., Sharp, P.A., and Young, R.A. (2015). Transcription factor trapping by RNA in gene regulatory elements. Science *350*, 978–991.

Snead, W.T., and Gladfelter, A.S. (2019). The Control Centers of Biomolecular Phase Separation: How Membrane Surfaces, PTMs, and Active Processes Regulate Condensation. Mol. Cell *0*.

Spitz, F., and Furlong, E.E.M.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.

Suter, D.M., Molina, N., Gatfield, D., Schneider, K., Schibler, U., and Naef, F. (2011). Mammalian genes are transcribed with widely different bursting kinetics. *332*, 472–474.

Thanos, D., and Maniatis, T. (1995). Virus induction of human IFNβ gene expression requires the assembly of an enhanceosome. Cell *83*, 1091–1100.

Tjian, R., and Maniatis, T. (1994). Transcriptional activation: A complex puzzle with few easy pieces. Cell *77*, 5–8.

Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol. Cell *10*, 1453–1465.

Wang, J., Choi, J.-M., Holehouse, A.S., Lee, H.O., Zhang, X., Jahnel, M., Maharana, S., Lemaitre, R., Pozniakovsky, A., Drechsel, D., et al. (2018). A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. Cell *174*, 688-699.e16.

Wang, Y., Zhang, T., Kwiatkowski, N., Abraham, B.J., Lee, T.I., Xie, S., Yuzugullu, H., Von, T., Li, H., Lin, Z., et al. (2015). CDK7-dependent transcriptional addiction in triple-negative breast cancer. Cell *163*, 174–186.

Wheeler, J.R., Matheny, T., Jain, S., Abrisch, R., and Parker, R. (2016). Distinct stages in stress granule assembly and disassembly. Elife *5*.

Wheeler, R.J., Lee, H.O., Poser, I., Pal, A., Doeleman, T., Kishigami, S., Kour, S., Anderson, E.N., Marrone, L., Murthy, A.C., et al. (2019). Small molecules for modulating protein driven liquid-liquid phase separation in treating neurodegenerative disease. BioRxiv 721001.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell *153*, 307–319.

Zhu, L., and Brangwynne, C.P. (2015). Nuclear bodies: The emerging biophysics of nucleoplasmic phases. Curr. Opin. Cell Biol. *34*, 23–30.

Zoller, B., Nicolas, D., Molina, N., and Naef, F. (2015). Structure of silent transcription intervals and noise characteristics of mammalian genes. Mol. Syst. Biol. *11*, 823.

Chapter 2: *Mechanism*

Genomic encoding of transcriptional regulation by phase separation

"Meaning lies as much
in the mind of the reader
as in the Haiku."
— Douglas R. Hofstadter,

Gödel, Escher, Bach: An Eternal Golden Braid

This chapter is primarily based on:
"Enhancer features that drive formation of transcriptional condensates"
K. Shrinivas[=], B. R. Sabari[=], E. L. Coffey, I. A. Klein, A. Boija, A. V. Zamudio, J. Schuijers, N. M. Hannett, P. A. Sharp[c], R. A. Young[c], A. K. Chakraborty[c],
Mol. Cell. **75**, 549-561.e7 (2019). ([=]equal contributions,[c]Corresponding author)

Enhancers are non-coding DNA elements that orchestrate the spatio-temporal recruitment and targeting of the transcriptional machinery to activate target genes. In the previous chapter, we suggested the phase separation may explain features of a class of enhancer elements known as super-enhancers and provided experimental evidence in favor of this model. How these collective assemblies (or condensates) form at specific parts of the genome (and not at others) and what features describe their interactions will be the key focus of this Chapter. Through the lens of phase separation, we will explore the implications of transcriptional condensate formation on the origins of enhancer formation and provide a general framework to describe condensates scaffolded on the genome.

Introduction

The precise regulation of gene transcription during development and in response to signals is established by the action of enhancer elements, which act as platforms for the recruitment of the gene control machinery at specific genomic loci (Levo and Segal, 2014; Long et al., 2016; Maniatis et al., 1998; Ptashne and Gann, 1997; Shlyueva et al., 2014; Spitz and Furlong, 2012). Imprecision in this process can cause disease, including cancer (Lee and Young, 2013; Smith and Shilatifard, 2014) Enhancer sequences contain short DNA motifs recognized by DNA-binding transcription factors (TFs), which recruit various coactivators that act together to engage RNA Polymerase II (RNAPII) resulting in transcriptional activity (Ptashne and Gann, 1997; Stampfel et al., 2015). Eukaryotic TFs typically recognize short DNA motifs of the order of 6-12 base pairs (Weirauch et al., 2014). There are many such similar affinity motifs in the genome[1] (Lambert et al., 2018; Wunderlich and Mirny, 2009). As a result, active enhancer regions represent only a small fraction of putative binding sites for any given TF (Levo and Segal, 2014; Slattery et al., 2014; Spitz and Furlong, 2012; Wunderlich and Mirny, 2009).

Determining whether a DNA motif participates in formation of an active enhancer element is thought to require defining a specific set of molecules and the mechanisms by which they act cooperatively to assemble the transcriptional machinery. Because this choice is made from

[1] In fact, (Wunderlich and Mirny, 2009) study this topic in some detail. Other than bacteria, most higher organisms contain TFs that lack the information content required to uniquely designate binding sites. One of the more high-information TFs in mammals is CTCF – which binds nearly all of the predicted binding sites and is thought to play a key role in maintaining certain genomic topologies. (Dekker and Mirny, 2016; Hnisz et al., 2016a)

a large set of possibilities, predicting enhancer elements is a significant challenge that has been referred to as the "futility theorem" (Wasserman and Sandelin, 2004).

Previous studies into the rules that govern enhancer formation[2] have focused on cooperativity between TFs, mediated through direct protein-protein interactions or indirectly through changes in chromatin accessibility, nucleosome occupancy, local changes in DNA shape upon binding, and motif organization (Jolma et al., 2015; Lambert et al., 2018; Levo and Segal, 2014; Long et al., 2016; Maniatis et al., 1998; Morgunova and Taipale, 2017; Spitz and Furlong, 2012). The presence of clusters of TF binding sites at a genomic locus has been found to be predictive of enhancer elements (Berman et al., 2002; Markstein et al., 2002; Rajewsky et al., 2002). Clusters of TF binding sites can also occur without producing enhancer activity, and enhancer function can be realized upon small insertions[3] (Mansour et al., 2014). The mechanisms by which TF binding site clusters enable the recruitment and stabilization of the appropriate transcriptional machinery at such loci are not well understood.

Recent studies suggest that the cooperative process of phase separation involving an ensemble of multivalent interactions among TFs, coactivators, and RNA Polymerase II can assemble these factors at specific enhancer elements as dynamic clusters, or condensates[4] (Boija et al., 2018; Cho et al., 2018; Chong et al., 2018; Fukaya et al., 2016; Hnisz et al., 2017; Sabari et al., 2018; Tsai et al., 2017). While transcriptional condensates have been observed at specific genomic loci and features of proteins with intrinsically disordered regions (IDRs) have been implicated in their formation (Boija et al., 2018; Cho et al., 2018; Sabari et al., 2018), the features encoded in the DNA elements that facilitate this process have not been explored. We reasoned that if transcriptional condensate formation contributes to assembling certain active enhancers, investigating how features encoded in the DNA element regulate this process should shed light on the cooperative mechanisms that enable the recruitment of the transcriptional machinery, and provide insights into how enhancer regions in the genome are defined[5].

Using a combination of computational modeling and *in vitro* reconstitutions, we first demonstrate that DNA elements with specific types of TF binding site valence, density, and specificity drive condensation of

[2] A phenomenal recent study combines genome editing, multiplexed imaging, and a simple learning model to predict various correlates of enhancer activity (Fulco et al., 2019)

[3] In some cases, small mutations can cause aberrant enhancer function by changing genome topology (Delaneau et al., 2019; Hnisz et al., 2016b). This area is a challenging and exciting area for future research. Reference in previous note also sheds light on this topic.

[4] See (Sabari et al., 2020) for a recent and up-to-date review

[5] Recently, studies have begun to report and observe the formation of multiple enhancer-assemblies that are dynamic, cooperative, and dependent on transient interactions (Chong et al., 2018; Mir et al., 2018). Variations of the framework here, which focuses on multivalent weak protein-protein interactions, and strong protein-DNA interactions, will likely shed further light on these observations.

TFs and coactivators. We show that modulating the affinities, number, or density of TF-DNA interactions and strength of IDR-IDR interactions impacts condensate formation[6]. Because of the cooperative nature of phase separation, condensates form above sharply defined values of these quantities. We then show that the DNA sequence features that promote condensation *in vitro*[7] also promote enhancer activity in cell-based reporter assays. Genome-wide bioinformatic analyses show that these features also characterize known enhancer regions. Importantly, we show that condensation localized to a specific genetic locus requires a combination of both weak multivalent IDR-mediated interactions and structured TF-DNA interactions. Our results also suggest that transcriptional condensate formation may contribute to long-range genomic interactions and organization, potentially promoting compartmentalization of actively transcribed regions.

Together, these results suggest that specific features encoded in DNA elements and the universal cooperative mechanism of phase separation contribute to localization of the transcriptional machinery at enhancers (especially, super-enhancers), and subsequent enhancer activity. Our studies provide a framework to understand how the genome can scaffold condensates at specific loci and how these condensates might be regulated.

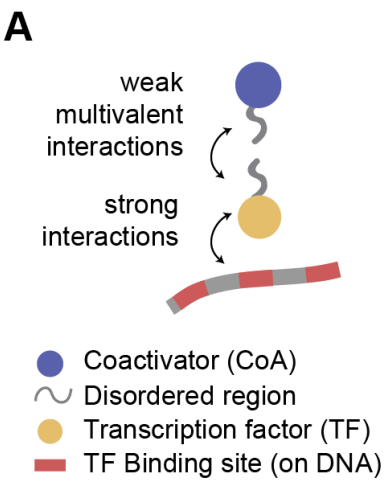
Results

Development of a computational model

To explore how the complex interactions among regulatory DNA elements, TFs, and coactivators impact formation of transcriptional condensates, we first developed a simplified computational model (Figure 1A, Figure S1A). Since enhancers are typically short regions of DNA that are bound by multiple TFs (Levo and Segal, 2014; Spitz and Furlong, 2012), we modeled regulatory DNA elements as a polymer with varying numbers of TF binding sites. Each TF binding site mimics a short (6-12bp) DNA sequence. Specific recognition of DNA motifs by TFs (Weirauch et al., 2014) is mediated by typical TF-DNA binding strengths corresponding to nanomolar dissociation equilibrium constants (Jung et al., 2018), which is the range of TF-DNA interaction en-

[6] Genome-wide bioinformatics in Figure 7 support this and more recent pre-prints have provided further support of the idea of non-linear thresholds based on local clustering/density of TF binding site features (Singh et al., 2020)

[7] Recently, (Gibson et al., 2019) studies the role of nucelosomal and epigenetic features of chromatin in phase separation. An interesting area of future research will be to tease out all the contributions that arise from TFs, histones, DNA and other species in this process.



1A. Schematic depiction of the stochastic computational model and key interactions between molecules. The model consists of a DNA polymer with variable number of TF binding sites, TFs, and coactivators. TFs bind TF binding sites with strong monovalent interactions, and TFs and coactivators interact via weak multivalent interactions between their flexible chains, which mimic the disordered regions of these proteins.

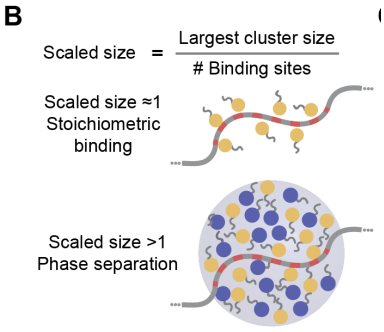
ergies that we have studied in our simulations (Methods). TFs and coactivators contain large IDRs that interact with each other (Boija et al., 2018; Sabari et al., 2018). Thus, we modeled IDRs of TFs and coactivators as flexible chains attached to their respective structured domains. The IDRs interact with each other via multiple low-affinity interactions[8]. The range of IDR-IDR interaction energies that we have studied in our simulations (Methods) corresponds to those that have been determined by in vitro studies of such systems (Brady et al., 2017; Nott et al., 2015; Wei et al., 2017). Our computational studies were focused on obtaining qualitative mechanistic insights that could then be tested by focused experiments.

We simulated this model using standard Langevin molecular dynamics methods to calculate spatiotemporal trajectories of the participating species (see methods, (Anderson et al., 2008)). To distinguish stoichiometrically bound complexes from larger assemblies of transcriptional molecules, we computed the size of the largest molecular cluster scaled by the number of TF binding sites present on DNA. Values of this scaled size greater than 1 represent super-stoichiometric assemblies, while values close to 1 correspond to stoichiometrically bound TFs (Figure 1B). The scaled size is a direct measure of recruitment of transcription machinery and captures finite-size effects, an important factor in characterizing transcriptional condensates, which have been shown to contain ~ 100s-1000s of molecules (Cho et al., 2018). To study whether the super-stoichiometric assemblies are phase separated condensates, we calculated fluctuations in the scaled size spectra when appropriate (see Methods). A characteristic signature of a phase transition is that the fluctuation spectrum exhibits a peak across the threshold value of the titrated parameter. Using the scaled size and its fluctuation spectrum as measures of transcriptional condensate formation, we studied how particular motif compositions on DNA, as well as TF-DNA interactions and interactions between TF and coactivator IDRs regulate transcriptional condensate formation at DNA loci.

Interactions between TFs and multivalent DNA drive formation of condensates of TFs and coactivators

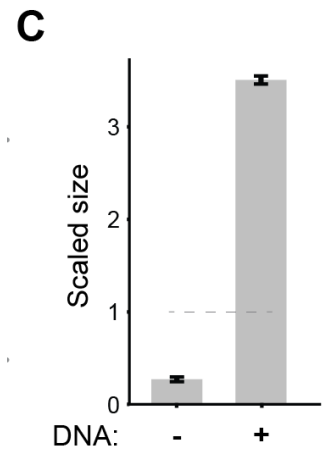
TFs and coactivators form condensates *in vitro* at supra-physiological concentrations (Boija et al., 2018; Lu et al., 2018; Sabari et al., 2018)[9].

[8] The molecular bases of specificity in interactions between IDRs of transcriptional activators and coactivators largely remains unknown. Some recent studies on activation domains have begun to shed light on this (Brodsky et al., 2020; Erijman et al., 2019).



1B. Scaled size is calculated from simulation trajectories, defined as the size of the largest cluster normalized by the number of DNA binding sites. This value is used as a proxy to differentiate stoichiometric binding of TFs to DNA (scaled size ≈ 1, top illustration) from phase-separated super-stoichiometric assemblies (scaled size >1, bottom illustration). For all reported simulation results, reported quantities are averaged over 10 replicate trajectories.

[9] In part, this could arise from the lack of appropriate conditions in the test-tube. While many studies, including this one, include crowding agents to mimic the cell's interiors, the exact quantification of this "crowding" mediated effects remains to be better studies. Preliminary explorations on Xenopus eggs seem to indicate the most *in vitro* studies over-estimate the role of crowders. In systems that capture cellular crowding better i.e. less crowding agents added, it is likely that TFs and coactivators are likely to have even higher threshold concentrations *in vitro*.



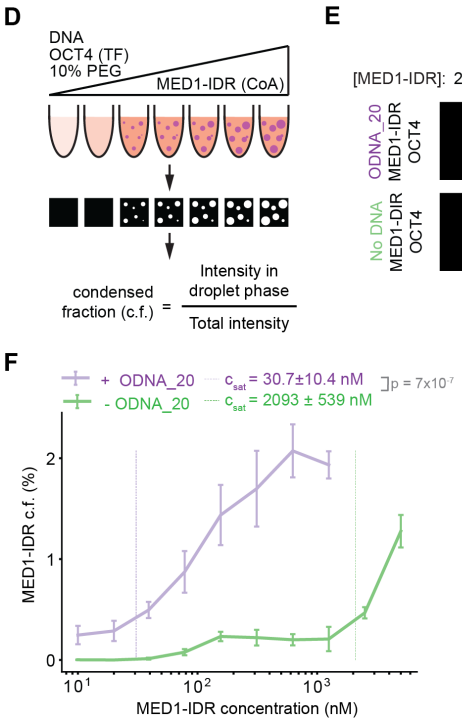
1C. Simulations predict that multivalent DNA-TF interactions result in phase separation of TF and coactivator at dilute concentrations, as shown by scaled size >1 upon addition of DNA.

1D. Schematic depiction of experimental workflow and image analysis for *in vitro* droplet assay. DNA, OCT4, and varying concentrations of MED1-IDR are incubated together in the presence of 10% PEG-8000 as a molecular crowder (illustrated with test tubes, see methods for detail). Fluorescence microscopy of these mixtures is used to detect droplet formation (illustrated by black square with or without white droplets). Multiple images per condition are then analyzed to calculate condensed fraction (c.f.) as intensity of fluorescence signal within droplets divided by total intensity in the image.

1E. Representative images of MED1-IDR droplets in the presence of OCT4 and ODNA_20 (top row) or with only OCT4 (bottom row) at indicated MED1-IDR concentrations. See Table S2 for sequence of DNAs used in droplet assays.

1F. Condensed fraction of MED1-IDR (in units of percentage) with DNA (purple) or without DNA (green) across a range of MED1-IDR concentrations (log scale). The respective inferred C_{sat} values are shown in dashed lines and p-values are estimated from a two-sided Welch's t-test. Higher condensed fraction implies higher fraction of total signal in droplet phase. Solid lines represent mean and error bars represent boundaries of mean±std from replicates.

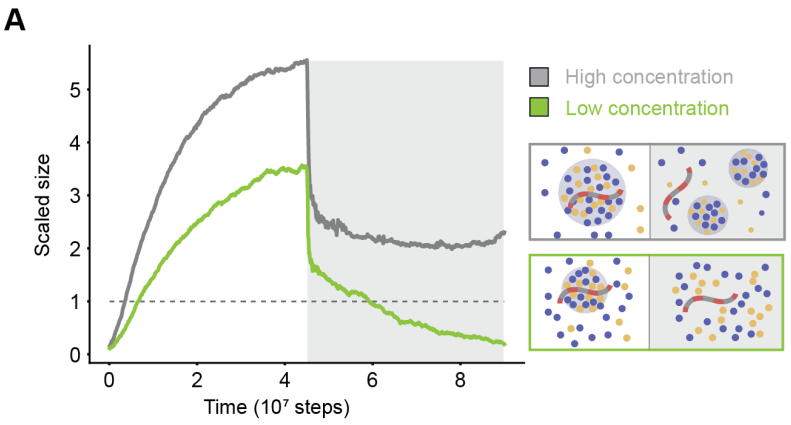
Our simulation results (Figure 1C) predict that a dilute solution of TFs and coactivators that does not phase separate by itself forms condensates (scaled size greater than 1) upon adding multivalent DNA (DNA with 30 TF binding sites). To test this prediction, we developed an *in vitro* phase separation droplet assay containing the three components present in our simulations: TF, coactivator, and DNA (Figure 1D). For TF and coactivator, we employed purified OCT4, a master transcription factor in murine embryonic stem cells (mESCs), and MED1-IDR, the intrinsically disordered region of the largest subunit of the Mediator coactivator complex. We have previously shown that these proteins phase separate together *in vitro* and *in vivo* (Boija et al., 2018; Sabari et al., 2018). For DNA, we used various synthetic DNA sequences containing varying numbers of OCT4 binding sites (see methods and Table S2). Each of the three components was fluorescently labeled either by fluorescent protein fusion, mEGFP-OCT4 and mCherry-MED1-IDR, or a fluorescent dye, Cy5-DNA.



Formation of phase-separated droplets was monitored over a range of MED1-IDR concentrations by fluorescence microscopy with a fixed concentration of OCT4 in the presence or absence of multivalent DNA (DNA with 20 OCT4 binding sites, 8bp motif with 8bp spacers, OD-

NA_20, see methods and Table S2). The fluorescence microscopy results were quantified by calculating the condensed fraction as a function of MED1-IDR concentration (Figure 1D, also see methods). From the condensed fraction, a saturation concentration (C_{sat}) is inferred (see methods under Image analysis and Statistical Analyses) to estimate the phase separation threshold under the specified experimental condition. Experimental variables with lower values of the inferred C_{sat} promote phase separation at lower MED1-IDR concentrations than ones with higher C_{sat}.

Consistent with model predictions, addition of DNA promoted phase separation at low MED1-IDR concentrations (Figure 1E). Addition of DNA lowered the inferred C_{sat} by ~68 fold from ~2100 nM to 30 nM (Figure 1F). These results demonstrate that multivalent DNA promotes the phase separation of TFs and coactivators at low protein concentrations, comparable to concentrations observed *in vivo* (Figure S1B).

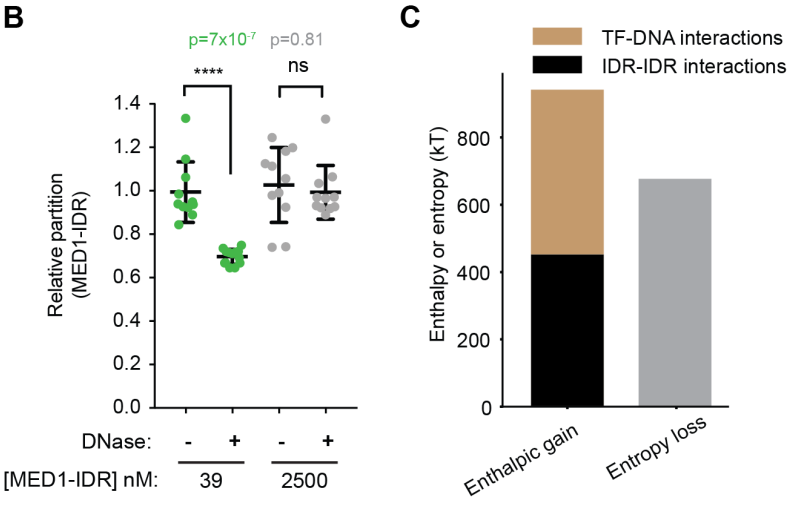


2A. Simulation results for dynamics of condensate assembly/disassembly at two different protein concentrations is represented by average scaled size on the ordinate, and time (in simulation steps after initialization) on the abscissa. TF-DNA interactions are disrupted after stable condensate assembly (shown by a dark gray background). Schematic depiction of phase behavior is shown enclosed in boxes whose colors match the respective lines. See Movies S1 and S1

To further study how DNA influences condensate stability, we performed simulations where TF and coactivator condensates were allowed to form in the presence of DNA, followed by a simulated disruption of TF-DNA interactions (grey box in Figure 2A). At dilute protein concentrations, disrupting TF-DNA interactions resulted in dissolution of condensates (Figure 2A green line, Movie S1), demonstrating that, under these conditions, DNA is required for both formation and stability of condensates. Computing the radial density function around DNA (see methods) confirmed that TFs and coactivators form a largely uniform dense phase dependent on TF-DNA interactions (Figure S1C). While addition of DNA at high protein concentrations increased the rate of

condensate assembly (Figure S1D; Movie S2), by reducing the nucleation barrier, disruption of TF-DNA interactions at these high concentrations did not lead to condensate dissolution (Figure 2A; grey line). We observed a drop in scaled size upon TF-DNA interaction disruption in this case, but this was primarily due to the condensate being broken into smaller droplets as the DNA was ejected from the condensate (as depicted in Figure 2A; grey box, Movie S2). Together, these results predict that, at dilute protein concentrations, specific TF-DNA interactions are required or both formation and stability of condensates at particular genomic loci.

To mimic disruption of TF-DNA interactions *in vitro*, we added DNase I to droplets formed at high or low concentrations of MED1-IDR in presence of OCT4 and ODNA_20 (see methods). As expected, DNA was significantly degraded in both conditions (Figure S1E). Consistent with our model predictions, droplets formed at the lower concentrations were more sensitive to the degradation of DNA than those formed at higher concentrations (Figure 2B). While enzymatic degradation of DNA did not completely dissolve droplets in our *in vitro* experiments, MED1-IDR enrichment within droplets was significantly diminished only at the lower protein concentration (Figure 2B). Together, the *in silico* and *in vitro* results indicate that DNA can nucleate and scaffold phase-separated condensates of TFs and coactivators at low protein concentrations.



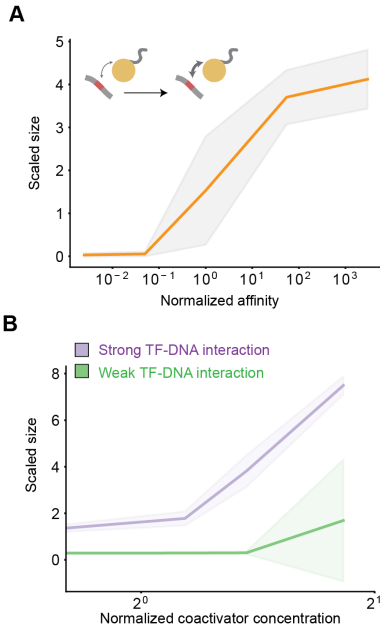
2B. Scatter plot depiction of experimentally determined MED1-IDR partition ratio (see methods) between condensate and background, at high (2500 nM, gray) and low concentrations (39nM, green) of MED1-IDR in the presence of OCT4 and ODNA_20, in the absence (-) or in the presence (+) of DNase I. The partition ratio is normalized to the (-) condition, and lower partition ratios imply lesser enrichment of MED1 in the droplet phase. Individual data points are presented with mean ± std, p-values represent Student's t-test.

2C. Energetic attractions, arising from a combination of TF-DNA (brown) and IDR (black) interactions, compensate for entropy loss (grey) of forming a condensate.

Physical mechanisms that underlie localized formation of transcriptional condensates

To understand the mechanisms driving DNA-mediated condensate formation, we cast our results in terms of the competing thermodynamic forces that govern phase separation. For computational efficiency, further characterization of our model was carried out with a simplified implicit-IDR model (Figure S2A), which recapitulated all features (Figure S2B-C) of the explicit-IDR model. Typically, condensate formation results in entropy loss because the molecules in the droplet are more confined than if they were in free solution. A condensate is stable only if this entropy loss is compensated by the energetic gain from enhanced attractive interaction energies between molecules confined in the condensate. We computed the energetic gain by summing up all pairwise molecular interactions in the condensate. Entropy loss due to confinement was calculated by adding a factor of $kT \ln\left(V_{droplet} / V_{system}\right)$ for each molecule in the condensate. This loss in free volume is the principal source of entropy loss in our coarse-grained model. Other sources of entropy loss like solvent/ion effects are effectively incorporated in our affinity parameters. Our simulations show that energetic gains arising from a sum of specific TF-DNA interactions and weak IDR interactions (TFs-coactivator interactions) are necessary to compensate for the entropy loss of forming condensates at low concentrations (Figure 2C). IDR interactions alone are insufficient to compensate for the entropy loss of condensate formation, thus disruption of TF-DNA interactions results in condensate dissolution (Figures 2 and S2B-S2D, dark grey background). Likewise, TF-DNA interactions alone are insufficient to compensate for the entropy loss of condensate formation and disruption of IDR-IDR interactions results in condensate dissolution (see next section). The same features are observed in explicit-IDR simulations (Figures 2A; orange line, and S2E), though our simplified calculation of the entropy loss in this case (see above) is an underestimate, as contributions from the change in configurational entropy of IDR chains is not accounted. These results provide a mechanistic framework to understand how the combination of TF-DNA interactions and weak IDR interactions determine assembly and stability of transcription condensates at low concentration[10].

[10] In a more coarse-grained setting such as the Flory-huggins formalism. This would reflect a parameterization of χ parameters such that heterotypic interactions b/w protein-DNA and homotypic protein-protein interactions work in concert to drive phase separation.



3A. Simulations predict a shift in scaled size from stoichiometric binding (≈1) to phase separation (>1) with increasing normalized affinity (darker arrow in schematic); affinity normalized to threshold affinity of E=12kT.

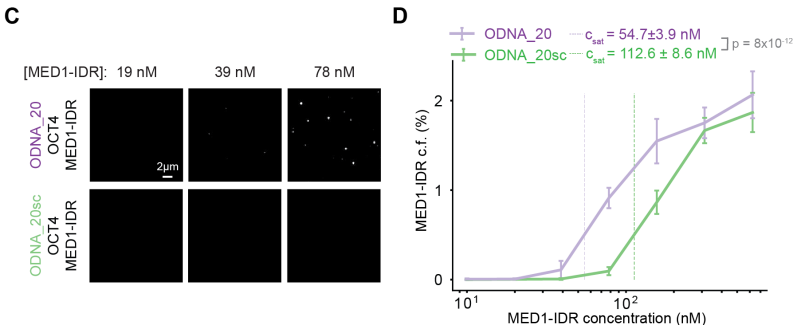
3B. Scaled size predictions for high (normalized affinity ≈50, purple) and low (normalized affinity ≈5e-2, green) TF-DNA affinities as a function of coactivator concentration. Coactivator concentrations are normalized to value of N_{coa}=150

3C. Representative images of MED1-IDR droplets with OCT4 and ODNA_20 (top row) or ODNA_20scramble (sc) (bottom row) at indicated MED1-IDR concentrations. See Table S2 for sequence of DNAs used in droplet assays.

3D. Condensed fraction of MED1-IDR (in units of percentage) for ODNA_20 (purple) and ODNA_20sc (green) across a range of MED1-IDR concentrations (log scale). The respective inferred C_{sat} values are shown in dashed lines and p-values are estimated from a two-sided Welch's t-test. Higher condensed fraction implies higher fraction of total signal in droplet phase.

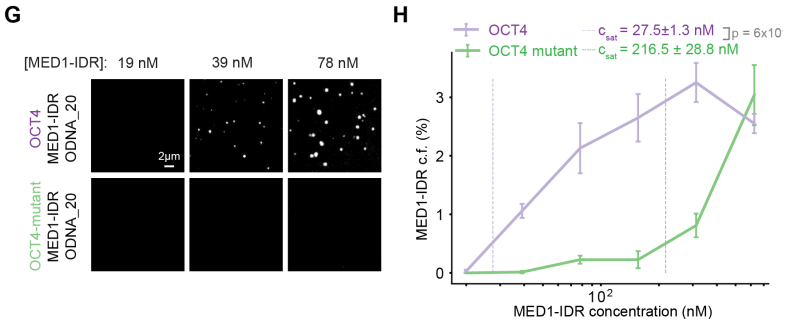
Specific TF-DNA interactions and weak multivalent IDR interactions regulate formation of transcriptional condensates

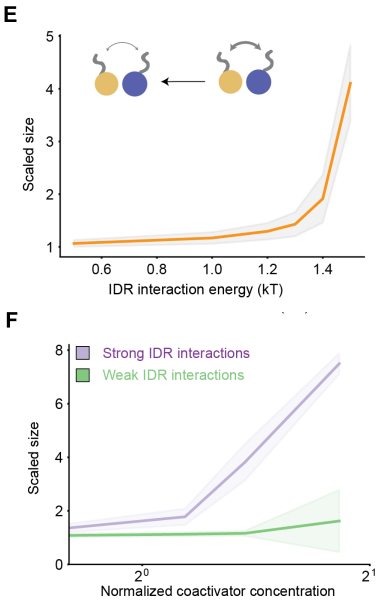
Given that TF-DNA interactions are necessary for condensate formation, we next investigated the effect of modulating TF-DNA affinity at dilute protein concentrations. Simulations predict that condensates form above a sharply defined affinity threshold (Figure 3A) and that high affinity TF-DNA interactions result in condensate formation at low coactivator concentration thresholds (Figure 3B). The normalized fluctuation spectrum of the scaled size (see methods for details) showed a peak across the threshold affinity value, characteristic of phase separation (Figure S3A-B). Using the *in vitro* droplet assay, we probed the effect of TF-DNA interactions by comparing phase separation of MED1-IDR over a range of concentrations, with fixed concentrations of both OCT4 and either ODNA_20 or a scrambled ODNA_20 which does not contain any consensus binding sites for OCT4 (ODNA_20sc, Table S2). High-affinity OCT4-ODNA_20 interactions promoted phase separation at lower MED1-IDR concentrations when compared to OCT4-ODNA_20sc (Figure 3C). Quantifying the MED1-IDR condensed fraction further corroborated our finding, showing a ~2-fold decrease in inferred saturation concentrations in presence of higher affinity OCT4-ODNA_20 interactions (Figure 3D). Similar results were obtained by quantifying the condensed fraction of OCT4 or DNA (Figure S4A-B). These results demonstrate that higher TF-DNA affinities promote phase separation above sharply defined thresholds. Therefore, TFs, which exhibit higher affinity for specific DNA binding sites compared to random DNA, can drive transcriptional condensate formation at specific DNA loci.



We next investigated the effect of modulating the affinities of multivalent IDR interactions, whose effective affinity can be regulated *in vivo* through post-translational modifications (Banani et al., 2017; Shin and Brangwynne, 2017). Reducing the strength of IDR interactions between TFs and coactivators in our simulations predicts that condensates dissolve below a sharply defined interaction threshold (Figure 3E), and strong IDR interactions result in condensate formation at lower coactivator concentration thresholds (Figure 3F). To test this prediction, we monitored MED1-IDR phase separation over a range of MED1-IDR concentrations with fixed concentrations of both ODNA_20 and either OCT4 or a previously characterized OCT4 activation-domain mutant (acidic to alanine mutant) with reduced interaction with MED1-IDR (Boija et al., 2018). Consistent with simulation predictions, the OCT4 mutant was much less effective at promoting phase separation at low concentrations, with a nearly 8-fold higher inferred C_{sat} as compared to OCT4 (Figures 3G-H). These results further highlight the importance of weak multivalent interactions between coactivators and TFs in the formation of transcriptional condensates.

Our results thus far suggest the following model. Specific TF-DNA interactions localize TFs to particular genomic loci. Transcriptional condensate formation is a cooperative process, which occurs at these loci when the weak multivalent interactions between TFs and coactivators exceed a threshold. While other processes may also be involved (e.g. DNA bending, removal and modification of nucleosomes, and interactions with RNA), this cooperative phenomenon of condensate formation by TF and coactivator phase separation contributes to assembling the transcriptional machinery at enhancers.



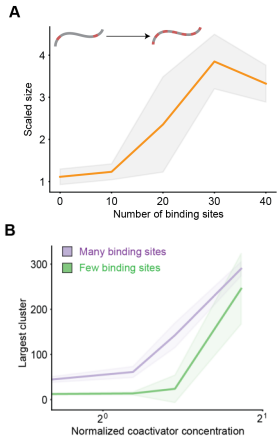


3E. Simulations predict a shift in scaled size from phase separation (>1) to stoichiometric binding (≈1) upon decreasing IDR interaction (from right to left, lighter arrow in schematic).

3F. Scaled size predictions for high (IDR = 1.5kT, purple) and low (IDR=1.0 KT, green) IDR interaction as a function of coactivator concentration (normalized as in 3B).

3G. Representative images of MED1-IDR droplets with ODNA_20 and OCT4 (top row) or an OCT4-mutant with reduced affinity for MED1-IDR (bottom row) at indicated MED1-IDR concentrations.

3H. Condensed fraction of MED1-IDR (in units of percentage) for OCT4 (purple) and OCT4-mutant (green) across a range of MED1-IDR concentrations (log scale). The respective inferred C_{sat} values are shown in dashed lines and p-values are estimated from a two-sided Welch's t-test. In all condensed fraction plots, solid lines represent mean and error bars represents



4A. Simulations predict a shift in scaled size from stoichiometric binding (≈1) to phase separation (>1) with increasing number of TF binding sites on DNA (schematic depicts increasing number of binding sites).

4B. Scaled size predictions for many (N =30, purple) and few (N=10, green) binding sites as a function of coactivator concentration (normalized as in 3B).

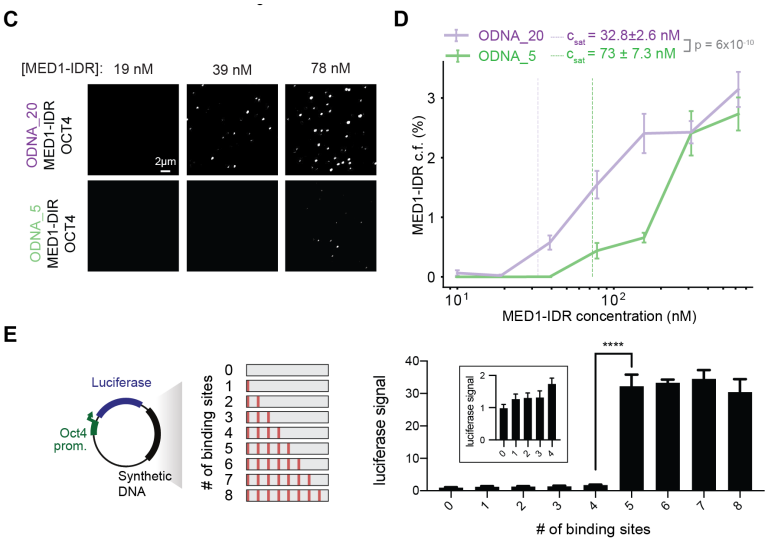
4C. Representative images of MED1-IDR droplets with OCT4 and ODNA_20 (top row) or ODNA_5 (bottom row) at indicated MED1-IDR concentrations. See Table S2 for sequence of DNAs used in assays

4D. Condensed fraction of MED1-IDR (in units of percentage) for ODNA_20 (purple) and ODNA_5 (green) across a range of MED1-IDR concentrations (log scale). The respective inferred C_{sat} values are shown in dashed lines and p-values are estimated from a two-sided Welch's t-test.

4E. Enhancer activity increases over a sharply defined TF binding site threshold. The left panel shows a schematic depiction of the luciferase reporter construct and the synthetic DNA sequences tested. The right panel shows the luciferase signal from constructs with the indicated number of binding sites transfected into murine embryonic stem cells (see methods). Inset presents data for 0-4 binding sites graphed on a different scale for the ordinate. Luciferase signal is normalized to the construct with 0 motifs. Data is graphed as average of three biological replicates ± std. **** = Student's t-test p-value < 0.0001.

Specific motif compositions encoded in DNA facilitate localized transcriptional condensate formation

To begin defining the specific DNA sequence features that result in condensate formation, we explored the effects of modulating the valence and density of TF binding sites with the same TF-DNA affinities. We reasoned that the same energetic compensation for entropy loss we observed by increasing TF-DNA affinities (Figures 2C and 3A-D) could be obtained instead through increasing the number of DNA binding sites (i.e. valence). Our simulations predict that, for the same TF-DNA binding energies, condensates form above a sharply defined valence threshold (Figure 4A, Figures S3C-D), and higher valence results in condensate formation at lower coactivator concentrations (Figure 4B). Consistent with this prediction, *in vitro* assays revealed that ODNA_20 promoted phase separation of MED1-IDR and OCT4 at lower concentrations, with an inferred C_{sat} ~2-fold lower than the threshold for DNA with fewer binding sites (ODNA_5) (Table S2) (Figures 4C-4D).



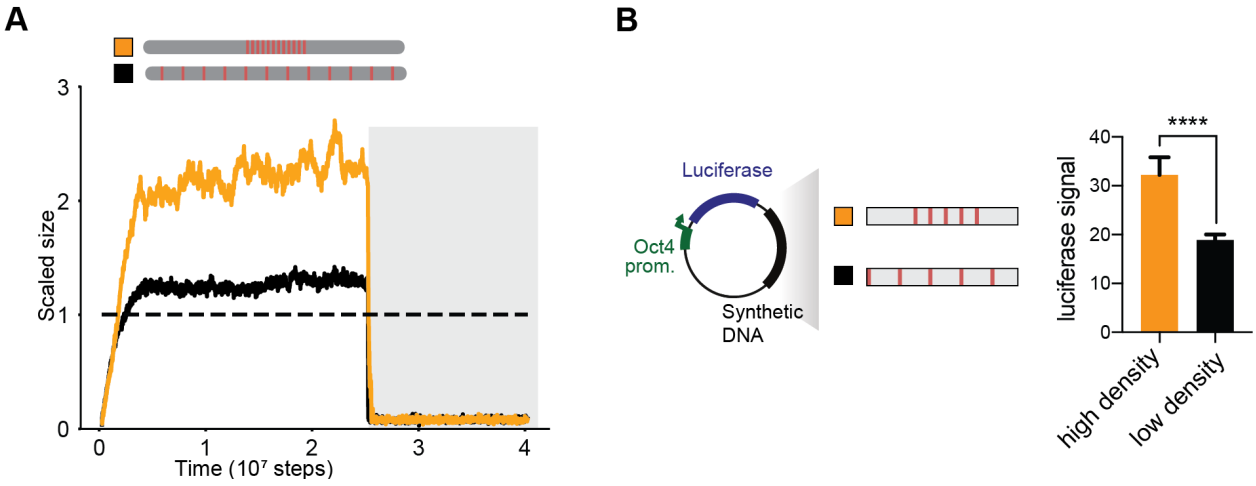
To test how motif valence impacts enhancer activity in cells, we cloned synthetic DNA sequences with varying number of OCT4 binding sites into previously characterized luciferase reporter constructs (Whyte et al., 2013) that were subsequently transfected into mESCs (see methods and figure 4E schematic). In these reporter assays, expression of the luciferase gene, read out as luminescence, is a measure of the strength of enhancer activity. Our computational studies and *in vitro* results show (Fig 4D) that for any concentration of MED1 less than C_{sat} of ODNA_5,

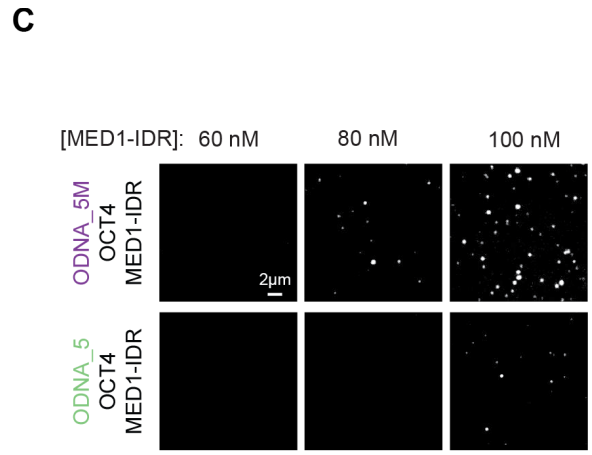
but higher than C_{sat} of ODNA_20, only DNA with valence greater than a threshold can drive condensate formation. Since cellular protein levels are tightly regulated, these results predict that condensate assembly, and thus enhancer function, will be a digital function above a threshold valence of binding sites. Using a series of DNAs with 0 to 8 binding sites (8bp motif with 24bp spacers, see Table S3) we found that enhancer activity increased above a sharply defined valence threshold (Figure 4E), in striking qualitative agreement with expectations from our computational and *in vitro* studies.

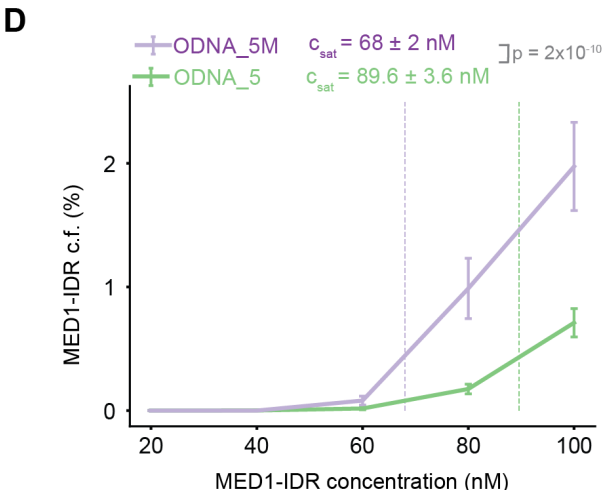
To distinguish whether this behavior stemmed from motif valence alone or local motif density, we carried out simulations of DNA chains with a fixed number of binding sites, but different distributions along the chain (Figure 5A). We found that high local density, as compared to the same number of binding sites at lower density, promoted condensate formation at low protein concentration (Figure 5A). In vitro experiments were carried out with DNA containing the same binding site number (5 binding sites), but different densities (DNA_5M with higher density than DNA_5, see methods and Table S2). Quantifying the microscopy data validated simulation predictions, evidenced by a ~30% increase in inferred Csat for DNA_5 over DNA_5M (Figure 5C-D). To test the effect of binding site density on enhancer activity in cells, we compared the enhancer activity of 5 binding sites with different densities (see Table S3) in luciferase assays in mESCs (Figure 5B). In agreement with the model predictions, reducing density of binding sites led to reduced enhancer activity.

5A. Scaled size versus simulation time steps comparing two different distribution of binding site densities (shown in the schematic below), but same overall number of binding sites. TF-DNA interactions are disrupted after stable condensate assembly (dark gray background).

5B. DNA sequences with the same number of binding sites, but higher density, shows increase in transcription activity. Left half shows a schematic depiction of the luciferase reporter construct and the synthetic DNA sequences tested. The right half shows the luciferase signal from constructs with indicated binding site density transfected into mouse embryonic stem cells. Data graphed as in E. **** = Student's t-test p-value < 0.0001.







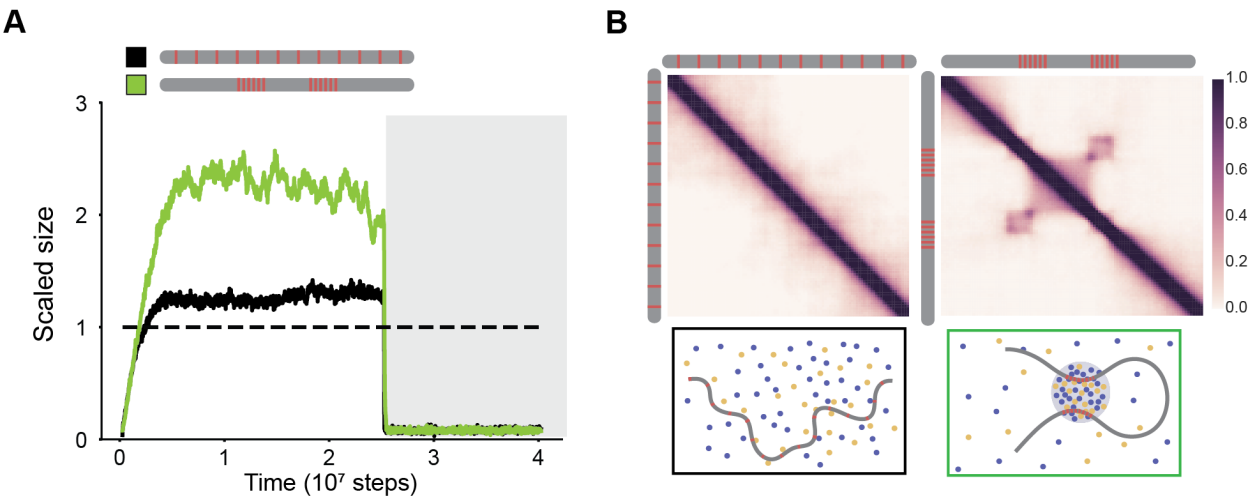
5C. Representative images of MED1-IDR droplets with OCT4 and high motif density (ODNA_5M) (top row) or low motif density (ODNA_5) (bottom row) at indicated MED1-IDR concentrations. See Table S2 for sequence of DNAs used in droplet assays.

5D. Condensed fraction of MED1-IDR (in units of percentage) for ODNA_5M (purple) and ODNA_5 (green) across a range of MED1-IDR concentrations. The respective inferred C_{sat} values are shown in dashed lines and p-values are estimated from a two-sided Welch's t-test. Solid lines represent mean and error bars represent boundaries of mean±std from replicates. See methods for details on calculation of condensed fraction and inferred C_{sat} .

The results in Figures 4 and 5 show that dense clusters of a particular TF's binding sites, with the valence of binding sites exceeding a sharply defined threshold, drive localized formation of transcriptional condensates, and that these same features influence enhancer activity in cells. The condensates form by the universal cooperative mechanism of phase separation which, in turn, requires weak cooperative interactions between the IDRs of TFs and coactivators (Figure 3). IDR-IDR interactions are relatively non-specific, and the same coactivator IDRs can assemble the transcriptional machinery in stable condensates at different enhancers upon cognate TF binding.

Transcriptional condensate formation may facilitate long-range interactions and higher-order genome organization

Given that regulatory elements often communicate over long linear distances, we next investigated whether two dense clusters of TF binding sites in DNA separated by a linker could assemble a single condensate. Our simulations show that this is indeed the case (Figure 6A; green line). Contact frequency maps, computed from the simulation data (see methods) show long-range interactions between the dense clusters of binding sites, which are absent (Figure 6B) at conditions with a low density of TF binding sites distributed uniformly (Figure 6A; black line). Further, removing a single cluster strongly diminished the ability of DNA to assemble a condensate (Figure S5), suggesting that both clusters of binding



sites worked cooperatively over intervening linker DNA to assemble a condensate. These results suggest that condensate formation could explain recent observations of CTCF/cohesin-independent long-range interactions between active regions of the genome (Rowley et al., 2017; Schwarzer et al., 2017). More generally, our results suggest that localized transcriptional condensate formation can facilitate higher-order organization of the 3D-genome and contribute to long-range communication between enhancer-promoter pairs.

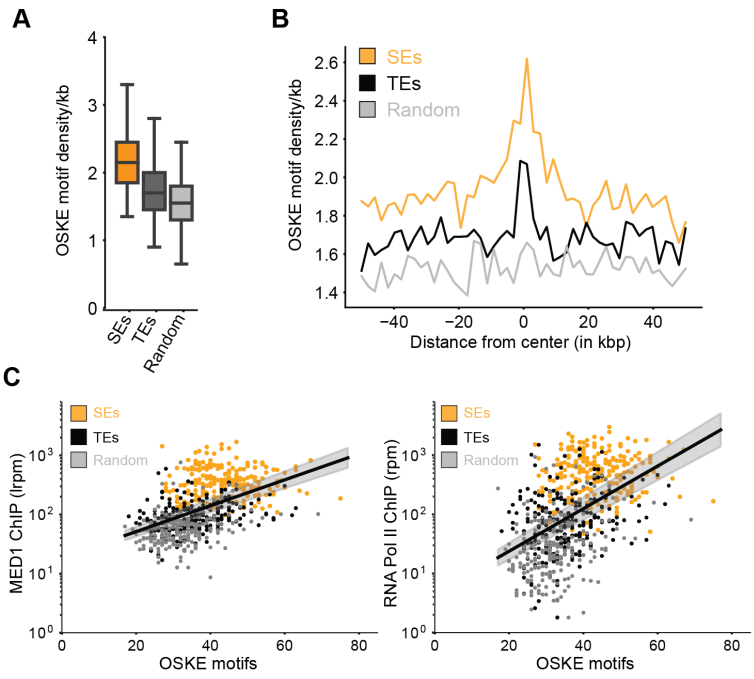
6A. Scaled size versus simulation time steps comparing two different distribution of binding site densities (as shown in the schematic legend). TF-DNA interactions are disrupted after stable condensate assembly (as shown in dark gray background).

6B. Contact frequency maps (see methods) show long-range interactions (right panel, checkerboard-like patterns) for high local motif density (computed for green line in Figure 6A), and not for low motif density (left panel, computed for black line in Figure 6A).

Mammalian genomes encode specific motif features in enhancers to assemble high densities of transcription apparatus

We next investigated whether enhancer features that our results suggest promote transcriptional condensate formation are present in mammalian genomes. Given that our results show that a linear increase in TF binding site valence can result in an exponential increase in coactivator recruitment by condensate formation (Figure 4), we investigated the relationship between TF binding site valence (i.e. occurrence of TF motifs) and coactivator recruitment in mESCs. We gathered genome-wide distribution of TF motif occurrence for highly expressed mESC master TFs – OCT4, SOX2, KLF4, ESRRB (OSKE). Super-enhancers, genomic regions with unusually high densities of transcriptional molecules (Whyte et al., 2013), where transcriptional condensates have recently been observed (Boija et al., 2018; Cho et al., 2018; Sabari et al., 2018), have higher OSKE motif densities when compared to typical enhancers or random loci (Figures 7A-B, methods). Consistent with our results, we

found a highly non-linear (roughly exponential) correlation between OSKE motif density and ChIP-Seq data for MED1, RNA Pol II (Figure 7C), and BRD4 (Figure S6A) across genetic regions including SEs, TEs, and random loci. This correlation was minimal when input control data was analyzed (Figure S6B). These results suggest that enhancer elements that encode specific DNA sequence features we have described can recruit unusually high densities of transcriptional apparatus by transcriptional condensate formation, consistent with our results. The same features enable recruitment of varied cofactors – BRD4, MED1, and Pol II, thus suggesting that phase separation contributes to stabilization of transcription machinery at specific genomic loci.



7A. Box-plot depiction of motif density (per kb) of master mESC TFs – OCT4 + SOX2 + KLF4 + ESRRB (OSKE), over 20kb regions centered on super-enhancers (SEs, orange), typical enhancers (TEs, black), and random loci (light gray).

7B OSKE motif density over a 100kb window centered at SEs (orange), TEs (black), and random loci (gray).

7C MED1 (left) and RNA Pol II (right) ChIP-Seq counts (ordinate, reads-per-million, log scale) against total OSKE motifs over 20kb regions centered on SEs (orange), TEs (black), and random loci (gray). The black line is a fit inferred between the logarithmic ChIP signal and the linearly graphed motif count across all regions, and so the fit represents a highly non-linear (exponential) correlation. The grey shaded regions represent 95% confidence intervals in the value of the inferred slope. The exponential fit describes a sizable fraction of the observed variance i.e. r=0.5 for both inferred lines.

Discussion

Enhancers are DNA elements that control gene expression by promoting assembly of transcriptional machinery at specific genomic loci. Recent studies have suggested that phase-separated condensates of molecules involved in transcription form at enhancers (Boija et al., 2018; Cho et al., 2018; Chong et al., 2018; Fukaya et al., 2016; Hnisz et al., 2017; Sabari et al., 2018; Tsai et al., 2017), providing a potential mechanism for concentrating transcriptional machinery at specific loci. Here we investigated how features encoded in DNA elements can regulate the formation of transcriptional condensates. Our results identify features of DNA sequences that can enable assembly of the transcriptional machinery at specific genomic loci by the general cooperative mechanism of phase separation.

We first demonstrated that interactions between TFs, co-activators, and multivalent DNA elements can form condensates at protein concentrations that are too low for such a phase transition in the absence of the DNA. We suggest that these results help explain why condensates of coactivators and TFs form at enhancers in cells wherein protein concentrations are much lower than that required for phase separation without DNA in vitro. We also found that at low protein concentrations, DNA elements with multiple TF binding sites serve as scaffolds for the phase separated transcriptional condensates. However, at high protein concentrations, the DNA elements act only as a nucleation seed, and are not necessary for condensate stability. These results suggest an explanation for why co-activator overexpression is often linked to pathological gene expression programs (Bouras et al., 2001; Zhu et al., 1999).

By considering the competing thermodynamic forces of entropy loss and energy gain that control phase separation, we described how a combination of specific TF-DNA interactions and weak cooperative interactions between IDRs of TFs and coactivators are required for transcriptional condensate formation. These parameters must be above sharply defined thresholds for phase separation to occur. The necessary sharp threshold for TF-DNA interactions results in formation of transcriptional condensates at specific genomic loci containing cognate TF binding sites. That there is a threshold affinity and valence between IDRs of

the interacting species for condensate formation implies that molecules with IDRs with complementary characteristics, such as those contained in TFs and coactivators, will be incorporated in transcriptional condensates. Therefore, different TFs with IDRs that are statistically matched with co-activator IDRs can mediate transcriptional condensate formation at different genomic loci via similar weak cooperative interaction[11]. This may be the reason underlying recent observations that TFs with different disordered activation domains can co-localize with MED-1 condensates (Boija et al., 2018). Biomolecular condensates can exhibit diverse material properties and phase behavior as a function of their specific IDR sequences (Banani et al., 2017; Dignon et al., 2018; Shin and Brangwynne, 2017). For example, recent studies focused on electrostatic interactions in IDRs have shown that particular statistical patterns of charged residues dictate overall phase behavior (Das and Pappu, 2013; Huihui et al., 2018; Lin et al., 2017) and enable specific protein interactions (Borgia et al., 2018; Sherry et al., 2017) . Similarly, the "spacer-sticker" framework (Harmon et al., 2017; Wang et al., 2018), which builds on previous mean-field models (Semenov and Rubinstein, 1998), has been successfully used to elucidate the interplay of gelation and phase separation in prion-like proteins (Wang et al., 2018). Leveraging these techniques to characterize IDRs of transcription-associated proteins will provide insights on the molecular grammar underlying their interactions and enable better understanding of the biophysical properties of transcription condensate[12].

Importantly, we find that DNA elements with dense clusters of TF binding sites that exceed a sharply defined valence threshold promote transcriptional condensate formation, and the same findings are mirrored for enhancer activity in cells.

Our results also provide insights on specific combinations of DNA features that facilitate transcription condensate formation. For example, low-affinity TF binding sites can contribute to scaffolding a transcriptional condensate, if present in sufficiently high valence and density, as the total energy gain comes from a combination of these parameters. This may explain recent intriguing descriptions of enhancer regulation through clusters of weak TF binding sites (Crocker et al., 2015; Tsai et al., 2017). In contrast, DNA sequences with many high affinity TF binding sites (high valence) distributed at low local density are not enhancer

[11]Interestingly, there are previous studies that explore how "statistical pattern matching" can mediate specificity – initially studied in context of developing sensors (Srebnik et al., 1996)

[12]Progress has been made on this topic with the development of learning-based approaches to identify relevant features in activation domains (Erijman et al., 2019) – yet a biophysical basis remains incomplete.

regions because they will not enable formation of transcriptional condensates. This may explain why many high-affinity sites that are not enhancers remain largely unbound and contribute to a deeper understanding of the futility theorem (Wasserman and Sandelin, 2004). Thus, our framework elucidates the key parameters, or specific combinations of these parameters, that must be above sharply defined thresholds for phase separated transcriptional condensates to form at specific genomic loci that function as enhancers.

Bioinformatic analyses reveal that the DNA sequence features that we have described as important for transcriptional condensate formation also characterize enhancer regions in mammalian genomes, and increases in the recruitment of transcriptional molecules at different loci are correlated in a highly non-linear way with motif density.

Taken together our results suggest the following model for a general cooperative mechanism that contributes to assembling the transcriptional machinery at enhancers, perhaps especially at super-enhancers. Dense clusters of a particular TF's binding sites, with the number of binding sites exceeding a sharply defined threshold, drive localized formation of transcriptional condensates at a specific genomic locus. The condensate, which recruits and stabilizes various transcriptional molecules, forms by the universal cooperative mechanism of phase separation. Thus, a threshold number of cooperative binding events have to occur at a particular genomic locus, before phase separation occurs to robustly assemble the transcription machinery. Although included only implicitly in our model, past data suggests that TF binding to DNA can be cooperative and sequential (for example, due to DNA bending)(Levo and Segal, 2014; Spitz and Furlong, 2012). Thus, a series of sequential steps occurs when TFs bind to a sufficiently large number of binding sites that serve as enhancers. This is analogous to kinetic proofreading in cell signaling processes(Hopfield, 1974; Ninio, 1975), such as T cell receptor signaling that discriminates between self and cognate ligands to mediate pathogen-specific immune responses. In the latter situation, a sequence of biochemical steps need to occur before productive downstream signaling can lead to activation; only the cognate ligands can complete these steps with high probability. In T cell signaling, once the kinetic proofreading steps are completed, a positive feedback loop amplifies signal levels to result in robust downstream signaling leading to

activation(Das et al., 2009). At enhancers, after TFs have bound to a sufficiently large number of cognate binding sites on DNA, amplification of the recruitment of transcriptional machinery occurs by condensate formation. Intriguingly, the mathematical description of a first order phase transition and a positive feedback loop's effect on signaling are isomorphic[13], suggesting that perhaps biological processes have evolved similar strategies in diverse contexts.

[13] Both, to first-approximation, are cubic polynomials that permit "multiple roots" or multiple steady-states, in the language of dynamical systems.

Condensate formation requires weak cooperative interactions between the IDRs of TFs and coactivators (Figure 3). Although different molecular grammars may describe different types of IDR-IDR interactions, these interactions are relatively non- specific, and the same coactivator IDRs can assemble within condensates at different enhancers. This model is consistent with the observation that clusters of TF binding can often correctly predict active enhancers because this feature of the DNA sequence drives formation of transcriptional condensates by a common mechanism (Berman et al., 2002; Markstein et al., 2002; Rajewsky et al., 2002).

Our model can also describe situations where insertion of a relatively small DNA element that binds to a master TF that regulates cell type specific gene expression programs can stabilize TFs that bind weakly to adjacent binding sites, and recruit the transcriptional machinery in condensates. We carried out simulations with a DNA sequence comprised of two types of binding sites – those that bind strongly to a TF and others that bind weakly. As Figures S3E-F show, a transcriptional condensate forms at such a locus beyond a threshold fraction of high-affinity (master) TF binding sites. This is because the cooperative process of condensate formation recruits and stabilizes the transcriptional machinery once the number of strong TF binding sites exceeds a certain value. This result may explain why a relatively small insertion of a TF binding site into a region that contained an inactive cluster of binding sites for other TFs, resulted in the formation of a super-enhancer in T-ALL cells (Mansour et al., 2014).

[14] It will be interesting to explore these models to structures such as the histone locus body and study their evolutionary trajectory. Early studies by fly biologists have documented many of the same important features we describe above for enhancers (Nizami et al., 2010; Salzler et al., 2013; White et al., 2011)

While our model explicitly incorporates enhancer DNA, TFs, and coactivators, the underlying mechanistic framework can be extended to understand diverse condensates that form at specific genomic loci[14]. Examples may include condensates in heterochromatin-organization (Larson

et al., 2017; Strom et al., 2017), histone locus body assembly (Nizami et al., 2010), lncRNA-mediated paraspeckle formation (Fox et al., 2018; Yamazaki et al., 2018), nucleolar formation (Feric et al., 2016; Pederson, 2011) and in polycomb-mediated transcriptional silencing (Tatavosian et al., 2018). Recent advances in microscopy at the nano-scales (Li et al., 2019) can potentially shed light into whether transcription-associated condensates form higher-order sub-structures, like the nucleolus (Feric et al., 2016).

Our study provides a framework towards understanding how the genome can scaffold condensates at specific loci and implicates particular TF binding site compositions. In addition to TF binding sites, processes that dynamically modulate valence and specificity of interacting species at specific genetic loci, such as local RNA synthesis or chromatin modifications, are likely to play a role in the formation of transcriptional condensates[15].

[15] For an example of such a study, consider the next chapter.

Bibliography

Anderson, J.A., Lorenz, C.D., and Travesset, A. (2008). General purpose molecular dynamics simulations fully implemented on graphics processing units. J. Comput. Phys. *227*, 5342–5359.

Banani, S.F., Lee, H.O., Hyman, A.A., and Rosen, M.K. (2017). Biomolecular condensates: organizers of cellular biochemistry. Nat. Rev. Mol. Cell Biol. *18*, 285–298.

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., and Eisen, M.B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc. Natl. Acad. Sci. U. S. A. *99*, 757–762.

Boija, A., Klein, I.A., Sabari, B.R., Dall'Agnese, A., Coffey, E.L., Zamudio, A. V., Li, C.H., Shrinivas, K., Manteiga, J.C., Hannett, N.M., et al. (2018). Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. Cell *175*, 1842-1855.e16.

Borgia, A., Borgia, M.B., Bugge, K., Kissling, V.M., Heidarsson, P.O., Fernandes, C.B., Sottini, A., Soranno, A., Buholzer, K.J., Nettels, D., et al. (2018). Extreme disorder in an ultrahigh-affinity protein complex. Nature.

Bouras, T., Southey, M.C., and Venter, D.J. (2001). Overexpression of the steroid receptor coactivator AIB1 in breast cancer correlates with the absence of estrogen and progesterone receptors and positivity for p53 and HER2/neu. Cancer Res. *61*, 903–907.

Brady, J.P., Farber, P.J., Sekhar, A., Lin, Y.-H., Huang, R., Bah, A., Nott, T.J., Chan, H.S., Baldwin, A.J., Forman-Kay, J.D., et al. (2017). Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. Proc. Natl. Acad. Sci. U. S. A. *114*, E8194–E8203.

Brodsky, S., Jana, T., Mittelman, K., Chapal, M., Kumar, D.K., Carmi, M., and Barkai, N. (2020). Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding Specificity. Mol. Cell *79*, 459-471.e4.

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

Chong, S., Dugast-Darzacq, C., Liu, Z., Dong, P., Dailey, G.M., Cattoglio, C., Heckert, A., Banala, S., Lavis, L., Darzacq, X., et al. (2018). Imaging dynamic and selective low-complexity domain interactions that control gene transcription. Science *361*.

Crocker, J., Abe, N., Rinaldi, L., McGregor, A.P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., et al. (2015). Low Affinity Binding Site Clusters Confer Hox Specificity and Regulatory Robustness. Cell *160*, 191–203.

Das, R.K., and Pappu, R. V (2013). Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proc. Natl. Acad. Sci. U. S. A. *110*, 13392–13397.

Das, J., Ho, M., Zikherman, J., Govern, C., Yang, M., Weiss, A., Chakraborty, A.K., and Roose, J.P. (2009). Digital Signaling and Hysteresis Characterize Ras Activation in Lymphoid Cells. Cell *136*, 337–351.

Dekker, J., and Mirny, L. (2016). The 3D Genome as Moderator of Chromosomal Communication. Cell *164*, 1110–1121.

Delaneau, O., Zazhytska, M., Borel, C., Giannuzzi, G., Rey, G., Howald, C., Kumar, S., Ongen, H., Popadin, K., Marbach, D., et al. (2019). Chromatin three-dimensional interactions mediate genetic effects on gene expression. Science *364*, eaat8266.

Dignon, G.L., Zheng, W., Best, R.B., Kim, Y.C., and Mittal, J. (2018). Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. Proc. Natl. Acad. Sci. U. S. A. *115*, 9929–9934.

Erijman, A., Kozlowski, L., Sohrabi-Jahromi, S., Fishburn, J., Warfield, L., Schreiber, J., Noble, W.S., Söding, J., and Hahn, S. (2019). A high-throughput screen for transcription activation domains reveals their sequence characteristics and permits reliable prediction by deep learning. BioRxiv 2019.12.11.872986.

Feric, M., Vaidya, N., Harmon, T.S., Mitrea, D.M., Zhu, L., Richardson, T.M., Kriwacki, R.W., Pappu, R. V., and Brangwynne, C.P. (2016). Coexisting Liquid Phases Underlie Nucleolar Subcompartments. Cell *165*, 1686–1697.

Fox, A.H., Nakagawa, S., Hirose, T., and Bond, C.S. (2018). Paraspeckles: Where Long Noncoding RNA Meets Phase Separation. Trends Biochem. Sci. *43*, 124–135.

Fukaya, T., Lim, B., and Levine, M. (2016). Enhancer Control of Transcriptional Bursting. Cell *166*, 358–368.

Fulco, C.P., Nasser, J., Jones, T.R., Munson, G., Bergman, D.T., Subramanian, V., Grossman, S.R., Anyoha, R., Doughty, B.R., Patwardhan, T.A., et al. (2019). Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nat. Genet. *51*, 1664–1669.

Gibson, B.A., Doolittle, L.K., Schneider, M.W.G., Jensen, L.E., Gamarra, N., Henry, L., Gerlich, D.W., Redding, S., and Rosen, M.K. (2019). Organization of Chromatin by Intrinsic and Regulated Phase Separation. Cell *0*.

Glaser, J., Nguyen, T.D., Anderson, J.A., Lui, P., Spiga, F., Millan, J.A., Morse, D.C., and Glotzer, S.C. (2015). Strong scaling of general-purpose molecular dynamics simulations on GPUs. Comput. Phys. Commun. *192*, 97–107.

Harmon, T.S., Holehouse, A.S., Rosen, M.K., and Pappu, R. V (2017). Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. Elife *6*.

Hnisz, D., Day, D.S., and Young, R.A. (2016a). Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. Cell *167*, 1188–1200.

Hnisz, D., Weintrau, A.S., Day, D.S., Valton, A.L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., et al. (2016b). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science *351*, 1454–1458.

Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., and Sharp, P.A. (2017). A Phase Separation Model for Transcriptional Control. Cell *169*, 13–23.

Hopfield, J.J. (1974). Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. Proc. Natl. Acad. Sci. U. S. A. *71*, 4135–4139.

Huihui, J., Firman, T., and Ghosh, K. (2018). Modulating charge patterning and ionic strength as a strategy to induce conformational changes in intrinsically disordered proteins. J. Chem. Phys. *149*, 085101.

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *527*, 384–388.

Jung, C., Bandilla, P., von Reutern, M., Schnepf, M., Rieder, S., Unnerstall, U., and Gaul, U. (2018). True equilibrium measurement of transcription factor-DNA binding affinities using automated polarization microscopy. Nat. Commun. *9*, 1605.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S.R., Tan, G., et al. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. *46*, D260–D266.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell *172*, 650–665.

Larson, A.G., Elnatan, D., Keenen, M.M., Trnka, M.J., Johnston, J.B., Burlingame, A.L., Agard, D.A., Redding, S., and Narlikar, G.J. (2017). Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. Nature 236–240.

Lee, T.I., and Young, R.A. (2013). Transcriptional Regulation and Its Misregulation in Disease. Cell *152*, 1237–1251.

Levo, M., and Segal, E. (2014). In pursuit of design principles of regulatory sequences. Nat. Rev. Genet. *15*, 453–468.

Li, J., Dong, A., Saydaminova, K., Chang, H., Wang, G., Ochiai, H., Yamamoto, T., and Pertsinidis, A. (2019). Single-Molecule Nanoscopy Elucidates RNA Polymerase II Transcription at Single Genes in Live Cells. Cell *0*.

Lin, C.Y., Lovén, J., Rahl, P.B., Paranal, R.M., Burge, C.B., Bradner, J.E., Lee, T.I., and Young, R.A. (2012). Transcriptional amplification in tumor cells with elevated c-Myc. Cell *151*, 56–67.

Lin, Y.-H., Brady, J.P., Forman-Kay, J.D., and Chan, H.S. (2017). Charge pattern matching as a 'fuzzy' mode of molecular recognition for the functional phase separations of intrinsically disordered proteins. New J. Phys. *19*, 115003.

Long, H.K., Prescott, S.L., and Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell *167*, 1170–1187.

Lu, H., Yu, D., Hansen, A.S., Ganguly, S., Liu, R., Heckert, A., Darzacq, X., and Zhou, Q. (2018). Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. Nature *558*, 318–323.

Maniatis, T., Falvo, J. V, Kim, T.H., Kim, T.K., Lin, C.H., Parekh, B.S., and Wathelet, M.G. (1998). Structure and function of the interferon-beta enhanceosome. Cold Spring Harb. Symp. Quant. Biol. *63*, 609–620.

Markstein, M., Markstein, P., Markstein, V., and Levine, M.S. (2002). Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. Proc. Natl. Acad. Sci. *99*, 763–768.

Mir, M., Stadler, M.R., Ortiz, S.A., Harrison, M.M., Darzacq, X., and Eisen, M.B. (2018). Dynamic multifactor hubs interact transiently with sites of active transcription in Drosophila embryos. BioRxiv 377812.

Morgunova, E., and Taipale, J. (2017). Structural perspective of cooperative transcription factor binding. Curr. Opin. Struct. Biol. *47*, 1–8.

Nguyen, T.D., Phillips, C.L., Anderson, J.A., and Glotzer, S.C. (2011). Rigid body constraints realized in massively-parallel molecular dynamics on graphics processing units. Comput. Phys. Commun. *182*, 2307–2313.

Ninio, J. (1975). Kinetic amplification of enzyme discrimination. Biochimie *57*, 587–595.

Nizami, Z., Deryusheva, S., and Gall, J.G. (2010). The Cajal Body and Histone Locus Body. Cold Spring Harb. Perspect. Biol. *2*, a000653.

Nott, T.J., Petsalaki, E., Farber, P., Jervis, D., Fussner, E., Plochowietz, A., Craggs, T.D., Bazett-Jones, D.P., Pawson, T., Forman-Kay, J.D., et al. (2015). Phase Transition of a Disordered Nuage Protein Generates Environmentally Responsive Membraneless Organelles. Mol. Cell *57*, 936–947.

Pederson, T. (2011). The nucleolus. Cold Spring Harb. Perspect. Biol. *3*, a000638.

Ptashne, M., and Gann, A. (1997). Transcriptional activation by recruitment. Nature *386*, 569–577.

Rajewsky, N., Vergassola, M., Gaul, U., and Siggia, E.D. (2002). Computational detection of genomic cis- regulatory modules applied to body patterning in the early Drosophila embryo. BMC Bioinformatics *3*, 30.

Rowley, M.J., Nichols, M.H., Lyu, X., Ando-Kuri, M., Rivera, I.S.M., Hermetz, K., Wang, P., Ruan, Y., and Corces, V.G. (2017). Evolutionarily Conserved Principles Predict 3D Chromatin Organization. Mol. Cell *67*, 837-852.e7.

Sabari, B.R., Dall'Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A. V., Manteiga, J.C., et al. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. Science *361*.

Sabari, B.R., Dall'Agnese, A., and Young, R.A. (2020). Biomolecular Condensates in the Nucleus. Trends Biochem. Sci. *0*.

Salzler, H.R., Tatomer, D.C., Malek, P.Y., McDaniel, S.L., Orlando, A.N., Marzluff, W.F., and Duronio, R.J. (2013). A Sequence in the Drosophila H3-H4 Promoter Triggers Histone Locus Body Assembly and Biosynthesis of Replication-Coupled Histone mRNAs. Dev. Cell *24*, 623–634.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., Haering, C., Mirny, L., et al. (2017). Two independent modes of chromatin organization revealed by cohesin removal. Nature *551*, 51.

Semenov, A.N., and Rubinstein, M. (1998). Thermoreversible Gelation in Solutions of Associative Polymers. 1. Statics. Macromolecules *31*, 1373–1385.

Sherry, K.P., Das, R.K., Pappu, R. V, and Barrick, D. (2017). Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the Notch receptor. Proc. Natl. Acad. Sci. U. S. A. *114*, E9243–E9252.

Shin, Y., and Brangwynne, C.P. (2017). Liquid phase condensation in cell physiology and disease. Science *357*, eaaf4382.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet. *15*, 272–286.

Singh, G., Mullany, S., Moorthy, S., Zhang, R., Mehdi, T., Shchuka, V., Tian, R., Moses, A., and Mitchell, J. (2020). A flexible repertoire of transcription factor binding sites and diversity threshold determines enhancer activity in embryonic stem cells. BioRxiv 2020.04.17.046664.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A.C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. Trends Biochem. Sci. *39*, 381–399.

Smith, E., and Shilatifard, A. (2014). Enhancer biology and enhanceropathies. Nat. Struct. Mol. Biol. *21*, 210–219.

Spitz, F., and Furlong, E.E.M.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.

Srebnik, S., Chakraborty, A.K., and Shakhnovich, E.I. (1996). Adsorption-Freezing Transition for Random Heteropolymers near Disordered 2D Manifolds due to "Pattern Matching." Phys. Rev. Lett. *77*, 3157–3160.

Stampfel, G., Kazmar, T., Frank, O., Wienerroither, S., Reiter, F., and Stark, A. (2015). Transcriptional regulators form diverse groups with context-dependent regulatory functions. Nature *528*, 147.

Strom, A.R., Emelyanov, A. V., Mir, M., Fyodorov, D. V., Darzacq, X., and Karpen, G.H. (2017). Phase separation drives heterochromatin domain formation. Nature *547*, 241–245.

Tatavosian, R., Kent, S., Brown, K., Yao, T., Duc, H.N., Huynh, T.N., Zhen, C.Y., Ma, B., Wang, H., and Ren, X. (2018). Nuclear condensates of the Polycomb protein chromobox 2 (CBX2) assemble through phase separation. J. Biol. Chem. jbc.RA118.006620.

Tsai, A., Muthusamy, A.K., Alves, M.R., Lavis, L.D., Singer, R.H., Stern, D.L., and Crocker, J. (2017). Nuclear microenvironments modulate transcription from low-affinity enhancers. Elife *6*.

Wang, J., Choi, J.-M., Holehouse, A.S., Lee, H.O., Zhang, X., Jahnel, M., Maharana, S., Lemaitre, R., Pozniakovsky, A., Drechsel, D., et al. (2018). A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. Cell *174*, 688-699.e16.

Wasserman, W.W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet. *5*, 276–287.

Wei, M.-T., Elbaum-Garfinkle, S., Holehouse, A.S., Chen, C.C.-H., Feric, M., Arnold, C.B., Priestley, R.D., Pappu, R. V., Brangwynne, C.P., Chih, C., et al. (2017). Phase behaviour of disordered proteins underlying low density and high permeability of liquid organelles. Nat. Chem. *9*, 1118–1125.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–1443.

White, A.E., Burch, B.D., Yang, X., Gasdaska, P.Y., Dominski, Z., Marzluff, W.F., and Duronio, R.J. (2011). Drosophila histone locus bodies form by hierarchical recruitment of components. J. Cell Biol. *193*, 677–694.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell *153*, 307–319.

Wunderlich, Z., and Mirny, L.A. (2009). Different gene regulation strategies revealed by analysis of binding motifs. Trends Genet. *25*, 434–440.

Yamazaki, T., Souquere, S., Chujo, T., Kobelke, S., Chong, Y.S., Fox, A.H., Bond, C.S., Nakagawa, S., Pierron, G., and Hirose, T. (2018). Functional Domains of NEAT1 Architectural lncRNA Induce Paraspeckle Assembly through Phase Separation. Mol. Cell *70*, 1038-1053.e7.

Zhu, Y., Qi, C., Jain, S., Le Beau, M.M., Espinosa, R., Atkins, G.B., Lazar, M.A., Yeldandi, A. V, Rao, M.S., and Reddy, J.K. (1999). Amplification and overexpression of peroxisome proliferator-activated receptor binding protein (PBP/PPARBP) gene in breast cancer. Proc. Natl. Acad. Sci. U. S. A. *96*, 10848–10853.

Chapter 3: *Control*

" An idea that is not dangerous is unworthy of being called an idea at all."

- Oscar Wilde

RNA-mediated non-equilibrium feedback regulation of transcriptional condensates

This chapter is primarily based on work *under review* titled
"RNA-mediated feedback control of transcriptional condensates "
Jonathan E. Henninger[=], Ozgur Oksuz[=], **Krishna Shrinivas[=]**, Ido Sagi, Gary LeRoy, Ming Zheng, James Owen Andrews, Alicia V. Zamudio, Charalampos Lazaris, Nancy M. Hannett, Tong Ihn Lee, Phillip A. Sharp, Ibrahim I. Cissé, Arup K. Chakraborty[c], Richard A. Young[c]
(=equal contributions. cco-corresponding authors)

Regulation of biological processes typically incorporate mechanisms that both initiate and terminate the process and, where understood, these mechanisms often involve feedback control. Condensates are increasingly implicated in various biological phenomena, yet, a framework to understand their regulation by feedback control is not well developed.

In this chapter, we'll study how transcriptional condensates are regulated by the outcome of transcription i.e. energy-dependent synthesis of RNA. By iterating theory, simulation, and experiments, we will propose that transcriptional regulation incorporates a feedback mechanism whereby small amounts of transcribed RNAs stimulate condensate formation but larger bursts of RNA ultimately arrest the process by causing dissolution. Towards the end, we'll discuss the ramifications on two enigmatic features of transcription – widespread production of non-coding RNAs and characteristics of discrete and bursty transcription.

Introduction

Diverse biological processes have evolved feedback mechanisms to enable both positive and negative regulation[1]. Examples of biological processes that are known to incorporate feedback regulation include signal transduction (Brandman and Meyer, 2008) , production of RNA splicing factors (Jangi and Sharp, 2014), circadian rhythms (Dunlap, 1999), red blood cell production (Ebert and Bunn, 1999), and response to DNA damage (Lahav et al., 2004). In transcription, some factors that regulate amino acid biosynthetic pathway genes can be allosterically regulated by intermediates produced by those pathways (Bergot et al., 1992; Bruhat et al., 2000; Sellick and Reece, 2003), but a general feedback mechanism has not been described. Evidence that feedback control is often mediated by the product of the process (Brandman and Meyer, 2008; Elowitz and Leibler, 2000; Gardner et al., 2000; Umbarger, 1956; Monod and Jacob, 1961) is one of the factors that led us to postulate that RNA may regulate transcription by a feedback mechanism.

Mammalian transcription produces diverse RNA species from regulatory elements and genes (Smith et al., 2019) and transcription of genes occurs in bursts of RNA synthesis (Chubb et al., 2006; Raj and van Oudenaarden, 2008; Raj et al., 2006). Transcription factors and coactivators recruit RNA polymerase II (Pol II) to enhancer and promoter ele-

[1]The first example of feedback and/or allosteric regulation comes from the seminal work Monod and Jacobs on the lac operon. See Judson, 2019 for an accessible overview of their discovery in the "golden-age" of molecular biology.

ment, where short (20-400 bp) RNAs are bidirectionally transcribed before Pol II pauses[2] (Adelman and Lis, 2012; Core and Adelman, 2019; Jin et al., 2017; Kim et al., 2010; Seila et al., 2008). These RNA species are short-lived and are reported to have various regulatory roles, although there isn't yet a consensus on their functions (Andersson et al., 2014; Catarino and Stark, 2018; Core et al., 2014; Henriques et al., 2018; Lai et al., 2013; Li et al., 2016; Mikhaylichenko et al., 2018; Nair et al., 2019, 2019; Pefanis et al., 2015; Rahnamoun et al., 2018; Schaukowitch et al., 2014; Scruggs et al., 2015; Sigova et al., 2015; Smith et al., 2019; Struhl, 2007). Pol II pause release leads to processive elongation, which occurs in periodic bursts (~1-10 minutes in duration), where multiple molecules of Pol II[3] can be released from promoters within a short timeframe and produce multiple molecules of mRNA (~1-100 molecules per burst) (Cisse et al., 2013; Fukaya et al., 2016; Larsson et al., 2019). How and whether the diverse RNA species produced during transcription, which differ in length, half-life, and number, impact or regulate transcription is currently unclear.

[2] An interesting study connects this antisense transcription and the spliceosome to evolutionary trajectories of promoter directionality (Almada et al., 2013)

[3] (Forero-Quintero et al., 2020) is a recent pre-print that tries to estimate the size of these transient clusters through live-cell imaging in an integrated read-out promoter – They estimate upto 75 polymerases transiently assembling, even if only ~1-10 continue for processive elongation

Recent studies have shown that transcriptional condensates can compartmentalize and concentrate large numbers of transcription factors, cofactors and Pol II at super-enhancers, which are clusters of enhancers that regulate genes with prominent roles in cell identity (Boija et al., 2018; Cho et al., 2018; Cramer, 2019; Hnisz et al., 2017; Sabari et al., 2018). The component enhancer elements of such genes promote transcriptional condensate formation by crowding transcription factors and Mediator at densities above sharply defined thresholds for condensate formation (Shrinivas et al., 2019). Transcriptional condensates are highly dynamic and can be observed in live cells to form and dissolve at timescales ranging from seconds to minutes (Cho et al., 2018). The periodic formation and dissolution of dynamic transcriptional condensates, coupled with evidence that different species and levels of RNAs are produced at different stages of transcription led us to wonder whether transcriptional condensates are regulated by a non-equilibrium feedback mechanism mediated by its RNA product.

RNA molecules are components of, and play regulatory roles in, diverse biomolecular condensates. These include the nucleolus, nuclear speckles, paraspeckles, and stress granules (Fay and Anderson, 2018; Strom and Brangwynne, 2019). RNA has a high negative charge density due to its

phosphate backbone, and the effective charge of a given RNA molecule is directly proportional to its length (Boeynaems et al., 2019). Condensates are thought to be formed by ensemble low-affinity molecular interactions, including electrostatic interactions, and RNA can be a powerful regulator of condensates that are formed and maintained by electrostatic forces (Banani et al., 2017; Peran and Mittag, 2020; Shin and Brangwynne, 2017). Indeed, RNA has been shown to enter and modify the properties of simple condensates formed by polyelectrolyte-rich molecules (Drobot et al., 2018; Frankel et al., 2016; Mountain and Keating, 2020). In a phenomenon called complex coacervation[4], a type of liquid-liquid phase separation mediated by electrostatic interactions between oppositely charged polyelectrolytes, low levels of RNA can enhance condensate formation whereas high levels can cause their dissolution (Lin et al., 2019; Overbeek and Voorn, 1957; Sing, 2017; Srivastava and Tirrell, 2016). Condensate formation and subsequent dissolution with increasing RNA concentration is an example of reentrant phase behavior, which is driven by favorable opposite-charge interactions at low RNA concentrations (formation) and repulsive like-charge interactions at high RNA concentrations (dissolution). We wondered whether such a reentrant equilibrium phase behavior coupled to the non-equilibrium processes that occur during transcription could regulate transcriptional output.

By combining physics-based modeling and experimental analysis, we propose and test a model whereby the products of transcription initiation stimulate condensate formation and those of a burst of elongation stimulate condensate dissolution. We provide experimental evidence that physiological RNA levels can enhance or dissolve transcriptional condensates. These results provide a mechanism by which the products of transcription regulate condensate behaviors and thus transcription, and suggest that this non-equilibrium process provides negative feedback to dissolve transcriptional condensates and arrest transcription.

[4]There are a number of studies, starting from early polymer chemists (Overbeek and Voorn, 1957) to more recent explorations in synthetic systems (Aumiller and Keating, 2016; Pak et al., 2016; Sing, 2017), along with in-text citations that explore this topic. Polymaths Oparin and Haldane (Oparin, 1953) speculated the coacervates may be a model for early origin-of-life events – with RNA as the center-piece moiety.

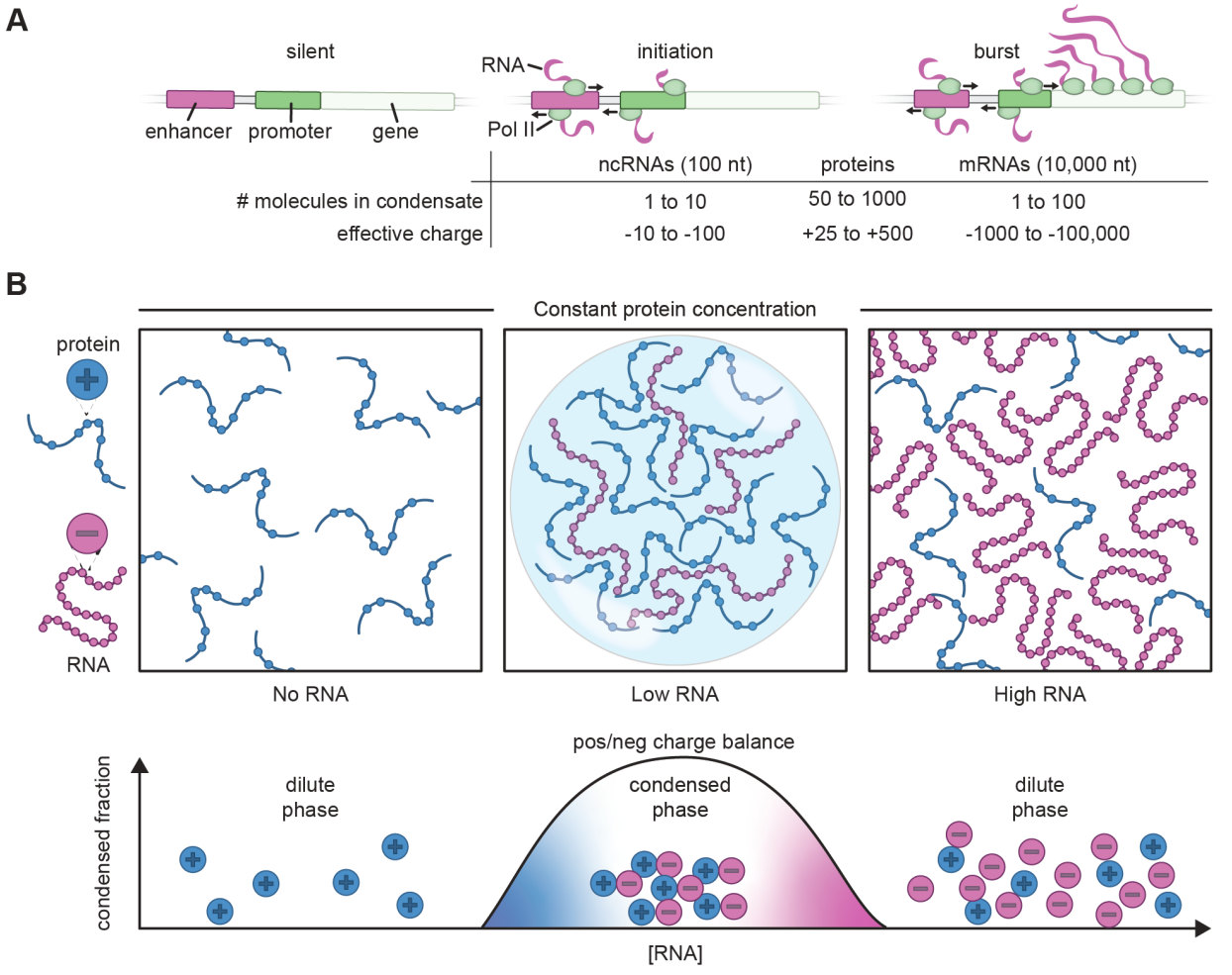
RESULTS

Reentrant equilibrium phase behavior of RNA and transcriptional proteins

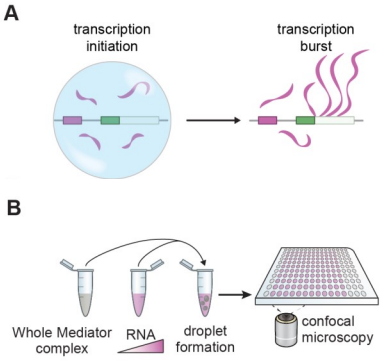
To explore the potential role of RNA in regulating transcriptional condensates, we sought to estimate the number and effective charge of RNA and protein molecules in a typical transcriptional condensate at different stages of transcription. In early stages of transcription, low levels of small noncoding RNAs are produced by Pol II at enhancers and promoter-proximal regions (Figure 1A) (Adelman and Lis, 2012; Core and Adelman, 2019; Kim et al., 2010; Seila et al., 2008). During pause release, Pol II produces longer genic RNAs during bursts of transcription elongation (Figure 1A) (Adelman and Lis, 2012; Core and Adelman, 2019).

1A Scheme of transcription states and the number of molecules and their corresponding effective charge in a typical transcriptional condensate during initiation and bursts of transcription (STAR Methods)

1B Diagram of reentrant phase transition in response to increasing concentration of RNA over constant protein concentration. The condensed fraction of protein peaks at the RNA concentration at which the charges between protein and RNA are balanced, while alteration of this charge balance in either direction decreases the condensed fraction.



These protein- and RNA-rich states can be thought of as mixtures of poly-electrolytes that may undergo complex coacervation (Figure 1B) (Lin et al., 2019; Overbeek and Voorn, 1957; Sing, 2017; Srivastava and Tirrell, 2016). This is likely to be relevant to transcriptional condensates because electrostatic interactions contribute to the formation of these condensates (Boija et al., 2018; Sabari et al., 2018). Complex coacervate formation through phase separation is promoted when poly-electrolytes are present at concentrations where their net charges are approximately balanced (Figure 1B). When the concentration of a poly-electrolyte, such as RNA, becomes sufficiently high, the domination of repulsive like-charge interactions can suppress phase separation (Lin et al., 2019; Milin and Deniz, 2018; Muthukumar, 2016; Overbeek and Voorn, 1957; Zhang et al., 2018). Thus, at constant protein concentration, titrating RNA levels results in reentrant phase behavior, by which low RNA levels promote and high RNA levels suppress condensate formation (Figure 1B) (Milin and Deniz, 2018; Zhang et al., 2018). We wondered whether the reentrant phase behavior can apply to the regulation of transcriptional condensates during transcription. Because the quantities of the diverse RNA species and proteins present in transcriptional condensates in populations of cells can be estimated (Figure 1A and S1), it is possible to conduct experimental tests to determine whether reentrant phase behavior occurs under physiologically-relevant conditions of these molecules.



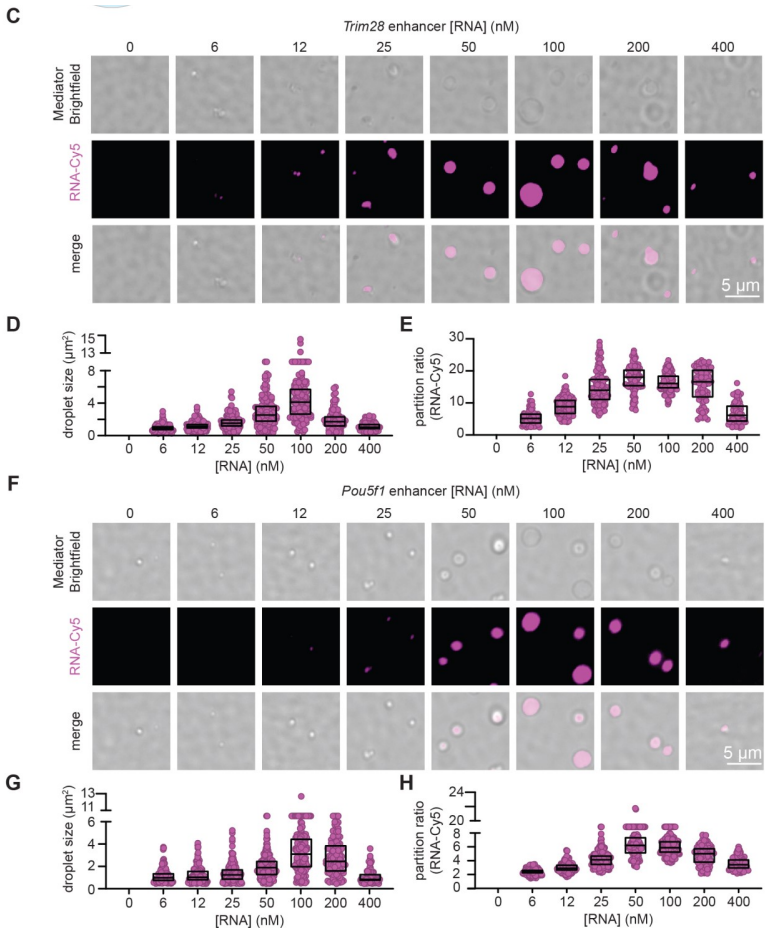
2A. Scheme for condensate dissolution upon increase in RNA levels during transcriptional burst.

2B. Experimental design for in vitro droplet formation assay. Whole Mediator complex is mixed with increasing concentrations of RNA under physiologically-relevant buffer conditions and droplets are imaged under confocal microscopy.

Low levels of RNA enhance and high levels dissolve Mediator condensates

As an initial test of whether low levels of RNA stimulate transcriptional condensate formation while high levels of RNA favor condensate dissolution (Figure 2A), we used an in vitro droplet assay (Figure 2B). Using components at physiologically-relevant conditions, we investigated whether an enhancer RNA transcribed from the Trim28 super-enhancer, which has previously been shown to form a transcriptional condensate in living cells (Boija et al., 2018; Guo et al., 2019), influences condensate formation of purified Mediator complex. Measurement of enhancer RNA levels in cells indicated that ~0.2 molecules of this enhancer RNA exist at steady-state in murine embryonic stem cells (mESCs) (Figure S1E). Given that multiple loci in a super-enhancer are

transcribed into enhancer RNAs, this roughly corresponds to ~100-1000 nM of RNA in a typical Mediator condensate in cells (STAR Methods). These condensates typically contain Mediator at a concentration of around 1-20 μM (STAR Methods). The results showed that addition of 6 -400 nM Trim28 enhancer RNA to 200 nM purified Mediator complex had a dose-dependent effect on the size of Mediator/RNA droplets (Figures 2C-2E). Droplet sizes peaked at 100 nM RNA (Figure 2D) and the relative enrichment of RNA in the droplets, as measured by the ratio of average intensity inside versus outside the droplet (partition ratio), followed a similar trend (Figure 2E). Similar results were obtained using an enhancer RNA transcribed from the Pou5f1 super-enhancer (Figures 2F-2H). Thus, within the range of physiological levels observed in cells, low levels of RNA can enhance condensate formation and high levels of RNA can reduce condensate formation by Mediator in vitro.



2C. Representative images of droplets formed by the unlabeled whole Mediator complex (200 nM) and Cy5-labeled *Pou5f1* enhancer RNA at increasing concentrations (0-400 nM).

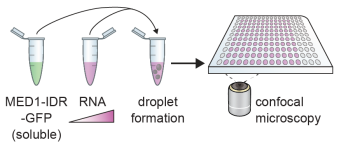
2D. Quantification of droplet sizes in (C).

2E. Quantification of partition ratios of Cy5-labeled RNA within the droplets in (C). Partition ratio is calculated as the mean intensity inside the droplets divided by the mean intensity outside the droplets.

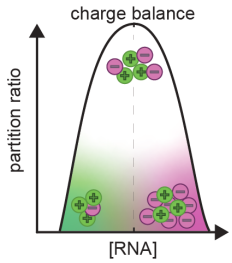
2F. Representative images of droplets formed by the unlabeled whole Mediator complex (200 nM) and Cy5-labeled *Trim28* enhancer RNA at increasing concentrations (0-400 nM).

2G.Quantification of droplet sizes in (F).

2H. Quantification of partition ratios of Cy5-labeled RNA within the droplets in (F).



3A. Experimental design for in vitro droplet formation assay. Soluble MED1-IDR-GFP is mixed with increasing concentrations of RNA under physiologically relevant buffer conditions and droplets are imaged with confocal microscopy.

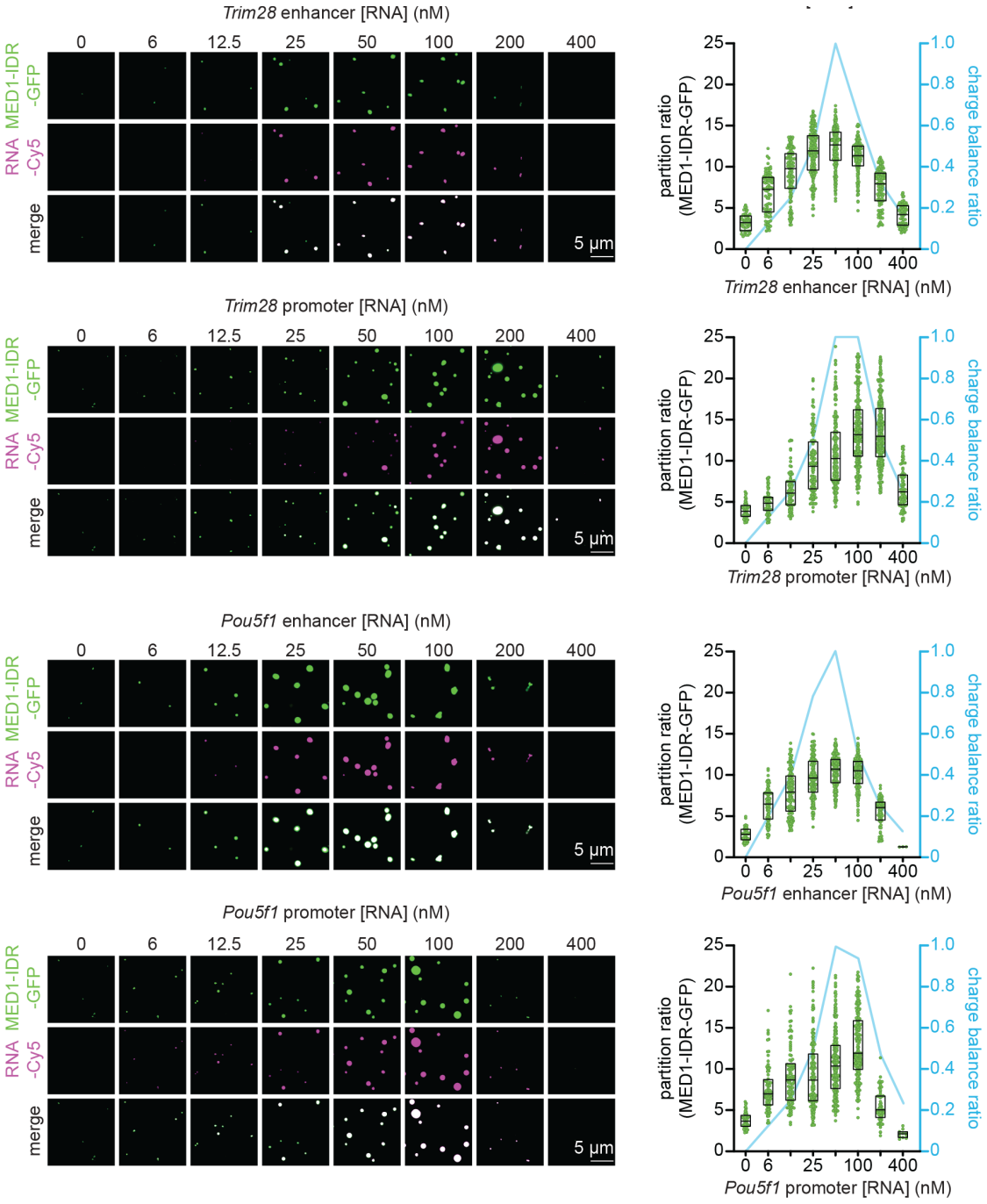


3B. Scheme of charge balance ratio between constant protein concentration and increasing RNA concentration.

RNA-mediated regulation of MED1-IDR condensates fits a charge balance model

We next sought to quantify how diverse RNAs regulate the reentrant phase behavior of transcriptional proteins. We performed in vitro droplet assays (Figure 3A) using the MED1 C-terminal intrinsically disordered region (MED1-IDR), which has proven to be a useful surrogate for the multisubunit Mediator complex, as it is not possible to purify sufficient amounts of this complex to test all the parameters of interest (Boija et al., 2018; Guo et al., 2019; Sabari et al., 2018; Shrinivas et al., 2019; Zamudio et al., 2019). The fusion of GFP to MED1-IDR allows quantification by fluorescence of a single species whose effective charge can be calculated to determine the charge ratio between protein and RNA. Noncoding and coding RNAs produced from three different super-enhancer loci and their associated genes (*Trim28, Pou5f1*, *Nanog*; Figure S1) were selected for this analysis based on prior studies of nascent RNA sequencing data in mESCs (Boija et al., 2018; Guo et al., 2019; Sabari et al., 2018; Sigova et al., 2015; Whyte et al., 2013). Addition of 6-400 nM of each of these RNAs to 1000 nM MED1-IDR (protein:RNA ratios = 167 to 2.5) stimulated formation of MED1-IDR condensates at low RNA concentrations and dissolved MED1-IDR condensates at higher RNA concentrations (Figures 3C,3D and S2). This effect was not observed when the RNA was tested with GFP alone or OCT4-GFP, which has a net negative charge (Figure S3). These results show that diverse RNAs are capable of stimulating MED1-IDR condensate formation when present at relatively low levels and dissolving MED1-IDR condensates at high levels.

We sought to further test whether the RNA-mediated effects on MED1-IDR condensates fit a charge balance model. MED1-IDR/RNA condensate formation should be enhanced when the protein and RNA polymers are balanced in charge, and they should be sensitive to disruption of this balance. To test this model, we quantified the relative charge of RNA and MED1-IDR (STAR Methods). As expected, RNA-mediated effects on MED1-IDR condensates fit a charge balance model (Figure 3D, blue lines). We would expect an RNA length-dependent shift in the RNA level required for peak MED1-IDR partitioning when RNAs of different length are introduced into the droplet assay in equal numbers.



C. (left) Representative images of droplets formed by increasing concentrations (0-400 nM) of the indicated RNAs mixed with 1 μM of MED1-IDR.

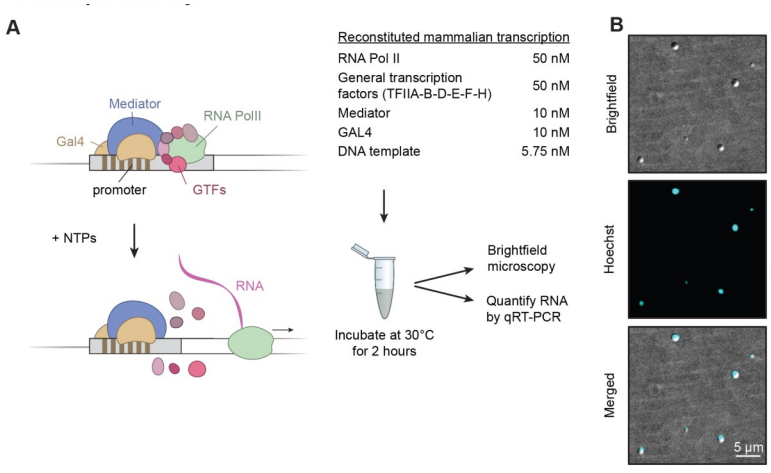
D (right) Quantification of partition ratios of MED1-IDR-GFP within the droplets in (C) (left y-axis). Charge balance ratios between MED1-IDR-GFP and increasing concentrations of the indicated RNAs are shown in blue lines (right y-axis).

This expectation that a higher concentration of shorter RNAs is needed to disrupt condensate formation was observed (Figures S4A and S4B). Another prediction of the charge balance model is that these interactions should be largely independent of RNA sequence, so antisense versions of any one of the RNA species should exhibit the same quantitative effects as the sense strand, and this was also observed (Figures S4B and S4C). As expected for a charge-balance model, MED1-IDR condensates

formed with RNA were sensitive to increasing monovalent salt, which screens charged interactions (Figure S4D). These results further support a charge balance model for the RNA-mediated effects on the equilibrium behavior of MED1-IDR condensates.

RNA-mediated effects on condensates in reconstituted in vitro transcription assays

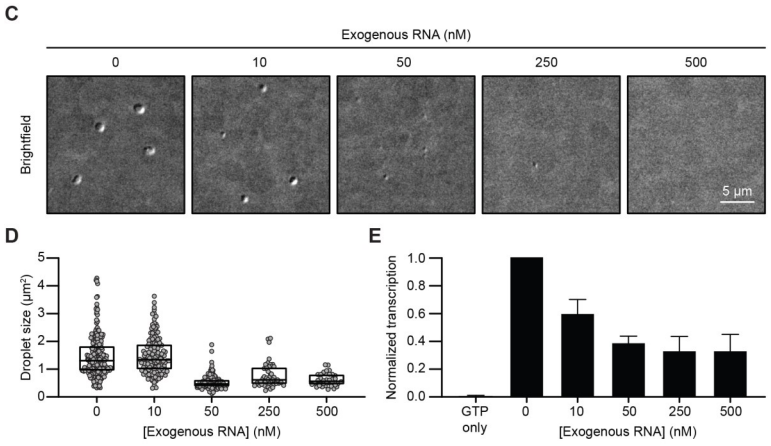
We sought to investigate the functional consequence of the RNA-mediated reentrant phase behavior on transcription. Pol II-dependent transcription can be reconstituted in vitro with purified components (Roeder, 2019), so we investigated whether droplets containing transcriptional components are formed in these assays and if conditions that alter droplet levels similarly alter transcriptional output. We used a classical reconstituted mammalian transcription system with purified components, including Pol II, general transcription factors, Mediator and a transcriptional activator (Gal4), where addition of nucleotides permits transcription of a linear DNA template (Figure 4A). We observed that component mixtures and buffer conditions that are optimal for transcriptional output (Flores et al., 1992; LeRoy et al., 2008; Orphanides et al., 1998) produced droplets containing the DNA template (Figure 4B). These results suggest that transcription may occur within condensate droplets in reconstituted systems under conditions optimal for transcription.



4A. Cartoon representation of the reconstituted in vitro mammalian transcription assay with purified components (left) and the design of the assay (right). General transcription factors (GTFs), Mediator, Gal4 and RNA Pol II are assembled on a template DNA containing a promoter with TATA-box and Gal4 binding sites. The transcription reaction is initiated by addition of NTPs. The transcription reaction is then subjected to indicated analyses.

4B. Brightfield images of droplets formed within the in vitro transcription reaction. Droplets are stained with DNA dye (Hoechst).

Quantification of the newly synthesized RNA in this system showed that 3.5 (± 0.5) pM RNA was produced in the transcription reaction (STAR Methods). The levels of RNA produced in the reconstituted system are unlikely to dissolve droplets formed within the in vitro transcription reaction. Thus, we tested whether increasing the levels of RNA through addition of purified RNA might simultaneously alter both droplet size and transcriptional output. Indeed, elevated concentrations of RNA led to a reduction in number and size of the droplets (Figures 4C and 4D), which correlated with a reduction in template-derived RNA synthesis as measured by qRT-PCR (Figure 4E). Thus, high levels of RNA can produce correlative negative effects on condensates and RNA synthesis in a reconstituted transcription assay.



4C. The effect of increasing exogenous RNA levels on transcriptional condensates. Representative images of droplets in the in vitro transcription reaction in the presence of indicated amounts of exogenous RNA.

4D. Quantification of droplet sizes in (C) (p=0.9309 0 vs. 10; p<0.001 for 0 vs. 50, 250, and 500, one-way ANOVA).

4E. The effect of increasing exogenous RNA levels on transcriptional output. Transcriptional output is measured by qRT-PCR. The relative levels of transcriptional output normalized to no RNA condition are indicated. The mean of 2 replicates are shown and error bars depict S.D. (p=0.0001 GTP only vs. 0; p=0.0111 0 vs. 10; p=0.0013 0 vs. 50; p=0.0008 0 vs. 250; p=0.008 0 vs. 500, one-way ANOVA)

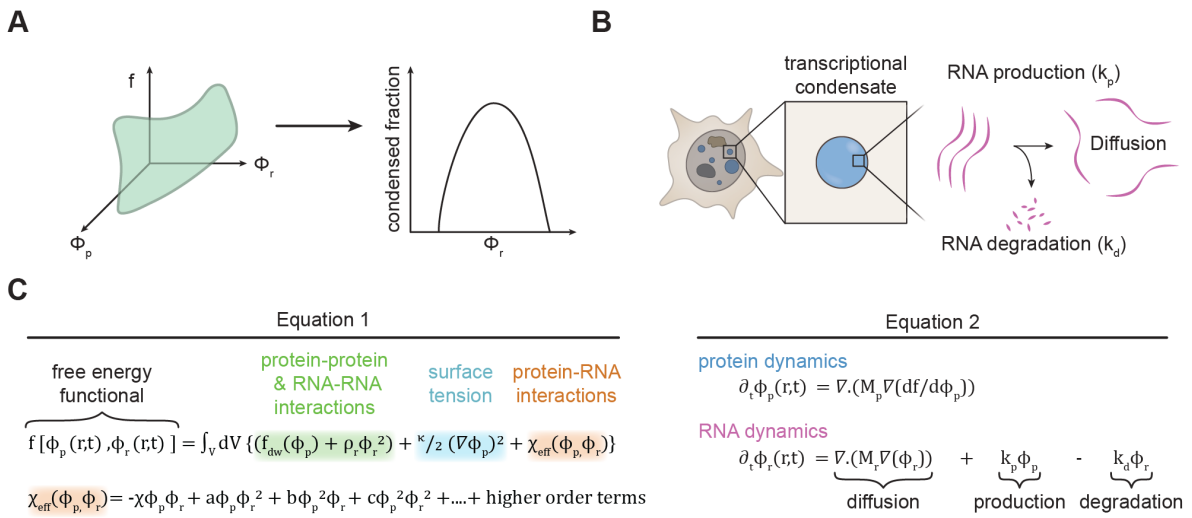
A model for RNA-mediated non-equilibrium feedback control of transcriptional condensates

The in vitro experiments, which provide evidence that key transcriptional proteins and RNA exhibit an electrostatics-driven, RNA-protein ratio dependent, reentrant phase transition, were performed under equilibrium conditions (Figures 2 and 3). However, in vivo, RNA is synthesized and degraded at specific genomic loci by dynamic, ATP-dependent, non-equilibrium processes (Azofeifa et al., 2018; Li et al., 2016; Pefanis et al., 2015). To investigate how non-equilibrium processes underlying transcription may regulate transcriptional condensates, we built a physics-based model with the goal of gaining mechanistic insights that could be tested experimentally. The model consists of two inter-linked parts: (1) A free-energy function (Figure 5A), which depends on the concentra-

tions of transcriptional proteins and RNA, that recapitulates the equilibrium reentrant phase behavior of RNA-protein mixtures (Figures 2 and 3). (2) A mathematical framework to study spatiotemporal evolution of condensates subject to dynamical processes of RNA synthesis, degradation, and diffusion (Figure 5B).

We first developed a free-energy function to recapitulate the experimentally observed reentrant phase behavior of RNA-protein mixtures (Figure 5A). The free energy function depends on the concentrations of transcriptional proteins and RNA, $\phi_P(r,t)$ and $\phi_R(r,t)$, which vary in space and time. For simplicity, all transcriptional proteins are combined into one pseudo-species. The free energy function describes repulsive RNA-RNA interactions and favorable interactions among the transcriptional proteins that drive condensate formation of transcriptional proteins in the absence of RNA (Figure 5C, Eq. 1, in green), as well as a surface tension term important for describing condensate formation (Figure 5C, Eq. 1 in blue, STAR Methods). The free energy function also includes protein-RNA interactions that are described by a concentration -dependent interaction term, which is expanded in the standard Landau fashion (Figure 5C, Eq. 1, in red) (Kardar, 2007). While symmetry arguments do not preclude any specific terms in this expansion, the choice of (χ>0,a,b ≈0, c≥0) ensures a reentrant phase transition (schematic in Figures 5B, S5A, STAR Methods) with a minimal number of higher order terms. Results using the simple Landau free energy (Figure 5C, Eq. 1) are recapitulated using a Flory-Huggins approach (Figure S5A and S5B, STAR Methods,). The free energy landscape described here enables us to subsequently study how the dynamics of transcriptional condensates is regulated by transcription.

We next developed a mathematical framework to study the temporal evolution of transcriptional condensates as transcription ensues. Most transcriptional proteins turn-over with a half-life of several hours (Cambridge et al., 2011; Chen et al., 2016), which is longer than timescales of transcription-associated events, which range from seconds to minutes (Chen and Larson, 2016; Fukaya et al., 2016; Rodriguez and Larson, 2020). Hence, the overall amount of protein is conserved in the timescales of interest. Thus the dynamics of the protein concentration (ϕ_P) are represented by standard Model B dynamics (Figure 5C, Eq. 2) (Hohenberg and Halperin, 1977). Since RNA concentrations vary over



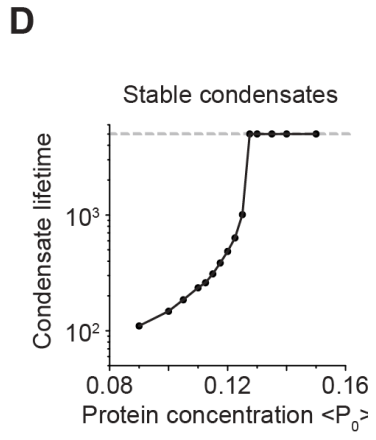
5A. Schematic of coarse-grained free-energy (f, green-surface) which depends on the transcriptional protein (ϕ_P) and RNA (ϕ_R) concentrations. This free-energy recapitulates in vitro observations of an equilibrium reentrant transition.

5B. Schematic of the non-equilibrium model coupling transcriptional activity with transcriptional condensate dynamics. In the model framework, we focus on a local micro-environment near a single transcriptional condensate (blue). RNA (magenta) is synthesized, degraded, and can diffuse.

5C. Equations underlying construction of the free-energy function (Equation 1) and dynamics of protein and RNA (Equation 2). The governing equations follow the outline in 5A-B, and are described in the text and STAR Methods.

transcription-associated time-scales, the dynamics of $\phi_R(r,t)$is explicitly governed by a reaction-diffusion equation. The key features (schematic in Figure 5B) are that RNA diffuses with mobility M_{RNA} and is synthesized and degraded with specific reaction rates k_P and k_d, respectively. Because the RNA dynamics are far from equilibrium and the free energy function noted above depends upon both protein and RNA concentrations, the coupled temporal evolution of transcriptional proteins and RNA (Figure 5C, Eqs. 1 and 2) cannot be obtained from near-equilibrium considerations of simply going downhill in free energy with time. We employ this mathematical framework to study non-equilibrium regulation of transcriptional condensates.

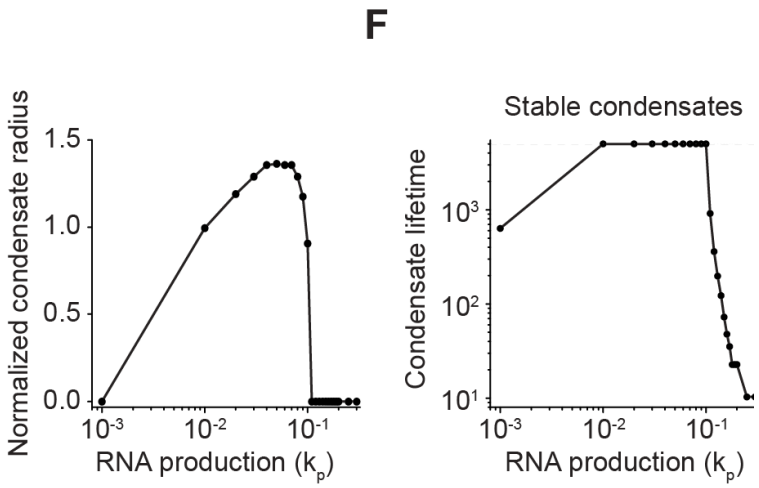
We first sought to determine whether this model is consistent with previous studies (Cho et al., 2016, 2018). These studies have shown that transcriptional condensates at different genomic loci recruit a varying number of transcriptional proteins, which in turn, correlates with condensate lifetimes. To explore this phenomenon, we numerically simulated Eq. 2 (Figure 5C) on 2 and 3-dimensional grids (STAR methods). Locus-dependent recruitment of the transcriptional machinery can be mimicked in our model by varying the total transcriptional protein amount ($<P_0>$) with all other parameters fixed, as our simulation volume represents a local micro-environment (Figure 5A). Our simulations predict that loci that can recruit more transcriptional proteins (higher) form relatively stable condensates ($<P_0>$), while condensates that recruit



5D. Simulation predictions of transcriptional condensate lifetime (ordinate) with varying total protein concentrations (abcissa) (2D simulation grid). The dashed -line represents the lifetime of condensates that don't dissolve at steady state, and the ordinate is presented in units of simulation time.

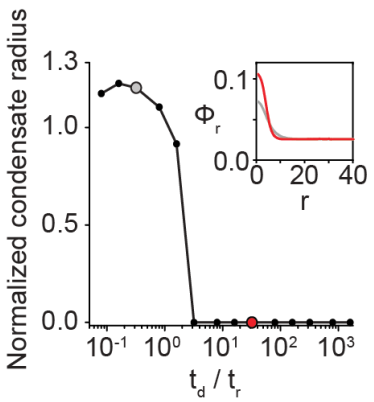
fewer proteins dissolve after a characteristic lifetime (Figure 5D). The model predictions for transcriptional condensate dynamics are qualitatively consistent with published data (Cho et al., 2016), and suggest that features encoded at genomic loci contribute to transcriptional condensate dynamics.

We next explored how non-equilibrium processes underlying transcription impact transcriptional condensates. First, we investigated how the sizes and lifetimes of transcriptional condensates change as a function of the effective rate of RNA synthesis k_P, while keeping all other parameters fixed. In these simulations, the size of condensates initially increases and subsequently decreases with increasing effective rates of RNA synthesis (Figure 5E). Above a threshold rate of RNA synthesis, condensates dissolve (Figure 5E). The underlying reason for this result is the reentrant phase behavior of mixtures of transcriptional molecules and RNA (Figure 1). We also find that condensates with higher transcriptional activity dissolve faster, as measured by condensate lifetimes (Figure 5F). Condensate lifetimes do not vary over a range of RNA transcription rates that reflect RNA-transcriptional protein ratios that roughly correspond to the charge balance conditions (Figure 5F). The same qualitative results are recapitulated in 3D simulations (Figure S5C) as well as simulations employing the Flory-Huggins free-energy (Figure S5D), and further reinforced by partition ratios computed from simulations (Figure S5E). Overall, our results suggest a model wherein low effective rates of RNA synthesis (or low transcription activity) stabilize transcriptional condensates while higher rates promote condensate dissolution.



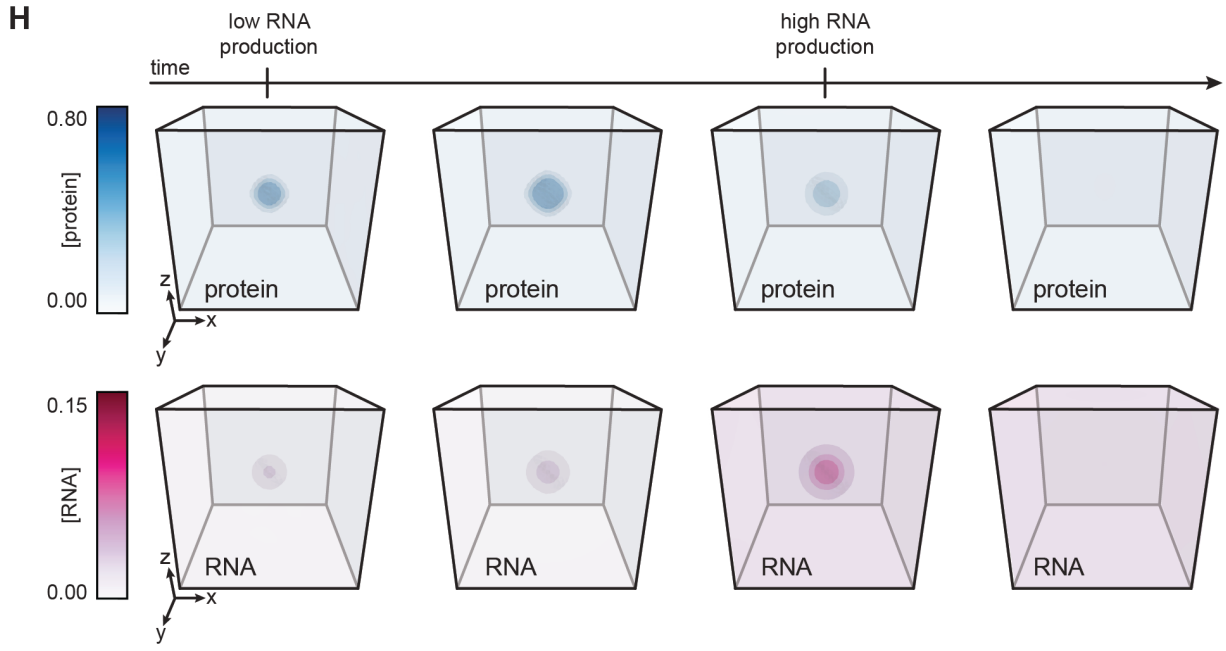
5E,F. Simulation predictions of transcriptional condensate radius (E) and lifetime (F) at varying effective rates of RNA synthesis (abcissa) (2D simulation grid). The radius values are normalized by the radius at k_p=0.01 and the dashed line in F represents lifetime of stable condensates, which are presented in units of simulation time (STAR Methods).

We then investigated the extent to which non-equilibrium effects underlying transcription regulate transcriptional condensate dynamics. RNA synthesis, degradation, and diffusion influence the spatial distribution of RNA, which in turn, may feedback on transcriptional condensates. To explore this, we varied the diffusivity of RNA and the effective rates of RNA synthesis and degradation, while holding the ratio of synthesis and degradation rates constant. The latter constraint ensures that the overall RNA concentration is constant in the condensate as other parameters are varied, thus any effect on condensate dynamics arises from purely non-equilibrium effects. Varying the parameters that control RNA synthesis/degradation rates and diffusion changes the relative time-scales of these processes (t_r and t_d, respectively) (STAR Methods), which in turn, influences the spatial distribution of RNA in the condensate. If diffusion is slower than synthesis/degradation ($t_r < t_d$), then RNA will accumulate near transcription sites, leading to a higher local RNA concentration in the condensate. Conversely, if diffusion is faster than synthesis/degradation ($t_r > t_d$), then RNA will diffuse away from transcription sites, leading to a lower uniform RNA concentration in the condensate. The spatial distribution of RNA will impact condensates according to local charge balance. To study how varying spatial distributions of RNA affect transcriptional condensates, we simulated conditions where the overall RNA concentration was fixed close to the charge-balance condition, thus promoting condensate formation at equilibrium. In these simulations, condensates that are stable when synthesis/degradation is slower than diffusion ($t_r > t_d$) dissolve when RNA synthesis/degradation is faster than diffusion (Figure 5G). When ($t_r > t_d$), RNA concentration is relatively uniform and low throughout the condensate and equilibrium effects dominate. Conversely, when ($t_r < t_d$), RNA is distributed non-uniformly with high local concentrations in the condensate and non-equilibrium effects dominate to result in condensate dissolution (Figure 5G). In the latter case, the localized high RNA concentrations exceed the charge balance condition due to non-equilibrium effects. Approximate estimates for the rates of RNA synthesis, degradation, and diffusion under physiological conditions ($t_d/t_r \approx$5-50, STAR Methods) suggest that transcriptional condensate dynamics are likely driven far off equilibrium.



5G. Variation of normalized condensate radius (y-axis, normalized to radius at k_p=0.01) with changing relative time-scales of reaction and diffusion (abcissa) (2D simulation grid). In these simulations, the total effective concentration of RNA produced is held constant (see text). The inset figure graphs the distribution of RNA concentrations at early simulation times (t=100) for two different values of t_D/t_R (highlighted in the main panel with corresponding colors).

We sought to synthesize our results so far to explore the effect of non-equilibrium dynamics on regulating transcriptional condensates across transcription initiation and productive elongation. Simulations were started at a relatively low effective rate of RNA synthesis, mimicking initiation, followed by an increase to a relatively high effective rate of RNA synthesis, mimicking productive elongation. The simulations predict that low effective rates of RNA synthesis enhance condensate formation, and these condensates subsequently dissolve upon ensuing higher effective rates of RNA synthesis (Figure 5H). These results suggest that non-equilibrium processes underlying RNA synthesis can potentially regulate the formation and dissolution of transcriptional condensates.

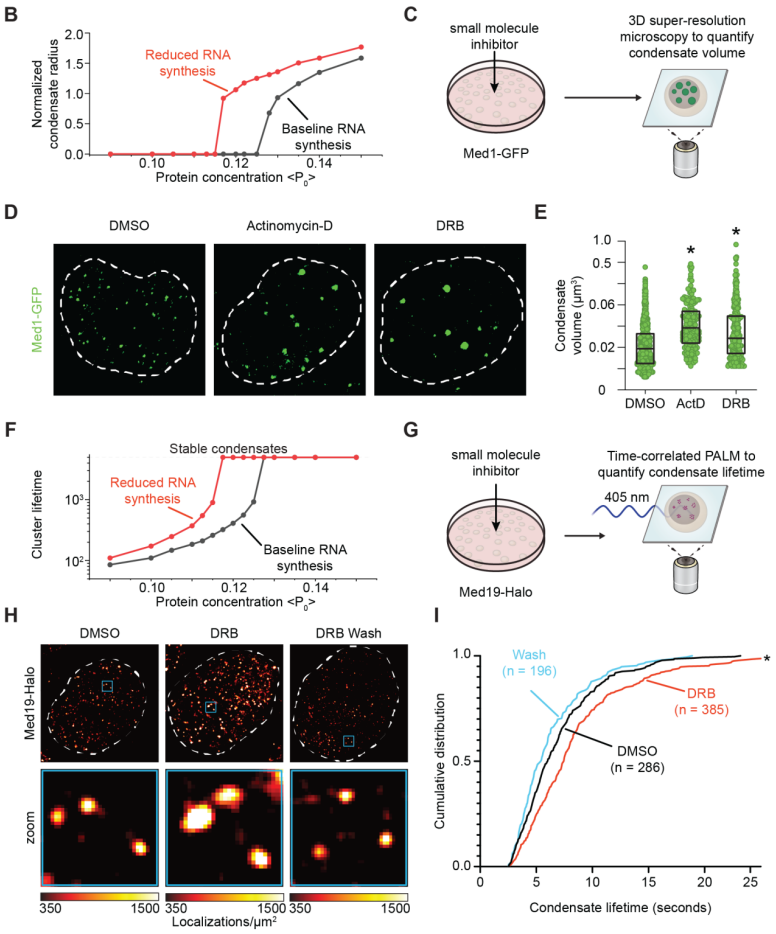


5H. Visualization of protein (blue) and RNA (magenta) concentration fields over simulation time for 3D simulations. The condensate is initialized (first panel) and then grows under low transcriptional activity (second panel). After a finite-time (t_{sim}=1000), the effective rate of RNA synthesis (k_p) is increased by 2.5-fold, which in turn, drives condensate shrinkage (third panel) and ultimately, dissolution (fourth panel) (STAR Methods).

Inhibition of RNA elongation leads to enhanced condensate size and lifetime in cells

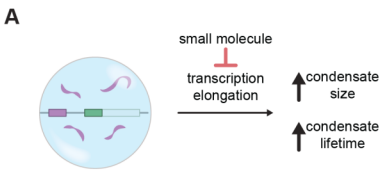
Transcriptional condensates in cells are highly dynamic, forming and dissolving at timescales ranging from seconds to minutes (Cho et al., 2018). We previously showed that condensate formation is associated with transcription activation and initiation (Cho et al., 2018). Once transcriptional condensates are formed, the RNA-mediated condensate dissolution model predicts that inhibition of elongation should increase the

size and lifetime of transcriptional condensates (Figure 6A). We used the physics-based model (Figure 5) to simulate the effects of elongation inhibition on transcriptional condensates and performed experiments to test the predictions from these simulations in cells (Figures 6B-6I). In order to account for the locus-dependent ability to recruit the transcriptional machinery and Pol II, we performed these simulations at a range of total protein concentrations (as in Figures 5D and 5E), but for conditions where the effective rate of RNA synthesis (k_p) was high (corresponding to elongation) and low (corresponding to inhibited elongation). The results of the simulations predict that a reduced effective rate of RNA synthesis should increase the size and lifetime of transcriptional condensates across a range of total protein concentrations (Figure 6B and 6F).



6H. Representative heatmap of Med19-Halo localizations in single nucleus upon addition of transcriptional inhibitor DRB, DRB wash or DMSO control.

6I. Cumulative distribution frequency plot of condensate lifetime in response to indicated treatments are shown ($p < 0.0001$, one-way ANOVA).



6A. Scheme for preventing condensate dissolution upon transcriptional burst by treatment with small molecules that inhibit transcriptional elongation.

6B. Simulation predictions show variation of normalized condensate radius with total protein amount (abscissa) in absence (black, k_p=0.1) and presence (red, k_p=0.05) of RNA synthesis inhibition (2D simulation grid). The radius is normalized by the radius at k_p=0.05,<P_0>=0.115.

6C. Experimental design to test the effect of transcriptional inhibition on the size of Mediator condensates. MED1-GFP mESCs are imaged by 3D super-resolution microscopy after treatment with small molecules.

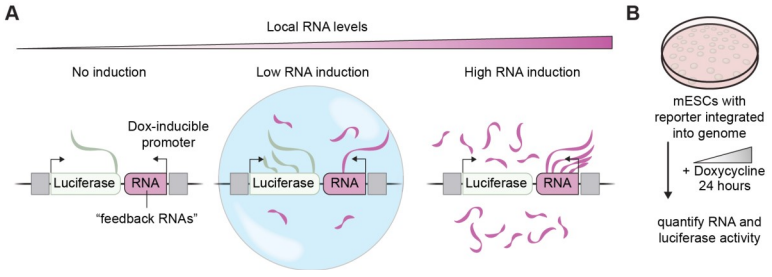
6D. Max intensity projection images of single nuclei tagged with endogenous Med1-GFP in the presence of indicated transcriptional inhibitors or DMSO control.

6E. Quantification of the volume of Med1-GFP condensates in (D). p-value for DMSO vs. AcD < 0.0001 and p-value for DMSO vs. DRB < 0.0001, one-way ANOVA.

6F. Simulation predictions show variation of condensate lifetime with total protein amount (abscissa) in absence (black, k_p=0.1) and presence (red,k_p=0.05) of RNA synthesis inhibition (2D simulation grid). The lifetime is presented in units of simulation time.

6G. Experimental design to test the effect of DRB on the lifetime of Mediator condensates in Med19-tagged mESCs. Lifetimes are quantified by time-correlated PALM.

To experimentally test these predictions from the simulations, mESCs engineered with an endogenous, GFP-tagged subunit of Mediator (Med1-GFP) (Sabari et al., 2018) were treated for 30 minutes with Actinomycin-D or DRB (Figure 6C), which disrupt transcription elongation through DNA intercalation and inhibition of CDK9-mediated Pol II pause release, respectively (Singh and Padgett, 2009; Sobell, 1985; Steurer et al., 2018). Consistent with the model predictions, after inhibition of elongation, Med1-GFP condensates increased in volume by ~2-fold as measured by 3D super-resolution microscopy (Figures 6D and 6E). Condensate lifetime could not be assessed in these cells due to the long duration of image acquisition and consequent photobleaching, so we turned to time-correlated PALM super-resolution microscopy (tcPALM) in mESCs with an endogenous Med19-Halo tag (Cho et al., 2018; Cisse et al., 2013) to investigate the effects of elongation inhibition on condensate lifetime (Figure 6G). Cells were treated for 30 minutes with DRB to disrupt transcription elongation, and the lifetime of Med19 condensates was quantified. When transcription elongation was inhibited by DRB treatment, Med19 condensates exhibited significantly longer lifetimes than mock-treated cells (Figures 6H and 6I), and when DRB-treated cells were washed with fresh media, the lifetimes of the Med19 condensates recovered to those of the mock-treated condition (Figures 6H and 6I). Taken together, the in silico and experimental results show that suppression of elongation in cells leads to increased condensate size and lifetime, consistent with the model that a burst of RNA synthesis can promote dissolution of transcriptional condensates in cells.



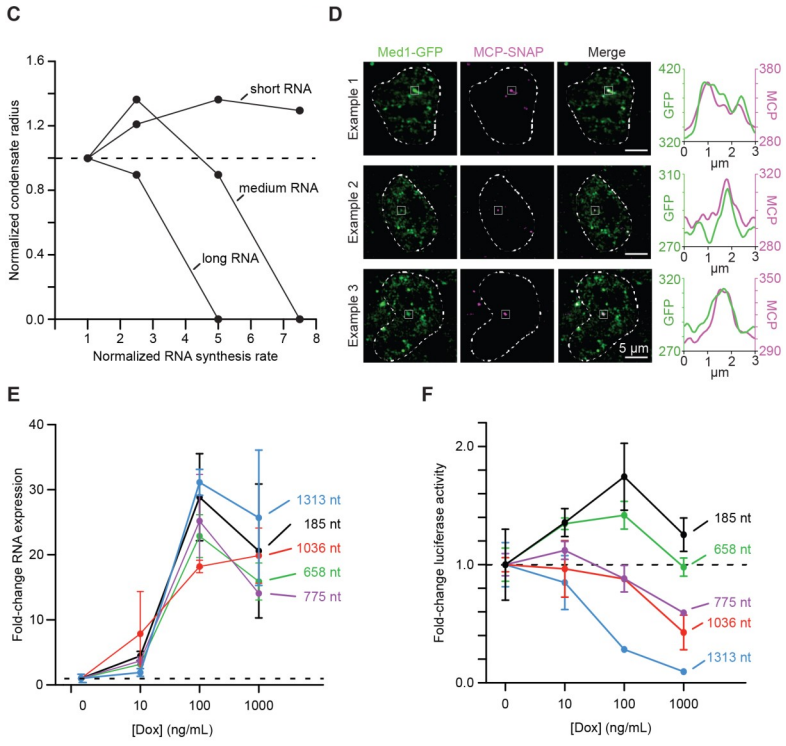
7A. Scheme depicting the effect of increasing local RNA levels on transcriptional condensates and transcriptional output in the indicated reporter system. In this system, local RNA expression near a luciferase reporter gene can be induced by doxycycline.

7B. The experimental design to test the effect of increasing local RNA levels on the reporter gene expression.

Increasing the levels of local RNA synthesis reduces transcription in cells

The RNA-mediated feedback model suggests that modifying the concentration or size of RNA molecules should have a predictable effect on transcriptional output. We developed complementary experimental and

simulation approaches (Figures 7A-7C) whereby the levels of putative "feedback RNAs" could artificially be increased. We first used the physics-based model (Figure 5) to simulate the effect of increasing effective rates of RNA synthesis as well as varying lengths for the synthesized RNA on condensates (STAR Methods). The simulations predicted that increases in the production rate of shorter RNAs initially enhance and subsequently suppress transcriptional condensate size, while increases in the production rate of longer RNAs lead to reduced condensate size with increasing synthesis rates (Figure 7C).



7C. Simulations predict the variation of condensate size with increasing effective rates of RNA synthesis (abscissa) (2D simulation grid). The condensate radius is normalized by value at rate=1 and RNA synthesis rates are normalized to k_p=0.02 (STAR Methods)

7D. Live-cell imaging showing localization of Mediator condensates and MS2-tagged RNA expressed near the reporter gene. Med1-GFP mESCs that have an integrated reporter system are transfected with MCP-SNAP to visualize MS2-tagged RNA. The line analysis of the diagonal of the depicted box is plotted on the right to demonstrate colocalization. Three independent examples are shown.

7E. Quantification of RNA levels near luciferase reporter by qRT-PCR with increasing dox concentrations with various RNA species. Markers show the mean of at least 3 replicates and error bars depict the S.D.

7F. Quantification of luciferase luminescence with increasing dox concentrations with various RNA species. Markers show the mean of at least 3 replicates and error bars depict the S.D.

To test this prediction, we investigated the effect of artificially increasing the levels of feedback RNAs on the transcription of an adjacent luciferase reporter gene in cells (Figure 7A and 7B) (Kirk et al., 2018). DNA molecules specifying RNAs of a range of sizes were cloned into this system to allow dox-inducible expression of these RNAs, and mESC lines were generated with clones of integrated constructs. Feedback RNAs were observed at loci of Mediator puncta under low dox stimulation, suggesting that these actively transcribed genes are associated with transcriptional condensates (Figure 7D). The cell lines were then treated with increasing doses of doxycycline to induce expression of diverse feedback RNAs (Figure 7E) and reporter expression was measured by

luminescence (Figure 7F). The results were consistent with model predictions (Figure 7C): increases in the levels of short feedback RNAs initially enhanced reporter expression and then suppressed this, while progressive increases in the levels of the longer feedback RNAs more strongly reduced reporter expression (Figure 7F).

DISCUSSION

The results described here indicate that transcription is a non-equilibrium process that provides dynamic feedback through its RNA product. The results support a model whereby RNA provides both positive and negative feedback on transcription via the regulation of transcriptional condensates. In this model, low levels of short RNAs produced during transcription initiation enhance formation of transcriptional condensates, while high levels of the longer RNAs produced during elongation promote condensate dissolution.

RNA-mediated feedback regulation of transcription is the result of coupling between the non-equilibrium processes of RNA synthesis, degradation, and diffusion with an underlying equilibrium phase behavior of RNA-transcriptional protein mixtures that exhibit a reentrant phase transition. Such phase transitions have been observed in prior studies of complex coacervation in mixtures of oppositely charged polyelectrolyte solutions, including RNA–polyelectrolyte mixtures (Lin et al., 2019; Overbeek and Voorn, 1957; Sing, 2017; Srivastava and Tirrell, 2016). We first present several lines of evidence to show that mixtures of RNA and transcriptional molecules undergo a reentrant phase transition at equilibrium. In droplet formation assays, low levels of RNA that occur at gene regulatory regions can stimulate condensate formation by the Mediator coactivator whereas high levels suppress condensates (Figure 2). When these experiments were repeated with the MED1-IDR, a disordered component of Mediator that contributes to transcriptional condensates, the effects of varying RNA molecules on droplets fit expectations for the charge balance model (Figure 3).

Using a physics-based model, we then studied how non-equilibrium processes of RNA synthesis, degradation, and diffusion are linked to equilibrium reentrant phase behavior to regulate the size and dynamics of transcriptional condensates in vivo. In agreement with predictions from the model, the dependence of these quantities and transcriptional

output on RNA synthesis rates and lengths were then positively tested in cell-free and in cellular systems (Figures 4-7). Together, our results suggest that the coupling of non-equilibrium processes inherent to RNA transcription with a phenomenon akin to complex coacervation plays an important role in regulating transcriptional condensates and their output in vivo. Previous theoretical efforts have explored how non-equilibrium processes may be coupled to condensate formation (Weber et al., 2019; Zwicker et al., 2017). Our study provides a framework to understand how non-equilibrium regulation of condensates is important for a specific biological process, namely transcription.

An RNA-mediated feedback model for transcriptional regulation provides a potential explanation for the roles of enhancer and promoter-associated RNAs, which are evolutionarily conserved features of eukaryotes. These low-abundance short RNAs, transcribed bidirectionally from enhancers and promoters, have been reported to affect transcription from their associated genes through diverse postulated mechanisms (Andersson et al., 2014; Catarino and Stark, 2018; Core et al., 2014; Gardini and Shiekhattar, 2015; Henriques et al., 2018; Lai et al., 2013; Li et al., 2016; Mikhaylichenko et al., 2018; Nair et al., 2019; Pefanis et al., 2015; Rahnamoun et al., 2018; Schaukowitch et al., 2014; Scruggs et al., 2015; Sigova et al., 2015; Smith et al., 2019). The diversity of sequences present in these short RNA species has made it difficult to postulate a common molecular mechanism for their effects on transcription. In this context, a model for RNA-mediated feedback regulation of condensates is attractive for several reasons. RNA molecules are known components of other biomolecular condensates, including the nucleolus, nuclear speckles, paraspeckles and stress granules, where they are known to play regulatory roles (Fay and Anderson, 2018). RNA is a powerful regulator of condensates that are formed by electrostatic forces because it has a high negative charge density due to its phosphate backbone (Drobot et al., 2018; Frankel et al., 2016), thus explaining why the effects of diverse RNAs on transcriptional condensates are sequence-independent.

Recent studies indicate that transcription occurs in periodic bursts (~1-10 minutes in duration), where multiple molecules of Pol II can be released from promoters within a short timeframe and produce multiple molecules of mRNA (~1-100 molecules per burst) (Cisse et al., 2013; Fukaya et al., 2016; Larsson et al., 2019). Multiple models explain such

periodic bursts through stochastic gene activation events (Chen and Larson, 2016; Larsson et al., 2019; Raj et al., 2006; Rodriguez and Larson, 2020; Suter et al., 2011; Tunnacliffe and Chubb, 2020) but are often agnostic to the underlying mechanism or attribute these to rate-limiting transcription factor binding events. We suggest that a rapid and spatially -localized change in charge balance, due to increased RNA synthesis at pause release of active Pol II, may contribute to dissolution of transcriptional condensates and thus dynamic loss of the pool of transcriptional apparatus in those condensates. This would provide a means to provide negative feedback to arrest transcription and a mechanism that may contribute to the dynamic bursty behavior observed for transcription.

REFERENCES

Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat Rev Genet *13*, 720–731.

Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B., and Sharp, P.A. (2013). Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature *499*, 360–363.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Andrews, J.O., Conway, W., Cho, W.-K., Narayanan, A., Spille, J.-H., Jayanth, N., Inoue, T., Mullen, S., Thaler, J., and Cissé, I.I. (2018). qSR: a quantitative super-resolution analysis tool reveals the cell-cycle dependent organization of RNA Polymerase I in live human cells. Scientific Reports *8*, 1–10.

Aumiller, W.M., and Keating, C.D. (2016). Phosphorylation-mediated RNA/peptide complex coacervation as a model for intracellular liquid organelles. Nature Chemistry *8*, 129–137.

Azofeifa, J.G., Allen, M.A., Hendrix, J.R., Read, T., Rubin, J.D., and Dowell, R.D. (2018). Enhancer RNA profiling predicts transcription factor activity. Genome Res *28*, 334–344.

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

Banerjee, P.R., Milin, A.N., Moosa, M.M., Onuchic, P.L., and Deniz, A.A. (2017). Reentrant Phase Transition Drives Dynamic Substructure Formation in Ribonucleoprotein Droplets. Angewandte Chemie International Edition *56*, 11354–11359.

Bergot, M.O., Diaz-Guerra, M.J., Puzenat, N., Raymondjean, M., and Kahn, A. (1992). Cis-regulation of the L-type pyruvate kinase gene promoter by glucose, insulin and cyclic AMP. Nucleic Acids Res *20*, 1871–1877.

Boeynaems, S., Holehouse, A.S., Weinhardt, V., Kovacs, D., Lindt, J.V., Larabell, C., Bosch, L.V.D., Das, R., Tompa, P.S., Pappu, R.V., et al. (2019). Spontaneous driving forces give rise to protein–RNA condensates with coexisting phases and complex material properties. PNAS *116*, 7889–7898.

Boija, A., Klein, I.A., Sabari, B.R., Dall'Agnese, A., Coffey, E.L., Zamudio, A.V., Li, C.H., Shrinivas, K., Manteiga, J.C., Hannett, N.M., et al. (2018). Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. Cell *175*, 1842 -1855 e16.

Brandman, O., and Meyer, T. (2008). Feedback Loops Shape Cellular Signals in Space and Time. Science *322*, 390–395.

Bruhat, A., Jousse, C., Carraro, V., Reimold, A.M., Ferrara, M., and Fafournoux, P. (2000). Amino Acids Control Mammalian Gene Transcription: Activating Transcription Factor 2 Is Essential for the Amino Acid Responsiveness of the CHOP Promoter. Mol Cell Biol *20*, 7192–7204.

Cambridge, S.B., Gnad, F., Nguyen, C., Bermejo, J.L., Krüger, M., and Mann, M. (2011). Systems-wide Proteomic Analysis in Mammalian Cells Reveals Conserved, Functional Protein Turnover. J. Proteome Res. *10*, 5275–5284.

Catarino, R.R., and Stark, A. (2018). Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. Genes Dev. *32*, 202–223.

Chen, H., and Larson, D.R. (2016). What have single-molecule studies taught us about gene expression? Genes Dev. *30*, 1796–1810.

Chen, W., Smeekens, J.M., and Wu, R. (2016). Systematic study of the dynamics and half-lives of newly synthesized proteins in human cells. Chem. Sci. *7*, 1393–1400.

Chiu, A.C., Suzuki, H.I., Wu, X., Mahat, D.B., Kriz, A.J., and Sharp, P.A. (2018). Transcriptional Pause Sites Delineate Stable Nucleosome-Associated Premature Polyadenylation Suppressed by U1 snRNP. Molecular Cell *69*, 648-663.e7.

Cho, W.-K., Jayanth, N., English, B.P., Inoue, T., Andrews, J.O., Conway, W., Grimm, J.B., Spille, J.-H., Lavis, L.D., Lionnet, T., et al. (2016). RNA Polymerase II cluster dynamics predict mRNA output in living cells. ELife *5*, e13617.

Cho, W.-K.K., Spille, J.-H.H., Hecht, M., Lee, C., Li, C., Grube, V., Cisse, I.I., Lee, C., Hecht, M., Cho, W.-K.K., et al. (2018). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. Science *361*, 412–415.

Chubb, J.R., Trcek, T., Shenoy, S.M., and Singer, R.H. (2006). Transcriptional Pulsing of a Developmental Gene. Current Biology *16*, 1018–1025.

Cisse, I.I., Izeddin, I., Causse, S.Z., Boudarene, L., Senecal, A., Muresan, L., Dugast-Darzacq, C., Hajj, B., Dahan, M., and Darzacq, X. (2013). Real-Time Dynamics of RNA Polymerase II Clustering in Live Human Cells. Science *341*, 664–667.

Core, L., and Adelman, K. (2019). Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. Genes Dev. *33*, 960–982.

Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat Genet *46*, 1311–1320.

Cramer, P. (2019). Organization and regulation of gene transcription. Nature *573*, 45–54.

Dignam, J.D., Lebovitz, R.M., and Roeder, R.G. (1983). Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. Nucleic Acids Res *11*, 1475–1489.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Drobot, B., Iglesias-Artola, J.M., Le Vay, K., Mayr, V., Kar, M., Kreysing, M., Mutschler, H., and Tang, T.-Y.D. (2018). Compartmentalised RNA catalysis in membrane-free coacervate protocells. Nature Communications *9*, 3643–3643.

Dunlap, J.C. (1999). Molecular Bases for Circadian Clocks. Cell *96*, 271–290.

Ebert, B.L., and Bunn, H.F. (1999). Regulation of the erythropoietin gene. Blood *94*, 1864–1877.

Elowitz, M.B., and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. Nature *403*, 335–338.

Ewels, P.A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M.U., Di Tommaso, P., and Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. Nature Biotechnology *38*, 276–278.

Fay, M.M., and Anderson, P.J. (2018). The role of RNA in biological phase separations. J Mol Biol *430*, 4685–4701.

Flores, O., Lu, H., and Reinberg, D. (1992). Factors involved in specific transcription by mammalian RNA polymerase II. Identification and characterization of factor IIH. J. Biol. Chem. *267*, 2786–2793.

Flory, P.J. (1942). Thermodynamics of High Polymer Solutions. J. Chem. Phys. *10*, 51–61.

Forero-Quintero, L., Raymond, W., Handa, T., Saxton, M., Morisaki, T., Kimura, H., Bertrand, E., Munsky, B., and Stasevich, T. (2020). Live-cell imaging reveals the spatiotemporal organization of endogenous RNA polymerase II phosphorylation at a single gene. BioRxiv 2020.04.03.024414.

Frankel, E.A., Bevilacqua, P.C., and Keating, C.D. (2016). Polyamine/Nucleotide Coacervates Provide Strong Compartmentalization of Mg2+, Nucleotides, and RNA. Langmuir *32*, 2041–2049.

Fukaya, T., Lim, B., and Levine, M. (2016). Enhancer Control of Transcriptional Bursting. Cell *166*, 358–368.

Gardini, A., and Shiekhattar, R. (2015). The many faces of long noncoding RNAs. The FEBS Journal *282*, 1647–1657.

Gardner, T.S., Cantor, C.R., and Collins, J.J. (2000). Construction of a genetic toggle switch in Escherichia coli. Nature *403*, 339–342.

Gu, B., Swigut, T., Spencley, A., Bauer, M.R., Chung, M., Meyer, T., and Wysocka, J. (2018). Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements. Science *359*, 1050–1055.

Guo, Y.E., Manteiga, J.C., Henninger, J.E., Sabari, B.R., Dall'Agnese, A., Hannett, N.M., Spille, J.-H., Afeyan, L.K., Zamudio, A.V., Shrinivas, K., et al. (2019). Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. Nature *572*, 543–548.

Guyer, J.E., Wheeler, D., and Warren, J.A. (2009). FiPy: Partial Differential Equations with Python. Comput. Sci. Eng. *11*, 6–15.

Henriques, T., Scruggs, B.S., Inouye, M.O., Muse, G.W., Williams, L.H., Burkholder, A.B., Lavender, C.A., Fargo, D.C., and Adelman, K. (2018). Widespread transcriptional pausing and elongation control at enhancers. Genes Dev. *32*, 26–41.

Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., and Sharp, P.A. (2017). A Phase Separation Model for Transcriptional Control. Cell *169*, 13–23.

Hohenberg, P.C., and Halperin, B.I. (1977). Theory of dynamic critical phenomena. Rev. Mod. Phys. *49*, 435–479.

Jain, A., and Vale, R.D. (2017). RNA Phase Transitions in Repeat Expansion Disorders. Nature *546*, 243–247.

Jangi, M., and Sharp, P.A. (2014). Building Robust Transcriptomes with Master Splicing Factors. Cell *159*, 487–498.

Jin, Y., Eser, U., Struhl, K., and Churchman, L.S. (2017). The Ground State and Evolution of Promoter Region Directionality. Cell *170*, 889-898.e10.

Kardar, M. (2007). Statistical physics of fields (Cambridge: Cambridge Univ. Press).

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res. *32*, D493-496.

Kim, T.-K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182–187.

Kirk, J.M., Kim, S.O., Inoue, K., Smola, M.J., Lee, D.M., Schertzer, M.D., Wooten, J.S., Baker, A.R., Sprague, D., Collins, D.W., et al. (2018). Functional classification of long non-coding RNAs by k-mer content. Nature Genetics *50*, 1474–1482.

Lahav, G., Rosenfeld, N., Sigal, A., Geva-Zatorsky, N., Levine, A.J., Elowitz, M.B., and Alon, U. (2004). Dynamics of the p53-Mdm2 feedback loop in individual cells. Nat Genet *36*, 147–150.

Lai, F., Orom, U.A., Cesaroni, M., Beringer, M., Taatjes, D.J., Blobel, G.A., and Shiekhattar, R. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature *494*, 497–501.

Larsson, A.J.M., Johnsson, P., Hagemann-Jensen, M., Hartmanis, L., Faridani, O.R., Reinius, B., Segerstolpe, Å., Rivera, C.M., Ren, B., and Sandberg, R. (2019). Genomic encoding of transcriptional burst kinetics. Nature *565*, 251–254.

LeRoy, G., Rickards, B., and Flint, S.J. (2008). The Double Bromodomain Proteins Brd2 and Brd3 Couple Histone Acetylation to Transcription. Molecular Cell.

Li, W., Notani, D., and Rosenfeld, M.G. (2016). Enhancers as non-coding RNA transcription units: recent insights and future perspectives. Nat Rev Genet *17*, 207–223.

Lin, Y., McCarty, J., Rauch, J.N., Delaney, K.T., Kosik, K.S., Fredrickson, G.H., Shea, J.-E., and Han, S. (2019). Narrow equilibrium window for complex coacervation of tau and RNA under cellular conditions. ELife *8*, e42571.

Maiuri, P., Knezevich, A., De Marco, A., Mazza, D., Kula, A., McNally, J.G., and Marcello, A. (2011). Fast transcription rates of RNA polymerase II in human cells. EMBO Rep *12*, 1280–1285.

Mikhaylichenko, O., Bondarenko, V., Harnett, D., Schor, I.E., Males, M., Viales, R.R., and Furlong, E.E.M. (2018). The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. Genes Dev *32*, 42–57.

Milin, A.N., and Deniz, A.A. (2018). Reentrant Phase Transitions and Non-Equilibrium Dynamics in Membraneless Organelles. Biochemistry *57*, 2470–2477.

Monod, J., and Jacob, F. (1961). General Conclusions: Teleonomic Mechanisms in Cellular Metabolism, Growth, and Differentiation. Cold Spring Harb Symp Quant Biol *26*, 389–401.

Mountain, G.A., and Keating, C.D. (2020). Formation of Multiphase Complex Coacervates and Partitioning of Biomolecules within them. Biomacromolecules *21*, 630–640.

Muthukumar, M. (2016). Electrostatic correlations in polyelectrolyte solutions. Polym. Sci. Ser. A *58*, 852–863.

Nair, S.J., Yang, L., Meluzzi, D., Oh, S., Yang, F., Friedman, M.J., Wang, S., Suter, T., Alshareedah, I., Gamliel, A., et al. (2019). Phase separation of ligand-activated enhancers licenses cooperative chromosomal enhancer assembly. Nature Structural & Molecular Biology *26*, 193–203.

Oparin, A. (1953). The origin of life.

Orphanides, G., LeRoy, G., Chang, C.-H., Luse, D.S., and Reinberg, D. (1998). FACT, a Factor that Facilitates Transcript Elongation through Nucleosomes. Cell *92*, 105–116.

Overbeek, J.T.G., and Voorn, M.J. (1957). Phase separation in polyelectrolyte solutions. Theory of complex coacervation. Journal of Cellular and Comparative Physiology *49*, 7–26.

Pak, C.W., Kosno, M., Holehouse, A.S., Padrick, S.B., Mittal, A., Ali, R., Yunus, A.A., Liu, D.R., Pappu, R.V., and Rosen, M.K. (2016). Sequence Determinants of Intracellular Phase Separation by Complex Coacervation of a Disordered Protein. Molecular Cell *63*, 72–85.

Pefanis, E., Wang, J., Rothschild, G., Lim, J., Kazadi, D., Sun, J., Federation, A., Chao, J., Elliott, O., Liu, Z.-P., et al. (2015). RNA Exosome-Regulated Long Non-Coding RNA Transcription Controls Super-Enhancer Activity. Cell *161*, 774–789.

Peran, I., and Mittag, T. (2020). Molecular structure in biomolecular condensates. Current Opinion in Structural Biology *60*, 17–26.

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

Rahnamoun, H., Lee, J., Sun, Z., Lu, H., Ramsey, K.M., Komives, E.A., and Lauberth, S.M. (2018). RNAs interact with BRD4 to promote enhanced chromatin engagement and transcription activation. Nature Structural & Molecular Biology *25*, 687–697.

Raj, A., and van Oudenaarden, A. (2008). Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. Cell *135*, 216–226.

Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., and Tyagi, S. (2006). Stochastic mRNA Synthesis in Mammalian Cells. PLoS Biol *4*.

Rodriguez, J., and Larson, D.R. (2020). Transcription in Living Cells: Molecular Mechanisms of Bursting. Annual Review of Biochemistry *89*, null.

Roeder, R.G. (2019). 50+ years of eukaryotic transcription: an expanding universe of factors and mechanisms. Nature Structural & Molecular Biology *26*, 783–791.

Sabari, B.R., Dall'Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A.V., Manteiga, J.C., et al. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. Science *361*, eaar3958.

Schaukowitch, K., Joo, J.-Y., Liu, X., Watts, J.K., Martinez, C., and Kim, T.-K. (2014). Enhancer RNA Facilitates NELF Release from Immediate Early Genes. Molecular Cell *56*, 29–42.

Scruggs, B.S., Gilchrist, D.A., Nechaev, S., Muse, G.W., Burkholder, A., Fargo, D.C., and Adelman, K. (2015). Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. Molecular Cell *58*, 1101–1112.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent Transcription from Active Promoters. Science *322*, 1849 –1851.

Sellick, C.A., and Reece, R.J. (2003). Modulation of transcription factor function by an amino acid: activation of Put3p by proline. EMBO J *22*, 5147–5153.

Shin, Y., and Brangwynne, C.P. (2017). Liquid phase condensation in cell physiology and disease. Science *357*, eaaf4382–eaaf4382.

Shrinivas, K., Sabari, B.R., Coffey, E.L., Klein, I.A., Boija, A., Zamudio, A.V., Schuijers, J., Hannett, N.M., Sharp, P.A., Young, R.A., et al. (2019). Enhancer Features that Drive Formation of Transcriptional Condensates. Molecular Cell *75*, 549-561.e7.

Sigova, A.A., Abraham, B.J., Ji, X., Molinie, B., Hannett, N.M., Guo, Y.E., Jangi, M., Giallourakis, C.C., Sharp, P.A., and Young, R.A. (2015). Transcription factor trapping by RNA in gene regulatory elements. Science *350*, 978–981.

Sing, C.E. (2017). Development of the modern theory of polymeric complex coacervation. Advances in Colloid and Interface Science *239*, 2–16.

Singh, J., and Padgett, R.A. (2009). Rates of in situ transcription and splicing in large human genes. Nat Struct Mol Biol *16*, 1128–1133.

Smith, K.N., Miller, S.C., Varani, G., Calabrese, J.M., and Magnuson, T. (2019). Multimodal Long Noncoding RNA Interaction Networks: Control Panels for Cell Fate Specification. Genetics *213*, 1093–1110.

Sobell, H.M. (1985). Actinomycin and DNA transcription. Proc Natl Acad Sci U S A *82*, 5328–5331.

Srivastava, S., and Tirrell, M.V. (2016). Polyelectrolyte complexation. Advances in Chemical Physics 499–544.

Steurer, B., Janssens, R.C., Geverts, B., Geijer, M.E., Wienholz, F., Theil, A.F., Chang, J., Dealy, S., Pothof, J., van Cappellen, W.A., et al. (2018). Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA Polymerase II. Proc Natl Acad Sci USA *115*, E4368–E4376.

Strom, A.R., and Brangwynne, C.P. (2019). The liquid nucleome – phase transitions in the nucleus at a glance. J Cell Sci *132*.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat. Struct. Mol. Biol. *14*, 103–105.

Suter, D.M., Molina, N., Gatfield, D., Schneider, K., Schibler, U., and Naef, F. (2011). Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. Science *332*, 472–474.

Tunnacliffe, E., and Chubb, J.R. (2020). What Is a Transcriptional Burst? Trends in Genetics *36*, 288–297.

Umbarger, H.E. (1956). Evidence for a Negative-Feedback Mechanism in the Biosynthesis of Isoleucine. Science *123*, 848–848.

Weber, C.A., Zwicker, D., Jülicher, F., and Lee, C.F. (2019). Physics of active emulsions. Rep. Prog. Phys. *82*, 064601.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell *153*, 307–319.

Zamudio, A.V., Dall'Agnese, A., Henninger, J.E., Manteiga, J.C., Afeyan, L.K., Hannett, N.M., Coffey, E.L., Li, C.H., Oksuz, O., Sabari, B.R., et al. (2019). Mediator Condensates Localize Signaling Factors to Key Cell Identity Genes. Molecular Cell.

Zhang, P., Shen, K., Alsaifi, N.M., and Wang, Z.-G. (2018). Salt Partitioning in Complex Coacervation of Symmetric Polyelectrolytes. Macromolecules *51*, 5586–5593.

Zwicker, D., Seyboldt, R., Weber, C.A., Hyman, A.A., and Jülicher, F. (2017). Growth and division of active droplets provides a model for protocells. Nature Physics *13*, 408–413.

Chapter 4: *Evolution*

Weak cooperative interactions for biological specificity

"What men are poets who can speak of Jupiter if he were a man, but if he is an immense spinning sphere of methane and ammonia must be silent?"
— Richard Feynman

This chapter is primarily based on work *published* in
Evolution of weak cooperative interactions for biological specificity.
Gao, A.[=], **Shrinivas, K[=].,** Lepeudry, P., Suzuki, H.I., Sharp, P.A., and Chakraborty, A.K.
(2018). Proc. Natl. Acad. Sci. *115*, E11053–E11060.

Functional specificity in biology is broadly mediated by two classes of mechanisms, "lock-key" interactions, and multivalent weak cooperative interactions (WCI). Despite growing evidence that WCIs are widely prevalent in higher organisms[2], little is known about the selection forces that drove their evolution and repeated positive selection for mediating biological specificity in metazoans. We report that multivalent WCI for mediating biological specificity evolved as the number of tasks that organisms had to perform with functional specificity became large (e.g., multicellular organisms). We find that the evolution of multivalent WCI confers enhanced and robust evolvability to organisms, and thus it has been repeatedly positively selected. Thus, we provide new insights on the evolution of WCI, and more broadly, on the evolution of evolvability.

[2] Including our previous efforts in Chapters 1-3 on role of WCIs in condensate formation

Introduction

Living organisms have evolved to perform diverse tasks with functional specificity using different mechanism[3] (Kirschner et al., 1998, 2006). Unlike highly specific enzymes that have structured recognition domains, many proteins have intrinsically disordered regions (IDRs) that do not fold into ordered structures (van der Lee et al., 2014). These proteins often mediate specific biological outcomes through multivalent weak cooperative interactions (WCI). For example, the highly disordered protein histone HP1-alpha binds to its chaperone prothymosin with specificity (Borgia et al., 2018) to enable chaperone function. This specificity is obtained not through structured "lock and key" interactions, but through multiple cooperative interactions based on coarse-grained associations of short tracks of amino acids of certain lengths and charge patterns, lack of aromatic side-chains[4], etc. Many cytoplasmic proteins contain multiple recognizable domains (such as SH2 and SH3) which contain low-affinity motifs in disordered backgrounds, which regulate specific biological outcomes via multivalent WCI (van der Lee et al., 2014; Su et al., 2016). Proteins with IDRs that interact through such interactions are common in liquid-like condensates (Banani et al., 2017; Brangwynne et al., 2015; Shin and Brangwynne, 2017) that form in the cytoplasm and the nucleus to mediate specific biological functions by compartmentalizing particular biochemical pathways.

[3] Read (Kirschner et al., 2006) for a fascinating overview of how evolution adapts and repurposes material to perform widely different tasks and respond to different environments. This book, and the perspective cited in-text, are truly seminal pieces of work.

[4] For more examples of this type of "coarse-grained" statistical matching, refer (Lin et al., 2017; Staller et al., 2017)

The most common and rapidly evolving molecular feature of biological

systems is changes in gene regulation. In prokaryotes, transcription is regulated by proteins that bind to promoters with high sequence specificity. In mammalian cells, activation of RNA Pol II at the transcription initiation site frequently depends on the binding of multiple proteins to distal non-coding DNA elements called enhancers. It is widely appreciated that the number of enhancers and their constituents change rapidly during evolution (Villar et al., 2015) and that this variation is critical for functional and morphological differences. Many enhancer binding proteins exhibit specificity of binding to a particular enhancer because of cooperative interactions with other proteins(Stampfel et al., 2015). Approximately 52% of DNA and 44% of RNA binding proteins in humans contain IDRs greater than 50 amino acids in length (nearly two-fold more common than in the entire proteome). Many of these have been shown to form liquid-like condensates at high concentrations (Banani et al., 2017), and at lower concentrations when mixed with RNA (Langdon et al., 2018; Maharana et al., 2018) . Clusters of enhancer elements in close physical proximity, known as super-enhancers, regulate the transcription of genes important for maintaining cell identity (Hnisz et al., 2013; Whyte et al., 2013). Recent evidence (Hnisz et al., 2017; Sabari et al., 2018) suggests that multivalent WCI between transcription factors, co-activators, and other transcriptional machinery result in their accumulation at genes regulated by SEs by forming a phase separated condensate. Because of the cooperative nature of phase transitions, this phenomenon occurs when upstream signals, valency of interactions, or concentration exceed a sharp threshold (i.e., with functional specificity).

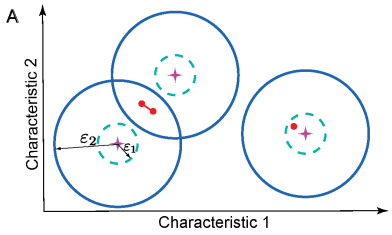
Specificity mediated by multivalent WCI is more prevalent in organisms that have evolved more recently (Kirschner et al., 1998). Examples that highlight this evolutionary trend include the observation that the fraction of the proteome containing IDRs is higher in more recently evolved organisms (Schlessinger et al., 2011), gene regulation in mammals versus prokaryotes noted above, and pathogen recognition mediated by multivalent WCI in vertebrate adaptive immunity (Flajnik and Kasahara, 2010). Other examples of WCI can be found in signal transduction pathways, extracellular matrix variation, and various cytoskeletal processes (Kirschner et al., 1998).

Despite these observations, little is known about the selection forces that drove the predominance of multivalent WCI in mediating biological

specificity in more recently evolved organisms. Here, we develop an easily interpretable model that is applicable to a broad class of biological processes and systems and use it to simulate evolution on a computer. Our results provide important new insights into why WCI evolved, why it has been repeatedly selected across metazoa, and more generally on the evolution of evolvability.

Model development & Methods

We consider a population of organisms that evolve as the number of tasks that they need to perform in order to function properly increases. Each organism has a number of genes, and the corresponding gene products can potentially perform the tasks. We ignore epistatic interactions between genes, but different gene products can potentially cooperate to perform functions together as described below.



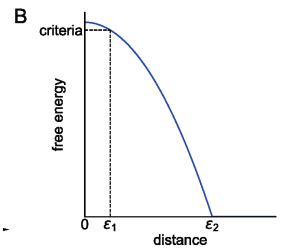
1A schematic depiction of the space that represents the gene products of organisms and the tasks that they need to perform to function properly. Each axis describes a particular characteristic of a task or matching characteristic in a gene product that determines their interactions (see text). Tasks are shown as stars and gene products as red dots. When a task and a gene product are within a distance equal to ε_1 , the gene product performs the corresponding task with high specificity. When a task and a gene product are within a distance equal to ε_2, the gene product performs the corresponding task incompletely. When two gene products have closely matched interaction characteristics, they can act cooperatively (indicated with a line connecting them above), to perform tasks together (see text).

The tasks that organisms must perform with functional specificity can be quite complex (e.g., gene regulation, stress responses, immune responses, etc.), but they are considered to be predicated on protein-protein recognition. Thus, our model is based on interactions between gene products and the tasks that they must perform. In performing a task with functional specificity the protein-protein interaction could have lock-key characteristics or be mediated by multivalent WCI between proteins. Inspired by models of protein-protein interactions where a few characteristics determine interaction strengths (Perelson and Oster, 1979), each task and gene product is associated with specific values of a set of characteristics important for their interactions. The value of each relevant characteristic (e.g., hydrophobicity) of a particular task is represented by its position on an axis (Fig. 1A). Thus, each task is specified by its positions on different axes that describe each characteristic; i.e., by the position of the task in the space spanned by the axes corresponding to interaction characteristics. For brevity, hereafter we will refer to this space as "characteristic space". The gene products that perform the tasks are also represented by positions in a characteristic space which describes interaction characteristics that match those that define the tasks. For example, if one axis in the task characteristic space corresponds to hydrophobicity of the tasks, the corresponding axis in the characteristic space in which gene products are represented define the latter's func-

tional hydrophobicity. Alternatively, if a particular axis in the characteristic space for tasks represents positive charge, the corresponding axis in the characteristic space for gene products represents negative charge. The position of each gene product in its characteristic space is specified by how well matched each of its interaction characteristics is with respect to the characteristics that define the tasks. So, given a set of interaction characteristics defining tasks, there is a known mapping between the task characteristic space and that in which the gene products are represented. Using this mapping, every axis in the gene product characteristic space can be made to coincide with the corresponding axis in the task characteristic space. For example, for charges, the two axes would coincide upon reversing the sign of the axis in the gene product characteristic space. So, in our model, we assume that the mapping has been applied and thus the closer a gene product and a task are on the same axis the more matched they are with respect to the corresponding characteristic, thus contributing to a favorable interaction. Considering all the characteristics together, the closer a task and a gene product are in characteristic space (Fig. 1A) the more favorable their functional interaction.

In order to construct a general model applicable to diverse examples where WCI have evolved to mediate specificity, we do not specify the particular characteristics that define the tasks. They could be different for each example, and given the way we have defined the model our results would still be applicable. The number of axes corresponds to the number of characteristics required to describe the protein-protein interactions that predicate tasks being performed. We assume that the number of axes needed is not large, since the strength of protein-protein interactions is usually determined by a small number of key relevant quantities (charge, charge distribution, hydrophobicity, etc.). Our qualitative results are insensitive to the particular choice of a finite number of axes (see SI Appendix, Fig. S9-11).

The fitness of an organism depends on how well its gene products perform the tasks. If a gene product is within a short distance,ε_1 , of a task (Fig. 1A), it is considered to perform this task with functional specificity via strong interactions. If the distance between a task and a gene product is within a larger distance,ε_2, then the task is considered to be done less completely via weak interactions. If a gene product is located a distance



1B The free energy of interaction between a task and a gene product is defined to be a function of the distance between a task and a single gene product as shown in the graph. The interaction free energy is parabolic when the task-gene distance is less than ε_2, and becomes 0 when the distance is larger than . As defined in the text, for cooperating gene products, their free energies with a given task are added up.

further away from the task than, then the interactions are too weak for the task to be done by this gene product.

If two or more gene products are within a short distance, ε_3, from each other, they can interact with each other and potentially act cooperatively to complete a task with functional specificity even though each gene product interacts weakly with the task (i.e., via multivalent WCI). The free energy of interaction between a task and a gene product is considered to be a function of distance as shown in Fig. 1B. We model the cooperative action of gene products within a distance ε_3 from each other by adding up their interaction free energies corresponding to a task (mathematical details in SI Appendix supplementary text). If the resulting number exceeds the value of the free energy corresponding to a distance (ε_1) of for a single gene product interacting with a task, then the gene products are considered to perform the task with functional specificity. Thus, multiple gene products can cooperatively perform a task with specificity via multivalent WCI if their interaction free energies with the task add up to be at least as favorable as that corresponding to a single gene product that performs a task in a lock-key fashion (located within a distance equal to ε_1 of the task).

Given a set of tasks, we define a function, F_j, for organism j, as follows:

$$F_j = \lambda_1 (M - \#\text{tasks done by } j) + \lambda_2 (M - \#\text{tasks done with functional specificity by } j) + \lambda_3 G^j \quad (1)$$

where G^j is the number of genes in organism j and M is the number of tasks to be performed for proper function. The fitness of organism j is defined as $f_j=\exp(-F_j)$. The first term in Eq. 1 makes organisms that perform the tasks, at least poorly, have a higher fitness than those that do not. The second term makes organisms that perform more tasks with functional specificity more fit. The third term makes organisms with bigger genomes less fit than their peers. The quantities λ_1, λ_2, and λ_3 represent the relative weights of these three factors, or selection forces, in determining an organism's fitness.

We initiate the evolutionary dynamics with a single task that must be

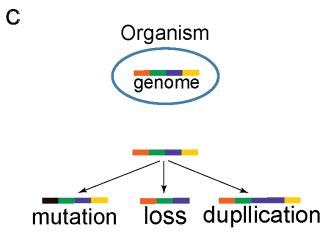
performed and each organism in the population has a single gene. The gene product of each organism is assigned to a randomly chosen point in characteristic space. The organisms evolve by mutation, gene duplication, and gene loss (Fig. 1C). Recombination is unlikely to affect the qualitative behavior of the model unless the recombination rate is unusually large. We are concerned here with the evolution of WCI as a mechanism for functional specificity, and this mechanism is more prevalent in higher organisms where horizontal gene transfer is less important. Therefore, we also do not consider horizontal gene transfer.

The organisms evolve according to standard Wright-Fisher evolutionary dynamics with a fixed number of organisms (N) in the population (Fig. 1D). At each time step, the genome of every organism can potentially undergo mutation, gene duplication, and gene loss. When a gene mutates, the location of its gene product in characteristic space is changed by translating it in a randomly chosen direction by a random distance whose average value is ε_1. The mutation rate is chosen such that, on average, in every organism, one gene is likely to mutate every two time steps. Gene duplication occurs at one tenth the rate of mutation. A duplicated gene makes a gene product that occupies exactly the same location in characteristic space as its copy. With time, the two genes, and hence their gene products, can diverge from each other and potentially perform different tasks (or functions). Gene loss occurs at the same rate as duplication. After mutation, gene duplication and loss are attempted with the probabilities specified above, each organism can acquire a potentially new genome, with new coordinates in characteristic space for its gene products (see Fig. 1D).

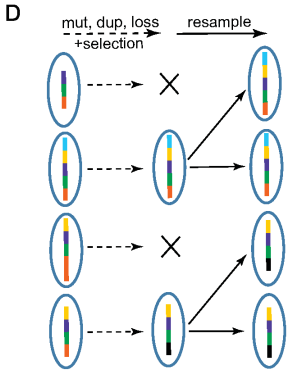
The probability that an organism will produce a progeny (or be positively selected) in this evolutionary time step is then calculated as follows:

$$p_s^j = \frac{f_j}{\Sigma_{j=1}^{N} f_j} \qquad (2)$$

where p_s^j is the probability that organism j will be present in the next time step of evolution. After this selection step, the number of organisms that produce a progeny is likely to be less N than because some organ-



1C Schematic depiction of the processes of gene mutation, loss, and duplication included in the evolutionary model. For example, in this schematic the orange gene has mutated to black, the yellow gene is lost, and the purple gene is duplicated.



1D Depiction of the model for evolutionary dynamics (only one generation of evolution is depicted.)

isms die without producing progeny as they are not sufficiently fit. To keep the population size constant as per Wright-Fisher dynamics, we rescale the total number of organisms to remain equal to N when the next time step begins (Fig. 1D). The proportion of organisms with a particular genome is kept the same as before rescaling (i.e. after selection). This evolutionary process continues in subsequent time steps. The stochastic processes described above are simulated using a Monte-Carlo computational procedure.

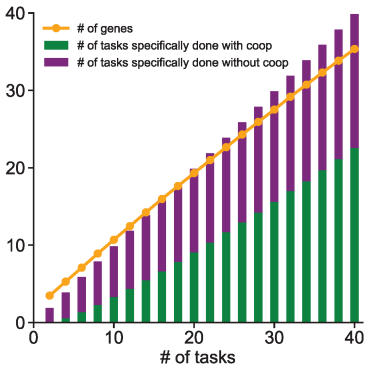
We characterize the system using the following variables: 1] the average number of genes in an organism in the population; 2] the number of tasks completed with functional specificity by an organism via WCI involving multiple gene products, averaged across the population; 3] the number of tasks completed with functional specificity by single gene products in an organism via lock-key interactions, averaged across the population. We carry out the evolutionary dynamics until a "steady state" is reached with respect to these variables; i.e., the system ceases to evolve further because a fitness peak has been reached (see SI Appendix, Fig. S1). We then introduce a new task in characteristic space (the value of M increases by one in Eq. 1) and carry out the evolutionary dynamics again until steady state, starting from the state of the organisms that were evolutionary fit for the previous tasks. Thus, the evolutionary history of the organisms is explicitly incorporated. This process is repeated as new tasks are introduced. Thus we study whether, and why, mechanisms for regulating specificity evolve as organisms have to perform more tasks specifically in order to function properly (e.g., as multicellular organisms became more complex). An important variable is the extent to which the newly introduced task is correlated, or similar, to the existing tasks. We have studied several cases that are described in the results.

The parameters in the model are λ_1, λ_2,λ_3, ε_1 ,ε_2 ,ε_3 and the extent to which newly introduced tasks are correlated with the existing ones. Based on parameter sensitivity studies (see SI Appendix, Figs. S2-S7), we note that the qualitative results that we report are robust as long as λ_1 and λ_2 are greater than λ_3. If λ_3 becomes too large, the introduction of new genes leads to severe fitness penalties. So, when the number of tasks that must be performed for proper function becomes large, the organisms prefer to have reduced fitness by not functioning properly (i.e., not completing the necessary tasks) rather than evolve new genes. This is tantamount to

being unable to evolve more complex multicellular organisms, and so we do not consider this case further. The values of the parameters used to obtain the results discussed below are $\varepsilon_1 = \varepsilon_3$(which equals the size of a single mutation step in our model), $\varepsilon_2 = 5\varepsilon_1$, $\lambda_1 = \lambda_2 = 1$, and $\lambda_3 = 0.1$. The dependence of the results on changing the value of λ_1 and ε_2 will be discussed below. Choosing ε_1 to be the same as the size of a single mutation implies that the condition for functional specificity via lock-key fit is stringent.

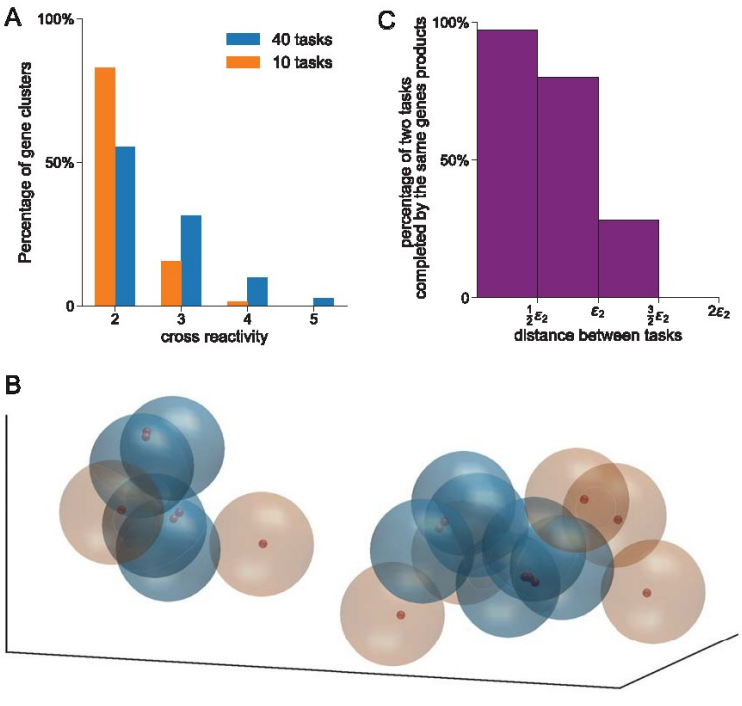
Results

We first studied a situation wherein each new task is introduced at a randomly chosen location in characteristic space that is at a distance equal to $1.8\varepsilon_2$ away from any one of the tasks that had to be previously performed. So, in terms of its interaction characteristics, the newly introduced task has some similarity with previous tasks. Our simulation results (Fig. 2) show that WCI evolve as a mechanism for mediating functional specificity as the number of tasks that organisms have to perform in order to function properly increases (or organisms become more complex). Furthermore, as organisms evolve to perform more tasks, the proportion of the tasks that they carry out via WCI increases (Fig. 2). These results are consistent with the observation that this mechanism for mediating functional specificity is prevalent in multicellular organisms. One reason that WCI evolved as a mechanism for biological specificity is because this allows similar tasks to be performed with some of the same cooperating components, and therefore, the number of genes required for organisms to function properly becomes less than the number of tasks to be performed (Fig. 2). This is consistent with the observation that proteins with similar IDRs (and even the same proteins) are involved in regulating different genes, and in forming condensates at different super-enhancers. The same is true for components that form condensates to mediate other biological functions in the cytoplasm and the nucleus. We have carried out calculations with different levels of correlation between new and old tasks (i.e, values of task-task distance other than $1.8\varepsilon_2$), and the qualitative behavior of our model is unchanged (see SI Appendix, Fig. S4) unless the new tasks become totally uncorrelated.



2 Weak cooperative interactions (WCI) evolve as organisms become more complex. Variation of the average number of genes in organisms and the number of tasks specifically done via WCI between gene products as the number of tasks required for an organism to function properly increases (or organisms become more complex). The number of tasks performed by single gene products is also shown. When the number of tasks equals 10, 33% of tasks are done via WCI, and when the number of tasks equals 40 this proportion is 56%. Three characteristics describe the interaction characteristics of tasks and gene products.

One implication of the results described so far is that as a greater proportion of tasks are performed via WCI (as the number of tasks increases), the extent to which gene products are cross-reactive to multiple tasks also increases. The results in Fig. 3A show that this is indeed the case. But, the cross-reactivity is limited to similar tasks. This can be seen clearly by considering a situation where a newly introduced task can either be closely related to one of the previous tasks or not. If new tasks that are related to at least one previous task are introduced more frequently than tasks that are unrelated (75% chance for a new task to be at a distance $1.8\varepsilon_2$ away from a previous task and 25% chance to be at least at a distance $3\varepsilon_2$ away from all previous tasks), the tasks will be distributed in characteristic space as disjoint groups of related tasks (Fig. 3B). One group may correspond to regulation of gene transcription, another could be signaling through SH2/SH3 domains in the cytoplasm, etc. Fig 3B illustrates that gene products that act via WCI are cross-reactive to a limited set of tasks that are closely related. Quantitatively, the number of tasks that are performed by the same gene products acting cooperatively rapidly declines as the interaction characteristics of the tasks become less related (Fig. 3C).



3(**A**) Variation of the extent of cross-reactivity with the evolution of WCI. The x -axis shows the number of tasks done by the same cluster of gene products, and the y-axis is the percentage of such clusters that are preforming 2, 3, 4 and 5 tasks in this cross reactive fashion.

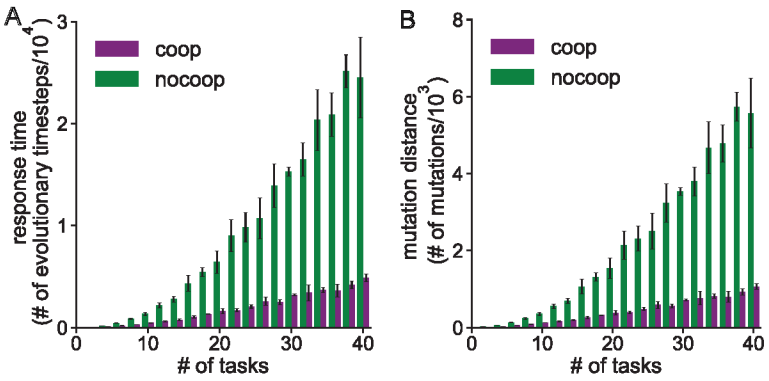
3(**B**) Snapshot of simulation results when new tasks are introduced such that they are either closely related to tasks from an earlier era or not. Two modules of such related tasks are depicted in characteristics space. Large spheres with radius ε_2 are drawn around each task. Brown spheres show tasks being performed by single gene products, blue spheres show closely related tasks being performed by clusters of cooperating gene products. Small spheres correspond to gene products.

3(**C**): The percentage of two tasks completed by the same gene products is high only for related tasks. Three characteristics describe the interaction characteristics of tasks and gene products.

Some cross-reactivity for similar tasks is an inherent property of the above cooperative model, but the extent of cross-reactivity is limited as otherwise task specificity would be lost. In cells, other mechanisms can be coupled to multivalent WCI to limit cross-reactivity. For example, master transcription factors bind with lock-key type specificity to particular DNA binding sites. Only then can interactions between transcription factor IDRs and that of transcriptional coactivators, chromatin remodelers, and RNA Polymerase II occur through multivalent WCI if specific upstream signals have modified the IDRs to have a valency exceeding a threshold. However, the co-activators, chromatin remodelers etc. can exhibit some cross-reactivity (as in Fig. 3) to regulate related functions, such as genes bound by different master transcription factors. The degree of cross-reactivity could also be limited by topological barriers such as chromosomal domains or localization in sub-cellular compartments. However, the cross-reactivity that accompanies the evolution of WCI for biological specificity could, when altered by mutation or modification, cause serious pathologies. For example, proto-oncogenes can be activated when DNA rearrangements create a fusion protein that targets transcriptional activation domains into their vicinity (Hnisz et al., 2016; Rabbitts, 1994). Also, cellular states that generate abnormally large condensates (Aguzzi and Altmeyer, 2016) formed by multivalent WCI could sequester high levels of client proteins important for the normal functioning of other genes.

Multivalent WCI as a mechanism underlying biological specificity is prevalent in many organisms across metazoa. Thus, we wondered whether the emergence of this mechanism makes organisms more evolvable, thus explaining why it has been repeatedly positively selected and its more prominent role in multi-cellular organisms. The properties of more evolvable systems (Kirschner et al., 1998) include: 1] Reduced constraints in maintaining old functions when a new function has to be evolved and 2] Fewer mutations required to produce novel phenotypic traits. Thus, to explore this question, we calculated the time required for organisms to evolve to perform the tasks required for proper function after a new task is introduced. We compared the results of simulations of our model to one where WCI are not allowed; i.e., regardless of the similarity between the interaction characteristics of gene products, they are not allowed to act in concert to perform tasks with functional speci-

ficity. As Fig. 4 shows, the model allowing the evolution of WCI exhibits shorter response times and requires fewer mutations to respond to new tasks. Increasing values of ε_2 also make mutations increasingly less likely to be lethal when WCI are allowed (see SI Appendix, Fig. S8). This is because a deleterious mutation in any one gene product involved in mediating functional specificity via multivalent WCI is less likely to result in loss of function (task not performed) compared to the effect of a similar mutation for lock-key interactions. We conclude that the evolution of WCI for functional specificity confers increased and robust evolvability to organisms, as they can evolve to perform new tasks while maintaining old functions with fewer mutations and increased tolerance to deleterious mutations. This result is reinforced by a recent report demonstrating rapid evolution of human IDR proteins (Afanasyeva et al., 2018). Notice that evolvability emerges in our model without violating causality – i.e., this mechanism evolves based on past selection forces and not a pathological knowledge of the future.



4(**A**) The response time for the organisms to evolve to function properly after a new task is introduced is shown as a function of the number of tasks (or complexity). Results are shown for both the full model and one wherein cooperative interactions between gene products is not allowed.

4(**B**) The number of mutations (which includes gene mutation, duplication and loss) that the average organism needs to acquire to function properly after a new task is introduced is shown as a function of the number of tasks (or complexity). Results are shown for both the full model and one wherein cooperative interactions between gene products is not allowed.

The qualitative results that we have described hold if there is no fitness advantage associated with an organism performing a task poorly (λ_1 equals zero in Eq. 1). The only difference is that the response times for organisms to evolve to perform new tasks increase (see SI Appendix, Fig. S3). That is, the system becomes less evolvable if there is no fitness advantage for performing tasks poorly. This is because the lack of ability to be positively selected while performing tasks poorly constrains the mutational trajectories that have to be followed to perform new tasks while not abrogating old functions. A similar observation has been made

in laboratory experiments following the mutational pathway of a kinase as it evolves to catalyze a new substrate (Raman et al., 2016). Mutations are first observed in the kinase's allosteric pocket resulting in conformational flexibility that enables it to act on the old and new substrates suboptimally. Then, a mutation in the catalytic site is acquired to change specificity. A similar effect has also been described during the evolution of cross-reactive antibodies during germinal center reactions (Ovchinnikov et al., 2018). Our qualitative results are also robust to changes in the finite number of characteristics required to describe interactions (see SI Appendix, Fig. S9-11 and associated discussion).

Discussion

Biological systems carry out tasks with functional specificity. We have considered a model where the ability of an organism to carry out tasks is predicated on protein-protein recognition mediated by either lock-key or multivalent WCI. The ability of an organism to function properly, or its fitness, depends upon whether or not it can carry out a set of tasks with functional specificity (described by the first term in Eq. 1). We simulated the evolution of a population of such organisms as the number of tasks that need to be carried out for organisms to function properly progressively increases (larger values of M). The fitness landscape, or the genotype-phenotype relationship, changes as M increases. Thus, the organisms mutate to try to achieve a phenotype that is more fit – viz., a phenotype that can carry out the larger number of required tasks with functional specificity. Our results show that WCI emerges as a prominent mechanism for mediating specificity as organisms have to carry out larger number of tasks with specificity. We argue that this is the reason that WCI is more prominent in multicellular organisms. The evolution of WCI as a mechanism for mediating specific biological functions allows higher organisms to carry out diverse tasks with a relatively small genome (in our model genome size is constrained by the third term in Eq. 1).

Our model also shows that the emergence of WCI makes organisms more evolvable in that, as new tasks are introduced, the population of organisms can evolve to higher fitness phenotypes faster and with fewer mutations. Furthermore, as organisms mutate to try and acquire higher

fitness, mutations are less likely to be lethal after WCI emerges as a mechanism that mediates functional specificity. In other words, the fitness landscape describing the genotype-phenotype relationship becomes less rugged when WCI evolves. It has been noted previously (1,2) that the more exact or precise the requirements for function, the less evolvable the system is. Performing tasks with functional specificity mediated by WCI does not require the level of biochemical precision characteristic of lock-key interactions, and thus the system becomes more evolvable. We argue that WCI has been repeatedly positively selected in higher organisms because of the enhanced evolvability conferred by this mechanism for functional specificity. Indeed, the evolution of WCI may have given metazoans the great capacity for change whose consequences we observe today.

Our model is consistent with the observation that weak interactions have evolved to be highly relevant for gene regulation in metazoa. The IDRs of transcription factors and coactivators leverage WCI to drive condensate formation at regulatory elements to mediate transcription in higher organisms. This is in contrast to prokaryotes, where gene regulation is largely dictated by lock-key interactions that promote localization of TFs to specific promoter sequences. The biochemical rules for the WCI that determine interactions between IDRs is not as precise as specific enzyme-substrate interactions. Thus, the same IDRs can be employed to perform related functions, and IDRs can evolve readily with few mutations to regulate new functions. Thus, these motifs have been conserved in higher organisms. In the future it will be interesting to decipher the molecular grammar that determine WCI in these contexts.

A much higher fraction of proteins (Schlessinger et al., 2011) in multicellular organisms, as compared to prokaryotes, possess IDRs, and these IDRs are strongly enriched in factors controlling regulatory processes. These IDR regions, frequently in combination with RNA and/or DNA binding, provide some of the valency necessary to form condensates and concentrate factors in regulatory pathways (Langdon et al., 2018; Maharana et al., 2018). Recent analysis (Afanasyeva et al., 2018) of the rate of evolutionary change in the IDR regions suggests that they are more tolerant of mutational variation than regions with structured domains, but are nevertheless under genetic constrain. Since regulatory variation is thought to be the most rapidly changing aspect of evolutionary change

in multicellular organisms, it is perhaps not surprising that WCI are concentrated in these networks (Wilson et al., 2017).

The same type of reasoning probably explains the common observation that many regulatory RNA binding proteins in multicellular organisms possess limited sequence specificity (3–4 nucleotides), while the total sequence complexity of expressed coding and non-coding RNAs in cells is enormous. Similar examples can be found in signal transduction pathways, extracellular matrix variation, and various cytoskeletal elements (Kirschner et al., 1998). In their discussions about evolvability and facilitated variation (Kirschner et al., 1998, 2006), Kirschner and Gerhart anticipated WCI as an important aspect of multicellular biology, proposing that "weak linkage," compartmentalization, and redundancy contribute to constraint reduction, thus resulting in the robustness and observed regulatory variability in these organisms. Our model predicts the evolution of these characteristics (i.e., WCI) in organisms when they are challenged with new tasks, under constraints that limit the unbounded growth of the number of genes. Simply stated, these features—and evolvability—emerge organically from the known physical structures and interactions of proteins, RNA, and DNA on which the model is based.

This model describes many types of specific biological functions beyond gene regulation. For example, unlike more ancient organisms, vertebrates have an adaptive immune system that can mount pathogen-specific responses against a diverse and evolving world of microbe (Flajnik and Kasahara, 2010). The immune system is routinely faced with performing new tasks (recognize foreign pathogens not encountered previously) with functional specificity. One way it achieves this goal is to generate diverse receptors of B and T lymphocytes that interact with pathogenic markers. Importantly, functional specificity for particular pathogenic markers is achieved by the receptors via multivalent WCI (Kosmrlj et al., 2008; Murphy and Weaver, 2016; Perelson and Oster, 1979; Stadinski et al., 2016). The receptor on a particular lymphocyte commonly exhibits cross-reactivity to a few pathogen-derived ligands (Glanville et al., 2017). Some degree of cross-reactivity helps with recognizing a vast space of antigens, but pathogen specificity requires that responses are not too broadly cross-reactive. This limited cross-reactivity naturally emerges from our model as cross-reactivity is limited

to similar tasks.

Many studies have considered the evolution of modularity when there is a frequently changing environment (Espinosa-Soto and Wagner, 2010; Kashtan and Alon, 2005; Parter et al., 2007; Wagner and Altenberg, 1996). Modules are units with highly interconnected moieties that interact with other modules via very few interactions. Modularity make biological systems more evolvable (Clune et al., 2013) because the modules can be combined with each other differently to carry out new functions, much like subroutines in computer programs can be reused for different computations. Our focus has been on multivalent WCI where the participating components interact with each other via numerous weak interactions to mediate functional specificity while making organisms more evolvable.

The need to efficiently perform new tasks while retaining the ability to functionally execute previously learned tasks is common in many biological systems. For example, this is characteristic of learning by the nervous system. It may also be a characteristic of how computational machine learning algorithms trained on large data sets to predict specific outcomes could be adapted to predict new outcomes. We suspect that the fundamental aspects of the model we have described maybe relevant to these situations as well.

References

Afanasyeva, A., Bockwoldt, M., Cooney, C.R., Heiland, I., and Gossmann, T.I. (2018). Human long intrinsically disordered protein regions are frequent targets of positive selection. Genome Res. *28*, 975–982.

Aguzzi, A., and Altmeyer, M. (2016). Phase Separation: Linking Cellular Compartmentalization to Disease. Trends Cell Biol. *26*, 547–558.

Banani, S.F., Lee, H.O., Hyman, A.A., and Rosen, M.K. (2017). Biomolecular condensates: organizers of cellular biochemistry. Nat. Rev. Mol. Cell Biol. *18*, 285–298.

Boehm, T. (2012). Evolution of Vertebrate Immunity. Curr. Biol. *22*, R722–R732.

Boehm, T., Hirano, M., Holland, S.J., Das, S., Schorpp, M., and Cooper, M.D. (2018). Evolution of Alternative Adaptive Immune Systems in Vertebrates. Annu. Rev. Immunol *36*, 1–24.

Borgia, A., Borgia, M.B., Bugge, K., Kissling, V.M., Heidarsson, P.O., Fernandes, C.B., Sottini, A., Soranno, A., Buholzer, K.J., Nettels, D., et al. (2018). Extreme disorder in an ultrahigh-affinity protein complex. Nature.

Brangwynne, C.P., Tompa, P., and Pappu, R. V. (2015). Polymer physics of intracellular phase transitions. Nat. Phys. *11*, 899–904.

Clune, J., Mouret, J.-B., and Lipson, H. (2013). The evolutionary origins of modularity. Proceedings. Biol. Sci. *280*, 20122863.

Espinosa-Soto, C., and Wagner, A. (2010). Specialization can drive the evolution of modularity. PLoS Comput. Biol. *6*.

Flajnik, M.F., and Kasahara, M. (2010). Origin and evolution of the adaptive immune system: genetic events and selective pressures. Nat. Rev. Genet. *11*, 47–59.

Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L.E., Rubelt, F., Ji, X., Han, A., Krams, S.M., Pettus, C., et al. (2017). Identifying specificity groups in the T cell receptor repertoire. Nature *547*, 94–98.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-Enhancers in the Control of Cell Identity and Disease (Cell Press).

Hnisz, D., Weintrau, A.S., Day, D.S., Valton, A.L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., et al. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science *351*, 1454–1458.

Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., and Sharp, P.A. (2017). A Phase Separation Model for Transcriptional Control. Cell *169*, 13–23.

Kashtan, N., and Alon, U. (2005). Spontaneous evolution of modularity and network motifs. Proc. Natl. Acad. Sci. *102*, 13773–13778.

Kirschner, M., Gerhart, J., Otey, C.R., and Arnold, F.H. (1998). Evolvability. Proc. Natl. Acad. Sci. U. S. A. *95*, 8420–8427.

Kirschner, M., Gerhart, J.C., and Norton, J. (2006). The Plausibility of Life: Resolving Darwin's Dilemma (Yale University Press).

Kosmrlj, A., Jha, A.K., Huseby, E.S., Kardar, M., and Chakraborty, A.K. (2008). How the thymus designs antigen-specific and self-tolerant T cell receptor sequences. *105*, 16671–16676.

Langdon, E.M., Qiu, Y., Ghanbari Niaki, A., McLaughlin, G.A., Weidmann, C.A., Gerbich, T.M., Smith, J.A., Crutchley, J.M., Termini, C.M., Weeks, K.M., et al. (2018). mRNA structure determines specificity of a polyQ-driven phase separation. Science *360*, 922–927.

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T., et al. (2014). Classification of Intrinsically Disordered Regions and Proteins. Chem. Rev. *114*, 6589–6631.

Lin, Y.-H., Brady, J.P., Forman-Kay, J.D., and Chan, H.S. (2017). Charge pattern matching as a 'fuzzy' mode of molecular recognition for the functional phase separations of intrinsically disordered proteins. New J. Phys. *19*, 115003.

Maharana, S., Wang, J., Papadopoulos, D.K., Richter, D., Pozniakovsky, A., Poser, I., Bickle, M., Rizk, S., Guillén-Boixet, J., Franzmann, T.M., et al. (2018). RNA buffers the phase separation behavior of prion-like RNA binding proteins. Science *360*, 918–921.

Murphy, K., and Weaver, C. (2016). Janeway's immunobiology (Garland Science).

Ovchinnikov, V., Louveau, J.E., Barton, J.P., Karplus, M., and Chakraborty, A.K. (2018). Role of framework mutations and antibody flexibility in the evolution of broadly neutralizing antibodies. Elife *7*, e33038.

Parter, M., Kashtan, N., and Alon, U. (2007). Environmental variability and modularity of bacterial metabolic networks. BMC Evol. Biol. *7*, 1–8.

Perelson, A.S., and Oster, G.F. (1979). Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self-non-self discrimination. J. Theor. Biol. *81*, 645–670.

Rabbitts, T.H. (1994). Chromosomal translocations in human cancer. Nature *372*, 143–149.

Raman, A.S., White, K.I., and Ranganathan, R. (2016). Origins of Allostery and Evolvability in Proteins: A Case Study. Cell *166*, 468–480.

Sabari, B.R., Dall'Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A. V., Manteiga, J.C., et al. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. Science *361*.

Schlessinger, A., Schaefer, C., Vicedo, E., Schmidberger, M., Punta, M., and Rost, B. (2011). Protein disorder — a breakthrough invention of evolution? Curr. Opin. Struct. Biol. *21*, 412–418.

Shin, Y., and Brangwynne, C.P. (2017). Liquid phase condensation in cell physiology and disease. Science *357*, eaaf4382.

Stadinski, B.D., Shekhar, K., Gómez-Touriño, I., Jung, J., Sasaki, K., Sewell, A.K., Peakman, M., Chakraborty, A.K., and Huseby, E.S. (2016). Hydrophobic CDR3 residues promote the development of self-reactive T cells. Nat. Immunol. *17*, 946–955.

Staller, M. V, Holehouse, A.S., Swain-Lenz, D., Das, R.K., Pappu, R. V, and Cohen, B.A. (2017). A deep mutational scan of an acidic activation domain. BioRxiv 230987.

Stampfel, G., Kazmar, T., Frank, O., Wienerroither, S., Reiter, F., and Stark, A. (2015). Transcriptional regulators form diverse groups with context-dependent regulatory functions. Nature *528*, 147.

Su, X., Ditlev, J.A., Hui, E., Xing, W., Banjade, S., Okrut, J., King, D.S., Taunton, J., Rosen, M.K., and Vale, R.D. (2016). Phase separation of signaling molecules promotes T cell receptor signal transduction. Science *352*, 595–599.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J., et al. (2015). Enhancer Evolution across 20 Mammalian Species. Cell *160*, 554–566.

Wagner, G.P., and Altenberg, L. (1996). Perspective: Complex Adaptations and the Evolution of Evolvability. Evolution (N. Y). *50*, 967.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell *153*, 307–319.

Wilson, B.A., Foy, S.G., Neme, R., and Masel, J. (2017). Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. Nat. Ecol. Evol. *1*.

Zarin, T., Strome, B., Nguyen Ba, A.N., Alberti, S., Forman-Kay, J.D., and Moses, A.M. (2019). Proteome-wide signatures of function in highly diverged intrinsically disordered regions. Elife *8*.

Summary and perspectives

"It is amateurs who have one big bright beautiful idea that they can never abandon. Professionals know that they have to produce theory after theory before they are likely to hit the jackpot."

Francis Crick

In this thesis, we are primarily interested with the following question – how does our genetic material encode the rules the determine physiology and development? We propose a phase separation model of transcriptional control which provides a single conceptual framework to explain and describe diverse features of gene regulation.

Employing a physical model, we hypothesized (Chapter 1; (Hnisz et al., 2017)) that the condensation of transcriptional proteins by phase separation explains various puzzling observations (Dowen et al., 2014; Hnisz et al., 2013) underlying formation and function of super-enhancers. The same model provides a framework to interpret how enhancers remotely regulate their target genes[1] and is consistent with emerging evidence that direct physical contact or molecular bridging is not required for enhancer activity (Benabdallah et al.; Heist et al., 2019; El Khattabi et al., 2019). We then highlight experimental evidence, in efforts led by our collaborators, that provide direct evidence of formation of transcriptional condensates (Chapter 1, (Sabari et al., 2018)).

[1] Referred often as "spooky-action at a distance"

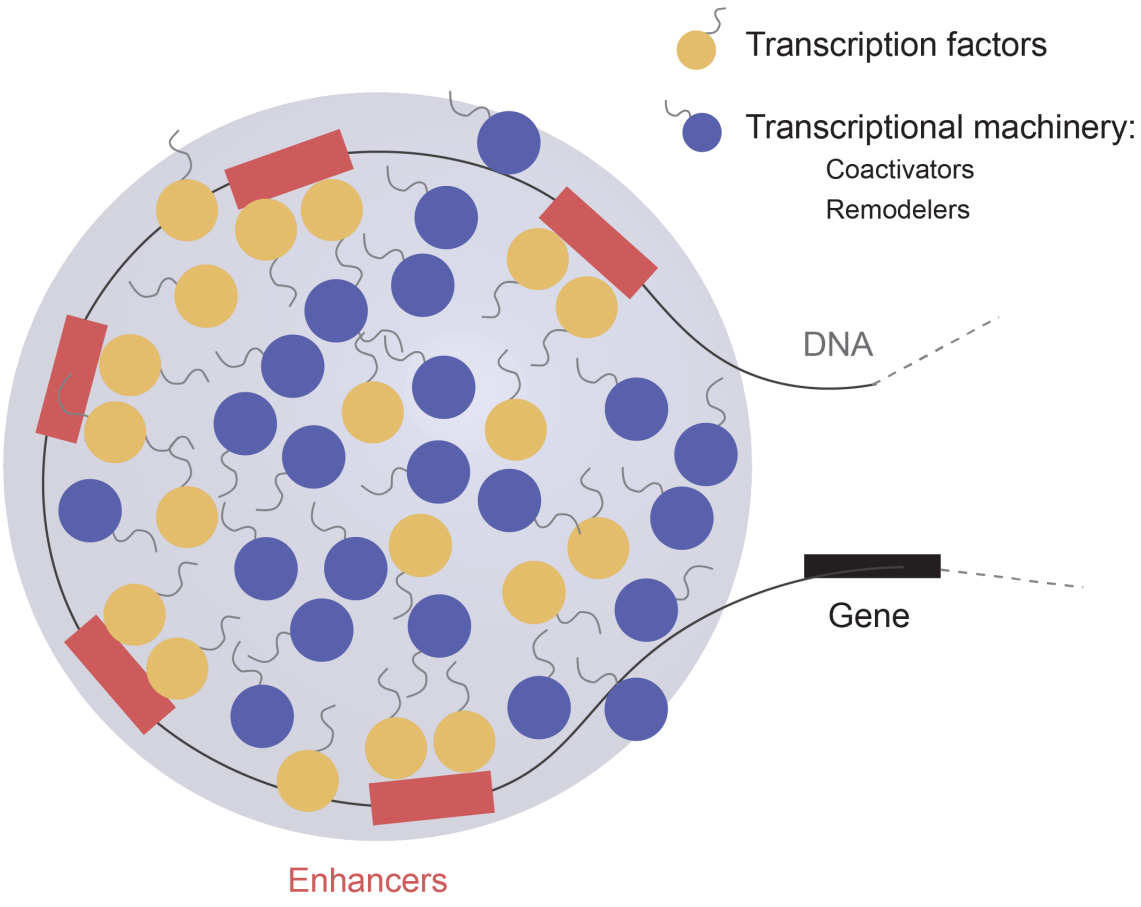
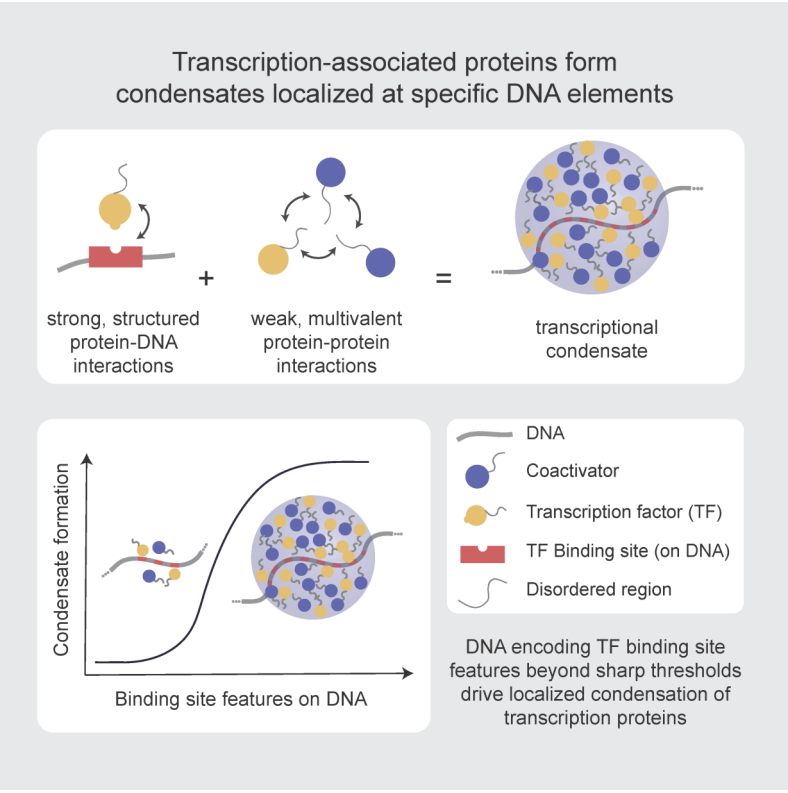


Figure 1 Model of a phase-separated assembly at super-enhancers (Chapter 1)

Experimental studies that validate our proposed model (Boija et al., 2018; Sabari et al., 2018) raise new questions: (1) how are transcriptional condensates organized at specific regions of our genome? (2) What are the nature of interactions that provide specificity to this process? (3) What is the role of DNA and are there features encoded in the genome that correlate with condensate formation? By combining molecular dynamics simulations, *in vitro* experiments, cell reporter assays, and bioinformatics analyses (Chapter 2; (Shrinivas et al., 2019)), we show that DNA elements encoding specific sequence features[2] beyond sharp thresholds can drive localized condensate formation[3]. We determine that a combination of specific, structured interactions[4] and weak, cooperative interactions[5] enables formation of condensates at specific loci. This chapter also discusses how localized condensation provides an attractive model for designation of enhancer activity along the non-coding genome and provides an additional regulatory axes[6].



[2] Such as high valency or abundance of high-affinity sites or clustering of binding sites for cognate transcription factors

[3] Thus shedding on light on Qs 1 & 3 defined above

[4] such as those between transcription factors and DNA)

[5] such as those between intrinsically disordered regions of transcriptional proteins. Incidentally, it has been long appreciated that many, if not most, transcription factors harbor large IDRs that facilitate gene activation – referred to as activation domains. How these unstructured domains contributed to activation was largely unknown, but is broadly consistent with the phase separation model

[6] Predicting which parts of our non-coding genome go on to become enhancers is a problem of great interest to the gene regulation community. There are many different known layers of regulation that contribute to this, including cooperativity at the molecular levels, nucleosome positions etc. See Introduction for Chapter 2 for a detailed description.

Figure 2 Genomic features that drive localized formation of transcriptional condensates (Chapter 2)

Having gained insight into the molecular interactions and correlates of transcriptional condensates, we next sought to explore how such assemblies are dynamically regulated in the cell? Here, we focused on exploring how synthesis of RNA, the energy-consuming non-equilibrium process that is the outcome of transcription, regulates transcriptional condensates. Motivated by recent studies on re-entrant phase transitions[7], we established that diverse RNA and transcriptional proteins largely fit a model of complex coacervation. By developing a dynamical model of non-equilibrium regulation and iterating with experiments in cells (Chapter 3), we predict that low levels of RNA produced during early stages of transcription[8] stimulate condensate formation (positive feedback) whereas the burst of RNAs produced during later stages[9] likely promotes dissolution (negative feedback). By connecting feedback regulation to condensate dynamics, our model provides additional insights (see Chapter 3 Discussion) on two long-standing mysteries in transcription: (1) What are the mechanisms of gene bursting i.e. production of RNA from genic DNA in discrete pulses? (2) Why are non-coding RNAs produced (and rapidly degraded) at enhancer loci?

[7] Such as (Banerjee et al., 2017; Lin et al., 2019; Milin and Deniz, 2018)

[8] Such as in initiation or non-coding RNAs at the enhancers

[9] Such as during transcriptional elongation – the process by which genic DNA is transcribed into the pre-mRNA.

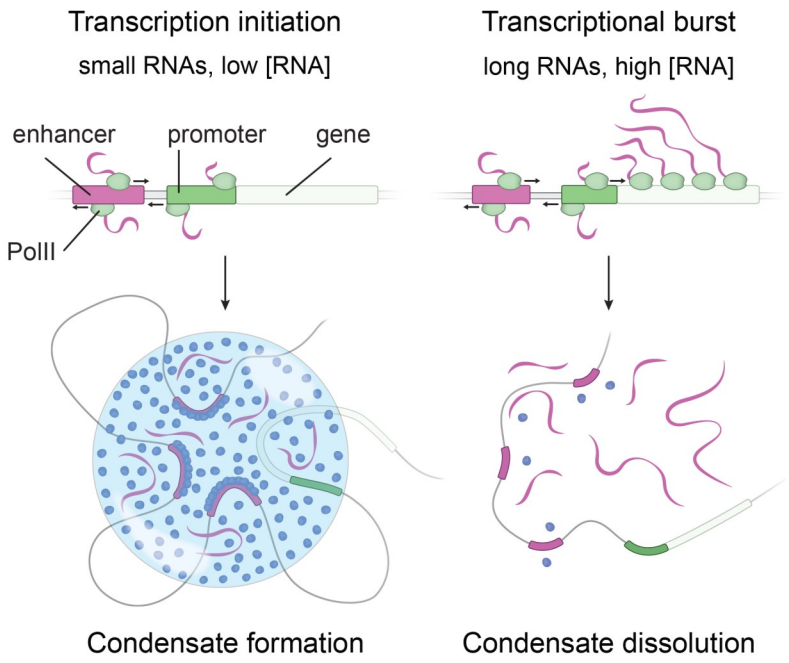


Figure 3 RNA-mediated non-equilibrium feedback control of transcriptional condensates (Chapter 3)

These studies on collective assemblies in transcription (Chapters 1-3 and (Boija et al., 2018; Guo et al., 2019; Hnisz et al., 2017; Sabari et al., 2018; Shrinivas et al., 2019)) inspired us to explore a more general question. Emil Fischer's classic model of finely-tuned lock-key type interactions between structured domains has ruled the roost in describing much of the molecular basis of biological specificity in the last century[10]. However, it is increasingly evident that another mechanism, one that relies on multivalent and weak cooperative interactions[11] (WCIs), is widely-prevalent in multi-cellular organisms[12]. Based on a computational model of evolution, we propose that mechanisms involving WCIs are likely to have become prominent as organisms needed to solve more tasks with specificity. Our model predicts that the emergence of WCIs leads to increased adaptability in organisms, pointing to an intriguing origin to the evolution of evolvability.

Together, this thesis combines approaches from theory, simulation, and experiment – refined in concert, to propose a role for phase separation in regulation of gene expression. Below, I'll provide perspectives on some promising areas of future investigation.

[10] Particularly for the class of the enzyme-substrate models and eventually led to the one gene - one enzyme – one specific task (Beadle and Tatum, 1941)

[11] Such as the ensemble of low-affinity interactions that drive condensate formation

[12] See an extraordinarily ahead-of-it's-time perspective by Kirschner (Kirschner et al., 1998) and a more accessible version on the topic of regulation by "weak linkage" (Kirschner et al., 2006)

Perspectives on some future avenues of work

The phase separation model of transcriptional control has garnered considerable attention from the community[13] and has evoked spirited debate on various aspects[14]. Below. I outline a few key areas of investigation which I believe hold promise for expanding our understanding of the condensate model of gene regulation.

[13] The experimental validation of our model was labeled as one of Top 10 scientific breakthroughs of 2018 by Science Magazine (Sabari et al., 2018)

[14] ranging from whether phase separation occurs *in vivo* in the transcriptional context, to understanding it's composition, the interplay between protein and polymer phase separation, and deciphering mechanisms of regulation

Searching for order in disorder

Intrinsically disordered regions (IDRs) of proteins are discussed oft in thesis (Chapters 1-4) and are gaining prominence in mediating many cellular processes (including condensate formation and dissolution) (Brangwynne et al., 2015; Choi et al., 2020; van der Lee et al., 2014). In the context of transcription, activation domains of transcription factors, historically recognized to be disordered and hard to crystallize[15] (Lech et

[15] Earning epithets such as acidic noodles

al., 1988; Ptashne and Gann, 1997; Struhl, 1988; Tjian and Maniatis, 1994) contributes to condensate formation through weak and multivalent interactions (Boija et al., 2018; Hahn, 2018; Shrinivas et al., 2019). Unlike the structure-function paradigm, however, our understanding of the molecular and evolutionary bases of IDR-mediated specificity remains very limited, despite their prominent roles in many biological processes.

With regards to sequence-features that correlate with condensate formation, extensions of sophisticated field theories (Firman and Ghosh, 2018; Lin et al., 2017, 2019; McCarty et al., 2019), as well as more coarse-grained approaches (Dignon et al., 2018; Sherry et al., 2017), have begun to connect the sequence of individual IDRs to their phase behavior *in vitro*. An exciting topic of investigation lies in understanding how ensembles of hetero-polymeric biomolecules can still exhibit compositional specificity. Older studies in the 1990s that explored how "randomly ordered" polymers exhibit specificity to "patterned" surfaces[16] (Srebnik et al., 1996) may be worth revisiting in the modern context.

[16] Employing replica-symmetry breaking approaches originally developed to probe glasses and other disordered media

A different route is to trace the evolutionary trajectory of IDRs with the goal of identifying conserved features (whether they are compositional or structural). A key challenge here is the rapid evolution of these regions (Schlessinger et al., 2011) which renders traditional alignment-based techniques with little to no power. Models that leverage deep learning techniques (Erijman et al., 2019), as well as those that infer evolutionary or compositional couplings (Toth-Petroczy et al., 2016; Zarin et al., 2019), have shown modest promise in uncovering features conserved in disordered domains. An exciting topic for the future would be to develop "interpretable" machine learning models[17] - approaches that will ultimately enable mechanistic interpretation of features learnt from data.

[17] Recently developed with limited success for identifying certain structural folds (Tubiana et al., 2019)

Genome structure and nuclear condensates

The genome consists of multiple chromosomes, which are long polymeric molecules of packaged DNA. In addition to the genetic code (sequence of DNA), it has been long appreciated that the structure (spatio-temporal conformations of this polymer) is central to cellular

fitness (Boveri, 1914). This structure informs how distant parts of the genome can relay information or communicate with each-other. Recent advent of the "Hi-C" class of techniques (Flyamer et al., 2017; Lieberman-Aiden et al., 2009), along with advances in multiplexed super-resolution imaging (Bintu et al., 2018; Boettiger et al., 2016; Wang et al., 2016), have enabled better characterization of the 4-dimensional conformation and dynamics of the genome. It is increasingly understood that a major outstanding question lies in understanding the nature of the feedback and interplay between the "4d-genome"[18] and different nuclear condensates. I will speculate on a few interesting areas of investigation below.

[18] 3d conformation across time

A central area of interest is the precise communication of information between enhancers and their target genes. Emerging evidence suggests that enhancers and their target promoters are often separated by distances larger than molecular-bridges (Benabdallah et al.; Chen et al., 2018a; Heist et al., 2019; El Khattabi et al., 2019) – a phenomena explained by the formation of a reservoir of polymerases and transcriptional material by phase separation. If this is the case, how is an enhancer able to specifically activate a target promoter and not other "off-target" regions which are sometimes closer along the genome by sequence? How does the genome sequence-structure paradigm interface with those of transcriptional condensates? A fascinating, and perhaps extreme, feat of coordination is the collation of genomic information across multiple chromosomes[19] in olfaction receptor neurons (Monahan et al., 2019). Theoretical and modeling approaches combining knowledge of phase behavior with large-scale polymer simulations, in concert with experiments and widely-available open-source data can shed further light on this area.

[19] i.e. regulation *in trans*

Over the last few years, two largely independent and major modes of genome organization have been delineated (1) the role of loop extrusion factors – molecules that loop out regions of DNA and (2) compartmentation (physically) of the chromatin into "active" (genes are expressed) and "silent" domains (genes are inactive or repressed) –with pertinent epigenetic correlates (Dekker and Mirny, 2016; Nora et al., 2017; Schwarzer et al., 2017; Vian et al., 2018). How these 2 modes of organization interact with condensates is largely unknown[20]. For example, loop-extruding proteins, as well as chromatin remodelers, are energy-

[20] Even the relative interplay of these 2 modes are active areas of investigation.

consuming motors that exert shearing forces – how these forces affect material and compositional properties of nuclear condensates remain unknown. There are two lines of evidence that suggest polymer microphase separation can contribute to compartmentation - first, from simulations and chromosome conformation data (Boettiger et al., 2016; Nuebler et al., 2018; Wang et al., 2016) and second, through *in vitro* observations of nucleosomal phase separation dependent on histone tail modifications (Gibson et al., 2019). In parallel, there is growing evidence that a number of protein-rich condensates form around particular genomic loci – including transcriptional condensates (Sabari et al., 2018), poly-comb mediated repression condensates (Tatavosian et al., 2018), heterochromatic condensates at silenced genes (Larson et al., 2017; Li et al., 2020; Sanulli et al., 2019; Strom et al., 2017), amongst other. How these processes intersect with each other, what sets the compositional specificity between particular genomic domains and protein condensates, and how these vary across cell-types are exciting areas of future research.

RNA processing and transcription

In higher organisms, the transcribed message (RNA) from the DNA is not continuous (Berget et al., 1977; Chow et al., 1977) but rather split. Much of RNA splicing (to split or excise the unwanted parts of the message i.e. introns) occurs co-transcriptionally i.e. as the message is being transcribed (Jangi and Sharp; Neugebauer, 2002; Pandya-Jones and Black, 2009). The mRNA undergoes further processing at distinct nuclear hubs – such as the nuclear speckles, which are splicing condensates (Mao et al., 2011; Nizami et al., 2010; Spector and Lamond, 2011; Staněk and Fox, 2017; Tatomer et al., 2016). Studying the interaction between transcriptional and splicing condensates will shed light on the mechanisms underlying gene control and co-transcriptional regulation.

Emerging evidence suggests a compositional gradient between condensates formed by transcriptional molecules and those by splicing factors (Guo et al., 2019) – with the latter being able to recruit more phosphorylated Polymerase. Splicing factors and co-factors often contain large disordered domains that are highly positively charged – as opposed to

highly negative charge carried by the multiply phosphorylated Pol II or transcribed RNA species (Plass et al., 2008). It is interesting to explore a model by which the biochemical rules of weak interactions between various moieties (sequence and structural features of RNA, transcriptional molecules, splicing proteins) give emergent organizational cues that link transcription to splicing.

Exploring the basis of how splicing speckles, which are themselves condensates enriched in RNA, are organized and regulated (Chen and Belmont, 2019; Chen et al., 2018b; Fei et al., 2017; Kim et al., 2019; Mao et al., 2011) is a related topic of interest- here observations suggest complex and layered organization of RNA/proteins (Fei et al., 2017), morphological sensitivity to phosphorylation and active transcription (Spector and Lamond, 2011).and non-trivial diffusive properties[21]. Investigation of these processes will require new approaches that integrate multi-phase phenomena, non-equilibrium statistical mechanics, and super-resolution imaging techniques.

21 For example, there is limited evidence for long-range correlated motion of speckles in response to heat shock (Chen and Belmont, 2019)

Bibliography

Banerjee, P.R., Milin, A.N., Moosa, M.M., Onuchic, P.L., and Deniz, A.A. (2017). Reentrant Phase Transition Drives Dynamic Substructure Formation in Ribonucleoprotein Droplets. Angew. Chemie Int. Ed. *56*, 11354–11359.

Beadle, G.W., and Tatum, E.L. (1941). Genetic Control of Biochemical Reactions in Neurospora. Proc. Natl. Acad. Sci. *27*, 499–506.

Benabdallah, N.S., Williamson, I., Illingworth, R.S., Kane, L., Boyle, S., Sengupta, D., Grimes, G.R., Therizols, P., and Bickmore, W.A. Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation. Mol. Cell *0*.

Berget, S.M., Moore, C., and Sharp, P.A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc. Natl. Acad. Sci. U. S. A. *74*, 3171–3175.

Bintu, B., Mateo, L.J., Su, J.-H., Sinnott-Armstrong, N.A., Parker, M., Kinrot, S., Yamaya, K., Boettiger, A.N., and Zhuang, X. (2018). Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. Science *362*, eaau1783.

Boettiger, A.N., Bintu, B., Moffitt, J.R., Wang, S., Beliveau, B.J., Fudenberg, G., Imakaev, M., Mirny, L.A., Wu, C., and Zhuang, X. (2016). Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. Nature *529*, 418–422.

Boija, A., Klein, I.A., Sabari, B.R., Dall'Agnese, A., Coffey, E.L., Zamudio, A. V., Li, C.H., Shrinivas, K., Manteiga, J.C., Hannett, N.M., et al. (2018). Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. Cell *175*, 1842 -1855.e16.

Boveri, T. (1914). Zur Frage der Entstehung maligner Tumoren. (Jena: Gustav Fischer).

Brangwynne, C.P., Tompa, P., and Pappu, R. V. (2015). Polymer physics of intracellular phase transitions. Nat. Phys. *11*, 899–904.

Chen, Y., and Belmont, A.S. (2019). Genome organization around nuclear speckles. Curr. Opin. Genet. Dev. *55*, 91–99.

Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J.B., and Gregor, T. (2018a). Dynamic interplay between enhancer–promoter topology and gene activity. Nat. Genet. 1.

Chen, Y., Zhang, Y., Wang, Y., Zhang, L., Brinkman, E.K., Adam, S.A., Goldman, R., van Steensel, B., Ma, J., and Belmont, A.S. (2018b). Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. J. Cell Biol. jcb.201807108.

Choi, J.-M., Holehouse, A.S., and Pappu, R. V. (2020). Physical Principles Underlying the Complex Biology of Intracellular Phase Transitions. Annu. Rev. Biophys. *49*.

Chow, L.T., Gelinas, R.E., Broker, T.R., and Roberts, R.J. (1977). An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. Cell *12*, 1–8.

Dekker, J., and Mirny, L. (2016). The 3D Genome as Moderator of Chromosomal Communication. Cell *164*, 1110–1121.

Dignon, G.L., Zheng, W., Best, R.B., Kim, Y.C., and Mittal, J. (2018). Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. Proc. Natl. Acad. Sci. U. S. A. *115*, 9929–9934.

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K., et al. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell *159*, 374–387.

Erijman, A., Kozlowski, L., Sohrabi-Jahromi, S., Fishburn, J., Warfield, L., Schreiber, J., Noble, W.S., Söding, J., and Hahn, S. (2019). A high-throughput screen for transcription activation domains reveals their sequence characteristics and permits reliable prediction by deep learning. BioRxiv 2019.12.11.872986.

Fei, J., Jadaliha, M., Harmon, T.S., Li, I.T.S., Hua, B., Hao, Q., Holehouse, A.S., Reyer, M., Sun, Q., Freier, S.M., et al. (2017). Quantitative analysis of multilayer organization of proteins and RNA in nuclear speckles at super resolution. J. Cell Sci. *130*, 4180–4192.

Firman, T., and Ghosh, K. (2018). Sequence charge decoration dictates coil-globule transition in intrinsically disordered proteins A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins Sequence charge decoration dictates coil-globul. J. Chem. Phys. J. Chem. Phys. J. Chem. Phys. J. Chem. Phys. *148*, 123305–123303.

Flyamer, I.M., Gassler, J., Imakaev, M., Brandão, H.B., Ulianov, S. V., Abdennur, N., Razin, S. V., Mirny, L.A., and Tachibana-Konwalski, K. (2017). Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature *544*, 110–114.

Gibson, B.A., Doolittle, L.K., Schneider, M.W.G., Jensen, L.E., Gamarra, N., Henry, L., Gerlich, D.W., Redding, S., and Rosen, M.K. (2019). Organization of Chromatin by Intrinsic and Regulated Phase Separation. Cell *0*.

Guo, Y.E., Manteiga, J.C., Henninger, J.E., Sabari, B.R., Dall'Agnese, A., Hannett, N.M., Spille, J.-H.J.-H., Afeyan, L.K., Zamudio, A. V., Shrinivas, K., et al. (2019). Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. Nature *572*, 543–548.

Hahn, S. (2018). Phase Separation, Protein Disorder, and Enhancer Function. Cell *175*, 1723–1725.

Heist, T., Fukaya, T., and Levine, M. (2019). Large distances separate coregulated genes in living Drosophila embryos. Proc. Natl. Acad. Sci. U. S. A. *116*, 15062–15067.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-Enhancers in the Control of Cell Identity and Disease (Cell Press).

Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., and Sharp, P.A. (2017). A Phase Separation Model for Transcriptional Control. Cell *169*, 13–23.

Jangi, M., and Sharp, P.A. Building Robust Transcriptomes with Master Splicing Factors. Cell *159*, 487–498.

El Khattabi, L., Zhao, H., Kalchschmidt, J., Young, N., Jung, S., Van Blerkom, P., Kieffer-Kwon, P., Kieffer-Kwon, K.-R., Park, S., Wang, X., et al. (2019). A Pliable Mediator Acts as a Functional Rather Than an Architectural Bridge between Promoters and Enhancers. Cell *0*.

Kim, J., Khanna, N., and Belmont, A.S. (2019). Transcription amplification by nuclear speckle association. BioRxiv 604298.

Kirschner, M., Gerhart, J., Otey, C.R., and Arnold, F.H. (1998). Evolvability. Proc. Natl. Acad. Sci. U. S. A. *95*, 8420–8427.

Kirschner, M., Gerhart, J.C., and Norton, J. (2006). The Plausibility of Life: Resolving Darwin's Dilemma (Yale University Press).

Larson, A.G., Elnatan, D., Keenen, M.M., Trnka, M.J., Johnston, J.B., Burlingame, A.L., Agard, D.A., Redding, S., and Narlikar, G.J. (2017). Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. Nature 236–240.

Lech, K., Anderson, K., and Brent, R. (1988). DNA-bound Fos proteins activate transcription in yeast. Cell *52*, 179–184.

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T., et al. (2014). Classification of Intrinsically Disordered Regions and Proteins. Chem. Rev. *114*, 6589–6631.

Li, C.H., Coffey, E.L., Dall'Agnese, A., Hannett, N.M., Tang, X., Henninger, J.E., Platt, J.M., Oksuz, O., Zamudio, A. V., Afeyan, L.K., et al. (2020). MeCP2 links heterochromatin condensates and neurodevelopmental disease. Nature 1–8.

Lieberman-Aiden, E., Berkum, N.L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science *326*, 289–293.

Lin, Y.-H., Brady, J.P., Forman-Kay, J.D., and Chan, H.S. (2017). Charge pattern matching as a 'fuzzy' mode of molecular recognition for the functional phase separations of intrinsically disordered proteins. New J. Phys. *19*, 115003.

Lin, Y., McCarty, J., Rauch, J.N., Delaney, K.T., Kosik, K.S., Fredrickson, G.H., Shea, J.-E.E., and Han, S. (2019). Narrow equilibrium window for complex coacervation of tau and RNA under cellular conditions. Elife *8*, 1–31.

Mao, Y.S., Zhang, B., and Spector, D.L. (2011). Biogenesis and function of nuclear bodies. Trends Genet. *27*, 295–306.

McCarty, J., Delaney, K.T., Danielsen, S.P.O., Fredrickson, G.H., and Shea, J.-E. (2019). Complete Phase Diagram for Liquid–Liquid Phase Separation of Intrinsically Disordered Proteins. J. Phys. Chem. Lett. *10*, 1644–1652.

Milin, A.N., and Deniz, A.A. (2018). Reentrant Phase Transitions and Non-Equilibrium Dynamics in Membraneless Organelles. Biochemistry *57*, 2470–2477.

Monahan, K., Horta, A., and Lomvardas, S. (2019). LHX2- and LDB1-mediated trans interactions regulate olfactory receptor choice. Nature *565*, 448–453.

Neugebauer, K.M. (2002). On the importance of being co-transcriptional. J. Cell Sci. *115*, 3865–3871.

Nizami, Z., Deryusheva, S., and Gall, J.G. (2010). The Cajal Body and Histone Locus Body. Cold Spring Harb. Perspect. Biol. *2*, a000653.

Nora, E.P., Goloborodko, A., Valton, A.-L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., and Bruneau, B.G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell *169*, 930-944.e22.

Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N., and Mirny, L.A. (2018). Chromatin organization by an interplay of loop extrusion and compartmental segregation. Proc. Natl. Acad. Sci. U. S. A. *115*, E6697–E6706.

Pandya-Jones, A., and Black, D.L. (2009). Co-transcriptional splicing of constitutive and alternative exons. RNA *15*, 1896–1908.

Plass, M., Agirre, E., Reyes, D., Camara, F., and Eyras, E. (2008). Co-evolution of the branch site and SR proteins in eukaryotes. Trends Genet. *24*, 590–594.

Ptashne, M., and Gann, A. (1997). Transcriptional activation by recruitment. Nature *386*, 569–577.

Sabari, B.R., Dall'Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A. V., Manteiga, J.C., et al. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. Science *361*.

Sanulli, S., Trnka, M.J., Dharmarajan, V., Tibble, R.W., Pascal, B.D., Burlingame, A.L., Griffin, P.R., Gross, J.D., and Narlikar, G.J. (2019). HP1 reshapes nucleosome core to promote heterochromatin phase separation. Nat. 2019 1–1.

Schlessinger, A., Schaefer, C., Vicedo, E., Schmidberger, M., Punta, M., and Rost, B. (2011). Protein disorder — a breakthrough invention of evolution? Curr. Opin. Struct. Biol. *21*, 412–418.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., Haering, C., Mirny, L., et al. (2017). Two independent modes of chromatin organization revealed by cohesin removal. Nature *551*, 51.

Sherry, K.P., Das, R.K., Pappu, R. V, and Barrick, D. (2017). Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the Notch receptor. Proc. Natl. Acad. Sci. U. S. A. *114*, E9243–E9252.

Shrinivas, K., Sabari, B.R., Coffey, E.L., Klein, I.A., Boija, A., Zamudio, A. V., Schuijers, J., Hannett, N.M., Sharp, P.A., Young, R.A., et al. (2019). Enhancer features that drive formation of transcriptional condensates. Mol. Cell *75*, 549-561.e7.

Spector, D.L., and Lamond, A.I. (2011). Nuclear speckles. Cold Spring Harb. Perspect. Biol. *3*, 1–12.

Srebnik, S., Chakraborty, A.K., and Shakhnovich, E.I. (1996). Adsorption-Freezing Transition for Random Heteropolymers near Disordered 2D Manifolds due to "Pattern Matching." Phys. Rev. Lett. *77*, 3157–3160.

Staněk, D., and Fox, A.H. (2017). Nuclear bodies: news insights into structure and function. Curr. Opin. Cell Biol. *46*, 94–101.

Strom, A.R., Emelyanov, A. V., Mir, M., Fyodorov, D. V., Darzacq, X., and Karpen, G.H. (2017). Phase separation drives heterochromatin domain formation. Nature *547*, 241–245.

Struhl, K. (1988). The JUN oncoprotein, a vertebrate transcription factor, activates transcription in yeast. Nature *332*, 649–650.

Tatavosian, R., Kent, S., Brown, K., Yao, T., Duc, H.N., Huynh, T.N., Zhen, C.Y., Ma, B., Wang, H., and Ren, X. (2018). Nuclear condensates of the Polycomb protein chromobox 2 (CBX2) assemble through phase separation. J. Biol. Chem. jbc.RA118.006620.

Tatomer, D.C., Terzo, E., Curry, K.P., Salzler, H., Sabath, I., Zapotoczny, G., McKay, D.J., Dominski, Z., Marzluff, W.F., and Duronio, R.J. (2016). Concentrating pre-mRNA pro-

cessing factors in the histone locus body facilitates efficient histone mRNA biogenesis. J Cell Biol jcb.201504043.

Tjian, R., and Maniatis, T. (1994). Transcriptional activation: A complex puzzle with few easy pieces. Cell *77*, 5–8.

Toth-Petroczy, A., Palmedo, P., Ingraham, J., Hopf, T.A., Berger, B., Sander, C., and Marks, D.S. (2016). Structured States of Disordered Proteins from Genomic Sequences. Cell *167*, 158-170.e12.

Tubiana, J., Cocco, S., and Monasson, R. (2019). Learning protein constitutive motifs from sequence data. Elife *8*.

Vian, L., Pękowska, A., Rao, S.S.P., Kieffer-Kwon, K.-R., Jung, S., Baranello, L., Huang, S.-C., El Khattabi, L., Dose, M., Pruett, N., et al. (2018). The Energetics and Physiological Impact of Cohesin Extrusion. Cell *0*.

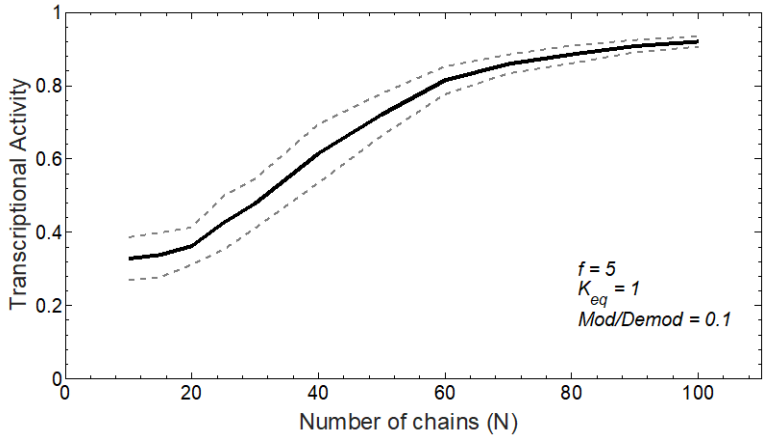
Wang, S., Su, J.-H., Beliveau, B.J., Bintu, B., Moffitt, J.R., Wu, C., and Zhuang, X. (2016). Spatial organization of chromatin domains and compartments in single chromosomes. Science *353*, 598–602.

Zarin, T., Strome, B., Nguyen Ba, A.N., Alberti, S., Forman-Kay, J.D., and Moses, A.M. (2019). Proteome-wide signatures of function in highly diverged intrinsically disordered regions. Elife *8*.

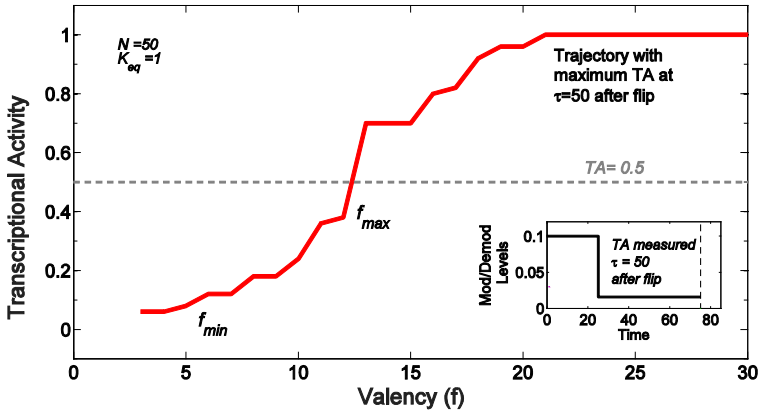
Appendices and supplementary figures

Chapter 1: Appendix

Supplemental Figures



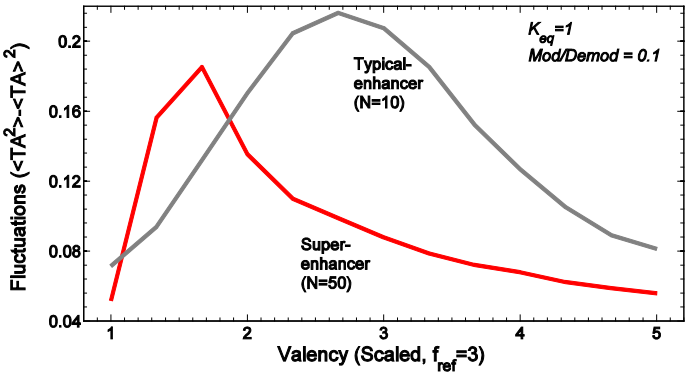
Supplemental Figure 1. Dependence of transcriptional activity (TA) on number of chains (N) is depicted above. The proxy for transcriptional activity (TA) is defined as the size of the largest cluster of cross-linked chains, scaled by the total number of chains. The solid lines indicate the mean and the dashed lines indicate twice the standard deviation in 50 simulations. All simulations are done at Modifier/Demodifier=0.1, K_{eq}=1 and *f*=5. TA levels are very different as long as the values of N (or concentration of components) for a SE and a typical enhancer are sufficiently different.



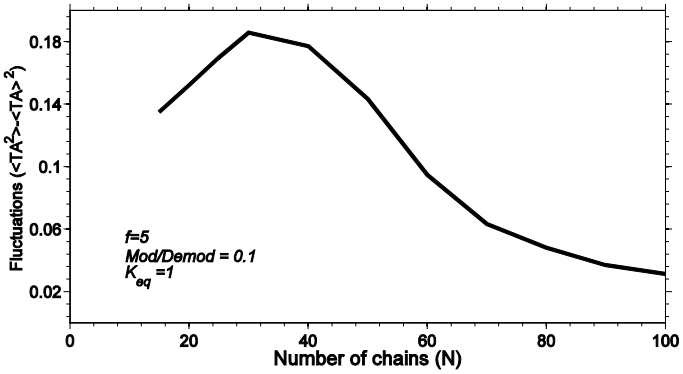
Supplemental Figure 2. Simulations carried out to study disassembly of the gel after a sharp change in the Modifier/Demodifier balance (mimics change in signals). The proxy for transcriptional activity (TA) is defined as the size of the largest cluster of cross-linked chains, scaled by the total number of chains. As depicted in the inset, the ratio of Modifier/Demodifier levels are flipped (at =25) from 0.1 to 0.016 and TA is calculated τ=50 time units *post change* in the Modifier/Demodifier balance. All simulations are done for N=50 (model for SE) and K_{eq}=1.The solid line represents the variation in the maximum value of the calculated TA in 250 replicate simulations as valency (f) is changed. Threshold valencies f_{min}, for ensuring cluster formation (see Figure 2C), and f_{max}, to ensure robust disassembly (defined as TA<0.5, dotted line) within τ=50 time units post change in Modifier/Demodifier levels are identified. The specific value of τ=50 time units post change in Modifier/Demodifier values is chosen for illustrative purposes, and

determines the value of f_{max}. The qualitative result that there exists a maximal valency above which the gel does not disassemble in a realistic time scale is robust to changes in the chosen value of this time scale.

A



B



Supplemental Figure 3. Noise characteristics of super-enhancers and typical enhancers

A. Dependence of fluctuations (or transcriptional noise), measured as variance in Transcriptional activity (TA), on valency are shown for SEs (N=50) and typical enhancers (N=10). The proxy for transcriptional activity (TA) is defined as the size of the largest cluster of cross-linked chains, scaled by the total number of chains. The angular brackets in the definition of the ordinate represent averages over 50 replicate simulations. All simulations are done at Modifier/Demodifier=0.1, K_{eq}=1. The normalized magnitude of the noise, and importantly the range of valencies over which the noise is manifested, are smaller for SEs compared to a typical enhancer. Note, however, that the absolute magnitude of the noise in the vicinity of the phase separation point is larger for bigger values of N.

B. Dependence of fluctuations (or transcriptional noise), measured as variance in Transcriptional activity (TA), on N is shown for f = 5 (the minimal valency required for cluster formation for N=50. All simulations are done at Modifier/Demodifier=0.1 and

K_{eq}=1. The proxy for transcriptional activity (TA) is defined as the size of the largest cluster of cross-linked chains, scaled by the total number of chains. The angular brackets in the definition of the ordinate represent averages over 50 replicate simulations.

Chapter 2 - Appendix

Methods

Cells

V6.5 murine embryonic stem cells were a gift from R. Jaenisch of the Whitehead Institute. V6.5 are male cells derived from a C57BL/6(F) x 129/sv(M) cross.

Cell culture conditions

V6.5 murine embryonic stem cells were grown in 2i + LIF conditions on 0.2% gelatinized (Sigma, G1890) tissue culture plates. 2i + LIF media contains the following: 967.5 mL DMEM/F12 (GIBCO 11320), 5 mL N2 supplement (GIBCO 17502048), 10 mL B27 supplement (GIBCO 17504044), 0.5mML-glutaminae (GIBCO 25030), 0.5X non-essential amino acids (GIBCO 11140), 100 U/mL Penicillin-Streptomycin (GIBCO 15140), 0.1 mM b-mercaptoethanol (Sigma), 1 uM PD0325901 (Stemgent 04- 0006), 3 uM CHIR99021 (Stemgent 04-0004), and 1000 U/mL recombinant LIF (ESGRO ESG1107). Cells were negative for mycoplasma.

Developing coarse-grained simulations of DNA, TFs, and coactivators

We set up a coarse-grained molecular-dynamics simulation to model 3 different components – TFs, DNA, and coactivators, employing the HOOMD simulation framework (Anderson et al., 2008; Glaser et al., 2015). Briefly, the DNA chain was modeled as beads on a string, with two types of monomers. "Active" DNA units were modeled by tessellating a sphere ($diameter = 1/3\ unit$), using the rigid-body feature(Nguyen et al., 2011), to form a roughly cubical monomer of unit side length (Fig 1A). Binding patches were modeled as rigid particles along the cubic face

centers, with as many patches added as number of binding sites per monomer. Tessellation of active DNA monomers enabled 1:1 binding interactions, facilitated by excluded volume interactions from other tessellated spheres. “Inactive” DNA monomers were modeled as spherical monomers of unit diameter without any binding patches. TFs and coactivators were modeled employing two different methods – explicit-IDR (Fig 1A) and implicit-IDR models (Fig S3A). In the explicit-IDR framework, TFs and coactivators were designed in a modular fashion (Fig S1A). The “structured” domain was modeled as a spherical monomer of diameter $d = 0.75$ units.s. IDRs were constructed by tethering a polymeric tail to the spherical domain, with TFs having shorter chains (4 monomers of $d = 1/3\ unit$) than coactivators (9 monomers of $d = 1/3\ unit$), to mimic the differential size of disordered regions. In the implicit-IDR model, TFs and coactivators were modeled as spherical monomers of unit diameter. All monomers had the same density. The sizes of the modeled monomers of DNA and proteins mimics the relative similarity in sizes between TFs/coactivators and nucleosomes. In both methods, DNA binding patches on proteins were modeled as rigid particles buried in the “structured” domains.

Non-bonding interactions between any two particles (including binding patches) were modeled using a truncated, shifted, and size-normalized LJ potential (U) with hard-core repulsion (particles don’t overlap), derived in the following form:

$$U_{ij}(\vec{r}) = \begin{pmatrix} P_{ij}(r) - P_{ij}(r^*) & r \le r^* \\ 0 & r > r^* \end{pmatrix}$$

$$P_{ij}(r) = 4\epsilon_{ij} \times ((\sigma/r)^{12} - (\sigma/r)^{6})$$
$$\sigma = 0.5 \times (d_i + d_j), r^* = 2.5 \times \sigma$$

Bonding interactions between neighboring monomers on a chain were modeled using a harmonic potential with hard-cores, with a spring constant $k = 1e4$. All energy units are scaled to kT units, with kT=1.

The strength of various interactions was set based on the rationale stated in main text. Typical TF-DNA binding affinities are strong and in the range of nanomolar (Jung et al., 2018) disassociation constants i.e. $K_D \approx 10^{-9}M$. The Gibbs free enthalpy change of binding can be approximately calculated as $\Delta G \approx -kTln(K_D) \approx 20kT$. Thus, specific monovalent DNA interactions were set to high affinities - for e.g. $\epsilon_{DNA-TF} = 20kT$ in fig 1B,2A, $\epsilon_{DNA-TF} = 16\ kT$ in fig S3A-B. IDR interactions were much weaker and individual interactions are often of the order of thermal fluctuations (Brady et al., 2017; Nott et al., 2015; Wei et al., 2017) i.e. order kT, though their energetic contributions can effectively multiply through multivalence. Thus, we set $\epsilon_{IDR} \sim\ kT$ between monomers on the IDR chain. For the implicit-IDR model reported in Fig S2, multivalent interactions were approximated by a weak LJ potential between particles, for e.g., $\epsilon_{TF-coA} = 1.5kT, \epsilon_{coA-coA} = 1.5kT, \epsilon_{TF-TF} = 1.0\ kT$. The key qualitative results i.e multivalent DNA acts as scaffold for phase separation at low protein concentrations and seed at higher protein levels, has been reproduced for different choices of interaction parameters guided by the rationale above.

Particles are randomly initialized in the periodic simulation box, and randomly re-seeded for each replicate trajectory, with the Langevin thermostat. Friction coefficients were $\gamma = 1$ for proteins and $\gamma = 100$ for DNA, to mimic chromosomal motion damping. Initial velocities were drawn from the Boltzmann distribution. First, simulations were run with small time steps ($dt = 5 \times 10^{-6}$) to prevent randomly generated "high-energy" configurations from blowing up and to relax the system to the thermostat temperature. These "warm-up" period ($t \sim 0.1\ units$) is much smaller than the time to reach steady-state $t_{ss} \sim (1000\ units)$, so these warm-up data points are not used in any analysis. All simulations are run with a single DNA chain.

Explicit-IDR simulations are run for at least $45e6$ steps to accurately recapitulate dynamics and reach steady-state, whilst implicit-IDR simulations are run for $5e6$ steps. The slowing down of explicit-IDR simulations (due to slower explicit-IDR dynamics), combined with

additional pair-wise interaction computations (explicit pairwise calls for all monomers, which are an order of magnitude more particles for explicit-IDR simulations, scale as $\sim N^2$ for N monomers), cause computation times for single trajectories to be ~50-100 times longer than the implicit-IDR version. Trajectory states were logged in the highly compressed, binarized GSD format every 50000 steps, while observables were logged every 20000 steps.

To probe the role of DNA in our simulations, after steady-state is reached, interactions with the DNA binding sites are switched off. Interactions are switched off by replacing all binding patches with "ghost" patches, with no energetic benefits. Simulations are typically continued for the same amount of steps before disrupting TF-DNA interactions to accurately sample steady-state. A brief overview of key parameters used in main/supplementary figures is found below in Table S1. The MD code for running analysis will be made freely available upon publication.

Analysis of simulation data

Broadly, analyses of simulation data were split into on-the-fly calculations employing the Freud package (https://freud.readthedocs.io/en/stable/installation.html), as well as post-simulation calculations that leverage a combination of various libraries which interface with python – including *numpy*, *scipy*, *freud*, *matplotlib*, and *fresnel*. On-the-fly calculations include:

1. In-built functions for logging potential energy, kinetic energy, and temperature.

2. Number of monomers in largest cluster and radius of largest cluster: A call-back routine was implemented that used Freud to estimate the size of the largest connected cluster with $r = 1.4d_{max}$ (d_{max} diameter of largest monomer) to identify largest cluster. This largest cluster size is relatively insensitive to studied

choices of parameter $r = 1.25,1.35,1.45\ d_{max}$. Every reported plot with scaled size at steady state, which is the number of molecules in the largest cluster divided by number of binding sites (Fig 1B), reports the mean in the dark line, and one standard deviation in the shaded background.

For post-simulation calculations, data was read from GSD formats using the gsd module. Explicit-IDR simulation trajectory data was parsed to convert from number of molecules to number of chains, while following the other steps as mentioned above. The entropy was calculated in Fig 2 and Fig S2 by identifying the number of molecules in the largest cluster (in the case of the explicit-IDR simulations, each polymer was counted as one molecule), and adding a value of $kTln(\frac{4/3\pi R_g^3}{V_{free}})$ for each molecule in the condensed phase. V_{free} was computed as the total volume minus the excluded volume occupied by all molecules.

For the fluctuation analysis in Fig S3, the variance in largest cluster size of individual stochastic trajectories was computed and averaged at steady state. This value was normalized by the scaled cluster size, to compute the scaling of fluctuations beyond the usual $\sqrt{N}$ finite-size effects.

Contact frequency analysis of simulation data

For contact frequency maps, which are similar to Hi-C maps, represented in Fig 6B and S5, the following analysis protocol was employed. After individual trajectories reached steady-state, the position of each DNA monomer along the chain was logged at every time step. Monomers closer than ($r = 3.0$ units) a distance at a time t are "cross-linked" i.e. they count as an interacting pair. The qualitative interaction maps reported in Fig 6B are robust to other tested values of crosslinking radius in the regime of $2.5 < r < 4$ units. The pairwise contact frequency matrix is then constructed by averaging over interactions over a time window at steady state per trajectory, as well as averaging over 10 replicate trajectories per simulation condition. The contact matrix is visualized

using the *seaborn* and *matplotlib* packages in python3.

Computing radial density profiles from simulation data

Simulations were analyzed at steady-state to estimate the radial density of TFs and coactivators around the DNA chain (g(r) from DNA). The freud rdf analysis package was used to compute the rdf around reference positions of DNA for both distributions of TFs and coactivator molecules. In case of explicit-IDR simulation, the structured domains of the respective molecules were used to probe their locations. The final g(r) from DNA is obtained by averaging over 50 distinct simulation frames (typically logged once every 50,000 steps) per trajectory, and over 10 trajectories. The g(r) is visualized for both explicit-IDR (Fig S1C) and implicit-IDR (Fig S2C) at low concentrations, before and after disruption of TF-DNA interactions, using *matplotlib* in *python3*.

Visualization of simulation data:

All simulation data-sets were analyzed in python3, with the aid of *matplotlib*, to generate publication-ready figures. Simulation movies were generated by stitching together down-sampled frames (once every 100000 steps) of individual stochastic trajectories, using Fresnel to render scenes with the same color palette used in Fig 1A, and PIL to store image arrays as gifs. After storing the gifs, these files were converted to .mp4 movies externally and subs are added at the frame at which TF-DNA interactions are turned off.

Quantitative immunoblot

Determination of number of MED1 molecules per cell and concentration by linear regression analysis. Quantitative Western Blotting was carried out as described in (Lin et al., 2012). Cell number was determined using a Countless II FL Automated Cell Counter (Thermo Fisher Scientific). Cells were lysed with Cell Lytic M (Sigma) with protease inhibitors at various

concentrations and denatured in DTT and XT Sample Buffer (Biorad) at 90°C for 5 minutes. Purified recombinant MED1-IDR was used as a standard and loaded in the amounts depicted in the figure in the same gel as the cell lysates. Lysates and standards were run on a 3%–8% Tris-acetate gel at 80 V for ~2 hrs, followed by 120 V until dye front reached the end of the gel. Protein was then wet transferred to a 0.45 μm PVDF membrane (Millipore, IPVH00010) in ice-cold transfer buffer (25 mM Tris, 192 mM glycine, 10% methanol) at 300 mA for 2 hours at 4°C. After transfer the membrane was blocked with 5% non-fat milk in TBS for 1 hour at room temperature, shaking. Membrane was then incubated with 1:1,000 anti-MED1 (Assay Biotech B0556) diluted in 5% non-fat milk in TBST and incubated overnight at 4°C, with shaking. The membrane was then washed three times with TBST for 5 minutes at room temperature shaking for each wash. Membrane was incubated with 1:10,000 secondary antibody conjugated to HRP for 1 hr at RT and washed three times in TBST for 5 minutes. Membranes were developed with ECL substrate (Thermo Scientific, 34080) and imaged using a CCD camera.(BioRad ChemiDoc). Band intensities were determined using ImageJ. Number of molecules per cell was determined by linear regression analysis through the origin using Prism 7. The concentration of MED1 was calculated using nuclear volumes obtained by analysis of Hoechst (Life Technologies)-stained mouse embryonic stem cells in ImageJ and assuming all MED1 molecules reside in the nucleus.

Protein purification

Proteins were purified as in (Boija et al., 2018; Sabari et al., 2018). cDNA encoding the genes of interest or their IDRs were cloned into a modified version of a T7 pET expression vector. The base vector was engineered to include a 5' 6xHIS followed by either mEGFP or mCherry and a 14 amino acid linker sequence "GAPGSAGSAAGGSG." NEBuilder® HiFi DNA Assembly Master Mix (NEB E2621S) was used to insert these sequences (generated by PCR) in-frame with the linker amino acids. Mutant sequences were synthesized as gBlocks (IDT) and inserted into the same

base vector as described above. All expression constructs were sequenced to ensure sequence identity. For protein expression, plasmids were transformed into LOBSTR cells (gift of Chessman Lab) and grown as follows. A fresh bacterial colony was inoculated into LB media containing kanamycin and chloramphenicol and grown overnight at 37°C. Cells containing the MED1-IDR constructs were diluted 1:30 in 500ml room temperature LB with freshly added kanamycin and chloramphenicol and grown 1.5 hours at 16°C. IPTG was added to 1mM and growth continued for 18 hours. Cells were collected and stored frozen at -80°C. Cells containing all other constructs were treated in a similar manner except they were grown for 5 hours at 37°C after IPTG induction. 500ml cell pellets were resuspended in 15ml of Buffer A (50mM Tris pH7.5, 500 mM NaCl) containing 10mM imidazole and cOmplete protease inhibitors, sonicated, lysates cleared by centrifugation at 12,000g for 30 minutes at 4°C, added to 1ml of pre-equilibrated Ni-NTA agarose, and rotated at 4°C for 1.5 hours. The slurry was poured into a column, washed with 15 volumes of Buffer A containing 10mM imidazole and protein was eluted 2 X with Buffer A containing 50mM imidazole, 2 X with Buffer A containing 100mM imidazole, and 3 X with Buffer A containing 250mM imidazole.

Production of fluorescent DNA

Gene fragments were synthesized by either GeneWiz or IDT and cloned into a pUC19 vector using HiFi Assembly (NEB) so that the sequence was immediately flanked by M13(-21) and M13 reverse primer sequences. 5'-fluorescently labeled (Cy5) M13(-21) (/5Cy5/ TGTAAAACGACGGCCAGT) and M13 reverse (/5Cy5/ CAGGAAACAGCTATGAC) primers (IDT) were used to PCR amplify the synthetic DNA sequence, yielding a fluorescently labeled PCR product. Fluorescent PCR products were gel-purified (Qiagen) and eluted products were further purified using NEB Monarch PCR purification to remove any residual contaminants. The octamer motif sequence "ATTTGCAT" from the immunoglobulin kappa promoter was used as

the TF binding site. All PCR products used are 377 bp. The sequences of PCR products are provided in Table S2.

In vitro droplet assay

Recombinant GFP or mCherry fusion proteins were concentrated and desalted to an appropriate protein concentration and 125mM NaCl using Amicon Ultra centrifugal filters (30K MWCO, Millipore) in Buffer D(125) (50mM Tris-HCl pH 7.5, 10% glycerol, 1mM DTT). Fluorescent PCR products were concentration normalized in Buffer D(0) (50mM Tris-HCl pH 7.5, 10% glycerol, 1mM DTT). For all droplet assays, DNA was included at 50nM, mEGFP-OCT4 at 1250nM, and mCherry-MED1-IDR at the indicated concentration. Recombinant proteins and DNA were mixed with 10% PEG-8000 as a crowding agent. The final buffer conditions were 50mM Tris-HCl pH 7.5, 100mM NaCl, 10% glycerol, 1mM DTT. The solution was immediately loaded onto a homemade chamber comprising a glass slide with a coverslip attached by two parallel strips of double-sided tape. Slides were then imaged with an Andor spinning disk confocal microscope with a 150x objective. Unless indicated, images presented are of droplets settled on the glass coverslip.

For DNase I experiment, MED1-IDR droplets were formed at indicated concentration in the presence of OCT4 (1250nM) and ODNA_20 (50nM). The solution containing droplets was split into two equal volumes, to one volume DNase I (Turbo DNase, Invitrogen, 3U) was added with manufacturer provided reaction buffer and to the second volume enzyme storage buffer and reaction buffer were added. These were loaded onto slides, incubated at 37° C for 2 hours and subsequently imaged as described above.

Image analysis for reconstructing experimental phase curves

A custom analysis pipeline was developed in MATLAB™, building on code described in (Boija et al., 2018). Briefly, droplets were identified by

employing a two-step thresholding procedure. First, the image was segmented in the MED1-IDR channel with an intensity threshold ($I_{pixel} > \mu^* + 3\sigma$, where μ^* is the most probable intensity, representative of background, and σ is the width of the distribution) to identify bright pixels. Subsequently, the identified bright pixels were labeled as "condensed" droplet phase after enforcing a minimum droplet size of 9 pixels i.e. atleast 9 clustered pixels had to simultaneously pass the intensity threshold to belong to the condensed phase. In the absence of phase separation, no pixels are identified as belonging to the condensed phase.

For each image, the total intensity in the condensed droplet phase was summed in each channel ($I_{channel,droplet}$), as well as the total background intensity outside droplets ($I_{channel,bulk}$). The condensed fraction in each channel was defined as :

$$c.f._{\cdot channel} = \frac{I_{channel,droplet}}{I_{channel,droplet} + I_{channel,bulk}}.$$

The condensed fraction was averaged over replicate images (≥10 per condition). At very low concentrations or in the absence of observable phase separation, c.f. is close to 0. We repeated the c.f. analysis with different intensity thresholds ($I > \mu^* + 2.5\sigma, I > \mu^* + 3.5\sigma, I > \mu^* + 4\sigma$) and size thresholds ($9,16,25\ pixels$). The qualitative results reported in main and supplementary figures did not change under these tested conditions.

In all plots of the c.f., solid lines represent the mean condensed fraction and error bars refer to values one standard deviation above and below the mean, computed from replicates (n≥10). Plots of the condensed fraction were generated by using the *matplotlib* library in python3. In all plots in the main figures (Figs 1F, 3D, 3H, 4D, 5B), the condensed fraction in the MED1-IDR channel is reported.

For inferring saturation concentrations from the condensed fraction curves, a linear interpolation was fit using the linear-least squares

approach to the data from the replicates across the data points above and below the threshold (0.4% - for all data reported). The apparent saturation concentration (C_{sat}) was estimated as the concentration at which the condensed fraction reached the threshold value. The standard deviation in inferred values were computed from the standard error of the regression.

The difference between the inferred values of saturation concentrations across any set of conditions (as measured by their ratio) was insensitive to other tested values of the threshold in the range 0.3-0.6 %. Lower values of the threshold (<0.3%) led to unreliable estimates, confounded by noise from replicates, as well as specking from background, and were thus not employed. A T-test (with unequal variances, Welch's test, refer - scipy.stats.ttest_ind_from_stats) was performed to test for significance between inferred saturation concentrations.

DNase I image analysis

Building on the above-described analysis, for each condition, the partition ratio for each replicate image is calculated as $p_{channel} = \frac{<Intensity>_{droplet}}{<Intensity>_{bulk}}$ in various channels for each image-set. The key difference is that a background intensity subtraction (of 80 pixel units) is performed to aid in droplet identification and partition calculation at low concentrations. The partition ratio is a proxy for the relative enrichment of molecules in the condensed phase over the bulk phase. For any given experimental condition, the sample of partition ratios are obtained over replicate images ($n \geq 10$).

Subsequently, the partition ratios for control (without DNAseI) and DNAseI experiments were normalized to the mean partition ratio for the control at same concentration of MED1-IDR. Scatter plots with mean +/- std were generated using the normalized partition ratios in the 561(MED1-IDR) channel for Fig. 2B, and in the 640 channel (DNA) for Fig S1D, using PRISM.

Luciferase reporter assays

For enhancer activity reporter assays, synthetic enhancer DNA sequences with varying valences or densities of OCT4 binding sites (see Table S3) were cloned into a previously characterized pGL3-basic construct containing a minimal OCT4 promoter (pGL3-pOCT4)(Whyte et al., 2013). The synthetic enhancer sequences were cloned into the SalI site of the pGL3-pOct4 vector by HiFi DNA Assembly (NEB E2621) with a SalI digested vector and PCR-amplified insert. All cloned constructs were sequenced to ensure sequence identity. 0.4µg of the pGL3-based enhancer plasmids were used to transfect $1x10^5$ murine ESCs in 24-well plates using Lipofectamine 3000 (Thermo Fisher L3000015) according to the manufacturer's instructions. 0.1µg of the pRL-SV40 plasmid was co-transfected in each condition as a luminescence control. Transfected cells were harvested after 24 hours, and luciferase activity was measured using the Dual-Glo Luciferase Assay System (Promega E2920). Luciferase signal was normalized to the signal measured in cells transfected with a construct containing zero OCT4 motifs. Experiments were performed in triplicates.

Bioinformatic analysis

Position-weight matrices (PWMS) for *Mus musculus* stem cell master TFs –SOX2 (MA0143), OCT4+SOX2 (MA0142), KLF4 (MA0039), and ESRRB (MA0141), were obtained from the JASPAR database (Khan et al., 2018). 100kb DNA sequences centered on super-enhancers (SEs, N=231), as annotated in (Whyte et al., 2013) were gathered. The same number and length of sequences were randomly subsampled from enhancers (typical enhancers, TEs) annotated in (Whyte et al., 2013), as well as from random genetic loci (Random) on the *mm9* reference genome. FIMO was used to predict individual motif instances in all sequences, against a background uniform random distribution, at a *p-value* threshold of 1e-4.

For the boxplots in Fig 6A, the average motif density is calculated as total number of motifs divided by length of sequence over a 20kb sequence region centered on SEs, TEs, and random loci, normalized in units of motifs/kb. For the line plots in Fig 6B, the whole distribution of motif density is represented along the 100 KB sequence, in bins of 2kb with similar units.

Published ChIP-Seq data-sets are gathered from (Sabari et al., 2018) for MED1, BRD4, RNA Pol II, and input control from GEO: GSE112808. Reads-per-million (rpm) are summed in previously defined regions for SEs, TEs, and random using BedTools. For Fig 6C, and supplementary figure S5, the summed rpm values are plotted on a log scale. On the x-axis, the total number of motifs calculated in a 20kb window centered on SEs, TEs, and random loci is plotted. Finally, a linear model is inferred between $\log(ChIP)$ signal and motif values using ordinary least squares regression. The inferred line is plotted in black and 95% confidence intervals are plotted as a shaded gray background. The data is visualized using the *matplotlib* library in python3.

Statistical analysis for simulation data:

Steady-state analysis of simulation data-sets in Fig 3 & 4 are reported with solid lines represented by the mean (μ) and averaged fluctuations at steady state (across trajectories) in the shaded background, whose boundaries are characterized by one standard deviation away from the mean on either side ($\mu \pm \sigma$). In all figures, the mean represents an average over 10 trajectories. In Fig 1B, the steady state value is reported for 2 specific conditions (+/- DNA at low protein concentrations), with mean and 1 standard deviance (n=10 trajectories). For dynamical plots reported in Figs 2,4,5, the mean trajectory (n=10) is reported.

Statistical analysis for bioinformatics:

The inferred linear lines in Fig 6C and S5 are generated between the logarithm of the ChIP signal and the motif density, and the R^2 reported

in the respective captions. The 95% confidence interval in the inferred slope of the linear fit is reported in the grey background, calculated from statsmodels.api in python.

Statistical analysis for in vitro condensate assays:

Condensed fraction reported at any given concentration in all figures are averaged over > 10 image-sets, with error bars representing one standard deviation from the mean condensed fraction. Saturation concentrations are inferred (mean and std error) from the above data (n>10 data sets, ref methods above for details). The T-test (with unequal variances, Welch's test, refer - scipy.stats.ttest_ind_from_stats) was performed to test for significance. Pairwise student's t-test for DNase experiment (Figure 2B) and luciferase experiments (Figure 4C,5B) were performed using PRISM 7 (GraphPad).

DATA AND SOFTWARE AVAILABILITY

All software and code generated in this project are publicly available at https://github.com/krishna-shrinivas/2019_Shrinivas_Sabari_enhancer_features . The raw experimental data can be found at https://dx.doi.org/10.17632/c36nyy79y4.1 .

Figure S1

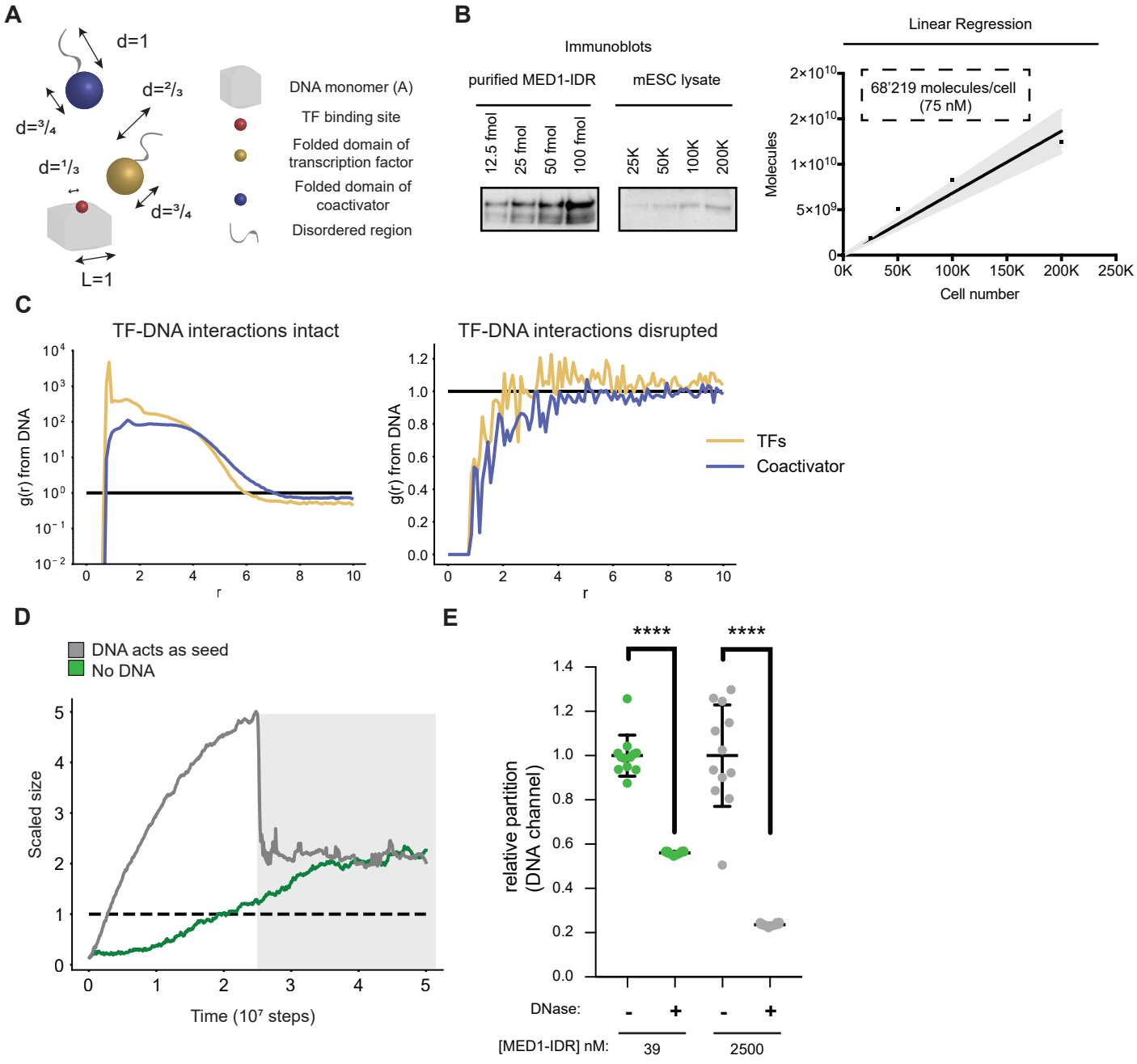


Figure S1: Multivalent TF-DNA interactions promote phase separation of TFs and coactivators, Related to Figures 1,2

A. Schematic illustrating geometries of the simulated molecules (in arbitrary units), including relative sizes of the folded and disordered domains. In typical simulations, the total length of the DNA chain is 10 DNA monomers.
B. Immunoblot of recombinant MED1-IDR at indicated concentrations or lysates from the indicated number of cells is shown on the top panel. Linear regression (bottom panel) is carried out to estimate number and concentration of MED1-IDR per cell (dashed box, bottom panel) (see methods for details).
C. The radial density function (g(r)) is computed around DNA at low concentrations for TFs (yellow) and coactivators (blue), before (left panel) and after (right panel) disruption of TF-DNA interactions. TFs and coactivators form a largely uniform dense phase incorporating DNA (high values and overlap of g(r)), which is lost upon disruption of TF-DNA interactions and condensate dissolution.
D. Dynamics of condensate assembly at conditions with (grey) and without DNA (green line) is represented by average scaled size on y-axis, and time (in simulation steps after initialization) on the x-axis. DNA promotes rate of assembly at high concentrations. However, DNA is not required for condensate stability, as evidenced by high values of scaled size after disruption of TF-DNA interactions (shown by a dark grey background).
E. Scatter-plot depiction of ODNA_20 partition ratio between condensate and background, at high (gray) and low MED1-IDR concentrations (orange) in conditions without DNase I addition (-) or with DNase I addition (+). The partition ratio is normalized to the (-) condition, showing that addition of DNase I degrades DNA.

Figure S2

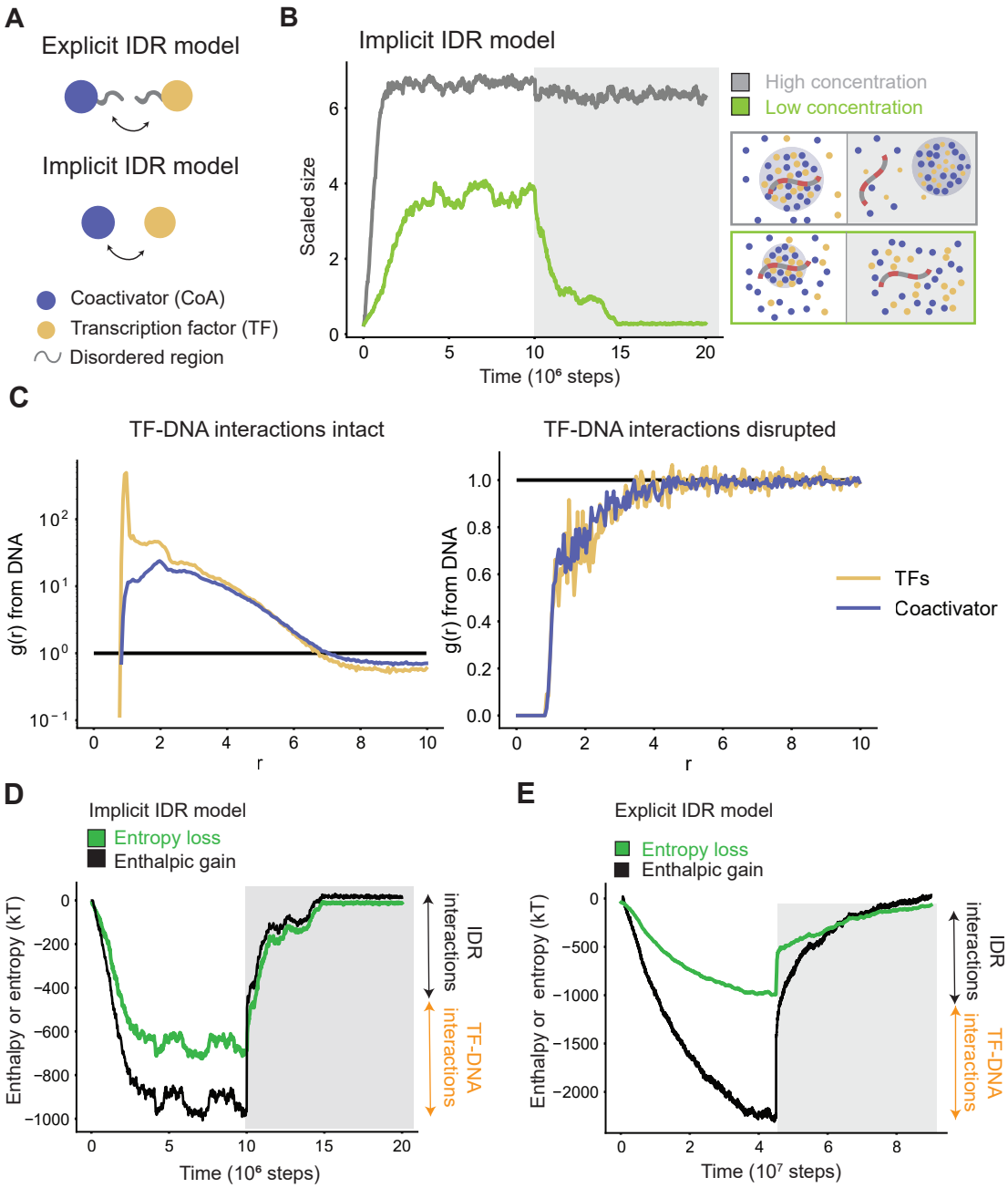


Figure S2: Simplified computational model recapitulates all features of explicit-IDR model, Related to Figure 2

A. Schematic cartoon of difference between explicit IDR model and implicit IDR model.
B. Dynamics of condensate assembly/disassembly at three different protein concentrations (gray = high concentration, orange = low concentration, black = lower concentration) is represented by average scaled size on the y-axis, and time (in simulation steps after initialization) on the x-axis. TF-DNA interactions are disrupted after steady state is reached (shown by a dark gray background). Schematic of phase behavior is presented next to simulation data, enclosed in boxes whose colors match the respective lines.
C. The radial density function (g(r)) is computed around DNA at low concentrations for TFs (yellow) and coactivators (blue), before (left panel) and after (right panel) disruption of TF-DNA interactions. TFs and coactivators form a largely uniform dense phase incorporating DNA (high values and overlap of g(r)), which is lost upon disruption of TF-DNA interactions and condensate dissolution.
D. Energetic attractions (black line) compensate entropic loss (green line) during condensate assembly, but disruption of TF-DNA interactions (magnitude =orange double arrow) causes dissolution at low concentrations.
E. Explicit-IDR simulations show a compensation of entropic loss (green line) by energetic attractions (black line) during condensate assembly, and disruption of TF-DNA interactions causes dissolution. However, the estimate of entropy loss from simulations is an under-count to the total loss of entropy, missing effects of configurational entropy.

Figure S3

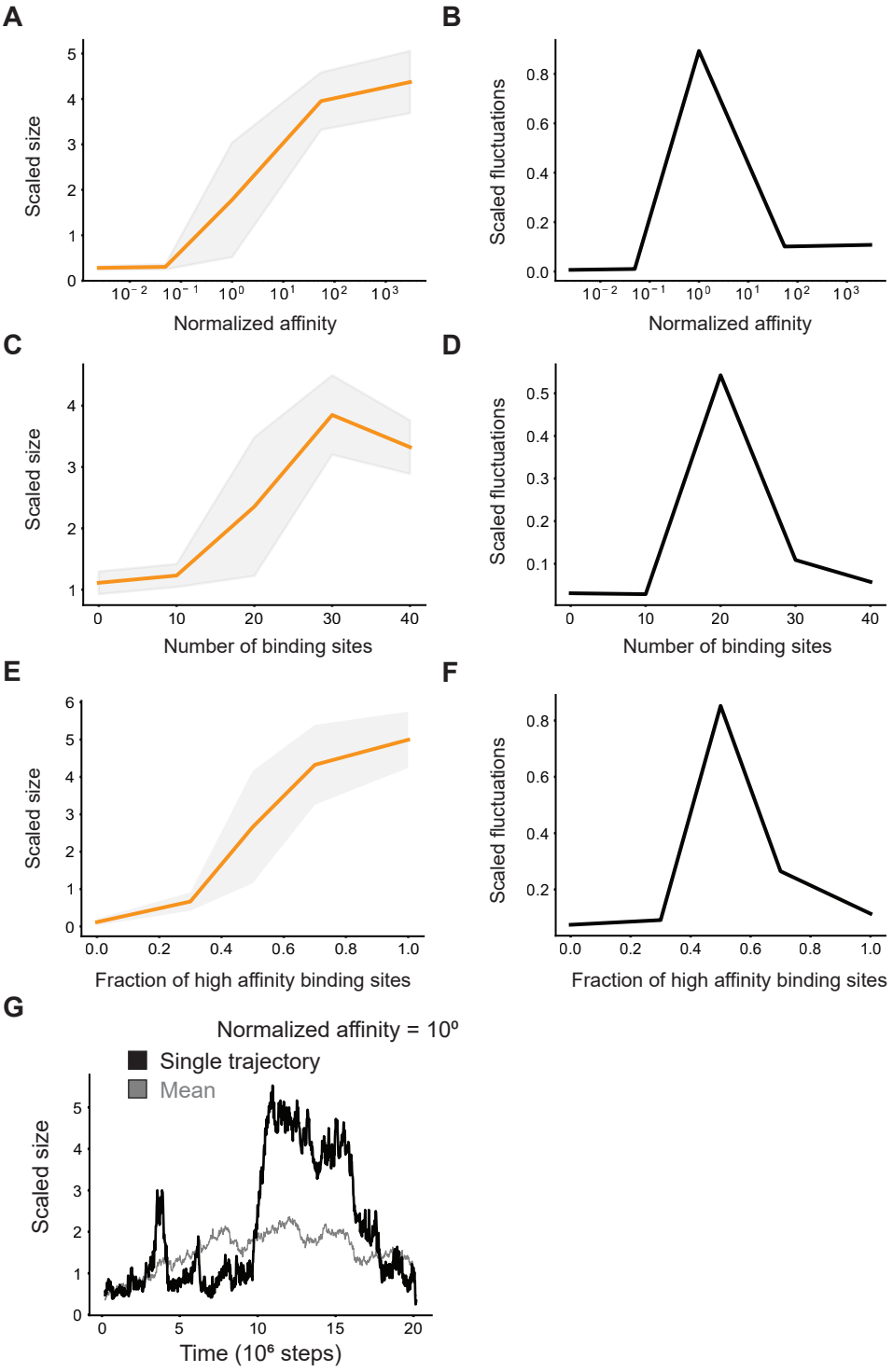


Figure S3: Normalized fluctuations exhibit a sharp peak across the transition point in scaled-size, characteristic of phase transitions, Related to Figures 3,4

Simulations predict a shift in scaled size from stoichiometric binding (≈1) to phase separation (>1) with increasing affinity for TF binding sites on DNA (**A**), valency of TF binding sites (**C**), or fraction of high-affinity binding sites (**E**).

Normalized fluctuations in scaled size (variance over mean) shows a peak near threshold affinity (**B**), valency (**D**), or fraction (**F**); affinity normalized to threshold affinity of E=12kT, fraction of binding sites normalized to total of 30.

G. Typical simulation trajectories show dynamic formation and disassembly of clusters (Transition between low and high scaled size) at threshold affinities.

Figure S4

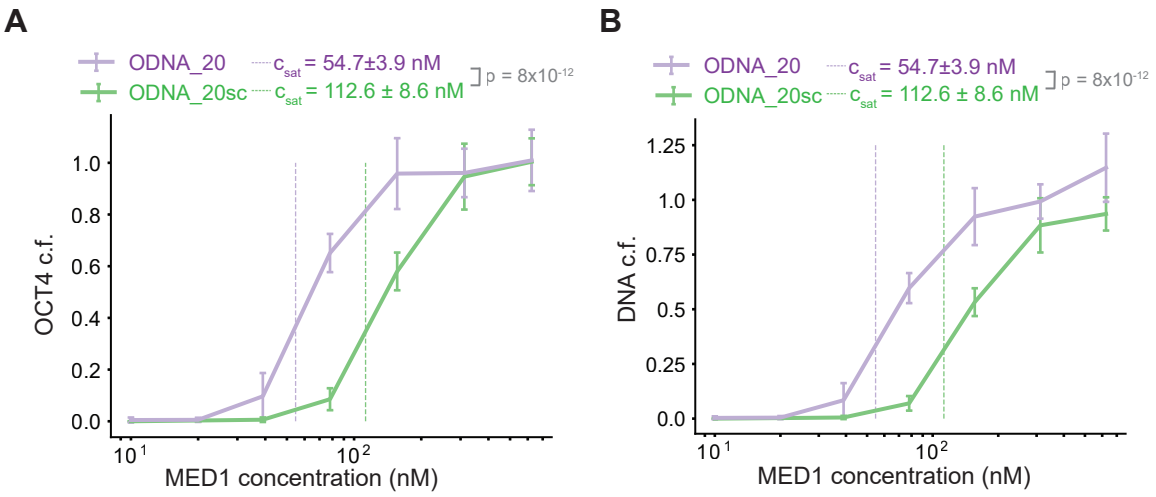


Figure S4: Phase separation of all components is promoted at lower coactivator concentrations by multivalent DNA with high density of TF binding sites, Related to Figures 3,4

A. Condensed fraction of OCT4 (in units of percentage) for ODNA_20 (purple) and ODNA_20sc (green) across a range of MED1-IDR concentrations.
B. Condensed fraction (in units of percentage) of ODNA_20 (purple) and ODNA_20sc (green) across a range of MED1-IDR concentrations.

Solid lines represent mean and error bars represent single standard deviations across replicates.

Figure S5

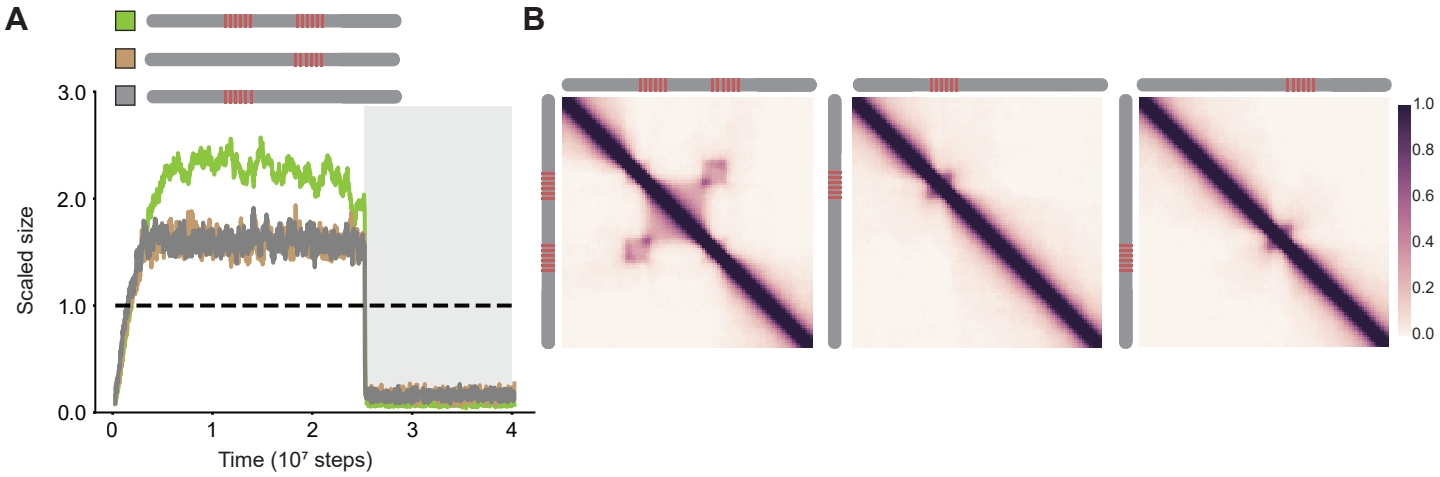


Figure S5: Patches of TF binding site interact over long distances to assemble the transcription machinery, Related to Figure 6

A. Scaled size versus simulation time steps comparing three different distribution of binding site number and distribution (as shown in the schematic legend). Dark gray background signifies disruption of TF-DNA interactions.
B. Contact frequency maps (see methods) show long-range interactions (right panel, checkerboard-like patterns) for DNA with different patch number and distribution.

Figure S6

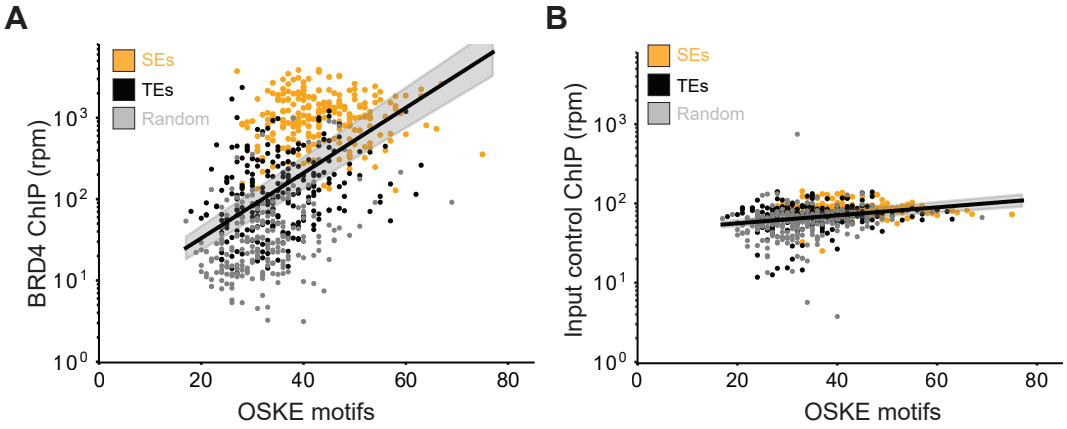


Fig S6. Mammalian genomes show correlation between high occupancy of coactivator and motif density at regulatory elements, Related to Figure 7

A. BRD4 ChIP-Seq counts (y-axis, reads-per-million) against total motifs of OCT4+SOX2+KLF4+ESRRB over 20kb regions centered on SEs (orange), TEs (black), and random loci (gray).
B. Same as **(A)** with data from sequenced input.

The black line represents a linear fit inferred between the logarithmic ChIP signal and motif count, and the grey shaded regions represent the 95% confidence intervals in the inferred slope. The linear model explains a sizable fraction of the observed variance ($R^2 \approx 0.28$) for the BRD4 signal, but not for input control ($R^2 \approx 0.07$).

Chapter 3 - Appendix

METHODS

Materials availability statement

All unique/stable reagents generated in this study are available from the Lead Contact upon reasonable request with a completed Materials Transfer Agreement.

Data/Code availability statement

The code generated during this study is available at:

https://github.com/krishna-shrinivas/2020_Henninger_Oksuz_Shrinivas_RNA_feedback.

Cell culture

The Jaenisch laboratory gifted the V6.5 mouse ES cells. ES cells were maintained at 37°C with 5% CO2 in a humidified incubator on 0.2% gelatinized (Sigma, G1890) tissue-culture plates in 2i medium with LIF, which was made according to the following recipe: 960 mL DMEM/F12 (Life Technologies, 11320082), 5 mL N2 supplement (Life Technologies, 17502048; stock 100X), 10 mL B27 supplement (Life Technologies, 17504044; stock 50X), 5 mL additional L-glutamine (Gibco 25030-081; stock 200 mM), 10 mL MEM nonessential amino acids (Gibco 11140076; stock 100X), 10 mL penicillin-streptomycin (Life Technologies, 15140163; stock 10^4 U/mL), 333 µL BSA fraction V (Gibco 15260037; stock 7.50%), 7 µL β-mercaptoethanol (Sigma M6250; stock 14.3 M), 100 µL LIF (Chemico, ESG1107; stock 10^7 U/mL), 100 µL PD0325901 (Stemgent, 04-0006-10; stock 10 mM), and 300 µL CHIR99021 (Stemgent, 04-0004-10; stock 10 mM). For confocal and PALM imaging, cells were grown on glass coverslips (Carolina Biological Supply, 633029) that had been coated

with the following: 5 μg/mL of poly-L-ornithine (Sigma P4957) at 37°C for at least 30 minutes followed by 5 μg/mL of laminin (Corning, 354232) at 37°C for at least 2 hours. Cells were passaged by washing once with 1X PBS (Life Technologies, AM9625) and incubating with TrypLE (Life Technologies, 12604021) for 3-5 minutes, then quenched with serum-containing media made by the following recipe: 500 mL DMEM KO (Gibco 10829-018), MEM nonessential amino acids (Gibco 11140076; stock 100X), penicillin-streptomycin (Life Technologies, 15140163; stock 10^4 U/mL), 5 mL L-glutamine (Gibco 25030-081; stock 100X), 4 μL β-mercaptoethanol (Sigma M6250; stock 14.3 M), 50 μL LIF (Chemico, ESG1107; stock 10^7 U/mL), and 75 mL of fetal bovine serum (Sigma, F4135). Cells were passaged every 2 days.

ChIP-seq analysis

ChIP-seq browser tracks for MED1, Pol II, BRD4, and OCT4 were generated as described (Sabari et al., 2018; Whyte et al., 2013). Briefly, reads were aligned to NCBI37/mm9 using Bowtie with the following settings: "-p 4 --best -k 1 -m 1 --sam -l 40". WIG files represent counts (in reads per million, floored at 0.1) of aligned reads within 50 bp bins. Each read was extended by 200nt in the direction of the alignment

(Source: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE112808)

GRO-seq analysis

For generation of the GRO-seq browser tracks, GRO-seq reads were processed as described in (Sigova et al., 2015). The GRO-seq .sra file corresponding to GEO accession number GSM1665566 (Sigova et al., 2015) was converted to .fastq using the SRA toolkit (Leinonen et al., 2011). Reads were aligned to the mouse genome (NCBI37/mm9) using Bowtie v1.2.2 (Langmead et al., 2009) with the following settings "-e 70 -k 1 -m 10 -n 2 --best". The reads corresponding to each one of the features

(super-enhancers, typical enhancers, proximal promoter regions, genes) were counted using featureCounts v1.6.2 (Liao et al., 2014) with default settings. The coordinates for typical enhancers and super-enhancers in mouse embryonic stem cells (mESCs) were acquired from (Whyte et al., 2013). The coordinates for genes (transcription start and end sites) were acquired using the UCSC Table Browser (Karolchik et al., 2004).The proximal promoter regions were defined as genomic areas containing 1 kb upstream of each TSS. Their coordinates were retrieved by using BEDTools v.2.26.0 (Quinlan and Hall, 2010) and the TSS coordinates as input (to the slop function). Reads were normalized with the size of the corresponding feature they aligned to.

RNA-seq analysis

The RNA-seq .sra file corresponding to GEO accession number GSM2686137b (Chiu et al., 2018) was converted to .fastq using the SRA Toolkit RNA-seq analysis was performed using the nf-core RNA-seq pipeline (v1.4.2) (Ewels et al., 2020) with default settings and NCBI37/mm9 as reference genome. Nextflow v20.01.0 was used as a workflow tool on an LSF High-Performance Computing environment (Di Tommaso et al., 2017). STAR v2.6.1d (Dobin et al., 2013) was used for the alignment of reads. Aligned reads were assigned to the aforementioned intervals (typical enhancers, super-enhancers, proximal promoter regions and genes) by using featureCounts v1.6.4, with the default settings.

Calculation number of RNA molecules in cells

Known concentrations of in vitro transcribed enhancer RNAs and pre-mRNAs from *Trim28* and *Pou5f1* loci are used as standards to approximate the number of molecules in cells. These RNAs are converted to cDNAs by reverse-transcription and mixed at equal concentrations. For each RNA species, a standard curve of qRT-PCR Ct value to RNA amount was generated using serial dilutions, with two different primer sets in technical duplicates. Next, qRT-PCR reactions using the same

primer sets were performed for biological duplicates of mESCs. Actb-normalized Ct values were then used to determine the amount of RNA species in the reaction based on the standard curves above. To calculate the number of RNA molecules per cell, the amount of RNA (g) was divided by the molar weight of each species (~350 (g mol^{-1} nt^{-1}) × length of in vitro transcribed RNA (nt)), multiplied by Avogadro's number (6.022×10^{23} mol^{-1}), and divided by the approximate number of cells used in each reaction (10,000 cells). Melting curves were analyzed to confirm primers specificity. Non-reverse-transcribed (–RT) controls were included to rule out the amplification of genomic DNA. Primer sequences are indicated in Table S1.

In vitro droplet assay

Recombinant GFP or mCherry fusion proteins were concentrated to a desired protein concentration using Amicon Ultra centrifugal filters (30K MWCO, Millipore). Droplet reactions with the recombinant proteins were performed in 10 ul volumes in PCR tubes under the following buffer condition: 30 mM Tris HCl pH 7.4, 100 mM NaCl, 2% Glycerol and 1 mM DTT. Droplet reactions with the Mediator complex were performed under the following buffer condition: 30 mM HEPES pH 7.4, 65 mM NaCl, 2% Glycerol and 1 mM DTT. For all droplet reactions, protein and buffer were mixed first and RNA was added later. The reactions were incubated at room temperature for 1 hr. The reactions were then transferred into 384 well-plate (Cellvis P384-1.5H-N) 5 minutes prior to imaging on a confocal microscopy at 150X magnification. The concentration of proteins and RNAs in the droplet reactions are indicated in the figure legends.

Synthesis of RNA by in vitro transcription

Enhancer and promoter sequences for RNAs were obtained from super-enhancer-regulated genes *Pou5f1*, *Nanog*, and *Trim28*. For promoter sequences, the first 475-490bp from the first exon were selected from

mm10. For enhancer sequences, GROseq reads (Sigova et al., 2015) from both + and - strands aligned to mm9 were overlapped with called super-enhancers (Whyte et al., 2013). Contiguous regions of read density above background were manually selected (Figure S1). Primers were designed to amplify the selected promoter and enhancer sequences from genomic DNA isolated from V6.5 mESCs (Table S1). The following sequences were added to the forward and reverse primers to add the bacterial polymerase promoters:

T7 (add to 5' of sense or forward primer): 5'-TAATACGACTCACTATAGGG-3'

SP6 (add to 5' of antisense or reverse primer): 5'-ATTTAGGTGACACTATAGAA-3'

Phusion polymerase (NEB) is used to amplify the products with the bacterial promoters, and products are run on a 1% agarose gel, gel-purified using the Qiaquick Gel Extraction Kit (Qiagen), and eluted in 40 µL H2O. Templates were sequenced to verify their identity. A volume of 8 µL of each template (10-40 ng/µL) was transcribed using the MEGAscript T7 (Invitrogen; sense) or MEGAscript SP6 (Invitrogen; antisense) kits according to the manufacturer's instructions. For visualization of the RNA by microscopy, reactions included a Cy5-labeled UTP (Enzo LifeSciences ENZ-42506) at a ratio of 1:10 labeled UTP:unlabeled UTP. The in vitro transcription was incubated overnight at 37°C, then 1 µL TURBO DNAse (supplied in kit) was added, and the reaction was incubated for 15 minutes at 37°C. The MEGAclear Transcription Clean-Up Kit (Invitrogen) was used to purify the RNA following the manufacturer's instructions and eluting in 40 µL H2O. RNA was diluted to 2 µM and aliquoted to limit freeze/thaw cycles, and RNA was run on 1% agarose gels in TBE buffer to verify a single band of correct size.

Recombinant protein purification

Recombinant protein purifications were performed as previously reported

(Boija et al., 2018, 2018; Guo et al., 2019; Sabari et al., 2018; Shrinivas et al., 2019; Zamudio et al., 2019). Briefly, pET expression plasmids containing 6xHIS tag and genes of interest or their IDRs tagged with either mEGFP or mCherry were transformed into LOBSTR cells (gift of I. Cheeseman Lab). Expression of proteins were induced by addition of 1mM IPTG either at 16°C for 18 hours or at 37°C for 5 hours. Extracts were prepared as previously described (Boija et al., 2018). Proteins were purified by Ni-NTA agarose beads (Invitrogen, R901-15), and eluted with 50mM Tris pH 7.4, 500mM NaCl, 250mM imidazole buffer containing complete protease inhibitors (Roche,11873580001). Proteins were dialyzed against 50mM Tris pH 7.4, 125 mM NaCl, 10% glycerol and 1mM DTT at 4°C for OCT4-GFP and GFP alone and the same buffer containing 500 mM NaCl for MED1-IDR-GFP.

Purification of human Mediator complex from HeLa nuclear extract.

HeLa nuclear protein extract (4g) was prepared as described in (Dignam et al., 1983). Nuclear extract was dialyzed against BC100: BC buffer, pH 7.5 + 100mM KCl (20 mM Tris-HCl, 20 mM β-Mercaptoethanol, 0.2 mM PMSF, 0.2 mM EDTA, 10% glycerol (v/v) and 100 mM KCl). The extract was fractionated on a phosphocellulose column (P11) with BC buffer containing 0.1, 0.3, 0.5 and 1M KCl. The Mediator complex eluted in the 0.5M KCL (BC500) fraction. This fraction was dialyzed against BC100 and loaded on a DEAE Cellulose column and sequentially fractionated with BC buffer containing 0.1, 0.3 and .5M KCl. The Mediator did not bind the DEAE Cellulose resin and was collected in the flow through fraction 0.1M KCl (BC100). This fraction was then directly loaded onto a DEAE-5PW column (TSK) and eluted with a linear KCl gradient from 0.1 to 1M KCl in BC buffer. The Mediator complex eluted between 0.4 and 0.6M KCl. The fractions containing Mediator were pooled and dialyzed against BD700: BD buffer, (20 mM Hepes pH 7.5, 20 mM β-Mercaptoethanol, 0.2 mM PMSF, 0.2 mM EDTA, 10% glycerol, and 700 mM (NH4)2SO4). This fraction was then loaded onto a Phenyl-

Sepharose Hydrophobic Interaction Chromatography (HIC) column and eluted with a linear reverse gradient from 0.7 to 0.025M (NH4)2SO4 in BD buffer. The Mediator complex eluted between 0.3 and 0.1M (NH4)2SO4. The Mediator-containing fractions were again pooled and dialyzed against BA100: BA buffer, pH 7.5 + 100 mM NaCl (20 mM Hepes, 20 mM β-Mercaptoethanol, 0.2 mM PMSF, 0.2 mM EDTA, 10% glycerol and 100 mM NaCl) and loaded onto a Heparin Agarose column. The column was washed with BA100 and step-eluted with BA buffer containing 0.25, 0.5, 1M and 1M NaCl. The Mediator complex eluted in the 0.5M NaCl (BA500) fraction. A portion of this fraction was then loaded on a Superose-6 (gel filtration column) that was equilibrated and run in BC100. The Mediator complex eluted from the gel filtration column with a mass range between 1-2MDa.

Reconstituted in vitro transcription assay

The reconstituted in vitro transcription by RNA polymerase II was performed as previously described (Flores et al., 1992; LeRoy et al., 2008; Orphanides et al., 1998) with some modifications. A template DNA containing adenovirus major late promoter, five Gal4 binding sites, TATA-box sequence and a 561 bp from eGFP sequence was used. First, pre-initiation complex was assembled at RT for 15 min by mixing the following components: 50 nM RNA polymerase II, 50 nM general transcription factors (TFIIA-B-D-E-F-H), and 5.75 nM template DNA, in a buffer containing 10 mM HEPES pH 7.5, 65 mM NaCl, 6.25 mM MgCl2, and 6.25 mM Sodium butyrate. Next, 10 nM Mediator complex and 10 nM Gal4 were added to the reaction. Last, nucleotide mix containing 0.375 mM ATP, CTP, UTP, CTP (Invitrogen), 0.01 U RNase Inhibitor (Invitrogen), 1.25 % PEG-8000 and various amounts of purified exogenous *Pou5f1* RNA (0-500 nM) are added. The reaction was incubated at 30°C for 2 hr. RNA isolation was performed using RNeasy kit (Qiagen) by including a spike-in RNA control and an RNA carrier. Purified RNAs were treated with ezDNase (Invitrogen) for 30 min at 37°C to eliminate the template DNA. Reverse transcription was performed

using Superscript IV (Invitrogen) and qPCR was performed with SYBR Green Real Time PCR master mix (Invitrogen) to quantify the template derived transcriptional output. The Ct values of the reactions were normalized to the spike-in RNA control. The concentration of template derived transcriptional output was calculated by using a standard curve of qRT-PCR Ct values generated by known amounts of serially diluted GFP RNA. The sequence of primers used for qRT-PCR are indicated in Table S1.

Constructing a free-energy for RNA-protein phase behavior

Our goal in this section is to develop a simplified and coarse-grained model that captures the qualitative physics of RNA-protein mixtures. Based on phenomenological observations of transcriptional proteins and RNA (Figure 2), such a model must recapitulate the following key features:

Transcriptional proteins phase separate in the absence of RNA through other types of interactions, albeit at higher concentrations.

At fixed protein concentrations, addition of RNA initially promotes de-mixing and at higher levels drive a re-entry into the mixed phase.

Motivated by the evidence that transcriptional condensates recruit diverse coactivators, transcription factors, and other proteins of the transcriptional apparatus (Boija et al., 2018; Guo et al., 2019; Sabari et al., 2018; Shrinivas et al., 2019), we define an effective protein component P that lumps together different transcriptional molecules. Similarly, while different species of RNA are likely present within these condensates, we define an effective RNA species (R).

Landau model

First, we approach this problem by constructing a phenomenological free-energy with 2 order-parameters that represent scaled concentrations of protein ($\phi_p(\vec{r},t)$) and RNA ($\phi_r(\vec{r},t)$). We define the free-energy (normalized to $k_BT = 1$) as:

$$f[\phi_p, \phi_r] = \int_V d^dV \left(f_{dw}\left(\phi_p(\vec{r}, t)\right) + \rho_r \phi_r^2 + \chi_{eff}\left(\phi_p(\vec{r}, t), \phi_r(\vec{r}, t)\right) + \frac{\kappa}{2}\left(\nabla \phi_p\right)^2 \right)$$

Here, $f_{dw}\left(\phi_p(\vec{r}, t)\right) = \rho_s\left(\phi_p - \alpha\right)^2\left(\phi_p - \beta\right)^2$ is a standard double-well potential that ensures protein components phase separate without RNA with co-existence concentrations specified by α, β. Choice of $\kappa > 0$ ensures that there is finite surface tension for the protein condensate. The second-order term for RNA ($\rho_r > 0$) states that within this model-framework, RNA cannot phase-separate in the absence of protein. Given that electrostatic interactions at physiological salt conditions are fairly short-ranged (Debye length ~ 1nm), we capture the non-linear nature of RNA-protein interactions in an effective interaction term χ_{eff}. We define this interaction term in the spirit of the Landau-Ginzburg approach as an expansion in powers of the order parameters:

$$\chi_{eff}\left(\phi_p, \phi_r\right) = -\chi \phi_p \phi_r + a\phi_p \phi_r^2 + b\phi_p^2 \phi_r + c\phi_p^2 \phi_r^2 + \cdots + H.O.T$$

While symmetry arguments often dictate or exclude certain types of terms (odd powers in Ising models for example) in such an expansion, there are no obvious symmetry constraints for this system. Hence, our modeling approach is to minimize the number of higher-order terms that need to be included to recapitulate the experimentally observed reentrant phase transition. Our experimental results suggest that low concentrations of RNA promote phase separation, and thus the lowest order term ($-\chi \phi_p \phi_r, \chi > 0$) lowers the free-energy. However, higher-order terms must counter this and below we outline how we determine which terms to include. In general, the stability of a mixture described by such a free-energy can be ascertained from the Jacobian matrix J. For our model, the elements of this 2×2 matrix are:

$$J_{pp} = \frac{\partial^2 f}{\partial \phi_p^2} = 2\rho_p\left(6\phi_p^2 - 6\phi_p(\beta + \alpha) + (\alpha - \beta)^2\right) + 2b\phi_r + 2c\phi_r^2 J_{pr}$$

$$= \frac{\partial^2 f}{\partial \phi_p \partial \phi_r} = -\chi + 2a\phi_r + 2b\phi_p + 4c\phi_p \phi_r$$

$$J_{rr} = \frac{\partial^2 f}{\partial \phi_r^2} = 2\rho_r + 2a\phi_p + 2c\phi_p^2$$

The mixed phase is no longer stable to perturbations when at least one eigen value of J becomes negative (spinodal instability). In the absence of RNA, the spinodal satisfies $J_{pp} = 0$. If only the pair-wise interaction terms were considered ($-\chi\phi_p\phi_r$), the spinodal region broadens i.e. phase separation is promoted at lower protein concentrations when RNA is present. We next characterized the effect of an additional higher-order term (only one of a, b or c is non-zero) on the Jacobian matrix. Briefly, we ascertained that:

$a > 0$: While the free-energy is dominated by repulsive interactions at higher RNA concentrations, the Jacobian matrix predicts a continuous underlying instability. Instead of suppressing phase separation at higher RNA concentrations and promoting re-entry to dilute phase, this term would instead change the composition of the demixed phases.

$b > 0$: While this term promotes a reentrant behavior, the resulting regions of instability demix RNA away from protein for most values of b.

$c > 0$: For values of c that are not too large (i.e. $c <\approx \rho_r$), the resulting phase diagram mirrors a reentrant shape with RNA enrichment in the protein condensate. If c is moderately large, then a second de-mixing transition (similar to case 2) is observed at high values of ϕ_p, ϕ_r. Since we are interested in the limit of relatively low protein/RNA concentrations, and the values of ϕ_p, ϕ_r represent qualitative proxies of protein/RNA concentrations, we choose to explore our model in this parameter regime.

While cubic and higher-order terms are required to recapitulate complete phase-behavior, we explored our model with $c > 0$, assuming the coefficients a, b are small. In the simulations reported in Figures 5-7, the free-energy parameters are $\alpha = 0.1, \beta = 0.7, \chi = 1.0, c = 10.0, \kappa = 0.5, \rho_s = 1.0, \rho_r = 10.0, a = b = 0$. All free-energy calculations were performed with *Python* and code is available at:

https://github.com/krishna-

shrinivas/2020_Henninger_Oksuz_Shrinivas_RNA_feedback .

Flory-Huggins model

In this approach, rather than employ a phenomenological model, we parametrize a mechanistic model motivated by Flory-Huggins polymer-solution theory (Flory, 1942). The simplified F-H model contains 3 components - protein, RNA, and the solvent (s), whose volume fractions are defined as $\phi_p(\vec{r},t), \phi_r(\vec{r},t), 1-\phi_p(\vec{r},t)-\phi_r(\vec{r},t)$ respectively. The free-energy (normalized as before) is defined as:

$$f = \sum_i \frac{\phi_i}{r_i} log(\phi_i) + \sum_{i,j>i} \chi_{ij}\phi_i\phi_j$$

Here, r_i are the solvent-equivalent polymerization lengths of the RNA & protein (assumed to be equal for simplicity) and χ_{ij} are the various pairwise interaction terms. As before, we assume these interactions to be short-ranged at physiological salt levels. Choice of $\chi_{pr} > \chi_{ps} > 0$ and $\chi_{rs} < 0$ recapitulate the attractive contributions of protein-protein/protein-RNA interactions and repulsive RNA-RNA interactions. With these choices of constraints, the resulting free-energy looks similar to the phase diagram from the Landau approach with $c > 0$ (Figure S5B) where the key F-H parameters are χ_{pr}=1.1,χ_{ps}=0.75,$\chi_{rs} = -0.6, and\ r_p = r_r = 30$.

Numerical phase-field simulations

Numerical investigations of the coupled-equations outlined in Figure 5C were performed with the FiPy package (Guyer et al., 2009). Simulations were performed on a 2-D/3-D square lattice ($L_x = L_y = 60, dx = 1.0; L_x = L_y = L_z = 40, dx = 1.0;$) and with adaptive time-stepping ($dt_{min} = 1e-8, dt_{max} = 5e-1$) until steady state is reached (which typically requires ~ 10000 simulation steps).

The chemical potential for the protein components is calculated as:

$$\mu_p = \frac{df}{d\phi_p} = 2\rho_s(\phi_p - \alpha)(\phi_p - \beta)(2\phi_p - \alpha - \beta) + \kappa\nabla^2\phi_p - \chi\phi_r + 2c\phi_r^2\phi_p$$

The radius of condensates was inferred from the volume of mesh regions where $\phi_p \geq \frac{\alpha+\beta}{2}$. The mobility of RNA and protein were chosen to be 1.0 unless mentioned elsewhere. The raw data for all figures from simulation data are provided along with the manuscript.

Design of Simulations to vary RNA features and rates of RNA synthesis

We designed simulations (Figure 7C) to study the effect of RNA features and rates of effective synthesis on condensate size. The rates of synthesis were changed by increasing k_p by multiplicative factors (see x-axis in Figure 7C). Since RNA length is not explicitly incorporated in the model framework, we defined the effective local synthesis rates of longer RNA as a product of k_p and an additional multiplicative factor (1,2, and 4x for short, medium, and long RNA respectively) to mimic increased local concentrations of RNA.

Calculation of number of charged molecules in condensates

In estimating the number and charge of transcriptional proteins (Figure 1A), we use previous estimates (Cho et al., 2018) that suggest key transcriptional proteins such as Mediator are present at 10-100 molecules in transcriptional condensates. Further, molecules such as MED1 or BRD4 contain large disordered domains with net positive charge of +5 to to+40. This provides a highly approximate estimate of 25-500 as the

effective positive charge. Since there are many more transcriptional proteins and most proteins tend to contain net positive charges, it is likely that this estimate represents lower bounds on the range. Steady-state levels of nascent eRNA (Figure S1) suggest a range of 0.2-10 molecules, and since super-enhancers typically contain clusters of such active enhancers, we approximate the typical range of eRNA molecules at a transcriptional condensate between 1-10. Since RNA carries a charge of around -1 per nt (Banerjee et al., 2017) and eRNAs are short (<1 kb), we estimate the effective negative charge during initiation to be in the range 10-1000. During productive elongation, mRNAs are produced in bursts ranging from few to tens (1-50) and are typically longer (>1kb), suggesting a conservative estimate of the effective charge to range from (1000-100,000). It is important to stress that our approximations are performed with the aim of obtaining order-of-magnitude estimates and do not account for factors such as local composition of different proteins or extent to which nascent mRNAs may be coated by RNA-binding proteins. With the above numbers, we estimate concentrations based on a typical transcriptional condensate of size r=0.25 μm (Cho et al., 2018) that suggests that eRNA concentrations range about 10-200 nM and transcriptional proteins range 1-20 μM within the condensate.

Reactive/diffusive time-scales and estimates in cells

As defined in the model (Figure 5B), the key rates of synthesis/degradation reactions are kp/kd, which have units of s^{-1}, and thus the relevant time-scales are $t_r = {k_p}^{-1} (or\ {k_d}^{-1})$. Timescales of RNA transport depend on both the diffusivity as well as the size of the condensate (L) and is defined as $t_d = L^2/M_{rna}$. We approximated the diffusivity of the nascent transcript by that of chromatin, which ranges from $10^{-3.5} - 10^{-2}\ \mu m^2/s$ (Gu et al., 2018). By assuming a typical eRNA of size 100nt and Pol II transcription rates as $\sim 20 - 70 \frac{nt}{s}$ (Maiuri et al., 2011) we inferred typical synthesis rates of $\sim 0.2 - 0.8$ s^{-1}Pol II^{-1}. In our previous work (Cho et al., 2018), we have seen that clusters that contain multiple polymerases (>5) are typically around $r \approx 200nm$. This

allows us to approximately obtain the ratio of diffusive and reactive time-scales as $\frac{t_d}{t_r} = \frac{kr^2}{M} \approx 5 - 50$.

Calculation of charge balance

Charge-balance calculations were performed (Figures 3, S4) employing the following method. Net protein charge per molecule was calculated as $C_p = \#(R, K) - \#(D, E)$ for the relevant sequence including the GFP tag. RNA charge per molecule was calculated as $C_r = -(\#\ of\ bp)$, assuming an approximate charge of -1 per nucleotide (Lin et al., 2019). Next, the charge balance ratio was computed at a particular RNA and protein concentration as:

$$Charge - balance\ ratio = \frac{\left(C_p[P], C_r[R]\right)}{(C_p[P], C_r[R])}$$

The effective concentration of MED1-IDR in our assays was $1000\ nM$. Our results were not quantitatively affected by inclusion/exclusion of the partial charge on Histidine residues, partly due to their low frequency on the protein sequences. The code for performing these calculations are available at:

https://github.com/krishna-shrinivas/2020_Henninger_Oksuz_Shrinivas_RNA_feedback.

Transcription inhibition by small molecules

For small molecule inhibition experiments, cells were treated with 100 μM DRB (Sigma), or 1 μM Actinomycin-D (Sigma) in 2i media (detailed

above) for 30 minutes, then imaged. For wash-out experiments, media was replaced with fresh 2i media and cells were allowed to recover for 1 hour, then the cells were imaged.

Condensate size

Cells with endogenously-tagged Med1-GFP (Sabari et al., 2018) were plated on 20 mm glass-bottom dishes (Mattek) coated with poly-L-ornithine (Sigma) and laminin (ThermoFisher). Mock (DMSO) and treated cells were imaged on a LSM 880 Confocal Microscope with Airyscan to obtain super-resolution z-stacks for at least 8 different fields containing multiple cells. For quantification, a manual threshold was applied equally across all conditions to remove background, and the size of Med1-GFP puncta was quantified in 3D using the 3D object counter plugin (Fiji/ImageJ).

Condensate lifetime

HaloTag was endogenously knocked into 5'-end of Med19 via homology-directed repair (HDR) in mouse embryonic stem cells (R1 mESCs). Three single-guide RNAs (sgRNAs) targeting +/- 100 bps from the start codons of Med19 gene were designed using the web-based CRISPR Design tool (Andrews et al., 2018) and integrated into a *Streptococcus pyogenes* Cas9 vector (Addgene #62988) for standard CRISPR/Cas9 editing. Single positive colonies were sorted by fluorescence-activated cell sorting (FACS) and validated under the microscope.

Cells were cultured in serum-free 2i medium on poly-L-ornithine (PLO) and Laminin-coated flasks for more than two days and then were transferred onto coated imaging dishes for another day. Before imaging, cells were stained with (PA)-JF549-HaloTAG dye (a gift from Luke Lavis Lab, Janelia Research Campus) of 100nM concentration for 2 hours followed by a 60-minute wash in fresh 2i medium. Lastly, dishes were filled in with 2ml Leibovitz's L-15 Medium (no phenol red, Thermo

Fisher) and brought to the microscope for imaging.

Photo-activation localization microscopy (PALM) imaging was performed using a Nikon Eclipse Ti microscope with a 100x oil immersion objective (NA 1.40) (Nikon, Tokyo, Japan). A 405nm beam of 100mW power (attenuated with 25% AOTF) and a 561nm beam of 500mW power were columnated and superposed to perform simultaneous activation and excitation. The combined beam was expanded and re-collimated with an achromatic beam expander (AC254-040-A and AC508-300-A, THORLABS) to improve the uniformity of illumination across the whole region of interest (ROI 2562 pixels). Images were acquired with an Andor iXon Ultra 897 EMCCD camera (gain 1000, exposure time 50ms) interfaced through Micro Manager 1.4. 2400 frames were acquired for each imaging cycle. The cells were maintained at 37°C in a temperature-controlled platform (InVivo Scientific, St. Louis, MO) on the microscope stage during image acquisition. Med19-Halo cluster lifetimes were calculated as previously described using the qSR software (dark time tolerance = 20 frames, min cluster size=50) (Andrews et al., 2018), and a cumulative distribution was generated using Prism software (GraphPad).

Reporter assay to determine the effect of local RNA synthesis on transcription

Vectors used in the reporter assay are modified from pTETRIS-cargo vector, gift from J. M. Calabrese (Kirk et al., 2018). 6X STOP codon sequence was cloned into NotI digested pTETRIS-cargo vector using Gibson cloning strategy by following the manufacturer's instructions (NEB). This vector is called pTETRIS-cargo-STOP. Using Gibson cloning strategy (NEB), various RNA sequences were cloned downstream of the 6X STOP sequence to prevent translation of these RNAs. Stable cell lines for individual RNAs were generated by transfecting Med1-GFP mESCs with the following vectors: 0.5 µg pTETRIS-cargo-STOP containing individual RNAs, 0.5 µg rTTA-cargo, gift from J. M. Calabrese (Kirk et al.,

2018), and 1 µg piggyBAC transposase (Systems Biosciences). Cells were selected on puromycin (2 µg/ml) and G418 (200 µg/ml) for 1 week for successful integrations. For luciferase assays, $1x10^5$ cells of each genotype were plated in triplicate on 0.2%-gelatin-coated 24-well plates and allowed to settle overnight. Cells were treated with doxycycline (Sigma) and harvested after 24 h to measure either luciferase activity or to purify RNA. Luciferase activity was measured using the Luciferase Assay System (Promega) according to manufacturer instructions. Luciferase signal was normalized to total protein content, measured by BCA protein assay kit (Invitrogen, #23227), and then normalized to a control not treated with doxycycline. To measure RNA expression, RNA was purified using the Qiagen RNeasy Mini kit (Qiagen) according to manufacturer instructions, cDNA was generated by Superscript III (Invitrogen) according to manufacturer instructions, and 10 ng of cDNA was used in a qRT-PCR SYBR-green reaction (Life Technologies) with primers specific to a common sequence shared across the vectors (qPCR_Tetris, Table S1). Ct values were normalized to a housekeeping gene (qPCR_mActb, Table S1) and a control condition with no doxycycline treatment.

For imaging experiments in Figure 7D, the reporter construct was modified using Gibson cloning to include a 6X-MS2 hairpin (Jain and Vale, 2017) at the 3' end of the RNA sequence. Cell lines with this construct were generated as detailed above in a mESC background with endogenously-tagged Med1-GFP (Sabari et al., 2018). $1x10^6$ reporter cells were plated on 20 mm glass-bottom dishes (Mattek) coated with poly-L-ornithine (Sigma) and Laminin (ThermoFisher), then transfected with a vector that expresses an MCP-SNAP fusion (CMV-2xMCP-SNAP) using Lipofectamine 3000 (ThermoFisher) according to the manufacturer instructions. After overnight transfection, cells were treated with 10 ng/mL doxycycline for 4 hours, then incubated with 100 nM SNAP ligand JF-549 (Gift from Janelia Farm) for 30 minutes, then washed with fresh media. Cells were subsequently incubated with Hoechst-33342 (Life Technologies) for detection of nuclei in live cells and images were acquired on an RPI Spinning Disk Confocal. Images were Gaussian filtered (sigma=2), and for the SNAP channel, images were median

filtered (21 pixels) and then subtracted from the original image to remove background. Line plots were analyzed in the original, unprocessed images, and they were generated from the diagonal of the boxes shown in Figure 7D for a single z-slice.

SUPPLEMENTARY FIGURE TITLES AND LEGENDS

Figure S1 - Transcription machinery and RNA at active genes in murine embryonic stem cells (Related to Figure 1)

A. B. C. Enrichment of transcription machinery and RNA at *Trim28* (A), *Pou5f1* (B) and *Nanog* (C) super-enhancers in mESCs. Gene tracks of ChIP-seq and nascent RNA-seq data at the indicated super-enhancers are shown. The enhancer- and promoter-derived RNAs that are used in this study are annotated in the gene tracks.

D. Nascent (left) or steady-state (right) levels of indicated RNAs at super and typical enhancers.

E. Quantification of the number of enhancer RNA and pre-mRNA molecules in cells. Calculations are based on two biological replicates (STAR Methods).

Figure S2 - RNA-mediated stimulation and dissolution of pre-formed MED1-IDR droplets (Related to Figure 3)

A. Experimental design to test the effect of RNA on pre-formed MED1-IDR droplets.

B. Representative images of MED1-IDR droplets. Indicated concentrations of RNA were added after formation of droplets with 1 μM of MED1-IDR.

C. Quantification of partition ratios of MED1-IDR-GFP (left) and cy5-labeled RNA (right) within the droplets in (B).

Figure S3. Control experiments for the effect of RNA on droplet formation (Related to Figure 3)

A. B. Purified GFP (A) or OCT4-GFP (B) was incubated with an enhancer RNA from the *Pou5f1* locus. Whereas this RNA could stimulate MED1-IDR-GFP condensate formation, it was unable to form droplets with GFP alone or OCT4-GFP. Images were adjusted to show signal and lack of droplet formation.

Figure S4. Modulation of charge balance contributes to stimulation and dissolution of MED1-IDR condensates by RNA (Related to Figure 3)

A. Experimental design for testing diverse sense and antisense RNAs of different lengths on formation of MED1-IDR-GFP droplets.
B. Quantification of the partition ratios of MED1-IDR-GFP within the droplets when incubated with RNAs of different lengths and sequences.
C. Quantification of the partition ratios of MED1-IDR-GFP within the droplets when incubated with antisense versions of the RNAs in (B).
D. Representative images of MED1-IDR droplets (left), which are formed with or without RNA and are subjected to increasing concentration of monovalent salt (NaCl). Quantification of partition ratios of MED1-IDR-GFP within the droplets are indicated (right).

Figure S5 - Computational model for non-equilibrium RNA feedback on transcriptional condensates (Related to Figure 5)

A. Regions where mixtures of protein and RNA phase separate spontaneously (red, left panel) are calculated from the Landau free-energy (Figure 4C) by analyzing the Jacobian (spinodal analyses, STAR Methods). As expected from the re-entry

transition, increasing RNA concentration (abscissa) at fixed protein levels can start from a region promoting phase separation, and beyond a threshold, drive re-entry into dilute phase. The right panel shows the initial direction of the instability (STAR Methods), which indicates the RNA is enriched in protein condensates (value>0, green shade), while at higher concentrations, RNA de-densifies the condensed phase (value<0).

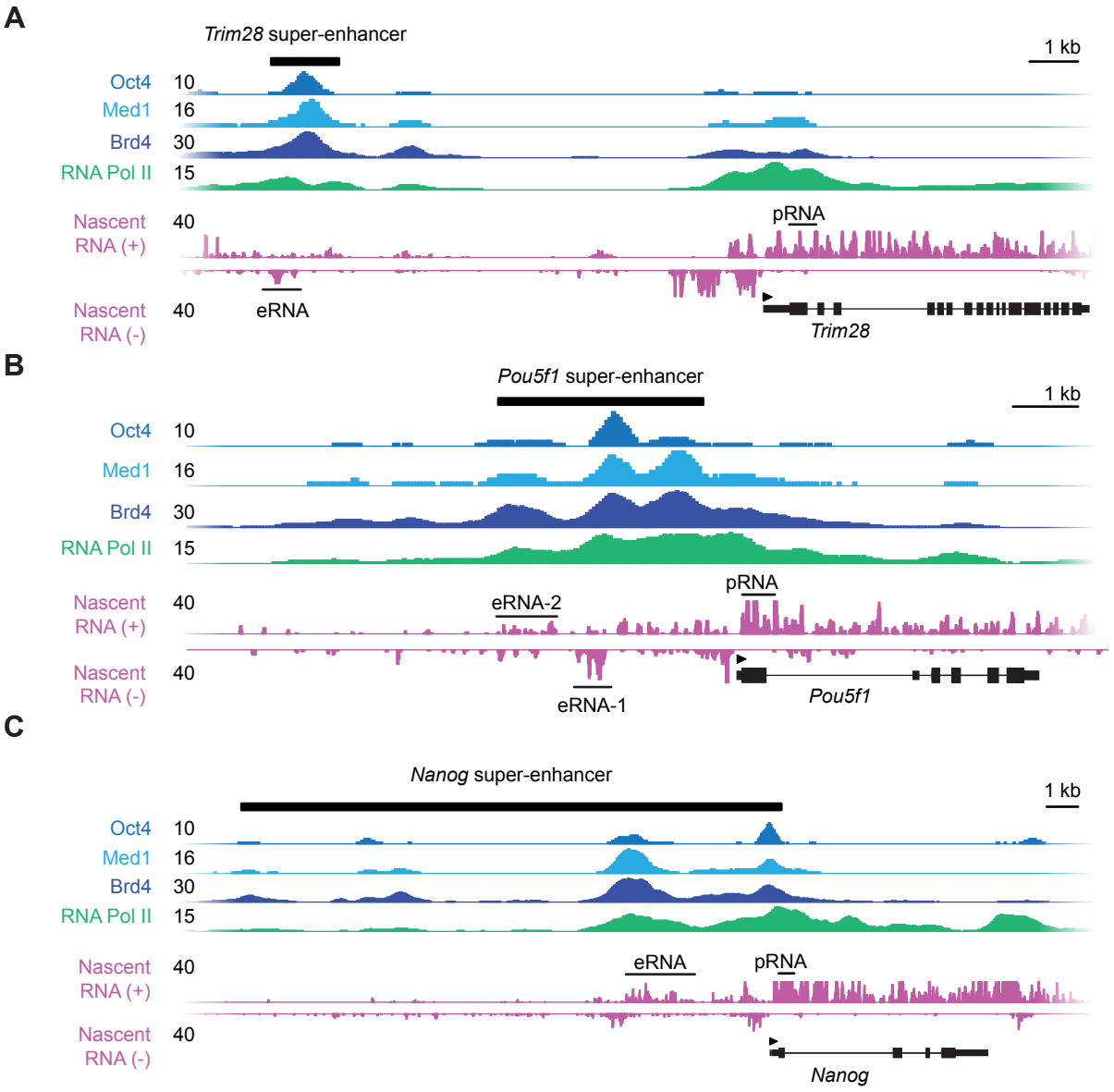
B. Similar analyses as in (A) are performed on a free-energy derived from Flory-Huggins model (STAR Methods).

C. Variation of condensate radius (left panel, normalized to value of R at k_p=0.02) and condensate lifetime (right panel) with effective rates of RNA synthesis (k_p, abscissa) for simulations performed in 3D employing the Landau free-energy (Figure 4C, STAR Methods). Low values of k_p promote condensate stability whereas higher rates drive dissolution. The dashed line in the right-panel represents the conditions under which condensates are stable in the simulations and condensate lifetime is presented in units of simulation time (STAR Methods).

D. Similar analyses as in (C) are performed on a free-energy derived from Flory-Huggins model on a 2D grid (STAR Methods). Values of the condensate radius are normalized to value of R at $k_p = 0.08$..

E. Partition ratio, computed as maximum RNA concentration in condensate divided by dilute phase concentrations, are presented for simulations employing the Landau free-energy in 2 & 3-D as well as those employing the Flory-Huggins model in 2-D (left to right). When condensates are dissolved, the expected value of this ratio is 1 (as depicted by dashed gray lines). These calculations correspond to simulation data from Figures 5D-E, S5C, and S5D respectively (left to right).

Figure S1 | Transcription machinery and RNA at active genes in murine embryonic stem cells (Related to Figure 1)



E

RNA	Number of molecules per cell
Trim28 enhancer RNA	0.2 +/- 0.08
Pou5f1 enhancer RNA	0.2 +/- 0.04
Trim28 pre-mRNA	14.7 +/- 5.3
Pou5f1 pre-mRNA	2.1 +/- 0.16

Figure S2 - RNA-mediated stimulation and dissolution of pre-formed MED1-IDR droplets (Related to Figure 3)

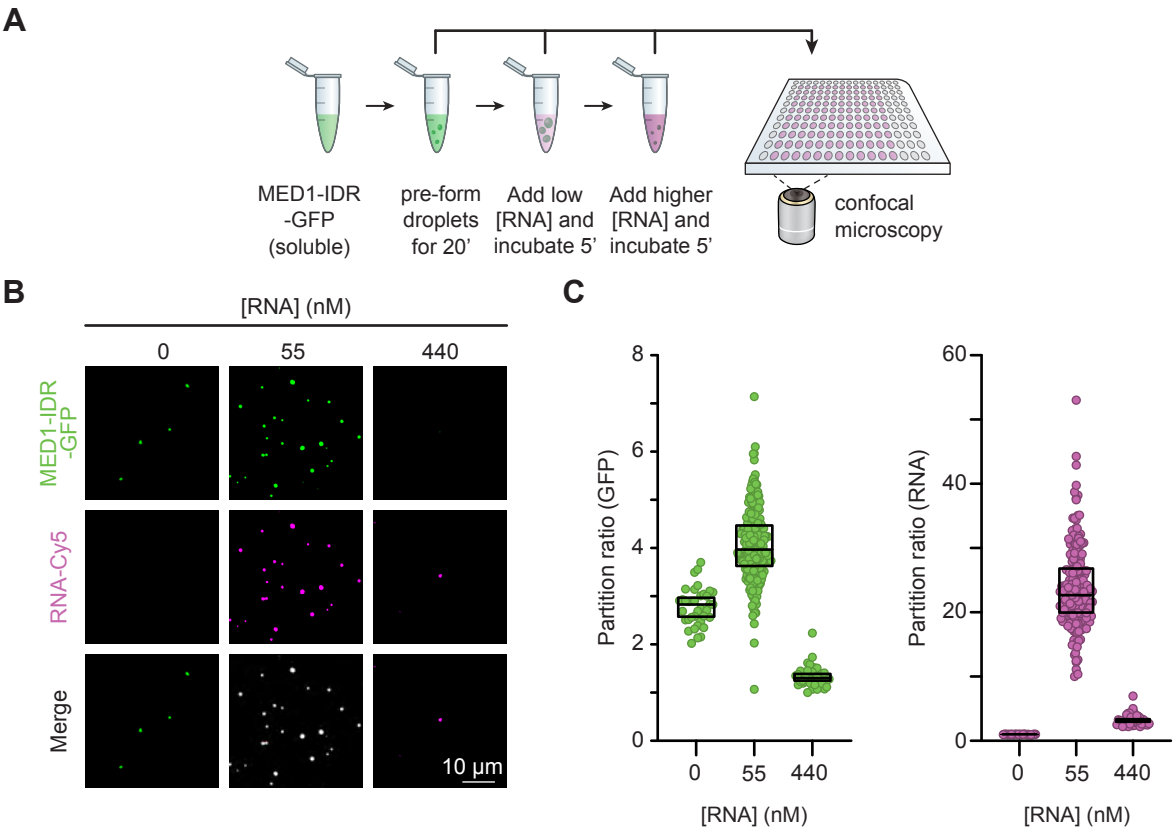


Figure S3 | Control experiments for the effect of RNA on droplet formation (Related to Figure 3)

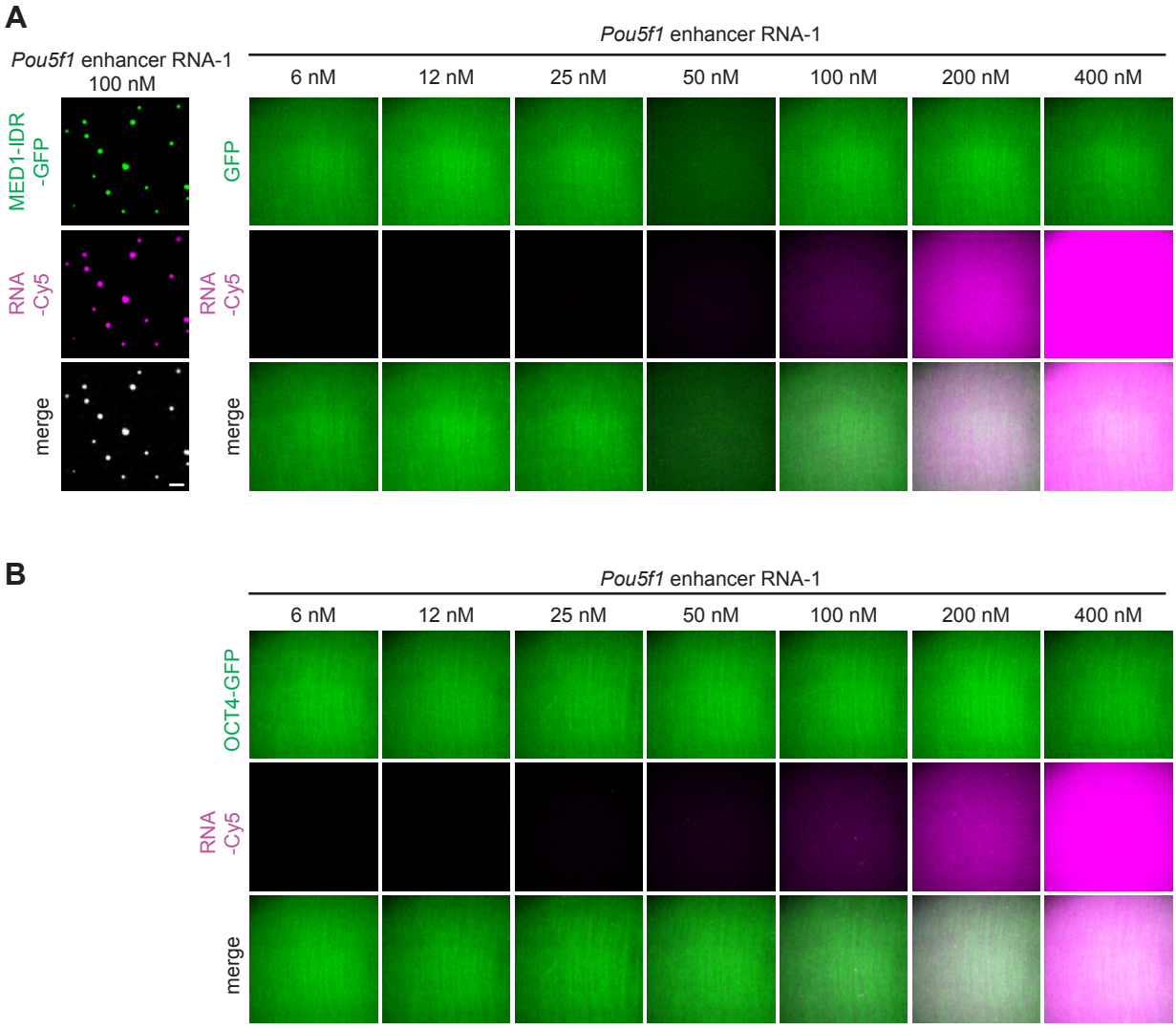


Figure S4 | Modulation of charge balance contributes to stimulation and dissolution of MED1-IDR condensates by RNA (Related to Figure 3)

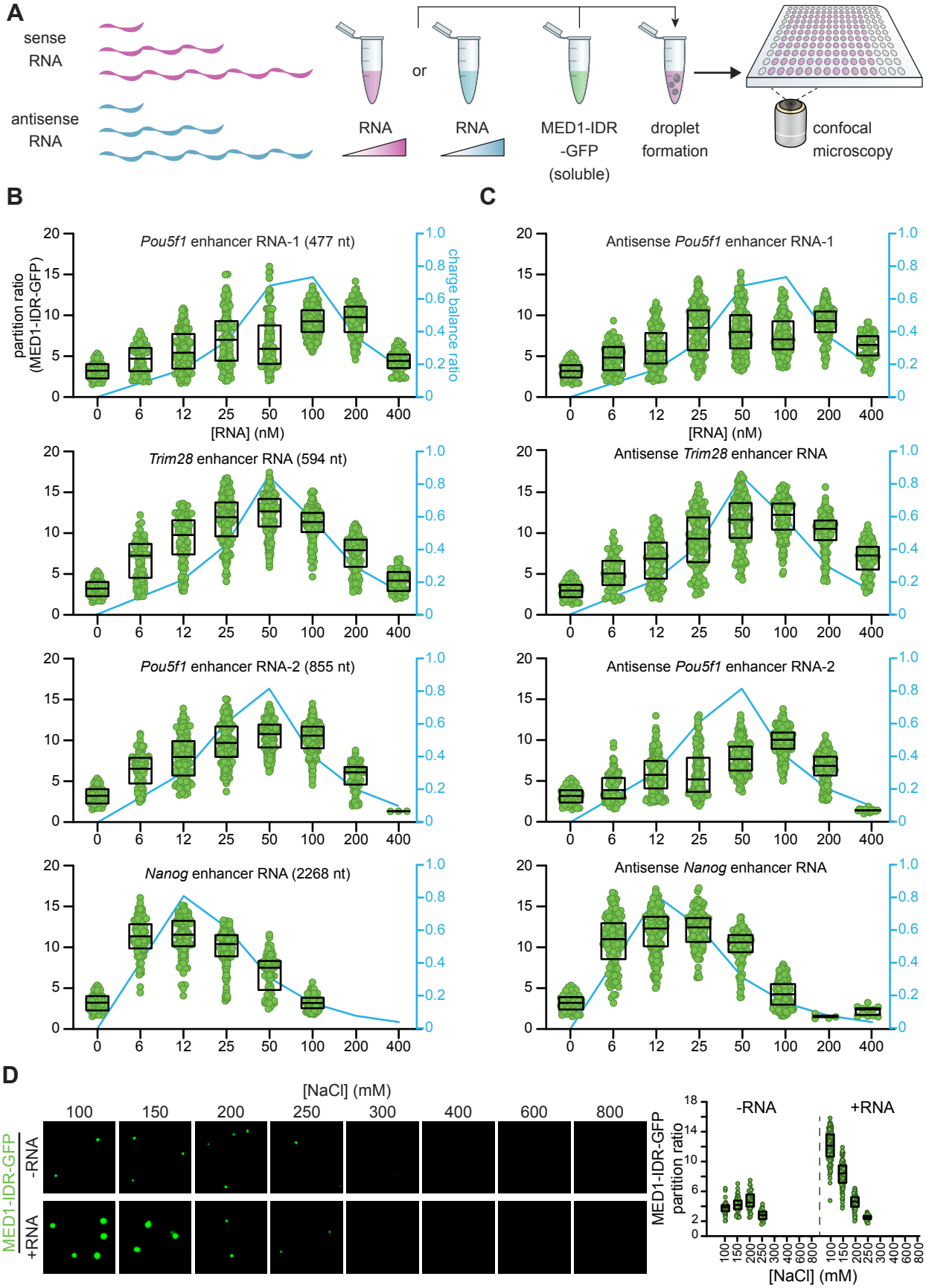
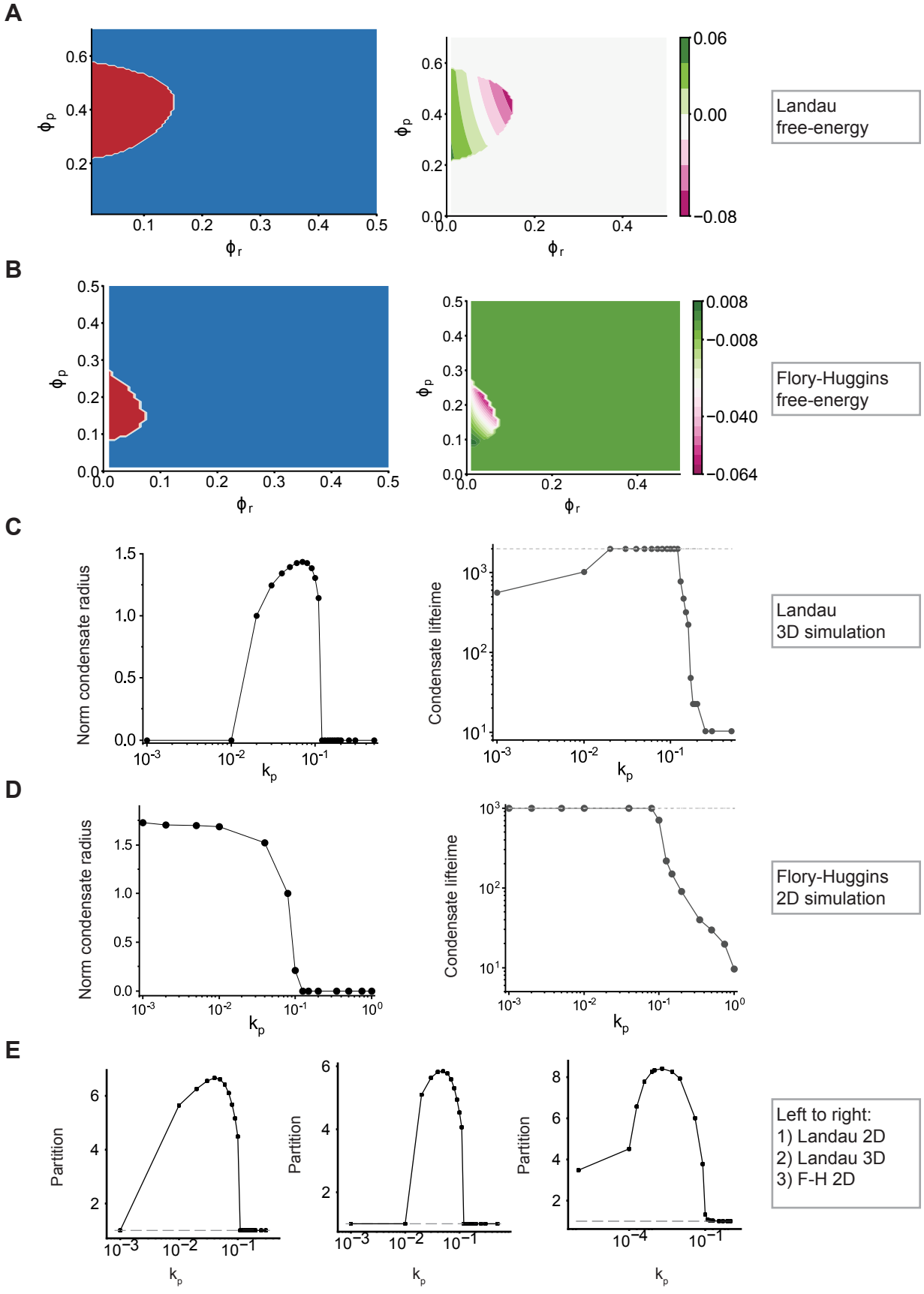


Figure S5 | Computational model for non-equilibrium RNA feedback on transcriptional condensates (Related to Figure 5)



Chapter 4: Appendix

Supplementary Information Text

Mathematical definition of the cooperation of gene products. We use $E(r)$ to denote the free energy of interaction between a task and a gene product. We use $1, \ldots, k$ to denote the gene products, and $r_1, \ldots, r_k$ to denote the distance between these gene products and a given task. This task is performed specifically by the cooperation of these gene products if $\sum_{i=1}^{k} E(r_i) \geq E(\varepsilon_1)$ and if for each gene product i there exists a gene product j such that $\left|\vec{r}_i - \vec{r}_j\right| < \varepsilon_3$. The definition of ε_1 and ε_3 can be found in Fig. 1A.

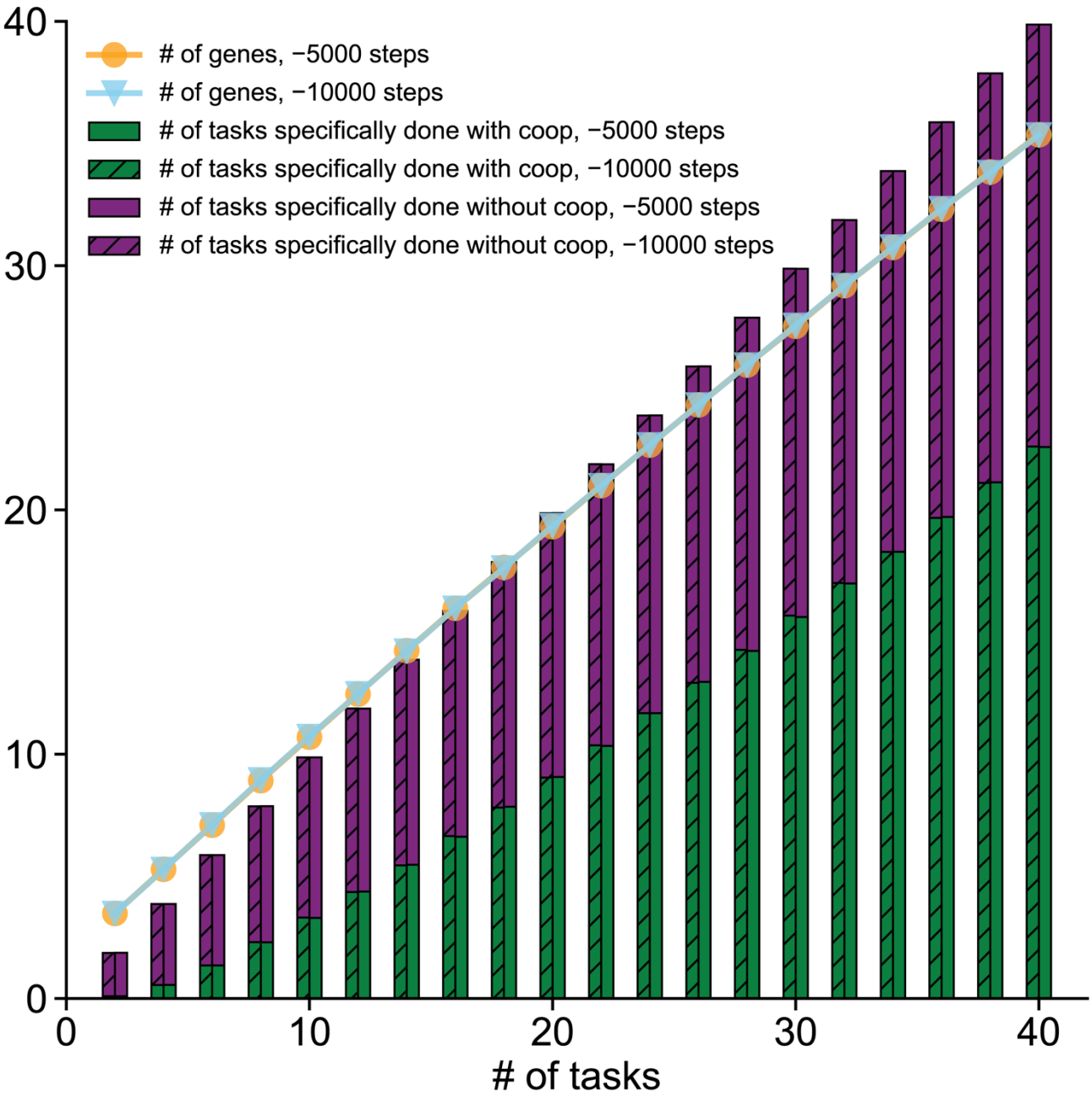


Fig. S1. Steady state is reached before introducing new tasks. Key variables (the number of genes and the number of tasks done with/without weak cooperation of gene products) of the reference system are measured at 5000 steps and at 10000 steps before each new task is introduced, and the results are almost exactly the same, which indicates that the steady state is reached before new task is introduced. The reference system is the one we discussed in the main text, where $\varepsilon_2 = 5\varepsilon_1 = 5\varepsilon_3,\ \lambda_1 = \lambda_2 = 1, \lambda_3 = 0.1$, the new task is located at distance $1.8\varepsilon_2$ away from one of the previous tasks, and three characteristics are used to describe the interaction characteristics between tasks and gene products.

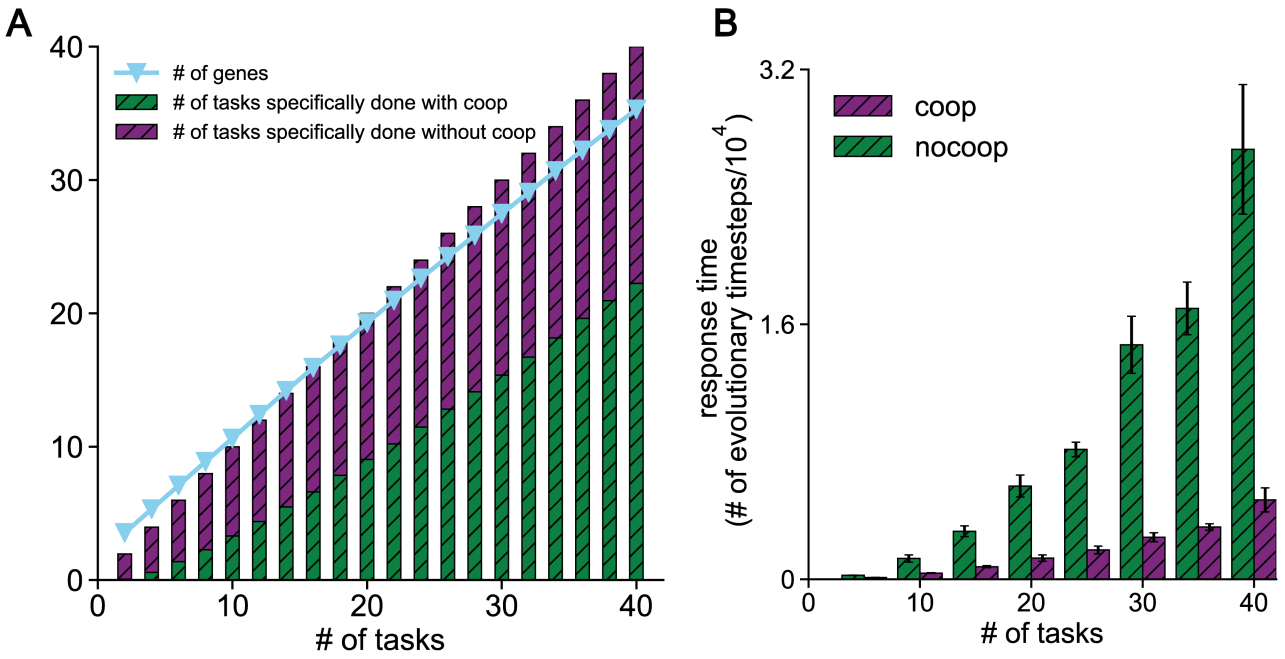


Fig. S2. Simulation results of the system where λ_1 and λ_2 is changed to 10. The other parameters of this system are kept the same as those of the "reference system". The reference system is the one we discussed in the main text, where $\varepsilon_2 = 5\varepsilon_1 = 5\varepsilon_3,\ \lambda_1 = \lambda_2 = 1, \lambda_3 = 0.1$, the new task is located at distance $1.8\varepsilon_2$ away from one of the previous tasks, and three characteristics are used to describe the interaction characteristics between tasks and gene products.
(A) Variation of the number of genes and the number of tasks specifically done with/without weak cooperation of gene products as the number of tasks required for an organism to function properly increases.
(B) The response time for the organisms to evolve to function properly after a new task is introduced is shown as a function of the number of tasks (or complexity). Results are shown for both the system where λ_1 and λ_2 is changed to 10 and its corresponding "nocoop system" wherein cooperative interactions between gene products are not allowed.

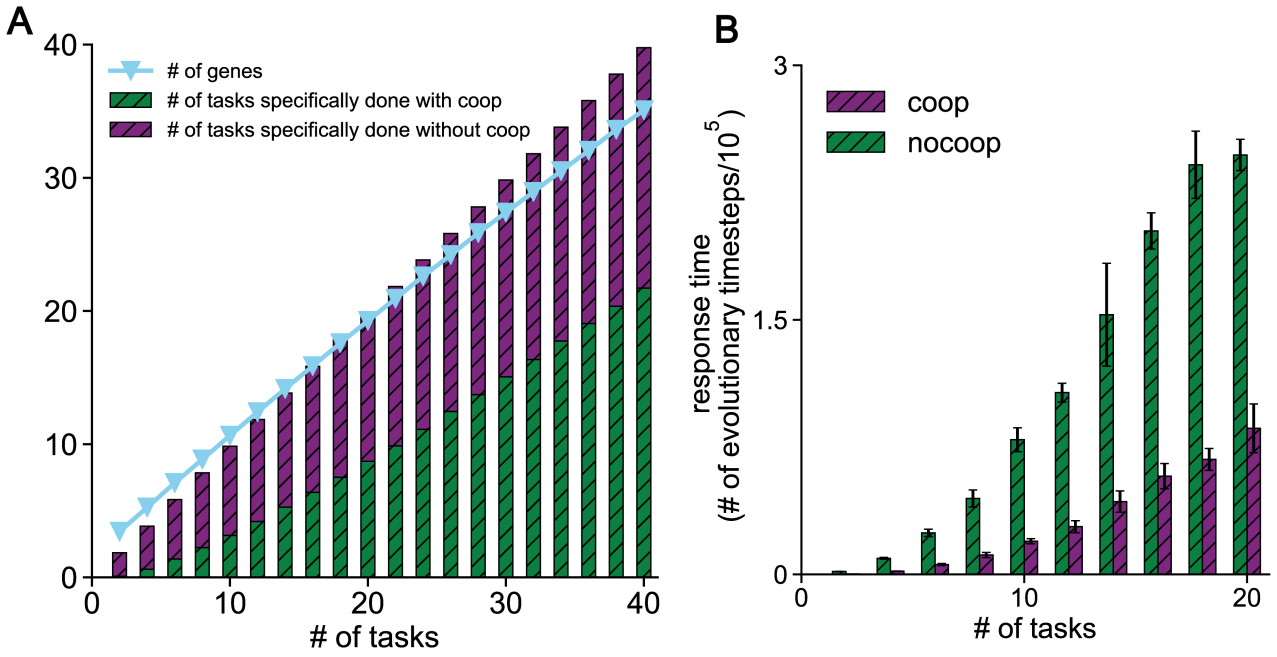


Fig. S3. Simulation results of the system where λ_1 is changed to 0. The other parameters of this system are kept the same as those of the reference system.
(A) Variation of the number of genes and the number of tasks specifically done with/without weak cooperation of gene products as the number of tasks required for an organism to function properly increases.
(B) The response time for the organisms to evolve to function properly after a new task is introduced is shown as a function of the number of tasks (or complexity). Results are shown for both the system where λ_1 is changed to 0 and its corresponding "nocoop system" wherein cooperative interactions between gene products are not allowed. The response time of this "$\lambda_1 = 0$" system is much larger than the response time of the reference system. For example, the response time for this "$\lambda_1 = 0$" system to complete the 20^{th} task is 86000, while the response time for the reference system to complete the 20^{th} task is 4900.

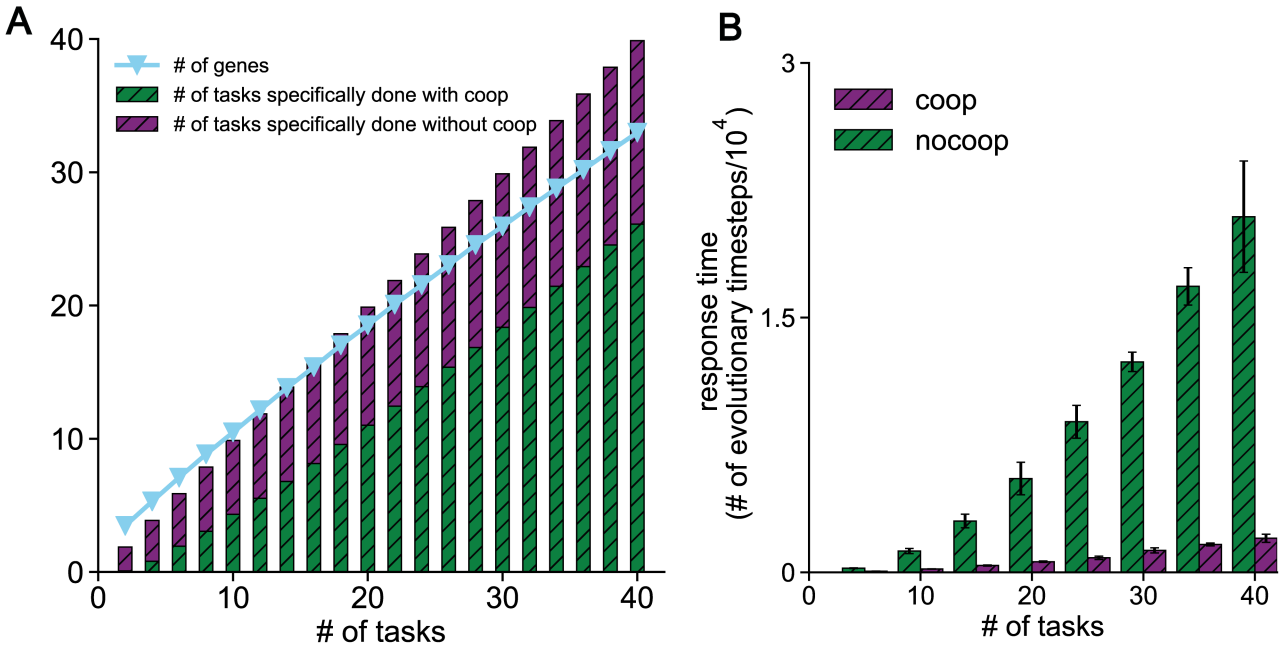


Fig. S4: Simulation results of the system where correlation between tasks are increased (the new task is located at distance $1.5\varepsilon_2$ away from one of the previous tasks.). The other parameters of this system are the same as those of the reference system.
(A) Variation of the number of genes and the number of tasks specifically done with/without weak cooperation of gene products as the number of tasks required for an organism to function properly increases.
(B) The response time for the organisms to evolve to function properly after a new task is introduced is shown as a function of the number of tasks (or complexity). Results are shown for both the system which has increased correlation between tasks and its corresponding "nocoop system" wherein cooperative interactions between gene products are not allowed.

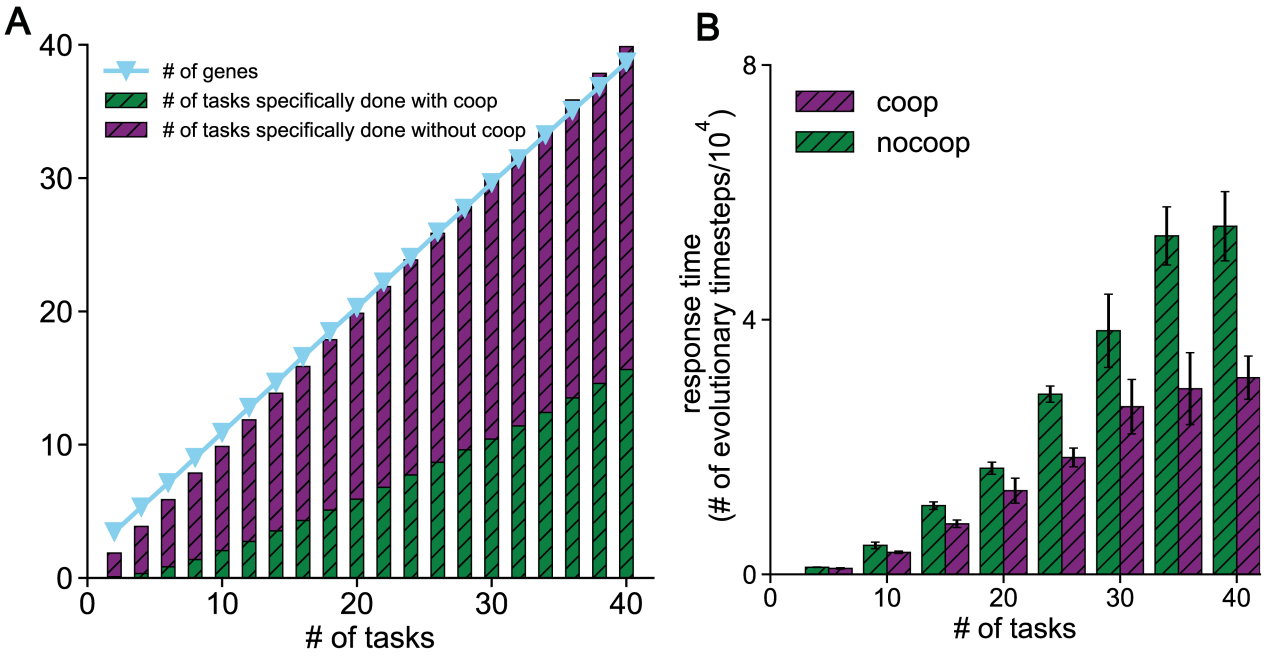


Fig. S5 Simulation results of the system where correlation between tasks are decreased (the new task is located at distance $2.5\varepsilon_2$ away from one of the previous tasks.). The other parameters are the same as those of the reference system.
(A) Variation of the number of genes and the number of tasks specifically done with/without weak cooperation of gene products as the number of tasks required for an organism to function properly increases.
(B) The response time for the organisms to evolve to function properly after a new task is introduced is shown as a function of the number of tasks (or complexity). Results are shown for both the system which has decreased correlation between tasks and its corresponding "nocoop system" wherein cooperative interactions between gene products are not allowed.

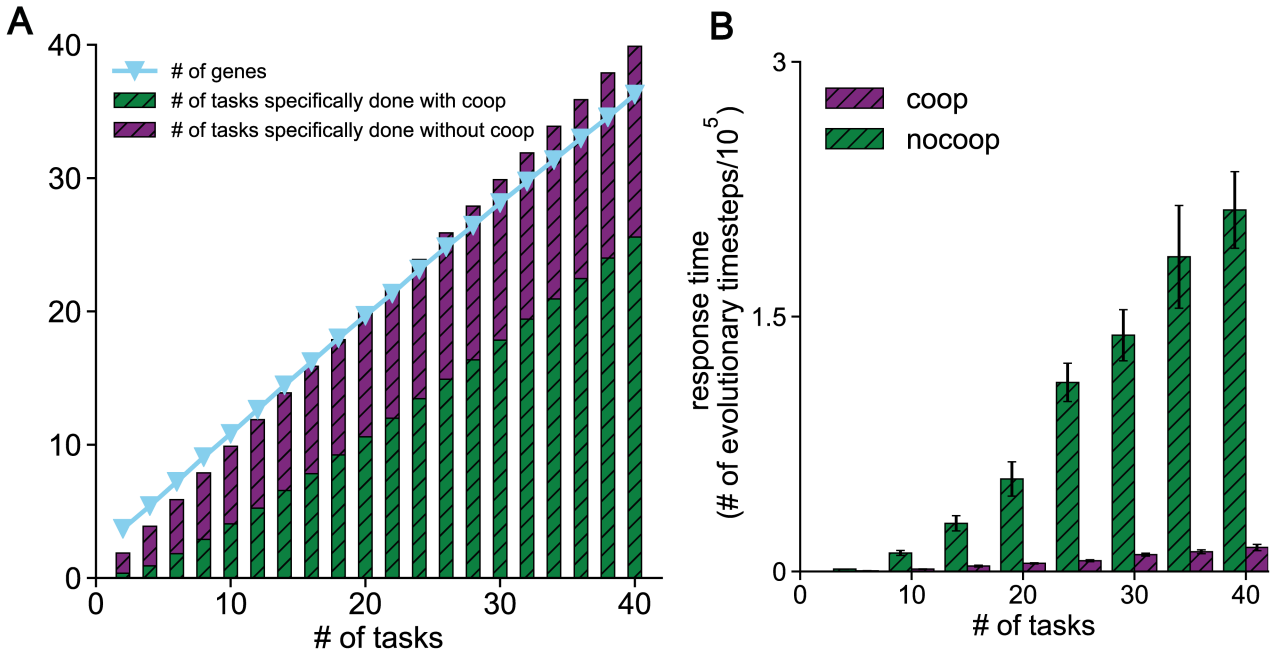


Fig. S6. Simulation results of the system where ε_1 is changed to $\frac{5}{8}$ that of the reference system. The other parameters of this system are the same as those of the reference system.
(A) Variation of the number of genes and the number of tasks specifically done with/without weak cooperation of gene products as the number of tasks required for an organism to function properly increases.
(B) The response time for the organisms to evolve to function properly after a new task is introduced is shown as a function of the number of tasks (or complexity). Results are shown for both the system where ε_1 is changed to $\frac{5}{8}$ that of the reference system and its corresponding "nocoop system" wherein cooperative interactions between gene products are not allowed.

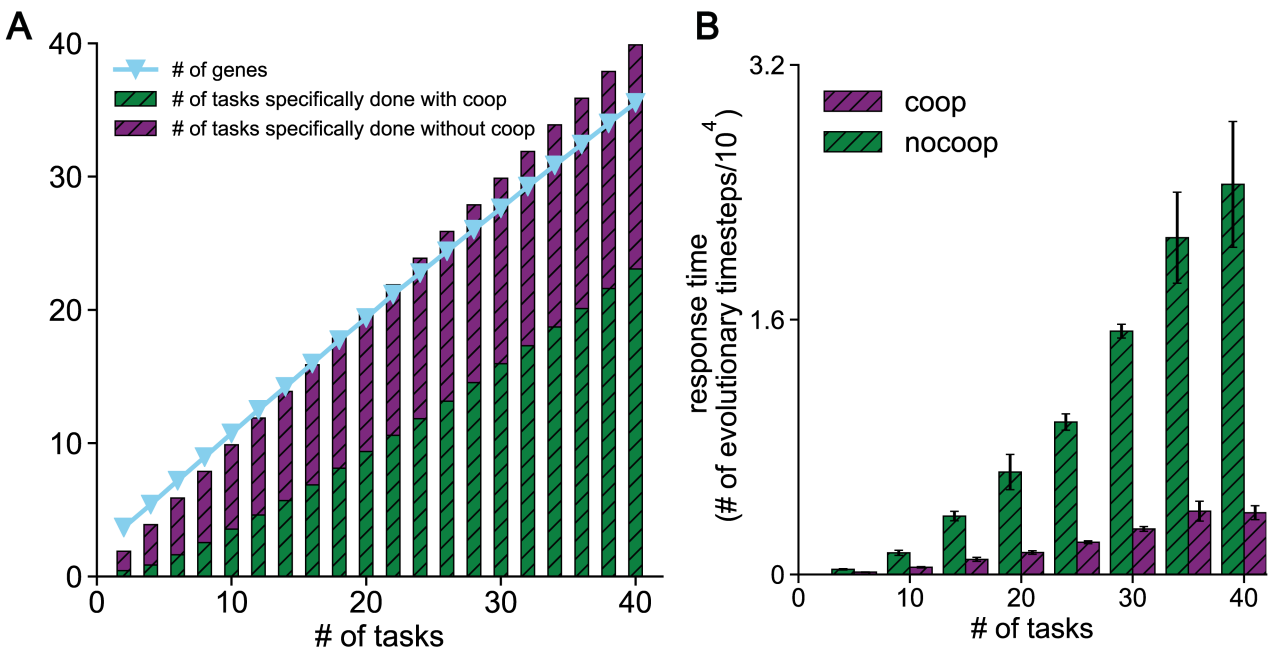


Fig. S7: Simulation results of the system where ε_3 is changed to twice that of the reference system. The other parameters of this system are the same as those of the reference system.
(A) Variation of the number of genes and the number of tasks specifically done with/without weak cooperation of gene products as the number of tasks required for an organism to function properly increases.
(B) The response time for the organisms to evolve to function properly after a new task is introduced is shown as a function of the number of tasks (or complexity). Results are shown for both the system where ε_3 is changed to twice that of the reference system and its corresponding "nocoop system" wherein cooperative interactions between gene products are not allowed.

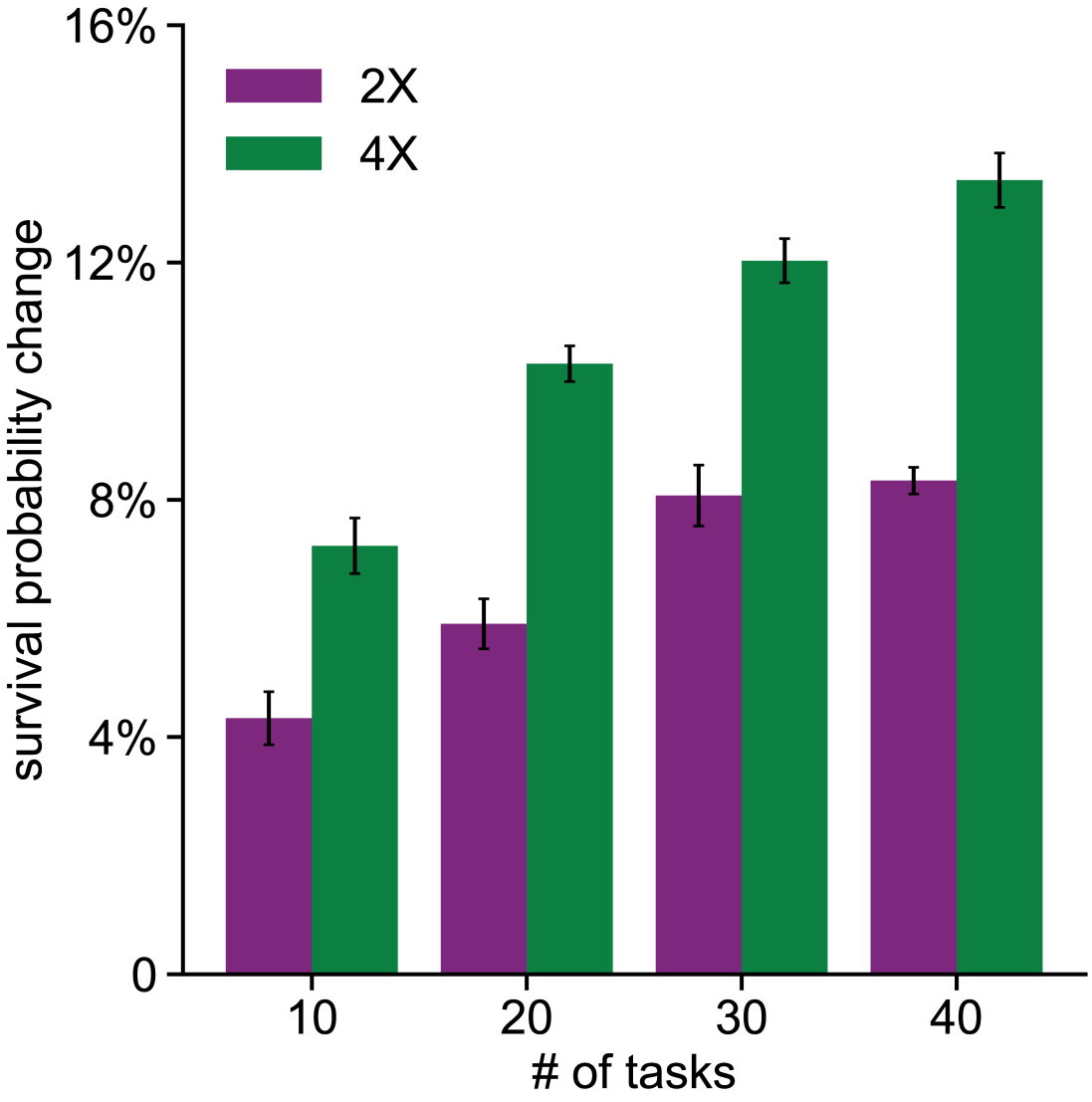


Fig. S8. The survival probability of mutated organisms is increased when ε_3 is increased. The survival probability of mutated organisms is the probability that the mutated organisms will be present in the next time step of evolution. The change of survival probability when ε_3 becomes twice that of the reference system as a function of the number of tasks that the organisms need to perform is shown as the purple bars. The change of survival probability when ε_3 becomes four times that of the reference system as a function of the number of tasks that the organisms need to perform is shown as the green bars. Further simulation results show that the survival probability of the mutated organism becomes larger than that of the corresponding "nocoop system" when ε_3 is more than twice larger than that of the reference system (i.e., when ε_3 is more than twice larger than the stepsize of the gene mutation).

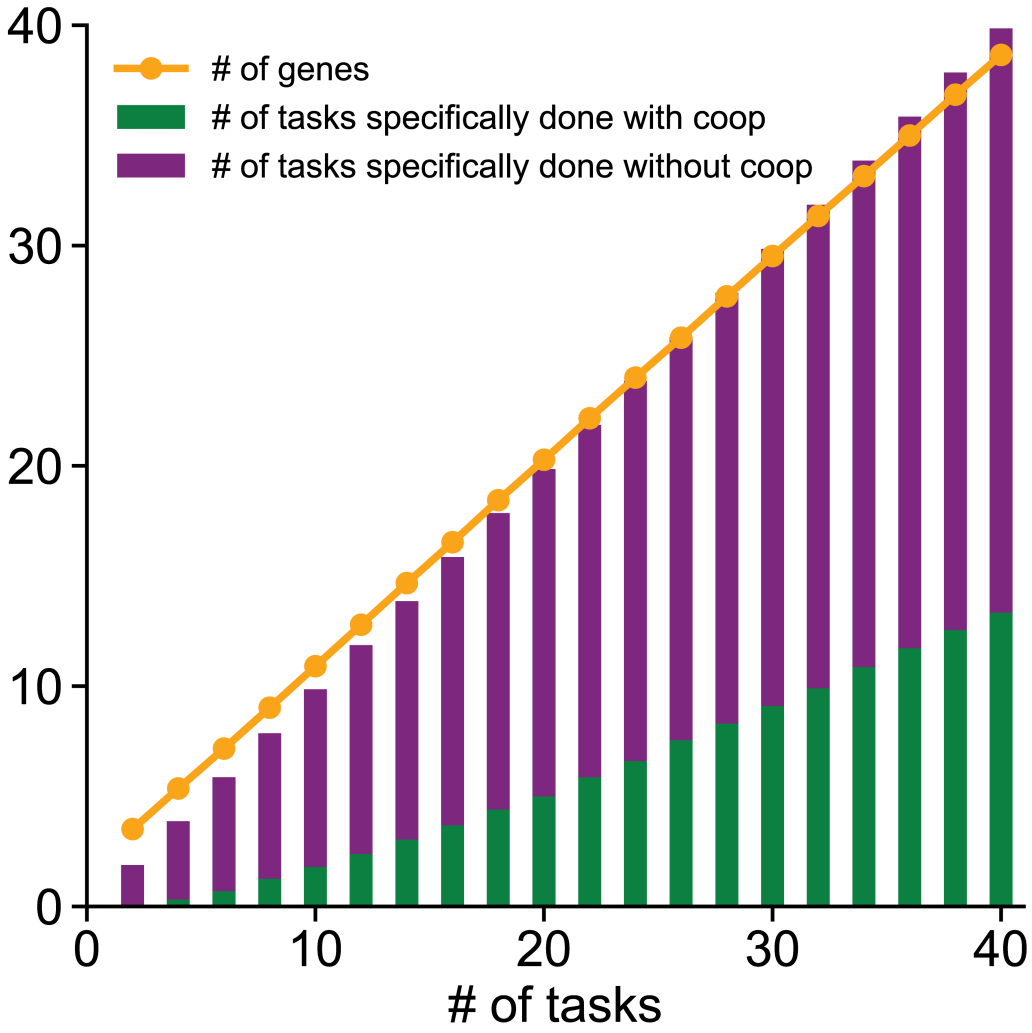


Fig. S9. The qualitative results discussed in the main text are robust to changing the number of characteristics required to describe the interaction from the value of three used to illustrate the results in the main text. Here, four characteristics are used to represent the characteristic space. ε_1 is changed to $0.3\varepsilon_2$ to make sure the ratio between the volume of the strong interaction regime (a 4D sphere of radius ε_1) and the volume of the weak interaction regime (a 4D sphere of radius ε_2) is the same as that of the reference system (the reference system is described by three characteristics). New tasks are introduced at a randomly chosen location that is $1.8\varepsilon_2$ away from any task from an earlier era, as was done for the reference system. The higher dimensionality of the characteristic space makes the new tasks less related to past ones compared to when three characteristics describe characteristics space. Therefore, weak cooperative interactions (WCI) evolve at a higher number of tasks. For example, when the number of tasks equals 40, 33% of tasks are done via WCI as shown in this figure, while this proportion is 56% for the reference system.

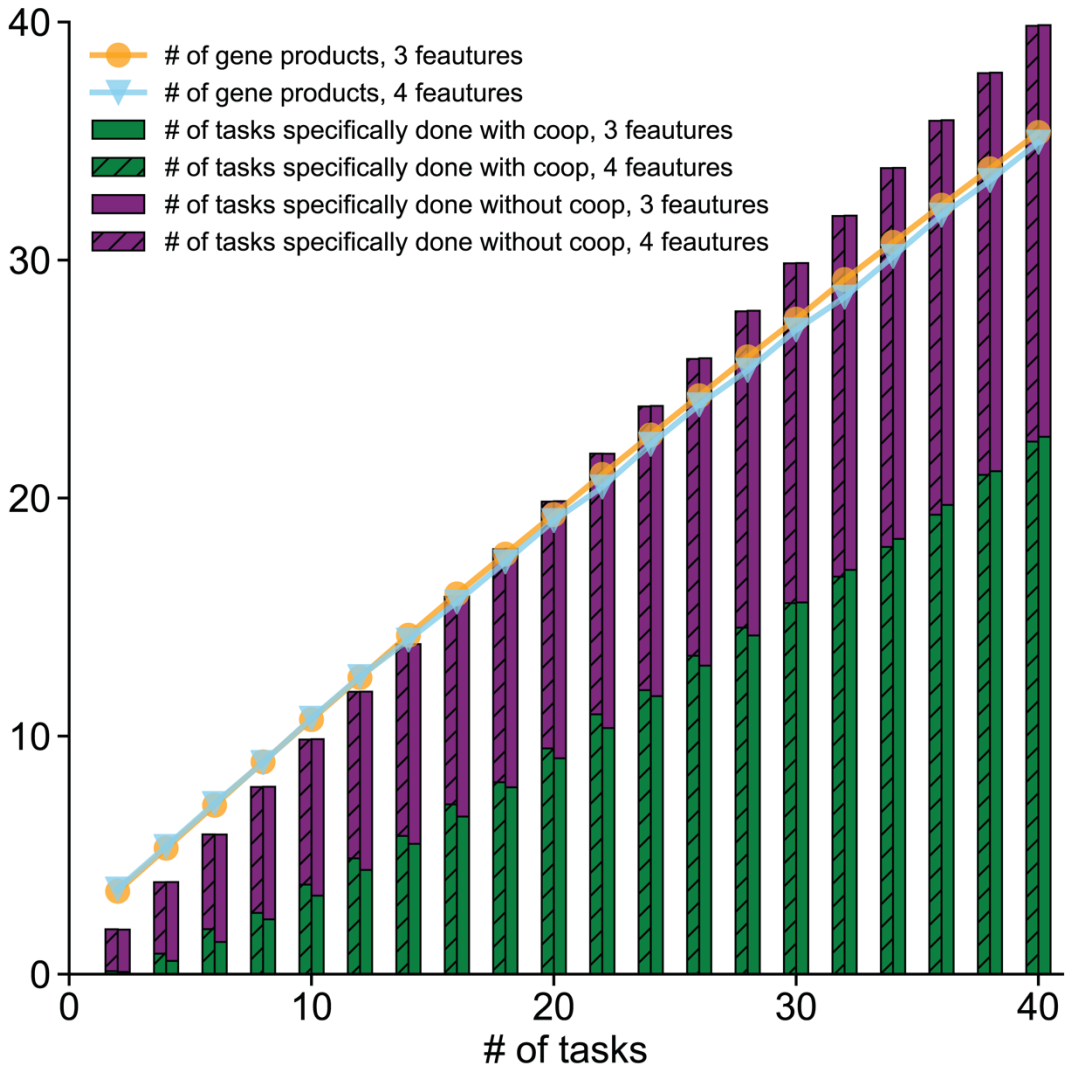


Fig. S10. Four characteristics are used to describe the characteristics space. New tasks are introduced at a closer distance ($1.4\varepsilon_2$) from a task from an earlier period (compared to when three characteristics suffice). The other parameters are the same as those used in Fig. S9. WCI evolve at a similar number of tasks compared to the reference system which is described by three characteristics.

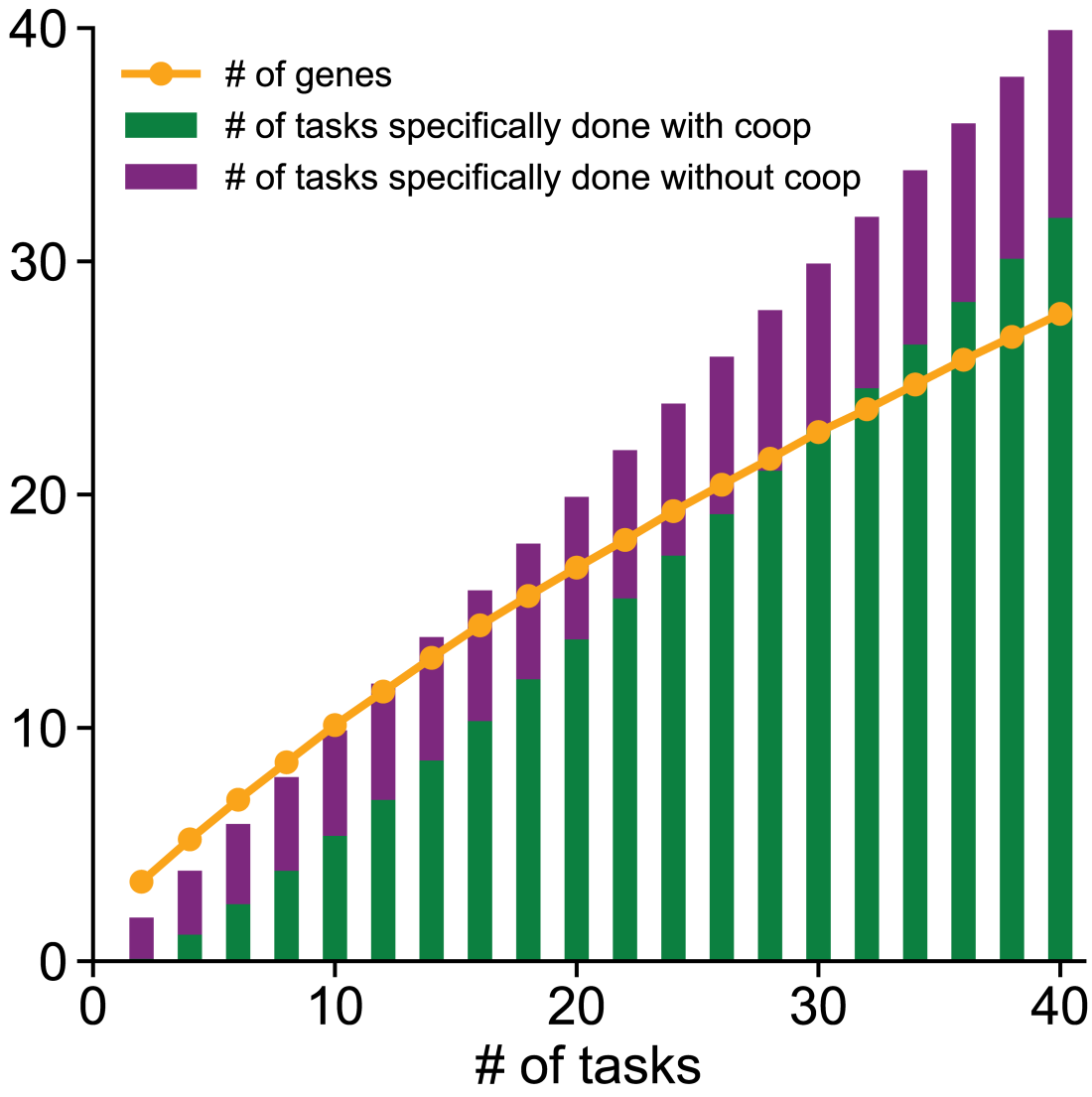


Fig. S11. The qualitative results discussed in the main text are robust to changing the number of characteristics required to describe the interaction from the value of three used to illustrate the results in the main text. Here, two characteristics are used to represent the characteristic space. ε_1 is changed to $0.09\varepsilon_2$ to make sure the ratio between the volume of the strong interaction regime (a 2D sphere of radius ε_1) and the volume of the weak interaction regime (a 2D sphere of radius ε_2) is the same as that of the reference system (the reference system is described by three characteristics). New tasks are introduced at a randomly chosen location that is $1.8\varepsilon_2$ away from any task from an earlier era, as was done for the reference system. The lower dimensionality of the characteristic space makes the new tasks more related to past ones compared to when three characteristics describe characteristics space. Therefore, WCI emerge at a smaller number of tasks. For example, when the number of tasks equals 40, 80% of tasks are done via WCI as shown in this figure, while this proportion is 56% for the reference system.